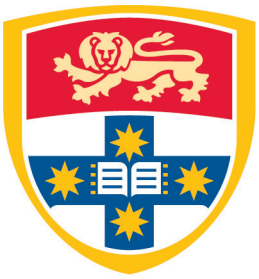
THE UNIVERSITY OF
SYDNEY

AUTOMATIC POPULATION OF STRUCTURED REPORTS FROM NARRATIVE PATHOLOGY REPORTS

A thesis submitted in fulfilment of the requirements for the degree of Doctor of Philosophy in the School of Information Technologies at The University of Sydney



THE UNIVERSITY OF SYDNEY

Ying Ou

2015

Abstract

Pathology reports provide vital information for the clinical management of cancer patients, allowing accurate diagnosis, staging and determination of treatment and prognosis. However, there are several issues resulting from traditional narrative reports compared to structured reports. For example, essential elements are occasionally omitted, especially negative results, which are not always reported clearly. As well, the referring doctors often find it difficult to identify the necessary elements in a free-text pathology report to justify a given diagnosis. There are a number of advantages for the use of structured pathology reports: they can ensure the accuracy and completeness of pathology reporting; it is easier for the referring doctors to glean pertinent information from them, thus improving the communication between pathologists and clinicians. Furthermore, they also facilitate efficient extraction of information for cancer registries, data collection and research purposes.

The goal of this thesis is to extract pertinent information from free-text pathology reports and automatically populate structured reports for three cancer diseases, namely melanoma, colorectal cancer, lymphoma and identify the commonalities and differences in processing principles to obtain maximum accuracy.

Unlike previous works that regard the task as automatic structuring of sentences of interest in narrative medical reports, this study aims to populate certain fields in structured reports based on the global view of the entire document. This is challenging, as it requires either inference from the entities or combination of various entities as well. The fields predefined in structured templates were determined mainly by utilizing three structured cancer reporting protocols from the Australia and the Royal College of Pathologists of Australia as well as advice from clinicians and pathologists.

A detailed corpus analysis was conducted on a set of pathology notes, with the objectives of identifying lexical and linguistic characteristics in the narratives, and the difficulties or challenges that may be encountered when processing these texts. Assessment of the level of completeness of original reports, and proposals for appropriate strategies for the establishment of structured templates were subsequently completed.

Three pathology corpora were annotated with entities and relationships between the entities in this study, namely the melanoma corpus, the colorectal cancer corpus and the lymphoma corpus. Detailed annotation schemas and guidelines were developed in an iterative process to ensure annotation consistency.

A supervised machine-learning based-approach was developed to recognise medical entities from the corpora. Specifically, the medical entity recognition system used conditional random fields (CRF) learners. The CRF-based models were able to capture a significant portion of the entity boundaries by

using contextual information. The application of rich feature sets provided useful clues for the classification of entity types. By feature engineering, the best feature configurations were attained, which boosted the F-scores significantly from 4.2% to 6.8% in 10-fold cross-validation experiments on the training sets. Several common effective features across the three corpora were identified, which can be beneficial for other medical entity recognition tasks.

Without proper negation and uncertainty detection, final outputs for several fields in the structured templates will be affected, and consequently the quality of the structured reports will be diminished. The negation and uncertainty detection modules were built to handle this problem. The modules obtained very good performance (with over 99% overall F-scores) on the training sets, which dropped on the test sets (where overall F-scores decreased to 76.6% – 91.0%).

A relation extraction system was presented to extract four relations from the lymphoma corpus. A rule-based approach was applied to classify Spatial Specialization relation, while a supervised machine learning-based approach was adopted to identify Result-Positive, Result- Negative and Result-Equivocal relations. Simple heuristic rules were applied in the rule-based module, while several useful features were prepared for the support vector machines (SVM) classifier. The system achieved very good performance on the training set, with 100% F-score obtained by the rule-based module and 97.2% micro-averaged F-score attained by the SVM classifier.

Predefined templates were designed based on a thorough review of the structured reporting protocols and analysis of the training corpora. Rule-based approaches were used to generate the structured outputs and populate them to the templates. The rule-based system attained over 97% F-scores on the training sets. A pipeline system was implemented with an assembly of all the components described above. It achieved promising results in the end-to-end evaluations, with 86.5%, 84.2% and 78.9% micro-averaged F-scores on the melanoma, colorectal cancer and lymphoma test sets respectively.

The pipeline system can be applied to cancer registries, clinical audits and epidemiology research. With further improvement, it can also significantly improve the quality of pathology reporting in the clinical setting.

Acknowledgements

First and foremost, I would like to sincerely thank my supervisor Prof Jon Patrick, who introduced me to the field of Natural Language Processing and afforded me valuable experiences during the past few years. He provided me with the opportunity to carry out this study and cultivate my research skills independently. His rich knowledge and logical thinking always inspired me. His support, encouragement and advice were of great significance to my study and this thesis.

I would like to express my sincere gratitude to Dr Stephen Crawshaw for his intelligence and patience. He not only provided medical knowledge for me to tackle clinical documents, but also spared no effort on various annotations, helped me when I had difficulty processing the documents, and proofread the draft of this thesis.

I would like to warmly thank other (or former) members in the Health Information Technologies Research Laboratory: Drs Yefeng Wang, Min Li, Nguyen Doang and Pooyan Asgari, for their great support on solving many technical issues during my study.

I would also like to thank the computational linguists in the annotation team and pathologists from the Royal Prince Alfred Hospital and Royal College of Pathologists of Australasia for giving me their assistances to obtain, annotate and understand the data.

Last but not least, I would like to give my special thanks to my family, especially my mother, Jiaying Liang. Without their understanding and encouragement, it would be impossible for me to finish this work.

Table of Contents

List of Tables

List of Figures

Chapter 1 Introduction

1.1 Structured Reporting

The treatment of cancer is often made based on input from a multidisciplinary team including surgeons, oncologists, radiologists, and pathologists. In particular, a pathology report on a cancer specimen can provide critical information related to diagnosis and prognosis, which should be accurate and complete. If a pathologist is not guided by a cancer-specific standardized template, but rather writes a non-organised prose report, it is easy to omit information that may be required for clinical decision making. The purpose of this research is to investigate models of how prose pathology reports might be automatically converted into structured reports and to build a technology that demonstrates the feasibility and accuracy at which the task can be performed.

Traditional narrative reports have significant variability since different pathologists use a multitude of different reporting styles to describe their findings. Such variability often results in missing important clinically relevant data elements such as margins, lymphatic invasion etc. Research indicate that the presence of both perineural invasion (Griffantibartoli et al., 1994; Nagakawa et al., 1993) and lymphovascular invasion (Tannapfel et al., 1992) which are poor prognostic indicators, were not reported in 16% and 34% respectively in free-text reports (Gill et al., 2009). As well, study points out that traditional pathology reporting underestimates the rate of margin positivity in pancreatic carcinoma (Verbeke et al., 2006). As a result, clinicians cannot make a proper management plan for an individual case, as they usually rely on the pathology reports to diagnose the patient. However, with the introduction of structured reporting, these issuses can be mitigated to a great extent. The completeness of information presented in pathology reports can be improved significantly. The structured format has been proven to result in more complete reports for patients with melanoma and breast cancer (Scolyer et al., 2004; Harvey et al., 2005). Moreover, the free-text reports affect the efficiency of clinical decision making, as clinicians sometimes find it difficult to pull out the relevant information from a long paragraph of continuous text, and they often have to search through paragraphs of information in various sections in order to find the key elements to manage clinical care (Srigley et al., 2009).

A great number of studies have emphasized the importance of structured reporting (also called synoptic reporting in some cases) for diverse cancers, including melanoma (Haydu et al., 2010; Karim et al., 2008), colorectal cancer (Beattie et al., 2003; Chan et al., 2008; Chapuis et al., 2007), haematologic and lymphoid neoplasms (Mohanty et al., 2007), etc.

Structured reports can ensure the accuracy and completeness of pathology reports. A structured report can ensure the pathologists avoid omission of all relevant data and necessary details, especially trainees and new pathologists. A survey on the reporting of colorectal cancer specimens in a tertiary

care pathology department (Chan et al., 2008) found that before synoptic reporting, macroscopic features such as the presence or absence of serosal involvement, and distance to the radial margin, microscopic parameters such as radial resection margin status, extramural venous invasion and host inflammatory response were underreported. Such features showed significant improvement after synoptic reporting (ranging from 50% to 80%).

It is easier for the referring doctors to glean pertinent information from structured reports, thus it is more user-friendly for both clinicians and pathologists since it can improve communication between them and thus reduce requests for repeat pathology and call-back for explanations from clinicians.

Furthermore, structured reports can also facilitate efficient extraction of information for cancer registries, data collection and research purposes. A lot of information regarding TNM and stage grouping is embedded in the pathology reports on resection specimens (Hammond and Henson, 1995). The information is critical for cancer surveillance systems and it is also used by cancer registrars, cancer agencies, and epidemiology researchers. Structured reports are much more readily usable than traditional narrative reports, since the information is populated in the reports or easy to be inferred from the reports. Structured pathology reports can also be used in quality improvement. One of the examples is the information about the retrieval of lymph nodes from colorectal cancer resection specimens (Wright et al., 2004). Lack of sufficient sampling nodes will lead to under-staging and consequently over-utilization of adjuvant chemotherapy. The total count of lymph nodes retrieved and the number involved by metastatic tumour can be derived with ease from a report in structured format. Likewise, clinical audits for tumour stage or margin positivity in resection specimens can also be facilitated by structured reporting.

Currently, there are many researchers focused on developing synoptic or structured reporting tools for pathologists to edit and standardize their reports, to simplify the process of routine reporting of pathology. For example, Qu et al. introduced a template-based tumour reporting system that minimized exhaustive list checking and extensive text editing to reduce reporting errors and improve work efficiency (Qu et al., 2007). However, these methods also have some potential disadvantages. Pathologists may be reluctant to change reporting practice with concerns that such tools may lack space for sufficient observations and flexibility for recording of microscopic details (Dworak, 1992; Nochomovitz, 1998). Some may regard this as a relatively cumbersome and time-consuming process, which requires additional steps to enter or edit the contents compared to traditional reporting formats (Mohanty et al., 2007). Besides, such tools cannot handle the free-text reports that have already been written or dictated, as, there is still much useful information in the reports which can be reused with other feasible and efficient approaches.

To avoid these issues and maintain the benefits of structured reporting, natural language processing (NLP) is one promising approach to extract critical findings and incorporate them into a predefined structured template, thereby achieving the goal of automatic population of structured reports.

1.2 Information Extraction

Generally, information extraction (IE) is a sub-discipline of NLP, which focuses on the identification of the specific facts and relations within unstructured texts, the extraction of the relevant values, and their transformation into standardized codes and/or structured information. The goal of IE is to extract significant information from unstructured data sources and transform it into structured data to facilitate access and retrieval of information.

Previous works on automatic structuring of free-text medical reports have attained some successes by classifying the relationships (e.g., dependencies) among medical entities with statistical methods at sentence level. In other words, they regard the task as automatic structuring of sentences of interest in the free texts. For example, useful information can be output in a frame from a radiology natural language processor with this method (Taira et al., 2001). An example is presented in Table 1.1:

Input: "A mass is seen in the right lower lobe that measures 5 cm in maximum diameter and is unchanged from the previous examination."		
has location	Value	right lower lobe
	Relation	in
has size	Value	5cm
	Dimension	maximum diameter
has size trend	Value	unchanged
	Reference event	previous examination

Table 1.1 Output knowledge frame from a radiology natural language processor.

However, such approaches would not be a best fit for this study, since the population of fields in a structured report should be based on the full view of the whole document rather than each sentence; besides, in many cases, it also requires inference from the entities or combination of various entities to instantiate a knowledge representation model. For instance, by using such approaches, the maximum dimension measurement (4.5cm) cannot be directly extracted from this example "Measurements: Length 4.5cm, width 4cm, depth 0.6cm" but rather only inferred by an added processing system. It suggests that more complex methods implemented in an IE system are preferable for this study.

From the author's point of view, a desirable clinical IE system for automatic structured reporting should consist of three major processes, which are medical entity recognition (MER), relation extraction (RE), and structured representation (SR).

Named entity recognition (NER) aims at identifying specific words or phrases ("entities") and categorizing them. In the general English domain, NER focuses on person, location, and organization; in the biomedical domain, genes and proteins are its main themes. In the clinical narrative, however, the semantic types for it are likely to be medical entities (e.g., disorders, signs or symptoms, anatomical sites, medications, and procedures). Moreover, in a particular sub-domain, the semantic

types can be problem-specific. For example, the Third i2b2 Workshop on Natural Language Processing Challenges for Clinical Records focused on the identification of medications in discharge summaries, including dosages, modes of administration, frequencies, durations, and reasons for administration (Uzuner et al., 2010). For this reason, although medical entity recognition (MER) is derived from NER, it can be more complicated than NER.

As in NER, several issues can make MER challenging, such as word/phrase order variation (varying order of words or phrases appears in instances), derivation (suffixes transform one part of speech to another), inflection (changes in number, tense, comparative/superlative forms), synonyms, homographs and abbreviations. Table 1.2 illustrates some examples of these issues collected from narrative pathology reports.

Relationship extraction (RE) focuses on determining relationships between entities or events. By extracting relations among the entities involved and how these entities are described, a clearer picture of the semantics is obtained. For instance, to answer questions like "what is the positive biomarker for the immunohistochemistry results on this patient?", not only "positive" and "CD20" should be extracted, but also the relation between these entities should be extracted from this sentence "CD20: most cells positive".

Issues	**Examples**
word/phrase order variation	Site of tumour: caecum *vs.* Tumour site: Caecum
derivation	caecum (noun) *vs.* caecal (adjective)
inflection	nodes (plural) *vs.* node (singular) identify (present tense) *vs.* identified (past tense) larger (comparative) *vs.* largest (superlative)
synonyms	right colon *vs.* ascending colon
homographs	MM has two expansions malignant melanoma *vs.* millimeter
abbreviations	ICV *vs.* ileocaecal valve

Table 1.2 Challenging issues and examples in a medical entity recognition task.

The relations embedded in a clinical document can be explicit as well as implicit, such as negation and uncertainty. Negation and uncertainty identification aims at inferring whether an entity is present or absent, and quantifying that inference's uncertainty. In fact, nearly half of all symptoms, diagnoses, and findings in clinical reports are estimated to be negative or uncertain (Chapman et al., 2001a). Take, for example, the clinical coding of medical concepts in a report, where the coding of a negative or uncertain finding or diagnosis may result in an over-coding financial penalty.

Without appropriate representation, the extracted entities and relations can be meaningless for the users. Structured representation involves construction of a predefined template and population of the template. The template is usually user-oriented, depends on the specific requirement of the users. For example, in the radiology domain, the radiologists often need to determine the size and the location of

a tumour. Thus there ought to be associated items (e.g., “tumour size”, “tumour location”) in the template.

1.3 The Obstacles to Information Extraction

Building such an IE system to fulfil this task is usually a slow and laborious process, because clinical reports often contain multiple sections, for instance, a typical pathology report may consist of these sections: Diagnosis, Macroscopic Description, Microscopic Description, and Comments, which often vary in narrative structure and uniformity; and there is also institutional or individual variation in reporting practices (e.g., the same clinical concept can be expressed in different ways).

Currently, there are known barriers to information extraction in the clinical domain.

On the one hand, it is difficult to access available clinical data for training and evaluation from many hospitals and clinics, due to concerns regarding patient privacy and revealing unfavourable institutional practices. As a result, annotated data are usually unique to a research group or laboratory that generated them for some tasks and that cannot be reused for other tasks without considerable translational effort.

On the other hand, clinical notes are more difficult to handle than newswire text, because it requires both linguistic expertise and understanding clinical domain knowledge. Moreover, in different sub-domains, there are different sub-languages used. For instance, abbreviations like “HMF (Hutchinson’s melanotic freckle)” and “SSM (superficial spreading melanoma)” appear frequently in melanoma pathology reports, but rarely in other pathology reports (e.g., colorectal cancer reports).

Moreover, unlike radiology reports, the information reported in pathology notes can differ from one disease to another. Thus a general template prepared for reports of all kinds of diseases seems inappropriate and cannot satisfy the requirements of the pathologists and clinicians. In the past, due to a lack of consensus and authorized materials as standards, it was very difficult to obtain such disease-specific templates. However, in recent years, the United Kingdom and the United States have defined processes for the development and review of structured reporting protocols. In line with these international developments the Royal College of Pathologists of Australasia (RCPA) is also implementing structured pathology reporting of cancer through its Cancer Services Advisory Committee to ensure that pathologists throughout Australasia have access to appropriate, nationally endorsed protocols (RCPA, 2013a). The RCPA structured cancer reporting protocols (RCPA, 2013b) served as the main sources for building standardized templates for this research.

1.4 Research Problems and Approaches

Since NLP techniques have been applied to different IE tasks and achieved state-of-art performance, this thesis addresses an issue as to whether these techniques can be utilised or modified to resolve automatic structured reporting from narrative notes.

The goal of this thesis is to extract pertinent information from free-text pathology reports and automatically populate structured reports. Specifically, the work focuses on medical entity recognition, negation and uncertainty detection, relation extraction, and structured output generation.

In this study, both rule-based approaches and machine learning methods are used, depending on the particular problem to be solved.

In order to utilize the machine learning approaches and evaluate system performance, it was necessary to acquire semantically annotated corpora for medical entity recognition, negation and uncertainty detection and relation extraction. Annotation schemas and annotation guidelines were also developed to ensure annotation consistency.

1.4.1 Corpus Analysis

A lexical analysis was completed on three pathology corpora of different cancers: melanoma, colorectal cancer and lymphoma.

The aim of the corpus analysis was to identify the characteristics and language phenomena of the pathology notes as well as identify necessary processing steps required to reduce noise in processing.

1.4.2 Corpus Annotation

Annotated corpora are prerequisites for IE systems with supervised machine learning methods. The process of annotation is mainly to add linguistic and semantic information to the raw texts. The annotation schema defines the types of information needed to be associated with the raw texts.

Three corpora were annotated in this study. Detailed annotation guidelines were developed in an iterative process to ensure annotation quality. The annotations were used as the training data to build a supervised machine learning system as well as gold-standards for evaluation on medical entity recognition, negation and uncertainty detection and relation extraction.

1.4.3 Medical Entity Recognition

Since supervised machine learning-based approaches have been widely adopted and achieved encouraging results in many tasks of MER (Patrick and Li, 2010; Patrick et al., 2011), they were also

adopted in this study. One of the advantages of using supervised machine learning-based approaches is that the entities can be classified into broader categories rather than restricted to a concept in a controlled terminology. Another advantage is that it can resolve lexical ambiguity to some extent.

1.4.4 Negation and Uncertainty Detection

Study of the training materials indicated that utilization of an existing algorithm without tuning could not attain satisfactory results in an IE system to detect negation in pathology reports. It was necessary to build a specific module to handle this problem.

There is comparatively smaller amount of research into solving uncertainty problems than negation detection in pathology reports, probably because it represents a reasoning process, which is usually hard to capture. However, it is a common language phenomenon in the reports. Resolving it can help users better understand the reports and direct their attention to the undetermined findings or diagnoses to make plans for further examinations or tests.

1.4.5 Relation Extraction

Relationship extraction is intended to find associations between medical entities. It may not be easy for a rule-based system to extract complex relations, but it is a suitable strategy when the training sample size is small. Machine learning methods are better at identifying the different forms of linguistic patterns to represent relationships between medical entities when there are sufficiently large training samples. Hence, given different training sample sizes, both rule-based and machine learning approaches have been adopted in this study.

1.4.6 Structured Output Generation

After a thorough review of the protocols and the training corpora, predefined templates were designed, and associated entities and relation types were mapped to the items in the templates.

The structured report components required either very large segments of text that would have been impossible to infer reliably, or inferences from the structure of a variety of medical entities to be properly constructed. Therefore, rule-based approaches were used to generate the structured output and populate the templates.

1.5 Contributions

As a result of a systematic study of IE problem in the clinical domain, one of the main contributions is the establishment of a methodology to make full use of NLP techniques to instantiate a knowledge representation model. Other primary contributions of this research can be concluded as follows:

- A comprehensive study of the language phenomena in narrative pathology notes.
- A detailed process to annotate pathology notes, as well as the annotation guidelines.
- Annotated corpora with both entities and relations suitable for training a supervised IE system.
- Ideas of how to achieve state-of -the-art performance on medical entity recognition, negation and uncertainty detection, relation extraction, and structured output generation.
- An implementation of a pipeline system which is able to achieve encouraging performance so that it can help pathologists to validate their reports and improve communication between pathologists and clinicians.

1.6 Thesis Organization

This thesis is organized in the order of the workflow in an automatic structured reporting system. Apart from chapters 2 and 9, each chapter presents one of the major components in the pipeline system, which is illustrated in Figure 1.1.

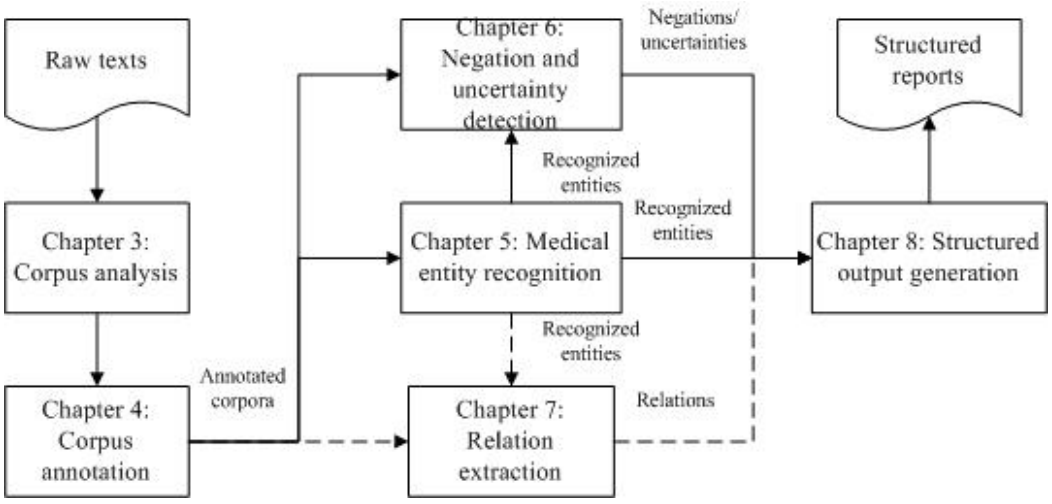


Figure 1.1 Organisation of the thesis. Relation extraction is performed on the lymphoma corpus.

Chapter 2 presents a survey of previous works on information extraction. The four major subtasks in information extraction: medical entity recognition, negation and uncertainty detection, relation extraction and structured output generation are reviewed.

Chapter 3 presents the detailed analysis of the training data in this study.

Chapter 4 describes a process for annotating pathology reports. Three semantically annotated corpora were developed, and annotation schemas and annotation guidelines are introduced.

In Chapter 5, a machine learning-based MER system is developed. Various features are explored, and integrated into the machine learner.

In Chapter 6, three different approaches including lexical-based approach, syntactic-based approach and supervised machine learning based approach have been experimented with for negation detection. To resolve uncertainty, a rule-based module is also proposed.

Chapter 7 shows a relationship extraction system, using both rule-based approaches and machine learning methods to extract relations from the lymphoma corpus.

Chapter 8 illustrates the design of predefined templates, ranking criteria and special rules tailored to the corpus, and development of a structured output generation system.

Chapter 9 sums up all the ideas presented in the thesis and suggests some directions for future work.

Chapter 2 Literature Review

2.1 Introduction

This chapter introduces previous works related to information extraction (IE). It gives a general overview about the current state of the art techniques in information extraction, and includes reviews of the four main subtasks: medical entity recognition (MER), negation and uncertainty detection, relation extraction (RE), and automatic structuring.

2.2 Medical Entity Recognition

2.2.1 Dictionary Look-up Approaches

A dictionary look-up approach also refers to concept mapping or encoding (Friedman et al., 2004) in some cases. It deals with the identification of relevant term(s) by mapping a concept from textual notes into a reference terminology, and it also links the concept with a referent identifier. Encoding information into a standard terminology can not only increase the accessibility of the information, facilitating storage and retrieval of it, but also achieve better interoperability between different information resources and enable the exchange and sharing of data.

Standard terminologies or a collection of terminologies with large lexical resources, such as the Unified Medical Language System (UMLS) (McCray, 2003), the Systematized Nomenclature of Medicine Clinical Terms (SNOMED CT) (IHTSDO, 2007-2014) and International Classification of Diseases (ICD) (WHO, 2009-2014), provide vast and rich lexical resources for these approaches to map the text in clinical notes to concepts.

The UMLS contains the greatest number of concepts in the medical domain, including three Knowledge Sources: Metathesaurus, Semantic Network and Specialist Lexicon (NLM, 2006-2014). There are more than 100 source vocabularies in the UMLS Metathesaurus, including terminologies designed for use in patient-record systems; large disease and procedure classifications used for statistical reporting and billing; vocabularies used to record data related to psychiatry, nursing, medical devices, and adverse drug reactions. The Semantic Network consists of a set of Semantic Types, that provide a consistent categorization of all concepts represented in the UMLS Metathesaurus, and a set of Semantic Relations between these concepts. The SPECIALIST Lexicon has been developed to provide the lexical information needed for the SPECIALIST natural language processing (NLP) System, including both common English words and biomedical vocabulary. The lexicon entry for each word records the syntactic, morphological, and orthographic information needed by the SPECIALIST NLP System.

The MetaMap Program is a well-known system that can parse free-text into simple noun phrases, and then map them to UMLS concepts (Aronson, 2001). In the program, first, the input text is parsed into simple noun phrases to limit the scope of further processing so that the mapping can be more tractable. For each phrase, the program generates rich variants, including acronyms, abbreviations, synonyms, derivational variants, meaningful combinations of these, and inflectional and spelling variants. Then it evaluates each candidate retrieved from UMLS Metathesaurus against the input text by first computing a mapping from the phrase words to the candidate's words and then calculating the strength of the mapping with four metrics: centrality, variation, coverage and cohesiveness. Complete mappings are constructed by combining candidates involved in disjoint parts of the phrase. The highest scoring complete mappings were chosen to represent the input.

The large number of source dictionaries in UMLS can help a dictionary look-up method to attain a high recall, however, they also bring some problems, e.g., indiscriminate use of a large set of overlapping terminologies in UMLS can introduce too much noise, and it can lower the efficiency of a dictionary-based system as well.

To overcome these issues, Huang et al proposed a context-based mapping method by restricting the concepts in different sections of the reports and mapping them to specific UMLS vocabularies (Huang et al., 2003). They found that this could increase precision effectively without a significant decrease in recall.

Another possible solution is to prune the mapping sources according to the focus of the research problem. For example, Long developed a program by using a small dictionary to divide disease statements into phrases for coding (Long, 2005). It was able to quickly identify the most specific codes available in SNOMED CT from UMLS for the statements in the discharge summaries. It was tested on 23 discharge summaries with 250 phrases to be coded, with only 19 false positives returned.

SNOMED CT is a comprehensive clinical terminology that provides clinical content and expressivity for clinical documentation and reporting. In 2004, the American Consolidated Health Informatics initiative identified SNOMED CT as the standard ontology for diagnoses, problem lists, and anatomy (Richesson and Krischer, 2007). There are over 400,000 atomic concepts and pre-coordinated concepts which are organized into 19 top-level hierarchies in SNOMED CT (in the January 2013 Release). It offers even more expressiveness through post-coordination (Elkin et al., 2003). Each concept in SNOMED CT is logically defined through its relationships to other concepts.

The Text to SNOMED CT (TTSCT) system was developed to detect SNOMED CT's concepts in free text and to annotate them with clinical reference terms (Patrick et al., 2007a). It generated an augmented lexicon composed of atomic words of SNOMED CT descriptions which were normalized with removal of all stop words, and each word was indexed with a list of associated DescriptionIDs (whose full names contain the word). A token matching algorithm was applied to the input text after

pre-processing (i.e., sentence boundary detection, stemming, lower case conversion, spelling variation generation, abbreviation expansion), the candidate DescriptionIDs are extracted by looking up the augmented lexicon. Then a dynamic programming algorithm checks all candidate combinations, and finds the combination with the maximum coverage of the text. In the evaluation of 487 clinical notes from an Intensive Care Service with 4,054 medical concepts, the system correctly identified 2,852 concepts, results in a precision of 70.4%, although the recall rate could not be fully evaluated (Wang and Patrick, 2008).

To reduce the ambiguities when mapping text to SNOMED CT core concepts, Hina et al first extracted 390023 SNOMED CT core concepts from SNOMED CT as a single gazetteer, and then developed separate gazetteers for each class defined in SNOMED CT (Hina et al., 2010). Evaluations on 300 discharge summaries showed that there is considerable improvement in reducing ambiguities by identifying the concepts in separate gazetteers rather than a single gazetteer.

The dictionary look-up component of the Mayo Clinic's Information Extraction system (renamed as clinical Text Analysis and Knowledge Extraction System (cTAKES) at present) was tested on a corpus of 160 free-text clinical notes (Kipper-Schuler et al., 2008), showing that such an approach had the potential to work well on MER, as it achieved an F-score of 0.56 for exact matches. However, it also indicated that the characteristics of clinical texts which include many lexical variations, disjoint concepts, and extensive use of abbreviations and acronyms could make the task more complex.

To address these issues, Wang performed some dictionary look-up experiments to match the concepts in the clinical notes with the SNOMED CT concepts (Wang, 2009). Removing unrelated terms and concept categories in the lexicon could reduce the ambiguity of the lexicon, leading to higher precision and recall. Proofreading such as correcting spelling errors or irregular conventions, resolution and expansion of abbreviations and acronyms in the notes, further increased both precision and recall. The dictionary look-up method finally achieved a relatively high precision with 74.81%, suggesting that it is an effective method to identify clinical concepts, but the extensive efforts it requires on pre-processing may not be easy to adapt to other corpora. In addition, it fails to identify some long and complex terms and resolve term disambiguation.

There are also other issues that should be noticed about dictionary look-up approaches. For example, the number of extracted entities is definitely hindered by the coverage of the terminologies in the medical corpora. For instance, a study pointed out that the Specialist Lexicon of UMLS only had about 79% coverage for syntactic information and 38% coverage for semantic information in a corpus of discharge summaries (Johnson, 1999). Thus, it is very likely that some entities will be overlooked in extraction with these approaches.

2.2.2 Rule-based Approaches

Rule-based methods were also prevalent in many early works on MER, and usually comprised hand-crafted rules and regular expressions to define patterns.

One of the typical rule-based systems is the medical language extraction and encoding (MedLEE) system (Friedman et al., 1994). It consisted of a pre-processor, a parser, a compositional regularizer, an encoder and a recovery component, which are described below: the pre-processor first identified sentences and abbreviations with rules, recognized and categorized words and phrases with lexical lookup; the parser used a grammar, a set of rules based on semantic and syntactic co-occurrence patterns to identify the structure of a sentence and to generate an intermediate structure that consisted of primary findings and different types of modifiers for the sentence; the compositional regularizer used a table of structural mappings to compose individual words into phrases when applicable; the encoder mapped words and phrases into codes with an encoding table; the recovery component allowed the parser to choose alternative strategies to structure the text if the initial one failed.

It was originally developed for the domain of radiological chest reports, but has subsequently been extended to mammography, discharge summaries, electrocardiography, echocardiography, and pathology (Friedman, 2000).

In general, rule-based methods are suitable to extract entities with explicit lexical, morphologic or orthographic patterns, e.g., medication information, test results and scores.

Turchin et al designed software with regular expressions to identify and extract blood pressure values and anti-hypertensive treatment intensification from physician notes with high accuracy (Turchin et al., 2006). This approach has both advantages and disadvantages. A set of regular expressions can be developed much faster than a full-fledged natural language processor, however, it lacks generality so that a new set of regular expressions has to be developed and validated for another particular task. Its applications are limited to data items with a constrained lexical scope, and variants like synonyms have to be manually generated as well.

To extract medication information from discharge summaries, Yang developed a relatively simple rule-based approach with manually curated feature term lists and token-based regular expressions (Yang, 2010). This approach performed reasonably well with a micro-averaged F-score of 80% for the term-level results and 81% for the token-level results. It was based on few annotated data, and was competitive without using pre-existing domain-specific tools and resources.

Similarly, without any deep NLP, such as part-of-speech (POS) tagging, chunking or syntactic parsing, a medication detection system first recognized drug names with a semantic lexicon, and then explored the context of these names to extract related information (mode, dosage, etc) according to rules capturing the document structure and the syntax of each kind of information (Deleger et al.,

2010). It initially obtained an F-measure of 77% in the 2009 i2b2/VA Challenge and increased to 81% with lexicon filtering and rule refinement. It demonstrated that a simple NLP system with surface rules could achieve high performance to capture medication-related information in clinical records.

The above methods could not perform well when encountering complicated medications that contain multiple signatures or contextual level information. MedEx adopted a more sophisticated method to cope with the task (Xu et al., 2010). It consisted of a sequential tagger and a combined parser: the sequential tagger combined lexical look-up, regular expression, and rule-based disambiguation components, could sigificantly improve the accuracy of semantic labeling of drug names and signatures. The parser combined a Chart parser and a regular expression-based Chunker to improve the ability to parse more complicated medications. The evaluation showed that MedEx can accurately extract not only drug names, but also associated information, such as strength, route, and frequency, with high F-scores (from 93.2% to 96.0%).

There are a number of research works indicating that dictionary look-up methods can also benefit from integrating with rules.

An NLP engine was developed to measure the quality of colonoscopy procedures, based on rules and dictionary look-ups (Harkema et al., 2011). Firstly, it utilized the MetaMap Transfer (MMTx) program (NLM, 2008-2014) to map words and phrases in each sentence in the report to a subset of concepts in the UMLS Metathesaurus, including the following semantic categories: Anatomical Structure, Neoplastic Process, and Sign or Symptom. Then a set of regular expression patterns was prepared to parse and interpret the temporal expressions and measurements of size. The ConText algorithm was used to identify the clinical and linguistic properties of the extracted concepts (Harkema et al., 2009). The values of the target variables were established by using some simple rules. The NLP engine attained encouraging results with 0.89 of average accuracy and 0.74 of average F-score.

A more detailed study was performed on a MER task with two dictionary-based systems MetaMap and Peregrine (Schuemie et al., 2007), by comparing with and without the use of a rule-based NLP module, which was composed of a set of post-processing rules that utilized POS and chunking information (Kang et al., 2013). It revealed that with this module, the F-scores of MetaMap and Peregrine improved by 12.3% and 14.1% respectively for exact boundary matching, and by 11.1% and 12.9% respectively for concept identifier matching, compared to those without this module.

However, it should be noted that the limitations of the rules can also propagate to dictionary look-up methods when they are integrated. Schadow et al used a rule-based parser by employing regular expressions to search for specimen headers in the diagnosis section of pathology reports, and guide the coding process by accepting only certain UMLS semantic types that fit the expected meaning of the input phrase (Schadow and McDonald, 2003). They found that 91% of 275 reviewed reports were

coded by this approach with the parser relying on regular expressions to attain structural clues. Thus it faced great difficulties when processing less consistently formatted reports.

2.2.3 Statistical Approaches

Recently, with more advanced machine learning algorithms being released, statistical methods have become more popular. One of the main characteristics of these methods is that they can utilize contextual information to predict the entities, which is especially useful when the coverage of the dictionary is quite limited, or rules or patterns are too difficult to be captured from the training examples.

According to the strategies by which machine learners generate their models, they can be classified as either generative techniques or discriminative techniques. Generative techniques seek to create rich models of probability distributions, and with such models, they can generate synthetic data; discriminative techniques are usually thought to be more utilitarian, since they directly estimate posterior probabilities based on observations. Besides, discriminative models often allow using more features than generative models, since when many features are used, generative models can become intractable. Bayesian networks (Ben-Gal, 2007) and hidden Markov models (HMM) (Stamp, 2004) are examples of generative methods, while support vector machines (SVM) (Cristianini and Shawe-Taylor, 2000) and conditional random fields (CRF) (Sutton et al., 2007) are examples of discriminative methods.

SymText uses probabilistic Bayesian networks to represent semantic types and relations (Haug et al., 1994). Syntactic knowledge comes from augmented transition networks. The system depends on a set of reports to train the network for a specific medical domain. For instance, when extracting pneumonia-related concepts from chest x-ray reports, it utilized three Bayesian networks: the first Bayesian network represented radiographic findings; the second one modeled the diseases that can be described in the reports; the third one modeled the devices that are frequently described in the chest x-ray reports (Fiszman et al., 2000). The performance of SymText was compared against four physicians, two different keyword searches, and three lay persons. The accuracy of SymText was similar to that of physicians and better than that of lay persons and keyword searches.

In its successor M+ (Christensen et al., 2002), Bayesian networks were represented in an object-oriented format and a bottom-up chart parser provided syntactic analysis. In addition, M+ used an abstract semantic language to link Bayesian network types to each other in a predication format. The advantages of using Bayesian networks for semantic representation include tolerance of noise and partial matches and sensitivity of context to the recognition of semantic patterns. In addition, its ability to guess the semantic types of unknown words is very valuable for bootstrapping semantic knowledge.

Nonetheless, both systems are limited to the domain covered by the semantic knowledge that is stored within the Bayesian networks. The creation of training cases is also a time-consuming task.
MER is one the tasks of the 2009 and 2010 i2b2/VA Challenges. The 2009 Challenge focused on the identification of seven types of medication information in discharge summaries, including dosages, modes of administration, frequencies, durations, and reasons for administration (Uzuner et al., 2010). The winner used a CRF model to extract the entities and achieved the best micro-averaged F-score of 85.65% (Patrick and Li, 2010). The 2010 Challenge defined three classes which were Test, Problem, and Treatment were extracted (Uzuner et al., 2011). Most of participants used CRF as the framework together with feature engineering specific to these classes in the challenge. For example, Patrick et al adopted a CRF learner to identify the entities and attained a micro-averaged F-score of 81.79% (Patrick et al., 2011).

Theoretically, CRF is a representative sequence labelling algorithm, which is suitable for the MER tasks, while SVM is based on large margin theory and has difficulty handling sequence labelling problems as it ignores the relationships between neighbouring tokens in sequences. For instance, Li et al proved that CRF outperformed SVM on MER of disorders in clinical notes (Li et al., 2008). However, CRF also has its own weaknesses, e.g., it is unable to utilize the global information in the notes.

To combine the advantages of both CRF and SVM, Tsochantaridis et al proposed a new machine learning algorithm named Structural Support Vector Machines (SSVM) for structural data, which is an SVM-based discriminative algorithm for structural prediction (Tsochantaridis et al., 2005). Tang et al applied SSVM to recognize clinical entities in discharge summaries and compared the performance of CRF and SSVM on the 2010 i2b2/VA Challenge data (Tang et al., 2012). Their evaluation showed that the SSVM-based system required less training time, but achieved better performance than the CRF-based system with the same features.

A combination of machine learners can also overcome some shortcomings of a stand-alone classifier.

Wang and Patrick presented a machine learning approach to MER using a combination of machine learners (Wang and Patrick, 2009). They firstly built a CRF based model to identify the entities and then reclassified the identified entities by Maximum Entropy and SVM models. A voting strategy was employed between the three classifiers to determine the class of the recognized entities. The results showed that the reclassifier effectively boosted the F-score by 3.35% over the stand-alone CRF model.

Xu et al purposed a novel method for MER of follow-up and time information in radiology reports which combined a labeled sequential pattern (LSP) classifier with a CRF recognizer (Xu et al., 2012). The LSP classifier was an SVM classifier that used binary features, each of which corresponds to a set of patterns mined for positive set and negative set respectively. In the training phase, the LSP classifier disregarded a large number of negative examples and thereby improved the consistency of

local contexts of positive examples; in the test stage, it used global patterns in a sentence to narrow down a set of candidate sentences. The experiment showed significant improvement of the performance of the CRF recognizer, due to the process of cleaning-up the training data and compensation for CRF's inability to use global information by the LSP classifier.

Feature Selection

Choosing proper features for machine learners is as important as selecting an appropriate machine learning technique, as features can provide critical clues for the learners to construct the models and make prediction on the test data. Usually, features can be categorized as follows:

- Contextual features: local context features (e.g., the bag-of-words (BOW), also called context window of words), global context features (e.g., section context, document types).
- Lexical features: such as lemma, lowercase of words.
- Syntactic Features: part-of-speech (POS) tags, chunking information, etc.
- Morphological features: affixes, orthography and so on.
- Semantic features: for example, exploring external resources can provide additional domain knowledge.

There is much research addressing the importance of feature selection or feature engineering.

The results from Li and Martinez's experiments on the extraction of a large number of categories from pathology reports (Li and Martinez, 2010) showed that a high level of accuracy could be attained on predicting nominal categories by using BOW feature. This indicated that pathology reports contain similar lexical items that can be captured by a BOW model. It also revealed that for numeric categories, richer features were required to improve the performance.

To minimize this limitation by combining supervised machine learning with empirical learning of semantic relatedness from the distribution of the relevant words in unannotated text, Jonnalagadda et al used a feature of distributional semantics with words that appear in similar contexts to the word in question, in addition to the traditional features such as dictionary matching, pattern matching and POS tags (Jonnalagadda et al., 2012). The evaluation of this approach on the i2b2/VA concept extraction corpus showed that incorporating this feature significantly aids MER.

The semantic domain knowledge from terminologies can be very helpful for determining the correct concept boundary and the semantic category of these concepts. Wang boosted the performance of a machine learning-based system significantly by using the results from the dictionary look-up on SNOMED CT as a feature, particularly on the recall. On another MER task, Abacha and Zweigenbaum obtained the best results from a CRF classifier combined with semantic features obtained from UMLS, apart from lexical and morpho-syntactic features (Abacha and Zweigenbaum, 2011).

2.2.4 Hybrid Approaches

On one hand, rule-based systems tend to provide reliable results with a relatively small amount of training data, and the hand-crafted rules are comprehensible for the developers or domain experts, thus it is easier to detect and correct errors during error-analysis. However, it is difficult for a rule-based system to deal with unfamiliar or erroneous input data, and the hand-crafted rules are usually tailored for a specific domain or task, which are not readily reusable. On the other hand, it might be possible for an IE system based on statistical methods to cope with problematic data by learning from available examples through training. Besides, statistical methods achieve comparable or better performance by simply adapting features from the corpus, hence they can be easily adapted to other domains. For example, BioTagger-GM was a machine learning tagger, originally developed for the detection of gene/protein names in the biology domain (Torii et al., 2009). To extract concepts from clinical documents, Torii et al replaced one of its components – BioThesaurus (Liu et al., 2006) with a collection of clinical terms extracted from discharge summaries, supplemented the section header as an additional feature, adjusted the context window size to derive features by encoding nearby tokens, and removed the hand-coded rules for post-processing. Evaluation on the 2010 i2b2/VA Challenge indicated its portability to the clinical domain and achieved good performance with 0.890 of F-score. However, statistical methods usually require a large amount of training data to create a gold-standard, which are typically expensive to obtain. Furthermore, statistical methods can benefit from the utilization of the rule-based NLP module as it may capture particular patterns that cannot be well handled by machine learning, especially for entities with low frequencies. Therefore in recent years, there has been a trend for using hybrid approaches with a combination of statistical methods and rule-based approaches when designing an IE system for MER.

Entity recognition in the biomedical domain has mainly used rule-based modules in post-processing to fix errors produced by the machine learning techniques. For example, Zhou and Su used an HMM model as the machine learner in their system to recognize biomedical entities, and employed some rule-based methods in post-processing, including named alias resolution, rule construction to handle classification errors in nested named entities, expansion of abbreviations and utilization of open and closed dictionaries to detect unknown words (Zhou and Su, 2004). The baseline HMM-based learner only achieved an F-score of 60.3%, but after post-processing, it boosted to 72.5%. The possible reason for the improvement was that the rules could bridge the gap when the system encountered unseen words, and handle complex entities better.

Nevertheless, rule-based components were usually prepared to generate additional features in pre-processing rather than post-processing for machine learning-based systems to allow the systems to self-optimize better. For instance, de Bruijn et al applied upper-case/lower-case patterns seen across the document and case-folding patterns in sentences as features for the feature space for the machine learner (de Bruijn et al., 2011).

A hybrid system named Textractor was developed for the 2009 i2b2/VA Challenge (Meystre et al., 2010), where two modules were based on machine learning algorithms, while other modules employed regular expressions, rules, and dictionaries, and another module embedded MMTx. It achieved satisfactory performance with F-score of 77%, recall of 72% and precision of 83%, which made it one of the top 10 best performing systems in the challenge.

Another hybrid system was established to automatically classify the surgical margin status from pathology reports following prostate cancer surgery (D'Avolio et al., 2007). By the preliminary pilot analysis of a small subset of reports, heuristics were designed for capturing potential margin sentences. With five simple rules based on keyword appearance, it was able to capture positive sentences from 780 of 782 reports, with accuracy over than 97%. The extracted sentences were tokenized into vectors of lowercase words and then passed to an implementation of an SVM classifier to classify to three classes: “positive (involved) margins”, “negative (uninvolved) margins”, and “not-applicable or definitive”, with an overall accuracy of 97.18%.

Roberts et al compared a lexical look-up method with a statistical method, and to a method which combined the two approaches for MER (Roberts et al., 2008a). The lexical look-up method based on UMLS had good recall, but poor precision, which was largely due to the ambiguity between domain terms and general language words. As much of the ambiguity was caused by a small number of terms, the ambiguous terms were filtered with a filter list using simple heuristics to improve precision. The SVM classifier trained on lexico-syntactic features alone yielded higher precision, but lower recall than the lexical look-up method. When these two approaches were combined, it gained higher recall with little loss of precision. The results suggested that they could compensate each other, and attain better performance than each on their own.

To extract medication information from the 2009 i2b2/VA Challenge data, Tikk and Solt (Tikk and Solt, 2010) first used a rule-based method by creating a custom grammar that combines the benefits of using vocabularies and regular expression rules, and the submission ranked fifth in the challenge. Then they used CRF models with vocabularies and typical entity patterns taken from the rule-based method as one of the features. They found that the standard CRF-based approach did not improve upon the rule-based approach with a limited amount of training data. However, when additional training data were made available, the approach resulted in considerably better performance.

They concluded that rule-based methods are easier to comprehend with a less time-consuming training phase, in favour of iterative trial-and-error development; since feature definition was more straightforward and less error-prone, it is more convenient for machine learning approaches to create models, and the performance can be improved with less manual effort with these approaches by adding more features, compared with cumbersome rule tuning with rule-based methods.

Nevertheless, there is also research indicating that rules can contribute to significant improvement over machine learning-based systems through post-processing in the clinical domain.

To extract numeric categories from pathology reports, Martinez and Li applied a two-step process (Martinez and Li, 2011). At first, machine learning-based sentence classifiers for each class were built to identify the positive sentences, then the numeric values were extracted, and the number closest to the median is assigned as output. However, when the target sentence was correctly identified, the simple strategy of using the median was not good enough to identify the right number for "Nodes positive". By manual analysis of the sentences, they found that in most cases the number of positive nodes is given together with the examined nodes, thus they devised some simple rules to identify the number of positive nodes, and achieved significant improvement over the machine learning approach alone.

It is noticeable that although the hybrid approaches can have the advantages of both rule-based and statistical methods, the limitations from each method can still maintain in hybrid approaches. For example, heuristic rules developed based on syntax structures in the training data, may not work as expected in the test set, if they cannot cover all possibilities of expressing the information to be extracted.

2.3 Negation and Uncertainty Detection

2.3.1 Negation Detection

In clinical reports, the presence of a medical term does not indicate the presence of the clinical condition represented by that term is certain. In fact, a large portion of clinical findings mentioned in the reports (e.g., discharge summaries, radiology reports) are negated. Accurately identifying whether these findings are present or absent is critical to extracting pertinent information from the reports and indexing them.

In the pathology domain, the narrative reports usually contain negative findings or diagnoses as well as positive ones. To detect whether particular findings are negated is of great significance to make a proper decision on diagnosis and prognosis. For example, constitutional symptoms such as fever, weight loss and night sweats are known to be of prognostic value in non-Hodgkin lymphoma (NHL) (Edge et al., 2010; Sobin et al., 2009). The presence or absence of these symptoms can be used to define two categories for each stage of NHL: A (if symptoms absent) and B (if symptoms present) (Sobin et al., 2009).

Generally, negation detection includes the detection of negation cues (specific terms to indicate negation) and their scope (the text negated by the terms). In the following example, "No evidence of

malignancy", where "No evidence of" is the negation cue and "malignancy" is in the scope negated by the cue.

Rule-based approaches can be sub-classified to lexical pattern matching methods and syntax-based approaches depend on whether they utilize syntactic information in the texts.

Previous work suggests that a small set of words cover a large portion of negation cues. It is evident that "no", "denies/denied", "without", and "not" are the most frequently used terms to indicate the absence of clinical observations (Chapman et al., 2001a). Several rule-based approaches that utilized lexical pattern matching have been widely applied to the clinical domain.

Negfinder used a Left-to-right Rightmost-derivation parser to detect negations in surgical notes and discharge summaries and achieved sensitivity of 95.7% and specificity of 91.8% without extracting syntactical structures of sentences and phrases (Mutalik et al., 2001). However, it could not detect negated concepts correctly if the negation cue was far away from negated concepts, since it terminated a concept list or negation if there are more than three intervening words between concepts or between a negating phrase and a concept.

NegEx, a regular expression-based algorithm, which is simple to implement, has shown success in detecting negations in discharge summaries with recall of 77.8%, and precision of 84.5% (Chapman et al., 2001b). It relied on three types of terms to determine whether a condition is negated, namely trigger terms, pseudo-trigger terms, and termination terms. Trigger terms like "no" and "not" indicate that the clinical conditions within the scope of the trigger term should be negated. Pseudo-trigger terms, such as "not rule out" and "gram-negative", which appear to indicate negation but identify double negatives or modified meanings instead. Termination terms, e.g., "but" and "though", can terminate the scope of the negation before the end of the window. Since it did not take into account any syntactic clue to determine the negation scope, it had faced difficulty in determining the scope of the negation phrase in some complex cases. Similar to Negfinder, a rigid window might lead to omission of some negated UMLS terms in long lists of terms, or when the negation phrase and a term were separated with a distance larger than the window size. Thus, the algorithm was likely to only negate part of high-level composite concepts.

Without any customization, its application to the pathology domain had lower performance (Mitchell et al., 2004), probably because the negation and pseudo-negation phrases that were used by NegEx may not adequately cover the spectrum of phrases in pathology reports. It also revealed that failure to correctly map the text phrases to UMLS concepts is one major source of errors.

PyConTextNLP was an extension of the NegEx algorithm that included modifications of the scoping rules, and more functionality for defining user- and task-specific rules (Chapman et al., 2011a).

Instead of a fixed window, the algorithm operated on the whole sentence to get the scope, unless it found user-defined conjunctions.

A unique ontology developed for negation is another solution purposed by Elkin et al. They extended the work of Mutalik et al (Mutalik et al., 2001) and Chapman et al (Chapman et al., 2001b) by performing their study using SNOMED CT and by utilizing a second independently developed ontology for negation. The negation ontology contained two sets of terms and their associated rules, with one set starting negations and another set stopping the propagation of negations. The system first mapped the text in a sentence to SNOMED CT to attain SNOMED CT concepts, and then assigned one of the three possible assertion attributes according to the negation ontology. The recall of the assignment of negation was 97.2% and the precision was 91.2%. The most common reason for failure was the inability of SNOMED CT to represent the negative concepts, as the human reviewer identified that 205 of 2028 negative concepts were not mapped by SNOMED CT, revealing that the terminology had 88.7% of coverage of the negative concepts.

For complicated negation cases, defining negation scope is still a challenging task. The above approaches could perform reliably when a negated concept is close to a negation cue and it can be mapped to a controlled terminology (e.g., UMLS, SNOMED CT) , but unsatisfactorily when they are separated with multiple words or they fail to be mapped to a controlled terminology. Syntactic information is a useful clue to resolve this problem.

NegExpander identified negated UMLS terms by constructing conjunctive phrases to define the negation boundaries (Aronow et al., 1999). Conjunctive phrases were referred to a group of noun phrases connected with conjunctions such as “and,” “or” and “,”. NegExpander did not take the conditional possibility of phrases such as “rule out” into account, hence it could not distinguish from uncertainty in some cases. As well, it could not distinguish between pre-UMLS and post-UMLS negation phrases inside conjunctive phrases. This might result in incorrectly negated UMLS terms preceding the pre-UMLS negation phrases or succeeding the post-UMLS negation phrases inside conjunctive phrases, consequently reducing the overall algorithm’s specificity.

A hybrid approach by combining regular expression matching with grammatical parsing has been proposed to detect negations (Huang and Lowe, 2007). The results showed that the structured grammar rules developed using linguistic principles were more powerful than detecting negated concepts at a fixed distance from negation cues. It did not rely on concept mapping to cluster words before detecting negations, thus it is more intuitive in understanding a complex sentence, which is very helpful to locate the negated phrase in the sentence. One of the limitations was reflected in the coverage of the negation grammar, which was not as comprehensive as expected during the test. It was also limited by the parsing performance of the NLP parser, especially the errors in noun phrase identification. Another limitation was that they only evaluated radiology reports. Radiology reports might contain significantly fewer negation phrases frequently used by non-radiology reports.

Therefore, it should be further validated on other types of clinical reports. Another approach to detect negation is dependency parser-based negation (DepNeg) (Sohn et al., 2012), using dependency parses which directly encode thematic roles like subject and object performed quite well and was able to identify complicated negations that were wrong in the cTAKES (Savova et al., 2010) by a limited set of dependency rules compiled from a small data set.

SynNeg is a negation scoping tool that uses morphological and syntactic information provided by the MaltParser (Nivre et al., 2007). It assumes that a negation scope does not cross the boundary of a sentence unit (i.e., subject + verb phrase). The MaltParser assigns the ES (logical subject), FS (dummy subject) or SS (other subject) and DEPREL (Dependency Relation) tag to a subject of a sentence unit. When a cue is found, SynNeg checks the DEPREL tags of either the following token or the preceding token from the cue to find a subject DEPREL tag. It also checks the POS tags for coordinating conjunction, minor delimiter and subordinating conjunction. Every time one of these POS tags is found, the position of the token is stored as a boundary candidate. Once a subject DEPREL tag was found, the nearest boundary candidate from the subject DEPREL tag was set as the boundary for the scope.

Both lexical pattern matching methods and syntax-based approaches have their merits and limitations on negation detection. Tanushi et al compared three different tools: NegEx, PyConTextNLP and SynNeg for determining negation scope in Swedish clinical text and achieved similar results with around 80% of F-scores (Tanushi et al., 2013). The pros and cons for these tools on negation detection were described as follows:
NegEx was efficient and simple, but it was not able to handle longer or complex sentences, or sentences with contradictory statements. PyConTextNLP was possible to improve results if lexical phrases that defined the boundaries for the scopes are determined. However, it was also likely to fail in some ambiguous cases. SynNeg was more generalizable and easier to port to another domain or language by using syntactic information, but its performance was hindered by the syntactic parser to a great extent.

There are different opinions on whether rule-based or machine learning-based approaches are more suitable for negation detection in the clinical domain. Some studies pointed out those rule-based methods outperformed machine learning-based approaches, while other researches suggested that machine learning-based systems have better performance than their rule-based counterparts.

Goryachev et al implemented and modified two existing rule-based algorithms: NegEx and NegExpander, and created two classification models based on SVM and Naive Bayes (Goryachev et al., 2006). All four methods were evaluated on 100 randomly selected outpatient notes, and the results revealed that NegEx and NegExpander did slightly better than SVM, and Naive Bayes has the worst performance.

As part of the assertion task, negation detected by machine learning techniques has been the state of the art in the 2010 i2b2/VA Challenge. For example, Patrick et al converted a baseline rule-based method to a statistical method trained with CRF, and gained more than 92% of F-score on the "absent" category (which stands for negated medical problems) in the task (Patrick et al., 2011).

2.3.2 Uncertainty Detection

Compared to negation detection, there are fewer studies addressing uncertainty detection. One of the possible reasons is that it is harder to determine uncertainty assertions than negations, sometimes even for human experts. To produce a freely available resource for research on handling negation and uncertainty in biomedical texts, Vincze et al annotated a corpus called the BioScope corpus (Vincze et al., 2008). In the study, they found that uncertainty detection was a more difficult task than identifying negation because of a higher level of keyword/non-keyword ambiguity. This was confirmed by the agreement rates of the human annotations for these two tasks, where the agreement rates for the keywords of uncertainty were about 3.4-5.7% lower than those of negation.

The MedLEE system made a distinction between negated and uncertain concepts through encoding negated concepts and certainty modifiers (Friedman et al., 1994). It defined five concepts to represent certainty information related to the finding: no, low certainty, moderate certainty, high certainty, and cannot evaluate, and therefore the words and phrases in the reports relating to this type of information would be mapped into one of these concepts. This limitation could greatly facilitate the subsequent retrieval of the structured findings in the reports. Since this type of information usually hedged information concerning the certainty of the findings in the reports and was basically vague, it was thought to be too hard to represent this type of information more precisely in other ways.

Negation and uncertainty detection was one of the emphases of the 2010 i2b2/VA Challenge, named as an "assertion classification" task (Uzuner et al., 2011). This task extended traditional negation and uncertainty extraction to conditional and hypothetical medical problems, and brought in information about the person to whom the medical problem belonged.

Uzuner et al presented two different approaches for the assertion classification task (Uzuner et al., 2009). One was extension of NegEx algorithm (ENegEx) to cover alter-association assertions; the other was a machine learning solution that applied SVM to build a Statistical Assertion Classifier (StAC). It turned out that StAC outperformed ENegEx, which benefited the most from a four-word context window.

A hybrid system was designed by Clark et al for the task (Clark et al., 2011). To combine machine learning algorithms with linguistic knowledge, regular expression-based patterns, and scope enclosure rules, they first fed the output from a statistical scope module to a rule-based status module, and input the results from that module, as well as other features derived from linguistic knowledge, to a final

statistical classifier. They thought this was a feasible way to leverage rule-based and statistical techniques, as rule-derived information could weight features automatically with respect to its contribution to the overall accuracy and the degree to which the information correlated with other features, when it was converted to features as an input to a machine learner.

Their study also revealed another finding that the choice of features was more important than the choice of the classifier for this task. Rather than exploring a large number of features, they only selected a small number of features based on the analysis of the data and linguistic intuitions. They used the same features as input to three machine learning classifiers: a Maximum Entropy classifier, a SVM with linear kernel and a CRF classifier, and the results did not differ significantly amongst each other.

It can be seen from the above studies that uncertainty detection is commonly accompanied with negation detection, probably because both negation and uncertainty indicate the non-factual information, which should be distinct from the positive clinical findings or diagnoses. The approaches to resolving negation are also applicable to resolve uncertainty in most cases.

2.4 Relation Extraction

A relation represents the link between two entities. In the general domain, the relations of interest usually are quite explicit, for instance, EMPLOYEE_OF, PRODUCT_OF, and LOCATION_OF, relations defined in the seventh Message Understanding Conference (MUC-7) (Chinchor, 1998). Therefore, pattern-based approaches can work fairly well for this task. In a pattern-based RE system, a template or frame is defined to hold relations between two entities, which is a table with slots that can be instantiated with the fragments or segments of information extracted from a given document. A set of pattern matching rules was used to assign entities to the slots of such templates. However, RE in biomedical or clinical domain can be challenging as it needs more domain knowledge to tackle, in addition to ambiguity and complexity that embed in the texts.

In the biomedical domain, RE systems usually focus on extracting interactions or relationships between biomedical entities. For example, extracting relations between genes and diseases (Chun et al., 2006), identifying relations between genes and proteins (Bundschus et al., 2008; Fundel et al., 2007), determining treatment relations between drugs and diseases (Rosario and Hearst, 2004).

In the clinical domain, the relations to be extracted are relationships that hold between medical entities, requiring a lot of domain knowledge as well. For example, the 2010 i2b2/VA Challenge proposed a relation classification task, which aimed to assign relation types between medical problems, tests, and treatments, including a treatment that improves or worsens a problem, a treatment that causes a problem, a treatment that is administered or not administered because of a problem, a test revealing a problem, a test conducted to investigate a problem, and a problem that indicates a problem

(Uzuner et al., 2011). RE between entities in clinical reports can improve accessibility to the high level of information in these reports.

2.4.1 Rule-based Approaches

Typical approaches to RE in most early work in the clinical domain were usually rule-based and relied on full parses, domain-specific grammars, or large domain knowledge bases.

A full syntactic and clinical sub-language parser was used to fill template data structures of medical statements in the Linguistic String Project (Sager et al., 1994), which were mapped to a database model that incorporated medical facts and the relationships between them.

Both MedLEE (Friedman et al., 1994) and BioMedLEE (Lussier et al., 2006) made use of a semantic lexicon and grammar of domain-specific semantic patterns. The patterns encoded potential relationships between entities, allowing direct matching of entities and the relationships between them in the text.

There were also some systems incorporating large-scale domain-specific knowledge bases. For example, MEDSYNDIKATE used a dependency parse of texts and a description logic knowledge base re-engineered from existing terminologies to build a rich discourse model of entities and their relationships (Hahn et al., 2002). A similar method also adopted by MENELAS, included a full parse, a conceptual representation of the text, and a large scale knowledge base (Zweigenbaum et al., 1995).

Rule-based approaches were still considered to be simple and reliable to apply in some recent RE tasks.

Halgrim et al used simple heuristic rules to associate each medication name with its related fields (Halgrim et al., 2011), which can be processed in three steps: firstly, they identified the closest prior and subsequent names for each field; secondly, they linked each field to one of those two names and in most cases, usually the prior name unless the distance to the subsequent name was much shorter than the one to the prior name by more than two lines; thirdly, they applied a few rules to assemble the pairs if more than one field of the same type was linked to the same name.

Nikolova and Angelova presented research work on automatic extraction of relations between medical concepts with rule-based methods (Nikolova and Angelova, 2011). Due to a lack of conceptual resources with a Bulgarian ontological vocabulary, they formed a terminological dictionary of Bulgarian terms, translated them to English and extracted their UMLS definitions, which were processed automatically by a semantic parser named RelEx (RelEx Developers, 2007-2014). They also applied additional rules, built a set of new relations such as IS-A and AFFECTS and inserted them into the conceptual resource. The accuracy of the system was between 81% and 89%. The major

source of errors was due to wrong parsing trees from the parser, thus better parses would lead to improvement of the system.

Abacha and Zweigenbaum proposed a knowledge and linguistic pattern-based approach for the extraction of medical entities and the semantic relations linking them (Abacha and Zweigenbaum, 2011). Specifically, for every pair of medical entities, they collected the possible relations between the semantic types in the UMLS Semantic Network. Then they constructed patterns for each relation type and matched them with the sentences to identify the correct relation. They obtained good results in precision and F-score compared to other semantic RE approaches, which suggested that such methods give good control on the extraction precision by testing and improving manual patterns but with an expensive cost needed to attain a good recall. A possible improvement is to integrate such methods in hybrid approaches to balance their qualities with that of statistical methods.

2.4.2 Statistical Approaches

The first effort at applying statistical methods to extraction of relationships from clinical texts was made by Roberts et al (Roberts et al., 2008b). They designed and implemented a machine learning-based system for RE from a clinical corpus annotated with seven types of clinically important relationships. There are several important findings from their experiments:

- Both lexical and syntactic features were assigned to tokens and entity pairs for the SVM classifiers prior to classification. For most relation types, the classifier with syntactic features outperformed the one with non-syntactic features with a higher macro-averaged F-score by 2-4%.
- The system achieved an overall F-score of 72%, only just 3% below the score of human inter-annotator agreement, showing that it is possible to extract important clinical relationships from free text using supervised machine learning methods, at the level of accuracy approaching to that achieved by human annotators.
- The precision for relation recognition over extracted entities remained close to that over gold-standard entities (64% *vs.* 63%), however, the recall decreased significantly from 76% to 40%, resulting in a dramatic drop of F-score by 22%. Apparently, good RE depends on accurate MER for an end-to-end RE system.

Patrick and Li classified the relations among six medication entities defined in the 2009 i2b2/VA Challenge with a machine learning approach trained with SVM (Patrick and Li, 2010). The features that they employed could be an important reference for feature engineering for other tasks in the clinical domain, including:

- Contextual features: words in an optimal window size before and after each entity; words between the two entities; words inside of each entity.
- Semantic features: entity types of each entity; entity types between the two entities.

Whereas the features for RE are not limited to the above, more features were also explored in other works, such as syntactic features (e.g., POS tags, chunk information, parse trees and dependency paths from a parser).

However, it should be also noted that over-inclusion of complex features may harm the performance. Jiang and Zhai conducted some experiments to evaluate the effectiveness of different feature subspaces for RE (Jiang and Zhai, 2007). They explored three different representations of sentences: sequences, syntactic parse trees, and dependency trees, and found that the performance improved only slightly by combining the three feature subspaces. Their experiments also showed that over-inclusion of complex features may not improve the performance much and hurt the performance instead. They concluded that a combination of features of different levels of complexity, coupled with feature pruning for particular tasks, can give better performance for RE.

A large amount of research has presented how different features and combinations or representations of features affected the performance of a machine learning-based system for RE.

Rink et al developed a state of the art system that automatically extracted relations between medical concepts with a supervised machine learning approach (Rink et al., 2011). A single SVM classifier and several knowledge resources such as Wikipedia (Wikipedia community, 2001-2014), WordNet (Fellbaum, 1998), and General Inquirer (Stone et al., 1966) were used in the system.

They assumed that relations that had similar contexts should also have similar relation types, but conventional lexical features like a string of words between the relation arguments was unable to directly capture this, as they could not reflect minor lexical variations. To overcome this issue, they used a sequence similarity metric known as Levenshtein distance (Levenshtein, 1966) to obtain similarity features to indicate the percentage of similar relations of each relation type. They found that the RE system benefited mostly from lexical, syntactic, semantic context features and similarity features. In addition, the knowledge resources were proved to improve RE performance by providing information about whether two entities in the candidate pairs are strongly associated.

Frunza and Inkpen also adopted SVM as a classification algorithm to train models to extract relations between diseases, treatments, and tests from clinical notes (Frunza and Inkpen, 2011). The best results with 86.15% of overall F-score were obtained by using rich features, including BOW, entity types, verb phrases identified by the GENIA tagger (Mitsumori et al., 2006), contextual information attained from ConText tool (Chapman et al., 2007a), entities extracted from the training data, semantic vectors that captured the distributional semantic correlation between the entities and each relation of interest.

Unlike the traditional BOW representation, Dogan et al represented a relationship between medical problems, treatments and tests with a scheme of five distinct context-blocks determined by the position of concepts in the text: the introductory, first concept, connective, second concept, and

conclusive block (Dogan et al., 2011). They thought this scheme could have better management of the word position information, which may be critical in certain relationships.

Taking into account variability in the sentences, Minard et al used different features, including those specific to the domain (e.g., entity types), and those similar to the general domain (e.g., the words and stems which constitute the entities and the headword of each entity) (Minard et al., 2011). They obtained reasonable results with an overall F-score of about 0.70. They believed that the features they selected were general enough to be ported to other corpora, with an adaptation of the features to the corpora.

A machine learning-based classifier usually performs better on the larger classes than on the smaller classes. To recognize the less prevalent classes, one possible solution is augmentation with handcraft rules as mentioned above, another way is down-sampling the larger classes. As one of the participants of the 2010 i2b2/VA Challenge, de Bruijn et al observed that some of the relationship types were much more frequent than others (e.g., negative problem-problem relations were about eight times more than positive ones) (de Bruijn et al., 2011). To address the imbalance of the category distribution, they down-sampled the training set to a positive/negative ratio between 1:2 and 1:4, selected as a development set, which reduced a classifier's bias towards the majority class. Moreover, this down sampling was especially important for the semi-supervised training on the supplied unlabelled data, which boosted the system with 0.74% F-score.

2.5 Automatic Structuring

This automatic structuring of pathology reports presented in the thesis requires nearly all the processes to be performed automatically by the system, thus minimizing manual interference. This makes it different from the work of Chen et al of semi-automatic structuring of clinical documents (Chen et al., 2010). In their work, the system used a keyword-based and semantic-driven data matching methodology to extract the specific information from the textual clinical documents. When the clinician started the matching operation based on the selected keyword, the information matching modules applied the matching operations based on the matching profile of the keyword retrieved from the matching metadata database. Through the extraction verification interface, clinicians could verify and correct the matched information. The extracted data were filled into predefined case-oriented templates, which were designed for collecting the necessary information for different diseases or research purposes.

Although one of the goals of this project is to facilitate medical informatics in cancer registries, the focus is on the detail structured report fields that can provide sufficient information for cancer staging rather than inference of the stage factors from the narratives (e.g., T, N, and M stages). Thus, the system should be also distinguished from the work of McCowan et al and the medical text extraction (MEDTEX) system that targeted on automatic extraction of cancer staging information from medical

reports. On the one hand, McCowan et al developed a prototype software system to automatically extract cancer staging information from medical reports of lung cancer patients (McCowan et al., 2007). The system trained SVMs to classify T and N stages' relevance of each report, and then sentences from relevant reports were analyzed by a series of SVM-based or rule-based classifiers according to specific contributing factors defined in the staging guidelines. Results from the classifiers were post-processed to determine the final T and N stages. The system achieved an overall accuracy of 74% for T staging and 87% for N staging. M staging was omitted in the system, thus proper stage group information could not be obtained, as it is computed with the combinations of T, N, M staging. On the other hand, a symbolic rule-based system named MEDTEX was proposed to extract TNM staging factors automatically from free-text pathology reports by subsuming items specified in a structured report (Nguyen et al., 2010). SNOMED CT was used as a base ontology to provide the semantics and relationships between concepts for subsumption querying. SNOMED CT expressions were used to populate a structured report according to the College of American Pathologists' surgical lung resection cancer checklist (CAP, 1991-2014), which could consist of a single concept or a combination of concepts post-coordinated by the user according to SNOMED CT's compositional grammar. TNM stages were classified by building logic from relevant structured report items. However, the structured report items other than stage were not evaluated due to the lack of readily available validation data.

Automatic structuring of medical reports in other clinical sub-domains such as radiology has been studied as well, where the issues addressed are likely to be resolved in this project. For example, automatic structuring of radiology reports is a difficult task for the following reasons:

- Automatic structuring requires deep understanding of the domain because it is desirable to translate all relevant information in the free text into a structured form.
- Automatic structuring must deal with ungrammatical writing styles as shorthand and telegraphic writing styles are common in radiology reports. Moreover, each subspecialty of radiology may have different language models. In addition, there are many stylistic variations between radiologists.
- The vocabulary is large. Large numbers of complex medical terms, proper names, product names, abbreviations, and staging codes are used in radiology reports. Hundreds of descriptive adjectives are used that are not found in any common electronic medical glossaries.

To cope with the issues above, Taira et al (Taira et al., 2001) did not use existing lexical sources such as the UMLS to build the specific lexicon for lexical analysis, because these lexical sources do not contain a sufficient number of semantic categories to support the statistical parsing and semantic interpretation algorithms in their system. In addition, the coverage of descriptive adjectives in the radiology domain is not yet adequate. Lexical terms were gathered from two distinct types of sources: published sources and actual radiology reports. The terms from the published sources can ensure the

generality of the concepts covered by the lexicon; the collection of words and phrases from actual radiology reports ensures that most of the string representations for the concepts are included.

Statistical and machine learning methods were used extensively throughout their system, which consisted of the following collaborative modules:

1. The structural analyzer divides the documents into sections and individual sentences within the sections.
2. The lexical analyzer extracts semantic and syntactic features of words with use of a specific lexicon.
3. The parser determines the modifier-head relations between words in a sentence.
4. The semantic interpreter interprets the links of the parser-generated dependency diagram and outputs a set of logical relations.
5. The frame constructor integrates the individual logical relations into structured frames.

A more recent and similar work to this project has been presented by Coden et al to automatically instantiate a knowledge representation model from free-text pathology reports (Coden et al., 2009). They introduced Medical Text Analysis System/Pathology (MedTAS/P) system that was based on an open-source framework and used NLP principles including machine learning and rules to discover and populate elements of the Cancer Disease Knowledge Representation Model (CDKRM). CDKRM is like a structured template in this project, storing cancer characteristics and their relations. Each node in the model is referred to as a class and each class can have multiple attributes. There are two types of classes: leaf classes and container classes. Leaf classes are defined as classes whose attributes are only values. The model has five leaf classes to describe cancer characteristics: anatomical site, histology, grade value, dimension and stage. Container classes are those whose attributes can be either values or other classes. For instance, a tumour class can contain multiple instances of tumour reading classes to capture the notion of multiple interpretations of the same tissue sample.

The pipeline of MedTAS/P could be broken into several components:

1. Ingestion: to extract implicit meaning from the structure of a document.
2. General NLP: to perform tokenization, sentence boundary detection, POS tagging and shallow parsing.
3. Concept finding: to determine concepts based on the International Classification of Diseases for Oncology (ICD-O) (Fritz, 2000) and determine negation.
4. Cancer-specific annotation: to annotate grade, stage, size, margin, date and tumour blocks.
5. Relation finding: to populate CDKRM and resolve co-referent relations.

Particularly, the concept recognition is handled by ConceptMapper and ConceptFilter. ConceptMapper maps the texts to the ICD-O to create candidate matches. ConceptFilter filters out the matches based on a set of rules.

Regular expressions are used to discover entities describing dimensions and sizes, dates, number of excised and positive lymph nodes and stage. Pattern matching is used to identify instances of the grade value class. It also discovers concepts by building machine-learning models.

RelationFinder extracts the relationships between the appropriate leaf classes to populate container classes. First, it determines which section should be considered for instantiating a container class. Second, certain classes are categorized according to multiple criteria. Third, it identifies co-referent instances. Fourth, it determines which instances of leaf classes to be populated to the container classes. Fifth, container classes are merged or split according to specific rules.

MedTAS/P achieved F-scores of 0.97–1.0 for most classes such as histologies or anatomical sites, 0.82–0.93 for primary tumors or lymph nodes, and 0.65 for metastatic tumors. The lower score for metastatic tumors is mainly due to two factors:

- There are relatively few metastatic tumor instances in the reports.
- Metastatic tumor class contains one more leaf class than primary tumor class. Since the correct population of a container class requires all members in the class must match the gold-standard, the additional leaf class greatly decreases the chance for concordance but increases the possibility for disagreement instead.

Except for the limitation mentioned above, another issue is noted in their study. Pathology reports have their own conventions and styles, which should be taken into account when adapting other NLP tools that originate in the general domain to the pathology domain. Although the grammar for general English for the shallow parser had been modified for pathology reports in their study, certain out-of-vocabulary words were still mislabelled. For instance ‘‘nodes” was labelled as a verb instead of a noun in the context of ‘‘lymph nodes”. Such a wrong POS tag can consequently cause the incorrect determination of context for a certain term or concept.

2.5.1 Structured Template

It can be seen from the above literature that there are three feasible ways to construct a template to present the extracted information in a structured format:

1. Predefine case-oriented templates for different diseases according to some standard or consensus reporting conventions of the diseases (e.g., the College of American Pathologists’ surgical lung resection cancer checklist for MEDTEX (Nguyen et al., 2010)).
2. Use appropriate frames to bundle all the logical relations that were found in the previous processes. Each frame represents knowledge discovery about a specific topic, together with descriptions of associated properties. For instance, there were three classes of topics prepared for frame construction in Taira et al’s system: abnormal findings, anatomy, and medical procedures, each with 11, 4 and 3 types of properties respectively (Taira et al., 2001).

3. Build a hierarchical knowledge representation model to store the entities and their relations. For example, low level concepts are represented by leaf classes, and high level ones are represented by container classes in CDKRM of Coden et al's work (Coden et al., 2009).

Each way has its own advantages and disadvantages:

- Predefined case-oriented templates can reveal important pathological features for a specific disease, distinct from other diseases, but it requires that a standard consensus for this disease is available;
- Structured frames are ad hoc, based on the logical relations that can be parsed, however, the topics or properties they represent may be too general to satisfy a pathologist's requirements for a particular disease;
- The correlations among each class are very clear and comprehensible in a knowledge representation model, as the classes are arranged in a hierarchy, whereas the errors in the leaf classes can also propagate to the associated container classes.

Given the considerations above, the author decided that structured templates in this project were established based on three structured cancer reporting protocols from the RCPA, where each field could be represented by a type of entity, a combination of several types of entity, or relationships among the entities, and population of them is separate, thus a decision made on one field would not affect decisions on other fields. Specifically, the three protocols are Primary Cutaneous Melanoma Structured Reporting Protocol (Scolyer et al., 2010), Colorectal Cancer Structured Reporting Protocol (Eckstein et al., 2010) and Tumours of Haematopoietic and Lymphoid Tissue Structured Reporting Protocol (Norris et al., 2010). These protocols contain standards and guidelines for the preparation of structured reports for these three types of cancer. They contain information from multiple international publications and datasets, and they have been developed in consultation with local practicing pathologists, oncologists, surgeons, radiologists and interested national bodies. They provide the frameworks for the reporting of these three types of cancer, whether as minimum data sets or fully comprehensive reports.

2.6 Conclusion

Diverse and substantial work on information extraction (IE) has been reviewed in this chapter. Four main topics are involved: medical entity recognition (MER), negation and uncertainty detection, relation extraction (RE) and automatic structuring. Two main streams are proposed for MER, negation and uncertainty detection and RE: rule-based approaches and statistical methods.

2.6.1 Medical Entity Recognition

On the one hand, rule-based approaches tend to provide reliable results with a relatively small amount of training data, and the hand-crafted rules ease error-analysis for the developers or domain experts,

however, they face difficulty when dealing with unfamiliar or erroneous input data, and the hand-crafted rules may not be reusable. On the other hand, statistical methods are better at coping with problematic data by learning from available examples through training. As well, statistical methods can achieve comparable or better performance by simply adjusting features, hence they are portable to other domains. However, statistical methods usually require a large and so expensive gold-standard for training. Therefore, a better solution is to use hybrid approaches with a combination of statistical methods and rule-based approaches when designing an IE system for MER.

2.6.2 Negation and Uncertainty Detection

Rule-based approaches can be sub-classified into lexical pattern matching methods and syntax-based approaches depending on whether they utilize syntactic information in the texts. Lexical pattern matching methods are efficient and simple, but they are not able to handle longer or complex sentences, or sentences with contradictory statements. Syntax-based approaches are more generalizable and easier to port to another domain, but their performance is greatly hindered by the limitations of the syntactic parsers.

Since some studies pointed out that rule-based methods outperformed the machine learning-based approach, while other research suggested that machine learning-based systems have better performance than their rule-based counterparts, both approaches are attempted in this research.

2.6.3 Relation Extraction

Typical approaches to relationship extraction in most early works in the clinical domain usually were rule-based and relied on full parses, domain-specific grammars, or large domain knowledge bases. Statistical methods have become more and more popular in recent years. Feature engineering is of great importance for a machine learning-based system. Augmentation with handcraft rules and down-sampling the larger classes are two possible ways to improve the performance for recognizing the less prevalent relation types.

2.6.4 Automatic Structuring

There are three feasible ways to construct a template to present the extracted information in a structured format:

1. Predefine case-oriented templates for different diseases according to some standard or consensus reporting conventions of the diseases.
2. Use appropriate frames to bundle all the logical relations that were found in the previous processes.
3. Build a hierarchical knowledge representation model to store the entities and their relations.

Each method has its own advantages and disadvantages. Structured templates in this project were drawn from three structured cancer-reporting protocols issued by the RCPA. The population of each field in the templates could be a type of entity, a combination of several types of entity, or relationships among the entities.

Chapter 3 Corpus Analysis

A detailed corpus analysis was conducted on the three corpora in this study, with the following two objectives:

- To identify lexical and linguistic characteristics in the pathology narratives, and address the difficulties or challenges that may be encountered when processing these texts.
- To assess the level of completeness of the original reports, and propose appropriate strategies for their conversion to structured templates.

3.1 Corpus Overview

The study protocol was approved by Royal Prince Alfred Hospital (RPAH), Sydney, Australia and the Royal College of Pathologists of Australia (RCPA).

The melanoma corpus consists of 477 prose pathology reports of primary cutaneous melanomas from patients referred to the Sydney Melanoma Unit at the RPAH in 2002; there are 612 free-text colorectal cancer pathology reports collected from the RCPA's members serviced in 2011 which constitute the colorectal cancer corpus; the lymphoma corpus is composed of 284 narrative pathology reports of lymphomas from patients serviced in the Anatomical Pathology Department at the RPAH from 2004 to 2008. Most of the reports are from Australia, only 20 free-text colorectal cancer pathology reports come from other countries or regions (e.g., Malaysia, Hong Kong, New Zealand, South Africa, Namibia, and UAE). They were scanned and optical character recognized (OCR-ed). The melanoma corpus and lymphoma corpus were de-identified, while the colorectal cancer corpus was not de-identified as the personal information was reserved for other projects.

Corpus	No. of training set documents	No. of test set documents
Melanoma corpus	380	97
Colorectal cancer corpus	397	215
Lymphoma corpus	277	57

Table 3.1 Distribution of training sets and test sets on each corpus.

The three corpora were divided into training sets and test sets by random selection. Table 3.1 outlines the distribution of training sets and test sets on each corpus.

The following analyses were all carried out on the training data, thus none of the information from the test data would be compromised in the training stage. This ensured the integrity and reliability of the system performance in the test stage.

3.2 Lexical Analysis

The lexical analysis was performed on each token in the training data, and the tokenizer used was a white space based tokenizer. Each token was separated by white space, unless there is a punctuation mark (e.g., full stop (.), comma (,), semicolon (;), and colon (:)) at the start or end of the token, which was separated from the token as well. Here is an example:

There is epidermal invasion (no ulceration), and several foci of papillary dermal invasion (level II, depth 0.45mm - block E).

can be tokenised to

There is epidermal invasion (no ulceration) , and several foci of papillary dermal invasion (level II , depth 0.45mm - block E) .

The basic token statistics of the three corpora are tabulated in Table 3.2. The melanoma corpus had the smallest number of overall tokens (only 71786), and the smallest average count for each note (less than 190), and smallest number of unique case sensitive tokens and unique case insensitive tokens, which are 4801 and 3783 respectively. The colorectal cancer corpus had the largest number of overall tokens (up to 224660), and the largest average count for each note (more than 565), and the largest number of unique case sensitive tokens and unique case insensitive tokens, which were 11072 and 9077 respectively. The statistics of the lymphoma corpus were between the above two corpora. The number of overall tokens, average count for each note, number of unique case sensitive tokens, and unique case insensitive tokens were 113413, 409.4, 7127 and 5919 respectively.

	Melanoma corpus	Colorectal cancer corpus	Lymphoma corpus
Total No.	71786	224660	113413
No. of unique case sensitive tokens	4801	11072	7127
No. of unique case insensitive tokens	3783	9077	5919
No. of alphabetic words	2263	4351	3513
No. of non-alphabetic tokens	1520	4726	2406

Table 3.2 Basic token statistics of each corpus.

The possible reason for these statistics is that the colorectal cancer corpus has the largest number of reports and most reports with considerably lengthy texts; the lymphoma corpus had the smallest number of reports, but each report was complete without missing contents, and there were more sections in some reports than those in the melanoma corpus, although it also had a larger amount of reports than the lymphoma corpus, where some of them were incomplete and most of them with shorter length texts than those in the lymphoma corpus.

The tokens could be classified into two main categories: alphabetic word and non-alphabetic token. Alphabetic words are tokens that only consist of alphabetic letters, while non-alphabetic tokens are those which contain digits and punctuation marks other than alphabetic letters. For the melanoma

corpus, in the unique case insensitive tokens, there were 2263 tokens of alphabetic words and 1520 were non-alphabetic tokens; for the colorectal cancer corpus, more than half were non-alphabetic tokens, the amount of which was larger than that of alphabetic words with 375; for the lymphoma corpus, the amount of alphabetic words exceeded that of non-alphabetic tokens by 1107.

3.2.1 Alphabetic Words

The alphabetic words were verified against a dictionary, constructed by the union of three lexical resources: MOBY, SNOMED CT and UMLS. SNOMED CT and UMLS are two standard terminologies, described previously in Chapter 2. MOBY's thesaurus (Ward, 1996-2000) was released as part of the MOBY lexicon project by Grady Ward in June 1996. It contains a single large synonym list for each headword and ordered alphabetically. However, it is an American English thesaurus, hence it is not able to recognize variant spelling of certain words in Australian English. The dictionary comprises 354992 lexical entries from MOBY, 99860 from SNOMED CT, 427578 from UMLS. Not only the word itself but also its base form (lemma) is verified against the dictionary, if a match of the word cannot be found in the dictionary. Since the workload for manual lemmatization would be very heavy, lemmatization is performed automatically by the GENIA tagger (Tsuruoka et al., 2005).
The frequencies of alphabetic words and the distributions in each lexical resource are listed in Table 3.3.

Lexical resource	Melanoma corpus		Colorectal cancer corpus		Lymphoma corpus	
	No. of tokens	No. of unique words	No. of tokens	No. of unique words	No. of tokens	No. of unique words
SNOMED CT	55562	1930	162637	3186	73664	2928
UMLS	55623	2112	162022	3581	74368	3219
MOBY	55852	2079	162368	3449	73557	3058
Any of three	56776	2193	166234	3760	75805	3315
None of three	120	70	1665	591	1368	198

Table 3.3 Frequencies of alphabetic words and the distributions in each lexical resource.

The results show that the dictionary captures most of the alphabetic words and tokens in the corpora. The highest coverage of tokens from the lexical resources varies between each corpus, which is 77.8% from MOBY for the melanoma corpus, 72.4% from SNOMED CT for the colorectal cancer corpus, and 65.6% from UMLS for the lymphoma corpus. However, UMLS captures the greatest proportion of alphabetic words in each corpus, accounting for 93.3% for the melanoma corpus, 82.3% for the colorectal cancer corpus, and 91.6% for the lymphoma corpus, as it is the largest dictionary.

It can be seen from the results that there are more tokens that are general English words in the melanoma corpus than the other two corpora, thus MOBY, as a general English lexicon, captures the largest ratio of tokens. The lowest coverage of alphabetic words from the dictionary is in the colorectal cancer corpus (86.5%) suggesting that there are more unknown words needing to be

resolved than in the other two corpora; the dictionary captures the smallest proportion of the tokens in the lymphoma corpus (66.8%), indicating that there are considerable proportions of non-alphabetic tokens and unknown words in that corpus.

About 97% of the alphabetic words in the melanoma corpus have been recognized in any of three lexical resources, with about 3.1% of the words being unknown. There are about 86% of alphabetic words in the colorectal cancer corpus identified in any of three lexical resources, leaving 13.6% words unknown. At least one match has been found in any of three lexical resources for about 94% of the alphabetic words in the lymphoma corpus, with no match for the remaining 5.6% words.

Unknown category	Melanoma corpus	Colorectal cancer corpus	Lymphoma corpus
Correct words	angioplasia, guantitation, traumatised, angiofibroplasia, lymphovascular	albicantia, biopsied, oedematous, nonperitonealised, mesocolonic	alkomas, angiotropism, immunoperoxidases, squamoproliferative, macrosteatosis
Abbreviations	amm, iescc, ipx, wepc, snb	emr, fhx, trg, lvi, drm	bll, dlbcl, nlphl, tblb, tjlb, faa
Shorthand	histopath, immunohisto, sebk	revd	btw, chemoth, lge, wrk, exc
Misspelling	albow, clikical, diagosis, dimention, kxcision	absen, abdominoperitinial, circumferance, ceacum, ccomments	aaplastic, agessive, architectrue, centrablasts, concensus,
Missing space	havepleomorphic, macronucleoli, midcalf, sqmm	aminor, columnarepithelium, furthertests, ofextracellular, predominantlylymphocytic	datetime, newbone, antibodydescriptionresult, lambdapositive
Named entities	darlinghurst, gosford, iml	albury, alexy, crgh, mandard, sswahs	bayfield, dutcher, fuhrmann, ivac, rnsh, temno
Complex	pjanchjbiopsy	nomx, tubulocribriform, tubuloadenoma	ileoresection

Table 3.4 Examples of unknown words.

The unknown words were analysed and manually resolved by a medical expert. They can be divided into seven categories: correct words, abbreviations, shorthand, misspelling, missing space, named entity and complex. Limited coverage of the resources was the main reason that the correct words could not be identified by the lexical resources (e.g., “lymphovascular” is a frequently used domain-specific word that is not recognized by the dictionary), variant spelling (e.g., “nonperitonealised” *vs.* “nonperitonealized”), and lemmatization errors from the tagger (e.g., “biopsied” is lemmatized as “biopsie”. Abbreviations and shorthand are words presented in a compact form, deliberately used by pathologists under time pressure. Misspelling and missing space can be caused by typing or errors from the OCR. Named entities are proper names such as geographic gazetteers (e.g., “darlinghurst” and “albury” are two places), names of institutions (e.g., “rnsh” stands for the name of a hospital: “Royal North Shore Hospital”), names of people (e.g., “alexy” and “bayfield”), medical named

entities (e.g., “temno” is a band of biopsy needle). The complex category is a combination of the above categories (e.g., the error of “pjanchjbiopsy” results from misspelling and missing space). Table 3.4 displays some examples of the unknown words.

The biggest contributions of unknown words are from correct words, abbreviations, misspelling and named entities. The method proposed by Patrick et al (Patrick et al., 2010) was applied to misspelling correction, which is a combination of a rule-based suggestion generation system and a context-sensitive ranking algorithm based on word frequencies and trigram probabilities. The results from the misspelling corrector were manually verified by the medical expert. Table 3.5 shows some results from misspelling correction. The medical expert also tried to resolve words in abbreviations, shorthand, missing space, and complex categories as well.

Unknown category	Original word	Correction
Misspelling	architectrue	architecture
Missing space	aminor	a minor
Complex	pjanchjbiopsy	punch biopsy

Table 3.5 Examples of misspelling correction.

After misspelling correction and manual verification, the unique unknown word size shrank to 20 for the melanoma corpus, 117 for the colorectal cancer corpus, and 25 for the lymphoma corpus.

Melanoma corpus		Colorectal cancer corpus		Lymphoma corpus	
Noun	Frequency	Noun	Frequency	Noun	Frequency
lesion	945	tumour	4734	cells	1547
melanoma	902	lymph	2164	procedure	1465
skin	728	nodes	2064	lymphoma	1289
specimen	539	margin	1999	node	1188
margin	448	resection	1951	lymph	1055
level	445	invasion	1395	t	694
dermis	361	specimen	1243	cell	612
sections	359	colon	1188	tissue	581
tumour	329	bowel	1136	b	578
invasion	323	adenocarcinoma	1018	specimen	483
ellipse	314	margins	876	nodes	426
microscopic	311	fat	868	tumour	384
mm	295	node	798	nk	383
cells	295	sections	780	biopsy	358
thickness, excision	282	surface	645	cd20	355
component	267	muscularis	623	cd3	286
report	237	tissue	613	sections	283
clark	236	length	606	flow	267
melanocytes	231	propria	602	report	255
surface, depth	215	mucosa	575	codes	245
histopathology	198	diameter	573	description, cd5, cd10	240

Table 3.6 Twenty most common nouns and their frequencies in each corpus.

Nouns are one of major elements of a medical entity. The set of nouns in the corpora were identified using the GENIA tagger (Tsuruoka et al., 2005) according to their part-of-speech (POS) tags. They could be singular, plural and proper nouns, and those converted for anonymization were filtered out from the results. Table 3.6 lists the twenty most common nouns and their frequencies in each corpus.

From Table 3.6, we can see that “tumour”, “specimen” and “sections” are frequently used in all three corpora, due to the nature of their genre: pathology reports of specific cancers. Different nouns are used to represent the major diagnoses in each corpus: “melanoma” for the melanoma corpus, “adenocarcinoma” for the colorectal cancer corpus, “lymphoma” for the lymphoma corpus. Various nouns are used to describe the usual sites or locations of the tumour(s) or specimen(s) from patients: “skin” for the melanoma corpus, “colon” and “bowel” for the colorectal cancer corpus, “lymph”, “node” and “nodes” for the lymphoma corpus. There are also some nouns related to the characteristics of the corpus. For example, for the melanoma corpus, “ellipse” and “excision” are two main specimen types; “melanocytes” is the primary cell type of melanoma. For the colorectal cancer corpus, “mucosa” and “muscularis propria” are two distinct layers of the bowel; “length” and “diameter” are two frequently used measurements to describe tumour or specimen sizes. For the lymphoma corpus, “t”, “b” and “nk” are descriptors of lineage; “cd20”, “cd3”, “cd5”, and “cd10” are biomarkers commonly employed in immunohistochemistry tests or flow cytometry.

3.2.2 Non-alphabetic Tokens

The non-alphabetic tokens are further categorized according to their orthographic features, including single punctuation, multiple punctuation, numeric values, dimension, alphanumeric, with slash, with hyphen, with apostrophe, with question mark, percentage, and other forms. The descriptions and some examples for each category are presented in Table 3.7.

There are various meanings for these tokens. Single and multiple punctuations are usually separators or indicators. For example, “.” is the most frequently used for sentence delimitation; “****” separates the paragraphs; “+”, “++” and “+++” indicate the severity or intensity. Dimension tokens such as “20mm”, “30x7mm” and “3x3x2mm” represent one-, two- and three-dimensional size respectively. Alphanumeric tokens like “1a” and “1e” are specimen block identifiers; “cd3” and “ck20” are biomarkers; “pn1” represents N staging information. The question mark in the beginning of the token, usually stands for “suspicious for” or “maybe”. Nonetheless, other punctuation symbols can have polysemous functions in the tokens, such as slash, hyphen, and apostrophe. An initial survey shows that there are up to 3 different functions for apostrophe, 7 for slash and 10 for hyphen. Some examples are displayed in Table 3.8. Undoubtedly, polysemia of punctuation in the tokens increases the difficulty in both tokenisation and disambiguation. Furthermore, the patterns in some categories are of great significance for recognizing potential entities (e.g., tokens in dimension category can compose entities about specimen or tumour sizes). Therefore, it requires more sophisticated strategies in tokenisation, disambiguation and pattern recognition to tackle these tokens.

Non-alphabetical token category	Description	Example		
		Melanoma corpus	Colorectal cancer corpus	Lymphoma corpus
Single punctuation	Token is punctuation such as period, comma, colon and bracket.	.; ,; (;)	#; *; @	/; &; %
Multiple punctuation	Token consists of multiple punctuations.	->; ++;…	****; -->; ----------	+/-; +-++
Numeric values	Token is a digit or decimal number.	27; 3.1	0.3; 1993; 24	16.0; 2000; 23
Dimension	Token describes a size or dimension value.	20mm; 30x7mm; 3x3x2mm	1.1mm; 10x10x2mm; 15cm; 7x6x3; 80x30mm	13.8cm; 45x40x25mm; 0.2mm
Alphanumeric	Token contains both numbers and alphabetic letters.	1a; f36; hmb45	1e; 2xdonuts; pn1; msh2	1780g; 2mths; cd3; ck20
With slash	Token contains slash (/).	22/02/01; 3/mm2; white/pink	0/20; 04/08/11; a/prof; ascending/transverse	17/12/2004; ae1/ae3; b/g; kappa/lambda
With hyphen	Token contains hyphen (-).	1b-1c; -1; ii-iii; band-like	01-jan-1999; acps-a; 2-3mm; chemo-radiotherapy; well-clear	15-20cm; b-cell; centrocyte-like; intra-abdominal; ki-67; m-00100; cd10-
With apostrophe	Token contains apostrophe (').	breslow's; o'clock; hutchinson's	carnoy's; crohn's; duke's	burkitt's; bx's; hodgkin's; tumour's
With question mark	Token contains question mark (?).	??melanoma; ?hmf	?adenoma; ?perforated; ?ulcer	??lymphoma; ?malt; ?transformation
Percentage	Token contains digits and %.	5%; >5%	20%; >10%; <3%	100%; 15-25%; ~95%
Other	Token cannot be categorized above.	20+; h.d.f; 13.2.01	03:48pm; 1,2; 1e&1f; margins:15mm	+ve; 5:18; <5mm;

Table 3.7 Descriptions and some examples for each non-alphabetic token category.

Punctuation	Function	Example
Apostrophe	Shorthand	bx's
	Of	tumour's
	Term named after someone	hodgkin's
Slash	Or	ascending/transverse
	Divide	kappa/lambda
	Ratio	0/20
	Per	3/mm2
	Abbreviation	b/g
	Date	17/12/2004
	Mixture	ae1/ae3
Hyphen	Identifier	1b-1c
	Range	15-20cm
	Prefix	intra-abdominal

	Joined words	well-clear
	Date	01-jan-1999
	Colon (:)	acps-a
	Biomarker	ki-67
	Code	m-00100
	Listing	-1
	Negative	cd10-

Table 3.8 Multiple functions and examples of apostrophe, slash and hyphen.

3.3 Language Phenomena in Pathology Reports

After the detailed lexical analysis, the following language phenomena identified in the pathology notes may be barriers to further processing. They are summarized as follows:

Unknown words: Pathology notes contain more unknown words than newswire documents. As demonstrated in the previous section, the unknown word rate is very high when only using a general-purpose dictionary like MOBY. Even with the combination of medical standard terminologies like SNOMED CT and UMLS, there are still a considerable number of unknown words. These unknown words are an obstacle for the application of dictionary look-up approaches.

Abbreviations: Abbreviations are prevalent in pathology notes. Some abbreviations have standard forms and naming conventions, but most of them do not, which makes the expansion of them to full terms quite difficult. The abbreviations can be divided into three categories: abbreviation (e.g., opening letter initialization and syllabic initialization), acronym (letter capitalisation) and shorthand (including end truncation and syllabic contraction). Some examples of abbreviations with their expansions are listed in Table 3.9.

Misspellings: Misspellings are mainly caused by typing errors, such as keyboard incompetence; another contribution is from non-native English speaking staff, who are more likely to miss syllables, substitute syllables and repeat syllables in the words when writing the reports. Most of the misspellings can be corrected using a misspelling corrector, but some complex ones also require manual verification.

Abbreviation category	Original form	Expansion
Abbreviation	FHx	family history
Acronym	SNB	sentinel node biopsy
Shorthand	btw	between

Table 3.9 Examples of abbreviations with their expansions.

Non-alphabetic tokens: The non-alphabetic tokens make up a great proportion of the overall tokens. Some of them may represent special meanings, which should be discriminated from those for layout and formatting guidance. As well, punctuation such as hyphen, apostrophe and slash can have

multiple functions in different contexts. Disambiguating the tokens with these punctuations and capturing the patterns in the tokens is of significant importance in recognizing potential entities containing them.

Lexical variants: Lexical variants are usually created by productive morphology and stylistic writing. Firstly, staff in the pathology laboratories is from a variety of countries or regions, and they have their preferred spelling. Secondly, pathologists tend to create their own ad hoc forms of frequently used phrases and sometimes these terms may not follow any naming conventions and can be rarely found in the standard terminologies. These created terms are personally idiosyncratic, or represent a local community accepted de facto standard. For example, lots of lexical variants are observed in the expressions of abdominoperineal resection: abdo peri resection, abdo-perineal resection, AP resection, abdomino-perineal resection, abdominoperineal resection, abdominal perineal resection and APR (NB. Misspellings like abdo-peritoneal resection, abdominal perineural resection and abdomino-peritoneal resection are not included).

Complex vocabulary: Pathology notes have more complex vocabulary than the texts in the general domain, which is mainly due to the prevalence of abbreviations, unknown words, misspellings, non-alphabetic tokens and lexical variants. Such complex vocabulary forms a special sub-language in the clinical domain. Fully understanding the divergence between this sub-language and the common language is critical to adopting suitable natural language processing (NLP) techniques to process the notes.

3.4 Completeness Analysis

One of the main objectives of the project is to help pathologists to validate their reports and improve the accuracy and completeness of them. At first, it is necessary to evaluate the quality of the original narrative reports, thus, as the most important indicator – completeness was analysed on each corpus to achieve this goal.

Completeness is reported below as quantitative measures of adherence to the *standards* and *guidelines* in the structured reporting protocols (Eckstein et al., 2010; Norris et al., 2010; Scolyer et al., 2010). According to the protocols, *standards are defined as mandatory fields*, reserved for core items essential for the clinical management, staging or prognosis of the cancer; *guidelines are defined as recommended fields*, covering items that are not essential for clinical management, staging or prognosis of a cancer, but are recommended. The following statistics do not include the measure for all standards or guidelines in the protocols because:

1. Fields that involve personal information were not reported. For example, in Colorectal Cancer Structured Reporting Protocol: “G1.01 The patient’s health identifiers should be recorded where provided.” and “S1.02 The principal clinician involved in the patient’s care

and responsible for investigating the patient must be identified.", as such information has already been removed in de-identification or out of the scope of this study.

2. Fields without associated information in the original reports were ruled out. For instance, in Primary Cutaneous Melanoma Structured Reporting Protocol, for "S5.01 The AJCC melanoma tumour–node (pTN) subcategories according to the current AJCC staging system must be recorded.", there is no associated information in the melanoma corpus.
3. Fields that are recommendations for clinical staff or pathologists to deliver or process the specimens rather than record the information about the specimens, which are too complicated or ambiguous to compute, were not included, such as in Tumours of Haematopoietic and Lymphoid Tissue Structured Reporting Protocol, "S1.05 Where lymphoma is suspected, the specimen must be sent immediately, intact and unfixed in a closed sterile container to the anatomical pathology laboratory."
4. Fields that are not presented in the structured report examples of the protocols were excluded. An example is in Primary Cutaneous Melanoma Structured Reporting Protocol, "S2.01 The tissue block(s) must be selected to facilitate microscopic assessment of the thickest or most suspicious portion of the tumour, and determination of the relationship of the tumour to the surgical margins."

Standards	**Abbreviation**	**No. of documents provided**	**Percentage provided**
S2.02 The specimen must be described.	Specimen description	361	95.00%
S2.03 The specimen dimensions must be measured and recorded.	Specimen dimensions	360	94.74%
S2.05 The primary lesion must be described.	Primary lesion	312	82.11%
S2.06 The presence of other lesions must be noted, and their features recorded.	Other lesions	14	3.68%
S3.01 The diagnosis of primary melanoma must be recorded.	Diagnosis	317	83.42%
S3.02 The Breslow thickness must be recorded.	Breslow thickness	309	81.32%
S3.03 The pathology report must indicate whether or not the invasive or in situ melanoma involves the surgical margins.	Margin involvement	181	47.63%
S3.04 The presence or absence of ulceration must be reported.	Ulceration	210	55.26%
S3.05 The mitotic rate per square millimetre of the invasive melanoma must be recorded.	Mitotic rate	123	33.68%
S3.06 The presence or absence of microsatellites must be recorded.	Microsatellites	24	6.32%
S5.02 The pathology report must include a field for free text in which the reporting pathologist can give overarching case comment if required.	Comment	23	6.05%

Table 3.10 Completeness measures of *standards* on the melanoma corpus.

Guidelines	Abbreviation	No. of documents provided	Percentage provided
G1.03 The anatomical site of the melanoma should be recorded; G1.04 The laterality of the melanoma should be recorded.	Site and laterality	248	65.26%
G1.05 The clinical diagnosis or differential diagnosis should be recorded.	Clinical diagnosis	170	44.74%
G1.06 The description of the type of specimen should be recorded.	Specimen type	67	17.63%
G1.08 The history and timing of lesional trauma, biopsy, irritation or treatment with topical agent should be recorded.	Lesional trauma	8	2.11%
G1.09 A history of previous primary melanoma, at this or any other site, should be recorded.	Previous melanoma	5	1.32%
G1.10 Evidence of metastatic disease should be recorded.	Metastatic disease	1	0.26%
G1.12 Other relevant history should be recorded.	Other relevant history	86	22.63%
G3.02 The pathology report should document the distance of invasive and in situ melanoma from peripheral and deep margins.	Margin distance	15	3.95%
G3.04 The level of invasion (Clark) should be recorded.	Clark level	280	73.68%
G3.05 The presence or absence of lymphovascular invasion should be recorded.	Lymphovascular invasion	204	53.68%
G3.06 The distribution and density of tumour-infiltrating lymphocytes (TILs) should be recorded.	TILs	3	0.79%
G3.07 The presence or absence of intermediate or late regression should be recorded.	Regression	47	12.37%
G3.08 The absence or presence and extent of desmoplasia (% of invasive component) should be recorded.	Desmoplasia	14	3.68%
G3.09 The presence or absence of neurotropism should be recorded.	Neurotropism	132	34.74%
G3.10 Any associated benign melanocytic lesion should be recorded.	Associated benign lesion	112	29.47%
G3.11 The intra-epidermal growth pattern of the melanoma should be recorded.	Growth pattern	15	3.95%
G3.12 The subtype of melanoma should be recorded.	Subtype	225	59.21%

Table 3.11 Completeness measures of *guidelines* on the melanoma corpus.

Melanoma corpus: The completeness measures of standards and guidelines are displayed in Table 3.10 and Table 3.11 respectively. For standards, the coverage in about half of the fields reaches to more than 50%, wherein “S2.02 Specimen description”, and “S2.03 Specimen dimensions” are mentioned in most reports (over 94%), while “S2.06 Other lesions”, “S3.06 Microsatellites”, and “S5.02 Comment” are seldom referred to (lower than 7%).

For guidelines, the coverage in less than a quarter of the fields exceeds 50%, the most frequently mentioned field is “G3.04 Clark level” with around 74%, and up to 7 fields are barely described (less than 4%), including “G1.08 Lesional trauma”, “G1.09 Previous melanoma”, “G1.10 Metastatic disease”, “G3.02 Margin distance”, “G3.06 TILs”, “G3.08 Desmoplasia”, and “G3.11 Growth pattern”.

Colorectal cancer corpus: The completeness measures of standards and guidelines are shown in Table 3.12 and Table 3.13 respectively. For standards, 15 out of 35 fields have more than 50% coverage, which is close to 100% in “S2.04 Specimen length” and “ S2.12 Macroscopic information”, but declines to less than 4% in 5 other fields: “S1.04 Presentation”, “S1.06 Distance from anal verge”, “S1.09 Local residual cancer”, “S1.10 Adjacent organ involvement” and “S2.11 Mesorectum intactness”, wherein only one case stated “S1.09 Local residual cancer”; for guidelines, most fields have poor coverage (less than 24%), and “G4.02 KRAS mutation testing” is only stated in two cases.

Lymphoma corpus: The completeness measures of standards and guidelines are tabulated in Table 3.14 and Table 3.15 respectively. For standards, 75% of fields have more than 50% coverage, and all documents describe “S2.04 Specimen size”, while “S2.01 Fluid” has the poorest coverage with 28.19%; for guidelines, 6 out of 16 fields are referred to in more than 50% of documents, while “G5.04 Stage” is only mentioned in 2 documents, “G1.05 Disease spread” and “G1.06 Extent of disease” are also rarely mentioned (in 8 and 7 documents respectively).

Standards	Abbreviation	No. of documents provided	Percentage provided
S1.04 Patient presentation at surgery must be recorded, in particular whether perforation is present.	Presentation	11	2.77%
S1.05 The tumour location must be recorded.	Tumour location	288	72.54%
S1.06 The distance from the anal verge must be recorded (for rectal tumours only).	Distance from anal verge	13	3.27%
S1.07 The type of operation performed must be recorded.	Operation type	282	71.03%
S1.08 If pre-operative radiotherapy has been administered, this must be recorded.	Pre-operative radiotherapy	27	6.80%
S1.09 The surgeon's opinion on the existence of local residual cancer following the operative procedure must be recorded.	Local residual cancer	1	0.25%
S1.10 The involvement of adjacent organs must be recorded.	Adjacent organ involvement	13	3.27%
S1.11 The presence of any distant metastases must be recorded.	Distant metastases (Clinical)	36	9.07%
S2.02 The nature and sites of all blocks must be recorded.	Blocks	390	98.24%
S2.03 All regional lymph nodes must be harvested from the specimen and examined histologically.	Lymph nodes	233	58.69%
S2.04 The specimen length must be recorded.	Specimen length	396	99.75%
S2.05 The site of the tumour must be recorded.	Tumour site	238	59.95%
S2.06 The maximum tumour diameter must be recorded.	Tumour diameter	383	96.47%
S2.07 The distance of the tumour to the nearer proximal or distal 'cut end' margin must be recorded.	Distance to proximal/distal margin	355	89.42%
S2.08 The distance of the tumour to the circumferential margin must be recorded.	Distance to circumferential margin	77	19.40%
S2.09 The presence or absence of tumour perforation must be recorded.	Perforation	38	9.57%
S2.10 For rectal tumours the relationship of the tumour to the anterior peritoneal reflection must be recorded.	Relationship to anterior peritoneal reflection	33	8.31%
S2.11 For rectal resections the intactness of the mesorectum must be recorded.	Mesorectum intactness	12	3.02%
S2.12 A descriptive or narrative field must be provided to record any macroscopic information that is not recorded in the above standards and guidelines, and that would normally form part of the macroscopic description.	Macroscopic information	396	99.75%
S3.01 The tumour type must be recorded.	Tumour type	301	75.82%
S3.02 The histological grading of the tumour must be recorded.	Histological grading	257	64.74%
S3.03 The maximum degree of local invasion into or through the bowel wall must be recorded.	Local invasion	330	83.12%
S3.04 Involvement of the proximal or distal resection margins ('cut-end' margins) by tumour must be recorded. If the margin is less than 10 mm, the clearance must be recorded.	Involvement of proximal/distal margin	202	50.88%

S3.05 The status of the nonperitonealised circumferential margin in rectal tumours must be recorded.	Status of circumferential margin in rectal tumours	47	11.84%
S3.06 The status of the nonperitonealised circumferential margin in colon tumours must be recorded.	Status of circumferential margin in colon tumours	93	23.43%
S3.07 Results of lymph node histopathology must be recorded.	Lymph node histopathology	333	83.88%
S3.08 For all tumours, venous and small vessel invasion must be reported and its anatomic location specified as mural or extramural.	Venous and small vessel invasion	298	75.06%
S3.09 The presence of histologically confirmed distant metastases and their site must be recorded.	Distant metastases (Microscopic)	72	18.14%
S3.10 The presence of any relevant coexistent pathological abnormalities in the bowel must be recorded.	Coexistent pathological abnormalities	224	56.42%
S3.11 The microscopic residual tumour status must be recorded (i.e., the completeness of resection).	Residual tumour (Microscopic)	31	7.81%
S3.12 The response of the tumour to neoadjuvant treatment must be recorded.	Response to neoadjuvant treatment	49	12.34%
S5.01 The tumour stage and stage grouping must be recorded, incorporating clinical and pathological data, based on the TNM staging system of the AJCC Cancer Staging Manual (7th Edition)	Tumour stage and stage grouping	128	32.24%
S5.02 The residual tumour status must be recorded according to the AJCC Cancer Staging Manual (7th Edition).	Residual tumour (Synthesis)	63	15.87%
S5.03 A field for free text or narrative in which the reporting pathologist can give overarching case comment must be provided.	Comment (Synthesis)	140	35.26%

Table 3.12 Completeness measures of *standards* on the colorectal cancer corpus. Note that the percentage for S1.06 was computed on documents with rectal tumours.

Guidelines	**Abbreviation**	**No. of documents provided**	**Percentage provided**
G1.03 Any additional relevant information should be recorded	Clinical information	66	16.62%
G2.01 Pathologists may be asked to provide tissue samples from fresh specimens for tissue banking or research purposes.	Tissue banking	37	9.32%
G2.02 Images of the gross specimen showing the overall conformation of the tumour and, especially in the case of rectal resections, images showing the relation of the tumour to the resection margins, are desirable, and useful for multidisciplinary meetings.	Specimen imaging	12	3.02%
G3.01 Involvement of the apical lymph node should be recorded, if required where staging systems additional to TNM staging are in use	Involvement of apical lymph node	62	15.62%
G3.02 Perineural invasion should be assessed using routine histology and reported as present or absent	Perineural invasion	176	44.33%
G3.03 Any additional relevant information should be recorded	Microscopic information	221	55.67%
G4.01 Immunohistochemistry tests should be performed to test mismatch repair deficiency status and the results recorded in the pathology report	Immunohistochemistry tests	52	13.10%
G4.02 The result of KRAS mutation testing should be recorded	KRAS mutation testing	2	0.50%
G5.01 The “Diagnostic summary” section of the final formatted report should include: a. specimen type (S1.01) b. tumour site (S2.05) c. tumour type (S3.01) d. tumour stage (S5.01) e. completeness of excision (S5.02)	Elements of The “Diagnostic summary” section	95	23.93%

Table 3.13 Completeness measures of *guidelines* on the colorectal cancer corpus.

Standards	Abbreviation	No. of documents provided	Percentage provided
S1.03 The site of biopsy must be recorded.	Biopsy site	206	90.75%
S1.04 The laterality must be recorded.	Laterality	184	81.06%
S2.01 The fluid in which the specimen is delivered to the laboratory must be reported.	Fluid	64	28.19%
S2.02 Specimen handling or triage must be reported.	Triage	153	67.40%
S2.03 The specimen type must be reported.	Specimen type	134	59.03%
S2.04 The specimen size must be reported.	Specimen size	227	100.00%
S3.02 The grade (for follicular lymphoma) must be reported.	Grade	65	28.63%
S4.01 All ancillary studies which have been performed, and which are pending, must be reported.	Ancillary studies	121	53.30%
S4.02 For ancillary studies performed in the reporting anatomical pathology laboratory (e.g., immunohistochemistry) test results and interpretation must be reported in full, including all positive, negative and indeterminate results.	Ancillary study results and interpretation	121	53.30%
S5.01 Lineage must be reported.	Lineage	154	67.84%
S5.02 The WHO disease subtype must be recorded.	WHO disease subtype	221	97.36%
S5.03 Facility for overall case comment must be provided.	Comment	71	31.28%

Table 3.14 Completeness measures of *standards* on the lymphoma corpus.

There are several important findings from the above results:

1. The completeness of standards is significantly better than that of guidelines, probably because standards are compulsory, while guidelines are suggestions, demonstrating clinical staff and pathologists usually pay more attention to the former.
2. Fields with specific conditions may have poor coverage. For example, rectal resection is the requisite procedure for recoding "S2.11 Mesorectum intactness". If the conditions are not satisfied, pathologists are prone to ignore them when writing the reports.
3. The coverage of fields with regard to the presence or absence of findings is also low (e.g., "G3.08 Desmoplasia"), since if a finding is absent, pathologists often omit to report it.
4. The coverage of fields that require co-occurrence of multiple elements is usually poor as well. For instance, "G3.02 Margin distance" requires four elements to co-occur in the document: the distance of invasive melanoma from peripheral margin, the distance of invasive melanoma from deep margin, the distance of in situ melanoma from the peripheral margin, and the distance of in situ melanoma from deep margin. However, in most documents, only one or some of the elements occur, therefore, they are excluded from the valid count. Similarly, "G3.06 TILs" consists of three elements: TILs, the distribution of TILs and the density of TILs, but most documents do not present all the elements, resulting in the small valid count.

5. The possible reason for the low coverage of several fields is that its definition or requirement is inconsistent with the facts in the documents. For example, "G3.11 Growth pattern" defines the intra-epidermal growth pattern to be recorded; in fact, many documents do refer to cell growth patterns, but they may not be intra-epidermal, but located in some layers of the skin (e.g., dermis) instead.

Guideline	**Abbreviation**	**No. of documents provided**	**Percentage provided**
G1.03 The reason for the biopsy should be recorded.	Reason for biopsy	25	11.01%
G1.04 The clinical diagnosis or differential diagnosis should be recorded.	Clinical diagnosis	129	56.83%
G1.05 Involved sites or pattern of disease spread and whether disease is nodal or extranodal should be recorded if known.	Disease spread	8	3.52%
G1.06 An estimation of stage or extent of disease should be given if possible.	Extent of disease	7	3.08%
G1.07 All relevant constitutional symptoms should be recorded.	Constitutional symptom	16	7.05%
G1.09 Any previous lymphoma, leukaemia or other relevant haematological disease should be recorded.	Previous relevant disease	64	28.19%
G1.10 Any previous relevant treatment should be recorded; G1.11 Predisposing factors such as immunocompromised states (immunodeficiency associated lymphoproliferative disorders) and autoimmune conditions should be recorded; G1.12 Predisposing factors such as infective agents should be recorded.	Predisposing factors	27	11.89%
G3.01 The pattern of infiltration or architecture of abnormal cells should be reported.	Architecture	188	82.82%
G3.02 The size of abnormal cells should be reported.	Cell size	201	88.55%
G3.03 The cytomorphology of abnormal cells should be reported.	Cytomorphology	119	52.42%
G3.05 Host cells and tissue reactions should be reported.	Tissue reactions	123	54.19%
G5.03 The 'Diagnostic summary' section of the final formatted report should include: a. specimen type (S2.03) b. tumour site and laterality (S1.03, S1.04) c. WHO diagnosis (S5.02) d. grade where relevant (S3.02)	Elements of The "Diagnostic summary" section	15	6.61%
G5.04 Stage should be recorded if known.	Stage	2	0.88%
G5.05 A supplementary report (or equivalent) should be added to the pathology report if further diagnostic information is subsequently obtained.	Supplementary report	34	14.98%

Table 3.15 Completeness measures of *guidelines* on the lymphoma corpus.

The issues discussed above should be addressed in the following processes, especially in the construction of the structured templates. Although the fields in the structured templates are supposed to be based on the standards and guidelines, they also need slight adjustments to the corpora. For instance, it is preferable to prepare three fields to depict "G3.06 TILs" in the template: TILs, the distribution of TILs and the density of TILs; broaden the scope of "G3.11 Growth pattern" to include other cell growth patterns to cut down the loss of useful information embedded in the texts.

3.5 Conclusion

This chapter presents detailed lexical and completeness analyses of the corpora.

Analyzing the characteristics of tokens is very important since tokens are the foundations for constructing more complex structures such as phrases and sentences in the documents. The lexical analysis has demonstrated that the texts in pathology notes have specific characteristics which are quite different from other genres of texts. For instance, there is extensive use of biomedical terms or concepts that can be captured by SNOMED CT and UMLS, and, high frequencies for some nouns according to the report types and diseases. The significant language phenomena observed in the corpora including abbreviations, unknown words, misspellings, non-alphabetic tokens, lexical variants and complex vocabulary indicate the difficulties or challenges that may be encountered when processing these texts, which require sophisticated NLP techniques to resolve.

A quantitative completeness analysis has been conducted, and the coverage of most fields is unsatisfactory, though certain fields achieve very high coverage. It reveals several issues to be addressed in the following processes, especially in the construction of the structured templates, which should be slightly adjusted to the corpora.

For these reasons above, the information extraction (IE) task for pathology notes is more complicated than that in a general English domain. Using current existing IE systems will not be capable of addressing the challenges properly; therefore, novel or appropriate techniques have to be developed to deal with them.

Chapter 4 Corpus Annotation

This chapter presents three semantically annotated corpora, namely melanoma corpus, colorectal cancer corpus and lymphoma corpus. These corpora are to be used for extracting entities and relations in free-text pathology reports. Medical or related entities are annotated in the melanoma corpus and colorectal cancer corpus, and relations between them are annotated as well in the lymphoma corpus. As far as we know, these corpora are the most specific and detailed cancer corpora prepared for automatic conversion to structured reports in the clinical domain. Most of the entities or relations have not been well studied previously.

This chapter begins with the overview of some existing annotated corpora, and then follows the design of the annotation schema and guidelines, the main annotation process as well as detailing the distribution of entity and relations across the corpora.

4.1 Introduction

4.1.1 Overview of Existing Annotated Corpora

Many corpora have been designed for information extraction (IE) in the biomedical domain. They vary from syntactic annotation (e.g., part-of-speech (POS) tags) to semantic annotation of named entities and complex relations between the entities. For example, the GENIA corpus has been widely used in a lot of biomedical natural language processing (NLP) research (Kim et al., 2003). It consists of 2000 biomedical abstracts extracted from the MEDLINE database, semantically annotated with rich biological named entities such as DNA and protein, and up to 47 related biological types defined in the GENIA ontology. Apart from semantic information, syntactic information such as POS tags was annotated (Tateisi and Tsujii, 2004) with a scheme based on the Penn Treebank corpus (Marcus et al., 1994). These annotated corpora made a great contribution to promoting the application of machine learning techniques in that domain. However, they are not suitable for NLP research in the clinical domain, as the materials they used are the biomedical literature, which is a different genre, usually well-formatted and with less noise than clinical notes; they focus on biological named entities, which are out of the scope of NLP research in the clinical domain.

In the clinical domain, only a few annotated corpora are publicly available. One probable reason is lack of access to the data, as hospitals, clinics and other health agencies strictly restrict the access to clinical data for researchers outside the associated institutions, given concerns about the possibilities of compromising patient privacy and institutional practices (Chapman et al., 2011b). Another reason is that the annotations require specific medical knowledge, and the recruitment, training and co-ordination of annotators also requires significant effort.

Several annotated corpora have been reported for IE in the clinical domain:

2010 i2b2/VA Challenge Corpus

This is composed of 1748 discharge summaries and progress reports received from Partners Healthcare, Beth Israel Deaconess Medical Centre, and the University of Pittsburgh Medical Centre. Three entity types were annotated: medical problems, tests, and treatments (Uzuner et al., 2011). Medical problems were defined as phrases that contain observations made by patients or clinicians about the patient's body or mind that are thought to be abnormal or caused by a disease; treatments were defined as phrases that describe procedures, interventions, and substances given to a patient in an effort to resolve a medical problem; the definition of tests is phrases that describe procedures, tests, and measures that are done to a patient or a body fluid or sample in order to discover, rule out, or find more information about a medical problem. They were all loosely based on the associated UMLS semantic types such as disease or syndrome, sign or symptom, medical device, clinical drug, laboratory procedure, diagnostic procedure and so on, but also included some instances not covered by UMLS.

The challenge set three tasks regarding these entities: the concept extraction task focused on the extraction of these entities; the assertion classification task was to assign assertion types for medical problems, including present, absent, possible, conditional, hypothetical, and not associated with the patient; the relation classification task aimed to assign relation types that hold between medical problems, tests, and treatments, e.g., treatment improves medical problem, test reveals medical problem and medical problem indicates problem. Therefore, the gold-standard data also included assertions and relations besides entities.

One of the potential benefits when using the corpus is that each record is de-identified, tokenised and broken into sentences, which can save much time on some NLP pre-processes: de-identification, tokenization and sentence boundary detection.

ODIE Corpus

The theme of the 2011 i2b2/VA Challenges was the resolution of coreference in medical records (Uzuner et al., 2012). The Ontology Development and Information Extraction (ODIE) corpus is part of the data for the challenge. It contained de-identified clinical and pathology reports from Mayo Clinic, discharge records, radiology reports, surgical pathology reports, and other reports from the University of Pittsburgh Medical Centre. Ten types of entities were annotated: anatomical site, disease or syndrome, indicator/reagent/diagnostic aid, laboratory or test result, none, organ or tissue function, other, people, procedure, and sign or symptom (Savova et al., 2011). Except for people defined in MUC-7 coreference task (Chinchor, 1998), other medical entity types were based on UMLS. "Other" and "none" were two special entity types prepared for the task, wherein "other" was assigned for entities that cannot be classified as any of the above and "none" served mostly as pronouns that inherit

one of entity types through coreference. Apart from the entities, the coreferent or anaphoric relations were also annotated between them.

However, the challenge required only entities that participated in anaphoric relations to be annotated. Thus, some entities were deliberately ignored for this reason. Moreover, some complex entities were also annotated, e.g., nested entities. The nested entities referred to entities with overlapping spans. For example, in the sentence "The tumor is 4.0 cm from the distal margin of resection.", "the distal margin of resection" was annotated as "other" and "resection" was annotated as procedure. These entities may increase the difficulty for the application of machine learning techniques (e.g., conditional random fields), since these techniques assume the entities to be predicted appear in sequence rather than nested. To identify such entities, rule-based methods are essential for post-processing.

CDKRM Corpus

Coden et al developed a detailed manually annotated corpus (Coden et al., 2009) to train and test the Medical Text Analysis System/Pathology version (MedTAS/P) for populating the Cancer Disease Knowledge Representation Model (CDKRM) from free-text pathology reports,. It consists of 302 pathology reports of 222 patients who could be assigned ICD-9 CM codes for diagnoses of colon cancer, including 153.0, 153.1, 153.2, 153.3, 153.4, 153.7, 153.7, 154.0, and 154.1. They only presented some types of entities to be evaluated: anatomical site, histology, grade value, dimension, date, gross description, primary tumour, metastatic tumour, and lymph node status. Except for anatomical site and histology it was pointed out that they were based on ICD-O, the definitions of other entity types were unclear, although their attributes have been interpreted in the descriptions of the model. They also annotated coreferences for anatomical site and histology.

According to the results they displayed, most entity types have achieved strong inter-annotator agreements, except for positive lymph nodes. They concluded that the corpus was of good quality and paved the way for automation.

C311 Corpus

The C311corpus comprises 311 clinical notes drawn from Royal Prince Alfred Hospital's Intensive Care Service, including admission notes, clinician notes, physiotherapy notes, echocardiogram reports, nursing notes, dietary reports and operating theatre reports (Wang, 2009). There were eleven entity types derived from the SNOMED CT concept hierarchy (IHTSDO, 2007-2014): abnormality, body, finding, health profile, object, observable, occupation, organism, procedure, qualifier, and substance. Nested entity was one of the emphases in the corpus. For example, the procedure entity "left cavernous carotid aneurysm embolisation" is the outermost entity, contains several inner concepts: the qualifier entity "left", the finding entity "cavernous carotid aneurysm", the body entity "cavernous carotid" and the abnormality entity "aneurysm". Though it has been pointed out that the recognition of nested entities is crucial for other tasks such as coreference resolution, relation

extraction, and ontology construction, the evaluation of these entities could be difficult, as they resulted in multi-label for a single token. Thus only the outermost entities were evaluated in this work.

MIMIC II Clinical Corpus

The MIMIC II Clinical Database (Goldberger et al., 2000; Moody and Lehman, 2009) contains clinical records for 32,536 subjects. These records include results of laboratory tests, medications, ICD-9 diagnoses, admitting notes, and discharge summaries. Each record consists of various data for a single subject, such as ICD-9 diagnosis codes, physician's orders, census events (e.g., patient admissions and transfers), solution (fluids given to the patient), and chart events (a set of observations, e.g., raw measurements come from Intensive Care Unit monitors and other instruments, and representative measurements).

Although the database has a great number of records, these records do not suit to this task, as they are presented in a semi-structured style, where each instance is listed with an abbreviated subheading.

In summary, these corpora are not suitable for the tasks in this study, for reasons as follows:

1. Different report types. The 2010 i2b2/VA challenge corpus comprised discharge summaries and progress reports; the C311corpus consists of clinical notes from Intensive Care Service; MIMIC II clinical corpus is composed of admitting notes and discharge summaries. They are distinct genres from pathology notes.
2. Small sample size. There were only 48 pathology reports and 18 surgical pathology reports in the ODIE corpus, which were insufficient for training and testing.
3. Sections in the reports. There were only two sections in the notes of the ODIE corpus and CDKRM corpus: the final diagnosis section and gross description section, which were far from the requirements of the structured reporting protocols, where other sections like clinical history and microscopic analysis were also required to present in the documents.
4. Annotated entity types. The advantages of using standard terminologies as reference to determine the entity types are evident: one can save much time on defining the entity types to be annotated, because the referring terminologies have explicit definitions for them; it eases the application of dictionary look-up approaches and encoding of the entities. For example, Wang purposed a lexicon look-up method for medical entity recognition on the C311 corpus (Wang, 2009); Coden et al mapped the entities to ICD-O (Coden et al., 2009). However, it will not work for this study, as the concept categories provided by standard terminologies like UMLS and SNOMED CT are too comprehensive or deficient to be used to annotate enough of the text to populate a structured report. For instance, both “3. 01 Diagnosis” and “G3.08 Desmoplasia” in Primary Cutaneous Melanoma Structured Reporting Protocol (Scolyer et al., 2010) belong to “body structure” in SNOMED CT, thus using “body structure” as an entity type cannot discriminate them from each other. There is no associated category in UMLS or SNOMED CT that can capture information about some standards and guidelines in the protocols, e.g., “G3.11 Growth pattern” in Primary Cutaneous Melanoma

Structured Reporting Protocol (Scolyer et al., 2010), "S2.02 Blocks" in Colorectal Cancer Structured Reporting Protocol (Eckstein et al., 2010) and "G3.02 Cell size" in Tumours of Haematopoietic and Lymphoid Tissue Structured Reporting Protocol (Norris et al., 2010) as well.

5. Complicated constructions of some entities. Nested entities were included in the ODIE corpus and C311 corpus, and the issues they created have been discussed above.
6. Accessibility of the corpus. The 2010 i2b2/VA challenge corpus, ODIE corpus, MIMIC II clinical corpus and their annotation guidelines are publicly available under data agreements, in contrast to the CDKRM corpus and C311 corpus.

Given the above reasons, none of these corpora is suitable for this study, but the annotation workflow and guidelines they introduced can provide very useful information in creating the annotation workflow and guidelines in this study.

4.1.2 Objective

Although a number of researchers have achieved some successes with unsupervised machine learning algorithms, more practical or sophisticated IE systems rely on annotated data to support learning or extraction of rules. Therefore, annotated data are very important for IE in the clinical domain, especially for supervised machine learning approaches which require such data to train the machine learners. These data are also useful as a gold-standard for evaluation of the IE systems. Creating a semantically annotated corpus can enable the performance of the IE system to be fully and automatically evaluated, as has been proven in the past i2b2/VA challenges (Uzuner et al., 2012; Uzuner et al., 2011). Furthermore, with these data, it is possible to tune an IE system to achieve better performance through comparisons across multiple versions of the system.

The aims to create semantically annotated corpora in this study include:

- To create training data for the application of supervised machine-learning approaches.
- To prepare test data for evaluation of the components of the IE system.
- To make use of the training data for tuning the components of the IE system to improve performance during development.

4.2 Annotation Schema

Annotation schemas act as knowledge representation tools regarding semantic categories and their specialized lexicon. Chapman et al have indicated that using annotation schema to train annotators could significantly increase agreement and decrease variability in annotations (Chapman et al., 2008). Hence, before starting the annotation progress, an annotation schema has to be acquired first.

The goal of the annotation schema is to identify the entities associated with fields in the structured reporting protocols, and determine the relationships between particular entities. Annotators can mark spans of text with an entity type, such as “De:Size”, “De:Specimen Type” and so on, also mark relationships as links between these spans. The entity types and relation types are presented in detail in the following sections.

Some researchers tend to develop their schemas based on standard terminologies. For example, for the Clinical E-Science Framework (CLEF) project, Roberts et al developed an annotation schema based on the UMLS semantic network, with the goal of utilising UMLS vocabularies in the following entity recognition task (Roberts et al., 2009). However, the annotation schemas purposed below specify the types of entities and relationships to be annotated, without adaption from standard terminologies, but are tailored to the structured reporting protocols. Specifically, most of the entity types are derived from the protocols, but some of them are defined by referring to published sources related to the project (e.g., pathology textbooks, colorectal cancer textbooks) or consulting the clinical staff or pathologists; a few of them regarding linguistic information are prepared by computational linguists; and, relation types are designed to classify relationships between particular entities. Accordingly, the definitions of entities and relations were developed by six computational linguists, a clinician and two pathologists. The schemas were developed and refined using an iterative process.

4.2.1 Entity Type

At first, an initial set of medical entities types was defined in each annotation schema, and new entity types were added to the schema if it was necessary to capture additional semantic or linguistic information. Most of the types have their own associated fields in the protocols, but some complex fields can be separated into more than one related type and ambiguous fields are combined together to be represented by the same entity types. Entity types without associated fields in the protocols were also created to capture some useful information or facilitate subsequent processes. For example, “Li:Lexical Polarity Negative”, “Li:Lexical Polarity Positive” and “Li:Modality” were created to reveal the assertion of an entity; “St:Clinical History Heading”, “St:Specimen Heading”, “St:Macroscopic Heading”, “St:Microscopic Heading”, “St:Diagnosis Heading” and “St:Comment Heading” were added to represent the section headings in the corpus, which consequently facilitated section context detection in the corpus.

There are some generic categories and entities types defined in the three corpora, as well as corpus-specific categories and entity types (please refer to Figure 4.1).

Two generic categories Synthesis and Structural are described as follows (definitions for other generic categories are tailored to each corpus, which are depicted in the following sub-sections):

- Synthesis (Sy): to reveal information required inference from the author(s) of the report.
- Structural (St): to depict the structure of the report.

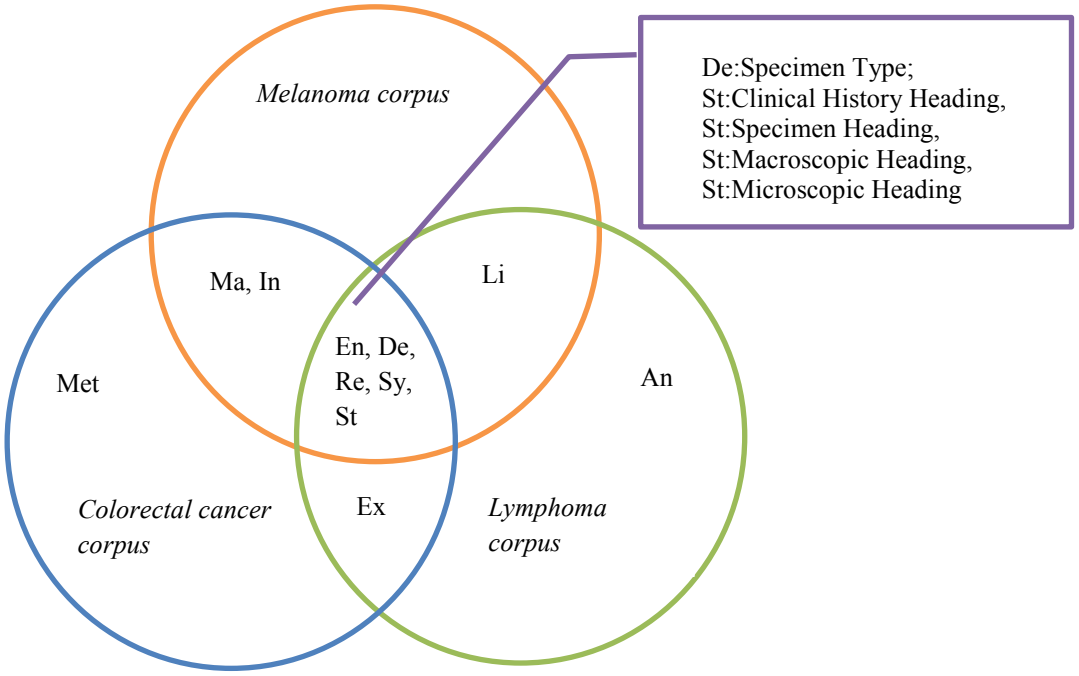


Figure 4.1 Generic and corpus-specific categories and entities types in the three corpora. En: Entity, De: Descriptor, Ma: Margin, Re: Reaction, In: Invasion, Sy: Synthesis, Li: Linguistic, St: Structural, Ex: Extent, Met: Metadata, An: Ancillary.

Five generic entities types are defined as (with examples are indicated with underlined texts):

De:Specimen Type – Specimen type captures the surgical procedure or site used to obtain the specimen.

Examples for melanoma corpus:

- (R) axillary SNB.
- The biopsy shows lentigo maligna melanoma.
- An ellipse of skin measuring up to 46mm x 25mm to a depth of 5mm…

Examples for colorectal cancer corpus:

- Specimen type: Extended right hemicolectomy
- Right colon: A right hemicolectomy comprising terminal ileum 110x15mm and proximal colon 150x35mm.

Examples for lymphoma corpus:

- Incision biopsy ?lymphoma.
- Two tan cores 18 and 16mm in length.
- "Distal gastrectomy + right hemicolectomy".

St:Clinical History Heading / St:Title- Clinical History – Any heading that pertains to the history of the patient.

Examples for melanoma corpus:

- CLINICAL NOTES
- Clinical History:

Examples for colorectal cancer corpus:

- CLINICAL NOTES:
- CLINICAL HISTORY: AP resection plus transverse colectomy.

Examples for lymphoma corpus:

- Procedure: Clinical Notes
- Procedure: CLINICAL DETAILS

St:Macroscopic Heading / St:Title- Macroscopic Description – Any heading that pertains to the macroscopic examination.

Examples for melanoma corpus:

- MACROSCOPIC
- MACROSCOPIC EXAMINATION
- MACROSCOPIC REPORT

Examples for colorectal cancer corpus:

- SPECIMEN: 1. Labelled - AP resection:
- MACROSCOPIC:
- MACROSCOPIC DESCRIPTION.

Examples for lymphoma corpus:

- Procedure: Macroscopic Description

St:Microscopic Heading / St:Title- Microscopic Description – Any heading that pertains to the microscopic examination.

Examples for melanoma corpus:

- MICROSCOPIC
- MICROSCOPIC REPORT
- MICROSCOPIC EXAMINATION

Examples for colorectal cancer corpus:

- MICROSCOPIC DESCRIPTION:

- MICROSCOPY: Sections of the caecal tumour…

Examples for lymphoma corpus:

- Procedure: Microscopic Report

The detailed corpus-specific categories and entity types are described below (with examples are indicated with underlined texts).

Melanoma Corpus

The entities types can be divided into eight main categories. Except two generic categories decribed above, others include:

- Entity (En): to identify the specimen and malignant or benign lesions on it.
- Descriptor (De): to describe the properties or characteristics of the Entity.
- Margin (Ma): to indicate whether the excision margins are clear and any descriptive material relating to the margins.
- Reaction (Re): to represent any lymphocytic or inflammatory reaction in the skin.
- Invasion (In): to illuminate local or distant invasion of the lesions.
- Linguistic (Li): to mark linguistic information about the above categories (except for Margin).

En:Associated Naevus (type) – References to any pre-existing or associated naevus with the melanoma.

- Features suggestive of a pre-existing naevus are seen.
- The sections show an unusual nodular malignant melanoma associated with a compound melanocytic naevus.

En:Primary Lesion – The primary lesion is typically the reason why the report was prepared. It can be usually described in various ways, which should be identified during annotation.

- The lesion is formed by small melanocytic nests…
- The dermal component consists of nests of moderately atypical naevoid melanocytes.
- On the surface is a variably pigmented grey nodule measuring 6 x 5 mm with a pale halo.

En:Lesion (other) – Other lesions mentioned in the report.

- The second lesion (block B) is a benign compound naevus.
- No pre-existing benign lesion is seen.

En:Satellites – References to any satellite lesion associated with the melanoma.

- No satellites.
- Microsatellites are absent.

En:Specimen Identifier – The specimen identifier is used to identify the specimen.

- The specimen consists of an ellipse of skin measuring 40mm x 20mm x 5mm.
- Sections show a nodular malignant melanoma…
- 1. The lesion is an acanthotic seborrhoeic keratosis.
- Specimen A: ?Dysplastic naevus

De:Specimen Type – The manner in which the specimen has been obtained.

- (R) axillary SNB.
- The biopsy shows lentigo maligna melanoma.
- An ellipse of skin measuring up to 46mm x 25mm to a depth of 5mm…

De:Cell Type – Descriptions of the primary cell type.

- The dominant cell type is epithelioid.
- Single malignant melanocytes are noted…
- …composed of large cells with epitheloid and balloon cell features.

De:Cell Growth Pattern – The cell growth pattern contributes to the identification of the sub-type of melanoma.

- The lesion is vertical growth phase …
- There is an asymmetrical poorly circumscribed proliferation of atypical melanocytes arranged in confluent units and nests at the dermal epidermal junction.
- Very occasional cells show intraepidermal Pagetoid spread.

De:Cosmetic Changes – Changes in appearance to the surrounding area that may be noted in the report and may be relevant to the diagnosis or the prognosis.

- There is a healing scar in the centre of the ellipse.
- Changed size following trauma.
- Focal balloon cell change is present.

De:Shape – Descriptions of the entity including colour, border and contour.

- Centrally there is an irregular mottled brown lesion 20 x 12 mm.
- A skin ellipse 12 x 6mm with an irregular tan nodule 7mm in greatest diameter with a pale centre and irregular outline.

De:Site and Laterality – The body part and side on which the lesion is located. This may also include finer locating information such as upper, lower, and mid.

- Changing lesion central abdo.
- Specimen labelled "Left upper thigh".

- Irregular pigmented lesion on the back.

De:Size – The sizes of the specimen, the primary lesion and any other lesions or noteworthy entities.

- The specimen consists of an ellipse of skin measuring 25mm x 15mm x 10mm.
- Situated on the surface there is a slightly raised grey lesion measuring 12mm in maximum extent.
- An ellipse of skin 15x7 mm bearing a slightly raised pale lesion 5x5mm.
- An ellipse of skin measuring up to 9mm X 4mm to a depth of 5mm from the nodule on the scalp.
- Specimen comprises a large oval piece of skin 65mm in length, width of 30mm and maximum thickness of 13mm.

De:Ulceration (mm) – References to any ulceration of a lesion and the size should be included if it is mentioned.

- The melanoma is not ulcerated.
- The tumour shows superficial ulceration.
- There is some surface ulceration but this is less than 6mm in extent.

De:Dermal Mitoses – Level of dermal mitosis and/or the mitotic rate.

- Occasional intradermal mitoses are seen, numbering less than 1 per square mm.
- Mitotic rate is 15 to 18 per mm2.
- Mitoses are approximately 2 per 10 hpf.

Ma:Excision Clear – Statement that the excision margins are clear or any descriptive material relating to the excision margins that doesn't belong under other Margin types.

- Excision appears complete.
- The lines of excision are clear of the lesion.

Ma:Excision Deep – The distance from the lesion to the deep margin.

- … and a deep margin of 2.3mm.
- Distance from deep margin = 4.0mm

Ma:Excision In Situ – The distance from the in-situ or junctional component to the lateral margins.

- The closest peripheral margin to in situ component measures 0.9mm…
- The in situ component is situated 0.2mm while …

Ma:Excision Invasive – The distance from the dermal component to the lateral margins (default category for lateral margins).

- Closest peripheral margin to invasive component = 1.0mm

- The melanoma is excised by a lateral margin of 0.5mm.
- The invasive component is 2.0mm clear of the nearest margin.

Re:Desmoplasia – References to the presence or absence of desmoplasia.

- It shows pleomorphic spindled malignant cells with strong S100-positivity permeating the dermis, inciting a desmoplastic reaction.
- Although there is no obvious desmoplasia…

Re:TILS – References to the presence or absence of tumour infiltrating lymphocytes (TILs).

- There is a patchy lymphocytic infiltrate around some of the deeper parts of the lesion.
- Tumour infiltrating lymphocytes (TILS) is seen at the base in block 3 especially in the subcutis nodule.

Re:Solar Elastosis – Evidence of skin reaction to the sun. It will make the skin appear leathery and can impact on diagnosis and prognosis.

- No significant solar elastosis is detected.
- Sections show sun-damaged skin.

Re:Fibrosis – Evidence of reaction in connective tissue or scarring.

- There is a narrow band of fibrosis between epidermis and the dermal deposit of tumour.
- Dermis beneath the lesion shows angiofibroplasia indicative of regression…

In:Vascular/Lymphatic – References to any infiltration of the blood vessels and lymphatic system.

- No lymphatic, vascular or perineural invasion is seen.
- There is no ulceration or vascular/lymphatic invasion.

In:Neurotropism – References to any neurotropism present or absent.

- Vascular, lymphatic or perineural invasion is not identified.
- There is no vascular invasion or neurotropism seen.

In:Breslow Thickness (mm) – Primary tumour thickness.

- The Breslow thickness is 0.8mm.
- The depth of invasion (Breslow) is 0.6mm, Clark level 3.
- Depth 0.55mm, level III.

In: Clark level – The layer of the skin into which the tumour has permeated.

- The tumour extends to Clark level IV with a Breslow thickness of 3.6mm.
- There is spread within the epidermis and within the dermis with a lesion contains a thickness of 1.4mm.

- Breslow thickness is about 1.2mm, Clark level 3.

Li:Lexical Polarity Negative – Lexically bound polarity related to negation.

- There is no evidence of dermal invasiveness.
- No perineural or lymphovascular invasion, neurotropism or desmoplasia is identified.
- Vascular invasion is not noted.
- The tumour lacks epidermal invasion…

Li:Lexical Polarity Positive – Lexically bound polarity related to confirmation.

- The biopsy shows lentigo maligna melanoma.
- The features are consistent with malignant melanoma…
- Patchy regression is present.
- A preexisting benign dermal naevus is also noted as in the previous biopsy.

Li:Modality – Lexically bound modality related to uncertainty.

- This possibly represents pre-existing dysplastic naevus.
- The lesion is probably malignant melanoma.
- This focus may be separate from the main tumour…
- No definite dermal mitoses are seen.

Li:Mood and Comment Adjuncts – Indication of degree or intensity.

- There is some dermal scarring, consistent with regression.
- There is a patchy, moderate lymphocytic infiltrate around the edges of the lesion.
- There is no significant atypia present.

Li:Temporality – References to any temporal indicator.

- … the subcutaneous tissue from the previous surgical procedure.
- There is evidence of both early and late regression.

Sy:Diagnosis – The diagnosis of the lesion within the specimen.

- The lesion is an invasive superficial spreading malignant melanoma.
- Sections show an ulcerated Level IV amelanotic melanoma measuring 3.1mm in maximum thickness.
- The section shows nests of basal cell carcinoma.

Sy:Regression – References to any regression within the lesion.

- There is vascular fibrous tissue up to 3.3mm deep consistent with a zone of late regression.
- There has been regressive activity and it is mostly complete.
- The lesion is a partially-regressed level II naevoid malignant melanoma…

Sy:Subtype – The histological type and classification of the melanoma.

- It is of superficial spreading type and is not ulcerated.
- The central nodule is ulcerated nodular melanoma.
- Sections show foci of invasive malignant melanoma arising in a Hutchinson's melanotic freckle (lentigo maligna).

St:Specimen Heading – Any heading that pertains to the specimen.

- SPECIMEN
- NATURE OF SPECIMEN
- Specimen(s) Received

St:Diagnosis Heading – Any heading pertaining to the diagnostic summary, generally present at the end of the report.

- DIAGNOSIS
- SUMMARY
- CONCLUSIONS

St:Comment Heading – Any heading that pertains to comments made by the pathologists.

- Comment
- COMMENT
- Further report

St:Sub Heading – Any miscellaneous subheading that does not fall under the aforementioned structural headings.

- Growth pattern:
- Lines of Excision:
- Cytological features:

Colorectal cancer corpus

The entities types can be broadly classified into nine categories. Apart from two generic categories decribed above, others include:

- Descriptor (De): to describe the specimen, the tumour's shape, structure, behaviour and spatial location.
- Entity (En): to identify any noteworthy structures composing the specimen or abnormal findings regarding to the tumour.
- Extent (Ex): to indicate the spread of the tumour within the adjacent organs, tissues, structures or the body.
- Invasion (In): to illuminate the malignant involvement caused by the tumour.

- Margin (Ma): to provide the distance of the tumour from specified margins or indicate whether they are involved by the tumour.
- Metadata (Met): to mark the staging information of the tumour.
- Reaction (Re): to reveal the reactive changes due to tumour growth or treatment.

De:Ancillary Studies – Ancillary modality identifies any supporting tests performed (and their findings), usually of diagnostic or prognostic significance.

- There is no loss of mismatch repair protein expression in tumour cells with immunohistochemical stains for MLH1, PMS2, MSH2 and MSH6.
- In view of the plasmacytoid morphology in part of the tumour, CD138 will be performed and neuroendocrine markers CD56, chromogranin and synaptophysin.
- MIH1 - normal mucosa positive, tumour positive.

De:Specimen Blocks – Specimen block describes how the specimen was sliced into sections for testing. An identifier and the composition of each block will be given and the annotation applies to each pair of these.

- Block 1: Proximal margin.
- Blocks: A-C: terminal ileum lesion.
- 1I-1L - multiple lymph nodes.

De:Specimen Images – Specimen image gives information about any imaging of the tumour.

- Macroscopic Photos - not taken.
- Mid rectal cancer - T2 on MRI.

De:Specimen Size – Specimen size gives the measured dimension(s) of the resected colorectal tract.

- Colon: 340 x 30 mm.
- With patient details only: A length of large bowel 200mm with attached mesocolon 70mm wide.
- Left colon: A left sided resection measuring 350mm x30mm.

De:Specimen Type – Specimen type captures the surgical procedure or site used to obtain the specimen.

- Specimen type: Extended right hemicolectomy
- Right colon: A right hemicolectomy comprising terminal ileum 110x15mm and proximal colon 150x35mm.

De:Tissue Banking – Tissue banking records whether any tissue, normal or cancerous, has been added to a tissue bank for research.

- Tissue bank specimen: TB1 - A piece of brown tumour 7 x 5 x 3mm.

- Tissue Bank 1 - Colon normal: all in TB1.

De:Mesorectal Integrity – The integrity of the mesorectum in the specimen is described as complete, nearly complete or incomplete.

- The mesorectum is complete and indurated measuring up to 70mm in thickness.
- Mesorectal integrity: Complete.

De:Perforation – Perforation details whether the tumour itself or the colorectal tract adjacent to the tumour appears perforated.

- Perforation: Nil seen.
- The serosal surface over this tumour appears ragged and ulcerated, possibly representing perforation of tumour onto the peritoneal surface.
- These sections confirm an area of perforation through an ulcerated poorly differentiated adenocarcinoma…

De:Serosa Description – Descriptions of the surrounding serosa (sometimes called "visceral peritoneum").

- The rest of the serosa is pink and smooth.
- There is puckering of serosa of the bowel at 50mm from distal resection margin.

De:Tumour Description – Descriptions of the tumour.

- 80mm from the nearest colonic resection margin there is an annular constricting ulcerated hard lesion involving a segment of bowel 30mm long.
- Appearance - sessile pale pink polypoid tumour
- 1-2. Sections from the rectum show residual moderately differentiated adenocarcinoma composed irregular glands and some nests of pleomorphic cells.

De:Tumour Size – Tumour size gives the measured dimension(s) of a tumour.

- Size: 53 mm in diameter
- At the caecum, there is an almost circumferential centrally ulcerated tumour, 40mm in axial length and 40mm in width…
- Within the caecum, immediately inferior to the ileocaecal valve, is a fungating mass measuring 45x30x10mm.

De:Peritoneal Reflection – Peritoneal Reflection locates the tumour as completely above, astride, or completely below the anterior peritoneal reflection.

- The tumour extends close to the serosal surface approximately 10mm above the anterior peritoneal reflection.

- Just below the peritoneal reflection margin mainly posterior there is a large ulcerating tumour…
- Relationship to anterior peritoneal reflection: Astride

De:Tumour Site –Tumour site gives the part of the colorectal tract in which the tumour was located. For rectal tumours, tumour site can include a measured distance from the anal verge.

- At the lower rectum, there is a centrally ulcerated cream tumour…
- Site of tumour: Right colon.
- CLINICAL NOTES: Rectal ca 8cm from anal verge.

En:Coexistent Pathology – Relevant coexistent pathological abnormalities, e.g., Polyps (describe type, number, etc), Ulcerative colitis (with dysplasia/without dysplasia), Crohn's disease (with dysplasia/without dysplasia).

- Polyps: Associated tubulovillous adenoma with high grade dysplasia.
- The specimen shows severe ulcerative colitis of the left colon with a mass associated dysplasia arising in the rectum…
- A non-caseating granuloma has been identified and the appearances are consistent with active Crohn`s disease which is much less in the section taken from the terminal ileum.

En:Distant Spread or Metastases – Metastases identify any distant spread of the cancer.

- There is a deposit of tumour in the omentum.
- Distant Spread - present, liver.
- Metastases (colon/other): Not identified

En:Lymph Nodes – Lymph nodes give the number of lymph nodes identified.

- LYMPH NODES: 10 lymph nodes identified, 5 of which show evidence of metastatic tumour.
- 16 additional mesenteric lymph nodes are retrieved 3-10mm in greatest dimension.
- Total number of nodes identified: 25

En:Residual Tumour – Residual Tumour identifies whether any tumour was left as residual by the surgical excision.

- No residual invasive tumour is identified.
- Residual tumour: R0

Ex:Donut Involvement – Donut involvement refers to the presence of tumour cells within the proximal/distant donuts (samples of supposedly normal tissue at the ends of the resection).

- 2. Distal rectal donut: No tumour identified.
- 2. Sections show unremarkable donuts of large bowel. No tumour is identified.

Ex:Extent – Extent indicates direct spread to adjacent organs, tissues or structures.

- Other organ invasion: Not present.
- The tumour extends to the serosal surface and shows focal lymphatic channel permeation with extrinsic infiltration into the adherent small bowel, reaching the the small bowel mucosa…

Ex:Extramuscular Spread – Extramuscular spread gives the measured distance of spread beyond the muscularis propria (in mm). This is often referred to as the tumour having infiltrative margins.

- Tumour Border: Infiltrative
- Distance beyond outer edge of muscularis propria: 3.0mm

Ex:Lymph Node Involvement – Lymph nodes involvement reports the number of extracted lymph nodes which are shown to be malignantly involved/un involved.

- Twenty lymph nodes were identified, five of which show evidence of metastatic adenocarcinoma.
- 12 lymph nodes in which no tumour is found.
- Lymph node status: 7 of 19 lymph nodes show metastatic carcinoma.
- (Node summary 1/57)

Ex:Serosal Involvement – Serosal involvement identifies whether there is malignant involvement of the serosa/outer layer of the colon.

- The free serosal surface appears clear of tumour.
- It involves the serosa and infiltrates the adjacent omentum.

In:Depth of Invasion – Depth of invasion records how far into the colorectal tissue the tumour has invaded. This may be reported as a measured distance or as the specific layer of colorectal tissue which the tumour reaches (excluding the serosa, see Ex:Serosal Involvement).

- The tumour infiltrates through the muscularis propria and extends into the subserosa.
- Depth of invasion: Subserosa (pT3)
- Sectioning through the tumour reveals depth of invasion is up to 10mm…

In:Perineural Invasion – Perineural invasion describes whether there is malignant involvement of the spaces near nerve cells.

- No peritoneal invasion is identified.
- Perineural invasion: Present

In:Venous and Small Vessel Invasion – Small vessel invasion refers to the malignant involvement of small vessels including lymphovascular or capillary involvement. Venous invasion refers to the malignant involvement of large vessels, including both veins and arteries.

- No extramural vascular invasion is identified.
- Lymphovascular invasion: Present.
- possible small vessel invasion in submucosa.
- There appears to be a focal tumour invasion of a medium sized vein within perirectal fat.

Ma:Circumferential Margin – Circumferential margin gives the measured distance between the tumour and the radial or circumferential margin. Also included is the mesenteric resection margin, which is an additional "radial" margin, considered in cancers of particular areas of the colon.

- …invading into pericolic fat, 0.4mm from the radial margin, without serosal involvement…
- Tumour extends to 4.4 mm of the non-peritonealised margin.
- Minimum distance between tumour and circumferential margin: 8mm

Ma:Proximal or Distal Margin – Proximal and distal margins gives the measured distance between the tumour and the cut-end margins.

- Arising in the mid colon is a circumferential tumour 132mm from the proximal resection margin, 60mm from the ileocaecal valve and 80mm from the distal resection margin.
- 35mm from one longitudinal resection margin is an ulcerated lesion 15mm in maximum dimension which penetrates into submucosa.

Ma:Clear – Clear indicates that the tumour is clear of its margins, in place of other Margin types if no numerical value is given.

- d. all surgical margins clear.
- Circumferential margin involved: No
- Tumour is well clear of longitudinal resection margins.

Met:Anatomic Stage – Anatomic stage represents the extent or severity of the cancer.

- Pathological stage: ACPS: A2
- This is a Dukes C colonic carcinoma.
- AJCC stage I (T1 N0 MX V0 R0)

Met:M Value – M Value identifies whether distant (to other parts of the body) metastasis (M) has occurred.

- T N M: pT3 pN1b pM0
- Pathological stage: ACPS: Stage D, TMN: pT4b N2b M1b, Stage IIIC

Met:N Value – N Value identifies whether cancer cells have spread to nearby (regional) lymph nodes (N).

- STAGE: pT3 N1
- Pathological stage: ACPS: Stage C, pT4, pN2a

Met:T Value – T Value gives the extent or spread of the tumour (T).

- AJCC stage 3B (T3 N1 M0).
- TNM staging pT4a, pN0, pMx.

Re:Desmoplasia and Fibrosis – Desmoplasia identifies whether any fibrous or connective tissue has grown as a reaction to tumour growth. Fibrosis identifies any hardening or scarring of tissue as a reaction, usually as a healing reaction.

- The carcinoma is associated with moderate desmoplastic reaction and a mild peritumoural chronic inflamatory cell infiltrate.
- (B) POSTERIOR VAGINAL WALL: PATCHY FIBROSIS, NO TUMOUR IDENTIFIED.

Re:Response to Rx – Response to Rx describes the patient's response to previous radiotherapy treatment.

- Radiation induced mucosal changes adjacent to residual adenocarcinoma and including distal donut.
- Response to neoadjuvant therapy – moderate
- …* TRG 2 (Mandard)…

Re:TILS and Peritumoural Lymphocytes – TILS and peritumoural lymphocytes refer to the presence, density distribution and severity of lymphocytic response, which are immune reactions to cancer.

- Intratumoural/peritumoural lymphocytic response - minimal.
- There are mild numbers of tumour infiltrating lymphocytes…
- The carcinoma is associated with a moderate peritumoural chronic inflammatory cell infiltrate.
- There is no significant diffuse or nodular Crohn`s-like lymphocytic chronic inflammatory infiltrate around the advancing tumour margin.

Sy:Comment – Comment is relevant to the comment fields of the structured template. This includes other issues noted by the pathologists.

- 2.8 Background abnormalities: No
- Appendix not seen.
- The colon proximal to the tumour is dilated.
- The left colon is abnormal, with the greatest abnormality identified in the distal 320mm of the specimen.

Sy:Histological Grade – Histological grade gives the level of differentiation, or percentage of the tumour composed of glandular structures. This maybe reported as a numerical grade.

- Histological grade: Moderately differentiated.
- The majority of the tumour is poorly differentiated consisting of ragged clusters…
- AJCC Stage IIIB (pT3 pN2a G2 R0)

Sy:Histological Type – Histological type identifies the type of cancer the tumour represents. This is almost always a form of adenocarcinoma.

- The tumour is a moderately differentiated adenocarcinoma.
- Tumour type: Mucinous adenocarcinoma.
- Tumour type: Signet ring adenocarcinoma.

Sy:Medical History – Previous medical history reported within the request form. This includes treatment, past disease, age, etc.

- CLINICAL NOTES: 60 year old female.
- No XRT.
- Past history of breast ca.

St:Ancillary Studies Heading – Any heading that designates the following content is concerned with ancillary tests performed and their results.

- SUPPLEMENTARY REPORT : Ckae1/3 immunostains are unremarkable (block 1G).
- ANCILLARY STUDIES

St:Synthesis Heading – Same as St:Diagnosis Heading for the melanoma corpus.

- DIAGNOSIS:
- SYNOPTIC REPORT: Colorectal carcinoma
- CONCLUSION: SITE: Caecum.

St:Subheading – Any miscellaneous subheading, which may be further divided into some reportable field. Note that it includes specimen identifiers.

- CONCLUSION: 1. Abdominoperineal resection (post adjuvant chemo-radiotherapy).
- MICROSCOPIC: SPECIMEN 1. Sections show the tumour to be a moderately differentiated mucinous adenocarcinoma.
- TUMOUR STAGING: AJCC STAGING: T3;N0;MX (STAGE IIA); DUKES STAGE: B.
- Tumour type/differentiation

Lymphoma corpus

The entities types can be categorized to eight kinds. Except two generic categories described above, others are illuminated as follows:

- Ancillary (An): to illuminate any ancillary study result and interpretation.
- Descriptor (De): to describe the specimen and the tumour.
- Entity (En): to identify the specimen and any note-worthy structures composing the specimen.
- Extent (Ex): to indicate the spread of the tumour within the body.
- Reaction (Re): to reveal the reactive changes due to tumour growth.
- Linguistic (Li): to mark linguistic information about the above categories (except for Ancillary).

An:Biomarker: Indicator that is usually used in immunohistochemistry tests or flow cytometry studies.

- The atypical lymphoid cells are positive for CD20, CD79a, cyclin D1, CD5 and bcl-2.
- They are negative for CD3, CD10 and CD23.

An:Immunohistochemistry- Positive: The positive results of the immunohistochemistry tests.

- Positive - CD20 +++, CD30 +
- The tumour stains strongly for CD45, CD20 and CD79a.

An:Immunohistochemistry- Negative: The negative results of the immunohistochemistry tests.

- CD23 – negative
- Immunohistochemistry shows that the neoplastic cells are CD79a+, CD20+, CD10+, CD23+, CD5-, cyclin D1- and CD30-.

An:Immunohistochemistry-Equivocal: The equivocal results of the immunohistochemistry tests.

- Staining for CD15 is equivocal.

An:Immunohistochemistry-Comment: The free text expressions for interpretive comment of the immunohistochemistry tests.

- The necrotic centre of the nodule also has some cell-outlining staining for CD56 and CD45RO.
- The CD5 staining appears to correlate with the CD3.

An:Flow Cytometry- Positive: The positive results of the flow cytometry studies.

- Positive for: Kappa, CD19, CD10, CD45, CD38

An:Flow Cytometry- Negative: The negative results of the flow cytometry studies.

- Negative for : Lambda, CD2, CD3, CD4, CD5, CD8, CD7, CD14, CD16, CD56

An:Flow Cytometry- Comment: The free text expressions for interpretive comment of the flow cytometry studies.

- INCREASED CD3+CD4+/CD3+CD8+ T CELL RATIO.
- NON DIAGNOSTIC FINDINGS.

An:FISH Results: The results of fluorescence *in situ* hybridization (FISH) tests.

- The remaining 58% of cells showed a normal diploid MYC signal pattern.
- nuc ish(MYC x3)

An:Cytogenetics Comment: The free text expressions for interpretive comment of FISH tests.

- Interphase FISH analysis revealed the presence of an abnormal signal pattern.
- Three intact MYC fusion signals were observed.

An:IgH Test: The results of polymerase chain reaction (PCR) analysis with immunoglobulin heavy chain (IgH) tests.

- Two MONOCLONAL bands of 104 and 121 bp were detected against a polyclonal background.
- An irregular POLYCLONAL smear was detected.

An:TCRgamma Test: The results of PCR analysis with T-cell receptor gamma chain (TCRgamma) tests.

- Tube 1: A POLYCLONAL smear was detected.

An:PCR Comment: The free text expressions for interpretive comment of PCR analyses.

- PCR amplification of DNA with primers flanking the region of gene rearrangement.
- IgH gene rearrangement studies were performed after the method of Brisco et al (1990) Br J Haem 75: 163-167 using the LJH and FR3A primer set.

De:Topography: Specified anatomical site and laterality of biopsy.

- Tumour (L) cubital fossa
- Right lower neck lymph nodes.
- FNAB (R) axilla- suspicions for Hodgkin's Lymphoma

De:Tissue Source: The source of the specimen or tissue.

- A lymph node 33 x 30 x 18mm.
- The infiltrate extends into perinodal fat.

De:Anatomical Structure: Unspecified anatomical site derived from names of Level 3 codes defined in topography axis of ICD-O-3 (WHO - World Health Organization, 1976-2000).

- "CA stomach".
- Cervical lymphadenopathy FNA - atypical cells.

De:Laterality: The laterality of biopsy.

- "SUPRACLAVICULAR CYST (LEFT)".
- Bilateral melanomas mets (R) parotid bed…

De:Specimen Type: Same as De:Specimen Type for melanoma corpus.

- Incision biopsy ?lymphoma.
- Two tan cores 18 and 16mm in length.
- "Distal gastrectomy + right hemicolectomy".

De:Lineage: Descriptions of the primary cell type.

- Lymph node, L cubital fossa: Malignant lymphoma, diffuse and follicular, large B cells predominating
- The nodules have a background of mainly T cells…

De:Architecture: The pattern of infiltration or architecture of abnormal cells.

- Sections show a lymph node with architecture totally effaced by a diffuse proliferation of large cells admixed with small lymphocytes and histiocytes.
- A few lymphoepithelial lesions are present.
- The nodal portion is follicular in pattern and consistent with grade 2 (of 3, WHO).

De:Cell Size: The size of abnormal cells.

- The lymph node shows a diffuse infiltrate of malignant large lymphoid cells with a few scattered multinucleated cells.
- Sections show effacement of the nodal architecture by a diffuse infiltrate of small to medium sized lymphoid cells.

De:Cytomorphology: Characteristic cytological features of individual tumour cells.

- Occasional binucleate Reed-Sternberg cells are identified.
- A few scattered centroblasts are also present.

De:Specimen Size: The measured dimension(s) of specimen(s).

- Two pieces of pale tissue each 5mm.
- Fatty tissue 55 x 40 x 15mm in aggregate with multiple enlarged lymph nodes.

De:Preservative Fluid: The fluid in which the specimen is delivered to the laboratory.

- "LYMPH NODE RIGHT NECK" (lymph node received fresh).
- Received in formalin, a small pale tan node 8mm across.

De:Sample Triage: The distribution of biopsy material to different laboratories (internal and/or external) and for different investigational modalities.

- Smears, imprints, tissue for flow cytometry, mollecular biology and frozen section were taken.

De:Specimen Blocks: Same as De:Specimen Blocks for the colorectal cancer corpus.

- A. Residual of frozen section.
- B. Remainder of specimen.

De:Tumour Size: Same as De:Tumour Size for the colorectal cancer corpus.

- Much of this area tumour, at least 80 x 50 x 15mm.
- Tumour size: - 50mm maximum diameter (macroscopically).

En:Specimen Identifier: Specimen identifier is used to identify the specimen. Unlike En:Specimen Identifier for the melanoma corpus, general terms like "specimen" and "sections" are excluded.

- 1. "Biopsy of liver".
- 2-5. The (L) axillary sentinel node biopsies show no evidence of melanoma in any of 11 nodes.

En:Coexistent Pathology: Same as En:Coexistent Pathology for the colorectal cancer corpus.

- Surface mucosal ulceration is extensive.
- 2. Gastric (Prepyloric) biopsy- Chronic gastritis.

Ex:Disease Extent: The extent of disease.

- Pt has generalised lymphadenopathy otherwise well.
- PET (30.1.06) showed it to be a solitary lesion, and glucose-avid.

Ex:Other Sites of Disease: Indicates involved sites or pattern of disease spread and whether disease is nodal or extranodal.

- The mediastinal mass was 13.8cm dia, involving superior and anterior mediastinum, down to the diaphragm.
- Focal extranodal spread is seen.

Re:Tissue Reaction: Host cells and tissue reactions.

- There are numerous eosinophils, plasma cells and some macrophages in the background.

- There are areas of necrosis.

Li:Lexical Modality: Same as Li:Modality for the melanoma corpus.

- Incision biopsy ?lymphoma.
- Small intestine and retroperitoneum - diffuse large B cell malignant lymphoma, probably follicular centre cell origin

Li:Lexical Polarity Negative: Same as Li:Lexical Polarity Negative for the melanoma corpus.

- There is no evidence of dysplasia.
- Although one resembles a RS cell variant, the morphology overall does not suggest Hodgkin lymphoma.

Li:Lexical Polarity Positive: Same as Li:Lexical Polarity Positive for the melanoma corpus.

- The sections show diffuse malignant lymphoma of large B-cells.
- Small lymphocytes and eosinophils are also present.

Li:Mood and Comment Adjuncts: Same as Li:Mood and Comment Adjuncts for the melanoma corpus.

- The lymph node shows a diffuse infiltrate of malignant large lymphoid cells with a few scattered multinucleated cells.

Li:Temporality: Same as Li:Temporality for the melanoma corpus.

- The appearances are similar to the previous biopsy.
- Mass left humerus increasing pain for the past 6-8 weeks.

Sy:Diagnosis: Same as Sy:Diagnosis for the melanoma corpus.

- Cervical lymph node - Nodular sclerosis Hodgkin lymphoma.
- Lymph node of neck - T cell-rich, large B cell malignant lymphoma
- Scapula (acromion) - malignant lymphoma, diffuse large B cell type

Sy:WHO Grade: The grade for follicular lymphoma.

- 2. Lymph node, site not stated - follicular lymphoma, WHO grade 1.
- I agree with the diagnosis of follicular lymphoma, grade 1/3, predominantly follicular.

Sy:Medical History: Any previous lymphoma, leukaemia or other relevant haematological disease.

- NHL 2 years ago.
- 1998 treated for Hodgkins disease.

Sy:Presentation: Clinical presentation of current relevant disease status.

- Presented with bilateral painless cervical lymphadenopathy.
- 3/12 history, pain.

Sy:Indication for Biopsy: The reason for the biopsy.

- Axillary lymph node ?recurrence.
- ?Transformation to high grade.

Sy:Clinical Impression: The clinical diagnosis or differential diagnosis.

- ?Lymphoma.
- ?SCC tonsil.

Sy:Constitutional Symptoms: References to any relevant constitutional symptom.

- Night sweats, weight loss ?lymphoma

Sy:Predisposing Factors: Immunocompromised states (immunodeficiency associated lymphoproliferative disorders), autoimmune conditions, and infective agents.

- Longterm HIV +ve.
- Post CTx.

Sy:Stage: The extent or severity of the lymphoma.

- ACP substage: C1

Sy:SNOMED RT Codes: SNOMED RT Codes and terms for the diagnosis.

- M-95903 Lymphoma, NOS
- T-C4000 Lymph nodes

Sy:Diagnosis Subtype: The sub-classification of lymphoma.

- ?MALT.
- The nodular sclerosis subtype is favoured.

Sy:Comment: Comment is relevant to the comment field in "SYNTHESIS" section.

- Consistent with large B-cell lymphoma, Please see report and comment.
- Please correlate with clinical and peripheral blood findings.

St:Title- Nature and Specimen Type: Same as St:Specimen Heading for the melanoma corpus.

- Procedure: Nature and Site of Specimen

St:Title- Summary: Same as St:Diagnosis Heading for the melanoma corpus.

- Procedure: Summary

St:Title-Pathologist Notes: Any heading pertaining to the Pathologist Notes.

- Procedure: Pathologist Notes - Not for Publication

St:Title- Frozen Section Report: Any heading pertaining to the Frozen Section Report.

- Procedure: FROZEN SECTION REPORT

St:Title- Supplementary Report: Any heading pertaining to the Supplementary Report

- Procedure: Supplementary Report

St:Title- Supplementary Summary: Any heading pertaining to the Supplementary Summary.

- Procedure: Supplementary Summary

St:Title- Special Investigations: Any heading pertaining to the Special Investigations.

- Procedure: SPECIAL INVESTIGATIONS

St:Title- Comment: Same as St:Comment Heading for the melanoma corpus.

- Procedure: Comment

St:Title-Subheading: Same as St:Sub Heading for the melanoma corpus.

- Immunohistochemical profile:
- SNOMED CODES:

Summary

It can be seen from the above annotation schemas that the design of the annotation schema for each corpus has a number of differences due to the distinguishable standards and guidelines in the protocols. For example, In:Clark Level is a medical entity type in the melanoma corpus; De:Peritoneal Reflection can only be annotated in the colorectal cancer corpus; De:Cell Size is in the annotation schema for the lymphoma corpus. Except for the medical entity types, the annotation schemas for melanoma and lymphoma corpora are augmented with Linguistic categories, while the colorectal cancer corpus is not, probably because the definition of a medical entity type has implicitly included the lexical polarity or modality, mood and comment adjuncts, and temporality (e.g., Ex:Lymph Node Involvement, Re:TILS and Peritumoural Lymphocytes, Sy:Medical History). There are some synoptic fields present in the colorectal cancer pathology notes, such as “Serosal involvement: Absent” and “TILs: MILD”, where the Linguistic category instances are not to be annotated separately from the entities. Thus, the computational linguists decided to include lexicons regarding linguistic information in the annotations of these entity types, and Linguistic categories are exempted in the schema.

However, there are still some common types among the three corpora (e.g., some Structural categories). The same category may be presented with different names. For instance, the heading for "Diagnostic Summary" section is named as St:Diagnosis Heading, St:Synthesis Heading, St:Title-Summary in the melanoma corpus, colorectal cancer corpus and lymphoma corpus respectively. De:Specimen Type is the only medical entity type that is defined with the same name and same description in all three corpora, suggesting that this is a general reporting standard in pathology notes.

4.2.2 Relation Type

There are 5 relation types in the final annotation schema. The relation annotation schema was targeted on the lymphoma corpus.

Relation types were grouped into three border categories: Negate, Result and Spatial Specialization. They are described in detailbelow (the examples are highlighted with bold and underlined texts):

Negate

The Negate relation represents the absence of a clinical or pathological finding and diagnosis. It is designed for negation detection in the following progress. Note that only pertinent negations within a sentence are annotated.

- A **lower grade** ["Sy:WHO Grade"] **MALT lymphoma** ["Sy:Diagnosis"] is **not** ["Li:Lexical Polarity Negative"] identified.
- **No** ["Li:Lexical Polarity Negative"] **Helicobacter** ["Sy:Predisposing Factors"] or **malignancy** ["Sy:Diagnosis"] seen.

If uncertainty is expressed in the sentence, it should be excluded from the annotation. Here is an example:

- A paraffin block was sent to the Department of Anatomical Pathology, HOSP_NAME, for FISH studies to *exclude* ["Li:Lexical Polarity Negative"] *mantle cell lymphoma* ["Sy:Diagnosis"].

In the above example, "exclude" and "mantle cell lymphoma" cannot be annotated as a Negate instance.

Result

The Result relation indicates whether a biomarker is positive, negative or equivocal indicator for an ancillary study. It can be sub-classified into three types:

1. Result-Equivocal

This relation should be annotated if a biomarker is an equivocal indicator for an ancillary study.

- **Equivocal** ["An:Immunohistochemistry-Equivocal"] - **Kappa and lambda light chains** ["An:Biomarker"], **immunoglobulins A, G and M** ["An:Biomarker"] .
- Staining for **CD15** ["An:Biomarker"] is **equivocal** ["An:Immunohistochemistry-Equivocal"].

2. Result-Negative

This relation should be annotated if a biomarker is a negative indicator for an ancillary study.

- CD5+ and **CD23** ["An:Biomarker"] – ["An:Immunohistochemistry- Negative"], suggesting mantle cell lymphoma on flow cytometry, but **cyclin Dl** ["An:Biomarker"] **negative** ["An:Immunohistochemistry- Negative"].
- **Negative** ["An:Immunohistochemistry- Negative"] - **CD3** ["An:Biomarker"], **CD10** ["An:Biomarker"], **CD56** ["An:Biomarker"], **ALK-1**["An:Biomarker"]

3. Result-Positive

This relation should be annotated if a biomarker is a positive indicator for an ancillary study. Note that instances that indicate the intensity of the positivity (e.g., "+++", "++", "+") have precedence over others to be connected to the biomarkers. In the following first example, "+++" is connected to "CD20", and "+" is connected to "CD30", rather than "Positive" to "CD20" or "CD30".

- Positive - **CD20** ["An:Biomarker"] +++ ["An:Immunohistochemistry- Positive"], **CD30** ["An:Biomarker"] + ["An:Immunohistochemistry- Positive"]
- Tumour cells are **strongly positive** ["An:Immunohistochemistry- Positive"] for **CD20** ["An:Biomarker"] & **BCL-2** ["An:Biomarker"].

Spatial Specialization

This relation depicts the specific laterality of an anatomical site.

- Right superficial parotidectomy & **bilateral** ["De:Laterality"] modified radical **neck** ["De:Anatomical Structure"] dissection.

Each relation has particular arguments; in other words, it can only connect certain types of entities according to these relations. For example, a Spatial Specialization relation can only exist between a De:Anatomical Structure and a De:Laterality entity, whereas up to 11 entity types can be linked to a Li:Lexical Polarity Negative for Negate relations. Table 4.1 lists the relationships and their potential arguments.

Relation type	Argument #1 type	Argument #2 type
Negate	Li:Lexical Polarity Negative	De:Architecture, De:Cytomorphology, Sy:Diagnosis, Sy:WHO Grade, Sy:Clinical Impression, Ex:Other Sites of Disease, Sy:Constitutional Symptoms, Sy:Predisposing Factors, Re:Tissue Reaction, Sy:Diagnosis Subtype, En:Coexistent Pathology
Spatial Specialization	De:Anatomical Structure	De:Laterality
Result-Positive	An:Immunohistochemistry-Positive, An:Flow Cytometry-Positive	An:Biomarker
Result-Negative	An:Immunohistochemistry-Negative, An:Flow Cytometry-Negative	An:Biomarker
Result-Equivocal	An:Immunohistochemistry-Equivocal	An:Biomarker

Table 4.1 Potential arguments for each relation type.

4.3 Methods

4.3.1 Annotation Tool

The Visual Annotator (VA) [1] was used as an annotation tool in this study. It is a Python-based annotator that supports both entity and relation annotations, which was developed and maintained by Health Language Analytics staff members. Figure 4.2 shows a screen shot of the VA working environment.

On one hand, Cohen et al indicated that most of the available annotated corpora were difficult to be restored or reused as they were stored in non-standard formats (Cohen et al., 2005). On the other hand, there are advantages for storing annotation in standoff format, such as optical recoverability and reusability.

Therefore, a standoff annotation style was adopted in this study. It stores annotation and raw text separately to prevent any loss of the structure information of the original text (Leech, 1993). VA stores files in its own "ann" format, which can be converted into other appropriate formats for further processing.

Each annotated entity and relation has some associated properties, which can be manually annotated or automatically generated by the tool.

An entity has the following properties:

- Type - The semantic type of the entity.

[1] http://www.icims.com.au/VisualAnnotator/

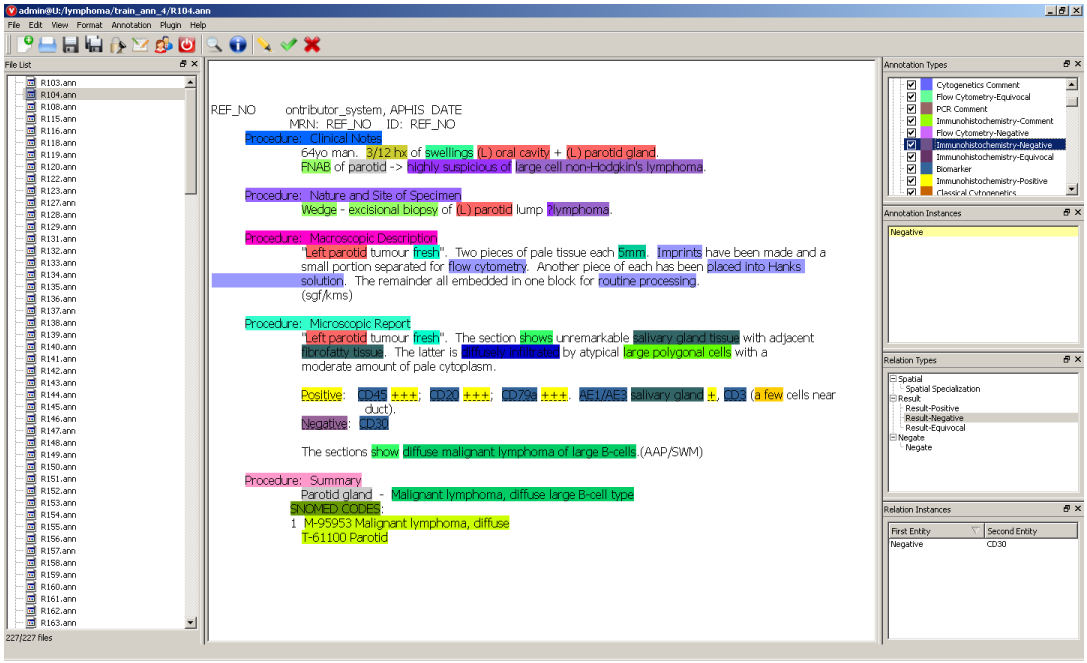


Figure 4.2 The Working environment of the Visual Annotator.

- Extent - The text span of the entity.
- Text - The string of text in the entity.
- Id - The internal identifier of the entity.

A relation has attributes as follows:

- Type - The semantic type of the relation.
- First Entity Type - The semantic type of the first entity.
- First Entity Extent - The text span of the first entity.
- First Entity Text - The string of text in the first entity.
- Second Entity Type - The semantic type of the second entity.
- Second Entity Extent - The text span of the second entity.
- Second Entity Text - The string of text in the second entity.
- First Argument Id - The internal identifier of the first entity.
- Second Argument Id - The internal identifier of the second entity.

During annotation, an annotator can annotate an entity by marking a span of text with its semantic type, and can also highlight a relation by marking links of two annotated entities.

4.3.2 Annotation Guidelines

Consistency is critical to the quality of a gold-standard corpus. It is important that each annotation must conform to the same standard. A large number of annotation tasks require direct annotation of

words in the text that relies on fairly consistent boundary segmentation by the annotators. However questions are frequently encountered while annotating a document. For example, should "2mm in diameter and 2mm in depth" be annotated as a De:Size instance, or as two instances? Should "extending focally onto the ragged serosal surface" be annotated as a In:Depth of Invasion instance or a Ex:Serosal Involvement instance? Should "diffuse malignant lymphoma of large and small cells" be annotated as a Sy:Diagnosis entity, or as a Sy:Diagnosis entity and a De:Cell Size entity? To ensure consistency, a set of guidelines ought to be provided to the annotators. Several issues are described in the guidelines: what should and should not be annotated; how to decide the boundary of a particular type of instance; how to decide whether two instances should be connected; and some special cases. The guidelines should also provide a sequence of steps, an instruction, which annotators should follow when annotating a document, in order to minimise errors of omission.

Roberts et al have done an analysis of annotation difference between computational linguists and clinicians, which revealed that the computational linguists could find more pronominal co-references and verbally signalled relations, while the clinicians could find more relations requiring domain knowledge to resolve (Roberts et al., 2007). Hence they believed that a combination of both linguistic and medical knowledge were preferable for developing the guidelines. In this study, both computational linguists and medical consultants were involved in the guidelines development to capture the greatest amount of information. Since the annotation categories cannot be mapped directly to an existing lexicon in any controlled vocabulary, sometimes, the guidelines were developed according to the text in the corpora, so that they could reflect what actually occurs in the text.

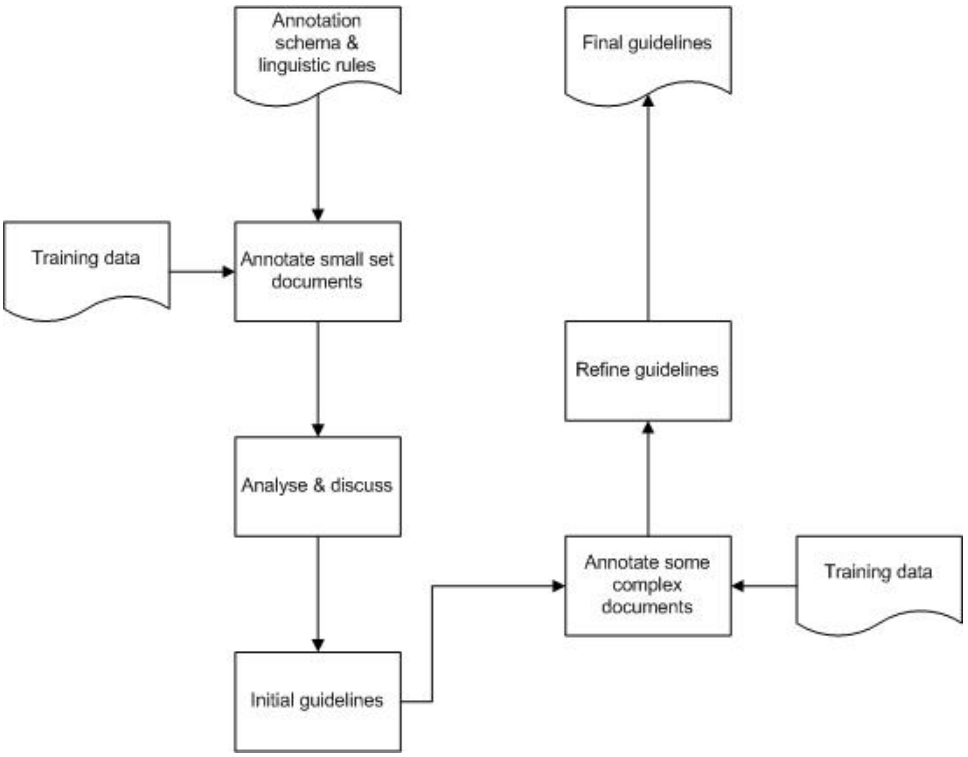


Figure 4.3 Annotation guidelines development process.

The guidelines were developed in the following processes, including guidelines creation and guidelines refinement, as illustrated in Figure 4.3.

Firstly, the computational linguists and medical consultants specify the annotation categories in the pathology notes that have been introduced into the annotation schemas. The computational linguists also defined a set of linguistic rules to indicate what constituents should be annotated. The computational linguists randomly selected a small set of documents from the notes and annotated them using these rules. The annotations were analysed where every instance annotated was discussed and then more specific rules were compiled and merged into the initial rules. Based on these annotations and rules, the initial guidelines were created.

Secondly, the six computational linguists individually annotated another small set of complex documents from the notes. They discussed every instance in their annotations with medical consultants, and made any change to the guidelines if necessary. A change to the guidelines can be either an addition of new rules to the guidelines to represent an unseen example, or modification of rules for ambiguous definitions. For instance, if the annotations of a boundary of a certain entity type occur inconsistently, a new rule will be added into the guidelines to resolve the ambiguity. After the guidelines refinement, the final guidelines were used as the references in the following main annotation exercise.

The initial guidelines specified a set of linguistic rules for the annotation convention of the noun phrase boundaries, that is, an entity should not cross noun-phrase boundaries. However, based on the thorough analysis of the documents, these did not fit for most cases. Thus, more specific rules for defining the boundaries of the entity types were presented in the final guidelines (see Table 4.2, 4.3, 4.4 for more details, wherein an empty boundary specification field suggests that the entity may cross the boundary of a noun phrase).

Entity type	**Protocol standard/guideline**	**Boundary specification**
De:Cell Growth Pattern	G3.11	
De:Cell Type		
De:Cosmetic Changes	G1.08	
De:Dermal Mitoses	S3.05	
De:Shape	S2.02	
De:Site and Laterality	G1.03	
De:Size	S2.03	
De:Specimen Type	G1.06	
De:Ulceration	S3.04	
En:Associated Naevus (type)	G3.10	
En:Lesion (other)	S2.06	NP
En:Primary Lesion	S2.05	
En:Satellites	S3.06	NP
En:Specimen Identifier		
In:Breslow Thickness (mm)	S3.02	
In:Clark Level	G3.04	

In:Neurotropism	G3.09	
In:Vascular/Lymphatic	G3.05	
Li:Lexical Polarity Positive		
Li:Lexical Polarity Negative		NP, VP, PP, ADJP
Li:Modality		
Li:Mood and Comment Adjuncts		
Li:Temporality		
Ma:Excision Clear	S3.03	
Ma:Excision Deep	G3.02	
Ma:Excision Invasive	G3.02	
Ma:Excision In Situ	G3.02	
Re:Desmoplasia	G3.08	NP
Re:Fibrosis		
Re:Solar Elastosis		
Re:Tils	G3.06	
St:Clinical History Heading		
St:Comment Heading		
St:Diagnosis Heading		
St:Macroscopic Heading		NP
St:Microscopic Heading		NP
St:Specimen		
St:Subheading		
Sy:Diagnosis	S3.01, G1.05, G1.10	
Sy:Regression	G3.07	
Sy:Subtype	G3.12	

Table 4.2 Correspondence and boundary specification for entity types in the melanoma corpus. NP: noun phrase, VP: verb phrase, PP: prepositional phrase, ADJP: adjective phrase.

Entity type	**Protocol standard/guideline**	**Boundary specification**
De:Ancillary Studies	G4.01, G4.02	
De:Mesorectal Integrity	S2.11	
De:Perforation	S1.04, S2.09	
De:Peritoneal Reflection	S2.10	
De:Serosa Description	S2.12	
De:Specimen Blocks	S2.02	
De:Specimen Images	G2.02	
De:Specimen Size	S2.04	
De:Specimen Type	S1.07	
De:Tissue Banking	G2.01	
De:Tumour Description	S2.12	
De:Tumour Site	S1.06, S2.05	
De:Tumour Size	S2.06	
En:Coexistent Pathology	S3.10	
En:Distant Spread or Metastases	S1.11, S3.09	
En:Lymph Nodes	S2.03	
En:Residual Tumour	S1.09, S3.11, S5.02	
Ex:Donut Involvement		
Ex:Extent	S1.10	
Ex:Extramuscular Spread	S3.03	
Ex:Lymph Node Involvement	S3.07, G3.01	
Ex:Serosal Involvement	S3.03	
In:Depth of Invasion	S3.03	
In:Perineural Invasion	G3.02	
In:Venous and Small Vessel Invasion	S3.08	
Ma:Circumferential Margin	S2.08, S3.05, S3.06	
Ma:Clear	S3.04, S3.05, S3.06	

Ma:Proximal or Distal Margin	S2.07, S3.04	
Met:Anatomic Stage	S5.01	
Met:M Value	S5.01	
Met:N Value	S5.01	
Met:T Value	S5.01	
Re:Desmoplasia and Fibrosis		
Re:Response to Rx	S3.12	
Re:Tils and Peritumoural Lymphocytes		
St:Ancillary Studies Heading		
St:Clinical History Heading		NP
St:Macroscopic Heading		NP
St:Microscopic Heading		NP
St:Subheading		
St:Synthesis Heading		
Sy:Comment	S2.12, S5.03, G1.03, G3.03	
Sy:Histological Grade	S3.02	
Sy:Histological Type	S3.01	
Sy:Medical History	S1.08, G1.03	

Table 4.3 Correspondence and boundary specification for entity types in the colorectal cancer corpus. NP: noun phrase.

Entity type	**Protocol standard/guideline**	**Boundary specification**
An:Biomarker		
An:Cytogenetics Comment	S4.02	
An:Fish Results	S4.02	
An:Flow Cytometry-Comment	S4.02	
An:Flow Cytometry-Negative	S4.02	ADJP
An:Flow Cytometry-Positive	S4.02	ADJP
An:IgH Test	S4.01, S4.02	
An:Immunohistochemistry-Comment	S4.02	
An:Immunohistochemistry-Equivocal	S4.02	NP, ADJP
An:Immunohistochemistry-Negative	S4.02	NP, ADJP
An:Immunohistochemistry-Positive	S4.02	
An:PCR Comment	S4.01, S4.02	
An:TCRgamma Test	S4.02	
De:Anatomical Structure	S1.03	
De:Architecture	G3.01	
De:Cell Clonality		ADJP
De:Cell Size	G3.02	
De:Cytomorphology	G3.03	
De:Laterality	S1.04	NP, ADJP
De:Lineage	S5.01	NP
De:Other Size		
De:Preservative Fluid	S2.01	NP, ADJP
De:Sample Triage	S2.02	
De:Specimen Blocks		
De:Specimen Size	S2.04	
De:Specimen Type	S2.03	NP
De:Tissue Source	S1.03	
De:Topography	S1.03	
De:Tumour Size		
En:Coexistent Pathology		
En:Specimen Identifier		
Ex:Disease Extent	G1.05	NP
Ex:Other Sites of Disease	G1.05	
Li:Lexical Modality		

Li:Lexical Polarity Negative		
Li:Lexical Polarity Positive		
Li:Mood and Comment Adjuncts		
Li:Temporality		
Re:Tissue Reaction	G3.05	
St:Title-Clinical History		
St:Title-Comment		
St:Title-Frozen Section Report		
St:Title-Macroscopic Description		
St:Title-Microscopic Description		
St:Title-Nature and Specimen Type		
St:Title-Pathologist Notes		
St:Title-Special Investigations		
St:Title-Subheading		
St:Title-Summary		
St:Title-Supplementary Report		
St:Title-Supplementary Summary		
Sy:Clinical Impression	G1.04	NP, ADJP
Sy:Comment	S5.03	
Sy:Constitutional Symptoms	G1.07	
Sy:Diagnosis	S5.02	
Sy:Diagnosis Subtype	S5.02	
Sy:Indication for Biopsy	G1.03	
Sy:Medical History	G1.09	
Sy:Predisposing Factors	G1.10	
Sy:Presentation		
Sy:SNOMED RT Codes		
Sy:Stage	G5.04	
Sy:WHO Grade	S3.02	

Table 4.4 Correspondence and boundary specification for entity types in the lymphoma corpus. NP: noun phrase, ADJP: adjective phrase.

4.3.3 Main Annotation Exercise

Previous works have reported that the levels of expertise are not the critical factor for annotation consistency if explicit guidelines or sufficient training are provided.

Roberts et al have pointed out in their work that much of clinical text can be understood by a non-clinician armed with a medical dictionary, as it can be exposed by the linguistic constructs of the text, although some relationships between entities may require more domain knowledge to understand (Roberts et al., 2007).

Patrick et al's study also indicated that computational linguists can reliably achieve higher consistency than pathologists in annotating a large corpus of pathology reports (Patrick and Scolyer, 2008). Especially, the consistency of linguists between each other is consistently higher than that between the pathologists, or between a linguist and a pathologist. It suggested that once trained to understand the linguistic features and extent of pathology concepts, linguists are also capable of annotating pathology notes reliably.

Therefore, the following annotation process was mainly carried out by an annotation team composed of six linguists. They were given a few hours of a training session, focused on the annotation instructions and the guidelines before they started to do the main annotation exercise. If difficult cases were encountered, they could also turn to medical consultants (clinician and pathologist) for advice.

Double annotation is widely used in an annotation task, where each document is independently annotated by two annotators, and the sets of annotations compared for agreements. Agreements from the original annotators are accepted into a consensus set, and the third annotator adjudicates on differences between the original annotators and resolves them. Obviously, this method has several advantages compared to single annotation, such as reducing the idiosyncrasies of an individual annotator and avoiding one-off errors made by a single annotator. Nevertheless, it also requires more annotator labour and effort to be devoted to the task.

To address the difficulty of the annotation task, Roberts et al suggested some approaches (Roberts et al., 2009), which include:

- Active learning or mixed initiative approaches can be explored to utilize annotator effort most effectively. Since the annotation and system learning stages are integrated in these approaches so that except in the early stage, annotators only need to correct and augment the annotations that the system has added to a document rather than annotating the whole document from scratch. These approaches can reduce the amount of human annotator input so that human effort can be concentrated on more difficult cases.
- Another approach is to adopt a distributed and collaborative annotation framework in which the grain size of annotation instances is reduced to a snippet. This approach has many advantages, such as smaller annotation grain size indicates smaller levels of effort can be exploited and reduces the difficulty for annotators by focusing effort on single-decision types over small snippets of text; the annotation of individual instances can be repeated until it reaches a satisfactory level of agreement, or they can be eliminated if they turn out to be problematic.

Inspired by these approaches, a mixed conveyor method with a two phase validation was purposed to accomplish the task. The annotation team was divided into two groups. Each Group had a subset of the total categories to annotate. Each team member annotated the documents for those categories assigned to them. The team leader reviewed each annotation, as a validator for the development of the first gold standards. There are several benefits of using this method:

- It requires less time for an annotator to process a document, as each annotator only needs to annotate particular information of interests in the document rather than the whole document.
- It reduces the difficulty for annotation, since annotators can focus effort on the categories assigned to them instead of the total categories to make their decisions.

- The quality of annotation is reflected by not only the consistency, but also the correct application of the guidelines by annotators. Since the validator reviewed the annotations strictly followed the guidelines, ensuring the guidelines were applied correctly in most cases.

The annotators conferred and reconciled the differences where possible. Unresolved differences were passed to a pathologist to resolve. It is assumed that over 95% consistency was attained at last, as less than 5% differences remained unresolved.

4.3.4 Recursive Validation

The first gold standards were created manually, and thus may still contain minor errors and inconsistencies. These errors and inconsistencies can be identified by performing recursive validation on the training data with a 100% train and test strategy, which involves using 100% of the training set to build a simple model and then test on the same set until no improvement of the performance can be made.

With this recursive validation process, more than 80, 2200 and 700 potential errors were detected in the training data of the melanoma corpus, colorectal cancer corpus and lymphoma corpus respectively. The potential errors can either represent erroneous annotations or weaknesses in the computational processing. Erroneous annotations can be text that should not be annotated, omitting what should be annotated, assigning an instance with incorrect category and marking an instance with the wrong span. The weaknesses in the computational processing can indicate the issues to be addressed in the various processes (e.g., features prepared for machine learners). Some error types of entity annotation and their corrections are listed in Table 4.5. Each potential error was manually identified if it was an erroneous annotation, then it was corrected so that the model would not learn from the incorrect examples. This process improved the micro-averaged scores by about 0.3%, 3.1% and 0.4% for the melanoma, colorectal cancer and lymphoma corpora respectively.

4.4 Results

4.4.1 Entity Annotation

The frequency of annotations for each entity type after recursive validation is detailed in Tables 4.6, 4.7, 4.8.

There are in total 17470 annotated entities in the melanoma corpus, 3440 are in Linguistic categories and 1489 are in Structural categories. Medical entities account for 71.79% of all entities in the corpus. The highest frequency medical entities are En:Primary Lesion and Sy:Diagnosis, which account for 22.73% of medical entities. About 4.73% of medical entities are distributed into 8 rare entity types: En:Lesion (other), En:Satellites, In:Neurotropism, Ma:Excision Deep, Ma:Excision In Situ, Re:Desmoplasia, Re:Fibrosis, and Re:Solar Elastosis.

Error type	Comment for correction	Incorrect example	Correct example
Including or excluding punctuation	Except for the entailed punctuation of an instance (e.g., "." for "En:Specimen Identifier" or "Specimen Identifier"), "?" representing lexical modality, punctuation that constitute a abbreviation, punctuation should be excluded from the spans.	2 ["En:Specimen Identifier"]. The sections confirm the clinical diagnosis of malignant melanoma.	2. ["En:Specimen Identifier"] The sections confirm the clinical diagnosis of malignant melanoma.
Inconsistent annotations between "De:Specimen Type" and "De:Tumour Site"	Annotate the instances according to the local context. If keywords like "tumour" and "carcinoma" appear in to the local context, then the instance should be annotated as "De:Tumour Site" , otherwise, it should be assigned to "De:Specimen Type".	1. Left colon ["De:Tumour Site"]: A left sided resection measuring 350mm x30mm.	1. Left colon ["De:Specimen Type"] : A left sided resection measuring 350mm x30mm.
Inconsistent spans for instances of "Ex:Lymph node involvement"	Separate a long span to instances of "En:Lymph Nodes" and "Ex:Lymph node involvement" respectively if it is possible.	Eighteen definite lymph nodes are identified, with two containing metastatic tumour (= 2/18) ["Ex:Lymph node involvement"].	Eighteen definite lymph nodes are identified ["En:Lymph Nodes"], with two containing metastatic tumour (= 2/18) ["Ex:Lymph node involvement"].
Inconsistent spans for instances of the same type	Use shorter spans in most cases to get more atomic instances.	A16-A18 - Total of nine nodes (3 bisected nodes in each). A19 - Six lymph nodes (one was bisected) ["De:Specimen Blocks"].	A16-A18 - Total of nine nodes (3 bisected nodes in each) ["De:Specimen Blocks"]. A19 - Six lymph nodes (one was bisected) ["De:Specimen Blocks"].
Inconsistent annotations between "Presentation" and " Constitutional Symptoms"	Manually corrected by medical consultants.	Lethargy ["Sy:Constitutional Symptoms "].	Lethargy [" Sy:Presentation"].
Nested annotations	Do not annotate an instance nested in another instance with a longer span if it is possible.	CD3 and CD20 each stain a population of *small lymphocytes* *["De:Cell Size"]*["An:Immunohistochemistry-Comment"].	CD3 and CD20 each stain a population of small lymphocytes ["An:Immunohistochemistry-Comment"].

Table 4.5 Error types of entity annotation and their corrections.

Entity type	Number	Proportion	Entity type	Number	Proportion
De:Cell Growth Pattern	615	3.52%	Li:Mood and Comment Adjuncts	931	5.33%
De:Cell Type	694	3.97%	Li:Temporality	167	0.96%
De:Cosmetic Changes	266	1.52%	Ma:Excision Clear	241	1.38%
De:Dermal Mitoses	364	2.08%	Ma:Excision Deep	166	0.95%
De:Shape	555	3.18%	Ma:Excision Invasive	362	2.07%
De:Site and Laterality	817	4.68%	Ma:Excision In Situ	88	0.50%
De:Size	845	4.84%	Re:Desmoplasia	16	0.09%
De:Specimen Type	627	3.59%	Re:Fibrosis	68	0.39%
De:Ulceration	280	1.60%	Re:Solar Elastosis	25	0.14%
En:Associated Naevus (type)	222	1.27%	Re:Tils	212	1.21%
En:Lesion (other)	57	0.33%	St:Clinical History Heading	250	1.43%
En:Primary Lesion	1612	9.23%	St:Comment Heading	28	0.16%
En:Satellites	24	0.14%	St:Diagnosis Heading	258	1.48%
En:Specimen Identifier	842	4.82%	St:Macroscopic Heading	363	2.08%
In:Breslow Thickness (mm)	508	2.91%	St:Microscopic Heading	375	2.15%
In:Clark Level	742	4.25%	St:Specimen	164	0.94%
In:Neurotropism	149	0.85%	St:Subheading	51	0.29%
In:Vascular/Lymphatic	229	1.31%	Sy:Diagnosis	1238	7.09%
Li:Lexical Polarity Positive	1369	7.84%	Sy:Regression	201	1.15%
Li:Lexical Polarity Negative	676	3.87%	Sy:Subtype	476	2.72%
Li:Modality	297	1.70%	Overall	17470	

Table 4.6 Entity frequency for the melanoma corpus.

It is not surprising that the En:Primary Lesion and Sy:Diagnosis are at the top of the most frequent medical entity types, as most of the other findings are based on the identification of Primary Lesion; Diagnosis is usually the most important issue that should be addressed in a melanoma pathology report, and it can appear in any section of a note.

Up to 29807 entities are annotated in total in the colorectal cancer corpus, wherein medical entities account for 87.08%, and the remaining 12.92% are in Structural categories. De:Specimen Blocks and De:Specimen Type have the highest frequencies, accounting for 20.17% of medical entities. Eight entity types have the lowest frequencies (below 0.5% each): De:Mesorectal Integrity, De:Peritoneal Reflection, De:Specimen Images, De:Tissue Banking, En:Residual Tumour, Ex:Donut Involvement, Ex:Extramuscular Spread, and Re:Response to Rx.

Most of the documents contain more than one De:Specimen Blocks instance as it is recommended that the pathologist should take sufficient blocks (generally at least 4) to fully assess all the necessary parameters for staging and prognosis. De:Specimen Type could appear more than once in a document as the clinician or pathologist tended to repeat it in different sections. Thus, both these entity types have high frequencies in the corpus.

Entity type	Number	Proportion	Entity type	Number	Proportion
De:Ancillary Studies	272	0.91%	In:Perineural Invasion	396	1.33%
De:Mesorectal Integrity	15	0.05%	In:Venous and Small Vessel Invasion	855	2.87%
De:Perforation	154	0.52%	Ma:Circumferential Margin	254	0.85%
De:Peritoneal Reflection	138	0.46%	Ma:Clear	758	2.54%
De:Serosa Description	212	0.71%	Ma:Proximal or Distal Margin	699	2.35%
De:Specimen Blocks	3343	11.22%	Met:Anatomic Stage	291	0.98%
De:Specimen Images	30	0.10%	Met:M Value	257	0.86%
De:Specimen Size	1585	5.32%	Met:N Value	409	1.37%
De:Specimen Type	1892	6.35%	Met:T Value	414	1.39%
De:Tissue Banking	57	0.19%	Re:Desmoplasia and Fibrosis	271	0.91%
De:Tumour Description	1464	4.91%	Re:Response to Rx	106	0.36%
De:Tumour Site	1446	4.85%	Re:Tils and Peritumoural Lymphocytes	389	1.31%
De:Tumour Size	682	2.29%	St:Ancillary Studies Heading	52	0.17%
En:Coexistent Pathology	1181	3.96%	St:Clinical History Heading	378	1.27%
En:Distant Spread or Metastases	284	0.95%	St:Macroscopic Heading	387	1.30%
En:Lymph Nodes	629	2.11%	St:Microscopic Heading	388	1.30%
En:Residual Tumour	124	0.42%	St:Subheading	2121	7.12%
Ex:Donut Involvement	144	0.48%	St:Synthesis Heading	524	1.76%
Ex:Extent	610	2.05%	Sy:Comment	1558	5.23%
Ex:Extramuscular Spread	197	0.66%	Sy:Histological Grade	828	2.78%
Ex:Lymph Node Involvement	1133	3.80%	Sy:Histological Type	998	3.35%
Ex:Serosal Involvement	392	1.32%	Sy:Medical History	206	0.69%
In:Depth of Invasion	1284	4.31%	Overall	29807	

Table 4.7 Entity frequency for the colorectal cancer corpus.

The total amount of annotated entities in the lymphoma corpus is 19255, wherein 2553 are in Linguistic categories and 1765 are in Structural categories, and the remaining 77.57% are medical entities. The most frequently annotated medical entities are An:Biomarker, De:Tissue Source and De:Topography, with about 32.23%. There are 211 medical entities distributed sparsely into 15 entity types (each has no more than 50 instances), wherein De:Cell Clonality, Ex:Disease Extent, An:Fish Results, An:Flow Cytometry-Negative, An:Flow Cytometry-Positive, An:IgH Test, Sy:Stage and An:TCRgamma Test have the lowest frequencies (each with less than 10 occurrences).

Entity type	Number	Proportion	Entity type	Number	Proportion
An:Biomarker	1928	10.01%	Ex:Other Sites of Disease	53	0.28%
An:Cytogenetics Comment	18	0.09%	Li:Lexical Modality	322	1.67%
An:Fish Results	8	0.04%	Li:Lexical Polarity Negative	366	1.90%
An:Flow Cytometry-Comment	88	0.46%	Li:Lexical Polarity Positive	1125	5.84%
An:Flow Cytometry-Negative	3	0.02%	Li:Mood and Comment Adjuncts	607	3.15%
An:Flow Cytometry-Positive	3	0.02%	Li:Temporality	133	0.69%
An:IgH Test	7	0.04%	Re:Tissue Reaction	259	1.35%
An:Immunohistochemistry-Comment	246	1.28%	St:Title-Clinical History	224	1.16%
An:Immunohistochemistry-Equivocal	27	0.14%	St:Title-Comment	36	0.19%
An:Immunohistochemistry-Negative	284	1.47%	St:Title-Frozen Section Report	39	0.20%
An:Immunohistochemistry-Positive	594	3.08%	St:Title-Macroscopic Description	223	1.16%
An:PCR Comment	27	0.14%	St:Title-Microscopic Description	226	1.17%
An:TCRgamma Test	2	0.01%	St:Title-Nature and Specimen Type	123	0.64%
De:Anatomical Structure	396	2.06%	St:Title-Pathologist Notes	83	0.43%
De:Architecture	472	2.45%	St:Title-Special Investigations	47	0.24%
De:Cell Clonality	1	0.01%	St:Title-Subheading	463	2.40%
De:Cell Size	488	2.53%	St:Title-Summary	226	1.17%
De:Cytomorphology	178	0.92%	St:Title-Supplementary Report	38	0.20%
De:Laterality	18	0.09%	St:Title-Supplementary Summary	37	0.19%
De:Lineage	140	0.73%	Sy:Clinical Impression	185	0.96%
De:Other Size	123	0.64%	Sy:Comment	77	0.40%
De:Preservative Fluid	110	0.57%	Sy:Constitutional Symptoms	28	0.15%
De:Sample Triage	793	4.12%	Sy:Diagnosis	1056	5.48%
De:Specimen Blocks	847	4.40%	Sy:Diagnosis Subtype	30	0.16%
De:Specimen Size	467	2.43%	Sy:Indication for Biopsy	28	0.15%
De:Specimen Type	802	4.17%	Sy:Medical History	82	0.43%
De:Tissue Source	1482	7.70%	Sy:Predisposing	60	0.31%

			Factors		
De:Topography	1404	7.29%	Sy:Presentation	98	0.51%
De:Tumour Size	51	0.26%	Sy:SNOMED RT Codes	994	5.16%
En:Coexistent Pathology	160	0.83%	Sy:Stage	3	0.02%
En:Specimen Identifier	675	3.51%	Sy:WHO Grade	134	0.70%
Ex:Disease Extent	8	0.04%	overall	19255	

Table 4.8 Entity frequency for the lymphoma corpus.

The possible reasons for the high frequencies on An:Biomarker, De:Tissue Source and De:Topography are:

- Biomarkers are requisite for many ancillary studies, especially for immunohistochemistry tests; an ancillary study usually involves more than one biomarker.
- A specimen can consist of different kinds of tissues, the source of which is captured by De:Tissue Source.
- De:Topography represents the anatomical site of a biopsy or operation, which appears extensively in Clinical History, Specimen and Summary sections.

The protocol did not explicitly specify what ancillary study should be performed by the pathologist. It seems that the pathologist tended to perform more immunohistochemistry tests than other ancillary tests (e.g., FISH tests, Flow Cytometry, IgH tests and TCRgamma tests) according to the low frequencies on the associated entity types in the corpus.

There are 38592, 163293 and 49799 tokens annotated as entities in the melanoma, colorectal cancer and lymphoma corpora respectively, thus the average lengths of entities are 2.21, 5.48 and 2.59 respectively. It suggests that the pathologists preferred to use longer terms or descriptions to depict their findings, procedures or diagnoses in the colorectal cancer corpus. A deeper analysis shows that in this corpus, up to 18 entity types have an average length of over 6, wherein those of two types (De:Ancillary Studies and De:Tissue Banking) are more than 10. The colorectal cancer corpus also has the largest entity density, which is 72.68%, as the entities of which outnumbered those of other two corpora; whereas, the smallest entity density is not in the melanoma corpus (53.76%), but in the lymphoma corpus (43.91%) instead. The possible reason is that compared to the lymphoma corpus, although the melanoma corpus has smaller amount of tokens annotated as entities, it has much smaller amount of tokens in total (71786 *vs.*113413). A comparison of the entities among the three corpora is presented in Table 4.9.

	Melanoma corpus	Colorectal cancer corpus	Lymphoma corpus
No. of tokens	71786	224660	113413
No. of entity types	41	45	63
No. of entities	17470	29807	19255
Average length	2.21	5.48	2.59
Entity density	53.76%	72.68%	43.91%

Table 4.9 Comparison of entity densities among the three corpora.

4.4.2 Relation Annotation

The relations were annotated after the annotation of entities. Annotators could only mark a relationship between two existing entities in the lymphoma corpus. Table 4.10 lists the distribution of relation types in the corpus.

Relation type	Sentence distance		Number	Proportion
	0	1		
Negate	318	0	318	14.12%
Spatial Specialization	12	0	12	0.53%
Result-Positive	937	10	947	42.05%
Result-Negative	924	16	940	41.74%
Result-Equivocal	35	0	35	1.55%
Overall	2226	26	2252	

Table 4.10 Distribution of relation types and sentence distance between two entities in the lymphoma corpus.

The corpus consists of 2252 relations in total, of which 318 are Negate relations and 1922 are Result relations. Not all entity types were covered by a relation, as mentioned in the relation schema. There are 20 types of entities selected for relation annotation.

Table 4.10 also shows the number of sentence distance for each relation. Among all annotated relations, 26 relations are inter-sentential, which is 1.15%, and all inter-sentential relations are from Result relations, where each connected entities are located in adjacent sentences. This is probably because the annotations of Negate and Spatial Specialization relations could not cross the sentence boundaries according to the guidelines. It is assumed that the inter-sentential relations may be harder to recognise, due to their longer distances between entities.

4.5 Discussion

Several exsiting annotated corpora for IE in the clinical domain are not suitable for the tasks in this study, mainly due to entity types annotated by using standard terminologies. A preliminary study shows that less than half of the concepts in the corpora could be annotated with medical categories from SNOMED CT. Thus it is necessary to prepare semantically annotated corpora for this study.

Annotating pathology notes is quite difficult, as it not only needs linguistic knowledge, but also a considerable amount of medical knowledge. Explicit annotation guidelines are very important to ensure high quality annotation due to the high variability of the medical vocabulary used and personal writing styles that pathologists presented in the notes.

There are two kinds of distinguishable opinions for designing an annotation schema. Some advocate that to reduce the difficulty of the annotation task, the annotation schema should narrow the scope and

simply focus on fewer entities or relations (Roberts et al., 2009). Others argue that to achieve better consistency, less ambiguity and greater coverage of the concepts in the corpus, a possible solution is to create fine-grained categories by dividing the top categories into smaller classes along the terminology's hierarchy (Wang, 2009). The former strategy may not be possible for an intended application, as the entities or relations designed in the schema can be too general to meet the requirements of the application. The latter strategy requires the annotators to spend more effort and use deeper domain knowledge, consequently increase the difficulty of the annotation task. It is evident that appropriate granularity is a crucial factor for designing an annotation schema.

One of the indicators for assessing the quality of the annotation schema in this study is the correspondence between the entity types and standards or guidelines in the protocols. The detail correspondence for matching the entity types to the standards or guidelines is displayed in Tables 4.2, 4.3, 4.4. It can be seen that most medical entity types have at least one corresponding standard or guideline, suggesting that the annotation schemas have appropriate granularity that can captured most of structured template related information without too much ambiguity.

Entity types in the Linguistic and Structural categories cannot be made to correspond with any standard or guideline, as these were designed with linguistic knowledge rather than medical knowledge from the protocols. There may be several reasons why some medical entity types do not have corresponding standard or guideline as well:

- They represent structured template related information that is not described in the standards or guidelines (e.g., Re:Tils and Peritumoural Lymphocytes for the colorectal cancer corpus), revealing the deficiency of the protocols.
- They are critical for the consruction of a structured template (e.g., En:Specimen Identifier for the melanoma corpus and lymphoma corpus), although they are not indicated in the standards or guidelines.
- They were designed with the intention to facilitate a distinction from similar entity types. For example, for the lymphoma corpus, De:Tumour Size and De:Other Size were designed to distinguish the two.
- Medical consultants thought they may be of clinical significance, but not defined by any standard or guideline in the protocols (e.g., De:Cell Type for melanoma corpus).

Likewise, there are some standards or guidelines without associated entity types, possibly because:

1. These standards or guidelines need combinations of multiple entity types to represent. For instance, G5.01 in the Colorectal Cancer Structured Reporting Protocol (Eckstein et al., 2010) requires a combination of De:Specimen Type, De:Tumour Site, Sy:Histological Type, Met:Anatomic Stage, Met:M Value, Met:N Value, Met:T Value and En:Residual Tumour to satisfy the conditions.
2. These standards or guidelines indirectly relate to the entity types in fact. For example, for the melanoma corpus, St:Comment Heading does not have direct connection to S5.02 in the

Primary Cutaneous Melanoma Structured Reporting Protocol (Scolyer et al., 2010), but the recognition is definitely associated with it.

It can be also observed that most of the entity types have a one-to-one match to a standard or guideline in the protocol. However, certain entity types can be matched to more than one standard or guideline, such as Sy:Diagnosis matched to S3.01, G1.05 and G1.10 in the Primary Cutaneous Melanoma Structured Reporting Protocol (Scolyer et al., 2010), and Sy:Comment matched to "S2.12, S5.03, G1.03 and G3.03" in the Colorectal Cancer Structured Reporting Protocol (Eckstein et al., 2010), implying that the information captured by these entity types can embed extensively in different sections in a report; or the goal of designing these entity types is to represent broader scope of involved standards or guidelines. Similarly, some standards or guidelines have more than one matched entity type. For example, G3.02 in the Primary Cutaneous Melanoma Structured Reporting Protocol (Scolyer et al., 2010) can match to Ma:Excision Deep, Ma:Excision Invasive and Ma:Excision In Situ. As mentioned in the previous chapter, these standards or guidelines require co-occurrence of multiple elements to be correctly represented, so it is preferable to design multiple entity types rather than a single one for them. This can be also due to the over-specific definitions of the entity types that cannot cover all the aspects the standards or guidelines may involve. For instance, S3.03 in the Colorectal Cancer Structured Reporting Protocol (Eckstein et al., 2010) have three matchs: Ex:Extramuscular Spread, Ex:Serosal Involvement, and In:Depth of Invasion. Each of them represents one aspect of the standard, as the bowel wall can have several layers: mucosa, submucosa, muscularis propria, subserosa and serosa; Ex:Extramuscular Spread and Ex:Serosal Involvement focus on muscularis propria and serosa respectively, while In:Depth of Invasion can depict the local invasion of all layers except for serosa.

One of the difficulties for the annotation task is the annotation of modifiers. Determining the annotation definition for a modifier is not easy as it requires both linguistic and domain knowledge. Some modifiers play important roles in constituting medical entities, while others are attributes of medical entities, or indicators of the intensity or degree of medical entities. Given the roles the modifiers play, there can be two distinct annotations to be made: for the former role, the modifiers should be annotated as part of the medical entities; for the latter one, the modifiers should be annotated as Li:Mood and Comment Adjuncts for the melanoma corpus and the lymphoma corpus. However, since there are no Linguistic categories in the colorectal cancer corpus, the phrases referring to them should always be annotated as part of the medical entities if they are clinically significant for the entities. In the following examples, the modifier *small* is annotated as a linguistic entity in the first example and part of a medical entity in the remaining examples:

- Occasional **small** ["Li:Mood and Comment Adjuncts"] foci of TILs suggest early regression.
- The **small** polyp identified macroscopically is a moderately dysplastic tubular adenomas ["En:Coexistent Pathology"].
- Mediastinal mass - Favour **small** cell B cell lymphoma ["Sy:Diagnosis"], await second opinion

The annotation of negation or uncertainty phrases is another difficult issue encountered during the task.

Since there are no explicit or coherent guidelines released for the annotation of negation phrases at present, and one of our research questions is to find out how the annotation of negation phrases can affect negation detection, two main strategies were chosen to annotate the negation phrases:

- For the melanoma and lymphoma corpora, the negation phrases were annotated as Li:Lexical Polarity Negative instances. However, there are minor differences for the boundary of this Linguistic category in the corpora: in the melanoma corpus, Li:Lexical Polarity Negative excludes prepositions, while they can be included in the lymphoma corpus. For example, "no evidence of" would be annotated as "no ["Li:Lexical Polarity Negative"] evidence ["Li:Lexical Polarity Positive"] of" in the melanoma corpus, but "no evidence of ["Li:Lexical Polarity Negative"]" in the lymphoma corpus.
- For the colorectal cancer corpus, a negation phrase is annotated as part of a medical entity if only this entity is negated by the negation phrase in the sentence or the definition of the entity has implicitly indicated to include negations; otherwise, if a negation phrase negates more than one entity, it should not be annotated. Here are three examples:
    1. 14 lymph nodes are identified and they show no evidence of metastatic tumour ["Ex:Lymph Node Involvement"]. The negation phrase "no evidence of" suggests that the identified lymph nodes are malignantly uninvolved, thus it is annotated as part of an Ex:Lymph Node Involvement instance.
    2. There is no lymphovascular ["In:Venous and Small Vessel Invasion"] or perineural ["In:Perineural Invasion"] invasion. As the negation phrase "no" negates two medical entities, it is not annotated.
    3. Extramural venous involvement is not seen ["In:Venous and Small Vessel Invasion"]. The negation phrase "not" only negates one entity, hence, it is annotated as part of the entity.

Similarly, the strategies for annotating uncertainty phrases are: for melanoma and lymphoma corpora, the uncertainty phrases are annotated as Li:Modality or Li:Lexical Modality instances; for the colorectal cancer corpus, if only one medical entity is asserted by an uncertainty phrase in the sentence or uncertainty can be included in the definition of the entity, the uncertainty phrase should be annotated as part of the entity; otherwise, if more than one entity is asserted by the uncertainty phrase, it should not be annotated.

The annotation task reveals that pathology reports are distinguishable from other clinical notes. The melanoma corpus is compared to discharge summaries from the 2010 i2b2/VA Challenges (Uzuner et al., 2011) in the following aspects:

1. Section headers. Section headers in a pathology report are more fixed, which can be summarized to six types in the melanoma corpus: “Clinical History”, “Specimen”, “Macroscopic”, “Microscopic”, “Diagnosis” and “Comment”; section headers in a discharge summary are more diverse, including “Chief Complaint”, “Past Medical History”, “Discharge Medications”, “Discharge Diagnosis”, etc.
2. Scope of a medical entity. Annotation guidelines of 2010 i2b2/VA Challenges pointed out that “Only complete noun phrases and adjective phrases should be marked”. Nevertheless, the scope of an entity in a pathology report can be more flexible. For example, a verb “shows” for Li:Lexical Polarity Positive, multiple noun phrases like “Breslow thickness 1.6mm” for In:Breslow thickness (mm), a clause or sentence like “mitotic rate is 15 to 18 per mm2” for De:Dermal mitoses.
3. Focus of medical entity types. 2010 i2b2 Challenges mainly focused on three types of medical entity: Problem, Test and Treatment. Medical entity types of a pathology report can be more specific and detailed. For instance, De:Site and Laterality, De:Cell Growth Pattern, En:Primary Lesion, etc., and there are up to 41 entity types in a melanoma pathology report. Thus, it requires more domain and linguistic knowledge, and training to annotate a pathology report.

It can be learned from the annotation task that although the annotators for the main annotation process were linguists instead of pathologists, they were competent to accomplish the task. Since most of the clinical text can be understood by the linguists and the meaning of most entities can be determined by the linguistic constructs of the texts once the linguists are trained, they were able to annotate most entities, using their linguistic knowledge rather than medical knowledge. For some difficult cases, such as abbreviations or acronyms, they also needed to resolve with medical knowledge. They could either look up the dictionaries or textbooks (e.g., NCI Dictionary of Cancer Terms Histology (NCI, 2008-2014) and Cell Biology: An Introduction to Pathology (Kierszenbaum and Tres, 2011), and Colorectal Cancer: Multimodality (Saltz, 2002)) or ask the medical consultants for advice. Since the annotators have similar background, their annotations are relatively consistent with the guidelines.

4.6 Conclusion

This chapter described three semantically annotated corpora: melanoma corpus, colorectal cancer corpus and lymphoma corpus, which were annotated with entities and relationships between the entities. The annotation process was described, including the design of the annotation schemas and guidelines, and main annotation process.

To represent structured templates with related information in the pathology notes, there are up to 29 and 39 types of medical entities annotated in the melanoma and colorectal cancer corpora respectively, 46 types of medical entities and 5 types of relationships annotated in the lymphoma corpus. Some Linguistic and Structural categories were added to the schemas as suggested by the

computational linguists, resulting in 12, 6 and 17 additional entity types annotated in the melanoma corpus, colorectal cancer corpus and lymphoma corpus respectively. The correspondence analysis shows that the annotation schemas have appropriate granularity that can capture most of the structured template related information without too much ambiguity.

A mixed conveyor method was used to improve the efficiency and reduce the difficulty of the annotation task. Linguists carried out the main annotation process, and they were capable of accomplishing the task. Furthermore, recursive validation was performed on the first gold standards to attain higher consistency among the annotations.

These corpora can be used as resources to support training and evaluating the information extraction systems built to extract information from pathology notes. Although the size of each corpus is small, owing to limited time and resources, it is believed that their unique annotations with high quality make them suitable for future experiments. They can be good supplementary materials to the clinical NLP research as well.

Chapter 5 Medical Entity Recognition

5.1 Introduction

Entity recognition (ER) is one of the key components of an information extraction (IE) system. As defined by the Message Understanding Conference-6 (MUC-6) (Grishman and Sundheim, 1996), it is a task that automatically locates references of interest in natural languages and classifies them into predefined categories. The predefined categories vary in different domains. In the general domain, they can be person, location, organization, date and so on; in the biomedical domain, they usually refer to proteins, genes, chemicals, etc.; in the clinical domain, they are likely to be disorders, signs or symptoms, anatomical sites, medications, and procedures. Moreover, in a particular sub-domain, they can be problem-specific. For example, identification of medication information from discharge summaries was the main theme of the 2009 i2b2/VA Challenge, where seven categories were defined: dosages, modes of administration, frequencies, durations, and reasons for administration (Uzuner et al., 2010). In the pathology domain, the predefined categories are more specific and detailed. As most categories to be classified are based on medical knowledge, they are named as medical entity types, and the task of recognising them is called medical entity recognition (MER).

MER should be discriminated from another similar task – named entity recognition (NER), as NER restricts the task to identify rigid designators as defined by Kripke (Kripke, 1980), including proper names, certain natural terms like biological species and substances. In the clinical domain, the named entities are terms in standard terminologies (e.g. UMLS, SNOMED CT and ICD-9). As mentioned in the previous chapter, the medical entities which this study attempts to identify are not restricted to them, thus MER can be more complicated than NER.

Medical entities are distributed extensively in the pathology notes according to the statistics of entity densities in the previous chapter, hence recognition of these entities provides opportunities to extract useful information embedded in the notes so that the information can be used to track the performance of pathologists and facilitate communication between the clinical staff and pathologists. MER can also improve the efficiency of reading the pathology notes, as the clinical staff or pathologists can identify the contents of interest rapidly through the highlighting of the entities. It can ease the data retrieval, automatic encoding and indexing by medical informaticians and researchers as well. Moreover, it is crucial for more advanced IE tasks such as negation and uncertainty detection, relation extraction, and structured output generation.

In the previous chapter, three semantically annotated corpora have been described along with the issues of developing training data to support learning of the machine learners and extraction of rules. In this chapter, a supervised machine-learning based-approach is developed to recognise medical entities from the corpora. Specifically, an MER system using Conditional Random Fields (CRF) and integration with various features is presented. This chapter firstly describes the overview of CRF and

evaluation methods, and then follows with some pre-processing (e.g., tokenisation, sentence boundary detection and proof reading), the descriptions of features, the experimental results and discussion.

5.2 Conditional Random Fields

Conditional Random Fields (CRF) is a framework for building probabilistic models to segment and label sequence data. Given a particular observation sequence, CRF defines a conditional probability distribution over label sequences rather than a joint distribution over both label and observation sequences. In CRF, models can be trained by learning the conditional distributions between the labels and features from the observations, and then they can be used to predict the most likely assignment to a new label.

It is a family of discriminative models first proposed by Lafferty et al (Lafferty et al., 2001). Its definition is as follows:
Let X be a random variable over the data, Y be the random variable over a label sequence, G = (V, E) be a graph where Y is indexed by the vertices of G, so that $Y = Y_v, v \in V$. The conditional random field (X, Y) is defined as when conditioned on X, the random variables Y_v has the Markov property with respect to the graph G: $p(Y_v \mid X, Y_w, w \neq v) = p(Y_v \mid X, Y_w, w \sim v)$, where w ~ v means that they are neighbours in G.

The simplest and important structure of G is a linear chain, which is very close to the nature of the text (a sequence of words), where X and Y are assumed to have the same length. The generic input sequence is denoted by $x = x_1, x_2, \cdots x_n$, and the label sequence is denoted by $y = y_1, y_2, \cdots y_n$. Figure 5.1 presents the graphical structure of linear chain CRF.

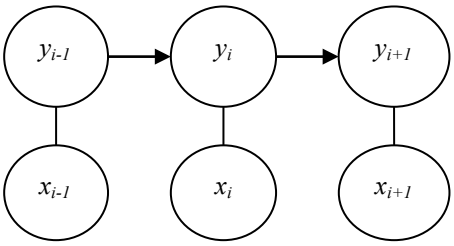


Figure 5.1 Graphical structure of linear chain CRF.

CRF (X, Y) is specified by two vectors: vector *f* stands for local features, and λ is the parameter vector learned weight for the feature vector.

The global feature vector *F* is given by

$$F(y, x) = \sum_i f(y, x, i)$$

where *i* is the current position.

Then the conditional probability of a state sequence given an input sequence in the linear chain CRF is

$$p_\lambda(Y|X) = \frac{\exp \lambda \cdot F(Y,X)}{Z_\lambda(X)}$$

where $Z_\lambda(X)$ is a normalisation factor of all label sequence, and given by

$$Z_\lambda(X) = \sum_y \exp \lambda \cdot F(y,x)$$

Since $Z_\lambda(X)$ does not depend on *y*, the best label sequence *y* for input sequence x can be found to maximise the following function:

$$\hat{y} = \arg\max_y p_\lambda(y|x) = \arg\max_y \lambda \cdot F(y,x)$$

It can be computed with the Viterbi algorithm.

The maximum likelihood principle is applied to estimate the weight vector λ. Given a set of training data $\{(x^{(1)}, y^{(1)}), (x^{(2)}, y^{(2)}), \cdots (x(k), y(k)\}$, the maximum likelihood principle finds its values by

$$L_\lambda = \sum_k \log p_\lambda(y_k|x_k) = \sum_k [\lambda \cdot F(y_k, x_k) - \log Z_\lambda(x_k)]$$

The optimization can be found by setting the partial derivative with respect to each parameter in λ to zero, which is represented as:

$$\nabla L_\lambda = \sum_k [F(y_k, x_k) - E_{p_\lambda(Y|x_k)} F(Y, x_k)]$$

Define the transition matrix *M* for *x* at position *i*:

$$M_i[y, y'] = \exp \lambda \cdot f(y, y', x, i)$$

where *y*, y' are labels.

The expectation $E_{p_\lambda(Y|x)} F(Y,x)$ can be computed efficiently using a variant of the forward-backward algorithm:

$$E_{p_\lambda(Y|x)} F(Y,x) = \sum_y p_\lambda(y|x) F(y,x) = \sum_i \frac{\alpha_{i-1}(f_i * M_i)\beta_i^T}{\alpha_n \cdot 1^T}$$

where $*$ stands for component-wise matrix product, α_i and β_i are the forward and backward state-cost vectors, denoted by

$$\alpha_i = \begin{cases} \alpha_{i-1} M_i & 0 < i \leq n \\ 1 & i = 0 \end{cases}$$

$$\beta_i^T = \begin{cases} M_{i+1}\beta_{i+1}^T & 1 \leq i < n \\ 1 & i = n \end{cases}$$

Lafferty et al described two iterative scaling algorithms for CRF training (Lafferty et al., 2001). They are very simple and guaranteed to converge, but the convergence is very slow when involving many correlated features, as pointed out in Minka and Malouf's works (Malouf, 2002; Minka, 2003). Conjugate gradient and second-order methods, such as preconditioned conjugate-gradient (Shewchuk, 1994) and limited-memory quasi-Newton (Nocedal and Wright, 1999) can speed up CRF training.

The CRF-based MER systems have achieved the state of art performance without further post-processing in the past i2b2 challenges (Uzuner et al., 2010; Uzuner et al., 2011), indicating that it is one of the best machine learners for MER tasks.

The MER task can be formulated as a sequential labelling task, where each token needs to be assigned with a label in a sequence. There were several representations to represent the associations between the token and the entity (Shen and Sarkar, 2005). IOB2 notation (Sang and Erik, 2002) is selected to represent entities in this task, which has been widely used in other ER tasks. Each token in a sentence is represented with one of the B, I, O tags, where tag B shows the current token is at the beginning of an entity, I denotes the current token is inside an entity, and O indicates the current token is outside any entity. Therefore, for *N* entity types, there will be *2N +1* BIO tag types in total. The input of a CRF learner is a sequence of observed instances, and the output of a CRF learner is a sequence of BIO tags. Figure 5.2 displays a BIO representation of a sentence in the melanoma corpus.

Token	BIO tag
The	O
appearances	O
are	O
those	O
of	O
in-situ	B-SY:DIAGNOSIS
melanoma	I-SY:DIAGNOSIS
of	O
superficial	B-SY:SUBTYPE
spreading	I-SY:SUBTYPE
type	I-SY:SUBTYPE
.	O

Figure 5.2 The BIO representation of the sentence: “The appearances are those of in-situ melanoma of superficial spreading type.”

5.3 Evaluation Methods

5.3.1 Evaluation Metrics

As in an information retrieval system, evaluation metrics such as precision, recall, and F-score are also widely employed in an IE system. True positive, false positive, false negative and true negative are four important elements for computing these metrics.

True positive (TP): the number of correctly predicted instances by the system.
False positive (FP): the number of incorrectly predicted instances by the system.
False negative (FN): the number of incorrectly rejected instances by the system.
True negative (TN): the number of correctly rejected instances by the system.

The following scoring functions can be calculated by the above elements:
Precision (P) is the ratio between the number of correctly predicted instances and the total amount of instances predicted by the system. It stands for the accuracy of the predictions made by the system.

$$P = \frac{TP}{TP + FP}$$

Recall (R) is the ratio between the number of correctly predicted instances and the total amount of instances. It assesses the coverage of the predictions made by the system.

$$R = \frac{TP}{TP + FN}$$

F-score (F) is the harmonic mean of precision and recall. Generally, it can be calculated as:

$$F_\beta = \frac{(\beta^2 + 1) \times P \times R}{\beta^2 \times P + R}$$

where β gives a weight to precision and recall. When β > 1, it weights recall higher than precision; β < 1, it puts more emphasis on precision than recall; β = 1, equally weights precision and recall. All the experiments in this thesis use β = 1, and the F-score, also known as the F1 measure, is calculated as:

$$F_1 = \frac{2 \times P \times R}{P + R}$$

The higher F-score the system attains, the better performance the system achieves.

5.3.2 Cross-validation

Cross-validation is a model validation technique for assessing the generality of the results of a statistical analysis to an independent data set (Stone, 1974). Its main purpose is for researchers to estimate how accurately a predictive model will perform in practice. By defining test sets to test the model in the training phase, it can limit some problems like over-fitting.

In *n*-fold cross-validation, the original data set is randomly partitioned into *n* sub-sets of equal size. Of the *n* sub-sets, a single sub-set is retained for testing the model, and the remaining $n - 1$ sub-sets are used as training data. The process is repeated *n* times, each time with different subset for testing, and then the results from each fold can be averaged to produce a single estimation. Ten-fold cross-validation was used in most of the experiments in this study.

5.3.3 Matching Criteria

Traditional evaluation in NER tasks like MUC, used exact match as the standard matching criterion. Exact match requires both the boundaries and type of the entity to be in agreement with the gold-standard. However, in some IE tasks, such as negation detection and relation extraction, the exact boundary match for entities is not essential for determining valid instances.

By studying some commonly used matching criteria in biomedical NER tasks, Tsai et al suggested that it is not necessary to apply exact boundary match in some cases, while left or right boundary match may be sufficient (Tsai et al., 2006). In the evaluations of past i2b2 challenges, partial match was also considered in the metrics (Uzuner et al., 2010; Uzuner et al., 2011).

In the following evaluation for MER, both exact match and partial match criteria would be applied, wherein partial match criteria can be sub-classified into left boundary match, right boundary match and sloppy match (if the boundary of system prediction overlaps with that of the gold-standard). Figure 5.3 displays examples of different matching criteria.

Entity	Gold-standard	Exact match	Left boundary match	Right boundary match	Sloppy match
diffuse	B-SY:DIAGNOSIS	B-SY:DIAGNOSIS	B-SY:DIAGNOSIS	O	O
malignant	I-SY:SUBTYPE	I-SY:SUBTYPE	I-SY:SUBTYPE	O	B-SY:SUBTYPE
lymphoma	I-SY:SUBTYPE	I-SY:SUBTYPE	I-SY:SUBTYPE	O	I-SY:SUBTYPE
of	I-SY:SUBTYPE	I-SY:SUBTYPE	O	O	O
large	I-SY:SUBTYPE	I-SY:SUBTYPE	O	B-SY:SUBTYPE	O
B-cells	I-SY:SUBTYPE	I-SY:SUBTYPE	O	I-SY:SUBTYPE	O

Figure 5.3 Examples of exact match, left boundary match, right boundary match and sloppy match.

5.4 Pre-processing

Like other IE tasks, pre-processing is also required by an MER task. Some typical pre-processing is described below.

5.4.1 Sentence Boundary Detection

Background

It is important to detect sentence boundaries because other tasks are performed at the sentence level. In the general domain, the Wall Street Journal (WSJ) corpus and Brown corpus were typically used as the training or evaluation data. For instance, Palmer and Hearst developed a trainable algorithm composed of a lexicon with part-of-speech probabilities and a feed-forward neural network, which attained 98.5% accuracy on the WSJ corpus (Palmer and Hearst, 1994). Reynar and Ratnaparkhi reported that they achieved accuracies of 98.8% on the WSJ corpus and 97.9% on the Brown corpus by designing a maximum entropy model based algorithm for sentence boundary detection (Reynar and Ratnaparkhi, 1997). In the biomedical or clinical domains, some researchers adopted rule-based methods to detect sentences. For example, a rule-based system was developed by Xuan et al using

specific dictionaries, and had error rates of below 0.3% on their evaluation (Xuan et al., 2007). There were also some clinical information systems such as Mayo clinical Text Analysis and Knowledge Extraction System (cTAKES) (Savova et al., 2010) that made use of third party tools (e.g., Gate[2] and OpenNLP[3]) to annotate sentences in clinical documents.

However, it is more difficult to detect the sentence boundaries in pathology notes, compared to those well-documented news articles or biomedical reports. Although most sentences end with period "." or colon ":", some end without any punctuation; abbreviations or acronyms (e.g., M. – malignant, W.E – wide excision), and entailed punctuation of some entities (e.g., "." for specimen identifiers, "?" representing lexical modality), can complicate the detection.

Methods and Results

Due to the characteristics of the corpora and the advantages of machine learning approaches, a sentence boundary detector was built by using the maximum entropy model based algorithm designed by Reynar and Ratnaparkhi (Reynar and Ratnaparkhi, 1997), implemented in Python and trained with the three corpora. The sentence boundary detector achieved accuracies of 98.77%, 98.86%, and 99.02% in 10-fold cross-validation experiments on the melanoma, colorectal cancer and lymphoma corpora respectively.

5.4.2 Tokenisation

Background

The raw texts in a document need to be split into sentences, and then each sentence needs to be separated into tokens. This basic task of splitting a sentence into a list of words and other symbols is called tokenisation. In the general English domain, most tokens can be separated straightforwardly by white space. However, this naïve approach does not suit clinical narratives in some cases, and the errors caused by it can propagate severely through a downstream processing pipeline. Punctuation in a word can lead to a prominent amount of ambiguity. For example, splitting a hyphen from any word as a separate token may break a medical term, a compound word, or a code in a standard terminology. A period inside a token but not at the end of the token suggests the token can be a measurement, specimen block notation, abbreviation or acronym. This should be identified when performing tokenisation.

[2] http://gate.ac.uk/
[3] http://opennlp.apache.org/

Methods and Results

Words were broken initially at white space and punctuation symbols. After analysing the performance of the white space tokeniser, a set of rules was added, with special attention to separating punctuation symbols from words. They are described as follows (note: examples are separated by semicolon ";"):

- Separate period "." at the end of a word from the word (e.g., Malignant polyp .; I . Anterior (R) forearm).
- Do not separate period "." from a word if it is a measurement, specimen block notation, abbreviation or acronym (e.g., 4.5cm; W.E; 1.1).
- Separate hyphen "-" and "+" at the end of a word from an alphanumeric word (e.g., CD3 +; CD5 -), but not a numeric word (e.g., 20+; 2+).
- Do not separate a hyphen "-" inside a word from the word if it is not at the end of the word (e.g., non-peritonealised; 1-2; B-cell; M-95913; BCL-2).
- Separate a question mark "?" at the beginning of a word from the word (e.g., ? lymphoma; ? MM).
- Do not separate numbers and letters (e.g., 3mm; CD3).
- Do not separate some consecutive punctuation (e.g., 1., 2.; ++).

5.4.3 Proof Reading

After tokenisation, each token is passed through the proof reading process. The main purpose of this process is to verify the token, identify the lexical resource of the token, and standardise the token. Besides three lexical resources: SNOMED CT (SCT), UMLS and Moby, which were introduced in Chapter 2 and 3, there are two other resources used in this process: abbreviation and misspelling lexicons. They were generated from the previous clinical notes (such as notes in the C311 corpus). The abbreviation lexicon contains about 1480 abbreviations or acronyms with their expansions, while the misspelling lexicon consists of over 75000 misspelt words and their correct spelling. Figure 5.4 illustrates the proof reading process.

1. First, the misspelling lexicon is used to verify whether the token is misspelt or not. If it is a misspelling, its correct spelling will be returned from the misspelling lexicon. Meanwhile, every word in the correct form of the token will be passed through the following steps.
2. The token is checked as to whether it belongs to the abbreviation lexicon. If it does, this abbreviation or acronym will be expanded to its full name. If it doesn't, it will be moved to the next step.
3. In the following steps, the token is checked against the other three lexical resources respectively. If it is an entry in these resources, then it will be tagged as "moby", "umls" and "sct" accordingly and exported.
4. If the token does not belong to any of the resources mentioned above, it is passed to manual verification.

5. All the results from the above steps are subsequently manually verified by the medical consultants, and then they are stored in a correction dictionary, expansion dictionary, moby dictionary and medical dictionary respectively.

Note that this process is only performed on single tokens, thus it cannot handle some complicated errors like missing letters and an extra white space (e.g., "fib nopurulent" should be "fibrinopurulent"). The frequency of these errors is not high in a single document, but for the whole corpus, it may be time-consuming to resolve them manually. Although a few of these errors have been found by the author and added to the correction dictionary, it is assumed that there are still some that require an additional process to identify.

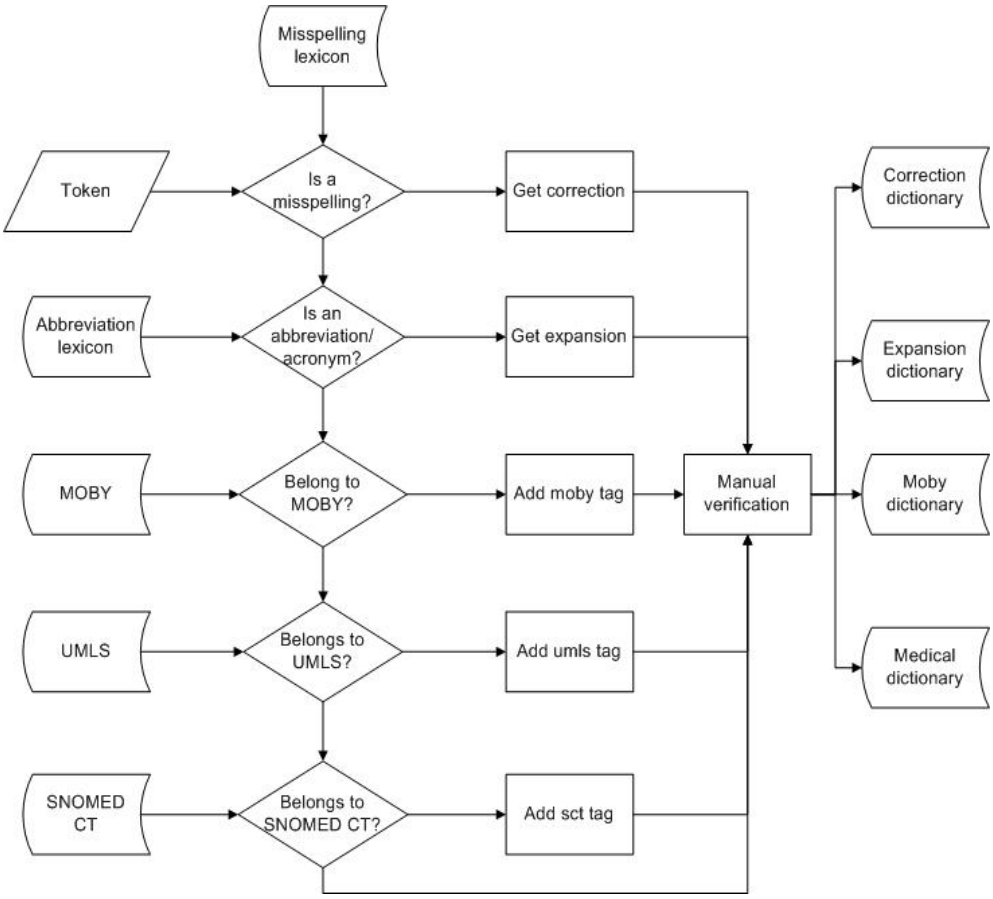


Figure 5.4 Proof reading process.

It is also worth metioning that the expansions for ambiguous abbreviations were based on the frequencies of the full forms occurred in the corpora; that is, the most frequent full form would be selected as the result. For some complex cases, the expansions were determined by medical consultants.

The proof reading process resulted in four types of dictionary, and the number of entries in the dictionaries for each corpus is presented in Table 5.1.

Dictionary	Melanoma corpus	Colorectal cancer corpus	Lymphoma corpus
Correction	34	332	39
Expansion	67	148	243
Moby	2890	4815	4024
Medical	189	476	437

Table 5.1 Entries in the four dictionaries for each corpus.

5.4.4 Part-of-speech Tagging and Shallow Parsing

Background

Part-of-speech (POS) tagging is the process of identifying a word in a text as corresponding to a particular part of speech. Shallow parsing, also called chunking is a process to identify phrases from constituent POS tagged tokens. For example, an adjective sequence followed by a noun can compose a noun phrase. Thus POS tagging is usually requisite for shallow parsing, which was recognised in 1994 in a preliminary investigation on mapping clinical terms to SNOMED III (Sager et al., 1994). The identification of noun phrases was a crucial factor for MER with dictionary lookup methods. It is, for example, one of the essential processes in MetaMap (Aronson, 2001). However, POS taggers for general purposes usually do not perform well in the clinical domain because the lexical characteristics of clinical documents are considerably different from those of articles in the general domain, which are often used as the training data for these taggers. This was addressed by Huang et al, using a statistical parser trained with a corpus in the general domain integrated with the UMLS Specialist Lexicon (Huang et al., 2005). They found that the integration with the UMLS Specialist Lexicon could boost precision and recall of the system by about 5% and 6% respectively.

Methods and Results

Likewise, the annotation guidelines presented in the previous chapter indicate that some entity types cannot cross the boundary of certain kinds of phrase, such as noun phrases and adjectival phrases. The identification of these phrases can contribute to the boundary detection in the MER task. The GENIA tagger is a robust POS tagger based on a cyclic dependency network with maximum entropy with inequality constraints, trained not only on the WSJ corpus, but also on the GENIA corpus and the PennBioIE corpus (Tsuruoka et al., 2005). The tagger has worked well on various types of biomedical documents. It achieved an accuracy of 98.49% on the GENIA corpus, and 91.2% on 332 abstracts of biomedical papers. Thus, it was adopted to perform POS tagging and shallow parsing in this sub-task. The results of POS tagging and shallow parsing on a sentence from the tagger are displayed in Table 5.2.

5.4.5 Lemmatisation

Background

Lemmatisation is a morphological transformation process that changes a given word with different inflected forms into the canonical form or lemma of the word, so that different morphological variants of a word can be analysed as a single item. By reducing the total number of distinct words in the text, it decreases the complexity of processing the text.

Stemming is closely related to lemmatisation, with a similar goal to map different forms of a word to a single form. It normalizes the morphological variants of a word into the same form, a stem, by stripping off the suffix of a word. Since it does not aim to generate a naturally occurring canonical form of a word, it often results in incorrect conflation of semantically distinct terms (Fuller and Zobel, 1998). For example, "excisions" and "excisional" would all be stemmed to "excision" by the Porter stemmer (Porter, 2006), while a lemmatiser would normalize into distinct base forms: "excision" and "excisional". A lemmatiser can yield the canonical form of "larger" as "large", while a stemmer cannot. Although most stemming algorithms are usually easier to implement and run faster, they fail to discriminate between words with different meanings depending on POS, as they don't consider knowledge of the context of the words. Compared to the truncated ambiguous stems, there are more advantages shown by lemmas in document clustering and information extraction (Korenius et al., 2004; Liu et al., 2011).

Methods and Results

Lemmatization may involve other processes such as understanding the context and determining the POS of a word in a sentence, thus a good POS tagger can also bring benefits to lemmatization. To be consistent with the results of POS tagging and shallow parsing, the GENIA tagger was used as a lemmatiser as well. Table 5.2 also presents some outputs of lemmatisation from the tagger.

Token	Lemma	POS tag	Chunk tag
The	The	DT	B-NP
appearances	appearance	NNS	I-NP
are	be	VBP	B-VP
those	those	DT	B-NP
of	of	IN	B-PP
in-situ	in-situ	FW	B-NP
melanoma	melanoma	NN	I-NP
of	of	IN	B-PP
superficial	superficial	JJ	B-NP
spreading	spread	VBG	I-NP
type	type	NN	I-NP
.	.	.	O

Table 5.2 Results of lemmatisation, POS tagging and shallow parsing on a sentence: "The appearances are those of in-situ melanoma of superficial spreading type." from the GENIA tagger.

5.4.6 Section Context Detection

Standard section headings are encouraged by health providers to write well-structured medical records (Fagan et al., 2003; Nieman et al., 2006). Accurate identification of section headings is critical for the detection of section contexts, which is a key step towards further automated or semi-automated clinical language processing. For instance, the diagnoses appearing in the "Clinical History" section are the clinical impressions or differential diagnoses provided by clinicians, while in the "Microscopic" or "Diagnosis Summary" sections, they are final diagnoses made by pathologists.

Several studies have addressed this issue. For instance, a pre-processor integrated into MedLee was able to recognise some common section headings in clinical records (Friedman et al., 1994), and it could achieve 92.0% precision and 91.0% recall in the evaluation conducted by Wang et al (Wang et al., 2010). Many methods purposed by other researchers are also based on rules. Meystre and Haug first implemented a regular expression section detector based on the heading morphology to analyse 200 cardiovascular records, and then augmented it with the section headings extracted from their whole corpus to improve their section detector (Meystre and Haug, 2005). Although they could obtain 100% precision and recall on the test set, the sample size was very small with only 20 records. Some researchers also purposed more sophisticated rule-based approaches to cope with this issue. For example, by applying post-processing, e.g. matching training data to UMLS concepts, correcting misspelling headings and removing stop words, SecTag started with the heading morphology, and finally obtained 95.6% precision and 99.0% recall on test records (Denny et al., 2009). Compared to those rule-based systems relying on more time-consuming and labour-intensive effort, statistical models have an apparent advantage that automatically learns from the predefined feature sets and labels the unseen data. A supervised machine learning approach was adopted by Guiasu and Shenitzer to identify section headings (Guiasu and Shenitzer, 1985).

5.4.6.1 Section Heading Detection

Section terminologies like Logical Observation Identifiers Names and Codes (LOINC) (LOINC Committee, 1994-2014) can facilitate Health Level 7 Clinical Document Architecture to provide a framework to represent and exchange clinical notes (Dolin et al., 2001). An hierarchical section header terminology developed by Denny et al, has also been utilised in SecTag as a reference terminology (Denny et al., 2008). However, these terminologies did not fit for this sub-task. As mentioned in the previous chapter, the section headers in pathology notes are distinct from other clinical documents (e.g. history and physical examination records, progress notes), which these terminologies were designed for. A preliminary study using rules in Meystre and Haug's work (Meystre and Haug, 2005) for this sub-task attained unsatisfactory results (lower than 80% accuracy). A CRF-based approach trained with annotated data, was adopted instead.

Feature Engineering

The preliminary study has shown that the surface heading lexicon is not enough for high quality learning, as there are many lexical variants for some section headings, especially the subheadings. To capture more characteristics of the headings, other features are studied, including the contextual window, lower case of word, standardisation, orthography, and bigram.

Contextual window: According to the analysis of the heading annotations, most of the headings are presented in a single line, which end with a new line separator; some headings are followed by punctuation like a colon ":" and hyphen "-". Hence, the new line separator and the punctuation might be effective learning features in the contextual window, also known as "*bag-of-word*" models. Different sizes of contextual windows have been tried on the corpora, and the optimal size for each corpus is different: seven for the melanoma corpus, nine for the colorectal cancer corpus and five for the lymphoma corpus.

Lower case of word: Lower case of the word is used to normalize the orthographic variants of the word. For example, "CLINICAL NOTES" and "Clinical Notes" can be normalized to "clinical notes".

Standardisation: Standardisation refers to misspelling correction and expansion of abbreviations. Specific lists for correcting the misspelt or abbreviated words inside the heading strings were generated manually through the analysis of the heading annotations. Note that there are no misspelt or abbreviated words found on the heading annotations in the lymphoma corpus.

Orthography: The rendition of words can be captured by the orthographic feature, which is a feature to indicate whether the predicates about the orthography of a word exist. Table 5.3 presents the predicates used in the experiments.

Predicate	Description	Example
IsUppercase	Is the token in uppercase?	CLINICAL; SPECIMEN
IsTitlecase	Does the token have initial capital?	Diagnosis; Immunoperoxidase
IsLowercase	Is the token in lowercase?	results; tumour
HasHyphen	Does the token contain any hyphen "-"?	1-2.; 1-3
IsHyphen	Is the token a hyphen?	-
IsSlash	Is the token a slash?	/
IsColon	Is the token a colon ":"?	:
IsBracket	Is the token a round bracket "(" or ")"?	(;)
IsDigit	Is the token a digit?	1; 4
HasPunctuation	Does the token contain any other punctuation?	T.N.M; ypT.N.M
IsPunctuation	Is the token punctuation except for those above?	& ; .
IsAlphanumeric	Is the token an alphanumeric word?	pTNM

Table 5.3 Predicates used for representing the orthographic features (examples from section heading instances). Note: examples are separated by semicolon ";".

Bigram: A bigram is every sequence of two adjacent elements in a text. In this study, it refers to the combination of two words in original form or standardised form: if the word is an entry in the lists for

standardisation, the standardised form of the word is used; otherwise, the original form of the word is adopted.

Results and Discussion

The 10-fold cross-validation experiments were performed, and the performances were evaluated by the evaluation metrics mentioned above. The scores from those experiments are presented in Tables 5.4, 5.5 and 5.6 with corresponding feature sets. From these tables, it is clear that the best feature sets vary between each corpus:

- For the melanoma corpus – seven-word contextual window, bag of lower case of word, and orthography.
- For the colorectal cancer corpus – nine-word contextual window, bag of lower case of word, standardisation, orthography and bigram.
- For the lymphoma corpus – five-word contextual window, lower case of word, and bag of orthography.

Model #	Feature	Precision	Recall	F-score
M1	seven-word contextual window	98.80%	93.96%	96.32%
M2	M1+ bag of lower case of word	98.81%	95.16%	96.96%
M3	M2 + orthography	99.09%	95.37%	97.19%

Table 5.4 Scores for section heading detection experiments on the melanoma corpus.

Model #	Feature	Precision	Recall	F-score
M1	nine -word contextual window	94.42%	85.27%	89.61%
M2	M1+ bag of lower case of word	94.11%	86.81%	90.31%
M3	M2 + standardisation	93.97%	87.01%	90.36%
M4	M3 + orthography	94.20%	88.10%	91.05%
M5	M4 + bigram	95.16%	88.78%	91.86%

Table 5.5 Scores for section heading detection experiments on the colorectal cancer corpus.

Model #	Feature	Precision	Recall	F-score
M1	five-word contextual window	99.76%	96.09%	97.89%
M2	M1+ lower case of word	99.77%	96.37%	98.04%
M3	M2 + bag of orthography	99.31%	97.17%	98.22%

Table 5.6 Scores for section heading detection experiments on the lymphoma corpus.

The baseline models performed best on the lymphoma corpus (97.89% F-score), and worst on the colorectal cancer corpus (89.61% F-score); the final models using the best feature sets also attained the best performance on the lymphoma corpus (98.22% F-score), and lowest F-score of 91.86% on the colorectal cancer corpus; performances on the melanoma corpus are close to those on the lymphoma corpus, which was 96.32% obtained with baseline model and improved to 97.12% with final model.

For the melanoma corpus, the F-score of the model was improved slightly by 0.64% by introducing *bag of lower case of word*; for the colorectal cancer corpus, the *bigram* feature made the most

contribution to boost the F-score by 0.81%, while a considerable improvement could be contributed by *bag of lower case words* and *orthography*; a relatively small gain (0.33%) was achieved by feature engineering on the lymphoma corpus.

Tables 5.7, 5.8 and 5.9 show the performance of each heading with the best models on each corpus. It can be seen from these tables that the lowest F-scores are on some rare headings, such as "St:Comment Heading" in the melanoma corpus and "St:Ancillary Studies Heading" in the colorectal cancer corpus. Subheadings also have relatively low F-scores due to their abundant lexical variants. The 100% F-scores achieved by most headings on the lymphoma corpus are due to their limited lexical variability (each with only one or two variants).

Heading	Number	Precision	Recall	F-score
St:Clinical History Heading	250	99.60%	98.80%	99.20%
St:Comment Heading	28	76.92%	35.71%	48.78%
St:Diagnosis Heading	258	99.61%	98.06%	98.83%
St:Macroscopic Heading	363	99.16%	97.80%	98.47%
St:Microscopic Heading	375	99.18%	97.33%	98.25%
St:Specimen	164	98.73%	95.12%	96.89%
St:Subheading	51	100.00%	66.67%	80.00%
Overall	1489	99.09%	95.37%	97.19%

Table 5.7 Performance of each heading with the best model on the melanoma corpus.

Heading	Number	Precision	Recall	F-score
St:Ancillary Studies Heading	52	80.95%	65.38%	72.34%
St:Clinical History Heading	378	99.73%	98.68%	99.20%
St:Macroscopic Heading	387	97.64%	96.38%	97.01%
St:Microscopic Heading	388	98.96%	97.94%	98.45%
St:Subheading	2121	92.34%	83.03%	87.44%
St:Synthesis Heading	524	98.81%	94.85%	96.79%
Overall	3850	95.16%	88.78%	91.86%

Table 5.8 Performance of each heading with the best model on the colorectal cancer corpus.

Heading	Number	Precision	Recall	F-score
Title-Clinical History	224	100.00%	100.00%	100.00%
Title-Comment	36	100.00%	100.00%	100.00%
Title-Frozen Section Report	39	100.00%	100.00%	100.00%
Title-Macroscopic Description	223	100.00%	100.00%	100.00%
Title-Microscopic Description	226	100.00%	100.00%	100.00%
Title-Nature And Specimen Type	123	100.00%	100.00%	100.00%
Title-Pathologist Notes	83	100.00%	100.00%	100.00%
Title-Special Investigations	47	100.00%	100.00%	100.00%
Title-Subheading	463	97.18%	89.20%	93.02%
Title-Summary	226	100.00%	100.00%	100.00%
Title-Supplementary Report	38	100.00%	100.00%	100.00%
Title-Supplementary Summary	37	100.00%	100.00%	100.00%
Overall	1765	99.31%	97.17%	98.22%

Table 5.9 Performance of each heading with the best model on the lymphoma corpus.

Through error analysis, the possible reasons for false negatives include:

- Unseen headings in the test set, e.g., "REQUEST FORM ["St:Clinical History Heading"]" and "ADDENDUM ["St:Comment Heading"]" in the melanoma corpus, "SYNTHESIS AND OVERVIEW ["St:Synthesis Heading"]" and "GROSS ["St:Macroscopic Heading"]" in the colorectal cancer corpus, "Architecturally ["St:Title-Subheading"]" and "Cytomorphology ["St:Title-Subheading"]" in the lymphoma corpus.
- Some long span instances exceed the contextual window size, e.g., "Dr X agrees with our diagnosis and his report is as follow ["St:Comment Heading"]" in the melanoma corpus, "The tumour cells show the following immunohistochemical staining pattern ["St:Title-Subheading"]" in the lymphoma corpus.

The following reasons brought not only some false negatives, but also several false positives:

- Polysemous usage of some instances. For example, in the colorectal cancer corpus, "SUPPLEMENTARY REPORT" can be a St:Ancillary Studies Heading if the following contents are regarding ancillary studies; otherwise, it would be a St:Subheading. "Macroscopic Description" at the beginning of a separate section should be regarded as a St:Macroscopic Heading ; whereas, when it is under the "Diagnostic Summary" section, it should be classified as St:Subheading.
- Ambiguity of annotations. In the annotation schema of the colorectal cancer corpus, some subheadings can be annotated as part of the entity if they only contain a single reportable field, e.g. "LYMPH NODES: 2/15 show metastatic adenocarcinoma ["Ex:Lymph Node Involvement"]."; else, they should be annotated as St:Subheading, if they can be further divided into multiple reportable fields, e.g. "LYMPH NODES ["St:Subheading"]: Nineteen (19) lymph nodes identified ["En:Lymph Nodes"], all of which show reactive changes ["Ex:Lymph Node Involvement"]."

Note that the results of the identification of subheadings is not be further used in the section context detection.

Compared to other existing heading detectors, e.g. MedLee (92.0% precision and 91.0% recall) and SecTag (95.6% precision and 99.0% recall), a similar or higher precision can be achieved by the present section heading detectors, except the recall is lower in the colorectal cancer corpus. The possible reasons for the lower recall is that subheadings including specimen identifiers increases the difficulty for detection due to their indistinguishable morphology and orthography; the larger number of variants for subheadings (about 350) make it difficult for the machine learner to learn and predict. The ratio between training set and test set in SecTag is 33.8:1, but is 9:1 in this study, hence the present detector models are encouragingly more successful.

5.4.6.2 Section Context Assignment

Based on the results from the above section heading detectors, the second stage of section context detection is to find the text span between each heading and assign a section context for the text span. For example, the text span between “St:Macroscopic Heading” and “St:Microscopic Heading” should be assigned a section context of “MACROSCOPIC”.

If a particular section heading is missing in the record, assign a section context for a text span according to the frequent sequences of heading appearances. For instance, in the melanoma corpus, headings usually appear in this order: St:Clinical History Heading, St:Macroscopic Heading, “St:Microscopic Heading”, and “St:Diagnosis Heading”. When a St:Clinical History Heading is omitted in a report, but St:Macroscopic Heading appears in the first place, it implies the text span preceding the St:Macroscopic Heading has a section context of “CLINICAL HISTORY”.

If multiple section headings are missing in the record, or in a poorer written report, no heading appears in it, specific rules are applied to resolve the problem:

1. Count the amount of newline separators and special combinations punctuation as section separators (e.g., “.,” and “,,” in the colorectal cancer corpus) and divide the texts in the document into several potential sections.
2. Try to assign section contexts according to the frequent sequences of heading appearance as mentioned above.
3. If the last step fails, assign a section context for the potential section arbitrarily based on the significance and frequency analysis of other well-organised reports. For example, in the colorectal cancer corpus, if a poorly written report has three potential sections, they can be assigned section contexts in this order: “MICROSCOPIC”, “MACROSCOPIC” and “DIAGNOSIS”; if there are two sections in it, “MICROSCOPIC” and “MACROSCOPIC” will be assigned to them respectively.

Finally, three section contexts are shared by the three corpora: “CLINICAL HISTORY”, “MACROSCOPIC” and “MICROSCOPIC”; there are two common section contexts for the melanoma corpus and lymphoma corpus: “SPECIMEN” and “COMMENT”; section contexts for the “Diagnostic Summary” section have different notations, which are “DIAGNOSIS” for the melanoma corpus, “CONCLUSION” and “SYNOPTIC” for the colorectal cancer corpus, “SUMMARY” for the lymphoma corpus; the colorectal cancer corpus has one more section context “ANCILLARY”; the lymphoma corpus has five more section contexts: “FROZEN SECTION”, “PATHOLOGIST NOTES”, “SPECIAL INVESTIGATIONS”, “SUPPLEMENTARY REPORT” and “SUPPLEMENTARY SUMMARY”. The main difference between “CONCLUSION” and “SYNOPTIC” is whether the following contents contain synoptic fields: if synoptic fields are present, the section context is assigned as “SYNOPTIC”; otherwise, it is assigned as “CONCLUSION”.

5.4.7 Ring-fenced Tagging

Background

Some chunks of a text may be of medical significance, like scores and measurements, or contain useful linguistic patterns, which have been dissembled in the above tokenisation. This requires an additional process of detection.

A simple way to resolve this issue is to employ regular expressions. However, there are several disadvantages of this method: while more rules are developed to capture new patterns, it is more difficult to handle the rules as any change of them may create the risk of losing previously recognised patterns or introducing some false positives; besides, it also requires exhaustive knowledge about regular expressions and a considerable amount of time to modify the rules.

A more efficient way is to automate the learning process to capture patterns. Patrick and Sabbagh developed a pattern-matching engine consisting of a trainable finite state automaton (FSA) as a solution (Patrick and Sabbagh, 2011). The trainable FSA can dynamically learn from training examples with high accuracy and efficient computational time. Additionally, the cascaded approach and removal of irrelevant words enhances the power of generalization of the engine and the active learning process increases the speed of the engine.

Methods and Results

Two lists of training patterns are prepared for the engine. One is the basic pattern file and the other is the complex pattern file. Both have the same format, with two columns, where the first one is the tag type and the second one is the text example. The training examples are then generalized by the FSA so that the engine can capture other similar forms of these patterns. Table 5.10 lists some training examples in a basic pattern file and complex pattern file.

Basic pattern file		**Complex pattern file**	
Tag_type	Pattern	Tag_type	Pattern
Digit	2; 1.16; 10; 110	Volume	20x14x9mm; 1.5x0.5x1.0 cm
mm	mm	Area	140x30mm; 40 mm x 35 mm
cm	cm	Measurement	50mm; 5.5cm
x	x		

Table 5.10 Some training examples in a basic pattern file and complex pattern file. Note: examples are separated by semicolon ";".

The results from the engine for a particular chunk of a text can be either the tag type defined in the training pattern list or the default output. The default output of the engine can have 22 semantic categories, e.g. "Date", "Time", "Range", "two word slash", "two word hyphen", etc. The default output for some chunks of text is presented in Table 5.11.

Semantic category	Chunk of text
Date	28/05/02
Range	1-2
two word slash	SPECIMEN/CLINICAL
two word hyphen	in-situ
two word apostrophe	Hutchinson's
Complexdigit	10/19
Operator	<
Punctuation	,
Plainword	component

Table 5.11 Default output for some chunks of text from the ring-fenced engine.

5.5 Methods

5.5.1 Overview of the System

A supervised machine learning approach is used for the Medical Entity Recognition (MER) System to identify medical entities in the free texts. To facilitate the learning and prediction of the machine learner, data should be converted into features so that the learner can distinguish them from each other. Features should represent the characteristics and empirical distribution of the training data. Furthermore, they should encode the most significant aspects of the data for the testing. In a CRF-based system, the generation of features is a crucial factor for the success of the system.

There are two stages involved in the feature generation process. The first stage is feature extraction, which is to extract potentially useful features from the corpus. However, not all the features extracted can be applied to the models, as some of them may be less informative or redundant, and will introduce noise and slow down the training process. Thus, the second stage, feature selection, is required to carefully remove the irrelevant features.

A detailed description of the features used in the MER task is presented as follows. The aims are to extract various features that can capture useful information about the entities, and then determine the optimal configuration of feature sets to yield the best performance. To discover the best feature sets, a selective incremental method was used: each feature was added progressively to identify its contribution to the model; if the performance of the model benefited from a feature, then this feature would be retained, else, it would be dropped.

5.5.2 Feature Sets

Features are descriptors of characteristic attributes of tokens prepared for the task. The features are usually represented in a vector string, which can be a Boolean, numeric or nominal value. For example, a Boolean feature assigned with "T" if the current token is in lowercase, else "F"; a nominal feature that represents the lemma of the token.

Various features have been experimented with in previous works, such as lexical features, contextual features and semantic features. The feature sets used in the MER experiments consist of lexical features, semantic features, contextual features, orthographic features, morphological features and syntactic features. They are not only focused on the identification of entity boundaries but also the classification of entity types.

Contextual Features

Contextual window: Tokens surrounding the target token provide useful contexts for predicting the entity type, as they are incorporated into the classifier to reveal the linguistic patterns of the entity type. Typically, a contextual window is used to represent this context feature, which is a sliding window around the target token. The Larger the window size, the more context information it can provide to the classifier. A nine-word contextual window was used in the following experiments, that is, four tokens preceding the target token and four tokens succeeding the target token.

Section context: The regional context information can be captured by a contextual window, but it cannot represent the global context information. Section context is a feature to represent the global context information, as the distribution of the entity types vary between each section. Tables 5.12, 5.13 and 5.14 show the details of each medical entity type present in the section with the highest frequency of the corpora. Figures 5.5, 5.6 and 5.7 present the proportion of medical entities held by each main section. Note that Linguistic categories and associated types of specimen identifiers are excluded, as it is assumed that they are insensitive to section context.

Entity type	Section context	Number	Proportion
De:Cell Growth Pattern	MICROSCOPIC	584	93.6%
De:Cell Type	MICROSCOPIC	659	94.8%
De:Cosmetic Changes	CLINICAL HISTORY	35	13.2%
De:Dermal Mitoses	MICROSCOPIC	355	97.3%
De:Shape	MACROSCOPIC	403	72.2%
De:Site and Laterality	CLINICAL HISTORY	272	29.6%
De:Size	MACROSCOPIC	801	92.5%
De:Specimen Type	MACROSCOPIC	437	66.1%
De:Ulceration	MICROSCOPIC	244	87.1%
En:Associated Naevus (type)	MICROSCOPIC	153	68.3%
En:Lesion (other)	MICROSCOPIC	36	64.3%
En:Primary Lesion	MICROSCOPIC	952	58.0%
En:Satellites	MICROSCOPIC	24	100.0%
In:Breslow Thickness (mm)	MICROSCOPIC	340	65.9%
In:Clark Level	MICROSCOPIC	553	74.2%
In:Neurotropism	MICROSCOPIC	144	96.6%
In:Vascular/Lymphatic	MICROSCOPIC	219	95.6%
Ma:Excision Clear	MICROSCOPIC	165	67.9%
Ma:Excision Deep	MICROSCOPIC	152	91.0%
Ma:Excision Invasive	MICROSCOPIC	285	78.3%
Ma:Excision In Situ	MICROSCOPIC	74	83.1%
Re:Desmoplasia	MICROSCOPIC	16	100.0%
Re:Fibrosis	MICROSCOPIC	65	95.6%
Re:Solar Elastosis	MICROSCOPIC	23	92.0%

Re:Tils	MICROSCOPIC	209	98.6%
Sy:Diagnosis	MICROSCOPIC	668	51.5%
Sy:Regression	MICROSCOPIC	173	86.1%
Sy:Subtype	MICROSCOPIC	301	61.7%

Table 5.12 Numbers of medical entities present in the sections with the highest frequency in the melanoma corpus.

Entity type	Section context	Number	Proportion
De:Ancillary Studies	ANCILLARY	178	63.8%
De:Mesorectal Integrity	MACROSCOPIC	12	80.0%
De:Perforation	MICROSCOPIC, MACROSCOPIC	41	26.6%
De:Peritoneal Reflection	MACROSCOPIC	125	90.6%
De:Serosa Description	MACROSCOPIC	165	77.8%
De:Specimen Blocks	MACROSCOPIC	3335	98.9%
De:Specimen Images	MACROSCOPIC	16	53.3%
De:Specimen Size	MACROSCOPIC	1551	97.7%
De:Specimen Type	MACROSCOPIC	794	41.9%
De:Tissue Banking	MACROSCOPIC	58	100.0%
De:Tumour Description	MICROSCOPIC	780	53.2%
De:Tumour Site	CONCLUSION	363	25.1%
De:Tumour Size	MACROSCOPIC	443	65.0%
En:Coexistent Pathology	MICROSCOPIC	470	39.7%
En:Distant Spread or Metastases	MICROSCOPIC	98	34.3%
En:Lymph Nodes	MACROSCOPIC	307	48.7%
En:Residual Tumour	CONCLUSION	61	49.2%
Ex:Donut Involvement	MICROSCOPIC	64	43.8%
Ex:Extent	MACROSCOPIC	210	34.4%
Ex:Extramuscular Spread	SYNOPTIC	112	56.9%
Ex:Lymph Node Involvement	MICROSCOPIC	479	42.2%
Ex:Serosal Involvement	SYNOPTIC	137	34.9%
In:Depth of Invasion	MICROSCOPIC	443	34.5%
In:Perineural Invasion	MICROSCOPIC	181	45.5%
In:Venous and Small Vessel Invasion	MICROSCOPIC	406	47.4%
Ma:Circumferential Margin	MACROSCOPIC	90	35.3%
Ma:Clear	MICROSCOPIC	335	43.9%
Ma:Proximal or Distal Margin	MACROSCOPIC	509	72.8%
Met:Anatomic Stage	CONCLUSION	164	56.4%
Met:M Value	CONCLUSION	176	68.5%
Met:N Value	CONCLUSION	247	60.4%
Met:T Value	CONCLUSION	164	56.4%
Re:Desmoplasia and Fibrosis	MICROSCOPIC	221	81.5%
Re:Response to Rx	MICROSCOPIC	69	64.5%
Re:Tils and Peritumoural Lymphocytes	SYNOPTIC	215	55.3%
Sy:Comment	MACROSCOPIC	821	52.7%
Sy:Histological Grade	CONCLUSION	366	44.1%
Sy:Histological Type	CONCLUSION	414	41.5%
Sy:Medical History	CLINICAL HISTORY	110	53.1%

Table 5.13 Numbers of medical entities present in the sections with the highest frequency in the colorectal cancer corpus.

Entity type	Section context	Number	Proportion
An:Biomarker	MICROSCOPIC	1748	90.50%
An:Cytogenetics Comment	SUPPLEMENTARY REPORT	16	88.90%
An:Fish Results	SUPPLEMENTARY REPORT	7	87.50%
An:Flow Cytometry-Comment	SPECIAL	71	77.20%
An:Flow Cytometry-Negative	SUPPLEMENTARY REPORT	3	100.00%
An:Flow Cytometry-Positive	SUPPLEMENTARY REPORT	3	100.00%
An:IgH Test	SUPPLEMENTARY REPORT	7	100.00%
An:Immunohistochemistry-Comment	MICROSCOPIC	226	90.40%
An:Immunohistochemistry-Equivocal	MICROSCOPIC	24	88.90%
An:Immunohistochemistry-Negative	MICROSCOPIC	248	87.30%
An:Immunohistochemistry-Positive	MICROSCOPIC	537	90.40%
An:PCR Comment	SUPPLEMENTARY REPORT	27	96.40%
An:TCRgamma Test	SUPPLEMENTARY REPORT	2	100.00%
De:Anatomical Structure	MICROSCOPIC	93	23.50%
De:Architecture	MICROSCOPIC	446	94.50%
De:Cell Clonality	SUPPLEMENTARY REPORT	1	100.00%
De:Cell Size	MICROSCOPIC	466	95.50%
De:Cytomorphology	MICROSCOPIC	168	94.40%
De:Laterality	CLINICAL HISTORY	6	33.30%
De:Lineage	MICROSCOPIC	120	85.70%
De:Other Size	MACROSCOPIC	103	83.70%
De:Preservative Fluid	MACROSCOPIC	77	70.00%
De:Sample Triage	MACROSCOPIC	538	67.80%
De:Specimen Blocks	MACROSCOPIC	847	100.00%
De:Specimen Size	MACROSCOPIC	462	98.90%
De:Specimen Type	MACROSCOPIC	200	24.90%
De:Tissue Source	MICROSCOPIC	771	52.00%
De:Topography	MACROSCOPIC	374	26.60%
De:Tumour Size	MACROSCOPIC	35	68.60%
En:Coexistent Pathology	MICROSCOPIC	111	68.90%
Ex:Disease Extent	CLINICAL HISTORY	7	87.50%
Ex:Other Sites of Disease	MICROSCOPIC	34	64.20%
Re:Tissue Reaction	MICROSCOPIC	245	94.60%
Sy:Clinical Impression	CLINICAL HISTORY	159	85.90%
Sy:Comment	SUMMARY	43	55.80%
Sy:Constitutional Symptoms	CLINICAL HISTORY	28	100.00%
Sy:Diagnosis	MICROSCOPIC	530	50.20%
Sy:Diagnosis Subtype	MICROSCOPIC	18	60.00%
Sy:Indication for Biopsy	CLINICAL HISTORY	26	92.90%
Sy:Medical History	CLINICAL HISTORY	78	95.10%
Sy:Predisposing Factors	CLINICAL HISTORY	35	58.30%
Sy:Presentation	CLINICAL HISTORY	86	87.80%
Sy:SNOMED RT Codes	SUMMARY	859	86.40%
Sy:Stage	SUMMARY	2	66.70%
Sy:WHO Grade	MICROSCOPIC	78	58.20%

Table 5.14 Numbers of medical entities present in the sections with the highest frequency in the lymphoma corpus.

From Figures 5.5, 5.6 and 5.7, it can be seen that “MACROSCOPIC” and “MICROSCOPIC” hold most of the entities, while a considerable number of entities located in the section contexts stand for the “Diagnostic Summary” section (e.g., “DIAGNOSIS” in the melanoma corpus) and “CLINICAL HISTORY”, and other sections account for a small proportion of the entities (e.g., “ANCILLARY” in the colorectal cancer corpus).

Figure 5.5 Proportions of medical entities contained in each main section of the melanoma corpus.

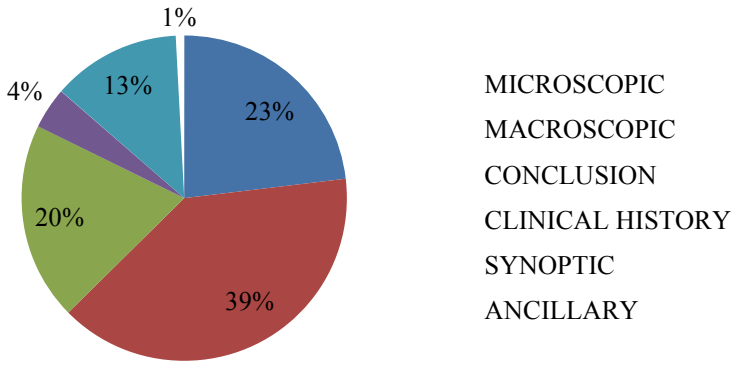


Figure 5.6 Proportions of medical entities contained in each main section of the colorectal cancer corpus.

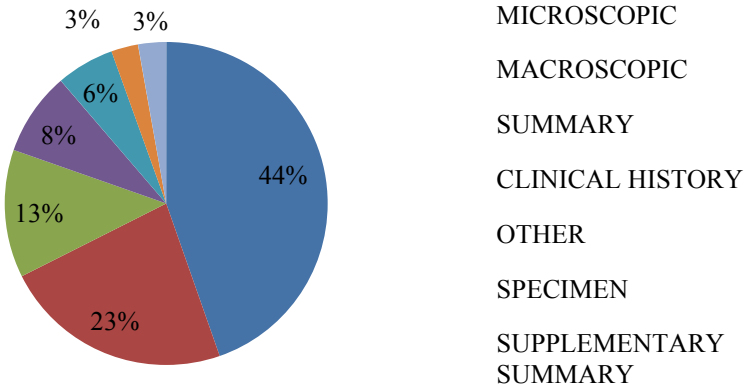


Figure 5.7 Proportions of medical entities contained in each main section of the lymphoma corpus.

Lexical Features

Lowercase of token: Every token in the training data can be used as a feature, since the frequency of the token is of significance to determine the entity type. To increase the recall, each token is converted to lowercase.

Lemma: By applying lemmatisation to a token, different morphological variants of a token can be normalized to its canonical form.

Correction of misspelling: Spelling errors detected in the proof reading process should be replaced by their correct forms stored in the correction dictionaries.

Expansion of abbreviations and acronyms: Similarly, abbreviations and acronyms identified in the proof reading process should be expanded to their full forms based on the abbreviation dictionaries.

Bigram: This feature refers to that in feature engineering for detecting section headings. It is assumed that some combination of two words compose phrases that are likely to be medical terms. Thus, it can be an informative supplement for the lexical information about certain entity types. Table 5.15 lists the ten most common bigrams and their frequencies in each corpus. The Bigram "of/the" frequently appears in the corpora, with second highest frequency in each corpus. Several of them are probably medical glossaries: "malignant/melanoma", "clark/level", "lymph/nodes", "resection/margin", "muscularis/propria", "lymph/node", "malignant/lymphoma", "t/cells", "snomed/codes", "flow/cytometry" and "microscopic/report".

Melanoma corpus		Colorectal cancer corpus		Lymphoma corpus	
Bigram	Frequency	Bigram	Frequency	Bigram	Frequency
malignant/melanoma	601	lymph/nodes	1736	lymph/node	844
of/the	591	of/the	1696	of/the	445
there/is	543	the/tumour	1193	malignant/lymphoma	279
the/lesion	391	there/is	1125	cell/lymphoma	266
of/skin	371	resection/margin	987	t/cells	258
clark/level	349	from/the	945	lymph/nodes	252
consists/of	256	of/tumour	701	snomed/codes	246
ellipse/of	253	up/to	649	flow/cytometry	241
from/the	249	muscularis/propria	601	1/m	235
the/specimen is/a	247	is/a	584	microscopic/report	226

Table 5.15 Ten most common bigrams and their frequencies in each corpus.

Morphological Features

Morphological information has been proven to be a good clue for recognising named entities in the biomedical domain. For example, Wang et al carried out some experiments, which showed that the performance of the NER system was greatly enhanced with prefix and suffix information, as this information can help a machine learner to predict whether an unseen token is an entity or not (Wang et al., 2008).

In the clinical domain, there are a large number of entities derived from Latin or Greek roots, and their affixes suggest special meanings. For instance, -omy suggests a surgical procedure, -oma indicates an abnormal structure, ade- implies or relates to a gland, cyt- associates with cell. These affixes do provide helpful hints for determining the entity types. The prefix and suffix features are focused on the characters of each word that begins and ends with respectively. A different number of characters from either the start or end of each alphabetic word were extracted as features. Affixes of length from two to four characters were used in the preliminary experiments. It turned out that the optimal sizes of affixes were 3 in most cases, except that of suffixes in the melanoma corpus was 2 instead. The ten most frequent suffixes and prefixes for the alphabetic tokens in each corpus are presented in Tables 5.16 and 5.17 respectively.

Melanoma corpus		Colorectal cancer corpus		Lymphoma corpus	
Suffix	Frequency	Suffix	Frequency	Suffix	Frequency
he	3025	ion	6808	ure	1802
on	2361	our	4964	ion	1698
al	2161	ing	4119	lls	1448
nt	1571	ted	2862	oma	1410
ed	1562	ent	2630	ing	1407
ng	1478	tal	2509	ent	1290
nd	1434	des	2405	ode	1238
re	1392	mph	2262	ive	1131
in	1231	gin	2023	mph	1088
es	1115	oma	1991	ith	1081

Table 5.16 Ten most frequent suffixes for the alphabetic tokens in each corpus.

Melanoma corpus		Colorectal cancer corpus		Lymphoma corpus	
Prefix	Frequency	Prefix	Frequency	Prefix	Frequency
the	3683	tum	4941	lym	3168
and	1234	lym	3286	cel	2141
mel	1210	nod	3221	nod	2017
les	928	mar	3165	pro	1886
wit	763	inv	2608	wit	1171
are	660	wit	2461	lar	816
mal	614	res	2320	sho	681
ski	612	pro	2133	sma	675
der	585	col	1978	spe	661
inv	559	per	1880	sec	641

Table 5.17 Ten most frequent prefixes for the alphabetic tokens in each corpus.

From Tables 5.16 and 5.17, the common frequent suffixes for the three corpora are “ion” or “on”, “ing” or “ng”, “ted” or “ed”, “ent” or “nt”, while “wit” is the most common frequent prefix. A detailed analysis reveals that this is likely because of the extensive use of past tense or gerund of verbs (e.g. “noted”, “ulcerated”, “measuring”, “infiltrating”, “extending”), nouns end with “ion” or “ent” (e.g. “component”, “involvement”, “resection”, “description”, “invasion”, “lesion”, “section”, “portion”), adjectives end with “ent” (e.g. “present”, “consistent”), prepositions “with” and “within” in the notes.

Semantic Features

Three kinds of semantic features are prepared: lexical resource, medical category, and ring-fenced tag.

Lexical resource: From the above proof reading process, the resource dictionaries are the references for determining the source of a token. The possible values for this feature are “moby”, “umls”, “sct” and “O” (the default value if the token is not an entry in the resource dictionaries).

Medical category: The medical category is one of the top categories of SNOMED CT, obtained by parsing the text to identify concepts of SNOMED CT using the TTSCT service (Patrick et al., 2007b), which was developed to detect SNOMED CT concepts in free texts and to annotate them with clinical reference terms. Note that in some cases, TTSCT can return more than one category as results; only the first order category is selected as the feature to reduce the complexity of the representation of the feature.

Ring-fenced tag: This is an internal semantic feature, attained from the pattern-matching engine mentioned above by providing training examples to the trainable FSA.

Syntactic Features

This feature set includes POS tag and chunk.

POS tag: This feature is able to generalise some tokens in an entity with relatively low frequency by representing them with a set of POS tags. Although it is low level syntactic information, it can help the machine learner to acquire the grammatical constructs of the entities and consequently, affect the determination of the boundaries of the entities, which has been proven in some biomedical NER systems, e.g. in Zhou and Su’s system (Zhou and Su, 2004).

Chunk: This feature results from shallow parsing by the GENIA tagger described above. It is also used as a clue to determine the boundary of an entity.

Orthographic Features

This feature set consists of the orthography, full word class and brief word class.

Orthography: This feature aims to capture the rendition of words, with the same description illuminated above. Most of the predicates are retained, but some of them (e.g., "*IsColon*" and "*IsBracket*") have been discarded after testing. More examples from some medical entities are displayed in Table 5.18.

Predicate	Example
IsUppercase	BOWEL; RM
IsTitlecase	Smears; Sigmoid
IsLowercase	flow; resection
HasHyphen	Non-Hodgkin; MLH-1
IsHyphen	-
IsSlash	/
IsDigit	20; 21
HasPunctuation	0.5mm; 0.8
IsPunctuation	(; :
IsAlphanumeric	CD10; 20mm

Table 5.18 Orthography feature with examples from some medical entities. Note: examples are separated by semicolon ";".

Full word class: To generalize the expression of the words, a feature named "full word class", similar to that in Collins's work (Collins, 2002), is used. It represents a token by replacing capital letters with "A", lowercase letters with "a", digits with "0", and all other characters with "_".

Brief word class: Like the full word class, the brief word class is also a generalized representation of the words. It collapses consecutive identical characters into one.

Table 5.19 presents the full word class and brief word class features for some tokens.

Token	Full word class	Brief word class
right	aaaaa	a
COLON	AAAAA	A
Large	Aaaaa	Aa
70mm	00aa	0a
4bp	0aa	0a
,	_	_
CD20	AA00	A0
1A	0A	0A
30	00	0
M-95903	A_00000	A_0
B-cell	A _aaaa	A_a
immunoblast-like	aaaaaaaaaaa_aaaa	a_a
CD79a	AA00a	A0a

Table 5.19 Full word class and brief word class features for some tokens.

Table 5.20 summarizes the feature sets used in the experiments, with all features generated for predicting the label of token *t* at position *i* in an input sequence.

Feature	Representation
Contextual window	t_{i-4}, t_{i-3}, t_{i-2}, t_{i-1}, t_i, t_{i+1}, t_{i+2}, t_{i+3}, t_{i+4}
Section context	*Context*(t_{i-4}), *Context*(t_{i-3}), *Context*(t_{i-2}), *Context*(t_{i-1}), *Context*(t_i), *Context*(t_{i+1}), *Context*(t_{i+2}), *Context*(t_{i+3}), *Context*(t_{i+4})
Lowercase of token	*Lower*(t_{i-4}), *Lower*(t_{i-3}), *Lower*(t_{i-2}), *Lower*(t_{i-1}), *Lower*(t_i), *Lower*(t_{i+1}), *Lower*(t_{i+2}), *Lower*(t_{i+3}), *Lower*(t_{i+4})
Lemma	*Lemma*(t_{i-4}), *Lemma*(t_{i-3}), *Lemma*(t_{i-2}), *Lemma*(t_{i-1}), *Lemma*(t_i), *Lemma*(t_{i+1}), *Lemma*(t_{i+2}), *Lemma*(t_{i+3}), *Lemma*(t_{i+4})
Correction of misspelling	*Correction*(t_{i-4}), *Correction*(t_{i-3}), *Correction*(t_{i-2}), *Correction*(t_{i-1}), *Correction*(t_i), *Correction*(t_{i+1}), *Correction*(t_{i+2}), *Correction*(t_{i+3}), *Correction*(t_{i+4}),
Expansion of abbreviations and acronyms	*Expansion*(t_{i-4}), *Expansion*(t_{i-3}), *Expansion*(t_{i-2}), *Expansion*(t_{i-1}), *Expansion*(t_i), *Expansion*(t_{i+1}), *Expansion*(t_{i+2}), *Expansion*(t_{i+3}), *Expansion*(t_{i+4})
Bigram	t_{i-4}/t_{i-3}, t_{i-3}/t_{i-2}, t_{i-2}/t_{i-1}, t_{i-1}/t_i, t_i/t_{i+1}, t_{i+1}/t_{i+2}, t_{i+2}/t_{i+3}, t_{i+3}/t_{i+4}
Prefix	*Prefix*(t_{i-4}), *Prefix*(t_{i-3}), *Prefix*(t_{i-2}), *Prefix*(t_{i-1}), *Prefix*(t_i), *Prefix*(t_{i+1}), *Prefix*(t_{i+2}), *Prefix*(t_{i+3}), *Prefix*(t_{i+4})
Suffix	*Suffix*(t_{i-4}), *Suffix*(t_{i-3}), *Suffix*(t_{i-2}), *Suffix*(t_{i-1}), *Suffix*(t_i), *Suffix*(t_{i+1}), *Suffix*(t_{i+2}), *Suffix*(t_{i+4})
Lexical resource	*Resource*(t_{i-4}), *Resource*(t_{i-3}), *Resource*(t_{i-2}), *Resource*(t_{i-1}), *Resource*(t_i), *Resource*(t_{i+1}), *Resource*(t_{i+2}), *Resource*(t_{i+3}), *Resource*(t_{i+4})
Medical category	*Category*(t_{i-4}), *Category*(t_{i-3}), *Category*(t_{i-2}), *Category*(t_{i-1}), *Category*(t_i), *Category*(t_{i+1}), *Category*(t_{i+2}), *Category*(t_{i+3}), *Category*(t_{i+4})
Ring-fenced tag	*Tag*(t_{i-4}), *Tag*(t_{i-3}), *Tag*(t_{i-2}), *Tag*(t_{i-1}), *Tag*(t_i), *Tag*(t_{i+1}), *Tag*(t_{i+2}), *Tag*(t_{i+3}), *Tag*(t_{i+4})
POS tag	*POS*(t_{i-4}), *POS*(t_{i-3}), *POS*(t_{i-2}), *POS*(t_{i-1}), *POS*(t_i), *POS*(t_{i+1}), *POS*(t_{i+2}), *POS*(t_{i+3}), *POS*(t_{i+4})
Chunk	*Chunk*(t_{i-4}), *Chunk*(t_{i-3}), *Chunk*(t_{i-2}), *Chunk*(t_{i-1}), *Chunk*(t_i), *Chunk*(t_{i+1}), *Chunk*(t_{i+2}), *Chunk*(t_{i+3}), *Chunk*(t_{i+4})
Orthography	*Orthography*(t_{i-4}), *Orthography*(t_{i-3}), *Orthography*(t_{i-2}), *Orthography*(t_{i-1}), *Orthography*(t_i), *Orthography*(t_{i+1}), *Orthography*(t_{i+2}), *Orthography*(t_{i+3}), *Orthography*(t_{i+4})
Full word class	*Full*(t_{i-4}), *Full*(t_{i-3}), *Full*(t_{i-2}), *Full*(t_{i-1}), *Full*(t_i), *Full*(t_{i+1}), *Full*(t_{i+2}), *Full*(t_{i+3}), *Full*(t_{i+4})
Brief word class	*Brief*(t_{i-4}), *Brief*(t_{i-3}), *Brief*(t_{i-2}), *Brief*(t_{i-1}), *Brief*(t_i), *Brief*(t_{i+1}), *Brief*(t_{i+2}), *Brief*(t_{i+3}), *Brief*(t_{i+4})

Table 5.20 Features generated for token *t* at position *i* used in the experiments.

It is noteworthy that the strategies for the identification of specimen identifiers vary between each corpus due to different definitions in the annotation schemas. Specimen identifiers are identified as subheadings in the colorectal cancer corpus, which have been discussed in the section on heading detection. They are recognised with other entities in the same model in the melanoma corpus, which are presented in the next section. They are detected in a separate model in the lymphoma corpus. Several experiments were performed for specimen identifier detection in the lymphoma corpus. The best model attained a high F-score of up to 99.04%, by using a combination of the features: five-word contextual window, bag of lowercase of token, bag of orthography, ring-fenced tag and bigram.

5.5.3 Experiment Setting

The toolkit used for applying CRF in this task is CRF++[4], currently one of the fastest and stable CRF toolkits, which is based on the algorithms proposed by Sha and Pereira, and Lafferty et al (Lafferty et al., 2001; Sha and Pereira, 2003). It provides a simple way to manage feature extraction. The input data file should be in a spread sheet-like format that each column is a potential feature such as the token itself and the POS of the token, except that the last column is the annotation category. It is necessary to specify a feature template to train a model, which indicates the combination of features customized by a user to train the model. Thus the user can combine features easily by modifying the template rather than changing the training data. The test data file has the same format as in the training data, and the results generated by the model are presented in a results file, with an additional column next to the last column in the test data file. This eases the evaluation, as there are existing evaluation scripts (e.g., the evaluation script for CoNLL 2000 shared task[5]) that can compute scores from the results file.

The experiments were carried out with 10-fold cross-validation, and each fold was stratified on a document level, and used the default parameter configuration of the toolkit. The standard evaluation metrics: Precision, Recall and F-score were used to measure the performance. The evaluation scripts were adapted from those provided by the JNLPBA 2004 shared task[6].

5.6 Results and Discussion

Baseline models were built using only the *bag-of-word* feature from the training corpora. A *contextual window size of nine* was used in all experiments. Further experimental analysis of the contribution of each feature was conducted by progressively adding features to the system. Note that only the combinations of features that improve the system performance are presented below.

5.6.1 System Performance on Melanoma Corpus

Table 5.21 shows the contribution of features to the system performance on the melanoma corpus. The baseline model achieved 78.95% F-score. The lexical feature set was the most effective, and improved the model by 3.72%, whereas the *lowercase of tokens* contributed 2.59%. Semantic and morphological feature sets yielded moderate improvements by 0.82% and 0.57% respectively. Minimal improvements were made by adding the syntactic feature set and the *section context* with 0.16% and 0.07%.

[4] http://crfpp.googlecode.com/svn/trunk/doc/index.html
[5] http://www.cnts.ua.ac.be/conll2000/chunking/
[6] http://www.nactem.ac.uk/tsujii/GENIA/ERtask/report.html

Model #	Features	Precision	Recall	F-score
1	Nine-word contextual window	85.79%	73.12%	78.95%
2	M1 + Lowercase of tokens	86.49%	77.12%	81.54%*
3	M2 + Lemma	86.46%	78.21%	82.13%
4	M3 + POS	86.12%	78.64%	82.21%
5	M4 + Chunk	86.07%	78.82%	82.29%
6	M5 + Medical category	86.35%	79.62%	82.85%
7	M6 + Expansions of abbreviations and acronyms	86.38%	79.64%	82.87%
8	M7 + Correction of misspelling	86.51%	80.16%	83.22%
9	M8 + Ring-fencing tag	86.41%	80.75%	83.48%
10	M9 + Suffixes	86.41%	81.15%	83.70%
11	M10 + Bag of prefixes	86.55%	81.69%	84.05%
12	M11 + Section context	86.63%	81.75%	84.12%
13	M12 + Bigram	87.11%	81.65%	84.29%

Table 5.21 Contribution of features to the system performance on the melanoma corpus. Scores marked with * suggests significant contribution within 95% confidence interval.

Entity type	Number	Precision	Recall	F-score
De:Cell Growth Pattern	615	71.06%	62.28%	66.38%
De:Cell Type	694	73.08%	71.18%	72.12%
De:Cosmetic Changes	266	67.76%	38.72%	49.28%
De:Dermal Mitoses	364	82.82%	80.77%	81.78%
De:Shape	555	76.69%	67.57%	71.84%
De:Site and Laterality	817	89.51%	82.50%	85.86%
De:Size	845	91.87%	88.28%	90.04%
De:Specimen Type	627	92.25%	83.57%	87.70%
De:Ulceration	280	93.12%	91.79%	92.45%
En:Associated Naevus (type)	222	70.59%	64.86%	67.61%
En:Lesion (other)	57	71.43%	8.77%	15.62%
En:Primary Lesion	1612	88.20%	90.45%	89.31%
En:Satellites	24	89.47%	70.83%	79.07%
En:Specimen Identifier	842	97.13%	88.36%	92.54%
In:Breslow Thickness (mm)	508	86.23%	85.04%	85.63%
In:Clark Level	742	88.52%	85.18%	86.81%
In:Neurotropism	149	97.93%	95.30%	96.60%
In:Vascular/Lymphatic	229	96.44%	94.76%	95.59%
Li:Lexical Polarity Positive	1369	95.31%	92.11%	93.68%
Li:Lexical Polarity Negative	676	96.60%	92.46%	94.48%
Li:Modality	297	90.00%	84.85%	87.35%
Li:Mood and Comment Adjuncts	931	77.84%	70.57%	74.03%
Li:Temporality	167	89.74%	62.87%	73.94%
Ma:Excision Clear	241	88.65%	84.23%	86.38%
Ma:Excision Deep	166	80.69%	70.48%	75.24%
Ma:Excision Invasive	362	64.51%	63.26%	63.88%
Ma:Excision In Situ	88	51.16%	25.00%	33.59%
Re:Desmoplasia	16	100.00%	50.00%	66.67%
Re:Fibrosis	68	80.36%	66.18%	72.58%
Re:Solar Elastosis	25	76.47%	52.00%	61.90%
Re:Tils	212	86.67%	85.85%	86.26%
Sy:Diagnosis	1238	92.86%	89.34%	91.07%
Sy:Regression	201	81.12%	79.10%	80.10%
Sy:Subtype	476	91.16%	88.87%	90.00%
Overall	15981	87.11%	81.65%	84.29%

Table 5.22 Performance of the best model by entity types on the melanoma corpus.

Table 5.22 displays the performance of the best model by entity types. From Table 5.22, most of the entity types attained F-scores of over 60%, whereas F-scores on seven medical entity types and two Linguistic categories were equal to or higher than 90%. However, there was poor performance on some medical entity types: De:Cosmetic Changes, En:Lesion (other) and Ma:Excision In Situ. Lexical variability is one of the possible reasons for this. Over 56% of the De:Cosmetic Changes instances only appear once in the corpus, making it difficult for the machine learner to learn from the training data, thus greatly decreasing the recall. Ambiguity is another possible reason for the low F-score. For example, "lesion" is a common word used by both En:Lesion (other) and En:Primary Lesion. The correct determination of the entity type is not only based on the local context, but also the global context of the whole document. For instance, if a primary lesion has been described in other specimens, the "lesion" appears in the specimen is likely to be a En:Lesion (other). Similarly, there are many similarities between Ma:Excision In Situ and Ma:Excision Invasive: the same words and same linguistic construction of the instances. Here are two examples:

- The in situ component is 1.4mm from the closest lateral resection margin.
- The tumour appears completely excised being 1.6mm from the closest lateral resection margin.

The first example is a Ma:Excision In Situ instance, while the second one is a Ma:Excision Invasive instance, as it can be inferred from the local context of the first instance "in situ component", and no such context can be detected for the second example. However, it is quite difficult to achieve the correct inference in some cases, as the average length of the Ma:Excision In Situ entities is over seven tokens, the local context may locate outside the contextual window. It is also more difficult if the context is situated in other sentences. For example, in these sentences:

> Superficial spreading melanoma extends to one lateral surgical margin. It is 2.5mm clear of the other edge. Invasive melanoma has a cutaneous clearance of 3.5mm and a deep clearance of over 8mm.

where the contexts that infer the entity type of the instance "2.5mm clear of the other edge" are "one lateral surgical margin" and "Invasive melanoma", are located in the previous and next sentences. Thus the machine learner would misclassify the instance to Ma:Excision Invasive, as the classification is at the sentence-level and Ma:Excision Invasive is the default category for lateral margins.

5.6.2 System Performance on Colorectal Cancer Corpus

The system performance on the colorectal cancer corpus according to the contribution of features is displayed in Table 5.23. The outcome of the experiments shows that the model achieved an improvement of about 2% on F-score by applying the lexical feature set, while semantic, morphological and syntactic feature sets also brought some gains by 0.86%, 0.65% and 0.51%. Orthographic features and *section context* improved the model slightly by 0.17% and 0.02%.

Model #	Features	Precision	Recall	F-score
1	Nine-word contextual window	77.96%	70.71%	74.16%
2	M1 + Lowercase of tokens	78.58%	72.27%	75.29%
3	M2 + Lemma	78.58%	72.55%	75.44%
4	M3 + Bag of POS	78.80%	73.30%	75.95%
5	M4 + Medical category	78.84%	73.70%	76.18%
6	M5+ Lexical resource	78.80%	73.80%	76.22%
7	M6 + Bag of expansions of abbreviations and acronyms	78.84%	73.89%	76.28%
8	M7 + Bag of ring-fencing tag	79.23%	74.65%	76.87%
9	M8 + Orthography	79.25%	74.94%	77.04%
10	M9 + Suffixes	79.34%	75.34%	77.29%
11	M10 + Bag of prefixes	79.66%	75.81%	77.69%
12	M11 + Section context	79.59%	75.91%	77.71%
13	M12 + Bigram	80.58%	76.33%	78.40%

Table 5.23 System performance on the colorectal cancer corpus according to the contribution of features.

The scores for the best model by entity types are presented in Table 5.24. From Table 5.24, most entity types achieved good performance with over 60% F-scores. The best performance is on eight entity types with F-scores of higher than 90%, while the worst performance is on five entity types: De:Mesorectal Integrity, De:Specimen Images, En:Coexistent Pathology, Re:Desmoplasia and Fibrosis and Re:Response to Rx, whose F-scores did not reach 60%. The poor performances on De:Mesorectal Integrity and De:Specimen Images were probably due to lack of sufficient training samples. A detailed error analysis on En:Coexistent Pathology shows that there were several possible reasons for the poor performance:

Entity type	Number	Precision	Recall	F-score
De:Ancillary Studies	272	76.49%	70.59%	73.42%
De:Mesorectal Integrity	15	83.33%	33.33%	47.62%
De:Perforation	154	78.50%	54.55%	64.37%
De:Peritoneal Reflection	138	81.54%	76.81%	79.10%
De:Serosa Description	212	73.45%	61.32%	66.84%
De:Specimen Blocks	3343	89.00%	88.33%	88.67%
De:Specimen Images	30	100.00%	30.00%	46.15%
De:Specimen Size	1585	79.82%	78.36%	79.08%
De:Specimen Type	1892	85.91%	81.87%	83.84%
De:Tissue Banking	57	100.00%	87.72%	93.46%
De:Tumour Description	1464	73.25%	68.65%	70.87%
De:Tumour Site	1446	81.60%	76.07%	78.74%
De:Tumour Size	682	80.34%	75.51%	77.85%
En:Coexistent Pathology	1181	56.39%	54.95%	55.66%
En:Distant spread or Metastases	284	69.86%	51.41%	59.23%
En:Lymph Nodes	629	82.30%	76.15%	79.11%
En:Residual Tumour	124	85.19%	55.65%	67.32%
Ex:Donut Involvement	144	72.07%	55.56%	62.75%
Ex:Extent	610	68.43%	55.08%	61.04%
Ex:Extramuscular Spread	197	95.24%	81.22%	87.67%
Ex:Lymph Node Involvement	1133	77.26%	77.05%	77.15%
Ex:Serosal Involvement	392	79.60%	70.66%	74.86%
In:Depth of Invasion	1284	69.45%	69.24%	69.34%
In:Perineural Invasion	396	95.26%	91.41%	93.30%

In:Venous and Small Vessel Invasion	855	88.26%	85.26%	86.73%
Ma:Circumferential Margin	254	68.66%	58.66%	63.27%
Ma:Clear	758	84.68%	83.11%	83.89%
Ma:Proximal or Distal Margin	699	82.62%	79.54%	81.05%
Met:Anatomic Stage	291	95.45%	86.60%	90.81%
Met:M Value	257	98.39%	94.94%	96.63%
Met:N Value	409	97.95%	93.64%	95.75%
Met:T Value	414	95.93%	91.06%	93.43%
Re:Desmoplasia and Fibrosis	271	60.00%	45.39%	51.68%
Re:Response to Rx	106	65.31%	30.19%	41.29%
Re:TILS and Peritumoural Lymphocytes	389	90.62%	86.89%	88.71%
Sy:Comment	1558	60.43%	62.64%	61.52%
Sy:Histological Grade	828	96.03%	93.60%	94.80%
Sy:Histological Type	998	92.87%	88.78%	90.78%
Sy:Medical History	206	79.07%	49.51%	60.90%
Overall	25957	80.58%	76.33%	78.40%

Table 5.24 Scores for the best model by entity types on the colorectal cancer corpus.

1. The machine learner could not detect the boundary correctly occasionally as the there are no specific requirements for its boundary in the annotation schema. In the annotations, there were various grammatical structures for En:Coexistent Pathology: a noun phrase (e.g., "villous adenoma"), multiple noun phrases connected by a preposition (e.g., "tubulovillous adenoma with high grade dysplasia"), a verb phrase and a noun phrase connected by a proposition (e.g., "arising within a moderately dysplastic tubulovillous adenoma"), a clause (e.g., "overlying adenoma seen in some of the sections"), a sentence (e.g., "Two polyps are identified within the ascending colon 7 and 12mm"), etc. Therefore, it was too difficult for the machine learner to learn from these structures and predict the potential instances as well.
2. The machine learner would be confused with other entity types in some case (e.g., Sy:Comment). The annotation schema defined that if a coexistent pathological abnormality is absent, it should be annotated as SY: COMMENT instead. For example, in the following sentence:

   > These sections show changes of diverticular disease with no evidence of diverticulitis.

   the first instance is a En:Coexistent Pathology, while the second one is a Sy:Comment instance. However, as negation phrases like "no evidence of" were not designed to be annotated separately in the schema, the machine learner could not classify the second instance correctly.
3. Ambiguity is another possible reason. For example, if a tumour is a polypoid lesion, the machine learner would usually misclassify it to En:Coexistent Pathology, as "polyp" is a familiar coexistent pathological abnormality. An example is presented below:
   The gold-standard is

   > Located 130mm from the distal resection margin ["Ma:Proximal or Distal Margin"] is a pedunculated polyp ["De:Tumour Description"] measuring 25x25x25mm ["De:Tumour Size"].

The system prediction is

> Located 130mm from the distal resection margin is a pedunculated polyp measuring 25x25x25mm ["En:Coexistent Pathology"].

Likewise, the above reasons also caused most of the classification errors on Re:Desmoplasia and Fibrosis and Re:Response to Rx, such as the incorrect boundary detection. The similar grammatical structures among Re:Desmoplasia and Fibrosis, Sy:Comment and En:Coexistent Pathology can confuse the learning of the machine learner in some cases. For instance, it tagged the following example: "OTHER FINDINGS: The appendix is fibrosed consistent with either old ischemia or previous inflammation" as Sy:Comment, while the correct type should be Re:Desmoplasia and Fibrosis. The machine learner could not discriminate the instances from Re:Desmoplasia and Fibrosis and Re:Response to Rx sometimes, if desmoplasia or fibrosis are the responses to the treatment. For example, the instance "extensive fibrosis suggesting at least moderate response to therapy" would be misclassified as Re:Desmoplasia and Fibrosis, but the correct assignment should be Re:Response to Rx instead, as the fibrosis is a manifestation of the response to the therapy.

5.6.3 System Performance on Lymphoma Corpus

Table 5.25 presents the system performance on the lymphoma corpus with the contribution of features. The performance of the baseline model was improved significantly by 3.3% by considering the lexical feature set, whereas a prominent improvement (2.40%) was made by introducing *lowercase of tokens*. Furthermore, a considerable improvement was contributed by the semantic feature set (nearly 2.83% gain), wherein more of the gain was achieved by the *bag of ring-fencing tag* (1.87%), as well as the *section context* feature (1.66% improvement). A relatively small gain (0.55%) was achieved by integrating the syntactic feature set, and orthographic and morphological features only accounted for minimal improvements of 0.13% and 0.16%.

Model #	Features	Precision	Recall	F-score
1	Nine-word contextual window	84.91%	69.66%	76.53%
2	M1 + Lowercase of tokens	85.33%	73.43%	78.93%*
3	M2 + Lemma	85.16%	74.74%	79.61%
4	M3 + POS	85.20%	75.68%	80.16%
5	M4 + Medical category	85.44%	76.49%	80.72%
6	M5+ Lexical resource	85.79%	76.93%	81.12%
7	M6 + Correction of misspelling	85.61%	77.47%	81.34%
8	M7 + Bag of ring-fencing tag	85.65%	80.90%	83.21%*
9	M8 + Suffixes	85.64%	81.11%	83.31%
10	M9 + Bag of prefixes	85.81%	81.07%	83.37%
11	M10 + Section context	87.49%	82.71%	85.03%*
12	M11 + Brief word class	87.52%	82.93%	85.16%

Table 5.25 System performance on the lymphoma corpus with the contribution of features. Scores marked with * suggest significant contribution within 95% confidence interval.

Entity type	Number	Precision	Recall	F-score
An:Biomarker	1928	93.07%	94.09%	93.58%
An:Cytogenetics Comment	18	40.00%	11.11%	17.39%
An:Fish Results	8	0.00%	0.00%	0.00%
An:Flow Cytometry-Comment	88	86.30%	71.59%	78.26%
An:Flow Cytometry-Negative	3	0.00%	0.00%	0.00%
An:Flow Cytometry-Positive	3	0.00%	0.00%	0.00%
An:IgH Test	7	50.00%	28.57%	36.36%
An:Immunohistochemistry-Comment	246	50.24%	42.68%	46.15%
An:Immunohistochemistry-Equivocal	27	86.96%	74.07%	80.00%
An:Immunohistochemistry-Negative	284	96.45%	95.77	96.11%
An:Immunohistochemistry-Positive	594	93.54%	92.59%	93.06%
An:PCR Comment	27	96.00%	88.89%	92.31%
An:TCRgamma Test	2	0.00%	0.00%	0.00%
De:Anatomical Structure	396	80.45%	72.73%	76.39%
De:Architecture	472	77.48%	72.88%	75.11%
De:Cell Clonality	1	0.00%	0.00%	0.00%
De:Cell Size	488	90.12%	89.75%	89.94%
De:Cytomorphology	178	86.16%	76.97%	81.31%
De:Laterality	18	0.00%	0.00%	0.00%
De:Lineage	140	83.90%	70.71%	76.74%
De:Other Size	123	56.94%	33.33%	42.05%
De:Preservative Fluid	110	94.39%	91.82%	93.09%
De:Sample Triage	793	92.43%	87.77%	90.04%
De:Specimen Blocks	847	94.02%	90.91%	92.44%
De:Specimen Size	467	81.38%	86.08%	83.66%
De:Specimen Type	802	91.23%	89.53%	90.37%
De:Tissue Source	1482	87.91%	81.92%	84.81%
De:Topography	1404	83.04%	79.84%	81.41%
De:Tumour Size	51	23.53%	7.84%	11.76%
En:Coexistent Pathology	160	83.61%	63.75%	72.34%
Ex:Disease Extent	8	100.00%	12.50%	22.22%
Ex:Other Sites of Disease	53	50.00%	22.64%	31.17%
Li:Lexical Modality	322	84.64%	70.19%	76.74%
Li:Lexical Polarity Negative	366	85.63%	79.78%	82.60%
Li:Lexical Polarity Positive	1125	87.45%	89.78%	88.60%
Li:Mood and Comment Adjuncts	607	81.77%	77.59%	79.63%
Li:Temporality	133	81.11%	54.89%	65.47%
Re:Tissue Reaction	259	78.88%	70.66%	74.54%
Sy:Clinical Impression	185	81.92%	78.38%	80.11%
Sy:Comment	77	90.28%	84.42%	87.25%
Sy:Constitutional Symptoms	28	81.82%	32.14%	46.15%
Sy:Diagnosis	1056	85.62%	84.00%	84.80%
Sy:Diagnosis Subtype	30	70.00%	23.33%	35.00%
Sy:Indication for Biopsy	28	75.00%	21.43%	33.33%
Sy:Medical History	82	69.44%	60.98%	64.94%
Sy:Predisposing Factors	60	68.57%	40.00%	50.53%
Sy:Presentation	98	71.23%	53.06%	60.82%
Sy:SNOMED RT Codes	994	99.40%	99.30%	99.35%
Sy:Stage	3	0.00%	0.00%	0.00%
Sy:WHO Grade	134	87.31%	87.31%	87.31%
Overall	16815	87.52%	82.93%	85.16%

Table 5.26 Performance of each entity type attained by the best model on lymphoma corpus.

Table 5.26 shows the performance of each entity type attained by the best model. From Table 5.26, although the overall micro-averaged F-score was up to 85.16%, and nine medical entity types obtained F-scores exceeding 90%, unsatisfactory results were presented on a considerable number of entity types, some with extremely low F- score (0%). The dramatic loss of the F-scores on these types was mostly caused by insufficient training examples, e.g., there is only one De:Cell Clonality instance and two An:TCRgamma Test instances in the training data. Besides the small sample size, massive lexical variants are another possible reason for the drop of F-score, e.g., there are 39 lexical variants of Ex:Other Sites of Disease in total, wherein 30 of them have a frequency of one, so it is difficult for the machine learner to learn from the training data effectively. Abbreviation seems to be a challenge for recognising Predisposing Factors. More than half of the Sy:Predisposing Factors instances consist of abbreviations or acronyms, such as chemo, CTx, and HIV +ve. It is believed that classifying abbreviations is harder than classifying full terms in the biomedical domain. Ambiguity also causes most of the classification errors on De:Other Size and De:Tumour Size. In this example: "The nodule ranges in size between 3 and 6mm", the gold-standard is De:Tumour Size as the "nodule" appears to be a tumour in this example. Nevertheless, "nodule" can represent other entities in different cases, such as a specimen or a coexistent pathological abnormality, and the associated entity types of the size should be De:Specimen Size and De:Other Size. Hence, it requires deep insight into the whole document to correctly identify these instances, and additional features need to be considered. Except for lexical variability, the long span of the instances is likely to be a problem for identifying An:Immunohistochemistry-Comment. The gold-standard of this example: "CD3, CD5 and CD 43 label moderate numbers of apparently small cells throughout the specimen" is An:Immunohistochemistry-Comment. However, as its span exceeds the nine-word contextual window, the machine learner could not identify it correctly and tagged it as follows:

> CD3 ["An:Biomarker"], CD5 ["An:Biomarker"] and CD43 ["An:Biomarker"] label moderate numbers ["Li:Mood and Comment Adjuncts"] of apparently small cells ["De:Cell Size"] throughout the specimen.

5.6.4 Discussion on Three Corpora

Although all features or their combinations with contextual windows (except for the *bigram* feature) depicted in Section 5.5.2 were attempted in feature engineering (which resulted in 32 models prepared for each corpus), not every one of them was effective on each corpus. Some only worked on two corpora or one corpus. The *full word class* feature was not helpful, which was discarded during feature engineering. Table 5.27 tabulates these common beneficial features across the three corpora, and specific features which were useful for two corpora or one corpus. Consequently, there were 19, 19 and 20 models discarded during feature engineering for the melanoma, colorectal cancer and lymphoma corpora respectively.

Beneficial feature	**Corpus**
Contextual window	Melanoma, colorectal cancer, lymphoma
Section context	Melanoma, colorectal cancer, lymphoma
Lowercase of token	Melanoma, colorectal cancer, lymphoma
Lemma	Melanoma, colorectal cancer, lymphoma
Correction of misspelling	Melanoma, lymphoma
Expansion of abbreviations and acronyms/Bag of expansions of abbreviations and acronyms	Melanoma, colorectal cancer
Bigram	Melanoma, lymphoma
Bag of prefixes	Melanoma, colorectal cancer, lymphoma
Suffixes	Melanoma, colorectal cancer, lymphoma
Lexical resource	Colorectal cancer, lymphoma
Medical category	Melanoma, colorectal cancer, lymphoma
Ring-fenced tag/Bag of ring-fencing tag	Melanoma, colorectal cancer, lymphoma
POS tag/Bag of POS	Melanoma, colorectal cancer, lymphoma
Chunk	Melanoma
Orthography	Colorectal cancer
Brief word class	Lymphoma

Table 5.27 Beneficial features and their contribution to the corpora.

It can be seen from the results that the scores of the baseline models are relatively high (all over 74%), which indicates that contextual and lexical information is very useful for recognising medical entities. The best feature configurations yielded prominent gains on F-scores from 4.24% to 6.84%. Most of the gains were brought by some common features: *lowercase of tokens*, *lemma*, *POS tag* (or *bag of POS tag*), *medical category*, *ring-fenced tag* (or *bag of ring-fenced tags*), *suffixes*, *bag of prefixes*, and *section context*.

The *lowercase of tokens* feature normalize the orthographic variants of a token, which is a good supplement for the basic lexical information, the token itself, and significantly increased both precision and recall. Given the larger proportions of unique case insensitive tokens in the token collection overall held by the melanoma and lymphoma corpora, the models of these two corpora benefited more from this feature. *Lemma* is another lexical feature that normalizes the morphological variants of a token, and *POS tag* is a simple syntactic feature to generalise the representation of the tokens. They both increased the recall to some extent. It is likely that the improvement they achieved was hindered by the accuracy of GENIA tagger on the texts. For example, the GENIA tagger assigned "VBG" as a POS tag for "Advancing" in this sentence "Advancing edge of tumour: Circumscribed.", but the correct tag should be "JJ" instead.

It was observed that a great number of the entities have very low frequency (some may have only one) so that the machine learner was unable to learn from the insufficient training examples. The medical categories generated by the TTSCT service utilising the SNOMED CT lexicon and the semantic tags provided by the ring-fenced tagging engine, were able to compensate for the drop of recall caused by unseen data in the test subset of each fold to some degree. The semantic knowledge provided by these features can benefit both the determination of the presence of an entity and the classification of the entity type. As the lymphoma corpus has a moderate number of tag types assigned in the basic pattern file (about 40) prepared for the ring-fenced tagging engine, the engine was likely to perform better on this corpus, which consequently led to a bigger gain on the system performance. Affix features boosted the F-scores by 0.16-0.65%, as the generalizability of the affixes in some entities may increase the recall. Note that *bag of prefixes* seems to be more informative than prefixes by combining prefixes with contextual windows. Given the different distribution of entities in the sections, *section context* was added to determine the presence of an entity. It was more powerful on the lymphoma corpus than on other corpora, probably because there are more types of section contexts in this corpus.

The orthographic features were supposed to be able to capture the capitalised information in abbreviations or acronyms and generalise tokens that contain numeric or punctuation in measurements and named entities such as “0.5mm” and “CD10”. But they were not as effective as expected, and only yielded very small gains in the colorectal cancer and lymphoma corpora (both less than 0.2%). This suggests that the pathologists might not follow consistent formation conventions and used the orthography of words arbitrarily when writing the reports, especially for melanoma pathology notes. It is observed that several notes were written in all uppercase format, and some uppercase words were intentionally used by the pathologists for emphasis in the sentences (e.g., “Sections show a MALIGNANT MELANOMA of superficial spreading type.”). This arbitrary variation of orthography could introduce noise in the learning of the models.

The syntactic feature *chunk* only worked on the melanoma corpus, yet with a minor improvement of 0.08% F-score, probably because of the specific boundary requirements of several entities (accounting for about 8.65% of total instances).

Surprisingly, the lexical feature *correction of misspelling* did not work on the colorectal cancer corpus, although it has more entries in the correction dictionary compared to the other two corpora. This is possibly due to the defects in the proof reading process: during proof reading, the correction was assigned for a misspelling according to the most frequent context it occurred in the corpus, thus only one form of correction would be considered for a misspelling even if there were other alternative forms of correction given in different contexts; moreover, some misspellings might not be identified during the process. Similarly, the lexical feature *expansion of abbreviations and acronyms* diminished the performance on the lymphoma corpus, probabaly due to erroneous expansion of ambiguous abbreviations. The *bigram* feature also had adverse effect on the lymphoma corpus, which was probably because of the relative lower frequencies of common bigrams in this corpus. It can be seen

from Table 5.15 that except for the bigram "lymph/node", the frequencies of other bigrams were significantly lower than those in the other two corpora.

The effectiveness of the semantic feature *lexical resource* seemed to be related to the ratio between the number of entries in the medical dictionary and that in Moby dictionary; it was more effective on the corpus with bigger ratio. For example, by using this feature, the lymphoma corpus had more gain on the F-score than the colorectal cancer corpus (0.40% *vs.* 0.04%), owing to its larger ratio (1:9.21 *vs.* 1:10.12). If this ratio is too small, the feature would be disadvantageous for the system performance instead. The application of this feature to the model on the melanoma corpus performed worse.

From Table 5.22, 5.24 and 5.26, there is a consistent gap between precision and recall, where the recall is 4.25 ~ 5.46% less than precision. The better performance is usually on the entity types with high frequency, such as Sy:Diagnosis in the melanoma corpus, Sy:Histological Type in the colorectal cancer corpus and An:Biomarker in the lymphoma corpus. This suggests that a sufficient training sample is a crucial factor in achieving both high precision and recall. However, besides the training sample size, consistent expressions of the instances are also important for the classification. For example, though the amount of De:Tissue Banking instances is relatively small, it still obtained very high F-score of 93.46% as its training data always contain lexical items like "tissue bank", "tissue banking", "TB", etc. Another example is Sy:SNOMED RT Codes. All of its instances follow a pattern: Code ID + Code name, such as "M-95903 Malignant lymphoma, NOS" and "T-C4480 Aortic lymph node". Such consistent expressions improve both learning and prediction by the models.

From the above analyses of each corpus, there are several common reasons accounting for the poorer performance on some entity types. Abundant lexical variants are the major one leading to low recall. Many variants have a frequency of only one. Though generalised information provided by certain features such as *POS tag*, *suffixes*, and *bag of prefixes* can partially cure this problem, there is still a lack of lexical information for the classification. Utilisation of the semantic feature sets like *medical category* and *ring-fenced tag* has shown its advantages in tackling this problem, but these resources are not exhaustive, and they may not cover the personally idiosyncratic writing styles of different pathologists in the notes. Ambiguity is another possible reason for the low F-scores. It manifests as similar use of the lexicons or grammatical structures, which can confuse the classification of the machine learner. More complicated contextual information involving the adjacent sentences may be helpful to solve this problem. The problems caused by the long span of instances may be a defect of CRF++, which restricts the maximal value of contextual window size to be nine. Using other machine learning algorithms (e.g., Support Vector Machines (SVM)) can be considered as a possible solution for this problem.

	Melanoma corpus			**Colorectal cancer corpus**			**Lymphoma corpus**		
Matching Criteria	Precision	Recall	F-score	Precision	Recall	F-score	Precision	Recall	F-score
Exact match	87.11%	81.65%	84.29%	80.58%	76.33%	78.40%	87.52%	82.93%	85.16%
Left boundary match	89.82%	84.19%	86.91%	85.53%	81.02%	83.21%	89.82%	85.11%	87.40%
Right boundary match	91.49%	85.75%	88.53%	86.37%	81.81%	84.03%	90.00%	85.28%	87.58%
Sloppy match	97.78%	91.65%	94.62%	98.85%	93.63%	96.17%	96.93%	91.85%	94.32%

Table 5.28 Partial match performance on the three corpora.

Table 5.28 lists the partial match performance on the three corpora. As suggested by the above analyses, many errors occurred at the boundary of the entities, and resulted in the poor performance of these entities. In the colorectal cancer corpus, the partial matching F-scores are significantly higher than that of the exact match by about 4.8%, 5.6% and 17.8% of the left boundary match, right boundary match and sloppy match respectively, which outperformed those in other corpora. This indicates that the system performance was hindered by the quality of the detection of entity boundaries to a greater extent in this corpus. This is possibly because:

- The average length of the entities is longer in this corpus than those in other corpora (more than twice).
- Some synoptic fields present in the corpus. As discussed in the section heading detection, they were annotated as part of the entities or subheadings in the gold-standard according to the amount of reportable fields they contained instead of linguistic structures.

Hence, the detection of entity boundaries seems to be a more difficult issue for the machine learner on this corpus.

Likewise, the smaller gaps between the F-scores of partial matches and exact match in the lymphoma corpus, implies that the misclassification of entity type is more likely to account for the classification errors. This is probably because the total entity types have larger numbers in this corpus than those in other corpora, and the machine learner is more confused when making decisions on entity types. Note that the inconsistent annotations of the entity boundary can also cause some faults on boundary detection.

5.6.5 Limitations

The overall results are promising, with micro-averaged F-scores ranging from 78.4% to 85.16%. Nonetheless, there are still some notable limitations in the experiments:

1. Additional features can be introduced into the models. For example, features that represent more complicated contextual information may remedy the ambiguity problem. The second

order category from TTSCT can also be considered as another medical category feature to enrich the semantic feature set.

2. Although nearly all other features were tried in combination with the contextual window to boost the system performance, the same nine window size was applied to them. This cannot rule out the possibility that the combination with different window sizes may yield better performance.
3. One of the disadvantages of using CRF is that CRF is likely to bias to the majority entity types in the classification, especially when ambiguous expressions occur in the minority counterparts. The voting or stacking strategies for aggregating the results from different machine leaning classifiers might be applied to resolve this issue (Dzeroski and Zenko, 2004; Wang and Patrick, 2009).

5.7 Conclusion

A supervised machine learning-based approach is proposed to recognise medical entities in the corpora. The spans of most entity types are smaller than nine, thus CRF-based models were able to capture a significant portion of the entity boundaries by using contextual information. The application of rich feature sets provides useful clues for the classification of entity types. By feature engineering, the best feature configurations were attained, which yielded prominent gains on F-scores from 4.24% to 6.84%. Several common beneficial features were identified, which can be helpful for other MER tasks using similar approaches.

The error analyses show that lexical variability and ambiguity are two main causes accounting for the poorer performance on some entity types. The limitation of the machine learning method can also result in some mistakes on the entities with a relatively long span. Future work will involve improving the performance of the boundary detection (e.g., using other machine learning algorithms), and classification of entity types (such as introducing additional features).

Chapter 6 Negation and Uncertainty Detection

6.1 Introduction

In the clinical domain, when a particular term appears in a patient record, it does not mean the clinical finding or condition it represents occurs in the patient, or the procedure it refers to has been performed on the patient. Actually, nearly half of all symptoms, diagnoses, and findings in clinical reports are estimated to be negative or uncertain (Chapman et al., 2001a). Without discrimination between the negative or uncertain information and the positive in an information retrieval and information extraction (IE) system, the reliability of the extracted information is diminished and this causes redundancy when indexing. For example, in the sentence: "CV - Ischemia ASA, lisinopril Pump no evidence of failure", the clinical finding "failure" is negated, which suggests that it can be ruled out for the patient. In another example: "Possible aspiration pneumonia", the clinical condition "aspiration pneumonia" is uncertain, which indicates that the patient may have it, but is not confirmed to have it. Negation and uncertainty detection was also part of the assertion classification task in the 2010 i2b2/VA Challenge, which was to determine what the clinical note asserts the medical problem to be based on and the context in which it is used (Uzuner et al., 2011) .

In pathology notes, negative or uncertain findings or diagnoses also appear frequently. To find out whether a finding is present, absent or uncertain is critical to making the correct diagnosis and prognosis for the patient. The presence or absence of a particular disease can influence the clinical management of the patient. For example, the treatment can be different for Hodgkin's lymphoma and non-Hodgkin's lymphoma. According to the protocols, there are several fields explicitly indicating that pathologists should record the findings whether they are present or not, such as "S3.04 Ulceration" in Primary Cutaneous Melanoma Structured Reporting Protocol (Scolyer et al., 2010). Thus, negation and uncertainty detection is an important component in the study and definitely can affect the final output for these fields in the structured templates.

In the previous chapter, potential medical entities have been identified by the medical entity recognition (MER) system, as well as Linguistic categories for the melanoma corpus and the lymphoma corpus. The study presented in this chapter focuses on the detection of absent and uncertain assertions for a selection of these entities. It begins with an overview of current methodologies for negation and uncertainty detection, and is then followed by a case study on the lymphoma corpus where three different approaches are experimented with, and the preferable method for the other two corpora. The associated results and discussions for these methods on the three corpora are presented as well.

6.2 Case Study on Lymphoma Corpus

A case study on the lymphoma corpus was carried out to find out a suitable method to be implemented for this project. Another objective of the case study is to discuss the advantages and disadvantages of different methods for negation detection on narrative pathology reports.

The lymphoma corpus was chosen ahead of the other corpora, because:

- It has the smallest number of training documents, thus it was likely to be less labour and time consuming for the annotation and evaluation of the gold-standards.
- It has a medium amount of tokens and entities, but the largest number of entity types amongst the corpora, hence it might be more representative than the other two corpora.

Besides the 227 reports mentioned in the previous chapters as a training set, an additional 57 reports were collected as a test set.

6.2.1 Negation Detection

In this study, only pertinent negations within a sentence are considered as valid instances. The pertinent negations indicate "completely absent", while partial negations such as "probably not" and "unlikely" were excluded. Normal or abnormal findings and test results, and related comments also were not considered. In the following sentence:

> The absence of CD15 expression ["An:Immunohistochemistry-Comment"] and the cellular arrangement of the large atypical cells ["De:Cell Size"] is much more in favour of ["Li:Lexical Modality"] a diffuse large B-cell lymphoma ["Sy:Diagnosis"].

"absence of CD15 expression" is a comment made by the pathologist based on the immunohistochemistry test results, thus it was not considered as negation. Negative prefixes or suffixes are also not considered, because they are often semantically ambiguous or they are part of an entity, e.g., "non-Hodgkin's malignant lymphoma" can represent several sub-types of malignant lymphoma, except for Hodgkin's lymphoma.

Motivated by the approaches mentioned above, three different methods were applied in this study to detect negation. Given a medical entity in a sentence, the methods seek to determine whether the entity is negated.

The processing components shown in Figure 6.1 include:

1. The MER system introduced in the previous chapter which annotates the medical entities and instances of Li:Lexical Polarity Negative in the test set. Not all types of entities were utilised, the selection of particular entity types was based on their definitions in the annotation schema, their associated fields in the protocol, and thorough analysis on the training data. The selected entity types were Sy:Clinical Impression, Ex:Other Sites of Disease, Sy:Constitutional Symptoms, Sy:Predisposing Factors, De:Architecture,

De:Cytomorphology, Re:Tissue Reaction, Sy:WHO Grade, En:Coexistent Pathology, Sy:Diagnosis and Sy:Diagnosis Subtype.

2. Different methods are applied to detect negated medical entities.
3. The results from the negation detection module are filtered by identification of pseudo-negations.
4. The final output is evaluated and compared to the performances of the other methods.

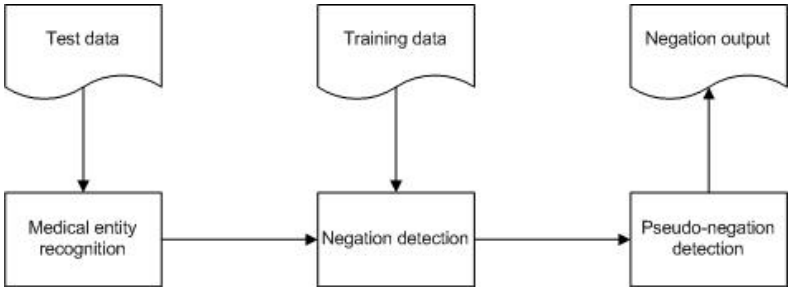


Figure 6.1 Processing components for negation detection on the lymphoma corpus.

Lexicon-based Approach

NegEx defined three types of terms: trigger terms, pseudo-trigger terms, and termination terms (Chapman et al., 2001b). Trigger terms are some negation phrases, also known as the negation cues; pseudo-trigger terms are phrases that indicate double negatives or modified meanings; termination terms are used to restrict the scope of the negation.

Similar to NegEx, trigger and termination terms were also adopted in this method. The instances of Li:Lexical Polarity Negative in the training data were used as trigger terms, and divided into three groups according to their positions relative to a medical entity: Group 1- the instance precedes a medical entity, Group 2 - the instance succeeds a medical entity, Group 3 - any of the above positions. Besides some specific words, particular punctuation was also considered as termination cues. Note that the termination punctuation varied between different section contexts. These terms or cues and some examples are compiled in Table 6.1. From Table 6.1, it can be seen that there are only four lexical entries of trigger terms the same as those used in NegEx: “no”, “no evidence of”, “not”, and “without”.

There are some differences between this approach and NegEx:

- In this approach, “not” is defined as a trigger term that can either precede or succeed a medical entity.
- Termination cues include specific punctuation.
- The negation scope is not constrained in a fixed context window.

Type of term/cue	Sub-category	Example
Trigger term	Group 1	devoid of, no features of, no morphological evidence of, not sufficient to, lacking, without, exclude, lack, no definite evidence of, none, rather than, no convincing evidence of, no evidence of, no
	Group 2	absent, not a feature
	Group 3	not
Termination cue	Termination term	but, which, though, although, however, so, whether, involved by, based on
	Termination punctuation	"CLINICAL HISTORY" and "SPECIMEN": , | ; | (|) | ->
		"SUMMARY" and "SUPPLEMENTARY SUMMARY": (|) | -
		Other section contexts: (|)

Table 6.1 Trigger terms and termination cues for negation detection in the lymphoma corpus. Note: word examples are separated by comma ","; punctuation examples are separated by pipe "|".

This rule-based method can be summarized in three steps, and is illustrated in Figure 6.2:

1. Find out whether there is at least one termination cue between the trigger term and the medical entity. If there is, filter out the entity.
2. Validate the position of the trigger term to the entity. If it is not the same as defined in the associated group, filter out the trigger term.
3. If there are multiple trigger terms, repeat the above two steps; if the trigger term is not filtered out, yield "absent" as the output.

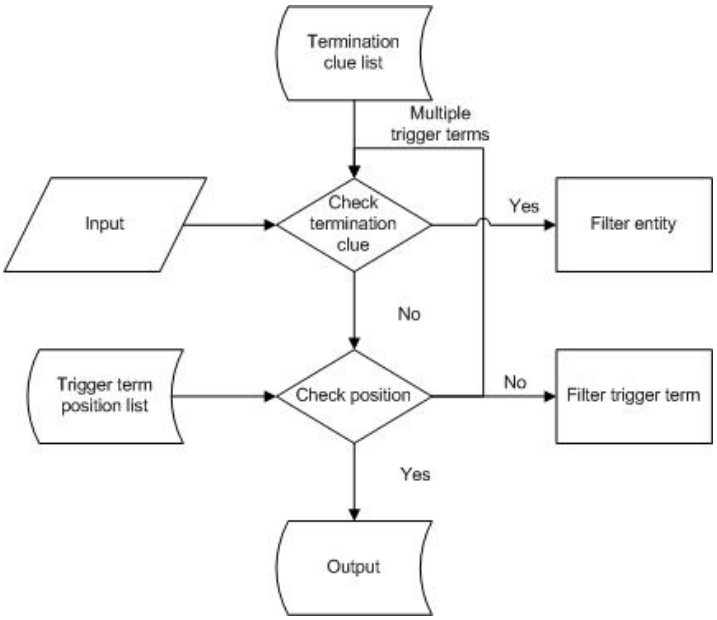


Figure 6.2 Workflow of the lexicon-based approach for negation detection on the lymphoma corpus.

Syntax-based Approach

The Stanford parser (Klein and Manning, 2003) is a well-known probabilistic natural language parser that computes the grammatical structure of new sentences based on knowledge of language gained from hand-parsed sentences. It can provide both phrase structure trees and Stanford dependencies as output. The usages of the output were different in a variety of syntax-based approaches for negation detection. In Huang et al's work, the parse tree output was utilised for the deriviation of a negation grammar (Huang and Lowe, 2007). Firstly, they constructed grammar rules from the parse trees, and then translated these rules into a structural rule to extract negated phrases: the classification of negations were firstly based on the syntactical category of negation signals, and further based on phrase patterns to locate negated phrases. DepNeg used several types of negation patterns based on dependency paths, which were computed from the dependency parse (Sohn et al., 2012). The negation patterns included:

- Negated Verbs – if a particular verb is negated, the whole verb phrase is negated as well, including the objects or complements of the verb.
- Negative Verbs – Particular verbs indicate exclusion of the direct object of the verbs.
- Negative Prepositions – Particular prepositions negate the object of the prepositions.
- Negated Nouns – Certain determiners negate the nouns they modify.
- Negative Adjectives – Certain adjectives negate the nouns they modify.
- Conjunction Expansion – A general rule can be applied to every other pattern to allow conjunctions or lists of the targets above.

The medical entities were identified in a named entity recognition module in DepNeg, which resembles the methods developed in this study. Therefore, the syntax-based approach prepared in this study also uses the dependency paths from the parser to extract the rules.

A set of grammatical relations was drawn from the Stanford dependencies (de Marneffe et al., 2006), which are all binary relations that hold between a governor and a dependent.

- Adverbial modifier (advmod): An adverbial modifier of a word.
- Adjectival modifier (amod): An adjectival modifier of a noun phrase.
- Appositional modifier (appos): An appositional modifier of a noun phrase.
- Conjunct (conj): The relation between two elements connected by a coordinating conjunction.
- Dependent (dep): When the parser is unable to determine a more precise relation between two words, it assigns this label to the words.
- Determiner (det): The relation between the head of a noun phrase and its determiner.
- Direct object (dobj): A noun phrase which is the accusative object of a verb.
- Infinitival modifier: An infinitive that serves to modify a noun phrase.
- Negation modifier (neg): The relation between a negation word and the word it modifies.
- Noun compound modifier (nn): Any noun that serves to modify the head noun.

- Nominal subject (nsubj): A noun phrase which is the syntactic subject of a clause.
- Passive nominal subject (nsubjpass): A noun phrase which is the syntactic subject of a passive clause.
- Participial modifier (partmod): A participial verb form that serves to modify a noun phrase or sentence.
- Object of a preposition (pobj): The head of a noun phrase following the preposition or the adverbs “here” and “there”.
- Prepositional modifier (prep): Any prepositional phrase that serves to modify a verb, adjective, noun, or another preposition.
- Relative clause modifier (rcmod): A relative clause modifying a noun phrase.
- Open clausal complement (xcomp): A clausal complement without its own subject, and is determined by an external subject.

In the collapsed representation, dependencies involving prepositions and conjuncts are collapsed to get direct dependencies between content words. Conjuncts involve conjunctions “and” and “or” are collapsed as “conj_and” and “conj_or”. Several variant conjunctions for “and not": “but not”, “instead of”, “rather than”, and “but rather” are collapsed as “conj_negcc”. Prepositional modifiers regarding prepositions “of”, “without “, and “such as” are collapsed as “prep_of”, “prep_without” and “prep_such_as” respectively.

Firstly, the dependency path between a medical entity and a Li:Lexical Polarity Negative instance can be computed from the result between the head words of the entity and the instance as two nodes in the dependency parse of the sentence. Figure 6.3 displays an example of dependency parse of the sentence: “No necrosis is identified.”

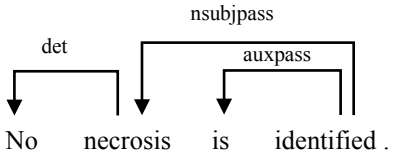


Figure 6.3 Dependency parse of the sentence: “No necrosis is identified.”

The dependency path between “No” and “necrosis” is

det (necrosis-2, No-1)

where “1” and “2” are the positions of “No” and “necrosis” in the sentence respectively.

Several rules were designed according to the dependency path based on manual analysis of the gold-standards in the training data. Other words except for the headwords in the path are called linkage words. The length of the dependency path is calculated as follows: if two nodes are connected directly with a grammatical relation, the length is zero; if two nodes are connected indirectly with two

grammatical relations and one linkage word, the length is one; if two nodes are connected indirectly with three grammatical relations and two linkage words, the length is two, etc. Figure 6.4 divides the rules into several categories according to the length of the dependency path, grammatical relations, the role of the headwords and linkage words, and the prerequisite conditions.

In total, seven negation patterns could be derived from the combinations of the rules:

Pattern 1: Rule #1

Pattern 2: Rule #3

Pattern 3: Rule #5

Pattern 4: Rule #2 / Rule #4 / Rule #6 + Rule #7 / Rule #9 / Rule #11

Pattern 5: Rule #2 / Rule #4 / Rule #6 + Rule #8 + Rule #13

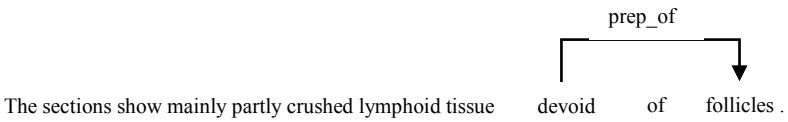
Pattern 6: Rule #2 / Rule #4 / Rule #6 + Rule #10 + Rule #14

Pattern 7: Rule #2 / Rule #4 / Rule #6 + Rule #12 + Rule #15

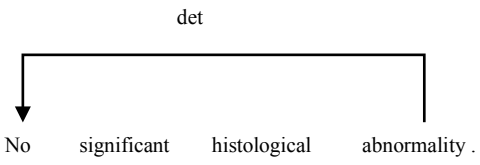
Note that for Pattern 6, grammatical relation “dep” cannot co-occur in Rule #10 and Rule #14. It is presumed that “dep” occurs once in Rule #10 or Rule #14, which may be due to a rare grammatical construction or an unresolved long distance dependency; whereas “dep” occurs both in Rule #10 and Rule #14, which is more likely because of an error from the parser.

Examples for each pattern are presented in graphical form as follows.

Pattern 1: Rule #1



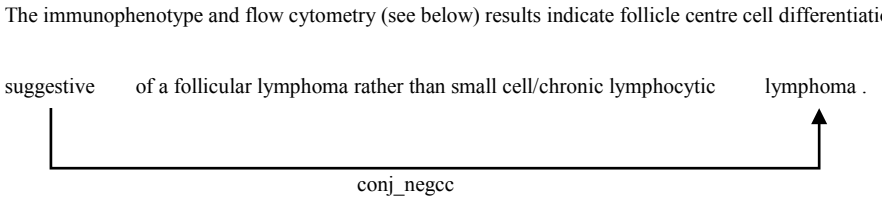
Pattern 2: Rule #3



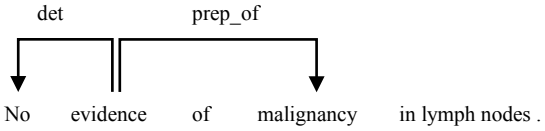
Rule	Length of the path	Grammatical relation (abbreviation)	The role of the headword of the "Lexical Polarity Negative" instance	The role of the first linkage word in the path	The role of the second linkage word in the path	The role of the headword of the medical entity	The order of appearance in the output from the parser	Condition
#1	0	GR 1	governor	N/A	N/A	dependent	N/S	N/A
#2	1	GR 1	governor	dependent	N/A	N/A	N/S	N/A
#3	0	GR 2	dependent	N/A	N/A	governor	N/S	N/A
#4	1	GR 2	dependent	governor	N/A	N/A	N/S	N/A
#5	0	GR 3	N/A	N/A	N/A	dependent	N/S	N/A
#6	1	GR 3	N/A	dependent	N/A	N/A	N/S	N/A
#7	1	GR 4	N/A	governor	N/A	dependent	Proceeding the match rule in the condition	Match one of the rules in #2, # 4 and #6
#8	2	GR 4	N/A	governor	dependent	N/A	Proceeding the match rule in the condition	Match one of the rules in #2, # 4 and #6
#9	1	GR 5	N/A	governor/ dependent	N/A	dependent/ governor	Succeeding the match rule in the condition	Match one of the rules in #2, # 4 and #6
#10	2	GR 5	N/A	governor/ dependent	dependent/ governor	N/A	Succeeding the match rule in the condition	Match one of the rules in #2, # 4 and #6
#11	1	"nsubj"	N/A	dependent	N/A	governor	Succeeding the match rule in the condition	Match one of the rules in #2, # 4 and #6
#12	2	"ccomp"	N/A	dependent	governor	N/A	Succeeding the match rule in the condition	Match one of the rules in #2, # 4 and #6
#13	2	GR 6	N/A	N/A	governor	dependent	Proceeding the match rule in the condition of #8, succeeding #8	Match # 8
#14	2	GR 6	N/A	N/A	governor	dependent	Succeeding the match rule in the condition of #10	Match #10
#15	2	GR 6	N/A	N/A	governor	dependent	Proceeding #12	Match # 12

Figure 6.4 Rules for constructing negation patterns. N/S: Not specified; N/A: Not applicable; GR 1: "dobj", "prep_of", "nsubj", "pobj", "dep" and "partmod"; GR 2: "det", "neg", "advmod", "nn" and "amod"; GR 3: "conj_negcc" and "prep_without"; GR 4: "nsubjpass", "nsubj" and "nn"; GR 5: "appos", "dobj", "prep_of", "conj_and", "nn", "conj_or", "infmod", "prep", "xcomp", "amod" and "dep"; GR 6: "nn", "prep_such_as", "pobj", "nsubj", "conj_or", "nsubjpass", "dobj", "prep_of", "rcmod" and "dep".

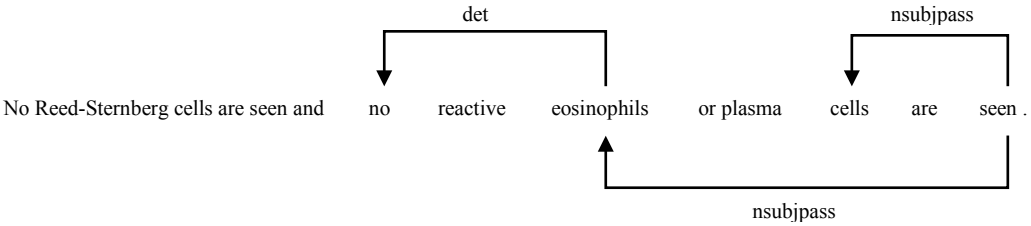
Pattern 3: Rule #5



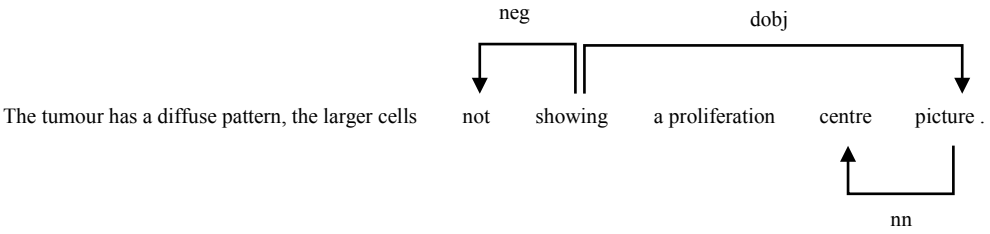
Pattern 4: Rule #4 + Rule #9



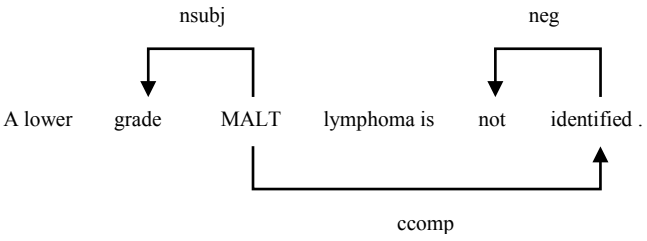
Pattern 5: Rule #4 + Rule #8 + Rule #13



Pattern 6: Rule #4 + Rule #10 + Rule #14



Pattern 7: Rule #4 + Rule #12 + Rule #15



This method consists of the following processes:

1. Find the headwords. For a medical entity, its headword is the head noun if it is a noun phrase; else, its headword is the last word of the entity. For a Li:Lexical Polarity Negative instance, the first word of the instance (e.g., "no" for "no evidence of", "not" for "not") is the headword.
2. Compute the path. The dependency path between the headwords of the Li:Lexical Polarity Negative instance and the medical entity is computed as described above. To reduce the complexity of the rules, only paths with length not larger than two are considered. Note that a special case exists: if the Li:Lexical Polarity Negative instance is "rather than" or "without", and the headword of the entity is a dependent with a grammatical relation of "conj_negcc" or "prep_without", the associated governor can be any word in the sentence, and the path is computed from the results between it and the headword of the entity.
3. Match the patterns. Evaluate whether the path matches the patterns as described above. If the length of the path is zero and it matches one of the patterns in Patterns 1 to 3, the entity is negated; if the length of the path is one and it matches Pattern 4, the entity is negated; if the length of the path is two and it matches one of the patterns in Patterns 5 to 7, the entity is negated.

Machine Learning-based Approach

Uzuner et al had argued that in the Statistical Assertion Classifier (StAC), contextual features could capture the information necessary for assertion classification, and syntactic information could make some contribution as well (Uzuner et al., 2009). Similar to their work, a machine learning-based approach applying a support vector machine (SVM) was built in this study for negation detection. However, the approach was different from their work, employing a pair-wise method instead, as Li:Lexical Polarity Negative instances were annotated in the training data, besides medical entities. Specifically, an instance of Li:Lexical Polarity Negative (the first concept) and a medical entity (the second concept) were paired and then passed into a SVM classifier to classify the negation relationships between them. Positive pairs were created for each pair with negation in a sentence, and negative pairs were created for each pair without negation in a sentence.

Besides contextual and syntax features, the SVM classifier was armed with other features, including semantic features, lexical features, grammatical features and positional features.

Contextual features: Four-word contextual window of each concept.

The contextual window size was determined by preliminary experiments on the training data. The optimal size found was four.

Grammatical features: Headwords of each concept.

Headwords of each concept are determined as described above. These features are general representatives for the concepts even if they consist of different lexicons. For example, the headword of the first concept can generalise two different Li:Lexical Polarity Negative instances "no definite evidence of" and "no convincing evidence of" to the same feature value: "no"; the headword of the second concept can yield the same feature value "lymphoma" for two Diagnosis entities "large B-cell lymphoma" and "classical Hodgkin lymphoma".

Syntax features: a) The grammatical relations of the shortest dependency path between the headwords of two concepts; b) The length of the shortest dependency path between the headwords of two concepts; c) Part-of -speech (POS) tags of the tokens between the two concepts.

The shortest dependency path between the headwords of two concepts can be computed as follows: First, compute the path between the root and the headword of each concept (path to root); compare the path to root for each concept, and the ones with least common nodes are selected from others; the selected paths to root are merged together into the shortest dependency path between two concepts.

In the syntax-based approach, grammatical relations are used to extract rules to construct negation patterns; with a), the classifier can automatically learn from the grammatical relations prepared for the model, and predict the unseen data based on probability.

For b), since most of the lengths of the shortest dependency paths for the positive pairs in the training data are not larger than two, assign the value for the feature as "C1" if the length is zero or one; "C2" if it is two; "F" if it is larger than two; "O" if the shortest dependency path cannot be found.

Semantic features: Annotation types of each concept.

Lexical features: a) Words inside each concept; b) Lowercase of words inside each concept; c) Tokens between the two concepts.

Tokens between the two concepts may contain termination cues; this feature provides an opportunity for the classifier to learn from these cues.

Positional features: a) Token distance between the two concepts; b) The order of appearance for the two concepts.

For a), as the average token distance for all positive pairs in the training data is smaller than one, so assign the feature value as "C" if the token distance between the two concepts is not larger than one; else, assign the feature value as "F".

For b), if the second concept is preceding the first concept, assign the feature value as "P"; else, assign the feature value as "S". As indicated in the lexicon-based approach, there are some patterns for the positions of particular trigger terms relative to a medical entity.

From the above features, it can be seen that several of them are adapted from the similar ideas in the two rule-based approaches. This is motivated by Patrick et al's work that converted a baseline rule-based method to a statistical approach based on the same idea for assertion classification, which produced better performance (Patrick et al., 2011).

Pseudo-negation Detection

A special module was implemented with regular expressions to handle pseudo-negations in sentences. They are triggered by some pseudo-negation phrases (e.g., "not possible", "not likely", "to exclude"). If a match is found, the related positive output from the negation detection module will be ruled out.

Evaluation Methods

All approaches were evaluated using single train-test cycles. The toolkit used for applying SVM in the machine learning-based approach is LIBSVM (Chang and Lin, 2011). Ten-fold cross-validation experiments were also carried out for this approach on the training set, and each fold was stratified at a document level, and used the default configuration for most parameters, except that parameter "cost" was set to 100, and "gamma" was set to 0.025.

The performances of the three methods on the training set were measured by the standard Precision, Recall and F-score. To compare the results to the work of Mitchell et al (Mitchell et al., 2004), three metrics were adopted to measure the performances of the three methods on the test set: strict, lenient and average metrics (Douthat, 1998). Strict metrics only consider exact match of the system predictions and gold-standards when they have the same boundaries; lenient metrics also consider partial match when they have any overlap of the boundaries; average metrics are the mean of the two above metrics. They are computed by true positive (TP), false positive (FP), false negative (FN) and partial positive (PP) as follows:

Strict Precision (SP) = TP/(TP+FP+½PP)
Strict Recall (SR) = TP/(TP+FN+½PP)
*Strict F-score (SF) = 2*SP*SR/(SP+SR)*
Lenient Precision (LP) = (TP+½PP)/(TP+FP+½PP)
Lenient Recall (LR) = (TP+½PP)/(TP+FN+½PP)
*Lenient F-score (LF) = 2*LP*LR/(LP+LR)*
Average Precision (AP) = (SP + LP)/2
Average Precision (AR) = (SR + LR)/2
*Average F-score (AF) = 2*AP*AR/(AP+AR)*

Results

Table 6.2 shows the contribution of features to the machine learning-based approach on the training set.

The section contexts were classified to four categories: Macroscopic (referring to "MACROSCOPIC"), Microscopic (referring to "MICROSCOPIC"), Summary (including "SUMMARY" and "SUPPLEMENTARY SUMMARY"), and Other (composed of "CLINICAL HISTORY", "SPECIMEN", "FROZEN SECTION REPORT" and "SUPPLEMENTARY REPORT").

Model #	Feature	Precision	Recall	F-score
1	Words inside each concept + annotation types of each concept	84.78%	89.31%	86.98%
2	M1 + Four-word contextual window of each concept	83.64%	99.69%	90.96%
3	M2 + Lowercase of words inside each concept	84.04%	99.37%	91.07%
4	M3 + Token distance between the two concepts	89.66%	98.11%	93.69%
5	M4 + Headwords of each concept	89.43%	98.43%	93.71%
6	M5 + The order of appearance for the two concepts	90.67%	97.80%	94.10%
7	M6 + Tokens between the two concepts	92.60%	98.43%	95.43%
8	M7 + The grammatical relations of the shortest dependency path between the headwords of two concepts	93.18%	98.74%	95.88%
9	M8 + The length of the shortest dependency path between the headwords of two concepts	94.28%	98.43%	96.31%
10	M9 + POS tags of the tokens between the two concepts	95.41%	98.11%	96.74%

Table 6.2 Contribution of features to the machine learning-based approach on the lymphoma *training set* (evaluated with the strict metric).

Method	Section	Number	Precision	Recall	F-score
Lexicon-based approach	Macroscopic	4	100.00%	100.00%	100.00%
	Microscopic	242	97.57%	99.59%	98.57%
	Summary	54	100.00%	98.15%	99.07%
	Other	18	94.74%	100.00%	97.30%
	Overall	318	97.83%	99.37%	98.60%
Syntax-based approach	Macroscopic	4	100.00%	100.00%	100.00%
	Microscopic	242	99.59%	99.17%	99.38%
	Summary	54	98.11%	96.30%	97.20%
	Other	18	94.74%	100.00%	97.30%
	Overall	318	**99.05%**	**98.74%**	**98.90%**
Machine learning-based approach (10-fold cross-validation)	Macroscopic	4	100.00%	100.00%	100.00%
	Microscopic	242	95.97%	98.35%	97.14%
	Summary	54	100.00%	100.00%	100.00%
	Other	18	76.19%	88.89%	82.05%
	Overall	318	95.41%	98.11%	96.74%

Table 6.3 Performance metrics across report sections for negation detection for the three methods on the lymphoma *training set* (evaluated with the strict metric).

Table 6.3 displays the results for the three methods by section and overall on the lymphoma training set, using the strict metric. From Table 6.3, the highest F-score of 98.90% was achieved by the syntax-based approach. The best performance for most sections varied from each method: the highest F-score of 99.38% for the "Microscopic" section was attained by the syntax-based approach; the highest F-

score of 100.00% for the "Summary" section was attained by the machine learning-based approach; the lexicon-based approach and syntax-based approach had the same performance for other sections with 97.30% F-score; no difference occurred on the performances of "Macroscopic" section with the three methods, all with 100.00% F-score.

The system performances across sections for the three methods on the test set are presented in Table 6.4. The overall micro-averaged F-scores decreased by 14.18 to 20.28%. The machine learning-based approach performed well on the Microscopic section, with 84.85% F-score; the syntax-based approach performed well within the Summary and other sections, with 100% and 61.54% F-scores respectively.

As shown in Table 6.5, the majority of errors on the test set can be attributed to MER. Note that the errors were categorized to incorrect MER in priority, hence it cannot rule out the possibility that some errors might be actually be due to mistakes from both MER and negation detection.

Discussion

From Table 6.2, the baseline model achieved 86.98% F-score. Contextual feature set yielded the biggest gain, and improved the model by 3.98%. Meanwhile, moderate improvements are contributed by the positional feature *token distance between the two concepts* and the lexical feature *tokens between the two concepts* (with 2.62% and 1.33% gain respectively). Three syntax features and the positional feature *the order of appearance for the two concepts* yielded some gains ranging from 0.39% to 0.45% respectively. Minimal improvements were made by adding the lexical feature *lowercase of words inside each concept* and *headwords of each concept* with 0.11% and 0.02%.

The results are consistent with those from the evaluation on StAC (Uzuner et al., 2009), which indicated that the contextual features were very effective at improving the model. Syntax features could correct some false positives when a Li:Lexical Polarity Negative instance occurs within the four-word window but does not in fact negate a medical entity. For example, in the sentence "The colonic wall is not involved by lymphoma.", "involved" is negated by "not" but not "lymphoma". Note that the syntax features in this study are different from those in StAC:

- StAC utilised the output of the Link Grammar Parser (LGP) (Sleator and Temperley, 1991), while Stanford parser was used to generate the dependency parse.
- StAC focused on the verbs preceding and succeeding the entity, while this was not emphasized in this study, as the training data for StAC were discharge summaries, which were pathology reports in this study; verbs were less frequently appearing in pathology reports, and they would be omitted in some cases, e.g., in this sentence "2. Lymph node, in transit sentinel - no evidence of malignancy."
- Unlike StAC, a few features that motivated by the rule-based approach (e.g., the positional feature *token distance between the two concepts* and the lexical feature *tokens between the two concepts*) were also adopted, which yield more gains than the syntax feature set.

Method	Section	TP	FP	FN	PP	SP	SR	SF	LP	LR	LF	AP	AR	AF
Lexicon-based approach	Macroscopic	1	0	1	0	100.00%	50.00%	66.67%	100.00%	50.00%	66.67%	100.00%	50.00%	66.67%
	Microscopic	53	3	15	8	88.33%	73.61%	80.30%	95.00%	79.17%	86.36%	91.67%	76.39%	83.33%
	Summary	10	1	0	0	90.91%	100.00%	95.24%	90.91%	100.00%	95.24%	90.91%	100.00%	95.24%
	Other	4	2	4	0	66.67%	50.00%	57.14%	66.67%	50.00%	57.14%	66.67%	50.00%	57.14%
	Overall	68	6	20	8	87.18%	73.91%	80.00%	92.31%	78.26%	84.71%	89.74%	76.09%	**82.35%**
Syntax-based approach	Macroscopic	1	0	1	0	100.00%	50.00%	66.67%	100.00%	50.00%	66.67%	100.00%	50.00%	66.67%
	Microscopic	46	2	23	7	90.20%	63.89%	74.80%	96.08%	68.06%	79.67%	93.14%	65.97%	77.24%
	Summary	10	0	0	0	100.00%	100.00%	100.00%	100.00%	100.00%	100.00%	100.00%	100.00%	100.00%
	Other	4	1	4	0	80.00%	50.00%	61.54%	80.00%	50.00%	61.54%	80.00%	50.00%	61.54%
	Overall	61	3	28	7	91.04%	66.30%	76.73%	95.52%	69.57%	80.50%	93.28%	67.93%	**78.62%**
Machine learning-based approach	Macroscopic	1	0	1	0	100.00%	50.00%	66.67%	100.00%	50.00%	66.67%	100.00%	50.00%	66.67%
	Microscopic	54	2	14	8	90.00%	75.00%	81.82%	96.67%	80.56%	87.88%	93.33%	77.78%	84.85%
	Summary	10	3	0	0	76.92%	100.00%	86.96%	76.92%	100.00%	86.96%	76.92%	100.00%	86.96%
	Other	4	2	4	0	66.67%	50.00%	57.14%	66.67%	50.00%	57.14%	66.67%	50.00%	57.14%
	Overall	69	7	19	8	86.25%	75.00%	80.23%	91.25%	79.35%	84.88%	88.75%	77.17%	**82.56**%

Table 6.4 *Test set* performance metrics across report sections for the three evaluation methods (Strict, Lenient and Average) for the lymphoma corpus.

- StAC restricted ±2 link window on the features, while there was no such restriction in this study.

It can be seen from Table 6.4 that the best performance on the lymphoma test set was on the Summary section. This is possibly because of the better grammatical structure in the Summary section. This section was present in almost every report, often in a well-structured and formalized format. This is also the main reason for the best performance in this section with the syntax-based approach (100% F-score), as the syntax-based approach relied on the output from the parser and the performance of the parser was hindered by the linguistic constructions of the input. Accurate parsing output can be generated due to the simple linguistic constructions in this section. The best performance for the Microscopic section was obtained by the machine learning-based approach, which indicates one of the advantages of using this method when the amount of training samples are sufficient: the training examples in the Microscopic section had the largest proportion (about 76%), that allowed the machine learner to learn more effectively, which led to more accurate prediction on the test set with the lowest drop from the training set result, compared to those in other sections.

Method	Total errors	Error from MER			Error from negation detection
		"Lexical Polarity Negative"	Other entity type	Both	
Lexicon-based approach	26	7	6	11	2
Syntax-based approach	31	6	6	11	8
Machine learning-based approach	26	7	6	10	3

Table 6.5 Summary of errors from medical entity recognition and negation detection on the lymphoma *test set*.

Table 6.5 shows that errors from negation detection directly accounted for about 8.7% to 25.8% of total errors depending on the method. The reasons for those errors with the syntax-based approach are mainly due to the poor parsing results from the parsers, where the rules do not work as expected if parse trees are problematic. For example, given the input "No vasculitis ["En:Coexistent Pathology"] with fibrinoid necrosis or leucocytoclastic debris, granulomas ["Re:Tissue Reaction"] or necrosis ["Re:Tissue Reaction"] are seen", the parser attached the noun phase "leucocytoclastic debris" to the incorrect location in the parse tree (see Figure 6.5). Consequently, the entities "granulomas" and "necrosis" could not be identified as negated concepts. This suggests that such errors can be amended by using a domain-specific parser or a parser trained with medical corpora to improve the parsing performance.

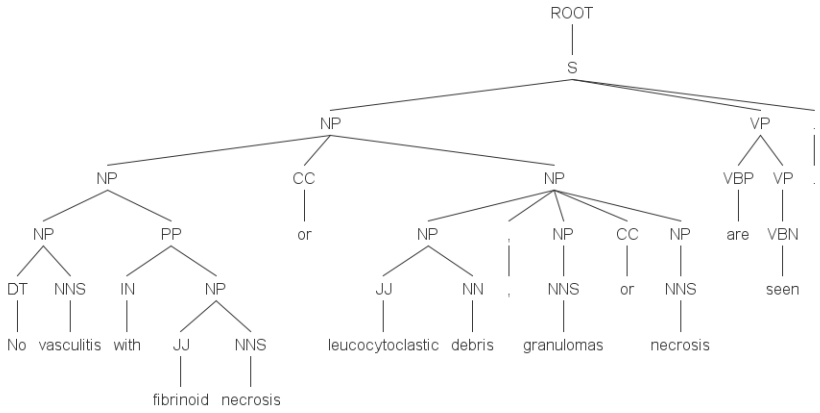


Figure 6.5 The incorrect parse tree of the text example as generated by the Stanford parser.

It is known that medical narratives often prefer to be presented in compact expressions, and therefore noun and prepositional phrases are more frequently used in a long sentence rather than complex verb structures or short sentences as in the general domain, and in some cases, they may be irregular grammatical structures. Errors from the other two methods are probably because of the abnormal structures presented in the notes. Here is an example:

> NO EVIDENCE OF ["Li:Lexical Polarity Negative"] METASTATIC MELANOMA IN 11 LYMPH NODES,
> SMALL LYMPHOCYTIC NON-HODGKINS LYMPHOMA ["Sy:Diagnosis"],
> Gene rearrangement studies pending, Please see report.

The instance of "SMALL LYMPHOCYTIC NON-HODGKINS LYMPHOMA" should not be tagged as negated in the above. However, due to the irregular grammatical structure of the sentence, it is too difficult for the negation detection module to generate a correct output. In fact, in regular grammatical structures, this sentence can be divided into four sentences: "NO EVIDENCE OF METASTATIC MELANOMA IN 11 LYMPH NODES.", "SMALL LYMPHOCYTIC NON-HODGKINS LYMPHOMA.", "Gene rearrangement studies pending." and "Please see report."; or some conjunctions like "and", "but" could be supplemented in the sentence for grammatical correction.

Similar work has been done by Mitchell et al on detecting and annotating UMLS concepts as well as annotating negation based on the NegEx algorithm (Mitchell et al., 2004). They reported that the overall precision and recall under average conditions were about 64% and 55% respectively, which are about 25.7% and 21.1% lower than the results obtained from the lexicon-based approach in this study. This is probably because the lexicon-based approach modified from NegEx has been adapted for this corpus. For example, the negation and pseudo-negation phrases were extracted from the training data. Except for some terms, particular punctuation for each section was also introduced as termination cues to improve the precision. Nevertheless, the comparisons are also subject to some notable differences between the Mitchell et al's work and this study:

1. The materials for evaluation. The previous work used surgical pathology reports; this study selected pathology reports of a specific disease (lymphoma).

2. The development of gold-standards. The Mitchell work used a modified Delphi technique to achieve consensus among the panel and pathologists participating in the manual annotation. This study only used a single pass panel and linguists were involved for manual annotation.
3. The test sample size. The previous work had a larger test sample size with 311 entities; our study only had 96 entities for testing.
4. The semantic types. The semantic types chosen by the previous work were up to 35 from five semantic categories relevant to surgical pathology reports in UMLS; this study defined 11 specific semantic types based on the protocol.
5. The methods for concept recognition. The previous work performed dictionary look-up to the NLM Knowledge Source Server, and matched phrases against the UMLS, therefore each extracted entity was a UMLS concept. This study used a supervised machine learning-based approach to extract medical entities, hence the extracted entities might not be UMLS concepts.
6. The results for the Comment section. The results for most sections reported in this study correspond to the previous work, e.g., Macroscopic *vs.* Gross Description, Microscopic *vs.* Microscopic Description, and Summary *vs.* Final Diagnosis. However, the results for the Comment section were not presented in this study, since all the contents in the Comment section are required to populate the structured templates instead of each entity type as in other parts of the project, and no entity was annotated in this section according to the annotation schema.

Tables 6.3 and 6.4 show that the syntax-based approach had the best overall performance on the training set but the poorest on the test set, in contrast to the machine learning-based approach. This is likely to be because: in the test set, there were unforeseen structures in the sentences which the rules or patterns designed for syntax-based approach cannot handle properly though they worked well on the training set. The model generated by the machine learning-based approach predicted the test data with features not only captured from the training set but also from the test set, which made it less vulnerable to the unforeseen structures. Nevertheless, the performance of the lexicon-based approach was more stable: on the training set, it only lagged behind the syntax-based approach by 0.3% F-score. On the test set, there was also a very small gap between the overall F-score and that of the machine learning-based approach of 0.21%.

The run time for applying each method is distinct: least for the lexicon-based approach, most for the machine learning-based approach and between the two for the syntax-based approach (see Table 6.6). Moreover, there is a prominent gap between the lexicon-based approach and the other two methods. It seems that the length of the sentence is a crucial factor in effecting the run time for applying the lexicon-based approach, while the number of medical entities in the sentence is more likely to effect the run time for the other two methods. The following two sentences were used for the comparisons shown in Table 6.6.

Input 1: No ["Li:Lexical Polarity Negative"] necrosis ["Re:Tissue Reaction"] or Reed-Sternberg cells ["De:Cytomorphology"] seen.
Input 2: Intraepithelial lymphocytes appear generally increased, however classic lymphoepithelial lesions ["De:Architecture"] are not ["Li:Lexical Polarity Negative"] identified ["Li:Lexical Polarity Positive"] (cytokeratin, CD3 and CD20 immunohistochemical stains).

Method	Input 1	Input 2
Lexicon-based approach	0.0007s	0.0250s
Syntax-based approach	12.35s	8.44s
Machine learning-based approach	14.74s	13.90s

Table 6.6 Run time for applying each method to the examples. s: seconds.

Considering the above factors and the simplicity of the lexicon-based approach to implement and tune to another corpus, it was adopted for the other two corpora to detect negation.

6.2.2 Uncertainty Detection

As described previously, uncertainty detection is more challenging than negation detection, for several reasons:

- The phrases or keywords indicating uncertainty are vaguer. For example, "unlikely" appears to be the antonym of "likely" and can be a candidate for a negation cue; in fact, it could be reclassified to uncertainty, as it indicates the lower possibility than "likely".
- Punctuation suggests uncertainty. Not only certain words, but also punctuation can be considered as an uncertainty cue. For instance, question mark(s) "?" and "??" frequently appear in "CLINICAL HISTORY" and "SPECIMEN".
- Uncertainty can be expressed explicitly or implicitly. An explicit expression like "suspicions for Hodgkin's Lymphoma" is easy to understand, while an implicit expression such as "The main differential diagnosis is between a T-cell rich large B-cell non-Hodgkin lymphoma and a nodular lymphocyte predominant Hodgkin lymphoma" requires inference from the domain knowledge to comprehend.

It is thus preferable to employ a rule-based method to resolve this issue. As natural language parsers like the Stanford parser cannot generate correct output when parsing some cases if it contains punctuation as an uncertainty cue, the rule-based method tends to be lexicon-based instead of syntax-based.

The lexicon-based approach designed to detect uncertainty is similar to that for negation detection. It also defines trigger terms and termination cues. The trigger terms refer to Li:Lexical Modality instances in the training data, and are categorized into three groups depending on their position relative to a medical entity: Group 1 – only preceding a medical entity; Group 2 – only succeeding a medical entity; Group 3 – any of the above positions. Termination cues include terms that suggest

cause, experiencer or transition, and particular punctuation. Table 6.7 shows some examples. Trigger terms in Group 1 and Group 3 outnumber those in Group 2, suggesting that the uncertainty phrases usually appear before the entity they assert in the notes; the positions of terms in Group 3 are more flexible, which makes determination for their scope more difficult. If there is more than one trigger term, the closest one to the entity is selected as the best candidate.

Type of term/cue	Sub-category	Example
Trigger terms	Group 1	probable, highly suspicious for, highly suspicious of, suspicious of, suspicious for, possibly, probably, perhaps, slightly favour, wondered about, wonder about, fit best for, more suggestive of, more in favour of, more in keeping with, possibilities, definite, definitive, possibility
	Group 2	cannot be excluded
	Group 3	certain, certainly, maybe, suspicious, whether, less likely, unlikely, convincing, likely, may, more likely, most likely, probable, possible, difficult to identify, reluctant, uncertain, ? | ??
Termination cues	Termination term	but, which, though, although, however, so, from
	Termination punctuation	"CLINICAL HISTORY" and "SPECIMEN": , | ; | (|) | ->
		Other section contexts: (|)

Table 6.7 Trigger terms and termination cues for uncertainty detection on the lymphoma corpus. Note: word examples are separated by comma ","; punctuation examples are separated by pipe "|".

The determination of the scope is distinguishable from that for negation detection. First, the text span between the trigger term and the entity is checked whether it contains any termination cue; then the following rules are applied to it:

If the token distance from the entity to the trigger term is not larger than four, the entity is assigned in the scope;

Else, it will be validated against these patterns:

The trigger term has a comparative modifier, such as "more suggestive of" and "more in favour of", and "than" appears in the text span;

The trigger term succeeds the entity and there is a conjunction in the text span.

If it has one of these patterns, the entity is included in the scope.

The lexicon-based approach can work well for explicit expressions, but cannot handle the implicit ones properly. An additional module that resembles the one for pseudo-negation detection was designed to cope with this problem. It was implemented with regular expressions to represent some patterns, which are described as follows:

Pattern 1: <n word> < diagnosis> <n word> <between/includes> <n word> <entity 1> <n word> <and> <n word> < entity 2> <n word>.

Pattern 2: <n word> < diagnosis> <n word> < includes> <n word> <entity 1> <n word> <but> < n word >.

Pattern 3: <n word> < entity 1> <n word> < however> < n word> < against this> < n word >.

Pattern 4: <n word> <differential diagnosis> < n word> < entity 1> < n word >.

Pattern 5: <n word> <more> < n word> <than> < n word> < entity 1> < n word >.

Pattern 6: <n word> <favour/favours> <n word> < entity 1> <n word> <over> <n word> < entity 2> <n word>.

Pattern 7: <n word> <either/ any Li:Lexical Polarity Positive instance > <n word> < entity 1> <n word> <or> <n word> < entity 2> <n word>.

Pattern 8: <n word> < entity 1> <n word> <favoured over> <n word> < entity 2> <n word>.

Where n stands for the number of tokens, without any restriction and can be 0. The entities in each pattern are assigned uncertainty. The pseudo-negations mentioned above would also be considered as candidates of the results.

Unlike the representation for negation that is quite clear that "absent" or "no" can be used to represent it, the ambiguous use of "possible", which was suggested in the assertion annotation guidelines of the 2010 i2b2/VA Challenge (i2b2, 2010a), cannot reveal the degree of certainty. Similar to MedLEE that used different certainty modifiers to indicate the degree of certainty, a standard dictionary was used to map each trigger term to four categories: "cannot exclude", "possible", "probable" and "definite", which stand for low certainty, low to moderate certainty, moderate to high certainty and high certainty, respectively. For example, "cannot be excluded" is mapped to "cannot exclude", "?" is standardized to "possible", "more suggestive of" is represented with "probable" and "certainly" is replaced with "definite".

The results for uncertainty detection on the lymphoma training set and test set are shown in Table 6.8. The uncertainty detection module attained very good performance on the training set with micro-averaged F-score over 97%, but dropped dramatically to about 67% on the test set. Two categories "cannot exclude" and "definite" could not be assessed fairly, due to lack of training and test examples.

Uncertainty type	Training set				Test set			
	Number	Precision	Recall	F-score	Number	Precision	Recall	F-score
cannot exclude	1	50.00%	100.00%	66.67%	0	0.00%	0.00%	0.00%
definite	2	100.00%	100.00%	100.00%	0	0.00%	0.00%	0.00%
possible	264	98.48%	98.48%	98.48%	62	77.55%	61.29%	68.47%
probable	30	96.55%	93.33%	94.92%	9	80.00%	44.44%	57.14%
Overall	297	97.98%	97.65%	97.81%	71	77.78%	59.15%	67.20%

Table 6.8 Results for uncertainty detection on the lymphoma training set and test set.

Error analysis shows that the errors on the training set can be categorized to:

- The fixed four-token window size may omit some entities far from the trigger term. Although in most cases, the entity asserted by the trigger term locates very close to the term, there are still some cases like "A high grade ["Sy:WHO Grade"] lymphoma of follicle centre cell

origin ["Sy:Diagnosis"] is also possible ["Li:Lexical Modality"] but considered less likely ["Li:Lexical Modality"]." where the entity "high grade" is situated distant from the trigger term "possible".

- The closest trigger term is not always the best candidate that asserts the entity. For example, in this sentence:

  Although the subtype is difficult to discern a nodular sclerosis Hodgkin's lymphoma ["Sy:Diagnosis"] or perhaps ["Li:Lexical Modality"] a lymphocyte depleted type ["Sy:Diagnosis Subtype"] would be the two most likely ["Li:Lexical Modality"].

  the closest trigger term to the entity "lymphocyte depleted type" is "perhaps", but the best candidate is "most likely".

There were 36 errors identified in the results on the test set, where errors in the MER accounted for most of them. Specifically, 8 were from incorrect recognition of Li:Lexical Modality instances, 16 were caused by misclassification of medical entities. The false positive recognition of both Li:Lexical Modality instances and medical entities led to 5 errors, and 7 errors owing to mistakes in the module. The main reason for the mistakes in the module was the incompetence of the uncertainty pattern. For instance, Pattern 7 relied on the appearance of the lexicon "either" or any Li:Lexical Polarity Positive instance, while in this sentence:

> The appearance of the lymph nodes ["De:Tissue Source"] and the skin ["De:Tissue Source"] infiltrate is of small lymphocytic lymphoma ["Sy:Diagnosis"] or involvement by chronic lymphocytic leukaemia ["Sy:Diagnosis"].

where "suggestive" was omitted by the author between "is" and "of" so that the uncertainty for the entities "small lymphocytic lymphoma" and "involvement by chronic lymphocytic leukaemia" cannot be detected.

The ambiguous usage of slash "/" can also cause some problems. Slash "/" can function as "separator", e.g., in this sentence:

> (L) para-aortic lymph node ["De:Topography"] / core bx ["De:Specimen Type"] - MALIGNANT LYMPHOMA ["Sy:Diagnosis"].

It can couple two entities, e.g., small lymphocytic lymphoma (SLL) and chronic lymphocytic leukaemia (CLL) are usually coupled together and expressed as: "SLL/CLL". It can stand for "per" in a measurement unit, such as "14/mm2". In this sentence "Low-grade ["Sy:WHO Grade"] follicular lymphoma ["Sy:Diagnosis"] / small cleaved lymphoma ["Sy:Diagnosis"].", "/" stands for "or". However, due to the ambiguity "/" may bring, it was not considered in the patterns.

The higher error rate of uncertainty detection compared to negation detection with the same method on the test set did suggest that uncertainty detection is more difficult to handle than negation detection. This is probably because:

- The lexical variants of trigger terms for uncertainty outnumbered that for negation. Additionally, many of them had an ad-hoc position that made their relationships with other lexical items more difficult to detect.
- The expressions of uncertainty were more diverse, hence predefined rules or patterns failed more often to correctly process.
- The sub-classification of uncertainty into four categories also increased the difficulty of resolving this issue.

Nevertheless, the good performance achieved by this method on the training set, and the simplicity it manifests, implies that it can be applied to the other two corpora as well.

6.3 Negation and Uncertainty Detection in the Other Two Corpora

As discussed above, the methods for negation and uncertainty detection on the other two corpora were also lexicon-based, but because of the idiosyncrasies or characterises of the corpora and the associated annotation schemas, the approach needs to be fine-tuned for each corpus.

6.3.1 Melanoma Corpus

The annotation schema for the melanoma corpus is similar to that for the lymphoma corpus, thus the main adjustment of the method for negation detection was to modify the entries of the trigger terms and termination cues, which are displayed in Table 6.9. Several medical entity types were involved: De:Ulceration, In:Vascular/Lymphatic, In:Neurotropism, En:Satellites, En:Associated naevus (type), Re:Desmoplasia, Sy:Diagnosis, Sy:Subtype, De:Cell Growth Pattern, En:Lesion (other), Re:TILs, and Sy:Regression.

Type of term/cue	Sub-category	Example
Trigger term	Group 1	no, nor, exclude, lack, rule out, non, neither, lacks, failure, precludes, obscures, rather than, obscure, without
	Group 2	not at all, nil, absent
	Group 3	not, unremarkable
Termination cue	Termination term	but, which, though, although, with, however, there is, so, it is, due to, this, and the, in the
	Termination punctuation	“CLINICAL HISTORY” and “SPECIMEN”: , | ; | (|)
		“DIAGNOSIS”:) | \n

Table 6.9 Trigger terms and termination cues for negation detection on the melanoma corpus. Note: word examples are separated by comma “,”; punctuation examples are separated by pipe “|”. “\n”: newline character.

Another adjustment was the utilisation of the Li:Lexical Polarity Positive instances to derive the negation rules. A detailed analysis on the corpus shows that a Li:Lexical Polarity Negative instance frequently occurs in the company of a Li:Lexical Polarity Positive instance, thus it is presumed that the utilisation of the Li:Lexical Polarity Positive instances can facilitate the detection.

The primary idea was similar to that presented in Patrick et al's work that the assertion of a medical entity was usually determined by the closest specific lexicon (e.g., "absent" lexicons, "possible" lexicons) (Patrick et al., 2011). The specific lexicon data were referred to Li:Lexical Polarity Negative and Li:Lexical Polarity Positive instances ("negative" and "positive" instances) in this study. For example, in this sentence "Mitoses are infrequent ["De:Dermal Mitoses"] and vascular space invasion ["In:Vascular/Lymphatic"] is not ["Li:Lexical Polarity Negative"] identified ["Li:Lexical Polarity Positive"].", "not" is the closest lexical polarity instance to the entity "vascular space invasion", thus the assertion of the entity is "absent".

The initial processes resemble steps 1 and 2 in the lexicon-based approach for the lymphoma corpus. The additional procedures are:

- Firstly, compute the token distances between an entity and each lexical polarity instance in the sentence, and sort them in ascending order.
- In most cases, let "negative" instances take precedence over "positive" ones: if a "negative" instance exists in the sentence and its position is valid, it can assert the entity no matter if it is the closest lexical polarity instance. For example, in this sentence:

  > There is no ["Li:Lexical Polarity Negative"] evidence ["Li:Lexical Polarity Positive"] of regression ["Sy:Regression"].

  although "no" locates farther than "evidence" to the entity "regression", the entity is asserted by "no". Likewise, "no" also negates the entity "perineural invasion" in the sentence:

  > There is no ["Li:Lexical Polarity Negative"] lymphovascular or perineural invasion ["In:Vascular/Lymphatic"] identified ["Li:Lexical Polarity Positive"].

  If multiple "negative" instances occur, the closest one accounts for the assertion, e.g., in this sentence:

  > No ["Li:Lexical Polarity Negative"] ulceration ["De:Ulceration"] is noted ["Li:Lexical Polarity Positive"] and no ["Li:Lexical Polarity Negative"] vascular invasion ["In:Vascular/Lymphatic"] is seen.

  the entity "ulceration" is negated by the first "negative" instance "No" and "vascular invasion" is negated by the second one "no".

However, there are some exceptions:

- If the negative" instance in Group 1 or Group 3, and the closest "positive" instance is "suggestive of", "evident", or "appears", the entity is asserted by the closest "positive" instance. An example is presented below:

  > Although there is no ["Li:Lexical Polarity Negative"] obvious desmoplasia ["Re:Desmoplasia"] , this growth pattern is suggestive of ["Li:Lexical Polarity Positive"] a desmoplastic melanoma ["Sy:Diagnosis"].

- If the "negative" instance is "non", it can only assert the adjacent succeeding word, which is "ulcerating" in this example:

The melanoma ["Sy:Diagnosis"] measures 3mm across ["De:Size"] and is non ["Li:Lexical Polarity Negative"] -ulcerating ["De:Ulceration"].

The pseudo-negation detection module is also similar to that for the lymphoma corpus, and focused on detecting pseudo-negation phrases such as "probably not" and "to rule out".

Uncertainty detection for the melanoma corpus also adopted the same module as for the lymphoma corpus, excluding the utilization of regular expressions to capture uncertainty patterns. A detailed manual analysis on the corpus shows that most of the expressions of uncertainty are quite explicit and uncertainty patterns prepared for the lymphoma corpus cannot fit the melanoma corpus. Table 6.10 presents the adapted trigger terms and termination cues for the corpus. Compared to the trigger terms for uncertainty detection in the lymphoma corpus, there are more terms identified in this corpus, many of which can be categorized to Group 2. In addition, the positions of question mark (?) and question marks (??) are more stable, which only occur before a medical entity.

Type of term/cue	Sub-category	Example
Trigger term	Group 1	probable, possibly, possible, probably, definite, convincing, possibility, suspicious for, suspicious of, if, whether, susp for, raise the possibility, most probably, cannot exclude, cannot determine ? | ??
	Group 2	cannot be excluded, cannot be completely ruled out, could not be entirely excluded, cannot be totally excluded, cannot be guaranteed, cannot be entirely excluded, cannot be determined, cannot be confidently excluded
	Group 3	may, likely, suspicious, presumably, borderline, unequivocally, uncertain, query, only just marginally, not absolutely certain, most likely, maybe, alternatively
Termination cue	Termination term	but, which, though, although, however, so, with, and
	Termination punctuation	"CLINICAL HISTORY" and "SPECIMEN": , | ; | (|)
		"DIAGNOSIS": (|) | \n
		Other section contexts: ,

Table 6.10 Trigger terms and termination cues for uncertainty detection in the melanoma corpus. Note: word examples are separated by comma ","; punctuation examples are separated by pipe "|". "\n": newline character.

6.3.2 Colorectal Cancer Corpus

Besides negation and uncertainty as in the other two corpora, there is another assertion that needs to be discriminated in this corpus: inapplicability, which is similar to the "cannot evaluate" category in MedLEE. As the annotation schema for the colorectal cancer corpus is different from those of the other two corpora, where no Linguistic category is available, a specific rule-based module named Negation/Uncertainty/Inapplicability (NUI) Detector was introduced to resolve this issue.

The input could be a medical entity or the sentence where the entity is located in a document. The related medical entity types are: De:Perforation, Ex:Serosal Involvement, In:Perineural Invasion, De:Tissue Banking, De:Specimen Images, Re:TILS and Peritumoural Lymphocytes, En:Distant Spread or Metastases and In:Venous and Small Vessel Invasion.

Type of term	Sub-category	Example
Trigger term	Negation	Group 1: without, nor, none, no, benign, negative, unremarkable, neither, clear of, short of, free of, free from, no evidence of , not sufficient for, not, negative for, clearance, rather than, non-involved, without evidence of, spares, absence of. Group 2: not, uninvolved, absent, nil, benign, negative, unremarkable, tumour free, no, none, clear.
	Uncertainty	alternatively, uncertain, equivocal, maybe, query, whether, if, possibility, possible, possibly, may, seems, susp, presumably, suspicious, likely, convincing, probably, probable, suspicion, certain, unequivocally, definite, definitive, ? | ??
	Inapplicability	Not applicable: not applicable, n/a, na. Unknown: not known, nil known, unknown, cannot be assessed, not given, not supplied, not assessed.
Pseudo-trigger term	Pseudo-negation	not applicable, not known, nil known, not through, not given, not supplied, no special type, no special-type, not otherwise specified, not assessed.
Termination term	--	but, although, though, despite, however, identified, there is. For preceding scope: and. For succeeding scope: which, further.

Table 6.11 Three types of terms and examples for the colorectal cancer corpus. Note that some negation phrases (e.g., not, benign, negative) occur in both groups, suggesting that they can precede or succeed the scope.

Similar to ConText (Chapman et al., 2007b), the module also relied on three types of terms to yield the output: trigger terms, pseudo-trigger terms, and termination terms. Trigger terms, as the cues, included negation phrases, uncertainty phrases and inapplicability phrases. Through a combination of manual scanning and semi-automated learning, 28 negation phrases, 26 modality phrases and 10 inapplicability phrases were identified respectively. These negation phrases could be divided into two groups according to their positions to the scope (Group 1: preceding the scope, Group 2: succeeding the scope) and inapplicability phrases were classified to two categories (Not applicable and Unknown). Pseudo-trigger terms particularly referred to pseudo-negation phrases, which contained a negation phrase but did not indicate negation of a medical entity. The text span between the trigger term and the entity (if the input is an entity), or the start or end of the sentence (if the input is a sentence) was called the potential scope. If the input was an entity, then the potential scope was limited to the instance; else, the potential scope was extended to the whole sentence, unless a termination term occurred. A termination term like "but" could terminate the potential scope before the end of the sentence. The termination terms were augmented with additional phrases depending on the positions of the potential scope to the trigger term. A similar approach to the one suggested in NegExpander (Aronow et al., 1999), was used to determine the scope by detecting conjunctions like

"and", "or", and ",", instead of a fixed five-word window size employed in ConText. These terms with examples are shown in Table 6.11.

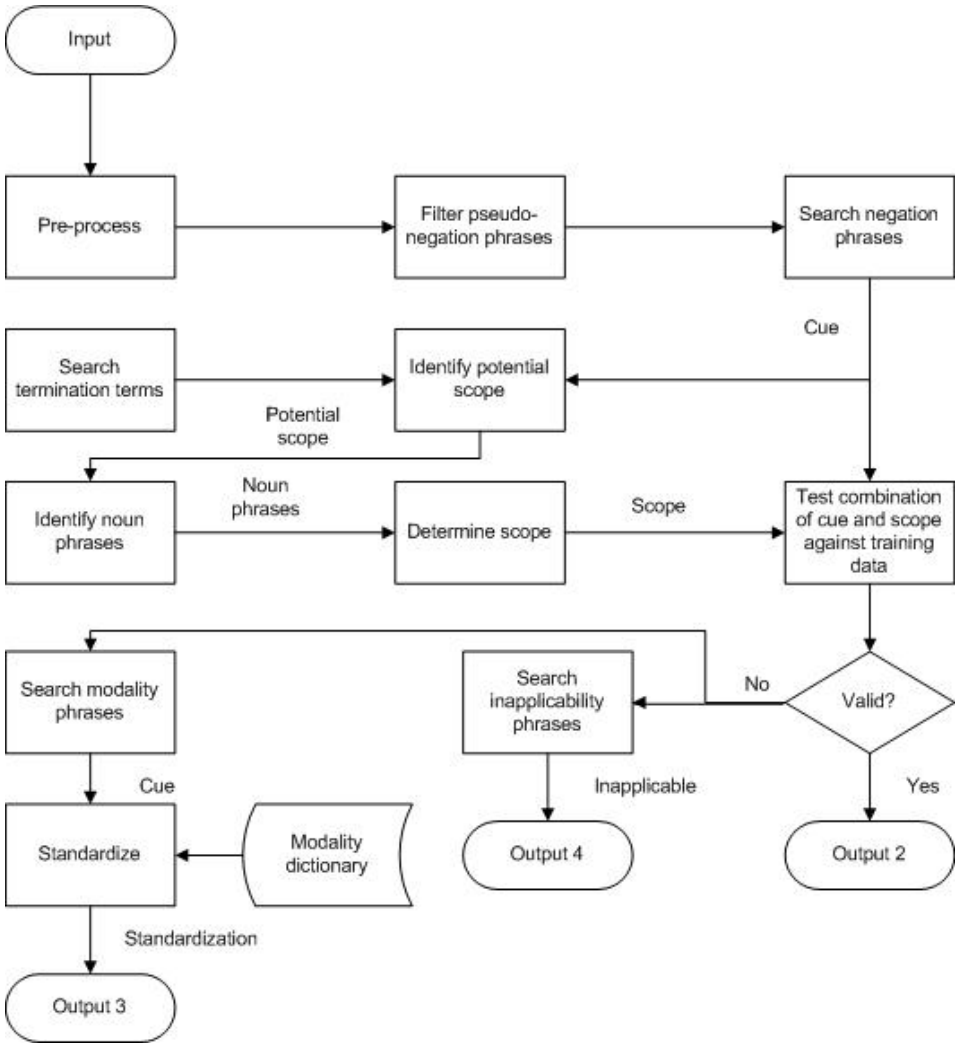


Figure 6.6 Workflow of Negation/Uncertainty/Inapplicability Detector.

The workflow of the module is illustrated in Figure 6.6, which includes the following processes:

1. Pre-process. The input is passed through the pre-processing engine to remove all punctuation and extra white spaces, and converted to lowercase.
2. Detect negation.
    1) Pseudo-negation phrases are filtered out from the input.
    2) Negation phrases are searched by a string match algorithm, and then the longest candidate is selected to be the cue. For example, in the sentence:

        "No evidence of vascular ["In:Venous and Small Vessel Invasion"], lymphatic ["In:Venous and Small Vessel Invasion"] or perineural ["In:Perineural Invasion"] invasion is seen.

    "no evidence of" is the cue rather than "no".

3) The GENIA tagger is used to identify noun phrases, adjective phrases and prepositional phrases in the potential scope. If there is no conjunction, then only the adjacent noun phrase proceeds the negation phrase (for Group 1 entries), and the adjacent noun phrase, adjectival phrase or prepositional phrase succeeds the negation phrase (for Group 2 entries) should be considered to be in the scope; else, the scope would be propagated to the conjunctive phrases as well.
4) To verify the validity, specific keywords and rules are applied to validate the cue and the scope.

   For example, a general rule is to check whether a termination cue occurs in the potential scope; if it occurs, the cue and the scope are verified as invalid. Except for termination terms described above, termination cues also include some punctuation, e.g., ": | - | " | ;" for "CONCLUSION", "SYNOPTIC" and "" | ;" for other sections. Specific rules were designed according to entity types. For instance, several keywords were defined as valid scope for En:Distant Spread or Metastases, such as "deposit", "spread" and "metastases"; if the scope does not contain one of these keywords, it will be filtered out as invalid output. This rule can filter some false positives, e.g., in this entity "deposit of tumour is present in mesenteric fat, with no residual vascular or lymph node architecture - considered an extranodal deposit (pN1c).", as the scope "residual vascular or lymph node architecture" does not consist of any of the keywords, the cue "no" and the scope are considered invalid. More examples are presented in Table 6.12. If both the cue and the scope are valid, it skips the following steps, and yields the output "absent".

Validity	Example
Valid	1. **no** obvious *lymphovascular invasion* ["In:Venous and Small Vessel Invasion"] 2. *tumour infiltrating lymphocytes* are **not** a feature ["Re:TILS and Peritumoural lymphocytes"]
Invalid	1. Tumour cells involve the subserosal layer (blocks 7 and 8), and block 8 shows *some microscopic involvement of the peritoneal surface* ["Ex:Serosal Involvement"], **without** obvious ulceration ["De:Serosa Description"]. 2. POORLY DIFFERENTIATED (HIGH GRADE) ADENOCARCINOMA WITH AREAS OF SIGNET RING DIFFERENTIATION AND FOCAL MUCINOUS DIFFERENTIATION, EXTENDING THROUGH COLON WALL TO ABUT THE SEROSAL SURFACE, EXTENDING TO THE LUMINAL SURFACE OF ADHERENT SMALL BOWEL, EXTENDING INTO ADHERENT ABDOMINAL WALL, INVOLVING OMENTUM, WITH *VASCULAR AND LYMPHATIC INVASION* ["In:Venous and Small Vessel Invasion"], **CLEAR OF** THE RESECTION MARGINS EXAMINED ["Ma:Clear"].

Table 6.12 Examples of valid and invalid negations. **Bold texts**: the cue, *Italic texts*: the scope, Underscore texts: the entity.

3. Detect uncertainty. Uncertainty phrases are searched by a string matching algorithm, and the matched entry is selected to be the cue. Next, a standard dictionary is used to map the cue to a standardized representation, which becomes the output.

4. Detect inapplicability. Inapplicability phrases are searched by a string matching algorithm. If an entry is found, the output is set to be "not applicable" (if it belongs to the "Not applicable" sub-category) or "unknown" (if it belongs to the "Unknown" sub-category).

If the output cannot be determined from the above processes, then "present" is assigned as the output.

6.4 Results and Discussion

The negation and uncertainty detection modules have to be combined together to determine the assertion for a medical entity. The combination of both modules for the colorectal cancer corpus has been described in the previous section. For the other two corpora, it is computed as follows:

1. If negation cannot be detected in the input, but "positive" instance(s) are found in the sentence, and uncertainty is detected as well, then the final output is set to be the output from uncertainty detection module.
2. If negation is detected in the input, the final output is "absent".
3. If uncertainty is detected in the input, the final output is the standardization of the uncertainty phrase.
4. If both negation and uncertainty cannot be detected in the input, the final output is "present".

6.4.1 Lymphoma Corpus

The combination of both modules yielded the following results for the lexicon-based method on the lymphoma training and test sets (see Table 6.13).

Category	Training set				Test set			
	Number	Precision	Recall	F-score	Number	Precision	Recall	F-score
absent	318	97.83%	99.37%	98.60%	96	83.95%	70.83%	76.84%
cannot exclude	1	50.00%	100.00%	66.67%	0	0.00%	0.00%	0.00%
definite	2	100.00%	100.00%	100.00%	0	0.00%	0.00%	0.00%
possible	264	98.48%	98.48%	98.48%	62	77.08%	59.68%	67.27%
present	2001	99.70%	99.45%	99.57%	543	81.04%	74.77%	77.78%
probable	30	96.55%	93.33%	94.92%	9	80.00%	44.44%	57.14%
Overall	2616	99.27%	99.27%	99.27%	710	81.10%	72.54%	76.58%

Table 6.13 Results for combination of negation and uncertainty detection modules on the lymphoma training and test sets.

It can be seen that there was a dramatic drop for the micro-averaged F-score by about 22.7% of the test set compared to training set. False positives and false negatives from "present" contributed to most of the errors, where most of them were caused by incorrect MER. Error analysis on the training set shows that some complicated cases require additional inference from the texts. For example, in the sentence:

> I am reluctant ["Li:Lexical Modality"] to diagnose mantle cell lymphoma ["Sy:Diagnosis"] in the absence of cyclin Dl ["An:Immunohistochemistry-Comment"] and t(11;14) without

["Li:Lexical Polarity Negative"] any strong morphological suggestion of ["Li:Lexical Polarity Positive"] MCL ["Sy:Diagnosis"].

the entity "MCL" was misclassified to "absent", and regarded "without" as the negation cue. Coreference resolution may be a helpful solution, as "MCL" co-referred to "mantle cell lymphoma", which was correctly identified as "possible".

6.4.2 Melanoma Corpus

Results for combining both modules on the training and test sets are shown in Table 6.14.

Category	Training set				Test set			
	Number	Precision	Recall	F-score	Number	Precision	Recall	F-score
absent	723	98.49%	99.45%	98.97%	160	97.42%	94.38%	95.87%
cannot exclude	11	90.91%	90.91%	90.91%	3	100.00%	66.67%	80.00%
definite	1	100.00%	100.00%	100.00%	0	0.00%	0.00%	0.00%
possible	210	97.17%	98.10%	97.63%	65	90.48%	87.69%	89.06%
present	2746	99.82%	99.45%	99.64%	696	90.23%	90.23%	90.23%
probable	25	96.15%	100.00%	98.04%	6	80.00%	66.67%	72.73%
Overall	3716	99.35%	99.35%	99.35%	930	91.42%	90.54%	90.98%

Table 6.14 Results for combining negation and uncertainty detection modules on the melanoma training and test sets.

From Table 6.14, the micro-averaged F-score decreased mildly from 99.35% to 90.98% on the test set. Most of the errors occurred in "present" due to incorrect MERs. There were 30 errors identified for negation or uncertainty, wherein 7 of them were due to the defects of the modules, and the rest were caused by poor MER performance. The defects of the modules include:

- The fixed window size for uncertainty detection can lead to the omission of some distant entities, e.g., the assertion "probable " of the entity "regression" in the sentence:

  Features were regarded as most probably ["Li:Modality"] representing ["Li:Lexical Polarity Positive"] a malignant melanoma ["Sy:Diagnosis"] with superficial dermal invasion ["In:Clark Level"] and regression ["Sy:Regression"].

  was not detected as it was nine tokens away from the uncertainty phrase "most probably".
- The scope involving prepositions were very difficult to tackle in some cases. Consider the following sentences:

  Sentence 1: "There is Pagetoid infiltration ["De:Cell Growth Pattern"] of the overlying-epidermis without ["Li:Lexical Polarity Negative"] epidermal ulceration ["De:Ulceration"], and an adjacent in-situ melanoma ["Sy:Diagnosis"] of superficial spreading type ["Sy:Subtype"]."

  Sentence 2: "The latter cells do not ["Li:Lexical Polarity Negative"] show ["Li:Lexical Polarity Positive"] the same mitotic activity ["De:Dermal Mitoses"] evident ["Li:Lexical Polarity Positive"] in the superficial portion of the tumour

["En:Primary Lesion"] but they are most probably ["Li:Modality"] also portion ["En:Primary Lesion"] of the malignant melanoma ["Sy:Diagnosis"]."

Where both the entities "in-situ melanoma" and "superficial spreading type" are negated by "without", while "portion" is asserted by "most probably", but not "malignant melanoma" itself. A simple lexicon-based algorithm has difficulty in determining the scope like these. The global context information would have to be taken into account to solve this problem.

Error analysis on the training set further revealed some weaknesses of the modules:

- The categorization of the trigger terms according to their positions in the training set does not cover all the possibilities of location in the test set. For instance, "cannot be excluded" was usually situated succeeding the entity; hence it was grouped to the Group 2 trigger terms. However, there were also exceptions like:

  In view of the evidence ["Li:Lexical Polarity Positive"] of extensive ["Li:Mood and Comment Adjuncts"] dermal regression ["Sy:Regression"], the possibility ["Li:Modality"] cannot be excluded ["Li:Modality"] that the melanoma ["Sy:Diagnosis"] , before ["Li:Temporality"] regression ["Sy:Regression"] , may ["Li:Modality"] have involved the superficial reticular dermis.

  where it occurred before the entity "melanoma", as its position was not matched to that in Group 2, it was ruled out by the uncertainty detection module.
- As addressed by Chapman et al, a simple lexicon-based algorithm could not handle complex cases which needed syntactic cues to resolve (Chapman et al., 2001a). Here is an example:

  It is of superficial spreading type ["Sy:Subtype"] and is not ["Li:Lexical Polarity Negative"] ulcerated ["De:Ulceration"] .

  where the entity "superficial spreading type" was misclassified as "absent".
- It seemed that explicit expressions of negation and uncertainty were harder to detect than implicit ones. In the following example:

  In view of the lack ["Li:Lexical Polarity Negative"] of a well developed ["Li:Mood and Comment Adjuncts"] junctional component ["En:Primary Lesion"], it is probably ["Li:Modality"] best classified ["Li:Lexical Polarity Positive"] as nodular ["Sy:Subtype"] melanoma ["Sy:Diagnosis"] in this material, although in this site acral lentiginous type ["Sy:Subtype"] is also considered.

  The correct assertion of the entity "acral lentiginous type" was "possible", whereas, there was no lexical nor syntactic information to indicate it, thus the uncertainty module failed to detect it. This issue requires domain knowledge to resolve.

6.4.3 Colorectal Cancer Corpus

Table 6.15 shows the performance of the NUI Detector on the training and test sets.

Category	Training set				Test set			
	Number	Precision	Recall	F-score	Number	Precision	Recall	F-score
absent	1509	99.80%	99.93%	99.87%	770	91.32%	87.40%	89.32%
cannot exclude	4	100.00%	100.00%	100.00%	4	100.00%	75.00%	85.71%
definite	1	100.00%	100.00%	100.00%	3	66.67%	66.67%	66.67%
not applicable	6	100.00%	100.00%	100.00%	0	0.00%	0.00%	0.00%
possible	22	91.67%	100.00%	95.65%	10	50.00%	40.00%	44.44%
present	974	99.90%	99.18%	99.54%	423	81.12%	72.10%	76.35%
probable	20	86.96%	100.00%	93.02%	13	75.00%	69.23%	72.00%
unknown	13	100.00%	100.00%	100.00%	1	0.00%	0.00%	0.00%
Overall	2549	99.65%	99.65%	99.65%	1224	87.45%	81.37%	84.30%

Table 6.15 Performance for Negation/Uncertainty/Inapplicability Detector on the colorectal cancer training and test sets.

From Table 6.15, the micro-averaged F-score for the test set declined by about 15.4% against the training set. The incorrect MER results for the entities in both "absent" and "present" were still the main reason for the drop of F-score. There were up to 180 errors in negation, uncertainty or inapplicability, wherein only 7 were created directly by the detector. These errors include:

- The efficiency of the lexicon-based approach was precluded by the limited predefined trigger terms. The trigger terms were obtained mainly based on the analysis of training data. They were not exhaustive, thus unseen terms in the test set could not be captured. For example, in the sentence:

  The possibility of vascular space invasion could not be excluded in the submucosa.

  where uncertainty phrase "could not be excluded" was not predefined in the trigger term lexicons, hence the detector omitted it.
- The determination of the scope involving preposition "of" has proven to be difficult in the melanoma corpus, while the scope involving the preposition "with" was also problematic. The preposition "with" was defined as a terminator for the detector, which satisfied most cases, whereas, in this sentence:

  These may represent discontinuous spread, venous invasion ["In:Venous and Small Vessel Invasion"] with extravascular spread ["In:Venous and Small Vessel Invasion"] or totally replaced nodes (TD).

  The asserted scope for "may" should be extended to the end of the sentence where "with" was not a correct terminator.
- The sequence for applying different detection modules can bring some problems. For example, the detector stated that negation detection takes precedence over uncertainty detection, which yielded the false output "absent" for the following In:Venous and Small Vessel Invasion entity: "Focal possible but not definite lymphovascular invasion is seen".

Additional deficiencies in the NUI detector were discovered through error analysis on the training set:

- The uncertainty trigger terms were not categorized according to their positions to the entities, which led to some false positives. For example, in the sentence:

Focal perineural invasion is seen ["In:Perineural Invasion"], and a focus of probable extramural lymphovascular invasion is identified ["In:Venous and Small Vessel Invasion"].

the uncertainty phrase "probable" only asserted the noun phrase "extramural lymphovascular invasion", but not "Focal perineural invasion".

- The keywords and rules used to validate the cue and the scope could not work well in some cases. For instance, in this En:Distant spread or Metastases entity: "more suggestive of consistent with this being a metastatic deposit rather than synchronous tumour", since the keywords list for the valid scope contained both "deposit" and "tumour", thus the detector could not verify the cue "rather than" and the scope "synchronous tumour" to be invalid.
- Another defect of the detector may be due to irregular grammatical structures of the text. Here is an example: "Perforated, likely secondary.", which was more likely to be a combination of several phrases rather than a regular sentence. The module could not handle these correctly.

6.4.4 Discussion of the Three Corpora

The evaluation performed on the training sets was to validate the competence of the rules. The high F-scores achieved by the lexicon-based method on the training sets indicate that the extracted rules were competent to cover most negation patterns in the corpora.

It can be seen from the above results that the method performed best on the melanoma test set, and worst on the lymphoma test set. The probable reasons for this are:

- The MER system had achieved encouraging performance on the recognition of associated Linguistic categories: Li:Lexical Polarity Positive, Li:Lexical Polarity Negative and Li:Modality, with 93.68%, 94.48% and 87.35% F-scores respectively in the 10-fold cross-validation experiments on the melanoma train set , which were significantly better than the counterparts on the lymphoma train set, which were 88.60%, 82.60% and 76.74% respectively.
- The performance of the recognition of most entity types by the MER system to be utilized for evaluation was also better in the 10-fold cross-validation experiments on the melanoma train set than on the lymphoma train set. Note that there are three entity types in the lymphoma train set: Sy:Constitutional Symptoms, Sy:Diagnosis Subtype and Ex:Other Sites of Disease which had attained F-scores of below 50%, while only one entity type: En:Lesion (other) in the melanoma corpus had achieved a very low F-score (15.62%) in the 10-fold cross-validation experiments.

Although the evaluation of the MER system on the test sets was not carried out, presumably, the performace would be similar to that in the 10-fold cross-validation experiments on the training sets.

Another reason for best performance attained on the the melanoma test set is that the melanoma corpus had the largest ratio between training set and test set (close to 4:1), which suggests that there would be more potential patterns and lexical information to be extracted for the rules in the training set.

The relatively poorer performance on the colorectal cancer test set also indicated that the utilization of associated linguistic categories could bring some advantages for negation and uncertainty detection. They could facilitate the acquisition of the trigger terms. The MER system would be able to identify them in unseen data by using the model trained on the training data. The ratio between training set and test set was smallest (about 2:1) in the colorectal cancer corpus, which could also hinder the extraction of rules or patterns to a great extent.

In addition, the reason for the poorest performance of uncertainty detection on the the lymphoma corpus lies in the fact that the accurate diagnosis of lymphoma is usually more difficult to be made by pathologists, thus the reports could contain more hedging information, which can be expressed in various forms, making it hard to capture the uncertainty patterns from the limited examples available in the training set.

Most of the errors on the test sets resulted from incorrect MER, accounting for 92.3%, 76.7% and 96.1% of the total errors in the lymphoma, melanoma and colorectal corpora respectively.

Nevertheless, the analysis of the errors types for the lexicon-based detection modules includes:

- Although most of the uncertainty trigger terms were very close to the asserted entities (usually within four-word window size), there were also some entities distant from these terms, thus the fixed window size for uncertainty detection would omit these entities.
- The lack of a positional cluster of the trigger terms would lead to some false positives; however a sloppy cluster would also bring some false negatives.
- It was not possible to determine the correct scope involving particular prepositions (e.g., “of”, “with”).
- It could not handle some complicated cases which needed syntactic cues or additional inference of the texts to resolve.
- Insufficient predefined trigger terms could affect the performance of the method.
- Current heuristic integration of different detection modules could also cause some problems.
- It had difficulty when facing implicit expressions of negation and uncertainty or irregular grammatical structures of the text.

Accordingly, some feasible improvements can be made to the method:

- A more flexible window size can be considered to determine the scope for uncertainty.

- A thorough analysis of the trigger terms, which is not only based on their positions relative to the entities, but also some conditions (e.g., whether a trigger term can be in the company of another trigger term).
- Resolving coreference or introducing syntactic cues would enable the method to cope with more complicated cases.
- Considering domain knowledge or global context information may be helpful to correctly determine the scope containing prepositional phrases or detect negation and uncertainty from implicit expressions.
- More lexical items could be considered to enrich the predefined trigger term lists.
- A more comprehensive integration of different detection modules to avoid some problems caused by weak integration.

6.5 Conclusion

The goal of negation and uncertainty detection on the corpora is to determine the assertion of the presence or absence of specific medical entities. In the case study of the lymphoma corpus, three different methods were experimented with: the lexicon-based approach was a rule-based method, modified from a known negation detection algorithm NegEx, relying on trigger terms and termination cues. The syntax-based approach was also a rule-based method, where the rules and negation patterns were designed according to the dependency output from the Stanford Parser. The machine learning-based approach used an SVM classifier to build models with a number of features. The syntax-based approach had the best overall performance on the training set, while the machine learning-based approach performed best on the test set. However, both of them were at the cost of very long run times. The lexicon-based approach was simple and efficient, and yielded more stable performance, thus it was preferable for the other two corpora. Given the challenges and characteristics of the corpora, a rule-based approach was created for uncertainty detection. The poorer performance for uncertainty detection suggests that uncertainty detection is much more difficult to handle than negation detection. The main adjustment for lexicon-based approaches applied to the other two corpora was to modify the entries of the trigger terms, pseudo-trigger terms and termination cues. There were also specific adaptations for each corpus. For example, the utilisation of the Li:Lexical Polarity Positive instances to derive the negation rules in the melanoma corpus, while specific keywords and rules were applied to validate the cue and the scope in the colorectal cancer corpus.

The good performances on the training sets are consistent with the finding of Mutalik et al's work that the language used in the medical domain is more restricted, so negation and uncertainty should be presented in much more direct and straightforward way in the texts (Mutalik et al., 2001).

Moreover, although the materials in this study are pathology reports of specific tumour streams, it still has some generalizability:

- The lexicon-based approach highlights the importance of adaptation to the corpus, which can boost the system performance markedly.
- Apparently, the negation rules and patterns purposed for the syntax-based approach can be reused to detect negation in other pathology notes, as they are not associated with semantic information.
- Since the methodology, classification strategies and algorithms adopted in the machine learning-based approach are general, they can be easily adapted for other negation detection tasks on clinical notes.

Though incorrect MER accounted for most of the errors on the test sets, error analyses that focused on the error types of the lexicon-based detection modules, reveals other problems, such as incorrect determination of the scope caused by the fixed window size for uncertainty detection, a sloppy cluster of the trigger terms and integration of different detection modules, difficulty in determining the correct scope involving particular prepositions, and restriction from insufficient samples of trigger terms. There are several possible solutions to improve them: utilization of more flexible window size, thorough clustering of the trigger terms, coreference resolution, introduction of syntactic cues, domain knowledge or global context information, additional lexical items considered as trigger terms, and more comprehensive integration of different detection modules.

The output from these modules was stored to populate values for associated fields in the structured templates, which will be described in Chapter 8 in detail.

Chapter 7 Relation Extraction

7.1 Introduction

Information extraction (IE) is a process to extract relevant information from unstructured text. As one of the major components in an IE system, entity recognition had been the focus in the early stage of IE. With the development of more and more complex IE systems, the significance of Relation Extraction (RE) was realized by more and more researchers. Extracting relations among entities is an efficient way to utilize the recognised entities so that the implicit connection among them can be revealed. It can help the users of the IE system to better understand the facts or events of interest without interpretation of irrelevant contexts.

In the clinical domain, RE is very important as not only medical entities themselves but also how they are related to each other are also of clinical significance. In the following sentences from Relation Annotation Guidelines of 2010 i2b2/VA Challenge (i2b2, 2010c):
Sentence 1: She has an elevated cholesterol ["Problem"] controlled with Zocor ["Treatment"].
Sentence 2: Penicillin ["Treatment"] causes rash ["Problem"].
Only recognizing the Treatment entities "Zocor" and "Penicillin", and the Problem entities "an elevated cholesterol" and "rash" from the examples does tell the differences between how "Zocor" affects "an elevated cholesterol" and how "Penicillin" is related to "rash". But with RE from the examples, they can be discriminated from each other: "Zocor" cures "an elevated cholesterol, while "Penicillin" causes "rash".

In the pathology domain, RE is also very important. Without RE, it is impossible to identify some crucial facts embedded in the texts, e.g., CD20, CD79a, CD10 and CD30 are the positive biomarkers in the following example, and CD3, cytokeratin and S100 are the negative biomarkers in the second example.
Example 1: On immunohistochemical stains the cells show diffuse strong membranous staining ["An:Immunohistochemistry-Positive"] for CD20 ["An:Biomarker"], CD79a ["An:Biomarker"] and CD10 ["An:Biomarker"] with moderate widespread membrane staining ["An:Immunohistochemistry-Positive"] for CD30 ["An:Biomarker"].
Example 2: The cells are negative ["An:Immunohistochemistry-Negative"] for CD3 ["An:Biomarker"], cytokeratin ["An:Biomarker"] and S100 ["An:Biomarker"].

Unlike other natural language processing (NLP) tasks, such as medical entity recognition, RE requires deeper analysis of the sentences because relationships often correspond to the grammatical structures of the sentences. In the above examples, the prepositional phrases composed of the preposition "for" and biomarkers modify the nouns "staining" or the adjective "negative".

The relation types are usually determined by the entity types involved. The involved entities are called the arguments of a relation. In the first example, “strong membranous staining” is the first argument of the Result-Positive relation, and “CD20” is the second argument of the relation; “negative” is the first argument of the Result- Negative relation, and “CD3” is the second argument of the relation. Only particular entity types can be connected with relations, which has been discussed in detail in Chapter 4.

In this chapter, an RE system for extracting relations from the lymphoma corpus is proposed. Specifically, the task attempts to identify relationships between eight types of medical entities and classify four relation types that occur amongst them. A rule-based approach was applied to classify Spatial Specialization relation, while a supervised machine learning-based approach was adopted to identify Result-Positive, Result- Negative and Result-Equivocal relations.

The rest of the chapter is organised as follows: firstly, it provides an overview of the mechanisms of Support Vector Machines (SVM), and then presents the classification strategy, system architecture and two proposed approaches. The Results and Discussion section illuminates the system performance and error analysis.

7.2 Support Vector Machines

Support Vector Machines (SVM) is a discriminative machine learning method that is based on the structural risk minimisation principle for binary classification. The basic idea is to find a decision hyper-plane to separate positive and negative examples by maximising the distance to the support vectors from each category.

Given *k* training examples $(x_i, y_i), i = 1, \cdots, k$, where each example has input data D ($x_i \in R^D$), and a category label with one of two values ($y_i \in \{-1, 1\}$). All hyper-planes in R^D can be parameterized by a vector (w) and a constant (b):

$$w \cdot x + b = 0$$

A canonical hyper-plane can be defined to separate the data from the hyper-plane by a distance of at least 1 (at least one example on both categories has a distance of exactly 1). It should satisfy

$$w \cdot x_i + b \geq +1, \text{ when } y_i = +1$$
$$w \cdot x_i + b \leq -1, \text{ when } y_i = -1$$

All such hyper-planes have a functional distance ≥ 1. For a given hyper-plane *(w, b)*, all pairs $\{\lambda w, \lambda b\}$ where $\lambda \in R^+$, define the exact same hyper-plane, but each can have a different functional distance to a given data point. The magnitude of *w* should be normalized to obtain the geometric distance from the hyper-plane to a data point by calculating

$$\frac{y_i\,(x_i \cdot w + b)}{\|w\|} \geq \frac{1}{\|w\|}$$

Intuitively, the hyper-plane is preferred to maximize the geometric distance to the closest data points (see Figure 7.1).

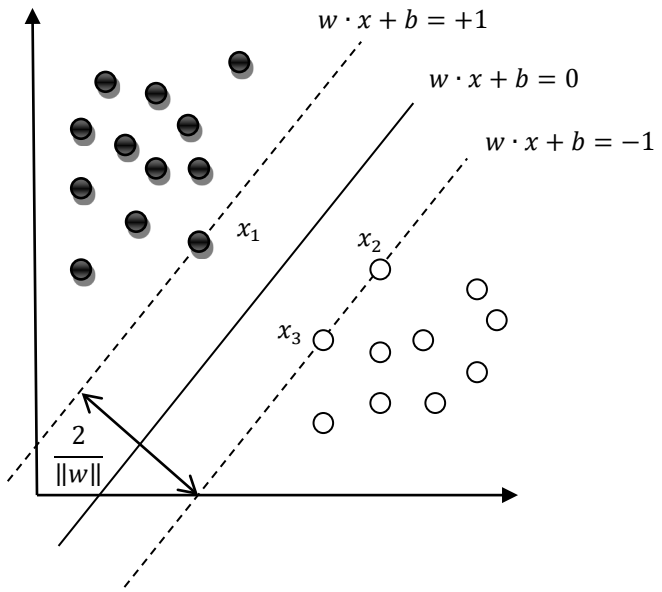


Figure 7.1 Support Vector Machines separate positive and negative examples. Note: x_1, x_2 and x_3 are support vectors.

Lagrange multiplier α is applied to minimizing $\|w\|$ (Burges, 1998), and the problem is transformed into:

$$\text{Minimize } W(\alpha) = -\sum_{i=1}^{k} \alpha_i + \frac{1}{2}\sum_{i=1}^{k}\sum_{j=1}^{k} \alpha_i \alpha_j y_i y_j\,(x_i \cdot x_j)$$

$$\text{subject to } \sum_{i=1}^{k} \alpha_i y_i = 0,\ 0 \leq \alpha_i \leq C\ (\forall i)$$

where α is the vector of k non-negative Lagrange multipliers to be determined, and C is a trade-off parameter between maximization of margin and minimization of error. Higher C weights more on classifying the training data correctly, while lower C results in a more flexible hyper-plane to minimize the margin error for each example (Alpaydin, 2004).

From the derivation of these equations, the optimal hyper-plane can be written as:

$$w = \sum_i \alpha_i\, y_i x_i$$

where w is a linear combination of the training examples.

According to the Karush-Kuhn-Tucker conditions, it shows that

$$\alpha_i\,(y_i\,(w \cdot x_i + b\,) - 1) = 0\ \ (\forall i)$$

Which suggests that when the functional distance of an example is greater than 1 ($y_i\ (x_i \cdot w + b\ ) > 1$), then $\alpha_i = 0$. Thus the training examples for $\alpha_i > 0$ are named support vectors, which are the only examples needed to define and find the optimal hyper-plane. Given any positive and negative support vector, x_p and x_n, it yields:

$$w \cdot x_p + b = +1$$

$$w \cdot x_n + b = -1$$

The constant *b* can be calculated by

$$b = -\frac{1}{2}\ (w \cdot x_p + w \cdot x_n)$$

The dual form of the SVM reduces to the following optimization problem:

$$\text{Maximize } W(\alpha) = \sum_{i=1}^{k} \alpha_i - \frac{1}{2}\sum_{i=1}^{k}\sum_{j=1}^{k} \alpha_i \alpha_j y_i y_j\ (x_i \cdot x_j)$$

$$\text{subject to } \alpha_i \geq 0, i = 1, \cdots, k \text{ and } \sum_{i=1}^{k} \alpha_i y_i = 0$$

When the input data are noisy, they are not easily separable. Cortes and Vapnik suggested a modified maximum margin idea to allow for mislabelled examples, which was known as the Soft Margin method (Cortes and Vapnik, 1995). It chooses a hyper-plane to split the examples as clearly as possible, by maximizing the distance to the nearest cleanly split examples. It introduces non-negative slack variables ε_i, which measure the degree of misclassification:

$$y_i\ (w \cdot x_i - b\ ) \geq 1 - \varepsilon_i, \qquad 1 \leq i \leq k$$

The optimization problem becomes a trade-off between a large margin and a small error penalty, which is

$$\text{Minimize } \frac{1}{2}\ \|w\|^2 + C \sum_{i=1}^{k} \varepsilon_i$$

$$\text{subject to } y_i\ (w \cdot x_i - b\ ) \geq 1 - \varepsilon_i,\ \varepsilon_i \geq 0$$

if the function penalizing non-zero ε_i is linear.

By introducing Lagrange multipliers α and β as done above, the problem becomes:

$$arg \min_{w,b,\varepsilon} \max_{\alpha,\beta} \left\{ \frac{1}{2}\|w\|^2 + C \sum_{i=1}^{k} \varepsilon_i - \sum_{i=1}^{k} \alpha_i [y_i\ (w \cdot x_i - b\ ) - 1 + \varepsilon_i] - \sum_{i=1}^{k} \beta_i \varepsilon_i \right\}$$

where $\alpha_i, \beta_i \geq 0$.

A linear classifier cannot separate data sets like those displayed in Figure 7.2, non-linear classifiers may resolve this issue (Hofmann et al., 2008). In non-linear classifiers, every dot product is replaced by a non-linear kernel function, so that the original input space can be transformed to higher dimensional space to find the optimal hyper-plane. For example, the data in Figure 7.2 can be separated by a non-linear classifier by transforming to a higher dimensional space in Figure 7.3.

There are several popular kernel functions, which are depicted below:

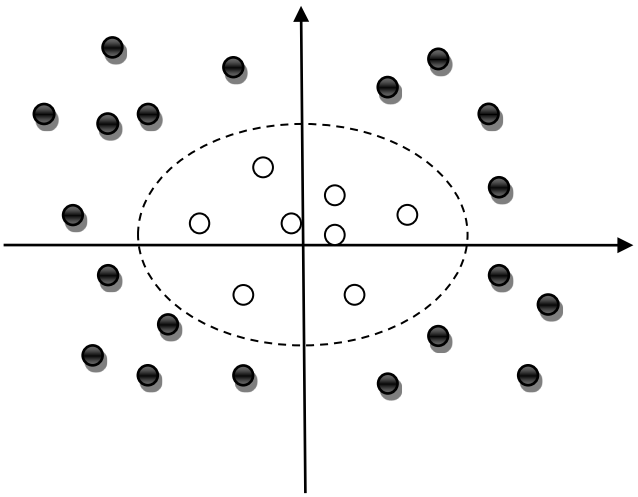


Figure 7.2 Data sets cannot be separated by a linear classifier.

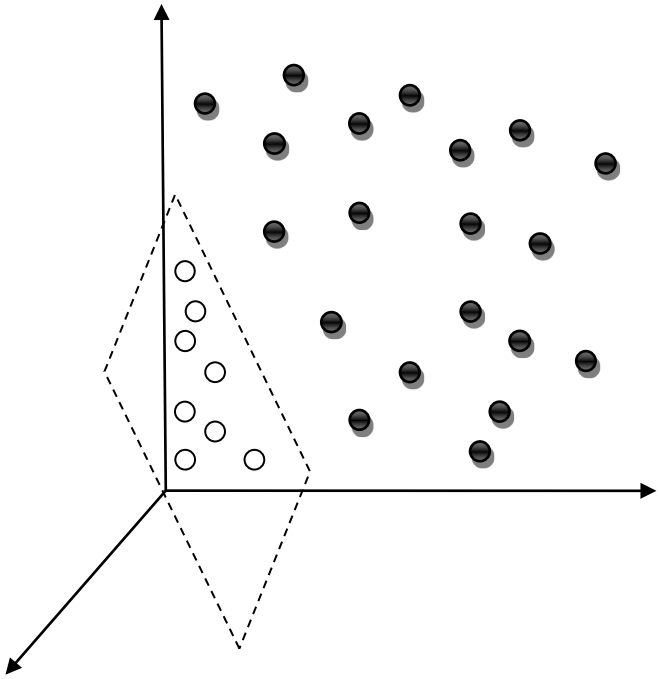


Figure 7.3 Higher dimensional space transformed from the original input space.

Linear Kernel

This is the simplest kernel that is used in a linear classifier, which is defined as

$$K(x_i, x_j) = x_i \cdot x_j$$

It can attain high accuracy if the data are linearly separable. It costs much less training time than non-linear kernels, especially when handling a very large number of features or training samples.

Polynomial Kernel

The polynomial kernel represents the similarity of training examples in a feature space over polynomials of the original variables, allowing learning of non-linear models. It is defined as

$$K(x_i, x_j) = (x_i \cdot x_j + 1)^d$$

where d is the degree of the kernel, which stands for the dimensionality of the feature space that the kernel transforms the data to.

Radial Basis Function

The Radial Basis Function (RBF) is defined as

$$K(x_i, x_j) = \exp(-\gamma \|x_i - x_j\|^2)$$

This kernel turns the hyper-plane into a Gaussian bell function. γ is related to the kernel width; a larger value for γ suggests the function is more specific to the training data, while a smaller value makes the function more generalised.

In summary, SVM has several advantages on classification tasks:

- The formulation of results in a global quadratic optimisation problem, which can be solved by interior point methods.
- The solution is obtained as a set of relevant support vectors, which lie on the boundary so that they can summarise the information to separate the data.
- The support vectors can be sparse, which is very useful for learning of the model, especially when only very small amount of training data are available.
- It can handle high dimensional feature spaces, which facilitate the integration of various features with it.
- The kernel functions provide several common model architectures, so that users can employ them in the classification tasks.

7.3 Relation Extraction System

7.3.1 Classification Strategy

In the following sections, a pair-wise method will be proposed to extract binary relationships between entities. Two entities e_1 and e_2 can be paired as (e_1, e_2), which can be instances of De:Anatomical Structure, De:Laterality, An:Immunohistochemistry-Positive, An:Flow Cytometry-Positive, An:Immunohistochemistry-Negative, An:Flow Cytometry- Negative, An:Immunohistochemistry-Equivocal or An:Biomarker. They can be connected via relations (*rel*), which can be one of the predefined relationships: Spatial Specialization, Result-Positive, Result-Negative and Result-Equivocal.

The RE task is formulated as a pair-wise classification task, which aims to classify every pair of entities (e_1, e_2) to the possible relation type *rel* between them or *None* if there is no relationship between them. Note that the order of appearance of arguments is not considered to affect a relation, e.g., if there is a Result-Positive relationship between e_1 and e_2, it makes no difference on whether e_1 occurs before or after e_2. But the order of argument types is confined to the annotation schema, e.g., the first argument of Result-Positive should be an An:Immunohistochemistry- Positive or An:Flow Cytometry- Positive entity, and the second argument should be an An:Biomarker entity.

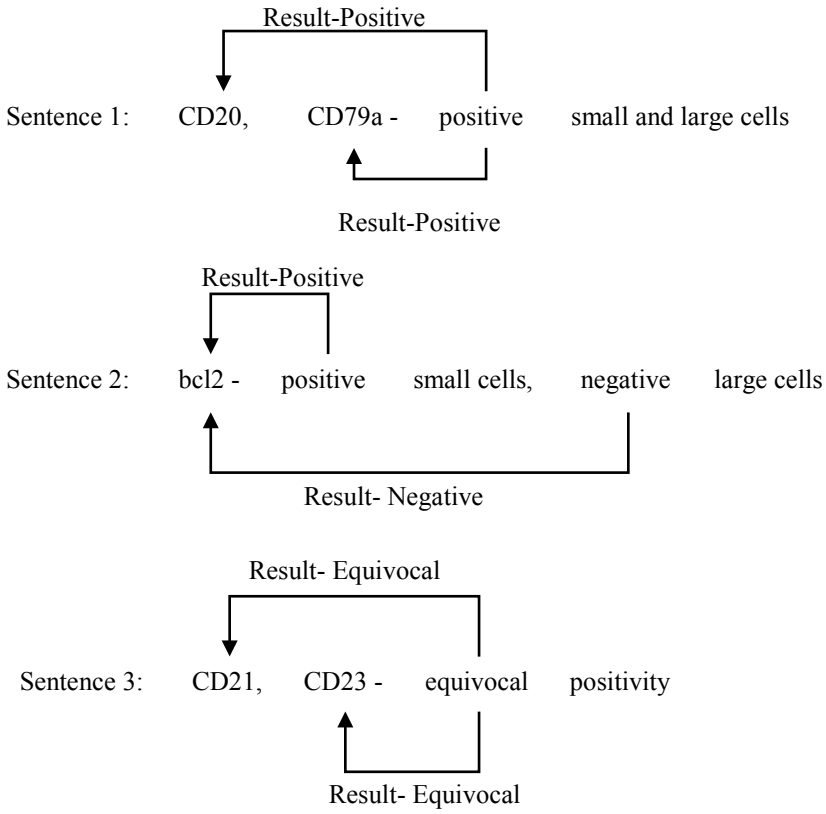


Figure 7.4 Three sentences with 9 entities and 3 relations hold between them.

To employ SVM for classification, both positive and negative examples are needed to be prepared. These examples are restricted within a ±1 sentence window. For example, the sentence in Figure 7.4 shows, there are three entities in each sentence. Three relations hold between them, namely Result-Positive, Result-Negative and Result-Equivocal. There are sixteen combinations of the entity pairs in total, wherein six are positive examples and ten are negative examples. These examples are listed in Table 7.1.

As mentioned in Chapter 2, there are two main streams for RE: rule-based approaches and statistical methods. Rule-based approaches are considered to be simple and reliable with limited amount of training data; machine learning-based approaches mainly focus on the feature pruning based on various levels of linguistic processing on the text. Given the sample size of different relation types in

the training data, a rule-based method is applied to extract Spatial Specialization relations, while a machine learning-based approach is used to extract other relations.

First entity	Position of first entity	Second entity	Position of second entity	Relation type
positive	Sentence 1	CD20	Sentence 1	Result-Positive
positive	Sentence 1	CD79a	Sentence 1	Result-Positive
positive	Sentence 1	bcl2	Sentence 2	None
positive	Sentence 2	CD20	Sentence 1	None
positive	Sentence 2	CD79a	Sentence 1	None
positive	Sentence 2	bcl2	Sentence 2	Result-Positive
positive	Sentence 2	CD21	Sentence 3	None
positive	Sentence 2	CD23	Sentence 3	None
negative	Sentence 2	bcl2	Sentence 2	Result-Negative
negative	Sentence 2	CD21	Sentence 1	None
negative	Sentence 2	CD23	Sentence 1	None
negative	Sentence 2	CD20	Sentence 1	None
negative	Sentence 2	CD79a	Sentence 1	None
equivocal	Sentence 3	bcl2	Sentence 2	None
equivocal	Sentence 3	CD21	Sentence 3	Result-Equivocal
equivocal	Sentence 3	CD23	Sentence 3	Result-Equivocal

Table 7.1 Entity pairs generated with their relation types in the sentences displayed in Figure 7.4.

7.3.2 System Architecture

The RE system architecture is illustrated in Figure 7.5.

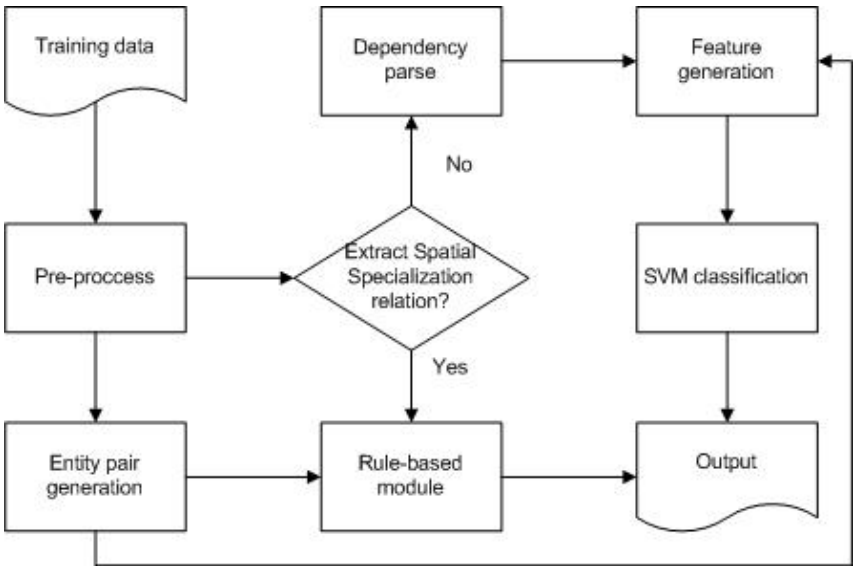


Figure 7.5 Architecture of the relation extraction system.

From Figure 7.5, first, the input data are passed to the pre-processing engine, which includes most of the pre-processes described in Chapter 5: sentence boundary detection, tokenisation, proofreading, part-of-speech (POS) tagging and shallow parsing.

The Entity Pair Generator generates the entity pairs according to the annotation schema. Note that for Spatial Specialization relation, only entities are paired; for other relations, the relation types connecting the entities are also included in the pairs to facilitate the learning of the statistical model. If it is to extract Spatial Specialization relation, a rule-based module will handle it; else, it will pass through to the subsequent procedures.

The Stanford parser (Klein and Manning, 2003) is used to perform dependency parses on the sentences, and the results are stored for further analysis. The feature generator prepares five features sets generated from the pre-processed texts and the dependency parse output for every entity pair. The SVM classifier classifies the relation type between each entity pair and yields the output.

7.3.3 Rule-based Module

Simple heuristic rules were applied in the module, which consists of three steps and is illustrated in Figure 7.6:

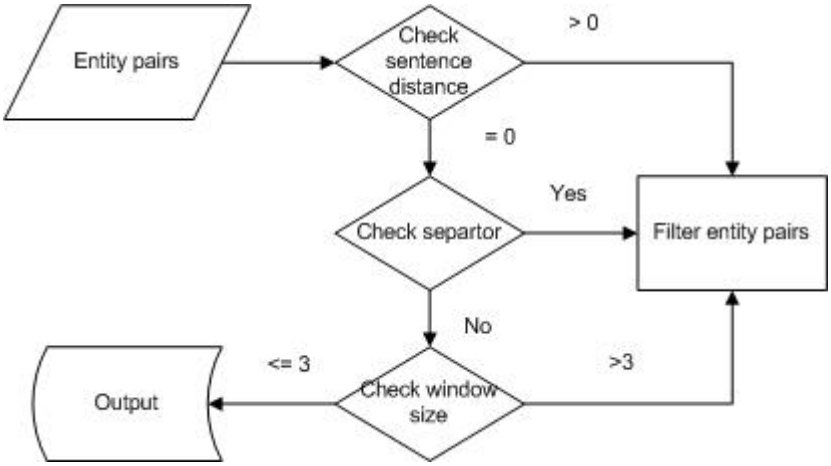


Figure 7.6 Workflow of the rule-based module in the relation extraction system.

1. Check whether the paired entities are in the same sentence. If they are not, they will be filtered out.
2. Check whether they are separated by particular punctuation or combination of punctuation (comma ",", semicolon ";", and arrow "->"). If they are, they will be filtered out.
3. Check whether they are inside a three-word window. If they are not, they will be filtered out.

If they are not filtered in the above processes, they will be linked with a Spatial Specialization relation.

7.3.4 Feature Sets

Various features were prepared for the SVM classifier, which can be categorised into five broader feature sets. They are a lexical feature set, semantic feature set, contextual feature set, syntactic feature set, and positional feature set, which are described in detail below.

Contextual Feature Set

Contextual window of the paired entities: As indicated in Giuliano'work (Giuliano et al., 2006), words surrounding the target entities often provide strong clues for RE. A ±4 token window of each entity in the pair was adopted in the task, which was determined by running some preliminary experiments on the training data. Given the entity at the i^{th} position in the sentence, this feature captured the tokens found in the $(i-1)^{th}$, $(i-2)^{th}$, $(i-3)^{th}$, $(i-4)^{th}$, $(i+1)^{th}$, $(i+2)^{th}$, $(i+3)^{th}$ and $(i+4)^{th}$ positions in the sentence. It treats each token at the above positions as an individual feature. ±1, ±2, and ±3 token windows are subsumed by the ±4 token window, such that any text string occurring in the smaller windows can be captured by the larger window as well.

Lexical Feature Set

To characterise the lexical nature of the local context of the involved entities, this feature set contains important lexical information about the entities or the text span between the entities.

Tokens inside each entity in the pair: Every token in the entities is used as a feature to represent the frequency of particular lexical items that make up the relations.

Lowercase of tokens inside each entity in the pair: The tokens are converted to lowercase in order to attain a higher recall.

Tokens between the paired entities: Helpful clues can be embedded in the tokens between the two entities.

- Some prepositions or verbs that express a state of being are indicators for the connected relation. For instance, the preposition "with" in the following sentence:

  They have strong staining ["An:Immunohistochemistry-Positive"] with fascin ["An:Biomarker"], and also with CD15 ["An:Biomarker"] and weaker ["An:Immunohistochemistry-Positive"] with CD30 ["An:Biomarker"].

  indicates the Result-Positive relation between "strong staining" and "fascin".
  Another example is the verb "is" as the 3rd person singular present tense to connect "CD15" with "equivocal", in the sentence:

  Staining for CD15 ["An:Biomarker"] is equivocal ["An:Immunohistochemistry-Equivocal"].

- Several punctuations provide hints for the construction of relations. For example, colon ":" is a strong hint for the connection of Result-Negative relation between "Negative" and "CD30" in the sentence:

  Negative ["An:Immunohistochemistry-Negative"]: CD30 ["An:Biomarker"].

Semantic Feature Set

Entity types of each entity in the pair: This feature explicitly indicates the argument types that comprise the associated relation type.

Entity types between the paired entities: The types of entity between the entities can function as an indicator to link two entities that are distant from each other or terminate the propagation of the relation span. For example, in the sentences below:

Positive ["An:Immunohistochemistry-Positive"] : CD20 ["An:Biomarker"], CD79a ["An:Biomarker"], CD23 ["An:Biomarker"], CD43["An:Biomarker"]
Negative ["An:Immunohistochemistry-Negative"] : CD5 ["An:Biomarker"], CD3["An:Biomarker"], cyclin D1 ["An:Biomarker"], CD30 ["An:Biomarker"], CD10 ["An:Biomarker"]

Although the distance from "Positive" to "CD23" and "CD43" exceeds ±4 token window, they can still be linked together, as there is only one unique entity type between them: An:Biomarker, suggesting that the span of the Result-Positive relation can be extended to the two An:Biomarker entities even if they are distant from "Positive". In contrast, the An:Immunohistochemistry-Negative entity "Negative" can terminate the span of the Result-Positive relation to subsume the succeeding An:Biomarker entities: "CD5", "CD3", "cyclin D1", "CD30" and "CD10".

Syntactic Feature Set

POS tags of each entity in the pair: These are the generalised representations of the paired entities, such as "JJ" for "Negative", "SYM" for "++", "NN" for "CD3" and "JJ NN" for "positive staining".

Shortest dependency path: To compute the shortest dependency path between the paired entities, first, the headwords of the entities need to be identified, which were determined as follows:

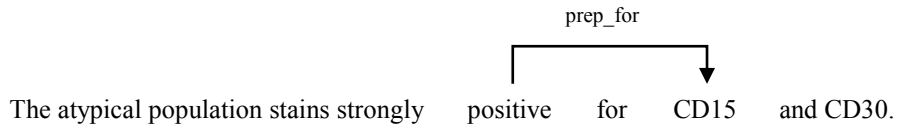
- For entity types of An:Immunohistochemistry-Positive and An:Flow Cytometry-Positive, two lists of specific lexicons are prepared, which are displayed in Table 7.2. A string match is used to search the lexicons in the text of the entity according to the order in the lists, and the matched entry becomes the headword of the entity; if no match can be found, the first token is the headword of the entity.
- For An:Biomarker entities, the last token of the entity is the headword of the entity.
- For other types of entity, the first token is supposed to be the headword of the entity.

List #	Lexicon
1	positively, positive, positivity
2	strongly, moderately, weakly, weaker, strong, moderate, weak

Table 7.2 Lists of lexicons for searching headwords of Immunohistochemistry-Positive and Flow Cytometry-Positive entities.

The shortest dependency path between the headwords of the entities can be computed as described in the previous chapter. Here are some examples to demonstrate the computation in graphical form:

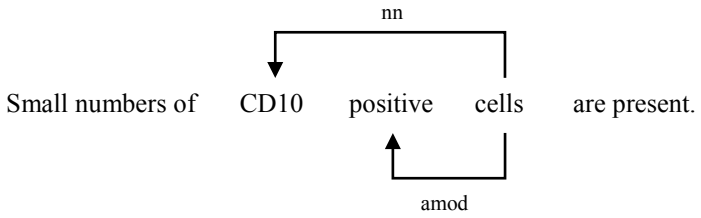
Example 1:



The shortest dependency path is:

prep_for (positive-6, CD15-8)

Example 2:

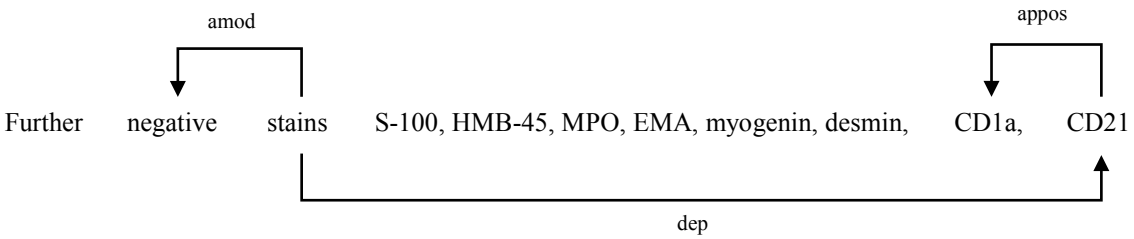


The shortest dependency path is:

nn (cells-6, CD10-4)

amod (cells-6, positive-5)

Example 3:



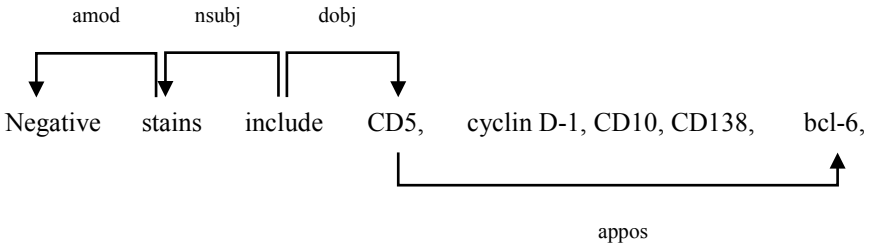
The shortest dependency path is:

amod (stains-3, negative-2)

appos (CD21-19, CD1a-17)

dep (stains-3, CD21-19)

Example 4:



lysozyme, TdT, CK, CD99, synaptophysin and CD56.

The shortest dependency path is:

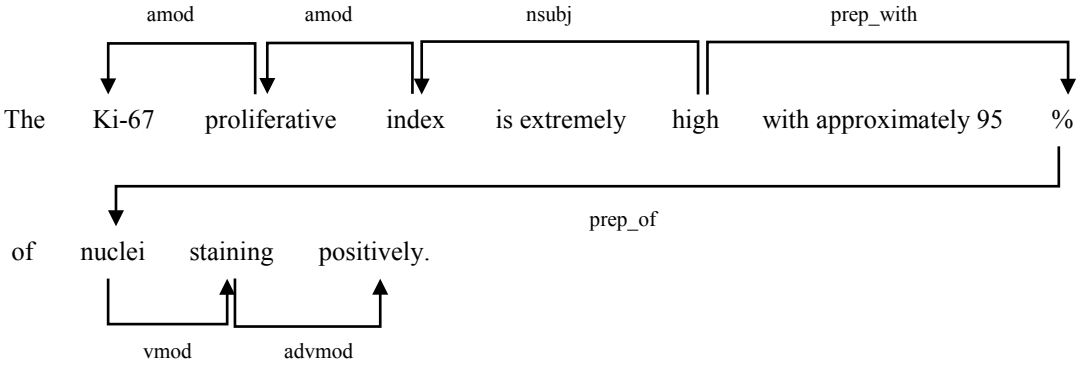
amod (stains-1, Negative- 0)

nsubj (include-2, stains-1)

dobj (include-2, CD5-3)

appos (CD5-3, bcl-6-12)

Example 5:



The shortest dependency path is:

amod (proliferative-3, Ki-67-2)

amod (index-4, proliferative-3)

nsubj (high-7, index-4)

prep_with (high-7, %-11)

prep_of (%-11, nuclei-13)

vmod (nuclei-13, staining-14)

advmod (staining-14, positively-15)

As the lengths of the shortest dependency paths for the positive pairs in the training data are not larger than three in most cases, so assign the value for the feature as “C1” if the length is zero or one; “C2” if it is two or three; “F” if it is larger than three; “O” if the shortest dependency path cannot be found.

Feature set	Feature	Abbreviation	Value
Contextual feature set	Contextual window of the first entity	BFW_FOUR_1 AFW_FOUR_1	-4 window: atypical|lymphoid|cells|a re +4 window: for|CD3,|CD4,|CD45RO
	Contextual window of the second entity	BFW_FOUR_2 AFW_FOUR_2	-4 window: are|positive|for|CD3, +4 window: ,|CD45RO|but|are
Lexical feature set	Tokens inside the first entity	C1_TOKENS	positive
	Tokens inside the second entity	C2_TOKENS	CD4
	Lowercase of tokens inside the first entity	C1_TOKENS_LOW	positive
	Lowercase of tokens inside the second entity	C2_TOKENS_LOW	cd4
	Tokens between the entities	BTW_TOKEN	for|CD3,
Semantic feature set	Entity types of the first entity	C1_CLASS	Immunohistochemistry-Positive
	Entity types of the second entity	C2_CLASS	Biomarker
	Entity types between the paired entities	BTW_TYPE	Biomarker
Syntactic feature set	POS tags of the first entity	C1_POS	JJ
	POS tags of the second entity	C2_POS	NN
	Shortest dependency path	DEP_LEN	C2
Positional feature set	Token distance between the paired entities	TOKEN_DIS	C1
	Sentence distance between the paired entities	SEN_DIS	0
	The order of appearance for the paired entities	POSITION	S

Table 7.3 Examples of features prepared for the entity pair *(positive, CD4)* in the sentence: "The atypical lymphoid cells are positive for CD3, CD4, CD45RO but are negative for CD20, CD8 and CD30.". Note: multiple values are separated by pipe "|".

Positional Feature Set

Token distance between the paired entities: This refers to the distance between the paired entities along a token path.

The average token distance for all positive pairs in the training data is smaller than four. There are three possible values to be assigned to this feature: "C1" if the token distance between the paired entities is not larger than two; "C2" if it is three or four; "F" if it is over four.

Sentence distance between the paired entities: This is a numeric value that was computed by the difference of the numbers of sentences between entities in the pair. Its possible values are 0 and 1.

The order of appearance for the paired entities: It seems that there is a pattern for the order of appearance for the entities with particular lexicons in the pair. If the first entity contains or consists of sign(s) such as "++", "-" and "1+", then it usually succeeds the second entity; if the first entity has an

initial capital , e.g., “Positive” and “Negative”, then it usually precedes the second entity. The feature value is assigned as “P” if the second entity is preceding the first entity; else, it is assigned as “S”.

Examples of the above features are presented in Table 7.3.

7.3.5 Vector Representation

To represent a relation, a binary feature vector is created by using the extracted features. Each feature has a unique index associated with it, which is stored in a feature index file. Table 7.4 displays part of the data extracted from the file for the above examples. Given a relation instance *R,* assign the index value with 1 if the associated feature is active, thus the feature vector representation for it is:

$$R = (index\_1{:}1, index\_2{:}1, \cdots, index\_n{:}1)$$

where n is the total number of active features.

According to Table 7.4, the feature vector representation of the examples in Table 7.3 is:
R = (17:1 19:1 21:1 22:1 23:1 35:1 36:1 37:1 38:1 40:1 84:1 116:1 133:1 140:1 163:1 164:1 165:1 176:1 179:1 183:1 199:1 473:1 499:1 721:1 722:1 2722:1 3520:1)

Feature	Value	Index
C2_CLASS	Biomarker	17
TOKEN_DIS	C1	19
C2_POS	NN	21
C1_POS	JJ	22
BTW_TYPE	Biomarker	23
C1_TOKENS	positive	35
C1_TOKENS_LOW	positive	36
C1_CLASS	Immunohistochemistry-Positive	37
POSITION	S	38
SEN_DIS	0	40
AFW_FOUR_2	,	84
BFW_FOUR_1	cells	116
BFW_FOUR_2	for	133
DEP_LEN	C2	140
BFW_FOUR_2	CD3,	163
AFW_FOUR_2	are	164
BFW_FOUR_1	are	165
AFW_FOUR_1	for	176
BTW_TOKEN	for	179
BTW_TOKEN	CD3	183
AFW_FOUR_1	CD3,	199
BFW_FOUR_1	lymphoid	473
BFW_FOUR_1	atypical	499
C2_TOKENS	CD4	721
C2_TOKENS_LOW	cd4	722
AFW_FOUR_2	but	2722
AFW_FOUR_2	CD45RO	3520

Table 7.4 Associated indices for the examples in Table 7.3.

7.4 Results and Discussion

7.4.1 Experimental Settings

The RE system was evaluated on the lymphoma corpus that was described in Chapter 3 and 4. The rule-based module was run on the corpus to evaluate the coverage of the rules. The experiments for evaluating the SVM classifier were carried out as follows:
First, the corpus was pre-processed, and positive and negative examples were generated from it. A relation can cross a sentence boundary, and a ±1 sentence window was used to generate entity pairs, as in the analysis described in Chapter 4 where the entities holding relations appear within the same sentence or the adjacent sentences. An over-sized window is not necessary and harmful for the classifier, since it will lead to the dramatic increase of negative examples, which can result in the bias of the classifier towards them, slow down the training speed, and even impair the quality of the learning of the model.

All experiments for evaluating the SVM classifier were conducted with 10-fold cross-validation, and each fold was stratified on a document level instead of instance level. As pointed out by Sætre et al (Sætre et al., 2007), it is likely for an RE system to gain an artificial boost of performance by evaluation at the instance level, since one sentence may generate many similar features for multiple entity pairs within it, which will be used in both the training and testing stage, however, it is supposed that the test set should remain to be unseen in the training stage. Therefore, it is preferable to evaluate the RE system at a document level to prevent the instances of the test data overlapping with those of the training data.

The Multiclass SVM implementation of LIBSVM (Chang and Lin, 2011) was used in the experiments. To compare the effects of different kernels on the classifier, the performance of three popular kernels: linear, polynomial and RBF kernels were evaluated. A grid search method (Hsu et al., 2010) was used with 10-fold cross-validation to find the optimal values of parameters C and γ. The parameter d for polynomial kernel is set to be 2.

The system performance is measured by the standard evaluation metrics: Precision, Recall and F-score.

7.4.2 System Performance

7.4.2.1 Rule-based Module

The 100% F-score obtained in the experiments suggests that the rules worked well, but it is likely to be limited by the small sample size (only 12 samples). More samples are needed to test it in future work.

7.4.2.2 SVM Classifier

Feature Contribution

The contribution of each individual feature to the model is reported in Table 7.5. A baseline model was built using the lexical feature *tokens inside each entity in the pair* and the semantic feature *entity types of each entity in the pair*. Feature engineering was conducted by progressively adding features to the classifier using the RBF kernel. The best feature configuration was obtained by using all the features described in Section 7.3.4. The best model achieved precision with 96.70%, recall with 97.66% and F-score with 97.18%, which outperformed the baseline model by F-score of about 45.5%.

Model #	Features	Precision	Recall	F-score
1	Tokens inside each entity in the pair + Entity types of each entity in the pair	63.65%	43.55%	51.71%
2	M1 + Contextual window of the paired entities	82.10%	95.68%	88.37%*
3	M2 + Token distance between the paired entities	88.84%	94.02%	91.35%
4	M3 + Sentence distance between the paired entities	92.77%	97.40%	95.03%*
5	M4 + Lowercase of tokens inside each entity in the pair	92.96%	97.55%	95.20%
6	M5 + Entity types between the paired entities	95.24%	97.76%	96.48%
7	M6 + The order of appearance for the paired entities	95.73%	97.92%	96.81%
8	M7 + Tokens between the paired entities	96.06%	97.71%	96.88%
9	M8 + Shortest dependency path	96.60%	97.71%	97.15%
10	M9 + POS tags of each entity in the pair	96.70%	97.66%	97.18%

Table 7.5 Contribution of each individual feature to the model. Score marked with * suggests significant contribution within 95% confidence interval.

The baseline features showed their power on gaining a relatively high precision (63.65%), as the major lexical information for connecting the entities could be captured by the lexical feature: tokens inside each entity in the pair, and the primary semantic information were revealed by the semantic feature: *entity types of each entity in the pair*. These two features integrated with each other, and defined the basic linguistic construct of a potential relation. Entity types restricted the semantic types of the arguments in the relations, since only certain types of argument can hold a particular relation according to the annotation schema. The lexical variability is relatively low in some entity types, which can assist the classifier to recognise the relations with them. For example, the An:Immunohistochemistry-Negative entities always contain the word “negative” or sign “-”.

The contextual feature *contextual window of the paired entities* is the most effective feature, which boosted the system by about 36.7% F-score, especially improved the recall by about 52.1%. It is consistent with the finding that a relation between two entities is generally correlated with the words surrounding the entities. The local contextual information about the entities was well-preserved in the contextual window feature, which could compensate for the weaknesses of the baseline features to a great extent, especially on the loss of recall.

The positional feature set yielded a moderate gain on the system with a total of about 7% F-score, wherein token distance and sentence distance made the biggest contribution with 2.98% and 3.68%

gains on F-score. The positional feature *token distance* mainly improved the precision by ruling out some entity pairs where the entities are distant from each other; sentence distance improved both the precision and recall, which corrected some mistakes made by using the *token distance* feature, e.g., some positive entity pairs within the same sentence, though with relative long token distance.

The semantic feature *entity types between the paired entities* were also very effective, and increased the F-score by 1.28%. It remedied part of the defects caused by introducing contextual window and token distance features.

The syntactic feature set and the remaining lexical features only improved the overall F-score slightly by 0.3% and 0.24% respectively.

The remaining lexical features only yielded small gains, probably because of their redundancy with several other features. For example, if the original text of an An:Immunohistochemistry-Positive instance is in lowercase (e.g., "positive") or consisting of punctuation (e.g., "++"), the value for the lexical feature *lowercase of tokens inside each entity in the pair* is the same as that for *tokens inside each entity in the pair*. Likewise, if the token distance between the entities is no more than four, then the lexical feature *tokens between the paired entities* can be replaced by the contextual window feature.

Since the shallow and dependency parsing results were not reliable, the errors generated in syntactic pre-processes will propagate to the associated feature generation and account for the limited improvement by the syntactic feature set. One of the prominent issues is that the POS tag for an An:Immunohistochemistry-Negative instance "-" is "HYPH" provided by the GENIA tagger (Tsuruoka et al., 2005), and the Stanford parser usually treats it as colon ":" , thus it will be ignored in the dependency parse output. It was likely that the parser would fail on the long distance dependencies as well. For instance, in the parse tree of the sentence:

> Immunohistochemical stains show positive staining ["An:Immunohistochemistry-Positive"] of the large atypical cells with CD30 ["An:Biomarker"] (on repeated stain), fascin ["An:Biomarker"] and to a lesser extent with CD15 ["An:Biomarker"].

the prepositional phrase "with CD15" was attached incorrectly by the parser to modify the noun "extent" (see Figure 7.7). Consequently, the parser generated an incorrect dependency output for the sentence, and the shortest dependency path between "CD15" and "positive staining" could not be computed from this result. Another problem is that due to missing verbs or prepositions in the sentences, the parser fails to parse the sentences correctly. Here is an example:

CD20, CD79a - positive small and large cells

Without the verb "are" and preposition "in", "CD20" and "CD79a" cannot be linked to "positive" via an explicit grammatical relation (see Figure 7.8 (a)). By revising the sentence as

CD20 and CD79a are positive in small and large cells

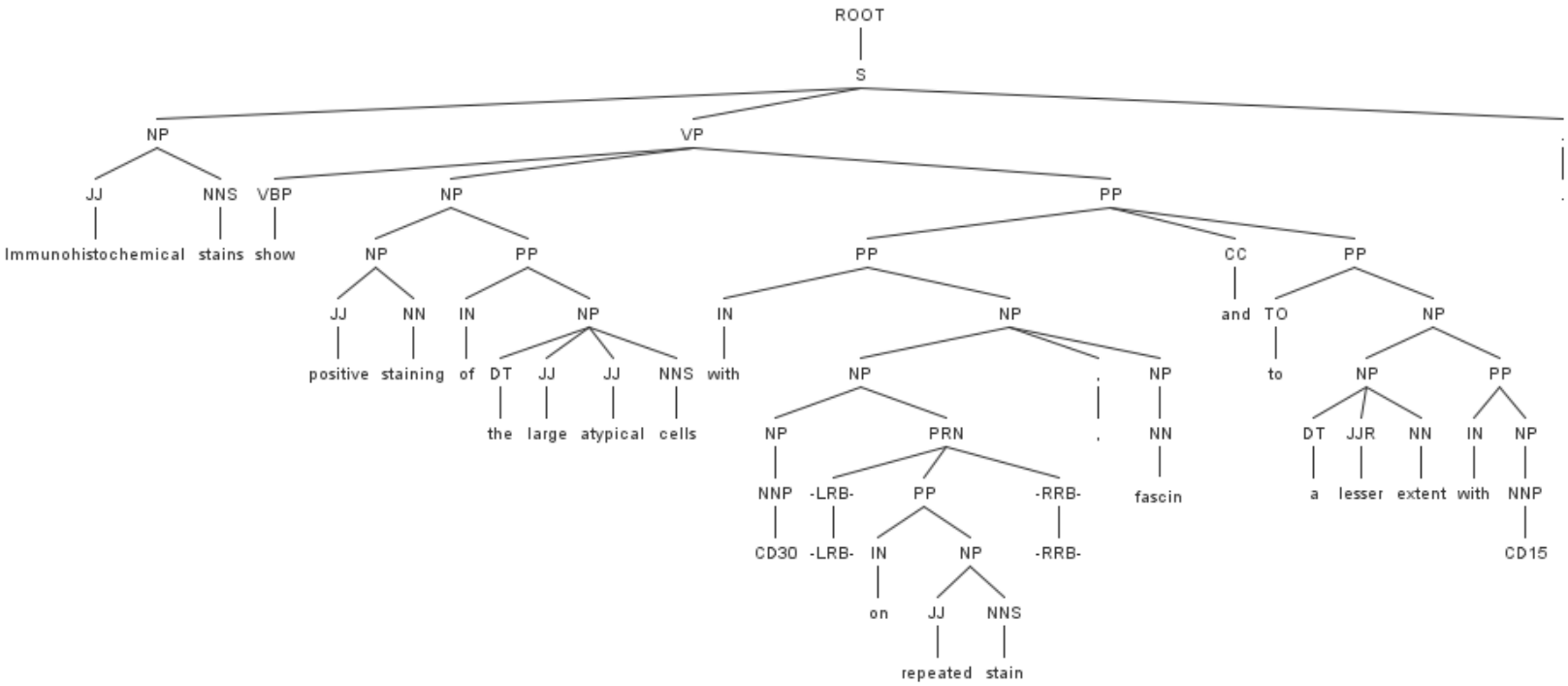
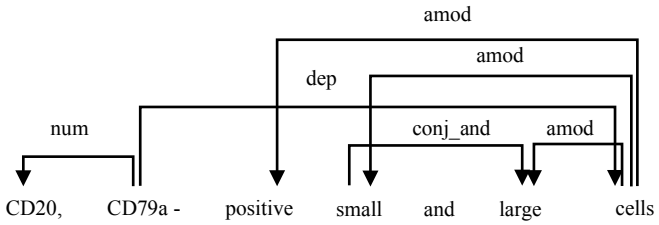
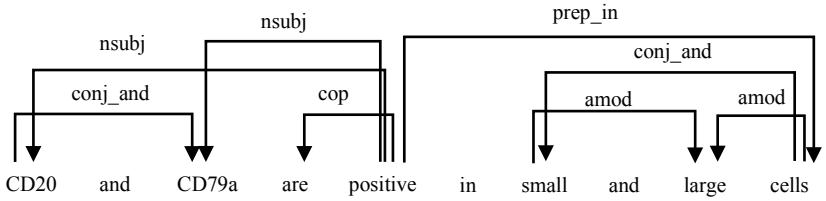


Figure 7.7 Parse tree of the sentence: “Immunohistochemical stains show positive staining of the large atypical cells with CD30 (on repeated stain), fascin and to a lesser extent with CD15.”



(a)



(b)

Figure 7.8 Dependency parse of the original sentence: “CD20, CD79a - positive small and large cells” and revised sentence “CD20 and CD79a are positive in small and large cells”.

the parser can yield correct dependency output for the sentence, where “CD20” and “CD79a” are linked to “positive” via “nsubj” (see Figure 7.8 (b)).

It is notable that this phenomenon is common in the corpus. It is necessary to adopt a parser trained on such ungrammatical texts in this domain, in order to fully utilise the syntactic structure information embedded in the texts. However, it requires much additional effort to develop or train such a domain-specific parser.

Unsurprisingly, the syntactic feature: shortest dependency path only contributes a slight boost of the system performance. In addition to the above reasons, the possible causes include:
Most dependency paths for the positive entity pairs are quite short, and the paired entities are probably located several tokens away, thus these short distance dependencies can be implicitly represented by some other features, e.g., the contextual window or token distance features. Therefore, the major effectiveness of the shortest dependency path feature should reflect on handling the longer distance dependencies. Nevertheless, as discussed above, the parser is more error-prone when coping with longer distance dependencies. Moreover, the dependency parse is performed at sentence level, so that for longer distance dependencies that cross sentence boundaries, the parser will fail to generate the dependency output for these cases.

Other features have been introduced to the model, such as lemmas and chunks of the paired entities, but they only introduced noise and decreased the overall F-score. This is possibly due to:

- They may represent some overlapping features. For example, *lemmas of the paired entities* are equivalent to *tokens inside each entity in the pair* if the tokens are in their canonical forms. According to the analysis on the training corpus, the morphological variants of lexical items inside the entities are very limited, thus *lemmas of the paired entities* can be replaced by *tokens inside each entity in the pair* in most cases.
- There may be unreliable results from the pre-processing. As mentioned above, there may be errors in the shallow parsing results provided by GENIA tagger. Hence using *chunks of the paired entities* as a feature with incorrect values may bring more harm rather than benefit for the learning of the model.

Performances of Individual Relation Type

The individual relation type performances obtained from the best model above are listed in Table 7.6.

Relation type	Number	Precision	Recall	F-score
Result-Equivocal	35	85.37%	100.00%	92.11%
Result-Negative	940	98.72%	98.19%	98.45%
Result-Positive	947	95.23%	97.04%	96.13%
Overall	1922	96.70%	97.66%	97.18%

Table 7.6 Performance for each individual relation type.

From Table 7.6, it can be seen that the micro-averaged F-score is over 97%, indicating that the features are sufficient to identify most of the relations. Result-Negative achieved the best performance with 98.45% F-score; Result-Positive attained the second highest F-score with 96.13%; the classifier performed worst on Result-Equivocal, with 92.11% F-score. The possible reasons for this are:

1. The sample size of Result-Equivocal instances (35 instances) is very small, and it is known that insufficient sample size can hinder the performance of a statistical classifier.
2. One of the constituent entity types for Result-Positive: An:Immunohistochemistry-Positive has more lexical variants than that of the counterpart for Result-Negative (An:Immunohistochemistry-Negative), the ratio being approximately 10:1. The An:Immunohistochemistry-Positive instances can be morphological variants of "positive", such as "positively" and "positivity"; phrases referring to the intensity of positivity, e.g., "strong", "moderately", and "weaker"; a combination of punctuation or numerals to indicate the intensity of positivity, such as "+" and "2+". It also has more syntactic variants, e.g., the POS tags, including "JJ", "NN", "RB", "JJR", "SYM", etc. The concurrence of different variants in the same sentences can also increase the difficulty for determination of the relation. For example, in this sentence:

   "Positive ["An:Immunohistochemistry-Positive"] : CD30 ["An:Biomarker"] +++ ["An:Immunohistochemistry-Positive"], Ki67 ["An:Biomarker"] (5% nuclei)"

there are two An:Immunohistochemistry-Positive instances: “Positive” and “+++”, where “Positive” should be connected to “Ki67”, while “+++”should be linked to “CD30”. This phenomenon often occurs in the Result-Positive instances, but does not exist in the Result-Negative ones. All of the above factors lead to greater variety in the linguistic patterns to determine the Result-Positive relations.

Error Analysis

Error analysis shows that most of the errors (about 67%) are probably due to the weaknesses of the features, while incorrect results from the pre-processing accounts for 28% of the errors.

Although the features have shown their advantages on recognising most of the relations, they still have weaknesses in several cases:

1. The possible values for token distance were determined based on the averaged count of the positive pairs in the corpus, which are not suitable for some cases. For example, it restricted the recognition of Result-Positive relations between the An:Immunohistochemistry-Positive entity “Positive” and some distant An:Biomarker entities (“CD45”, “CD138”, “CD30” and “CD3”) in this sentence:

   > Positive [“An:Immunohistochemistry-Positive”] - CD20 [“An:Biomarker”] and CD79a [“An:Biomarker”] (only a proportion of the large cells stain with each antibody), CD45 [“An:Biomarker”] (most cells), CD138 [“An:Biomarker”], EMA [“An:Biomarker”] (strong [“An:Immunohistochemistry-Positive”] and diffuse), CD30 [“An:Biomarker”] (moderate numbers of cells), Human Herpes Virus 8 [“An:Biomarker”] (HHV8 [“An:Biomarker”], strong staining [“An:Immunohistochemistry-Positive”]), CD3 [“An:Biomarker”] (scattered small lymphocytes).

   The argument for using four-tokens as a close distance also caused some misclassifications of the paired entities located at this distance. For instance, it brought a false Result-Positive relation between the An:Biomarker entity “CD30” and the An:Immunohistochemistry-Positive entity “positive” in this sentence:

   > The cells stain strongly [“An:Immunohistochemistry-Positive”] for CD30 [“An:Biomarker”], are also positive [“An:Immunohistochemistry-Positive”] for CD15 [“An:Biomarker”], but are negative [“An:Immunohistochemistry-Negative”] for CD20 [“An:Biomarker”].

2. The initiative for utilizing entity types between the paired entities was to try to extend or shrink the relation span through learning the possible entity types between the entities in positive pairs. However, this may allow some invalid constructions of the relations. For example, invalid Result-Positive relations were constructed among the An:Biomarker entities “CD10”, “CD20”, “CD79a”, “bcl-2” and the second An:Immunohistochemistry-Positive entity “positive”, as well as those among the An:Biomarker entities “CD3”, “CD43” and the first An:Immunohistochemistry-Positive entity “positive”, in the sentence:

Immunoperoxidase staining of these cells is positive ["An:Immunohistochemistry-Positive"] for CD10 ["An:Biomarker"], CD20 ["An:Biomarker"], CD79a ["An:Biomarker"] and bcl-2 ["An:Biomarker"] with CD3 ["An:Biomarker"], CD43 ["An:Biomarker"] positive ["An:Immunohistochemistry-Positive"] T cells mainly in the interfollicular regions.

as one of the frequent entity types between the entities in positive pairs is An:Biomarker. It also precluded some valid connections of the entities. For instance, the connection of "bcl2" and "negative" was excluded in the sentence:

bcl2 ["An:Biomarker"] - positive ["An:Immunohistochemistry-Positive"] small cells, negative ["An:Immunohistochemistry-Negative"] large cells

since an An:Immunohistochemistry-Positive entity "positive" occurred between them.

3. The tokens between the paired entities feature did not consider the implicit meaning of particular tokens, hence it failed to rule out some invalid relations. For example, the invalid Result-Negative relation between "CD21" and "negative" should be ruled out from the sentence:

Immunohistochemical stains show the large cells to be positive ["An:Immunohistochemistry-Positive"] for CD20 ["An:Biomarker"], CD79a ["An:Biomarker"], kappa ["An:Biomarker"], BCL-2 ["An:Biomarker"], BCL-6 ["An:Biomarker"] (scattered nuclei) and CD21 ["An:Biomarker"] (patchy cytoplasmic staining); and negative ["An:Immunohistochemistry-Negative"] for CD5 ["An:Biomarker"], lambda ["An:Biomarker"], CD10 ["An:Biomarker"] and CD30 ["An:Biomarker"].

considering the occurrence of the punctuation semicolon ";". This suggests that it needs discrimination from particular tokens. However, to obtain such particular lexicons requires additional effort not only in the investigation of the positive pairs but also negative pairs in the corpus.

4. The goal of using dependency paths as a feature is to identify some paired entities that may locate far along a token path but close along a dependency path. It classified the dependency path lengths of zero and one to the same category "C1", which may lead to some problems on classifying the entity pairs with short dependency distance. In the following example:

Despite the negative ["An:Immunohistochemistry-Negative"] CD23 ["An:Biomarker"], and strong ["An:Immunohistochemistry-Positive"] CD20 ["An:Biomarker"] on flow cytometry and immunohistochemistry, I favour an atypical chronic lymphocytic leukaemia (or small lymphocytic lymphoma), morphologically, and in view of the strong ["An:Immunohistochemistry-Positive"] CD5 ["An:Biomarker"] staining on immunoperoxidase stains.

the dependency path between "CD23" and the first An:Immunohistochemistry-Positive entity "strong" can be computed by

conj_and (CD23-3, CD20-7)

amod (CD20-7, strong-6)

where the length is one.

But the dependency path between “CD20” and “negative” can be computed by

amod (CD23-3, negative-2)

conj_and (CD23-3, CD20-7)

where the length is also one.

Thus, the classifier produced an incorrect Result-Positive relation between “CD23” and “strong”, as well as an invalid Result-Negative relation between “CD20” and “negative”.

The main defect of the pre-processing was incorrect results from sentence boundary detection, which affects the following dependency parse, and the effectiveness of the *sentence distance* feature. For example, the following sentence:

> The negative [“An:Immunohistochemistry-Negative”] CD23 [“An:Biomarker”] on immunostaining and flow cytometry does not support CLL/SLL and the negative [“An:Immunohistochemistry-Negative”] CD10 [“An:Biomarker”] and BCL6 [“An:Biomarker”] does not support a follicular lymphoma.

was incorrectly divided into two sentences by the sentence boundary detector:

> The negative CD23 on immunostaining and flow cytometry does not support CLL/SLL and the negative” and “CD10 and BCL6 does not support a follicular lymphoma.

which caused the omission of the Result-Negative relations among “CD10”, “BCL6” and the second An:Immunohistochemistry-Negative entity “negative”, because a Result-Negative relation seldom crossed sentence boundaries.

Likewise, the sentence boundary detector could not recognize that the following text is composed of two sentences:

> Larger cells -CD30 [“An:Biomarker”], CD15 [“An:Biomarker”], Fascin [“An:Biomarker”] positive [“An:Immunohistochemistry-Positive”]
> EMA [“An:Biomarker”], ALK1 [“An:Biomarker”], LCA [“An:Biomarker”], CD3 [“An:Biomarker”], CD20 [“An:Biomarker”] negative [“An:Immunohistochemistry-Negative”]

Consequently, a false Result-Positive relation was defined between “EMA” and “positive” by the classifier.

Another issue is about the difficulty of dependency parsing to correctly parse some ungrammatical sentences. Here is an example:

> CD5 [“An:Biomarker”] - Scattered small cells positive [“An:Immunohistochemistry-Positive”] consistent with reactive T-cells.

As it seems to be a combination of several phrases rather than a sentence, the parser yielded erroneous results for it. From Figure 7.9, there are several mistakes in the parse tree:

- The POS tag for “Scattered” should be “JJ” rather than “VBN”;

- The chunk tag for "Scattered small cells positive consistent with reactive T-cells" should be "S" instead of "VP";
- The adjective "consistent" and the prepositional phrase "with reactive T-cells" were attached to the wrong place.

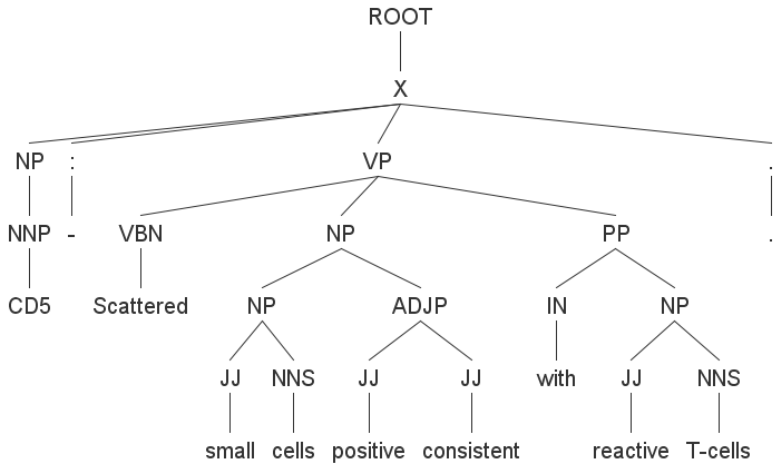


Figure 7.9 Parse tree of the sentence: "CD5 - Scattered small cells positive consistent with reactive T-cells."

As discussed above, the parser was not trained with pathology notes, thus it was unable to parse this ungrammatical sentence correctly.

The classifier also could not handle entity pairs situated in the sentences in irregular structures. For example, the classifier was confused by the sentence:

> CD5 ["An:Biomarker"] + ["An:Immunohistochemistry-Positive"] and CD23 ["An:Biomarker"] - ["An:Immunohistochemistry-Negative"], suggesting mantle cell lymphoma on flow cytometry, but cyclin Dl ["An:Biomarker"] negative ["An:Immunohistochemistry-Negative"].

Due to the misuse of the verb "suggesting" and omission of the verb "is" or the form "suggest" being used instead.

Comparison of Kernels

The comparative results among different kernels employed in the classifier are displayed in Figures 7.10, 7.11 and 7.12.

From Figure 7.10, 7.11 and 7.12, the polynomial kernels achieved significantly better overall F-score on model 1, probably due to the higher recall on the model; the gaps between each kernel was narrowed by increasing features, and remained stable for models 4 ~10.

The better performance attained by the polynomial kernel on model 1, suggests that there is a positive influence on the system performance by mapping the original feature space into a higher dimensional

feature space, especially with a limited feature size. But this influence will be reduced with features added to the model.

Figure 7.10 F-scores of three kernels on each language model.

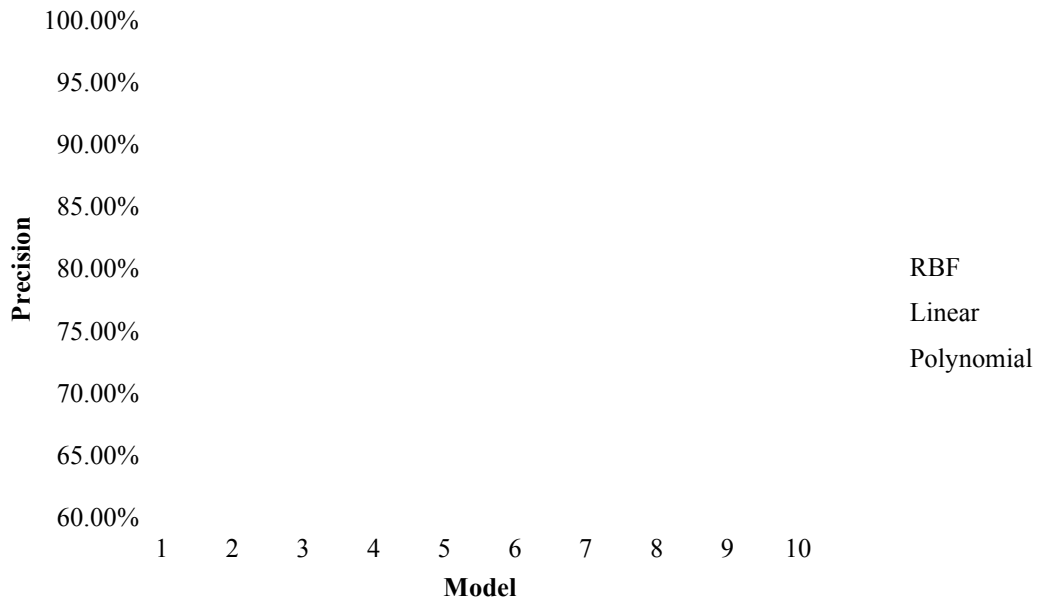


Figure 7.11 Precisions of three kernels on each language model.

It seems that the contribution of the features to the models was also affected by the kernels. The positional feature token distance yielded prominent larger gain on the precision of model 3 with the polynomial kernel, while the contextual window improved the recall to a less extent of model 2 with the linear kernel. This is possibly because in the models there were data that are not linearly separable,

and a non-linear kernel like the polynomial kernel was more suitable to separate them, as it increased the flexibility of the classifier; the RBF kernel could extend the feature space into an infinite number of dimensions, while the polynomial kernel could create combinations of features.

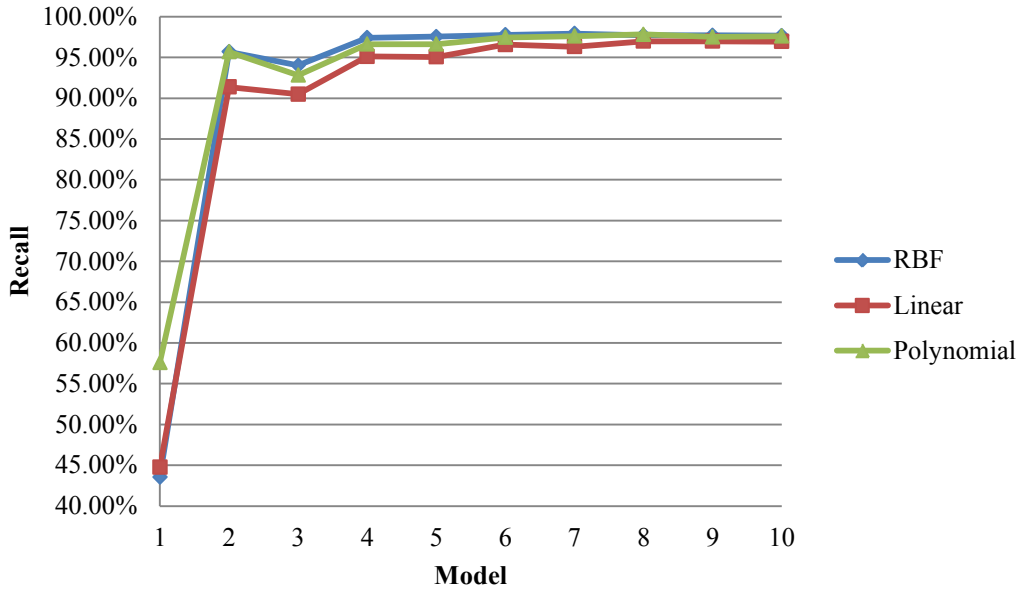


Figure 7.12 Recalls of three kernels on each language model.

There are different opinions on kernel selection in SVM. On the one hand, some researchers advocate that the linear kernel should be considered initially, as it has a simpler training algorithm that saves more time during training and scales well with the number of training examples (Bishop, 2007; Hastie et al., 2001); it has only one parameter to be tuned, that can prevent over-fitting to the training data, which the RBF and polynomial kernels may lead to with a small sample size. On the other hand, in many applications, SVM classifiers armed with non-linear kernels could provide better accuracy (Cristianini and Shawe-Taylor, 2000). In a simulation study done by Way et al, the polynomial kernel was more vulnerable to overtraining with large feature sets, and the RBF kernel was better than or comparable to the polynomial kernel under most conditions (Way et al., 2010). In this study, given the medium sample size and feature size, the RBF kernel was selected at first.

According to Figure 7.10, with a considerable amount of features, there was no significant difference between the performances of each kernel, which suggests that the choice of kernels has limited influence on the system with medium sample size when sufficient features were provided.

General Applicability

The relative value of various types of features has been demonstrated in this study, and researchers can use these features as a baseline for the development of more complex models in the future. Furthermore, the system performance was accomplished by using simple models and SVM, a classical

machine learning algorithm, therefore the system is likely to be durable and reusable and can be readily modified to meet different requirements for other clinical relation extraction tasks.

7.4.2.3 Limitations

Note that although the experiments have shown very good performance, there are also some issues addressed by them:

1. Incorrect results from the pre-processing, includes errors from sentence boundary detection and dependency parsing. To overcome this problem, it requires a more sophisticated sentence boundary detector and parser trained on the domain.
2. There are defects in feature extraction and construction. For example, there should be further consideration on particular tokens for the tokens between the paired entities feature; the categorization of the dependency path length may be too ambiguous, which needs to be tuned for short dependency distances.
3. The sample size for the rule-based module was too small so that the module could not be properly evaluated in the experiments.
4. There are disadvantages in the feature selection method. In this study, a "bottom-up" method (Whitney, 1971) was used, where features were progressively added to an initial empty feature set to find the best configuration. Its counterpart is the "top-down" method, where features are gradually removed from an initial full feature set to obtain the best feature set. Both the methods suffer from the nesting effect that features once added cannot be removed or once removed would not be re-considered. This can be overcome by the stepwise feature selection (Sahiner et al., 2000) and sequential forward floating search (Pudil et al., 1994) methods. Better still would be an investigation into the inter-relationships between the features so that the particular structures they are exploiting are clearly identified and redundancies between features removed.

7.5 Conclusion

This chapter presents a relation extraction system to extract four relations from the lymphoma corpus, including a rule-based module and a SVM classifier. Simple heuristic rules were applied in the rule-based module, while several useful features were prepared for the SVM classifier. The system has achieved very good performance, with 100% F-score obtained by the rule-based module and 97.18% micro-averaged F-score attained by the SVM classifier. The contextual, positional and semantic features were identified as the most effective features.

Error analysis shows that weaknesses of the features and incorrect results from the pre-processing were the main reasons for the loss of precision and recall. The small sample size for testing and the disadvantages of the feature selection method were also addressed in the evaluation of the system. Future work can be focused on adopting a more sophisticated sentence boundary detector and domain-

specific parser, improving feature extraction and construction, and using other methods for feature selection.

Chapter 8 Structured Output Generation

8.1 Introduction

Many information extraction (IE) shared tasks only focused on one aspect: extracting relevant information from unstructured text, but neglected the subsequent task: transforming the extracted information into structured data. Although they might have specifications for the output formats (in a structured or semi-structured style), the goal of proposing these specifications was to ease the evaluations of the extracted information rather than facilitate the users to access or understand them. For example, in the concept extraction task of 2010 i2b2/VA Challenge (i2b2, 2010b), the organizer required the system output should be a plain text file that contains entries in the form:

"c= concept text offset || t=concept type"

So that the evaluation scripts could run on the system output and compute the system performance.

Unlike these tasks, this study also emphasizes the importance of structured representation of the extracted information, so as to represent the information in a straightforward way that the users can understand and utilize easily and efficiently.

The targeted users of the system in this study are pathologists and clinical staff, although they did not participate in the validation of the system directly at present. Without a proper structured representation of the extracted information, they may be reluctant to use it; or worse, the inappropriate representation may affect the efficiency to make clinical decisions, and consequently diminish the quality of the clinical management of the patients.

Note that the structured representation in this work refers to the population of structured templates instead of structured generation of codes, thus it is distinguishable from other systems, e.g., MedLEE (Friedman et al., 2004) and cTAKES (Garla et al., 2011), which aimed to encode medical concepts in clinical documents.

The structured representation process includes construction of predefined templates and population of the templates, which will be described in this chapter. The rest of the chapter is organised as follows: firstly, it depicts the design of the predefined templates, and then presents the detailed mapping strategies and a particular sub-system for populating these templates. The results section illustrates the performances evaluated on the sub-system and those evaluated on the full system by assembling all components together.

8.2 Design of Structured Templates

As mentioned in Chapter 2, the structured templates were established based on three associated structured cancer reporting protocols from the Royal College of Pathologists of Australasia (RCPA).

Moreover, they were slightly modified for the corpora according to the detailed analyses presented in Chapter 3.

The reason for using structured templates rather than utilizing clinical entities identified in Chapter 4 directly is:
Most clinical entities are defined based on the standards and guidelines from the structured protocols, which are suitable to be repoted in structured checklists. The strucutred protocols are designed accoding to the disease and what has to eb reproted about the disease. This information varies from one disease to another so no single list of clinical entities can be defiend for use across all protocols. According to the practice of clinical and pathology staff, a good structured pathology report should be formatted to provide information clearly and unambiguously to the treating doctors, and should be organised with their use of the report in mind. In this sense, the report differs from the structured checklist, which is organised with the pathologists' workflow as a priority.

Therefore, the aim of designing structured templates is to systematically report cancer diseases, making it easier for treating doctors to understand the reports.

8.2.1 Structured Template of the Melanoma Corpus

The structured template of the melanoma corpus is displayed in Table 8.1. From this table, almost each section context can be mapped to its associated section in the template, except for "SPECIMEN". "SPECIMEN" was finally decided to correlate with "CLINICAL HISTORY", as they seemed to be complementary with each other: the reporting items were similar; when the contents in one of them were missing, those of the other could function as a replacement or supplement.

Major differences in the template from the sample template provided in Primary Cutaneous Melanoma Structured Reporting Protocol (Scolyer et al., 2010) include:

1. Item "Comment" in the "CLINICAL" section was replaced with another item "Description", with a broader scope to cover the contents in "CLINICAL HISTORY" and "SPECIMEN".
2. The default unit for item "Mitotic rate" is "per mm2" in the protocol. However, units like "per HPF (High Power Field)", "per 5 HPFs" are also frequently used by pathologists in the corpus. A study indicates that the number of mitoses in a one square millimetre area is equal to the count in approximately 5 full HPFs with an Olympus BH2 microscope at ×400 magnification (Scolyer et al., 2003). It suggests that without knowledge of which brand of the microscope or the magnification the pathologist used, the arbitrary conversion of the count in HPFs to the number in one square millimetre may be inappropriate. Therefore, it was decided to present the units aside from the numeric values to obtain flexible representation of this field in the template.
3. A supplementary field was prepared to report the presence or absence of tumour infiltrating lymphocytes (TILs) in the template, given the analysis on the corpus that pathologists do not

always record the distribution and density of TILs, but only whether TILs were present or not in some cases; the presence or absence of TILs is also of prognostic significance. For example, in a recent study of melanoma, absent TILs could predict sentinel lymph node positivity (Taylor et al., 2007).

4. Item "Int. / late regression" in "MICROSCOPIC" section was replaced with "Regression", as
    - Regression can be categorised into three stages: early, intermediate and late.
    - The focus on intermediate and late stage may omit the importance of characteristics of the regression. For example, one study of thin melanomas showed that past regression adversely affected survival in patients, while active regression without fibrotic area did not have significant influence on it (Sondergaard and Hou-Jensen, 1985).
5. Item "Intraepidermal growth" was substituted with "Cell growth". Researchers pointed out that in the radial growth phase, the melanoma tends to grow within the epidermis along the lines or radii of a circle and does not form any expansive nest or nodule, which does not indicate any metastatic potential (Guerry et al., 1993); in a vertical growth phase, the melanoma extends vertically into the underlying dermis, where melanoma cells form expansive and coalescent nests and nodules, which shows metastatic potential with possible invasion into dermal lymphatic and vascular channels (Oliveira Filho et al., 2003). Only reporting the cell growth patterns within the epidermis may neglect the abnormal ones in other skin layers, e.g., dermis.

8.2.2 Structured Template of the Colorectal Cancer Corpus

Table 8.2 illuminates the structured template designed for the colorectal cancer corpus. From this table, each section context can be mapped to its associated section in the template, wherein "CONCLUSION" was associated with "Diagnostic Summary" section, while "SYNOPTIC" was relevant to "SYNTHESIS" section, considering the potential advantages of utilizing the synoptic fields presented in "SYNOPTIC".

There are several notable variations from the sample template provided in the Colorectal Cancer Structured Reporting Protocol (Eckstein et al., 2010):

1. In the sample template, only a type of operation should be considered as a value for item "Specimen type" (the original name "Type" was thought to be ambiguous, hence "Specimen type" was used instead). However, the analysis of the corpus revealed that the site of operation was also used by pathologists to imply the surgical resection of it, thus it was also considered as a possible value in the structured template.

Template section		Template item	Section context	Possible value
Diagnostic Summary		Summary	"DIAGNOSIS"	Contents in associated section context(s)
		Comment	"COMMENT"	Contents in associated section context(s)
Supporting Information	CLINICAL	Description	"CLINICAL HISTORY", "SPECIMEN"	Contents in associated section context(s)
		Site and laterality		Anatomical site
		Clinical diagnosis		Diagnosis made by the clinician; for negative diagnosis: "no" + the diagnosis; for uncertain diagnosis: an entry in the uncertainty standard dictionary + the diagnosis
		Specimen type		Surgical procedure, biopsy type
		Prev. Rx / Trauma		Cosmetic change indicating trauma/treatment (and history or timing if applicable)
		Previous melanoma		Present, absent, an entry in the uncertainty standard dictionary
		Distant metastasis		Present, absent, an entry in the uncertainty standard dictionary
		Other medical history		History of the current lesion
	MACROSCOPIC	Description	"MACROSCOPIC"	Contents in associated section context(s)
		Size of specimen		Measurement of the specimen dimensions
		Other lesions		Present, absent, an entry in the uncertainty standard dictionary
	MICROSCOPIC	Description	"MICROSCOPIC"	Contents in associated section context(s)
		Diagnosis		Diagnosis made by the pathologist; for negative diagnosis: "no" + the diagnosis; for uncertain diagnosis: an entry in the uncertainty standard dictionary + the diagnosis
		Tumour thickness		Breslow thickness of the tumour
		Excision margins: Invasive		Numeric value - Distance of invasive melanoma from peripheral margin (and "clear" if the margin is uninvolved by the melanoma)
		Excision margins: In-situ		Numeric value - Distance of in-situ melanoma from peripheral margin (and "clear" if the margin is uninvolved by the melanoma)

		Excision margins: Deep		Numeric value - Distance of the melanoma from deep margin (and "clear" if the margin is uninvolved by the melanoma)
		Ulceration (mm diam)		Present (and measurement of the ulceration if applicable), absent, an entry in the uncertainty standard dictionary
		Mitotic rate		Mitotic rate of the melanoma
		Microsatellites		Present, absent, an entry in the uncertainty standard dictionary
		Level of invasion (Clark)		Classification of Clark level
		Lymphovascular invasion		Present, absent, an entry in the uncertainty standard dictionary
		TILs		Present, absent, an entry in the uncertainty standard dictionary
		TILs: Distribution		Phrase referring to the distribution of TILs
		TILs: Density		Phrase referring to the density of TILs
		Regression		Present (and stage or characteristic if applicable), absent, an entry in the uncertainty standard dictionary
		Desmoplasia		Present, absent, an entry in the uncertainty standard dictionary
		Neurotropism		Present, absent, an entry in the uncertainty standard dictionary
		Assoc. benign naevus		Type of associated naevus, present, absent
		Cell growth		Cell growth pattern of the melanoma
		Subtype		Sub-classification of the melanoma; for negative subtype: "no" + the subtype; for uncertain subtype: an entry in the uncertainty standard dictionary + the subtype

Table 8.1 Structured template of the melanoma corpus. The default possible value is "N/A" (not applicable).

Template section		Template item	Section context	Possible value
Diagnostic Summary		Summary	"CONCLUSION"	Contents in associated section context(s)
		Comment		Texts relating to other issues noted during the pathology reporting in associated section context(s)
Supporting Information	CLINICAL	Site	"CLINICAL HISTORY"	The part of the colorectal tract where the tumour was found by the clinician
		Other sites of disease		Relevant coexistent pathological abnormality
		Medical history		Previous medical history of the patient
	MACROSCOPIC	Specimen type	"MACROSCOPIC"	Surgical procedure or site
		Tissue banking		Yes, no
		Specimen images		Yes, no
		Specimen length		Numeric value - Measured length of resected colorectal tract
		Tumour site		The part of the colorectal tract where the tumour was located found by the pathologist
		Peritoneal reflection		Astride, above, below
		Mesorectal integrity		Complete, nearly complete, incomplete
		Tumour size		Numeric value - Measurement of the maximum dimension of a tumour
		Extramuscular spread		The measured distance of spread beyond the muscularis propria (in mm), the status of tumour border/margin (e.g., "infiltrative", "pushing")
		Tumour description		Description of the tumour
		Overlying serosa		Description of the surrounding serosa
		Perforation		Present, absent, an entry in the uncertainty standard dictionary
		Margins: Proximal		Numeric value - The measured distance between the tumour and the proximal margin in macroscopic examination (and "clear" if the margin is uninvolved by the tumour)

		Margins: Distal		Numeric value - The measured distance between the tumour and the distal margin in macroscopic examination (and “clear” if the margin is uninvolved by the tumour)
		Margins: Radial		Numeric value - The measured distance between the tumour and the radial or circumferential margin in macroscopic examination (and “clear” if the margin is uninvolved by the tumour)
		Lymph nodes		Numeric value - The total number of lymph nodes identified in macroscopic examination
		Metastases		Present, absent, an entry in the uncertainty standard dictionary
		Blocks selected		Description of how the specimen was sliced into sections for testing
		Comment		Texts relating to other issues noted during the pathology reporting in macroscopic examination
	MICROSCOPIC	Histological type (WHO)	“MICROSCOPIC”	The histological type of cancer the tumour represents
		Histological grade		The level of differentiation of the tumour
		Depth of invasion		The depth that the tumour has invaded into the colorectal tissue
		Serosal involvement		Present, absent, an entry in the uncertainty standard dictionary
		Small vessel invasion		Present, absent, an entry in the uncertainty standard dictionary
		Venous invasion		Present, absent, modality, an entry in the uncertainty standard dictionary
		Perineural invasion		Present, absent, modality, an entry in the uncertainty standard dictionary
		TILs		Present, absent, an entry in the uncertainty standard dictionary (and sub-classification of the lymphocytic response, phrase referring to density/distribution/degree if applicable), texts refers to other lymphocytic responses
		Margins: Proximal		Numeric value - The measured distance between the tumour and the proximal margin in microscopic examination (and “clear” if the margin is uninvolved by the tumour)
		Margins: Distal		Numeric value - The measured distance between the tumour and the distal margin in microscopic examination (and “clear” if the margin is uninvolved by the tumour)

		Margins: Radial		Numeric value - The measured distance between the tumour and the radial or circumferential margin in microscopic examination (and "clear" if the margin is uninvolved by the tumour)
		Lymph nodes		Numeric value - The total number of lymph nodes in microscopic examination
		Number involved		Numeric value - The number of extracted lymph nodes which are shown to be malignantly involved
		Distant spread		Present, absent, an entry in the uncertainty standard dictionary
		Response to Rx		Reaction to treatment
		Comment		Texts relating to other issues noted during the pathology in microscopic examination
	ANCILLARY STUDIES	Description	"ANCILLARY"	Supporting tests performed (and their findings)
	SYNTHESIS	TNM stage: T	"CONCLUSION", "SYNOPTIC"	T value
		TNM stage: N		N value
		TNM stage: M		M value
		Stage group		Pathological stage grouping for colorectal cancer
		Residual tumour (R)		R status, description of whether any tumour was left as residual
		Comment	"SYNOPTIC"	Texts relating to other issues noted during the pathology reporting in associated section context(s)

Table 8.2 Structured template of the colorectal cancer corpus. The default possible value is "N/A" (not applicable).

2. The possible value for item “Tumour size” was referred to the maximum dimension of a tumour instead of the maximum diameter of a tumour, as it was stated in the latest Macroscopic Cut-Up Manual for Colorectal tumour from RCPA (RCPA, 2013-2014), although it was traditionally defined as the greatest linear diameter by macroscopic examination (Miller et al., 1985); the pathologists might not indicate which dimension of the tumour was measured, such as “size: 40mm”, “75x50mm” and “38mm from proximal to distal”.
3. There was no specification about the value for item “Extramuscular spread” in the protocol. By seeking advice from the medical consultants, it was decided that the possible value for this item can be the measured distance of tumour spread beyond the muscularis propria or the tumour border configuration, as extramuscular spread is often present when the tumour has an infiltrative border; the tumour border configuration represents an important histomorphological prognostic indicator. As indicated in Koelzer’s work, infiltrative tumour border is associated with poor survival outcome and early disease recurrence of colorectal cancer patients; a “pushing” tumour border frequently occurs in colorectal cancer cases with low risk for nodal and distant metastasis (Koelzer and Lugli, 2014).
4. Not only the T stages were considered as possible values for item “Depth of invasion”, but also the narratives about the definitions of the stages were also taken into account. For example, our system can be draw a conclusion that the maximum degree of local invasion is pT2 from the texts “extending into but not through muscularis propria” and pT3 from the texts “tumour extends to the full thickness of the muscularis propria into the mesocolon” (Edge et al., 2010).
5. There was also no specification about the value for item “TILs” in the protocol. But given the experience of the same item in the structured template of the melanoma corpus, it was assumed that the presence or absence of TILs, density, distribution and degree of the TILs should be reported if applicable. However, the coverage of this item in the colorectal cancer corpus was broader than that in the structured template of the melanoma corpus, which represented the lymphoid host response to the tumour. Based on the analysis of the corpus, it can be classified to four categories: TILs, peritumoural lymphocytes, Crohn’s-like reaction and other. TILs was defined as at least four unequivocal intraepithelial lymphocytes found in a single ×40 field on haematoxylin and eosin-stained slides (Michael-Robinson et al., 2001); peritumoural lymphocytes were considered to be present as a cap or mantle of chronic inflammatory cells at the deepest point of invasive tumour border (Bosman et al., 2010); Crohn’s-like reaction was based on the finding that three or more nodular lymphoid aggregates deep to the advancing tumour margin within a single ×4 field (Graham and Appelman, 1990); the other category includes other lymphocytic reaction responses that cannot be classified to the categories above. The possible values for “TILs” should include the category of the lymphoid host response if applicable as each has its own prognostic impacts on colorectal cancer patients (Ogino et al., 2009).

6. Items “Margins: Other” and “Margins: Donuts” were removed from the sample template, as they were not described in the protocol, thus it was hard to define them based on the protocol; there were few instances about “Margins: Other” in the corpus; donuts from stapling devices do not need to be examined histologically if the tumour is more than 3 cm from the cut end of the main specimen (Cross et al., 1989).
7. Item “TNM stage” was divided into “TNM stage: T”, “TNM stage: N” and “TNM stage: M”, in order to gain a better granularity for the field, and ease the evaluation.
8. Besides the pathologic stages defined in the American Joint Committee on Cancer (AJCC) Cancer Staging Manual, stages defined in other staging systems were also considered to be populated for the item “Stage group”, such as Australian Clinico-Pathological Staging (ACPS) and Dukes classification, if they were available in the reports.
9. The AJCC Cancer Staging Manual (Edge et al., 2010) defined three R codes: R0, R1 and R2 to represent the residual tumour status. But if they cannot be found in a report, the descriptions regarding them would also be considered as the possible values for item “Residual tumour (R)”.

8.2.3 Structured template of the Lymphoma Corpus

The structured template of the lymphoma corpus is presented in Table 8.3. From this table most sections of the template have their own associated section contexts, except that:

- “SUPPLEMENTARY REPORT” was recognized as a frequently occurring supplementary section context if the related information about the fields in the primary section contexts were missing, or the reports lacked the primary section contexts.
- Both “SPECIMEN” and “MACROSCOPIC” were mapped to the “SPECIMEN” section, as they were categorized to the same chapter in Tumours of Haematopoietic and Lymphoid Tissue Structured Reporting Protocol (Norris et al., 2010): “Specimen handling and macroscopic findings”; both of them contained information about how to handle the specimen and gross examination of the specimen.
- “IMMUNOPHENOTYPING”, “CYTOGENETICS” and “MOLECULAR” sections have up to three associated section contexts, as the locations of the results and interpretations of these ancillary studies were unstable: sometimes they might be recorded with other microscopic findings in “MICROSCOPIC”; in some cases, they were recorded separately in “SUPPLEMENTARY REPORT” or “SPECIAL INVESTIGATIONS”; otherwise, they were recorded in multiple section contexts, especially for some complex cases.
- “SUPPLEMENTARY SUMMARY” was integrated with “SUMMARY” to represent “SYNTHESIS”, as a supplement for the primary diagnosis summary.

Most items were almost the same as those in the sample template provided in the protocol, except for some fields:

Template section		Template item	Section context	Possible value
Diagnostic Summary		Summary	"SUMMARY", "SUPPLEMENTARY SUMMARY"	Contents in associated section context(s)
		Comment	"COMMENT"	Contents in associated section context(s)
Supporting Information	CLINICAL	Site and laterality	"CLINICAL HISTORY", "SUPPLEMENTARY REPORT"	Anatomical site
		Presentation		Clinical presentation of the disease
		Indication for biopsy		Primary diagnosis, staging, relapse, assessment of transformation, the failure of another biopsy
		Clinical impression		Clinical diagnosis or differential diagnosis
		Disease extent		Solitary, localised, generalised
		Other sites of disease		Present, absent, an entry in the uncertainty standard dictionary
		Const. symptoms		Constitutional symptom
		Medical history		Previous relevant disease
		Predisposing factors		Previous relevant treatment, immunodeficiency–associated lymphoproliferative disorder, autoimmune disorder, infective agent
	SPECIMEN	Specimen type	"SPECIMEN", "MACROSCOPIC", "SUPPLEMENTARY REPORT"	Surgical procedure, biopsy type
		Size		Measurement of the specimen dimensions
		Received in		Fresh, formalin, saline
		Triage		Frozen section, imprints, cytology, flow cytometry, paraffin section, cytogenetics, molecular laboratory, microbiology laboratory, tissue bank, electron microscopy and macroscopic photography
		Description		Contents in associated section context(s)
	MICROSCOPIC	Pattern of infiltration	"MICROSCOPIC", "SUPPLEMENTARY REPORT"	Diffuse, follicular, marginal zone, mantle zone, interstitial, perivascular, nodular, superficial, deep, angiocentric, lymphoepithelial lesions, proliferation centres
		Cell size		Small, medium, large, mixed, indeterminate
		Cytomorphology		Pleomorphic, hyperlobate, anaplastic, clear cell, giant cell, spindle cell, signet ring cell, blastic, indeterminate, centroblastic, centrocytic, immunoblastic, plasmacytic, lymphoplasmacytic, lymphoplasmacytoid, prolymphocytic, paraimmunoblastic, plasmablastic, monocytoid, centrocyte-like,

				popcorn cell, reed-sternberg cell-like
		Tissue reactions		Host cell or tissue reaction
		Grade		Grade 1, 2, 3, low grade, high grade
		Description		Contents in associated section context(s)
	IMMUNOPHENOTYPING	Immunohistochemistry: Positive for	"MICROSCOPIC", "SUPPLEMENTARY REPORT", "SPECIAL INVESTIGATIONS"	Biomarker
		Immunohistochemistry: Negative for		Biomarker
		Immunohistochemistry: Equivocal for		Biomarker
		Immunohistochemistry: Comment		Interpretive comment of immunohistochemistry tests
		Flow cytometry: Positive for		Biomarker
		Flow cytometry: Negative for		Biomarker
		Flow cytometry: Comment		Interpretive comment of flow cytometry tests
	CYTOGENETICS	FISH		Result of FISH tests
		Cytogenetics: Comment		Interpretive comment of FISH tests
	MOLECULAR	PCR: IgH		Result of PCR analysis with IgH tests
		PCR: TCRgamma		Result of PCR analysis with TCRgamma tests
		PCR: Comment		Interpretive comment of PCR analyses
	SYNTHESIS	Lineage	"SUMMARY", "SUPPLEMENTARY SUMMARY"	B-cell, T-cell, NK-cell, NK/T-cell, histiocytic, dendritic cell, myeloid, Hodgkin-like
		Clonality		Monoclonal, polyclonal
		Diagnosis (WHO)		WHO category of lymphoma or leukaemia (includes subtype or grade if applicable), other relevant haematological disease
		SNOMED RT Codes		SNOMED RT Codes and terms for the diagnosis
		Stage		Pathological stage grouping for the diagnosis
		Comment		Texts relating to other issues noted during the pathology reporting in associated section context(s)

Table 8.3 Structured template of the lymphoma corpus. The default possible value is "N/A" (not applicable).

Template section		Template item	Medical entity type	Linguistic category
Diagnostic Summary		Summary		
		Comment		
Supporting Information	CLINICAL	Description		
		Site and laterality	De:Site and Laterality	
		Clinical diagnosis	Sy:Diagnosis, Sy:Subtype*, En:Associated naevus (type)	Li:Lexical Polarity Positive, Li:Lexical Polarity Negative, Li:Modality
		Specimen type	De:Specimen Type	
		Prev. Rx / Trauma	De:Cosmetic Changes, De:Specimen Type*	Li:Temporality
		Previous melanoma	Sy:Diagnosis	Li:Temporality
		Distant metastasis	Sy:Diagnosis	
		Other medical history	En:Primary Lesion, En:Associated naevus (type), De:Cosmetic Changes, De:Size*, En:Lesion (other)*	Li:Temporality*
	MACROSCOPIC	Description		
		Size of specimen	De:Size	
		Other lesions	En:Lesion (other)	Li:Lexical Polarity Positive, Li:Lexical Polarity Negative, Li:Modality
	MICROSCOPIC	Description		
		Diagnosis	Sy:Diagnosis	Li:Lexical Polarity Positive, Li:Lexical Polarity Negative, Li:Modality
		Tumour thickness	In:Breslow Thickness (mm)	
		Excision margins: Invasive	Ma:Excision Invasive, Ma:Excision Clear	
		Excision margins: In-situ	Ma:Excision In Situ, Ma:Excision Clear	
		Excision margins:	Ma:Excision Deep, Ma:Excision Clear	

		Deep		
		Ulceration (mm diam)	De:Ulceration	
		Mitotic rate	De:Dermal Mitoses	
		Microsatellites	En:Satellites	
		Level of invasion (Clark)	In:Clark Level	
		Lymphovascular invasion	In:Vascular/Lymphatic	Li:Lexical Polarity Positive, Li:Lexical Polarity Negative, Li:Modality
		TILs	Re:TILs	Li:Lexical Polarity Positive, Li:Lexical Polarity Negative, Li:Modality
		TILs: Distribution	Re:TILs	Li:Mood and Comment Adjuncts
		TILs: Density	Re:TILs	Li:Mood and Comment Adjuncts
		Regression	Sy:Regression	Li:Lexical Polarity Positive, Li:Lexical Polarity Negative, Li:Modality, Li:Temporality, Li:Mood and Comment Adjuncts
		Desmoplasia	Re:Desmoplasia	Li:Lexical Polarity Positive, Li:Lexical Polarity Negative, Li:Modality
		Neurotropism	In:Neurotropism	Li:Lexical Polarity Positive, Li:Lexical Polarity Negative, Li:Modality
		Assoc. benign naevus	En:Associated naevus (type)	Li:Lexical Polarity Positive, Li:Lexical Polarity Negative, Li:Modality
		Cell growth	De:Cell Growth Pattern	Li:Lexical Polarity Positive, Li:Lexical Polarity Negative, Li:Modality
		Subtype	Sy:Subtype	Li:Lexical Polarity Positive, Li:Lexical Polarity Negative, Li:Modality

Table 8.4 Mapping strategy for the melanoma corpus. Entity type marked with * suggests it was added after first round error analysis.

Template section		Template item	Medical entity type
Diagnostic Summary		Summary	
		Comment	Sy:Comment , En:Coexistent Pathology, De:Ancillary Studies
Supporting Information	CLINICAL	Site	De:Tumour Site
		Other sites of disease	En:Distant Spread or Metastases*, Ex:Extent*
		Medical history	Sy:Medical History
	MACROSCOPIC	Specimen type	De:Specimen Type
		Tissue banking	De:Tissue Banking
		Specimen images	De:Specimen Images
		Specimen length	De:Specimen Size
		Tumour site	De:Tumour Site
		Peritoneal reflection	De:Peritoneal Reflection
		Mesorectal integrity	De:Mesorectal Integrity
		Tumour size	De:Tumour Size
		Extramuscular spread	Ex:Extramuscular Spread
		Tumour description	De:Tumour Description
		Overlying serosa	De:Serosa Description
		Perforation	De:Perforation
		Margins:Proximal	Ma:Proximal or Distal Margin, Ma:Clear
		Margins:Distal	Ma:Proximal or Distal Margin, Ma:Clear
		Margins:Radial	Ma:Circumferential Margin, Ma:Clear
		Lymph nodes	En:Lymph Nodes
		Metastases	En:Distant Spread or Metastases
		Blocks selected	De:Specimen Blocks
		Comment	Sy:Comment, En:Coexistent Pathology, De:Ancillary Studies

	MICROSCOPIC	Histological type (WHO)	Sy:Histological Type
		Histological grade	Sy:Histological Grade
		Depth of invasion	In:Depth of Invasion
		Serosal involvement	Ex:Serosal Involvement
		Small vessel invasion	In:Venous and Small Vessel Invasion
		Venous invasion	In:Venous and Small Vessel Invasion
		Perineural invasion	In:Perineural Invasion
		TILs	Re:TILS and Peritumoural Lymphocytes
		Margins:Proximal	Ma:Proximal or Distal Margin, Ma:Clear
		Margins:Distal	Ma:Proximal or Distal Margin, Ma:Clear
		Margins:Radial	Ma:Circumferential Margin, Ma:Clear
		Lymph nodes	En:Lymph Nodes, Ex:Lymph Node Involvement
		Number involved	Ex:Lymph Node Involvement
		Distant spread	En:Distant Spread or Metastases
		Response to Rx	Re:Response to Rx
		Comment	Sy:Comment , En:Coexistent Pathology, De:Ancillary Studies
	ANCILLARY STUDIES	Description	De:Ancillary Studies
	SYNTHESIS	TNM stage:T	Met:T Value
		TNM stage:N	Met:N Value
		TNM stage:M	Met:M Value
		Stage group	Met:Anatomic Stage
		Residual tumour (R)	En:Residual Tumour
		Comment	Sy:Comment , En:Coexistent Pathology, De:Ancillary Studies

Table 8.5 Mapping strategy for the colorectal cancer corpus. Entity type marked with * suggests it was added after first round error analysis.

Template section		Template item	Medical entity type	Linguistic category	Relation Type
Diagnostic Summary		Summary			
		Comment			
Supporting	CLINICAL	Site and laterality	De:Topography, De:Anatomical		Spatial

Information			Structure, De:Laterality		Specialization
		Presentation	Sy:Presentation	Li:Temporality	
		Indication for biopsy	Sy:Indication for Biopsy, De:Specimen Type		
		Clinical impression	Sy:Clinical Impression, Sy:Diagnosis Subtype	Li:Lexical Modality	
		Disease extent	Ex:Disease Extent		
		Other sites of disease	Ex:Other Sites of Disease	Li:Lexical Polarity Negative, Li:Lexical Polarity Positive, Li:Lexical Modality	
		Const. symptoms	Sy:Constitutional Symptoms	Li:Lexical Polarity Negative, Li:Lexical Polarity Positive, Li:Lexical Modality	
		Medical history	Sy:Medical History	Li:Lexical Polarity Negative, Li:Lexical Polarity Positive, Li:Lexical Modality	
		Predisposing factors	Sy:Predisposing Factors		
	SPECIMEN	Specimen type	De:Specimen Type		
		Size	De:Specimen Size		
		Received in	De:Preservative Fluid		
		Triage	De:Sample Triage		
		Description			
	MICROSCOPIC	Pattern of infiltration	De:Architecture	Li:Lexical Polarity Negative, Li:Lexical Polarity Positive, Li:Lexical Modality	
		Cell size	De:Cell Size		
		Cytomorphology	De:Cytomorphology	Li:Lexical Polarity Negative, Li:Lexical Polarity Positive, Li:Lexical Modality	
		Tissue reactions	Re:Tissue Reaction	Li:Lexical Polarity Negative, Li:Lexical Polarity Positive, Li:Lexical Modality, Li:Mood and Comment Adjuncts	
		Grade	Sy:WHO Grade	Li:Lexical Polarity Negative,	

				Li:Lexical Polarity Positive, Li: Lexical Modality	
		Description			
	IMMUNOPHENOTYPING	Immunohistochemistry: Positive for	An:Biomarker, An:Immunohistochemistry-Positive		Result-Positive
		Immunohistochemistry: Negative for	An:Biomarker, An:Immunohistochemistry-Negative		Result-Negative
		Immunohistochemistry: Equivocal for	An:Biomarker, An:Immunohistochemistry-Equivocal		Result-Equivocal
		Immunohistochemistry: Comment	An:Immunohistochemistry-Comment		
		Flow cytometry: Positive for	An:Biomarker, An:Flow Cytometry-Positive		Result-Positive
		Flow cytometry: Negative for	An:Biomarker, An:Flow Cytometry-Negative		Result-Negative
		Flow cytometry: Comment	An:Flow Cytometry-Comment		
	CYTOGENETICS	FISH	An:FISH Results		
		Cytogenetics: Comment	An:Cytogenetics Comment		
	MOLECULAR	PCR: IgH	An:IgH Test		
		PCR: TCRgamma	An:TCRgamma Test		
		PCR: Comment	An:PCR Comment		
	SYNTHESIS	Lineage	De:Lineage, Sy:Diagnosis, Sy:Diagnosis Subtype		
		Clonality	De:Cell Clonality		
		Diagnosis (WHO)	Sy:Diagnosis, Coexistent Pathology, Sy:Diagnosis Subtype, Sy:WHO Grade	Li:Lexical Polarity Negative, Li:Lexical Polarity Positive, Li:Lexical Modality	
		SNOMED RT Codes	Sy:SNOMED RT Codes		
		Stage	Sy:Stage		
		Comment	Sy:Comment		

Table 8.6 Mapping strategy for the lymphoma corpus.

1. The "Description" field in both "SPECIMEN" and "MICROSCOPIC" sections would populate all contents in associated section context(s) rather than part of them in case of omission of any relevant information.
2. Item "Classical cytogenetics" was removed from the sample template, as only fluorescence *in situ* hybridization (FISH) tests were performed in the reports.
3. Item "ICD O-3" was replaced with "SNOMED RT Codes", since pathologists tended to use SNOMED Reference Terminology (SNOMED RT) (Spackman et al., 1997) to encode the diagnosis summary in the reports.
4. Possible values for several fields were restricted to the descriptors or examples indicated in the standards or guidelines in the protocol if provided, such as "Cell size" and "Cytomorphology". It is worth pointing out that:
    - Some descriptors or examples were excluded if they did not fit the training data.
    - Additional descriptors or examples were obtained from the training data if they represented a prominent proportion of the data.
    - Some fields could not be utilized with the descriptors even if they were provided in the protocol. For example, as the candidates to be populated for "Tissue reactions" had very high lexical variability, which makes the standardization of them to the specific descriptors quite difficult, thus there was no such restriction on the population of this field.

8.3 Mapping Strategies

8.3.1 Mapping Strategy for the Melanoma Corpus

The mapping strategy for the melanoma corpus is illustrated in Table 8.4. From this table, except for "Summary", "Comment", and "Description", the associated medical entity types and linguistic categories were identified for other fields. Note that since some medical entity types were mapped to multiple fields, during population process it holds that:

- Section context detection is critical for utilizing the medical entity type to populate the field.
- The lexical items of the entity can affect the population of a particular field.
- The medical entity is subject to the co-occurring linguistic category for the population of a field.

For example, if a Sy:Diagnosis entity is found in "MICROSCOPIC", it should be considered as a candidate to populate "Diagnosis"; if it is found in "CLINICAL HISTORY" or "SPECIMEN", and if it contains a lexicon about "metastasis", it is likely to be a value for "Distant metastasis"; if a Li:Temporality entity regarding "past history" co-occurs with it and it contains a lexicon about "melanoma", it is probably a value for "Previous melanoma"; otherwise, it matches to "Clinical diagnosis".

8.3.2 Mapping Strategy for the Colorectal Cancer Corpus

Table 8.5 displays the mapping strategy for the colorectal cancer corpus. In this table, except for "Summary", other fields have their associated medical entity types. There are some medical entity types mapped to multiple fields. For the reasons above, it also implies that they may need a sub-classification process to find the suitable candidate for a particular field. For instance, sub-classifying In:Venous and Small Vessel Invasion entities into "small vessel" and "venous" groups is suitable for the candidates for populating the associated fields "Small vessel invasion" and "Venous invasion". Some fields involve more than one medical entity type, which reveals that related information of these fields are distributed dispersedly, and a complete population of the field should take into account more than one associated medical entity type if applicable. For example, to populate "Lymph nodes" in "MICROSCOPIC", both En:Lymph Nodes and Ex:Lymph Node Involvement entities in "MICROSCOPIC" should be considered to be used to compute the number of lymph nodes identified in the microscopic examination, as pathologists sometimes would not record the number of identified lymph node separately, but record it implicitly as or wrapped with the number of malignantly involved ones.

8.2.3 Mapping Strategy for the Lymphoma Corpus

Table 8.6 presents the mapping strategy for the lymphoma corpus. Not only medical entity types and linguistic categories account for the population of most fields, but also relation types correspond to the population of some fields, such as Result-Positive for "Immunohistochemistry: Positive for" and Result-Negative for "Immunohistochemistry: Negative for". Most medical entity types are correlated to one particular field. There are also exceptions indicating multiple or repetitive roles they play in the fields. For instance, biomarkers are requisites for performing immunohistochemistry tests and flow cytometry, thus An:Biomarker is one of major medical entity types for the population of the result fields in "IMMUNOPHENOTYPING" section. The annotation schema required the annotation of Sy:Diagnosis should consider as long a span as possible to denote a WHO category of the disease, therefore lineage is often subsumed in the span of a Sy:Diagnosis entity. The Sy:Diagnosis entities can be directly populated to "Diagnosis (WHO)", while the lineages inside them need to be stripped from them to populate "Lineage". The reason for fields involving more than one medical entity type is similar to the one above, with a goal to attain complete populations of the fields by considering as many medical entity types as possible.

8.4 Rule-based System for Structured Output Generation

As can be seen from the above sections, the population of most fields requires extraction of very large segments of text in the reports, e.g., all the contents in "DIAGNOSIS" are needed to populate the "Summary" in the structured template of the melanoma corpus; or inferences from the associated

medical entities, e.g., “40mm” is the maximum measured dimension for the De:Tumour Size entity “40x30mm”. A statistical method will not be able to infer them reliably or construct them properly, thus rule-based approaches were used to generate the structured outputs.

Consequently, a rule-based system was established for structured output generation (SOG), including four main processes: document classification, specimen context detection, candidate preparation and extensible mark-up language (XML) generation.

8.4.1 Document Classification

A Document Classifier had inserted into a heuristic rule to classify the documents to multiple specimen/tumour documents (documents containing more than one specimen or tumour) or single specimen/tumour documents (documents containing only one specimen or tumour), based on the entity or subheading recognition results on En:Specimen Identifier from the melanoma corpus and the lymphoma corpus, St:Subheading from the colorectal cancer corpus.

The heuristic rule is to detect whether the identifier (id) has a lexicon that can be converted to a numeric value larger than 1, e.g., “2”, “iii”, “d”, etc; if it has, the document is classified to a multiple specimen/tumour document; else, it is a single specimen/tumour document.

8.4.2 Specimen Context Detection

A context detection engine was built to detect the section context information for each specimen for multiple specimen/tumour documents. Basically, it separates the sections by the positions of the specimen ids, e.g., the text span between the positions of specimen id 1 and specimen id 2 in “MICROSCOPIC” results in “MICROSCOPIC” for specimen id 1. There are also several rules to handle special cases. Here are some examples:

- If specimen id 1 is missing, but specimen id 2 is found in a section, the text span between the section start and the position of specimen id 2 can yield an output as the associated section context for specimen id 1.
- If both specimen id 1 and specimen id 2 are found in “MICROSCOPIC”, but none of them can be detected in “MACROSCOPIC”, it will yield a collective “MACROSCOPIC” for specimen id 1 and specimen id 2 as a result for the text span in “MACROSCOPIC”.

8.4.3 Candidate Preparation

The candidate preparation process was to find appropriate candidates for the population to the templates, implemented with a series of post-processing modules and ranking criteria. Concise descriptions about these post-processing modules and ranking criteria are presented in Tables 8.7 and 8.8 respectively.

Post-processing module	Entity type	Brief description
General process module	Most medical entity types listed in Tables 8.4, 8.5 and 8.6.	This module includes several general pre-processes that can be applied to most fields before ranking and some post-processes to handle candidates after ranking.
Measurement module	De:Ulceration (mm), In:Breslow Thickness (mm), Ma:Excision Deep, Ma:Excision In Situ, Ma:Excision Invasive, Ma:Excision Clear; In:Depth of Invasion, Ex:Extramuscular Spread, Ma:Circumferential Margin, Ma:Proximal or Distal Margin	This module extracts one-dimension size from candidates.
Naevus module	En:Associated Naevus (type)	This module extracts the type of naevus from candidates.
Level module	In: Clark level	This module extracts Roman numeral(s) from candidates. Arabic numeral (s) is converted to Roman numeral(s) if applicable.
Temporality module	Li:Temporality	This module extracts Li:Temporality entities within a context window.
Regress module	Sy:Regression	This module extracts stage or characteristic of regression from candidates.
Rate module	De:Dermal Mitoses	This module extracts mitotic rate from candidates.
Dimension processor	De:Size; De:Tumour Size; De:Specimen Size	This module extracts the size or maximum dimension of a specimen or tumour. It converts measurements in “cm” to “mm” if applicable.
Node number module	En:Lymph Nodes	This module extracts numeric value(s) from candidates.
Involvement number module	Ex:Lymph Node Involvement	This module extracts numeric value(s) from candidates. If no number is found, but negation is detected, then assigns the output to be “0”.
Sub-classification module	Ma:Excision Clear; Ma:Proximal or Distal Margin, Ma:Clear, In:Venous and Small Vessel Invasion, En:Distant Spread or Metastases, Re:TILS and Peritumoural Lymphocytes	This module classifies the candidate to a specific sub-type.
Tils module	Re:TILS, Li:Mood and Comment Adjuncts; Re:TILS and Peritumoural Lymphocytes	This module extracts any lexical items in the distribution gazetteer, density gazetteer and degree gazetteer from the input.
Mood degree module	Li:Mood and Comment Adjuncts	The module returns the Li:Mood and Comment Adjuncts entities around the candidate and their associated scores, which can be used by the mood degree

		criterion or contribute to the population of the associated field.
Subheading module	Multiple entity types listed in Table 8.5	This module verifies whether the candidate is a synoptic field with an associated subheading.
Tumour border status module	Ex:Extramuscular Spread	This module extracts the status of tumour border/ margin from candidates.
Tumour description module	De:Tumour Description	The module verifies whether the candidate matches a certain pattern and extracts the non-prepositional phrase from the candidate.
Clear processor	Ma:Excision Deep, Ma:Excision In Situ, Ma:Excision Invasive; Ma:Circumferential Margin, Ma:Proximal or Distal Margin	This module detects whether lexical items “clear” or “clearance” is in the candidate.
Tumour site processor	De:Tumour Site	This module extracts the anatomical site and laterality from the candidate.
Specimen length processor	De:Specimen Size	The module verifies whether the candidate matches a certain pattern.
Descriptor convertor	Sy:Indication for Biopsy, De:Specimen Type, De:Architecture, De:Cell Size, De:Cytomorphology, De:Lineage	The module standardizes the lexical variants in the candidates of associated entity types to the descriptors defined as the possible values for some fields if applicable.
Id validation module	De:Sample Triage, all entity types in Ancillary category	The module determines the specimen id(s) for the population of fields under some section contexts.
Special candidate selection module	Sy:Diagnosis, Sy:Subtype, En:Associated naevus (type), De:Cosmetic Changes, De:Specimen Type, En:Primary Lesion, De:Size, En:Lesion (other), De:Cell Growth Pattern	The module tackles special cases where the best candidate(s) cannot be determined by ranking.

Table 8.7 Entity types involved and brief descriptions of post-processing modules.

Criterion	Entity type	Condition considered	Possible score	Other module required
Span length criterion	De:Site and Laterality, De:Specimen Type, De:Ulceration (mm), In:Vascular/Lymphatic, En:Associated naevus (type)	Length of text span	0, 1	
Uppercase criterion	De:Site and Laterality, De:Ulceration (mm), In:Vascular/Lymphatic; De:Tumour Site,	Uppercase	-1, 0, 1	
Negation uncertainty inapplicability criterion	Multiple medical entity types listed in Tables 8.4, 8.5 and 8.6.	Assertion	-1, 0, 0.3, 0.5, 0.8, 1, 2	Negation and uncertainty detection modules
Measurement criterion	De:Ulceration (mm), In:Breslow Thickness (mm), Ma:Excision Deep, Ma:Excision In Situ, Ma:Excision Invasive, Ma:Excision Clear; In:Depth of Invasion, Ex:Extramuscular Spread, Ma:Circumferential Margin, Ma:Proximal or Distal Margin	Numeric value	0, 0.5, 0.8, 1	Measurement module
Clear criterion	Ma:Excision Deep, Ma:Excision In Situ, Ma:Excision Invasive	Sentence distance to Ma:Excision Clear entity	0, 1	
Frequency criterion	De:Site and Laterality, Sy:Diagnosis, En:Associated naevus (type); De:Specimen Type	Frequencies of overlapping tokens	Variable scores	
Primary criterion	Ma:Excision Deep, Ma:Excision In Situ, Ma:Excision Invasive	Information about the primary lesion	0, 1	
Temporality criterion	De:Specimen Type, Sy:Regression; De:Topography, De:Anatomical Structure, De:Laterality	Temporality or lexicons regarding it	-1, 0, 1	Temporality module
Body structure criterion	De:Site and Laterality	Medical category in SNOMED CT	0, 1	
Laterality criterion	De:Site and Laterality; De:Topography	Lexical entries in the laterality gazetteer	0, 1	
Melanoma criterion	Sy:Diagnosis	Lexical entries in the melanoma gazetteer	0, 1, 2	
Naevus type criterion	En:Associated naevus (type)	Naevus type	0, 1	Naevus module
Level criterion	In: Clark level	Roman numeral(s)	0, 1	Level module
Regress criterion	Sy:Regression	Adjective, Lexical entries in the regress gazetteer	0, 1	Regress module
Dimension criterion	De:Size	Dimension of the size	0, 1	

Position criterion	De:Size; De:Specimen Size	Position	0, 1	
Specimen distance criterion	De:Size; De:Specimen Size	Sentence distance to De:Specimen type entity or the lexicon “specimen”	0, 1	
Margin criterion	Ma:Excision Clear	Sentence distance to Ma:Excision In Situ, Ma:Excision Invasive, or Ma:Excision Deep entity	0, 1	
Distribution density degree criterion	Re:TILS; Re:TILS and Peritumoural Lymphocytes	Lexical entries in the distribution gazetteer, density gazetteer and degree gazetteer	Variable scores	Tils module
Rate criterion	De:Dermal Mitoses	Mitotic rate	0, 1	Rate module
Acronym criterion	Sy:Diagnosis, Sy:Subtype, Re:TILS	Acronym	-1, 0	
Margin type criterion	Ma:Excision Clear	Margin type	0, 1	Sub-classification module
Invasive criterion	Ma:Excision Invasive	Information about an invasive lesion	0, 1	
Mood degree criterion	En:Associated naevus (type); De:Architecture	Li:Mood and Comment Adjuncts entity	0, 0.5, 1, 1.5, 2, 2.5, 3	Mood degree module
Diagnosis criterion	Sy:Subtype	Token and sentence distance to Sy:Diagnosis entity	0, 1, 2	
Breslow criterion	In:Breslow Thickness (mm)	The lexicon “Breslow”	0, 1	
Specific criterion	De:Specimen Type	Specific biopsy type	0, 1	
Type criterion	Sy:Subtype	Lexical items “type” or “pattern”	0, 1	
Summary criterion	Multiple entity types listed in Table 8.5	Associated subheading	0, 1	Subheading module
Tumour site criterion	De:Tumour Site, De:Specimen Type	Lexical entries in the tumour site gazetteer	Variable scores	
Sub-classification criterion	Ma:Excision Clear; Ma:Proximal or Distal Margin, Ma:Clear, In:Venous and Small Vessel Invasion, En:Distant Spread or Metastases, Re:TILS and Peritumoural Lymphocytes	Specific sub-type	0, 1, 2	Sub-classification module
Maximum measurement criterion	De:Specimen Size, De:Tumour Size	Maximum value	0, 1	Dimension processor
Medical category criterion	De:Specimen Type	Medical category in SNOMED CT	0, 1	
Specimen length	De:Specimen Size	Measured length	0, 1, 2	Dimension processor

criterion				
Size criterion	De:Tumour Size	Measured volume or area	0, 1, 2	Dimension processor
Tumour description criterion	De:Tumour Description	Certain patterns	0, 1, 2	Tumour description module
Location criterion	De:Peritoneal Reflection	Relationship of the tumour to the anterior peritoneal reflection	0, 1	
Involvement number criterion	Ex:Lymph Node Involvement	Numeric value, total count	0, 1, 2	Involvement number module
Node number criterion	En:Lymph Nodes	Numeric value, total count	0, 1, 2	Node number module
T stage criterion	In:Depth of Invasion	Lexical entries in the T stage gazetteer	0, 1	
R status criterion	En:Residual Tumour	Lexical entries in the R status gazetteer	0, 1	
Tumour distance criterion	De:Tumour Size	Sentence distance to lexical items “tumour” or “it”	0, 1	
Procedure criterion	De:Specimen Type	Specific surgical procedure	0, 1	
Integrity criterion	De:Mesorectal Integrity	Intactness of the mesorectum	0, 1	
Tumour boarder status criterion	Ex:Extramuscular Spread	Lexical entries in the tumour boarder status gazetteer	0, 1	Tumour border status module
Depth criterion	In:Depth of Invasion	The lexicon “depth”	0, 1	
Regression grade criterion	Re:Response to Rx	Regression grade	0, 1	
Maximum dimension criterion	De:Tumour Size	Maximum measured dimension	0, 1	
Specimen id criterion	De:Specimen Size, De:Tumour Site	Specimen id’s (or ids’) context, De:Tumour Size or De:Tumour Description entity	0, 1	
Abbreviation criterion	De:Specimen Type	Abbreviation	0, 1	
Revision criterion	Met:Anatomic Stage, Met:M Value, Met:N Value, Met:T Value	Revised classification of the stage	0, 1	
Noun phrase criterion	De:Topography, De:Anatomical Structure	Noun	0, 1	
Grade criterion	Sy:WHO Grade	Expression of the grade	-1, 0, 1	
Total criterion	De:Specimen Size	Total size	0, 1	
Cell size criterion	De:Architecture	Sentence distance to De:Cell Size entity	0, 1	

Architecture criterion	De:Architecture	Lexical entries in the architecture gazetteer	0, 1	
Pattern criterion	De:Architecture	Lexical entries in the pattern gazetteer	0, 1	
Pos criterion	Re:Tissue Reaction	Part-of-speech tag	-1, 0, 1	
Classification criterion	Sy:WHO Grade	Grade from classification systems	-1, 0, 1	
Tissue reaction criterion	De:Cell Size	Token and sentence distance to Re:Tissue Reaction entity	0, 1	
Malignancy criterion	Sy:Medical History	Specific malignant disease	-1, 0	
Addition criterion	De:Specimen Size	Additional size	0, 1	

Table 8.8 Brief descriptions of ranking criteria. These include associated entity types, specific conditions to be considered, possible score to be returned and other modules as prerequisite.

Criterion/Module	Gazetteer	Lexical entry
Temporality module	Regression temporality gazetteer*	late, early, past, current, prior
Regress module	Regress gazetteer*	partly, completely, partial, complete, patchy, minor patchy
Tumour site processor	Tumour gazetteer*	carcinoma, cancer, mass, adenocarcinoma, lesion, neoplasm, carcinomas, polyp, tumour, tumour, tumours
Special candidate selection module	Trauma/treatment gazetteer*	trauma, surgical, incision, excision, biopsy, injured, treatment, graft, therapy
Laterality criterion	Laterality gazetteer*	l, r, (l), (r), left, right, mid, central, lt, rt, (lt), (rt), anterior, bilateral, middle, l., r., central, posterior, medial, upper, lower
Melanoma criterion	Melanoma gazetteer	melanoma, malignant, malignancy, tumour
TILs module	Distribution gazetteer*	band, band-like, extensive, diffuse, peripheral, scattered, focal, patchy, variably distributed, band like, focally
TILs module	Density gazetteer	sparse, dense, heavy, light, heavily, quite dense, moderately dense, low density
TILs module	Degree gazetteer	brisk, prominent, moderate, minimal, marked, mild to moderate, limited, little, mild, moderate to marked, modest, numerous, occasional, scanty, scant, inconspicuous, minor, significant, infrequent, small numbers, abundant, smaller numbers, conspicuous
Tumour site criterion	Tumour site gazetteer*	Preferable terms: colon, sigmoid, rectum, rectal, rectosigmoid, flexure, bowel, caecum, caecal, cecum, ascending, descending, lower, upper, transverse, low, left, right, splenic, hepatic, verge, mid, ileocaecal, recto-sigmoid, appendix Unfavourable terms: colonic, colorectal, wall, dentate, border, part, specimen, mucosa Additional unfavourable terms: margin, donut, donuts, small, liver, nodule, end, node, fallopian, tube, ring, stump

Tumour description module	General tumour gazetteer	tumour, mass, lesion, carcinoma, tumour
Tumour boarder status module	Tumour boarder status gazetteer	infiltrative, circumscribed, infiltrating, pushing, expanding, expansile, serpiginous
Specimen site criterion	Specimen site gazetteer*	Preferable Terms: colon, rectum, sigmoid, caecum, colorectal, colorectum, recto-sigmoid, rectosigmoid, ascending, descending, lower, upper, transverse, low, left, right, mid, large bowel Unfavourable terms: terminal, mesentery, accompanying, attached, appendix, fat, ileum, mesenteric, omentum, small, doughnut, pericolic, perirectal, anus, mesosigmoid, tissue, spleen, lymph, gallbladder, omental, pericolonic, skin, cervix, anal, annulus, vaginal, mesorectum, apron, stump, meso-colon, peri-colorectal, mesocolon, ileo-colic,ileocolic, meso-appendix, mesoappendix, mesocolic, peri-colic, bladder, both, donut, donuts, ring, rings, each, duct, tube, fragment, nodule, short, material, valve, ovary, liver, meso-rectum, end, one, other, smaller, separate, shorter, single, stalk, stoma, stomach, stomal, two, unremarkable, uterus, wedge, separate, structure, part
Architecture criterion	Architecture gazetteer*	diffuse, nodular, follicular, perivascular, angiocentric, deep, follicles, nodules
Pattern criterion	Pattern gazetteer*	Preferable terms: infiltrate, pattern, areas, proliferation, structures, infiltration, process, patterns, fashion, patterns, collections, sheets, formations, architecture Unfavourable terms: effacement, effaced, effaces, altered, normal, loss, effacing
T stage criterion	T stage gazetteer	T0, T1, T2, T3, T4, TX, Tis
R status criterion	R status gazetteer	R0, R1, R2, RO, RX
Mood degree module	Specific mood gazetteer*	some, occasional, several, numerous, rare, few, numbers, number, amounts, amount

Table 8.9 Lexical entries of each gazetteer. Gazetteer marked with * means it was modified after the first round error analysis.

Table 8.9 presents the lexical entries of each gazetteer adopted by some post-processing modules and ranking criteria. The details of each post-processing module and ranking criterion can be referred to Appendix I. The application of the post-processing modules and ranking criteria for the structured fields in each corpus is displayed in Appendix II. Note that general process module is applied to most structured fields except for "Summary", "Description" or "Comment" fields.

8.4.4 XML Generation

The XML generator generates the outputs in XML format with the candidates extracted from the above processes, as XML is one of the accepted standards for representing and distributing structured reports within a clinical environment. To increase semantic interoperability so as to ensure the representation of clinical information to be rich, detailed, and unambiguous, Health Level 7 Clinical Document Architecture (HL7 CDA) (Dolin et al., 2001) will be adopted in future work.

Generally, for most fields with associated medical entity types, if a field involves a post-processing module(s), the value to be populated is the result(s) from the module(s) (Result A); if it involves the negation uncertainty inapplicability criterion, the value to be populated is the result from the negation and uncertainty detection modules (Result B) or integrated with the texts of the candidate(s) after general processing; if both post-processing module(s) and the criterion are involved, the value to be populated is the combination of Result A and Result B; if none of them are involved, the value to be populated is the texts of the candidate(s) after the general processing. For those fields with associated relation types, the results from the relation extraction system should also be considered during population. For example, if an An:Biomarker entity has a Result-Positive relation with an An:Immunohistochemistry-Positive entity, the An:Biomarker entity is populated to the field "Immunohistochemistry: Positive for". More details for the population process are presented in Table 8.1, 8.2 and 8.3.

There are some additional processes to ensure or enhance the quality of the population process as well:

Template Construction

For a single specimen document, the template for each corpus is shown in Section 8.2. However, for a multiple specimen document, the template may need to be modified to facilitate the populating. There are three strategies to construct such a template:

1. The template should be separated by each specimen id, and each subset under an id is a copy of the fields in the template of a single specimen document.
2. No change is made to the template; the value is populated to the associated field as that in the template of a single specimen document without specification of specimen ids.

3. The template is flexible according to the combination of specimen ids detected under the associated section contexts. If the specimen ids are separate under a section, the section of the template can be reported by specimen id; otherwise, the section of the template can be reported by a set of specimen ids.

To find out which strategy is most suitable for the study, each of them was applied to the melanoma, colorectal cancer and lymphoma corpora respectively: Strategy 1 for the melanoma corpus, Strategy 2 for the colorectal cancer corpus and Strategy 3 for the lymphoma corpus.

Subtype Standardization

To follow the standard reporting convention for "Subtype", lexical items "type", "pattern", "component" (and "in" if present) are stripped from the candidates.

Severity Maximization

The severity of the tumour invasion can be revealed by the values in "Tumour thickness" and "Level of invasion (Clark)". Thus, the severity maximization module aims to find the greatest numeric value from the In:Breslow Thickness (mm) candidates and Roman numerals from the In:Clark Level candidates to be populated to the two associated fields.

Node Number Accumulation

As the fields "Lymph nodes" and "Number involved" require populating the total count of the lymph nodes identified or involved, thus a special module is needed to merge the count of each candidate if there are multiple best candidates after ranking.

Take merging the count of the candidates for "Number involved" as an example. Check the candidates against these predicates:

a. Whether they have any numeric value as a measurement;
b. Whether they have a ratio denoted with slash "/";
c. Which one has a smaller count from the Involvement Number module;
d. Whether comma "," or bracket "(" between them or lexical item "these" is inside one of them.

Rule out the candidate matches Predicate a, Predicates c and d, and include the one matches Predicate b. Finally, sum the count from the remaining best candidate(s).

The count merging process for "Lymph nodes" in the "MACROSCOPIC" section is quite simple, which is to sum the count of each candidate, except where the count of the candidate is fuzzy (e.g., a number starts with "<" or ">"), which will be populated directly to the structured report field.

The count merging process for “Lymph nodes” in the “MICROSCOPIC” section is more complicated, as it involves two entity types. For Ex:Lymph Node Involvement candidates, the process is similar to that described above, with slight differences: the involvement number module is replaced with the node number module in Predicate c; lexical item “addition” is added to Predicate d. The sum from the count of these candidates is called the extra count, while the sum from the count of En:Lymph Nodes candidates called the original count. The final count is the original count or the sum of both counts, determined by the following conditions:

- If the original count is smaller than the extra count, the final count is the sum of both counts.
- If the original count is larger or equal to the extra count, the final count is the original count.
- If the original count is larger than the extra count; the extra count is larger than zero; there is only one En:Lymph Nodes candidate and lexical items “additional” or “'further” is inside the candidate, the final count is the sum of both counts.
- If the two counts are in different specimen ids’ contexts, the final count is also the sum of both counts.

Convention Configuration

After the first round error analysis, an apparent issue arose that the population of some fields did not follow the standard conventions, especially the fields involving multiple entity types, which caused a considerable number of errors. A convention configuration module was developed to resolve this issue. For example, for “Diagnosis (WHO)” in the lymphoma corpus, the reporting convention is defined as

- The combination of the candidates in each associated section context should be in the following order: Sy:Diagnosis, Sy:Diagnosis Subtype, Sy:WHO Grade, and En:Coexistent Pathology.
- The result in “SUMMARY” is called primary diagnosis, and the one in “SUPPLEMENTARY SUMMARY” is called supplementary diagnosis. Like the overlapping candidate reduction process, the repetitive or less informative diagnosis should be removed. If both diagnoses are present, a prefix “Primary:” is added to primary diagnosis, while “Supplementary:” is added to supplementary diagnosis.

Given some sample input, output examples of the final values for the population of some structured fields are presented in Appendix III. Note that the effects of the id validation module and template construction are not reported in these tables, as they involve the detection of global contexts in a document.

8.5 Results

The performance of the system was measured by the standard Precision, Recall and F-score metrics.

Firstly, an initial evaluation was performed on the structured outputs to find out the competence of the rules. Then another evaluation was carried out to reflect the improvement by the refinement of the rules. Tables 8.10, 8.11 and 8.12 show the results from the Structured Output Generation (SOG) system of first and second round evaluations on the training sets in each corpus.

Field	Number	First round evaluation			Second round evaluation		
		Precision	Recall	F-score	Precision	Recall	F-score
Assoc. benign naevus	437	79.03%	83.76%	81.33%	100.00%	98.32%	99.15%
Cell growth	437	52.63%	53.40%	53.01%	96.64%	97.05%	96.84%
Clinical diagnosis	437	73.94%	72.19%	73.05%	97.73%	96.63%	97.18%
Desmoplasia	437	76.47%	86.67%	81.25%	93.33%	93.33%	93.33%
Diagnosis	437	95.01%	92.03%	93.50%	99.28%	99.04%	99.16%
Distant metastasis	437	68.42%	68.42%	68.42%	94.74%	94.74%	94.74%
Excision margins: Deep	437	99.15%	84.06%	90.98%	100.00%	98.56%	99.28%
Excision margins: In-situ	437	85.19%	85.19%	85.19%	98.31%	100.00%	99.15%
Excision margins: Invasive	437	94.78%	90.08%	92.37%	96.39%	97.17%	96.77%
Level of invasion (Clark)	437	99.09%	98.20%	98.65%	100.00%	99.70%	99.85%
Lymphovascular invasion	437	97.29%	97.29%	97.29%	99.55%	99.55%	99.55%
Microsatellites	437	95.83%	95.83%	95.83%	100.00%	100.00%	100.00%
Mitotic rate	437	98.88%	98.32%	98.60%	100.00%	100.00%	100.00%
Neurotropism	437	97.92%	97.92%	97.92%	100.00%	100.00%	100.00%
Other lesions	437	91.18%	93.94%	92.54%	100.00%	100.00%	100.00%
Other medical history	437	74.39%	69.71%	71.98%	90.96%	96.07%	93.44%
Prev. Rx/Trauma	437	66.67%	10.00%	17.39%	84.21%	88.89%	86.49%
Previous melanoma	437	85.71%	100.00%	92.31%	100.00%	100.00%	100.00%
Regression	437	89.63%	81.21%	85.21%	96.69%	97.33%	97.01%
Site and laterality	437	96.66%	94.00%	95.31%	98.74%	98.49%	98.62%
Size of specimen	437	94.87%	93.95%	94.40%	99.03%	98.31%	98.67%
Specimen type	437	95.44%	95.22%	95.33%	98.09%	97.62%	97.86%
Subtype	437	65.08%	82.00%	72.57%	96.98%	97.35%	97.16%
TILs	437	97.47%	99.48%	98.47%	99.48%	99.48%	99.48%
TILs: Density	437	100.00%	100.00%	100.00%	100.00%	100.00%	100.00%
TILs: Distribution	437	96.43%	84.38%	90.00%	96.77%	96.77%	96.77%
Tumour thickness	437	99.71%	97.73%	98.71%	99.72%	99.72%	99.72%
Ulceration(mm diam)	437	95.30%	96.54%	95.91%	99.13%	99.13%	99.13%
Overall	12236	90.73%	89.77%	90.25%	98.36%	98.42%	98.39%

Table 8.10 Scores from structured output generation system of first and second round evaluations on the melanoma training set.

Field	Number	First round evaluation			Second round evaluation		
		Precision	Recall	F-score	Precision	Recall	F-score
Ancillary Studies	397	98.28%	96.61%	97.44%	100.00%	98.31%	99.15%
Blocks selected	397	97.18%	98.70%	97.93%	100.00%	99.74%	99.87%
Comment (DIAGNOSTIC)	397	96.07%	97.16%	96.61%	99.44%	100.00%	99.72%

Comment (MACROSCOPIC)	397	99.17%	99.17%	99.17%	100.00%	99.45%	99.72%
Comment (MICROSCOPIC)	397	99.31%	98.30%	98.80%	100.00%	99.32%	99.66%
Comment (SYNTHESIS)	397	100.00%	99.23%	99.61%	100.00%	100.00%	100.00%
Depth of invasion	397	98.10%	96.56%	97.32%	99.06%	97.84%	98.45%
Distant spread	397	93.06%	93.06%	93.06%	97.22%	100.00%	98.59%
Extramuscular spread	397	100.00%	100.00%	100.00%	100.00%	100.00%	100.00%
Histological grade	397	99.61%	100.00%	99.81%	100.00%	100.00%	100.00%
Histological type (WHO)	397	99.67%	100.00%	99.83%	99.67%	100.00%	99.83%
Lymph nodes (MACROSCOPIC)	397	80.30%	89.08%	84.46%	99.16%	99.16%	99.16%
Lymph nodes (MICROSCOPIC)	397	97.87%	97.87%	97.87%	99.39%	99.39%	99.39%
Margins: Distal (MACROSCOPIC)	397	99.71%	97.99%	98.84%	99.71%	99.14%	99.42%
Margins: Distal (MICROSCOPIC)	397	100.00%	98.52%	99.25%	99.63%	100.00%	99.81%
Margins: Proximal (MACROSCOPIC)	397	99.07%	95.96%	97.49%	99.10%	98.65%	98.88%
Margins: Proximal (MICROSCOPIC)	397	99.62%	99.25%	99.44%	99.63%	100.00%	99.81%
Margins: Radial (MACROSCOPIC)	397	100.00%	91.67%	95.65%	100.00%	100.00%	100.00%
Margins: Radial (MICROSCOPIC)	397	100.00%	90.50%	95.01%	99.50%	99.50%	99.50%
Medical history	397	87.36%	88.37%	87.86%	100.00%	100.00%	100.00%
Mesorectal integrity	397	100.00%	100.00%	100.00%	100.00%	100.00%	100.00%
Metastases	397	100.00%	100.00%	100.00%	100.00%	100.00%	100.00%
Number involved	397	99.70%	99.70%	99.70%	99.70%	99.70%	99.70%
Other sites of disease	397	0.00%	0.00%	0.00%	93.02%	100.00%	96.39%
Overlying serosa	397	94.81%	99.22%	96.97%	99.26%	99.26%	99.26%
Perforation	397	100.00%	97.44%	98.70%	100.00%	97.44%	98.70%
Perineural invasion	397	99.43%	99.43%	99.43%	100.00%	100.00%	100.00%
Peritoneal reflection	397	96.97%	96.97%	96.97%	100.00%	100.00%	100.00%
Residual tumour (R)	397	100.00%	98.41%	99.20%	100.00%	98.41%	99.20%
Response to Rx	397	100.00%	100.00%	100.00%	100.00%	100.00%	100.00%
Serosal Involvement	397	99.18%	100.00%	99.59%	99.18%	100.00%	99.59%
Site	397	82.31%	83.16%	82.74%	100.00%	100.00%	100.00%
Small vessel invasion	397	98.63%	99.31%	98.97%	99.32%	100.00%	99.66%
Specimen images	397	100.00%	100.00%	100.00%	100.00%	100.00%	100.00%
Specimen length	397	57.61%	95.78%	71.95%	93.39%	95.36%	94.36%
Specimen type	397	84.66%	96.68%	90.27%	97.81%	97.01%	97.41%
Stage Group	397	99.53%	99.53%	99.53%	100.00%	100.00%	100.00%
TILs	397	100.00%	100.00%	100.00%	100.00%	100.00%	100.00%
TNM stage: M	397	99.11%	99.11%	99.11%	100.00%	99.56%	99.78%
TNM stage: N	397	98.83%	99.41%	99.12%	99.42%	100.00%	99.71%
TNM stage: T	397	98.26%	99.12%	98.69%	98.84%	100.00%	99.42%
Tissue banking	397	100.00%	100.00%	100.00%	100.00%	100.00%	100.00%

Tumour description	397	98.12%	98.38%	98.25%	98.40%	99.46%	98.93%
Tumour site	397	90.68%	94.27%	92.44%	97.84%	97.00%	97.41%
Tumour size	397	98.95%	98.69%	98.82%	99.21%	99.21%	99.21%
Venous invasion	397	96.48%	97.51%	96.99%	98.58%	98.93%	98.76%
Overall	18262	94.88%	96.96%	95.91%	99.21%	99.28%	99.24%

Table 8.11 Scores from structured output generation system of first and second round evaluations on the colorectal cancer training set.

Field	Number	First round evaluation			Second round evaluation		
		Precision	Recall	F-score	Precision	Recall	F-score
Cell size	321	93.15%	95.33%	94.23%	94.59%	96.77%	95.67%
Clinical impression	246	98.45%	92.03%	95.13%	100.00%	98.55%	99.27%
Comment	298	100.00%	100.00%	100.00%	100.00%	100.00%	100.00%
Const symptoms	246	21.88%	77.78%	34.15%	100.00%	100.00%	100.00%
Cytogenetics comment	294	37.50%	37.50%	37.50%	50.00%	80.00%	61.54%
Cytomorphology	321	90.68%	91.45%	91.06%	92.86%	99.15%	95.90%
Diagnosis (WHO)	298	97.57%	96.23%	96.90%	99.32%	99.66%	99.49%
Disease extent	246	100.00%	100.00%	100.00%	100.00%	100.00%	100.00%
FISH	294	60.00%	66.67%	63.16%	66.67%	85.71%	75.00%
Flow cytometry: Comment	321	95.31%	98.39%	96.83%	95.24%	96.77%	96.00%
Flow cytometry: Negative for	321	0.00%	0.00%	0.00%	40.00%	66.67%	50.00%
Flow cytometry: Positive for	321	0.00%	0.00%	0.00%	40.00%	66.67%	50.00%
Grade	321	85.51%	95.16%	90.08%	91.30%	98.44%	94.74%
IgH	232	80.00%	66.67%	72.73%	100.00%	100.00%	100.00%
Immunohistochemistry: Comment	321	91.94%	93.44%	92.68%	97.60%	98.39%	97.99%
Immunohistochemistry: Equivocal for	321	86.21%	96.15%	90.91%	89.29%	96.15%	92.59%
Immunohistochemistry: Negative for	321	97.78%	92.63%	95.14%	98.43%	99.47%	98.95%
Immunohistochemistry: Positive for	321	93.63%	89.25%	91.39%	97.30%	99.08%	98.18%
Indication for biopsy	246	100.00%	84.00%	91.30%	100.00%	100.00%	100.00%
Lineage	298	99.36%	98.11%	98.73%	99.38%	100.00%	99.69%
Medical history	246	97.10%	97.10%	97.10%	98.59%	100.00%	99.29%
Other sites of disease	246	81.82%	81.82%	81.82%	100.00%	100.00%	100.00%
PCR comment	232	80.00%	66.67%	72.73%	100.00%	100.00%	100.00%
Pattern of infiltration	321	88.41%	81.46%	84.80%	93.22%	93.75%	93.48%
Predisposing factors	246	86.21%	100.00%	92.59%	96.43%	100.00%	98.18%
Presentation	246	90.20%	66.67%	76.67%	100.00	95.71%	97.81%

					%		
Received in	371	98.65%	100.00 %	99.32%	98.65%	100.00 %	99.32%
SNOMED RT codes	227	100.00 %	100.00 %	100.00 %	100.00 %	100.00 %	100.00 %
Site and laterality	246	90.45%	95.67%	92.99%	93.67%	96.73%	95.17%
Specimen size	371	98.04%	96.16%	97.10%	99.45%	98.63%	99.04%
Specimen type	371	93.61%	94.91%	94.25%	94.12%	95.41%	94.76%
Stage	298	100.00 %	100.00 %	100.00 %	100.00 %	100.00 %	100.00 %
TCRgamma	232	50.00%	33.33%	40.00%	100.00 %	100.00 %	100.00 %
Tissue reactions	321	92.31%	78.69%	84.96%	93.55%	97.48%	95.47%
Triage	371	92.57%	98.42%	95.41%	98.48%	98.98%	98.73%
Overall	10253	93.57%	93.01%	93.29%	96.71%	98.17%	97.44%

Table 8.12 Scores from structured output generation system of first and second round evaluations on the lymphoma training set. Note that score for "Clonality" is not presented, as there is no valid sample in the training set.

From Tables 8.10, 8.11 and 8.12, the melanoma training set attained the biggest improvement by about 8.1%, while the smallest improvement was achieved by the colorectal cancer training set. This is probably because the performance on the first round evaluation was relatively low on the melanoma training set (90.25%) and that was very high on the colorectal cancer training set (95.91%). Finally, the rule-based system obtained over 97% F-score on all training sets in the second round evaluation, suggesting the rules worked well on the training sets.

Each error was manually inspected and summarized into several categories: incorrect annotations, errors from other processing engines, errors from mapping strategies, weaknesses in ranking criteria, weaknesses in post-processing modules, errors from negation and uncertainty detection, insufficiency of ranking criteria, insufficiency of post-processing modules, inappropriate application of ranking criteria or post-processing modules, usability problems and other errors.

8.5.1 First Round Evaluation on the Training Sets

In total, 714, 598 and 417 errors were identified in the first round evaluation on the melanoma, colorectal cancer and lymphoma training sets respectively. Table 8.13 presents the distribution of the errors in each category in the first round evaluation. From table 8.13, most errors in the melanoma training set were due to usability problems, incorrect annotations and weaknesses in post-processing modules; insufficiency of post-processing modules and weaknesses in ranking criteria accounted for most errors in the colorectal cancer training set; the majority of the errors in the lymphoma training set were due to incorrect annotations, weaknesses in ranking criteria, weaknesses in post-processing modules and errors from other processing engines.

Usability Problems

A standard language convention and format is indispensible for data analysis. Inappropriate language usage in some fields (e.g., "Clinical diagnosis" and "Subtype") is the main reason for lower performance of those fields in the melanoma training set. For instance, "superficial spreading" is the standard phrasal convention rather than "superficial spreading type" as the value for "Subtype" from the example "there is an intraepidermal element ["En:Primary Lesion"] of superficial spreading type ["Sy:Subtype"]". Likewise, a potential value "caecal" or "rectal" for "Site" or "Tumour site" in the colorectal cancer training set should be standardized to "caecum" or "rectum". Lack of this standardization is the major cause for lower F-score of "Site".

Error category	Melanoma training set (N=714)	Colorectal cancer training set (N=598)	Lymphoma training set (N=417)
Incorrect annotations	**22.13%**	4.52%	**21.34%**
Errors from other processing engines	3.50%	0.67%	17.51%
Errors from mapping strategies	6.02%	12.04%	0.00%
Weaknesses in ranking criteria	2.94%	**15.38%**	**19.18%**
Weaknesses in post-processing modules	16.67%	11.04%	17.99%
Errors from negation and uncertainty detection	8.12%	2.01%	5.28%
Insufficiency of ranking criteria	3.36%	1.84%	2.64%
Insufficiency of post-processing modules	6.86%	**33.78%**	4.08%
Inappropriate application of ranking criteria or post-processing modules	2.24%	3.85%	0.96%
Usability problems	**25.21%**	12.71%	1.92%
Other	2.94%	2.17%	9.11%

Table 8.13 Distribution of the errors in each category in the first round evaluation. The two largest error sources are shown in bold.

Another usability problem is misspelling correction and abbreviation/acronym expansion. For example, "kxcision" is needed to be corrected to "excision" for "Specimen type"; "t/v" is needed to be expanded to "transverse" for "Site". This was addressed before the second round evaluation.
After the first round evaluation, subtype standardization, severity maximization and convention configuration modules were developed for the purpose of improving the usability of the system.

Incorrect Annotations

Although during reflexive validation, most annotation errors have been identified and corrected, there were still a few errors detected in the first round evaluation, accounting for the lower F-scores of "Cell growth" and "Prev. Rx/Trauma" in the melanoma training set, and "Const symptoms" and "Presentation" in the lymphoma training set. For example, "nausea", "vomiting", "haematuria" and

"lethargy" should be annotated as Sy:Constitutional Symptoms entities instead of Sy:Presentation entities, according to the advice of medical consultants.

Insufficiency of Post-processing Modules

This is a notable issue for the colorectal cancer training set, especially for "Specimen length". This is probably because:

1. The specification of this field is not very clear in the protocol, which doesn't indicate whether the entire length of the specimen should be recorded or length of each segment in the specimen should be recorded. To simplify the problem, it was decided to only populate the entire length of the specimen (for single specimen documents) or the length of the main specimen(s) (for multiple specimen documents) in this study.
2. Some pathologists tended to report the lengths for each segment of the specimen rather than the length of the entire specimen, which can complicate the computation of the length. An example is shown below:

   Specimen Dimensions ["St:Subheading"]
   Colon - 270mm long ["De:Specimen Size"].
   Mesentery - 230 x 60mm ["De:Specimen Size"].
   Ileum - 120mm ["De:Specimen Size"].
   Appendix - 95mm long and 5mm wide ["De:Specimen Size"].
3. In some cases, the description of the measurement of a specimen or segment was quite ambiguous (e.g., "caecum and ascending colon measuring 80mm"), which did not always indicate it is a measurement of length.

As a result, a specific post-processing module named Specimen Length Processor was designed to cope with this issue after the first round evaluation.

The Margins Clear Processor was prepared to fix some errors on the three and six fields describing excision margins in the melanoma and colorectal cancer training sets respectively: "Excision margins: Deep", "Excision margins: In-situ" , "Excision margins: Invasive" , "Margins:Distal (MACROSCOPIC)", "Margins:Distal (MICROSCOPIC)", "Margins:Proximal (MACROSCOPIC)", "Margins:Proximal (MICROSCOPIC)", "Margins:Radial (MACROSCOPIC)" and "Margins:Radial (MICROSCOPIC)" after the first round evaluation.

Weaknesses in Ranking Criteria

The weaknesses in the ranking criteria consist of: the conditions for applying the criteria, lexical entries in the associated gazetteers of the criteria, and incorrect scores assigned for the criteria. For instance, in the original medical category criterion, both "Body structure" and "Procedure" were the preferred SNOMED CT categories. This brought several errors in populating values for "Specimen type". Both "Rectosigmoid colon" and "anterior resection" were populated from this example:

> 1. ["St:Subheading"] Labelled - Rectosigmoid colon ["De:Specimen Type"]: The anterior resection ["De:Specimen Type"] specimen….

By excluding "Body structure" from the criterion, only "anterior resection" would be populated, which is the correct value for the field.

Another example is the lexical entries in the pattern gazetteer of the pattern criterion. Adding the lexicon "collections" to the gazetteer, so that the system can yield "nodular" besides "diffuse" for "Pattern of infiltration" from this example:

> The sections show ["Li:Lexical Polarity Positive"] a mass composed of nodular collections of lymphocytes ["De:Architecture"] separated by densely sclerotic, hyalinised stroma ["Re:Tissue Reaction"] as well as more diffuse areas ["De:Architecture"] of atypical lymphocytes in a sclerotic stroma ["Re:Tissue Reaction"].

Incorrect weighting of the candidates with assertions of present and absent in the negation uncertainty inapplicability criterion resulted in a few errors as well. For example, "no malignancy" was produced as the value for "Diagnosis" rather than "dysplastic junctional naevus" in this example:

> This is a dysplastic junctional naevus ["Sy:Diagnosis"]. No ["Li:Lexical Polarity Negative"] evidence ["Li:Lexical Polarity Positive"] of malignancy ["Sy:Diagnosis"] seen.

Adjusting the weights in the criterion was one of the improvements after the first round evaluation.

Weaknesses in Post-processing Modules

In the initial design of the post-processing modules, some useful information failed to be captured from the candidates. For example, "preexisting" is occasionally missed for populating "Assoc. benign naevus" in examples like "A preexisting benign dysplastic naevus ["En:Associated naevus (type)"] is noted ["Li:Lexical Polarity Positive"]." The possible reason is that the results from the GENIA tagger were used to determine the boundary of a noun phrase in a En:Associated naevus (type) candidate; in some cases, the tagger would tag "preexisting" as "VBG", which is out of the scope of the noun phrase, hence it could not be populated correctly.

The Dimension Processor initially could only generate a value "15x0.5mm" from a Specimen Size candidate "12 and 15mm in length and up to 0.5mm in diameter"; by modifying the extraction and combination rules in the module, it could also generate another value "12x0.5mm" from the candidate. The order for choosing candidates in different section contexts were also be adjusted, especially for those to be populated to the fields under "IMMUNOPHENOTYPING", "CYTOGENETICS" and "MOLECULAR" sections in lymphoma corpus, as the locations of the candidates for these fields are very flexible, all of the associated section contexts should be considered at the same time rather than sequentially.

Errors from Other Processing Engines

The Errors from other processing engines, such as specimen context detection, section detection, and sentence boundary detection.

Incorrect results from specimen context detection could consequently affect the construction of the structured template. This caused up to 55 errors in a single record on the lymphoma training set, wherein 14 were false negatives that underreported for a set of specimen ids, while the others were false positives that over-generated for each specimen id.

All characters in the reports were assumed to be encoded in "utf-8". However, due to mistakes of scanning or OCR, some characters were actually not encoded in "utf-8", which led to several invalid outputs. An extra pre-process was required to resolve this issue.

Errors from Mapping Strategies

In the initial mapping strategies, the incorrect or insufficient mapping of entity types to the associated fields was the major cause for the drop of F-scores on "Other medical history" in the melanoma training set and "Other sites of disease" in the colorectal cancer training set. For example, initially, only En:Primary Lesion, En:Associated naevus (type) and De:Cosmetic Changes were mapped to "Other medical history", which led to occasional omission of some important information such as the history or duration of change (derived from Li:Temporality), and size of lesion (derived from De:Size); En:Coexistent Pathology was mapped incorrectly to "Other sites of disease" , which yielded no gain on the field. The mapping strategies were modified before second round evaluation (see Tables 8.4 and 8.5 for more details).

Errors from Negation and Uncertainty Detection

The results from the negation and uncertainty detection modules applied the negation uncertainty inapplicability criterion to determine the assertions of the candidates. First round evaluation revealed several problems in these modules, such as

- Deficient lexical entries of terminal terms led to the failure to delimitate the negation or uncertainty scope.
- The keywords and rules used to validate the cue and the scope would not work well for some special cases.

Most of these issues were resolved before the second round evaluation.

Insufficiency of Ranking Criteria

The first round evaluation also revealed that some fields might need additional ranking criteria to refine the candidates. For instance, a specific ranking criterion was designed to assign scores for

particular biopsy types (e.g., punch biopsy and wide excision), so that both "punch biopsy" and "ellipse of skin" for "Specimen type" can be populated from the example:

> The specimen ["En:Specimen Identifier"] consists of a punch biopsy ["De:Specimen Type"] of skin 6 x 9mm ["De:Size"] bearing a pigmented ["De:Shape"] lesion ["En:Primary Lesion"] 3mm in diameter ["De:Size"] and an ellipse of skin ["De:Specimen Type"] measuring 14 x 12 x 3mm ["De:Size"].

For the same reason, an acronym criterion was prepared to decrease the weight of a Sy:Subtype candidate "HMF", such that it can be ruled out to be the best candidate for "Subtype" from the example:

> CLARK LEVEL 3 ["In:Clark Level"] AND WHICH IS ARISING FROM A LENTIGO MALIGNA ["Sy:Subtype"] (HMF ["Sy:Subtype"]) WITH THE LATTER REACHING INTO EACH LATERAL EDGE.

Inappropriate Application of Ranking Criteria or Post-processing Modules

Inappropriate application of ranking criteria or post-processing modules includes deficient application or misuse of ranking criteria or post-processing modules. For example, due to the lack of application of a special candidate selection module (which restricts the assertion of a valid candidate), the false positive output "pre-existing dysplastic naevus" was populated to "Assoc. benign naevus" from the example:

> Sections ["En:Specimen Identifier"] show ["Li:Lexical Polarity Positive"] an invasive malignant melanoma ["Sy:Diagnosis"], possibly ["Li:Modality"] arising in a pre-existing dysplastic naevus ["En:Associated Naevus (type)"].

Initially, a site ranking criterion was applied to assign a score for the De:Specimen Size candidates with the same lexical items as De:Tumour Site items. Thus, "100mm" was populated to "Specimen length", instead of "205mm" from the example:

> An anterior resection specimen ["De:Specimen Type"] of total length 205mm in length ["De:Specimen Size"], consisting of 105mm length of sigmoid ["De:Specimen Size"] and 100mm length of rectum ["De:Specimen Size"].

This criterion was removed after the first round evaluation.

Other Errors

Other errors consist of missing specimen id(s) in the original reports, abnormal grammatical structures, irregular language usage, etc.

For example, as a specimen id "3" was missed in the following example:

> 2. ["En:Specimen Identifier"] R flank ["De:Site and Laterality"] ? ["Li:Modality"] acanthoma ["Sy:Diagnosis"]. Biopsy ["De:Specimen Type"] lesion ["En:Primary Lesion"]

central T spine ["De:Site and Laterality"]. ? ["Li:Modality"] Sup spreading BCC ["Sy:Diagnosis"].

false outputs for "Clinical diagnosis" and "Other medical history" were generated under specimen ids "2" and "3" respectively.

The unit "mm" was omitted in the example "There is a brown, circumferential and stenosing tumour ["De:Tumour Description"] 35 in length ["De:Tumour Size"] and 20mm circumference ["De:Tumour Size"]…", hence an incorrect value "20mm" rather than 35mm was produced for "Tumour size".

Inconsistent use of specimen ids in different sections seemed to be a prominent issue for the lymphoma training set. Here is an example:

In "SPECIMEN" section, the pathologist used "2" and "3" as specimen ids:

2. Right axillary lymph nodes ["De:Topography"] x 2 (SNB ["De:Specimen Type"] x 2).
3. Left intermuscular space lymph nodes ["De:Topography"] x 3 (SNB ["De:Specimen Type"] x 3).

However, in other sections (e.g., "MACROSCOPIC" and "MICROSCOPIC"), specimen ids "2" and "3" were used to refer to the two specimens descried under specimen id 2 in "SPECIMEN"; "4", "5" and "6" were used to represent the three specimens mentioned under specimen id 3 in "SPECIMEN".

Such arbitrary use of specimen ids had negative influence on specimen context detection, and consequently resulted in the errors of the fields under these specimen ids.

8.5.2 Second Round Evaluation on the Training Sets

After the first round evaluation, several measures were taken to resolve the issues above, such as corrected annotations, revised mapping strategies, modified ranking criteria and post-processing modules, the design of additional ranking criteria and post-processing modules, and representation of outputs to follow standard convention and format. Most measures have been illuminated in detail in Section 8.4.3. The performance of revised negation and uncertainty detection modules are displayed in Chapter 6 Section 6.4.

The second round evaluation revealed that a great number of errors had been amended, and the total amount of errors reduced dramatically to 118, 124 and 168 on the melanoma, colorectal cancer and lymphoma training sets respectively. From Tables 8.10, 8.11 and 8.12, the performances for most fields were quite good (with F-scores over 90%), except for "Prev. Rx/Trauma", "Cytogenetics comment", "FISH", "Flow cytometry: Negative for" and "Flow cytometry: Positive for", probably owing to their small sample sizes (none exceeded 20). The distribution of the errors in each category in the second round evaluation is displayed in Table 8.14.

Error category	Melanoma training set (N=118)	Colorectal cancer training set (N=124)	Lymphoma training set (N=168)
Errors from other processing engines	4.24%	2.42%	**36.31%**
Weaknesses in ranking criteria	**22.03%**	**30.65%**	17.86%
Weaknesses in post-processing modules	**21.19%**	**46.77%**	13.69%
Errors from negation and uncertainty detection	11.86%	0.81%	4.17%
Insufficiency of ranking criteria	3.39%	3.23%	1.79%
Insufficiency of post-processing modules	5.93%	4.84%	2.98%
Inappropriate application of ranking criteria or post-processing modules	1.69%	0.00%	0.00%
Usability problems	11.02%	0.81%	0.60%
Other	18.64%	10.48%	**22.62%**

Table 8.14 Distribution of the errors in each category in the second round evaluation. The two largest error sources are shown in bold.

From Table 8.14, the modification or augmentation of the ranking criteria and post-processing modules improved the performance significantly, but might also have had some adverse effects on the system, which is the possible reason for a large proportion of errors still being categorized as weaknesses in them.

Although there were a considerable amount of specimen context detection errors unfixed on the lymphoma training set, they were clustered in one document. Considering the adjustment of the specimen detection engine to one special document might affect the generality of the engine, the engine was not modified.

The errors in the Other category remained, as they were poor-writing of the original reports rather than defects of the system, which were thought to be too difficult to resolve at present.

8.5.3 End-to-End Evaluation on the Test Sets

Finally, the best models for medical entity recognition (MER) and relation extraction (described in Chapters 5 and 7) were utilized to predict the test sets, and then generated the structured outputs with the above refined rule-base system. Tables 8.15, 8.16 and 8.17 display the results of end-to-end evaluation on the melanoma, colorectal cancer and lymphoma test sets respectively. From Tables 8.15, 8.16 and 8.17, the best performance was on the melanoma test set (86.49% F-score), and the worst on the lymphoma test set (78.90% F-score). One possible reason is that the test sample sizes for most fields in the lymphoma test set were much smaller than those in the melanoma test set (the smallest was 57), which might lead to poorer scalability; in addition, some fields (up to 3), such as "IgH" and "TCRgamma", could not be evaluated in the lymphoma test set owing to lack of test samples.

Field	Number	TP	FP	FN	Precision	Recall	F-score
Assoc. benign naevus	108	27	5	5	84.38%	84.38%	84.38%
Cell growth	108	45	9	16	83.33%	73.77%	78.26%
Clinical diagnosis	108	44	4	10	91.67%	81.48%	86.27%
Desmoplasia	108	3	1	4	75.00%	42.86%	54.55%
Diagnosis	108	94	4	10	95.92%	90.38%	93.07%
Distant metastasis	108	1	1	4	50.00%	20.00%	28.57%
Excision margins: Deep	108	23	3	7	88.46%	76.67%	82.14%
Excision margins: In-situ	108	4	3	18	57.14%	18.18%	27.59%
Excision margins: Invasive	108	48	17	11	73.85%	81.36%	77.42%
Level of invasion (Clark)	108	81	3	9	96.43%	90.00%	93.10%
Lymphovascular invasion	108	47	2	4	95.92%	92.16%	94.00%
Microsatellites	108	3	1	5	75.00%	37.50%	50.00%
Mitotic rate	108	46	1	6	97.87%	88.46%	92.93%
Neurotropism	108	32	2	4	94.12%	88.89%	91.43%
Other lesions	108	1	1	4	50.00%	20.00%	28.57%
Other medical history	108	32	12	16	72.73%	66.67%	69.57%
Prev. Rx/Trauma	108	4	1	4	80.00%	50.00%	61.54%
Previous melanoma	108	0	1	3	0.00%	0.00%	0.00%
Regression	108	31	5	4	86.11%	88.57%	87.32%
Site and laterality	108	81	14	10	85.26%	89.01%	87.10%
Size of specimen	108	94	5	10	94.95%	90.38%	92.61%
Specimen type	108	89	12	10	88.12%	89.90%	89.00%
Subtype	108	64	7	7	90.14%	90.14%	90.14%
TILs	108	51	2	6	96.23%	89.47%	92.73%
TILs: Density	108	7	3	3	70.00%	70.00%	70.00%
TILs: Distribution	108	9	1	8	90.00%	52.94%	66.67%
Tumour thickness	108	84	5	12	94.38%	87.50%	90.81%
Ulceration(mm diam)	108	56	3	6	94.92%	90.32%	92.56%
Overall	3024	1101	128	216	89.59%	83.60%	86.49%

Table 8.15 Results of end-to-end evaluation on the melanoma test set. TP: true positive, FP: false positive, and FN: false negative.

Field	Number	TP	FP	FN	Precision	Recall	F-score
Ancillary Studies	215	13	7	5	65.00%	72.22%	68.42%
Blocks selected	215	137	39	52	77.84%	72.49%	75.07%
Comment (DIAGNOSTIC)	215	43	26	36	62.32%	54.43%	58.11%
Comment (MACROSCOPIC)	215	87	84	65	50.88%	57.24%	53.87%
Comment (MICROSCOPIC)	215	52	67	60	43.70%	46.43%	45.02%
Comment (SYNTHESIS)	215	50	19	22	72.46%	69.44%	70.92%
Depth of invasion	215	112	26	27	81.16%	80.58%	80.87%
Distant spread	215	19	3	7	86.36%	73.08%	79.17%
Histological grade	215	113	1	8	99.12%	93.39%	96.17%
Histological type (WHO)	215	136	3	7	97.84%	95.10%	96.45%
Lymph nodes (MACROSCOPIC)	215	60	9	11	86.96%	84.51%	85.71%
Lymph nodes (MICROSCOPIC)	215	136	18	26	88.31%	83.95%	86.08%
Margins: Distal (MACROSCOPIC)	215	144	19	21	88.34%	87.27%	87.80%
Margins: Distal (MICROSCOPIC)	215	124	4	14	96.88%	89.86%	93.23%
Margins: Proximal (MACROSCOPIC)	215	78	13	14	85.71%	84.78%	85.25%
Margins: Proximal	215	128	5	11	96.24%	92.09%	94.12%

(MICROSCOPIC)							
Margins: Radial (MACROSCOPIC)	215	46	2	10	95.83%	82.14%	88.46%
Margins: Radial (MICROSCOPIC)	215	77	11	15	87.50%	83.70%	85.56%
Medical history	215	16	4	21	80.00%	43.24%	56.14%
Mesorectal integrity	215	2	0	2	100.00%	50.00%	66.67%
Metastases	215	4	1	6	80.00%	40.00%	53.33%
Number involved	215	139	14	22	90.85%	86.34%	88.54%
Other sites of disease	215	13	8	7	61.90%	65.00%	63.41%
Overlying serosa	215	48	7	10	87.27%	82.76%	84.96%
Perforation	215	3	0	4	100.00%	42.86%	60.00%
Perineural invasion	215	84	1	7	98.82%	92.31%	95.45%
Peritoneal reflection	215	10	1	7	90.91%	58.82%	71.43%
Residual tumour (R)	215	18	0	7	100.00%	72.00%	83.72%
Response to Rx	215	14	4	5	77.78%	73.68%	75.68%
Serosal Involvement	215	41	3	12	93.18%	77.36%	84.54%
Site	215	113	31	29	78.47%	79.58%	79.02%
Small vessel invasion	215	134	2	13	98.53%	91.16%	94.70%
Specimen length	215	118	2	17	98.33%	87.41%	92.55%
Specimen type	215	179	18	18	90.86%	90.86%	90.86%
Stage Group	215	112	3	21	97.39%	84.21%	90.32%
TILs	215	48	13	8	78.69%	85.71%	82.05%
TNM stage: M	215	86	5	26	94.51%	76.79%	84.73%
TNM stage: N	215	150	3	22	98.04%	87.21%	92.31%
TNM stage: T	215	155	5	16	96.88%	90.64%	93.66%
Tissue banking	215	10	0	0	100.00%	100.00%	100.00%
Tumour description	215	167	24	24	87.43%	87.43%	87.43%
Tumour site	215	76	30	38	71.70%	66.67%	69.09%
Tumour size	215	163	13	34	92.61%	82.74%	87.40%
Venous invasion	215	132	2	14	98.51%	90.41%	94.29%
Overall	9460	3590	550	801	86.71%	81.76%	84.16%

Table 8.16 Results of end-to-end evaluation on the colorectal cancer test set. TP: true positive, FP: false positive, and FN: false negative. Note that scores for "Extramuscular spread" and "Specimen images" are not presented, as there are no test samples for them.

Field	Number	TP	FP	FN	Precision	Recall	F-score
Cell size	89	38	2	13	95.00%	74.51%	83.52%
Clinical impression	65	23	8	8	74.19%	74.19%	74.19%
Comment	76	16	0	1	100.00%	94.12%	96.97%
Const symptoms	65	3	2	2	60.00%	60.00%	60.00%
Cytogenetics comment	77	0	2	1	0.00%	0.00%	0.00%
Cytomorphology	89	27	0	2	100.00%	93.10%	96.43%
Diagnosis (WHO)	76	58	8	11	87.88%	84.06%	85.93%
Disease extent	65	2	1	2	66.67%	50.00%	57.14%
FISH	77	0	1	1	0.00%	0.00%	0.00%
Flow cytometry: Comment	89	6	2	4	75.00%	60.00%	66.67%
Flow cytometry: Negative for	89	0	0	1	0.00%	0.00%	0.00%
Flow cytometry: Positive for	89	0	0	1	0.00%	0.00%	0.00%
Grade	89	13	5	8	72.22%	61.90%	66.67%
Immunohistochemistry: Comment	89	4	19	23	17.39%	14.81%	16.00%
Immunohistochemistry:	89	3	0	3	100.00%	50.00%	66.67%

Equivocal for							
Immunohistochemistry: Negative for	89	32	5	13	86.49%	71.11%	78.05%
Immunohistochemistry: Positive for	89	38	5	12	88.37%	76.00%	81.72%
Indication for biopsy	65	1	1	7	50.00%	12.50%	20.00%
Lineage	76	33	0	1	100.00%	97.06%	98.51%
Medical history	65	7	3	7	70.00%	50.00%	58.33%
Other sites of disease	65	1	0	3	100.00%	25.00%	40.00%
PCR comment	60	0	0	1	0.00%	0.00%	0.00%
Pattern of infiltration	89	35	2	9	94.59%	79.55%	86.42%
Predisposing factors	65	0	2	3	0.00%	0.00%	0.00%
Presentation	65	6	4	11	60.00%	35.29%	44.44%
Received in	109	16	2	5	88.89%	76.19%	82.05%
SNOMED RT codes	57	54	1	2	98.18%	96.43%	97.30%
Site and laterality	65	44	11	13	80.00%	77.19%	78.57%
Specimen size	109	89	5	16	94.68%	84.76%	89.45%
Specimen type	109	47	16	27	74.60%	63.51%	68.61%
Tissue reactions	89	19	4	12	82.61%	61.29%	70.37%
Triage	109	30	2	9	93.75%	76.92%	84.51%
Overall	2588	645	113	232	85.09%	73.55%	78.90%

Table 8.17 Results of end-to-end evaluation on the lymphoma test set. TP: true positive, FP: false positive, and FN: false negative. Note that scores for "Stage", "IgH" and "TCRgamma" are not presented, as there are no test samples for them.

There were 306, 1164 and 300 errors identified in the melanoma, colorectal cancer and lymphoma test sets respectively. Table 8.18 summarizes the error types for each test set.

Error category	Melanoma test set	Colorectal cancer test set	Lymphoma test set
1. Errors from entity recognition	88.89%	84.79%	83.00%
1.1 Errors from specimen id detection	44.77%	0.00%	2.33%
1.2 Errors from section heading detection	2.61%	2.58%	0.00%
1.3 Errors from medical and linguistic entity recognition	41.50%	82.22%	80.67%
2. Errors from relation extraction	0.00%	0.00%	3.00%
3. Errors from structured output generation	7.52%	8.33%	6.33%
3.1 Weaknesses in ranking criteria	0.98%	1.72%	2.33%
3.2 Weaknesses in post-processing modules	2.61%	4.64%	1.67%
3.3 Errors from negation and uncertainty detection	1.31%	0.60%	0.33%
3.4 Inappropriate application or insufficiency of ranking criteria or post-processing modules	1.63%	0.34%	2.00%
3.5 Errors from other processing engines	0.00%	1.03%	0.00%
3.6 Usability problems	0.98%	0.00%	0.00%
4. Poor-writing of the original report	3.59%	6.87%	7.67%

Table 8.18 Error types for each test set.

The greatest contribution to the errors was due to the poor performance of MER, accounting for 83% or more of the total errors in each test set, wherein incorrect specimen id detection attributed to over half of the errors in MER on the melanoma test set. Only about 6.3-8.3% of the errors on the test sets were due to the weaknesses in the SOG components, which is consistent with the results in the

training sets. It is also worth pointing out that around 6.9% and 7.7% of the errors on the colorectal cancer and lymphoma test sets were caused by the poor-writing of the original reports. Except for those issues mentioned in the Other category, an occasional but notable issue on the colorectal cancer test set is misuse of the section headings by the pathologists in some documents (e.g., use "SPECIMEN" as a "MICROSCOPIC" heading; in fact, it stands for a "MACROSCOPIC" heading in the training set), which affects section context detection and eventually leads to the invalid outputs.

The micro-averaged F-scores in most fields were over 60%, except for six fields in the melanoma test set: "Desmoplasia", "Distant metastasis", "Excision margins: In-situ", "Microsatellites", "Other lesions" and "Previous melanoma"; five fields in the colorectal cancer test set: "Comment (DIAGNOSTIC)", "Comment (MACROSCOPIC)", "Comment (MICROSCOPIC)", "Medical history" and "Metastases"; ten fields in the lymphoma test set: "Cytogenetics comment", "FISH", "Flow cytometry: Negative for", "Flow cytometry: Positive for", "Immunohistochemistry: Comment", "Indication for biopsy", "Medical history", "Other sites of disease", "PCR comment" and "Predisposing factors".

Incorrect results from the MER prevented correct population of the structured fields in the first place. Most of the fields with poorer performances were because of the worse results on the associated medical entity types obtained from the MER, for four possible reasons:

1. The associated medical entity types were scanty in the training sets, such as "Desmoplasia", "Microsatellites", "Cytogenetics comment", "FISH", "Flow cytometry: Negative for", "Flow cytometry: Positive for", "Indication for biopsy" and "PCR comment", each with a frequency smaller than 30.
2. High lexical variability occurs in the associated medical entity types of some fields (e.g., "Other sites of disease" and "Immunohistochemistry: Comment"), which makes the MER model hard to identify the entities in the test sets.
3. Abbreviations proved to be a challenge for MER during the training phase, which also increased the difficulty for testing. This is highlighted on "Medical history" and "Predisposing factors", where entities presented as abbreviations or acronyms cannot be identified by the models in many cases.
4. Ambiguity is another possible reason for the lower F-scores on some fields, such as "Other lesions" and "Excision margins: In-situ". The instances of associated entity types En:Lesion (other) and Ma:Excision In Situ are frequently misclassified to other two entity types: En:Primary Lesion and Ma:Excision Invasive, as they have similar lexical items and linguistic constructions, and the machine learner tends to misclassify the instances to the dominant types: En:Primary Lesion and Ma:Excision Invasive.

Besides poor MER results, the lower F-scores on several fields may also be due to:

- The values for "Comment (DIAGNOSTIC)", "Comment (MICROSCOPIC)" and "Comment (MACROSCOPIC)" are combinations of the instances of three entity types ("Sy:Comment",

"En:Coexistent Pathology" and "De:Ancillary Studies") situated in "CONCLUSION", "MICROSCOPIC" and "MACROSCOPIC" sections respectively, which makes them more error-prone, since only when all these types of instances are recognized accurately, can the fields be populated correctly.

- The distribution of the entity type En:Distant Spread or Metastases in the colorectal cancer training set is not even: the largest portion (34.3%) locates in the "MICROSCOPIC" section. Thus the performance of "Distant spread" (derived from the entities in the "MICROSCOPIC" section) was much better than the performances of "Metastases" (derived from the entities in the "MACROSCOPIC" section), the F-scores of which were 79.17% and 53.33% respectively on the test set.
- In the melanoma training set, the distribution of the associated entity type Sy:Diagnosis is also uneven: the "MICROSCOPIC" section holds more than a half of the total amount, while only about 15.7% of these entities occur in "CLINICAL HISTORY" section. Additionally, the positive fields (where the values are not "N/A") of "Distant metastasis" and "Previous melanoma" are scarce (with training sample sizes of 19 and 6) so that the extraction rules derived from the training data may not fit for the test data.

The whole system has been released to the research community as a web page for testing (see http://www.icims.com.au/QUPPDemo for more details). Some examples are demonstrated in Appendix IV.

8.6 Discussion

8.6.1 Comparison of Different Template Construction Strategies

From Table 8.16, incorrect specimen id detection accounted for diverse proportion of total errors: largest for the melanoma test set; minimal for the lymphoma test set; none for the colorectal cancer test set, because of different template construction strategies applied to them. For a multiple specimen/tumour document, in the melanoma and lymphoma corpora, the template sections were reported by specimen id or a set of specimen ids; in the colorectal cancer corpus, the reporting did not rely on specimen ids. Therefore, accurate detection of specimen ids is much more important for the other two corpora than in the colorectal cancer corpus. Moreover, it is critical for the melanoma corpus, as the template sections are reported under each specimen id. If a specimen id is missed (a false negative on specimen id detection), a template section under that id will be underreported; if a specimen id is misclassified (a false positive on specimen id detection), a template section under that id will be over-generated. By modifying this strategy to allow the template sections to be reported under multiple ids if the ids share the same contexts, the effect of incorrect specimen id detection declined dramatically to a minimum on the lymphoma test set.

In terms of the usability, a structured report constructed by the strategy used for the melanoma corpus is clear and easy to follow, as the sections are separated by specimen id, but it is vulnerable to incorrect specimen id detection. That constructed by the strategy for the colorectal cancer corpus is stable, since the amount of reportable fields will not change according to the number of specimen or tumour ids, yet implicit, as it is hard to tell whether the value in a field is reported against a single specimen/tumour or multiple specimens/tumours. That constructed by the strategy for the lymphoma corpus is preferable: it is explicit, similar to the one in the melanoma corpus. It is more robust to the errors from specimen id detection, which diminishes the risks of underreporting or over-generation of template sections.

8.6.2 Comparison with Other Works

In comparison with MedTAS/P, the system achieved comparable performance (F-scores) in some fields, such as "Histological grade": 0.96 *vs.* "Grade value": 0.98, "Specimen length":0.93 *vs.* "Gross description part": 0.90, "Metastases": 0.53 and "Distant spread": 0. 79 *vs.* "Metastatic tumor": 0.65, and poorer performance in certain fields, for instance, "Site": 0.79 *vs.* "Anatomical site":0.97, "Tumour size": 0.87 *vs.* "Dimension": 1.00. Other fields can not be compared, as they are out of the scope of this study (e.g., "Date"), or the definitions of them are quite different, for example, in the evaluation of MedTAS/P, for "Lymph nodes", the total number of excised nodes is recorded as is the number of positive ones; in this study, it was devided into the total number of excised nodes in macroscopic examination ("Lymph nodes (MACROSCOPIC)") and microscopic examination ("Lymph nodes (MICROSCOPIC)"), and the number of positive ones ("Number involved"). Moreover, the system provides additional information of clinical significance, which MedTAS/P lacked, such as "Perineural invasion", "Small vessel invasion" and "Venous invasion".

The system performance is not compared to those of existing natural language processing (NLP) systems like MedLEE (Friedman et al., 2004) and cTAKES (Garla et al., 2011), since these systems have achieved relatively high performance on encoding of clinical documents or recognizing medical entities, at the cost of maintaining a lexically variant-rich encoding table or dictionary, but the goal of this study is different, that is to extract pertinent information from the free texts to populate structured templates rather than encoding.

The results presented above are also not compared to other works (e.g., the works of Qu et al (Qu et al., 2007) and Nguyen et al (Nguyen et al., 2012)), as those works either focused on the usability (e.g., the designs of user interface) , or the inference of TNM stage values from the narratives, and they did not report any error rate on populating structured fields.

8.6.3 Discussion of the Three Corpora

Although this study was focused on three specific cancer diseases, the rules still have some generality. For example, the negation uncertainty inapplicability criterion was frequently applied to many fields

across the three corpora, which demonstrates that this is a common criterion that fits a variety of cancer pathology notes. Likewise, the Clear Processor and Measurement module were used to post-process the fields involved in excision margins in both the melanoma and colorectal cancer corpora, which indicates that these modules can also be reused in other cancer pathology reports to extract information about excision margins if applicable. Table 8.19 presents these general criteria and the modules used in common across the corpora.

Corpus	Criterion/Module
Melanoma, colorectal cancer, lymphoma	Frequency criterion, negation uncertainty inapplicability criterion; Dimension Processor
Melanoma, colorectal cancer	Specimen distance criterion, measurement criterion, uppercase criterion, distribution density degree criterion; Clear Processor, measurement module, sub-classification module
Melanoma, lymphoma	Laterality criterion, position criterion, temporality criterion, specific criterion, mood degree criterion; Temporality module
Colorectal cancer, lymphoma	Medical category criterion, size criterion, procedure criterion

Table 8.19 General criteria and modules usage across the corpora.

Error analysis showed that a single specimen/tumour report with standard headings and the presence of simple and concise statements was significantly associated with correct populating. This is probably because:

The poorer performance of En:Specimen Identifier and Specimen Identifier on MER could affect the final populations in the melanoma corpus to a great extent and lesser extent to the lymphoma corpus. For example, “A.” can be presented as a block id rather than specimen id in some cases. If a specimen id is missed or misclassified, it could directly affect the results of document classification and specimen context detection, and then finally negatively influence the structured outputs. A detailed analysis shows that a good representation of a specimen id can start with a lexicon “Specimen”, include brackets or period for a numeral (e.g., “Specimen A”, “(1)”, “2.”), and the representation needs to remains consistent in the whole report.

Correct detection of section contexts is requisite for SOG, thus misuse or omission of section headings will hinder section context detection, and consequently affect the final transposition.
A simple and concise statement was also more likely to be detected by the machine learning algorithms. For instance, it seemed too difficult to populate correct values for “Excision margins: In-situ” and “Excision margins: Invasive” from the example:

> The nearest resection margins for the dysplastic junctional naevus [“En:Associated naevus (type)”], in situ and invasive melanoma [“Sy:Diagnosis”] measure 1.8, 2.5 and 3.5mm respectively [“Ma:Excision In Situ”, “Ma:Excision Invasive”].

Firstly, current machine learning methods such as Conditional Random Fields (CRF) cannot assign more than one tag for an overlap instance; secondly, even if the instances can be recognized correctly, it still needs more complex rules to extract values from them.

Error analysis also addressed a critical and yet general issue that MER appeared to be the bottleneck of the whole study, which resulted in most of the errors on end-to-end evaluation. Some solutions may be useful to improve the system performance, which have been discussed in Chapter 5, such as exploring other features for better feature selection, ensemble multiple classifiers or machine learning algorithms.

The system performed better on well-written reports than the poorly-written ones, as the poor-writing brought several issues that are difficult to handle (which have been discussed above). Some examples of these poorly-written reports are presented in Appendix V.

8.6.4 Limitations

One of the limitations of this work is that the sample size for testing is not enough to carry out a thorough evaluation, especially for the lymphoma corpus. Several fields cannot be evaluated in the test sets (e.g., “Specimen images”, “IgH” and “TCRgamma”) due to lack of test samples. It is likely that the system performance is limited by insufficient sample size.

At first, the structured outputs were displayed in a web page for pathologists to evaluate, but we have been unsuccessful in recruiting any pathologist to engage fully with the task although a number volunteered initially. The extracts were validated by computational linguists trained to do this task. While they may not have been able to interpret the extractions as precisely as pathologists, their work has face validity and is internally consistent, which has been shown in a previous work on the project (Patrick and Scolyer, 2008).

8.7 Conclusion

In this chapter, the detailed process of generating structured outputs was described, including the design of the predefined templates, mapping strategies and a rule-based system for populating these templates.

The good performances of the structured output generation system on the training sets in the second round evaluation (all F-scores exceeded 97%) revealed that the rules were competent at populating the structured outputs based on the gold-standard annotations. This was also consistent with the findings in end-to-end evaluation on the test sets, where weaknesses in the structured output generation system contributed a small number of errors.

Error analysis on end-to-end evaluation also demonstrated that medical entity recognition is the bottleneck of the whole study, as the majority of total errors were due to its incorrect results.

It is believed that the rules proposed have some generality, which are not limited to the cancer diseases in the study, as some general ranking criteria and post-processing modules can be reused or easily adapted across the three sets of cancer pathology notes.

Chapter 9 Conclusions

In this thesis the problem of automatic population of structured reports from narrative pathology reports was studied and several sub-tasks of the study presented. In the pathology domain, traditional narrative reports commonly have some problems. For example, essential elements are occasionally omitted, especially negative results which are not always reported clearly; the referring doctors often find it difficult to identify the necessary elements to justify a given diagnosis. Compared to free-text reports, there are a number of advantages for the use of structured reports, which can improve the communication between pathologists and clinicians. For instance, they can improve the completeness of pathology reporting; they are more concise and easy to read and, they can improve the efficiency for cancer registries, clinical audits and epidemiology research. Natural language processing (NLP) is one promising approach to extracting critical findings and diagnoses and incorporating them into a predefined structured template, thereby achieving the goal of automatic population of structured reports.

Generally, this application of NLP technology is an information extraction (IE) task in the clinical domain, but it is more difficult to achieve than IE tasks in the general domain, as it requires deep understanding of the domain knowledge. It has to deal with specific and complex lexicons that cannot be found in common medical terminologies. Moreover, in different clinical sub-domains, there are different sub-languages used, which need to be also considered in the task.

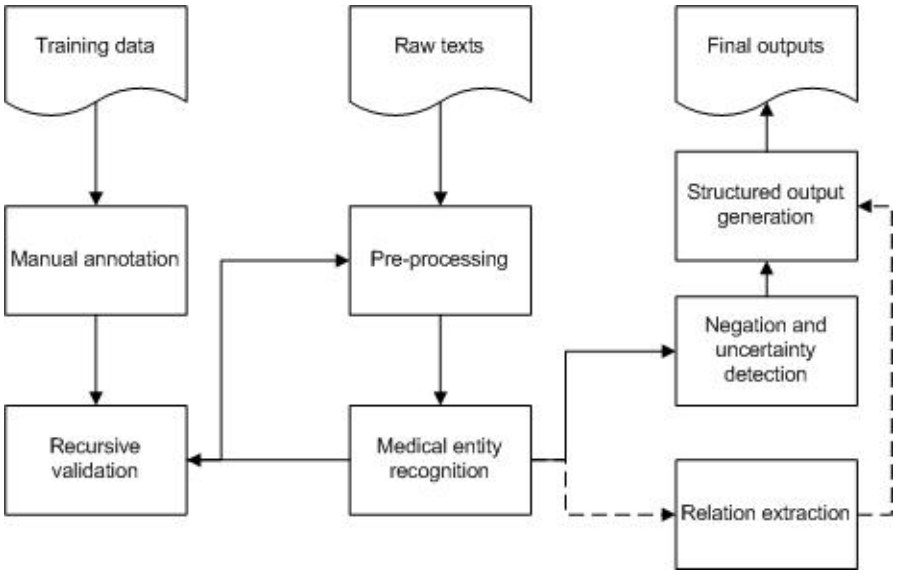


Figure 9.1 Pipeline system architecture for automatic structured reporting. Dashed lines indicate that the subsequent processes were applied to the lymphoma corpus.

The work presented in this thesis has demonstrated that an IE system that combined a supervised machine learning based approach enhancing by some rule-based methods was a feasible strategy for

accomplishing this task, and achieved promising results. The complete pipeline architecture is illustrated in Figure 9.1.

As shown in the diagram, raw records are passed to the pre-processing engine, including sentence boundary detection, tokenization, part-of-speech (POS) tagging, and section context detection. In a separate process, the training data are annotated manually to create gold-standards. Subsequently errors in the manual annotations are identified by performing recursive validation on the training data. The errors were corrected manually so that the model would not learn from the incorrect examples. After pre-processing, a supervised machine-learning based-approach is used to recognise medical entities from the corpora, using Conditional Random Fields (CRF) (Lafferty et al., 2001) and integrating various features. The negation and uncertainty detection modules are applied to detect the assertions of particular entities. For the lymphoma records, a relation extraction system is prepared to extract specific relations between entities, consisting of a rule-based module and a Support Vector Machine (SVM) (Cristianini and Shawe-Taylor, 2000) classifier. Then the rule-based structured output generator populates the final outputs conforming to the structured templates.

The main contribution of the research is that it is a pilot study investigating how narrative pathology reports can be automatically converted into structured reports by building a hybrid system that demonstrates the feasibility and accuracy of it. Researchers can use it as a baseline for the development of more complex models in the future. Accordingly, the development of the system can be divided into the developments of the following sub-systems or modules focusing on: medical entity recognition (MER), negation and uncertainty detection, relation extraction (RE), and structured output generation respectively.

9.1 Thesis Overview

This thesis began with an introduction of the work in Chapter 1, including the significance of the study, background knowledge about IE tasks in the clinical domain, the main barriers to the task, the research problems and proposed solutions.

Chapter 2 surveyed some previous work related to the field, which reviewed current state of the art techniques focusing on four main sub-tasks: MER, negation and uncertainty detection, RE and automatic structuring. Generally, there are two main steams: rule-based methods and statistical approaches. On one hand, rule-based methods tend to provide reliable results with a relatively small amount of training data, and the hand-crafted rules are comprehensible for the developers or domain experts, which eases the error analysis effort. However, these methods face difficulty when dealing with unfamiliar or erroneous input data, and the rules are usually tailored for a specific task, which may not be readily reusable for other tasks. On the other hand, statistical approaches can handle problematic data by learning from training examples. Also, statistical approaches can achieve comparable or better performance by simply adopting features extracted from the corpus, thus they are

usually more portable than their rule-based counterparts, but they also require a large amount of training data to create gold-standards. Therefore, a combination of statistical methods and rule-based approaches is preferable to build a hybrid system for a complex IE task.

Chapter 3 presented the detailed analyses conducted on the corpora. The lexical analysis demonstrated that specific characteristics of the texts in pathology notes are distinguishable from those in other genres of texts. Several significant language phenomena were observed, including abbreviations, unknown words, misspellings, non-alphabetic tokens, lexical variants and complex vocabulary, which indicated the challenges that may be encountered in the following processes, which require sophisticated NLP techniques to resolve. The quantitative completeness analysis showed the coverage of most fields is unsatisfactory, revealing several issues to be addressed in the following processes, especially in the construction of the structured templates.

Chapter 4 described three semantically annotated corpora: the melanoma corpus, the colorectal cancer corpus and the lymphoma corpus, which were annotated with entities or relationships between the entities. The whole annotation process was illuminated, including the design of the annotation schemas and guidelines. The correspondence analysis showed that the annotation schemas had appropriate granularity that could capture most information related to the structured fields without too much ambiguity. A mixed conveyor method with a two phase validation was adopted, which improved the efficiency and reduced the difficulty of the annotation process. The main annotation process was carried out by computational linguists, and they were competent to accomplish the task once they were properly trained and detailed annotation guidelines were provided. Moreover, recursive validation was performed on the initial gold-standard annotations to attain higher consistency among them. As a result, these unique annotations of high quality were suitable for future experiments.

A supervised machine learning-based approach was proposed in Chapter 5 to recognize medical entities in the corpora. CRF-based models were able to capture a significant portion of the entity boundaries by utilizing contextual features, since the spans of most entities were within a nine-token window. Rich feature sets provided a great number of useful clues for classifying the entity types. By feature engineering, the best feature configurations were attained and achieved significant gains on the models. Some common effective features were identified: *lowercase of tokens*, *lemma*, *POS tag* (or *bag of POS tags*), *medical category*, *ring-fenced tag* (or *bag of ring-fenced tags*), *suffixes*, *bag of prefixes*, and *section context*, which can also be beneficial for other MER tasks using similar approaches.

The negation and uncertainty detection modules were presented in Chapter 6. First, a case study of negation detection was performed on the lymphoma corpus, where three different methods were experimented with: the lexicon-based approach was a rule-based method, modified from NegEx (Chapman et al., 2001b), an existing negation detection algorithm, relying on the application of trigger

terms and termination clues; the syntax-based approach was a rule-based method as well, implemented with a set of rules and negation patterns designed according to the dependency output from the Stanford Parser (Klein and Manning, 2003); an SVM classifier armed with a number of features was adopted in the machine learning-based approach. The syntax-based approach and machine learning-based approach performed best on the training set and test set respectively, at the cost of very long run time, while the lexicon-based approach was simple and efficient, and yielded reliable performance, thus it was applied to the other two corpora. A similar approach was also prepared for uncertainty detection. The main adjustment of the approaches for the other two corpora was modifying the entries of trigger terms, pseudo-trigger terms and termination clues. The modules obtained very good performances on the training sets. The dramatic drop of F-scores on the test sets were mainly due to incorrect MER rather than errors from the modules.

A RE system was described in Chapter 7 to extract four relations from the lymphoma corpus. It included a rule-based module where simple heuristic rules were applied, and an SVM classifier that adopted several useful features. The system achieved very good performance on the training set, and the most effective features for the classifier were the contextual, positional and semantic features.

In Chapter 8, the process of generating structured outputs was described in detail. It illuminated the design of the predefined templates, mapping strategies and a rule-based system for the population of these templates. The rule-based system had four processes: document classification, specimen context detection, candidate preparation and XML generation. The main process was candidate preparation, implemented with a series of post-processing modules and ranking criteria. The rule-based system was improved significantly by the refinement of the rules, which performed very well on the training sets. MER was the bottleneck of the whole study, as incorrect results from it were the major cause of total errors in end-to-end evaluations on the test sets, while the structured output generation system contributed to a small number of errors. Although the rules were proposed based on three specific cancer diseases, they still had some generality, since several post-processing modules or ranking criteria can be reused or easily adapted for other cancer pathology notes.

9.2 Future Work

9.2.1 Further Improvement

In view of the complexity and variability of language embedded in narrative reports, coupled with the existing error rate of the system, the system is capable of further improvement.

Improvement for the Quality of Annotations

Both the quality and the size of training data can significantly affect the performance for machine learning-based approaches. Given limited time and resources, three corpora with relatively small sizes were annotated for the study.

The size of annotated corpora should be increased if more cancer pathology notes are available. The size of the annotated data is also subject to the time spent on the annotations. An active learning approach (Thompson et al., 1999) is a possible solution to reduce the annotation effort, as only the most informative instances are annotated in this approach.

The quality of annotated data can be improved by enhancing the clarity of the annotation schemas and the consistency of applying the annotation guidelines. For example, the annotation guidelines can be further refined by employing more medical and linguistic knowledge, adding more specifications and detailed examples.

Improvement for Medical Entity Recognition

The supervised machine learning-based system did not exhaust all potential useful features. Additional features could be introduced to the models. For instance, besides a local contextual window and section context, more complicated contextual clues involving the adjacent sentences could be utilised to remedy the ambiguous classification of entity types.

The combination of the features could be explored as well. Other features were tried in combination with the nine-token contextual window in the study, so the combination with different window sizes could be investigated to improve the system performance.

Other machine learning algorithms could be used to recognize entities with long spans (over nine tokens), which seems to be a drawback of CRF++. CRF tends to misclassify the minority entity types to the majority counterparts, especially when they have ambiguous expressions. One possible solution is to apply the stacking or voting strategies to the aggregation of the results from different machine leaning classifiers (Dzeroski and Zenko, 2004; Wang and Patrick, 2009). By overcoming the problems brought by a single classifier, the combination of multiple classifiers may yield better classification performance.

Improvement for Negation and Uncertainty Detection

The fixed four-word window size for uncertainty detection led to the occasional omission of some distant entities from the uncertainty trigger terms, thus a more flexible window size could be considered to determine the scope for uncertainty.

Inappropriate clustering of the trigger terms resulted in some false outputs. Simple positional clusters were not able to handle some cases properly, which require thorough cluster analysis to tackle. Specific conditions should be considered in complex clusters.

The lexicon-based detection approaches could not cope with complicated cases, which is likely to be overcome by resolving coreference or introducing syntactic clues. They also failed to determine the scope involving particular prepositions correctly. Employing domain knowledge or global context information may be helpful to resolve this issue.

Insufficient predefined trigger terms hindered the performances of the modules, hence the predefined trigger term lists should be augmented with additional lexical entries.

A more comprehensive integration of different detection modules should be taken into account in order to avoid some problems caused by current weak integration of the modules.

Improvement for Relation Extraction

One of the obstacles to better RE is incorrect results from sentence boundary detection and dependency parsing. This requires a more sophisticated sentence boundary detector and parser trained on the domain.

Deficient feature extraction and construction also brought some problems. For example, the implicit meaning implied by particular tokens between the paired entities was not considered, which caused the failure to exclude some invalid relations. There should be further consideration on these tokens when applying the related feature. The categorization of dependency distances may be too ambiguous, which needs to be split into finer grained categories in future work.

The current feature selection method is a “bottom-up” method (Whitney, 1971), where an initial empty feature set is incremented progressively with features to find the best configuration. It suffered from the nesting effect where features once added cannot be removed. Other advanced methods, such as the stepwise feature selection and sequential forward floating search methods (Pudil et al., 1994; Sahiner et al., 2000), can be used instead, where the system performance is assessed on more flexible combinations of features. For example, in a sequential forward floating search method (Pudil et al., 1994), the number of features is added or removed at each step dynamically, and the stopping criterion is controlled by a predefined amount of desired features.

Improvement for Structured Output Generation

The small sample size for testing could not support thorough evaluation on the system, especially for the lymphoma corpus. The scalability of the system performance is subject to the limited sample size. Therefore, besides increasing the size of training data, the size of test data also needs to be increased.

At present, the structured outputs have been validated by computational linguists with training. However, a stricter validation should be conducted by pathologists, for further application of the system to the clinical settings.

In addition, to assess the portability of the methodology in this work, the system can be modified and adapted for other cancer pathology notes, and then the system performance can be evaluated on them. To attain better usability, pathologists could take part in the testing, with their feedback taken into account for the improvement of the system.

9.2.2 Further Development

Given the structured outputs obtained from the system, there are several potential uses for them:

Screening Tool

Using structured reports can result in more complete and consistent pathology reports. The system can be implemented as a screening tool, which screens pathology reports prior to finalization by prompting pathologists about important findings that may be inadvertently left out and inconsistent with the structured fields. For example, in the AJCC Cancer Staging Manual (Edge et al., 2010), N stage is classified based on the number of malignantly involved regional lymph nodes. By validating the values in the fields "Number involved" and "TNM stage: N", it can tell pathologists whether they have provided the correct N stage values in the reports. Likewise, if "follicular" is populated to the field "Pattern of infiltration", while "Diffuse large B-cell lymphoma" is a value for the field "Diagnosis (WHO)", it can remind pathologists that they may have made a wrong diagnosis in the report. Additionally, it can also reduce the ambiguity of medical terms used, and decrease variability in their interpretation.

Decision Support

Some structured fields can help clinicians to decide further clinical management of the patients. Here are some examples:

1. Three fields about excision margins in the melanoma corpus: "Excision margins: Deep", "Excision margins: In-situ" and "Excision margins: Invasive". Given the thicknesses of the melanomas, where there are different requirements for the width of the surgical margin. For example, a 1-cm margin is recommended for the excisions of melanomas with thickness <1 mm; melanomas that are >2 mm thick should be excised with 2-cm surgical margins (Balch, 2002; Reintgen, 2001). Hence, if any of the numeric values populated to the three fields are smaller than those indicated in the requirements, re-excision may be advised in case of residual melanoma at the primary site.
2. Two fields about metastases: "Distant spread" and "Metastases" in the colorectal cancer corpus. According to the clinical practice guidelines, first- or second-line chemotherapy is standard treatment for colorectal cancer patients with metastases (Van Cutsem et al., 2010). Thus, if "present" occurs in any of the two fields, clinicians should consider chemotherapy as a preferable treatment for the patients.

Automated Encoding

Encoding the structured fields is highly significant. The structured outputs can be mapped to existing terminologies (such as SNOMED CT, ULMS, ICD-O) so that they can be accessed reliably by other automated clinical applications at different institutions and used for a broader range of purposes. Currently, only medical entities in the lymphoma corpus are encoded in SNOMED RT codes, which is limited to the diagnostic summary. By integrating with some existing automated encoding tools (e.g., Friedman et al's and Patrick et al's works (Friedman et al., 2004; Patrick et al., 2007a)) or designing new tools, the structured outputs from the narrative pathology notes can be converted into codes or concept identifiers defined in the terminologies, to ease the data storage and facilitate effective retrieval.

Appendix I Details of the Post-processing Modules and Ranking Criteria for the Structured Fields in Each Corpus

I.1 Post-processing Modules

General Process Module

This module includes several general pre-processes that can be applied to most fields before ranking, such as misspelling correction and expansion of acronyms or abbreviations.

During ranking, for those fields in a section with multiple associated section contexts, the candidates in the primary section context take precedence over those in the supplementary section context. For example, in the lymphoma corpus, if two candidates for "Site and laterality" locate in "CLINICAL HISTORY" and the "SUPPLEMENTARY REPORT" respectively; the one in "CLINICAL HISTORY" is considered first. The results from specimen context detection also affect the order for the selection. The candidates in a single specimen id's context are preferable to the one in a multiple specimen ids' context.

There are also some post-processes to handle the candidates after ranking. For example, a process called overlapping candidate reduction is to remove the repetitive or less informative candidates, by comparing the token length and similarity among the candidates.

Measurement Module

For the melanoma corpus, the module processes according to the following steps:

1. Check whether there is any measurement unit (e.g., mm, cm, and millimetre) in the candidate and extract it if it exists.
2. Check whether there is any numeral (e.g., 1, 20, and 0.8) in the candidate and extract it if it exists.
3. Check whether there is any alphabetic word that can be converted to an Arabic numeral (e.g., one, two, and three) in the candidate and convert it if it exists.
4. Check whether there is any keyword or punctuation that suggests the value is fuzzy (e.g., -, >, greater, less) in the candidate and extract it if it is a punctuation or convert it to an associated sign if it is a keyword, e.g., "less" is converted to "<".
5. Check whether there is any conjunction (e.g., and, or) in the candidate and extract it if it exists.

Finally, the module combines all the values from the steps above to yield the output.

After revision, a pre-process to remove extra white space inside a potential numeric value (e.g., "1. 5mm") was added in the beginning of all the steps.

For the colorectal cancer corpus, the fifth step was skipped, thus if multiple numeric values were extracted, the first one was used as the output in the initial design. This was modified to allow all the extracted values to be considered as appropriate output after first round error analysis.

Naevus Module

This module processes as follows: exclude any prepositional phrase from the candidate; find the noun phrase in the candidate; strip the determiner from the noun phrase if applicable; the remaining part of the phrase becomes the output.

After revision, the lexicon "cell" and "cells" were also stripped from the noun phrase to yield a standard output.

Level Module

There are five Clark's levels defined in the protocol: Level I, II, III, IV, and V. First, the module tries to identify whether the candidate has any of these Roman numerals, or any Arabic numeral that can be converted to these Roman numerals, and then extracts the numeral or converts it if applicable; detects lexical items "to" and "or" in the candidate, and convert them to "-"and "/" respectively if applicable.

Temporality Module

The module firstly detects whether there is a Li:Temporality entity inside the five-token window and the same sentence of the candidate; if there is, the entity is extracted as the result.

After the first round error analysis on the melanoma training set, the result would be ruled out if it is not an entry in the regression temporality gazetteer and the candidate is a Sy:Regression entity.

Regress Module

Initially, the module finds any adjective (except "regressive") in the candidate.

After the first round error analysis, the module also finds an adverb in the candidate or the surrounding Li:Mood and Comment Adjuncts entities if it is an entry in the regress gazetteer.

Rate Module

The processes in this module are similar to those in measurement module, replaced with different predefined units, such as "/mm2", "/sqmm", "per hpf", "per square millimetre" and "in a high power field". The default unit is "/mm2", if no particular unit can be detected from the candidate.

Dimension Processor

The module processes according to the following steps:

1. Unify the lexical variants of dimensions. There are four standard dimensions in total, which are “diameter”, “length”, “width”, and “depth”. For example, phrases referring to “length” can be expressed as “long”, “length”, “longitudinally”, “longitudinal”, “axially”, and so on; if any of them is detected in the candidate, it will be standardized to the dimension “length”.
2. Detect the dimensionality of the candidate. A simple rule is used for the detection: check the number of the multiply sign “×” between numeric value(s) in the candidate: for a two-dimensional size, there should be two “×”; three “×” for a three-dimensional size; no “×” for a one-dimensional size.
3. Extract the numeric value(s) from the candidate. Extract any numeral and unit from the candidate, and integrate with the detected dimension to yield dimension pairs. For a one-dimensional size, the dimension pair is denoted as {the standardization in Step 1: one-dimensional size}; for a two-dimensional size, the dimension pair is rendered as {“area”: two-dimensional size}; for a three-dimensional size, the dimension pair is shown as {“volume”: three-dimensional size}; the default one is {“dimension”: one-dimensional size}, if none of the standard dimensions, but a one-dimensional size is detected in above steps.
4. For the melanoma and lymphoma corpora, an extra step is used to handle special cases, e.g., multiple one-dimensional sizes with different keys in the pairs or hybrid dimensional sizes (a two-dimensional size and a one-dimensional size) are detected in the candidate. The module can merge them together with “×”. For the melanoma corpus, the dimension of “diameter” should always be indicated in the result if applicable (advised by the pathologist).
5. For the colorectal cancer corpus: If the candidate is a De:Specimen Size entity, the module will only generate a result for dimension pairs with keys of “length”, “dimension”, “area” and “volume”; for “area” and “volume”, the first numeric value is used to generate the result. If the candidate is a De:Tumour Size entity, the result is the maximum numeric value from a dimension pair with a key of “area” and “volume” or a standard dimension with multiple extracted one-dimensional sizes; the numeric value for a dimension pair with a key of a standard dimension with single extracted one-dimensional size.

After the first round error analysis, the module was slightly modified to tackle more complex cases.

Node Number Module

There are several steps in this module:

1. Extract any numeral or any alphabetic word which can be converted to an Arabic numeral (e.g., “twenty-five” and “eighteen”) from the candidate if applicable. Note that numeric values for measurements of the lymph nodes are ruled out, such as “2mm” and “3-10mm”.
2. If the candidate is an Ex:Lymph Node Involvement entity, the ratio between involved nodes and identified ones should also be detected, and the number of identified ones is extracted if applicable, e.g., “30” is extracted from “none of which are involved by metastatic adenocarcinoma (0/30) including the apical lymph node” and “16” is extracted from “4 out of 16 lymph nodes show metastatic carcinoma”.

3. If no numeral or alphabetic counterpart is detectable, check whether the candidate has negation inside and assign number "0" if it is negated (e.g., "no lymph node in the appendiceal area"). Note that this step cannot be applied to an Ex:Lymph Node Involvement candidate (e.g., the count for "Two proximal ileocaecal nodes are clear of tumour" is "2" instead of "0").
4. Check whether the candidate is the total count of nodes.
5. Special rules are applied to find extra counts of nodes. For example, assign number "1" for "apical lymph node is identified 4mm in diameter" and "single local lymph node".

Involvement Number Module

Most processes are similar to those in node number module, except that in Step 2, the number of involved nodes is extracted instead of that of identified ones; Step 3 is always applied to the candidate.

Sub-classification Module

This module tries to classify the candidate to a specific sub-type. The sub-type result will be used for ranking, or become part of the population to the associated field.

Specific sub-types of some entity types are listed below:

- Ma:Proximal or Distal Margin: "Proximal", "Distal"
- Ma:Clear: "Proximal", "Distal", "Radial"
- In:Venous and Small Vessel Invasion: "Large vessel", "Small vessel"
- En:Distant Spread or Metastases: "Distant spread", "Metastases"
- Re:TILS and Peritumoural Lymphocytes: "Tils", "Peritumoural lymphocytes ", "Crohn's-like reaction"
- Ma:Excision Clear: "Invasive", "In-situ", "Deep"

TILs Module

This module extracts any lexical items in particular gazetteers from the input as the output.

For the melanoma corpus, the input is the candidate or surrounding Li:Mood and Comment Adjuncts entities; particular gazetteers refer to the distribution and density gazetteers.

For the colorectal cancer corpus, the input is the candidate; particular gazetteers refer to the distribution, density or degree gazetteers; an additional step is to map the lexical items in the degree gazetteer to six predefined categories: "minimal", "mild", "mild to moderate", "moderate", "moderate to marked" and "marked".

Mood Degree Module

The Li:Mood and Comment Adjuncts entities were assigned with various scores according to the degree or intensity they indicated, ranging from 0.5 to 3 (e.g., "minimal": 0.5, "mild": 1, "moderate": 2, "prominent": 3). The module returns the Li:Mood and Comment Adjuncts entities around the candidate and their associated scores, which can be used by mood degree criterion or contribute to the population to the associated field. Note that for a Re:Tissue Reaction candidate, the lexical items of the extracted Li:Mood and Comment Adjuncts entities need to be verified against the specific mood gazetteer.

Subheading Module

This module verifies whether the candidate is a synoptic field with an associated subheading. First, the candidate is checked if it has one of these patterns: colon ":", hyphen "-", or two titlecase tokens inside the candidate. If it has, the potential subheading is extracted from the candidate, spanning from the first character to the previous character before the punctuation or the second titlecase character. The potential subheading is validated against particular subheading lexical items by entity types. For instance, "Site" is a valid subheading for a De:Tumour Site candidate "Site: Caecum"; whereas, "Polyps" is not a valid subheading for an En:Coexistent Pathology candidate "Polyps: Present, benign hyperplastic". The valid subheadings are used for ranking, and then stripped from the candidate to facilitate other processes.

Tumour Border Status Module

This module extracts any lexicon belonging to an entry in the tumour border status gazetteer.

Tumour Description Module

The module verifies whether the candidate matches one of these patterns:

Pattern 1: a non-propositional phrase + an entry in the general tumour gazetteer

Pattern 2: an adjectival phrase without any preposition

Pattern 3: a noun phrase without preposition "to" and plural nouns

Pattern 4: non-prepositional phrase 1 + an entry in the general tumour gazetteer + preposition "with" + non-propositional phrase 2

Pattern 5: an entry in the general tumour gazetteer + verb "is" + a non-prepositional phrase

For Pattern 1 and Pattern 5, the non-prepositional phrase is extracted as the result; for Pattern 4, both non-prepositional phrases are extracted as the result; for Pattern 2, the adjectival phrase is extracted as the result; for Pattern 3, the noun phrase is extracted as the result.

After revision, the main changes were the adjustment of lexical entries in the general tumour gazetteer.

Clear Processor

This module returns the result as "clear" if lexical items "clear" or "clearance" is detected in the candidate.

Tumour Site Processor

This module aims to extract the anatomical site and laterality from the candidate. It tries to remove lexical items from the candidate if they are the entries in the tumour gazetteer, the determiner "the" and prepositions like "in", "of" and "to".

After revision, the tumour gazetteer was introduced with more lexical entries, and the module also converted two very frequent terms: "rectal" and "caecal" to "rectum" and "caecum" if applicable, to enhance the usability.

Specimen Length Processor

After the first round error analysis, the module, Specimen Length Processor, was required to validate the candidate (the details are discussed in Section 8.5.1), with the processes below:

1. Validate the dimensions. By counting the keys in dimension pairs from Dimension Processor on all candidates, the candidate is validated against whether it is the only one candidate with a dimension of "length", "dimension", "area" or "volume".
2. Validate the lexical items. The candidate is validated against: a) whether it obtains a positive score from the specimen site criterion; b) whether it obtains a zero score from the specimen site criterion. The criterion is depicted below:

   It returns variable values depending on the amount of lexical entries in the specimen site gazetteer the candidate has; else 0. The entries in the specimen site gazetteer can be sub-classified to preferable terms and unfavourable terms. For a preferable term detected in the candidate, the criterion gain +1 score, while, a -1 score is obtained if the candidate has an unfavourable term.
3. Validate the position. The candidate is validated against whether it appears first in all candidates.
4. Validate the totality. The candidate is validated against whether it obtains a positive score from the total criterion (see Section I.2).
5. Validate the size. The candidate is validated against whether it obtains a positive score from the maximum measurement criterion (also see Section I.2).

The valid patterns are described as follows:

Pattern 1: the candidate satisfies Conditions 1, 2a, and 3.

Pattern 2: the candidate satisfies Conditions 3 and 4.

Pattern 3: the candidate satisfies Conditions 2b, 3 and 5.

Pattern 4: the candidate satisfies Conditions 1, 3 and 5.

If the candidate matches one of the above patterns, it is verified as valid.

Descriptor Convertor

Given the descriptors defined as the possible values for some fields (e.g., "Cell size" and "Cytomorphology"), the convertor standardizes the lexical variants in the candidates of associated entity types to these descriptors if applicable.

Id Validation Module

This module aims to solve an issue recognized in the first round error analysis that the system occasionally cannot assign a value to a specimen id under certain sections in a multiple specimen document. The basic idea is to determine the specimen id(s) for the population of fields under some section contexts without sufficient ids by checking whether the id(s) occur in other specific section contexts. For example, as the "SPECIMEN" section usually lacks specimen ids, the Sample Triage candidates are hard to assign for each specimen id. In contrast, the "FROZEN SECTION" often contains specimen ids. This module can determine whether a Sample Triage entity referred to in the "SPECIMEN" section can be populated for a particular specimen id by checking whether this id occurs in the "FROZEN SECTION".

Special Candidate Selection Module

This is a complicated module designed to handle special cases where the best candidate(s) cannot be found by ranking but rather by particular rules.

One of the examples is finding the appropriate De:Cell Growth Pattern candidates. Initially, all candidates were considered, unless their assertions are absent; after revision, their assertions were restricted to present and probable.

Another example is choosing candidates for fields "Prev. Rx / Trauma" and "Other medical history" from De:Cosmetic Changes entities. Firstly, an entity is checked against: a) if the lexical items inside it are in the trauma/treatment gazetteer; b) if there is a valid result from the temporality module or the entity has the lexicon "following"; c) if the entity does not have the lexicon "scar". If the entity satisfies Conditions a and b, it is classified to a candidate of "Prev. Rx / Trauma"; else, if it satisfies Condition c, it is categorized to a candidate of "Other medical history".

I.2 Ranking Criteria

The motivation for creating ranking criteria was when multiple candidates for a field were present, only one or some of them should be used to populate to the field. Here is an example:

> … a raised tan-brown tumour ["De:Tumour Description"] with a polypoid surface partially covered in fibrinous exudate ["De:Tumour Description"], measuring 45mm proximal to distal ["De:Tumour Size"], 30mm in height ["De:Tumour Size"], 60mm wide ["De:Tumour Size"] and occupying approximately 65% of the mucosal circumference ["Ex:Extent"]…

There are up to three candidates for "Tumour size": 45mm, 30mm and 60 mm, but apparently only 60mm should be populated to the field.

Ranking criteria were targeted to find the best candidate(s) by comparing the measure of each candidate resulting from a set of criteria; that is, a potential candidate was assigned a salience measure based on the criteria, and the one with the highest salience measure was selected as the best candidate.

Span Length Criterion

This criterion returns 1, if the candidate has the longest text span; else 0.

Uppercase Criterion

Thorough analyses show that pathologists tended to indicate the significance of certain terms by using uppercase in the melanoma corpus, while a De:Tumour Site entity in uppercase usually represents a non-specific location in the colorectal cancer corpus.

For the melanoma corpus: This criterion returns 1, if the candidate is in uppercase; else 0.

For the colorectal cancer corpus: This criterion returns -1, if the candidate is in uppercase; else 0.

Negation Uncertainty Inapplicability Criterion

If inapplicability is detected by the negation and uncertainty detection modules described in Chapter 6, this criterion returns -1; if uncertainty is detected by the modules, it returns various scores depending on the category of the uncertainty: 0.3 for "cannot exclude", 0.5 for "possible", 0.8 for "probable" and 1 for "definite".

For the colorectal corpus: If negation is detected by the modules, it returns -1; if the assertion of the candidate is present, it returns 1.

For the other two corpora: Initially, if the assertion of the candidate is present or absent, it returns 1. After revision, if the assertion of the candidate is absent, it returns -1. If the assertion of the candidate is present, it returns 1; for particular entity types, it returns 2, if a Li:Lexical Polarity Positive term is also present in the same sentence with the candidate, given the influence of the term. These entity types are Sy:Diagnosis and Sy:Subtype.

Measurement Criterion

Detailed analyses indicate that the representation of measurement in the melanoma corpus is simpler than that in the colorectal cancer corpus, thus extra weights of score were prepared for candidates in the colorectal cancer corpus.

For the melanoma corpus: This criterion returns 1, if there is a valid result from the measurement module; else 0.

For the colorectal cancer corpus: Initially, it returns 2, if there is an exact numeric value obtained from the module; it returns 1, if there is a fuzzy numeric value obtained from the module; else 0. After revision, it returns 1 and 0.5 for the above first and second conditions respectively; it returns 0.8, if more than one valid result is obtained from the module.

Clear Criterion

This criterion returns 1, if a Ma:Excision Clear entity and the candidate are in the same sentence; else 0.

Frequency Criterion

This criterion returns variable scores according to the frequencies of the tokens inside the candidate occurring in other candidates in the same specimen id's (or ids') context; if none of the tokens occur in other candidates, it returns 0.

Note that there may be restrictions for the section contexts. For example, it excludes other candidates located in "CLINICAL HISTORY" for the colorectal cancer corpus.

It should also be noticed that for the lymphoma corpus, when the candidate is a De:Topography entity, not only other candidates of the same type, but also the ones of De:Anatomical Structure should also be taken into consideration in the application of this criterion. After revision, tokens like "lymph" and "nodes" were ruled out during ranking, as these are general terms which reduce the specificity of the criterion.

After the first round error analysis, the full form of the token should be used if it is an acronym or abbreviation.

Primary Criterion

This criterion returns 1, if the candidate refers to information about the primary lesion; else 0.

Temporality Criterion

This criterion returns 1, if the candidate is a Sy:Regression entity, and there is a valid result from the temporality module; it returns -1, if the candidate is a De:Topography or De:Anatomical Structure entity, and the result does not contain any of the lexical items "now", "current" and "currently"; it returns -1, if the candidate is an entity of other entity types, and there is a valid result from temporality module; else 0.

Body Structure Criterion

This criterion returns 1, if the medical category in SNOMED CT of the candidate is "Body structure"; else 0.

Laterality Criterion

This criterion returns 1, if the candidate contains an entry in the laterality gazetteer; else 0.

Note that several lexical items were added to the gazetteer after first round error analysis.

Melanoma Criterion

Initially, this criterion returns 1, if the candidate contains an entry in the melanoma gazetteer; else 0.

After revision, it was adjusted as follows: it returns 2, if the candidate contains the lexicon "melanoma"; 1, if the candidate contains another lexical entry in the melanoma gazetteer; else 0.

Naevus Type Criterion

This criterion returns 1, if there is a valid result from the naevus module; else 0.

Level Criterion

This criterion returns 1, if there is a valid result from the level module; else 0.

Regress Criterion

This criterion returns 1, if there is a valid result from the regress module; else 0.

Dimension Criterion

This criterion returns 1, if the candidate is a two or three dimensional size; else 0.

Position Criterion

Thorough analyses suggest that the requirement for applying this criterion is more stringent for the melanoma corpus that those for the lymphoma corpus, hence the conditions are specified for the two corpora.

For the melanoma corpus: This criterion returns 1, if the candidate appears first in all candidates; else 0. After revision, it was applied under a specific condition that the sentence where the candidate locates should have only one De:Specimen Type entity.

For the lymphoma corpus: This criterion returns 1, if the candidate appears first or in the same sentence with the one that appears first; else 0.

Specimen Distance Criterion

Initially, this criterion returned 1, if a De:Specimen type entity or the lexicon "specimen" and the candidates are in the same sentence; else 0. After revision, a four-word window size of the candidate is a specific condition that restricts the application of it for the melanoma corpus.

Margin Criterion

This criterion returns 1, if the candidate is a Ma:Excision Clear entity and co-occurs with any entity of Ma:Excision In Situ, Ma:Excision Invasive, or Ma:Excision Deep in the same sentence; else 0.

Distribution Density Degree Criterion

This criterion returns variable scores depending on the amount of valid results from the TILs module; for each valid result, it gains a +1 score.

Rate Criterion

This criterion returns 1, if there is a valid result from the rate module; else 0.

Acronym Criterion

This criterion returns -1, if the candidate has an acronym; else 0.

Margin Type Criterion

This criterion returns 1, if there is a valid result from the sub-classification module; else 0.

Invasive Criterion

This criterion returns 1, if the candidate refers to information about an invasive lesion; else 0.

Mood Degree Criterion

This criterion returns a score > 0 according to the result from the mood degree module; else 0. Note that if the candidate is a De:Architecture entity, it returns 1, if there is a valid result from the mood degree module.

Diagnosis Criterion

This criterion returns 2, if the candidate is adjacent to a Sy:Diagnosis entity; it returns 1, if they are in the same sentence; else 0.

Breslow Criterion

This criterion returns 1, if the candidate is a "Breslow thickness"; else 0.

Specific Criterion

This criterion returns 1, if the candidate represents a specific biopsy type; else 0.

Type Criterion

This criterion returns 1, if the candidate ends with lexical items "type" or "pattern"; else 0.

Summary Criterion

This criterion returns 1, if there is a valid result from the subheading module; else 0.

Tumour Site Criterion

In this criterion, the lexical items inside the candidate are verified against the tumour site gazetteer, which includes preferable terms and unfavourable terms; for each preferable term, a +1 score is gained, while, a -1 score is assigned for each unfavourable term. If the candidate is a De:Specimen Type entity, additional unfavourable terms should be considered.

Sub-classification Criterion

This criterion returns 2, if the result from the sub-classification module meets the requirement of the field; it returns a score >=1, if part of the result meets the requirement of the field; else 0.

For example, for "Venous invasion", the sub-type requirement is "Large vessel". For a In:Venous and Small Vessel Invasion candidate "venous invasion: not identified", the criterion returns 2, as its sub-type result is "Large vessel"; for another In:Venous and Small Vessel Invasion candidate "lymphovascular invasion: present", the criterion returns 1, since it has a sub-type result: "Large vessel" and "Small vessel".

Maximum Measurement Criterion

This criterion returns 1, if the result from the Dimension Processor is the maximum one in all candidates; else 0.

After revision, this criterion cannot not be applied to a De:Tumour Size entity if the report is a multiple tumour document.

Medical Category Criterion

Initially, this criterion returns 1, if the medical category in SNOMED CT of the candidate is "Body structure" or "Procedure"; else 0.

After revision, "Body structure" was excluded from the criterion.

Specimen Length Criterion

This criterion returns 2, if there is a valid result from the Dimension Processor and the result is a measured length; it returns 1, if the result is another measured dimension instead of length; else 0.

Size Criterion

This criterion returns 2, if there is a valid result from the Dimension Processor and the result is a measured volume or area; it returns 1, if the result is another measured dimension; else 0.

Tumour Description Criterion

The five patterns for verifying the candidates have been described in the tumour description module. This criterion returns 3, if the candidate matches Patterns 1, 2, or 4; it returns 2, if the candidate matches Pattern 3; it returns 1, if the candidate matches Pattern 5; else 0.

After revision, additional lexical items "largest", "polyp" and "nodule" were used to verify against the lexical items inside the candidate if it matches Pattern 3 or 5; for each match of these lexical items, a +1 score is assigned.

Location Criterion

This criterion returns 1, if the candidate indicates the relationship of the tumour to the anterior peritoneal reflection; else 0.

Involvement number Criterion

This criterion returns 1, if there is a valid result from the involvement number module; It returns 2, if the result represents the total count; else 0.

Node number Criterion

This criterion returns 1, if there is a valid result from the node number module; it returns 2, if the result represents the total count; else 0.

T stage Criterion

This criterion returns 1, if the candidate has an entry in the T stage gazetteer; else 0.

R status Criterion

This criterion returns 1, if the candidate has an entry in the R status gazetteer; else 0.

Tumour Distance Criterion

This criterion returns 1, if a De:Tumour Description entity, lexical items "tumour" or "it", and the candidate is in the same sentence; else 0.

Procedure Criterion

This criterion returns 1, if the candidate is a specific surgical procedure, such as "right hemicolectomy", "anterior resection" and "Hartmann's procedure"; else 0.

Integrity Criterion

This criterion returns 1, if the candidate indicates the intactness of the mesorectum; else 0.

Tumour Boarder Status Criterion

This criterion returns 1, if there is a valid result from the tumour border status module; else 0.

Depth Criterion

This criterion returns 1, if the candidate has the keyword "thickness"; else 0.

The keyword was replaced with "depth" after revision.

Regression Grade Criterion

This criterion returns 1, if the candidate indicates the regression grade; else 0.

Maximum Dimension Criterion

This criterion returns 1, if the candidate represents the maximum measured dimension; else 0.

Specimen id Criterion

This criterion returns 1, if the candidate is in the same specimen id's (or ids') context with a De:Tumour Size or De:Tumour Description entity; else 0.

Abbreviation Criterion

This criterion returns 1, if the candidate is an abbreviation; else 0.

Revision Criterion

This criterion returns 1, if the candidate represents a revised classification of the stage; else 0.

Noun Phrase Criterion

This criterion returns 1, if the candidate has a noun (or nouns); else 0.

Grade Criterion

This criterion returns 1, if the candidate is an explicit expression of the grade, e.g., "grade 1"; it returns -1, if the candidate is an implicit expression of the grade, e.g., "lower grade"; else 0.

Total Criterion

This criterion returns 1, if the candidate represents the total size; else 0.

Cell Size Criterion

This criterion returns 1, if a De:Cell Size entity and the candidate are in the same sentence; else 0.

Architecture Criterion

This criterion returns 1, if the candidate has an entry in the architecture gazetteer; else 0.

The lexical entries in the gazetteer were modified after first round error analysis.

Pattern Criterion

Initially, this criterion returns 1, if the candidate has an entry in the pattern gazetteer; else 0.

After first round error analysis, the pattern gazetteer was divided into preferable terms and unfavourable terms, and the ranking process resembled that in specimen site criterion.

POS Criterion

This criterion returns 1, if the candidate has a noun or adjective; it returns -1, if the candidate has a verb; else 0.

Classification Criterion

This criterion returns 1, if the candidate is a grade from WHO ICD-10 classification system; it returns -1, if the candidate is a grade from other classification systems, such as the Revised European American Lymphoma Classification (REAL) and Working Formulation (WF); else 0.

Tissue Reaction Criterion

This criterion returns 1, if a Re:Tissue Reaction entity precedes the candidate, and they are in the same sentence, or a Re:Tissue Reaction entity succeeds the candidate within a three-token window; else 0.

Malignancy Criterion

This criterion returns -1, if the candidate does not indicate a specific malignant disease; else 0.

Addition Criterion

This criterion returns 1, if the candidate represents an additional size; else 0.

Appendix II Application of the Post-processing Modules and Ranking Criteria for the Structured Fields in Each Corpus

II.1 Melanoma Corpus

Template section		Template item	Post-processing module	Ranking criterion
Diagnostic Summary		Summary		
		Comment		
Supporting Information	CLINICAL	Description		
		Site and laterality		Span length criterion, uppercase criterion, frequency criterion, body structure criterion, laterality criterion
		Clinical diagnosis	Special candidate selection module	Negation uncertainty inapplicability criterion
		Specimen type		Span length criterion, temporality criterion*, specific criterion*
		Prev. Rx / Trauma	Special candidate selection module	
		Previous melanoma	Special candidate selection module	
		Distant metastasis	Special candidate selection module	Negation uncertainty inapplicability criterion
		Other medical history	Special candidate selection module	
	MACROSCOPIC	Description		
		Size of specimen	Dimension Processor	Dimension criterion, position criterion, specimen distance criterion
		Other lesions		Negation uncertainty inapplicability criterion
	MICROSCOPIC	Description		
		Diagnosis		Uppercase criterion, frequency criterion, negation uncertainty inapplicability criterion, acronym criterion*, melanoma criterion
		Tumour thickness	Measurement module	Measurement criterion, Breslow criterion*
		Excision margins: Invasive	Clear Processor*, measurement module, sub-classification module*	Measurement criterion, clear criterion, primary criterion*, margin criterion, margin type criterion*, invasive criterion*

		Excision margins: In-situ	Clear Processor, measurement module, sub-classification module*	Measurement criterion, clear criterion, primary criterion*, margin criterion, margin type criterion*
		Excision margins: Deep	Clear Processor, measurement module, sub-classification module*	Measurement criterion, clear criterion, primary criterion*, margin criterion, margin type criterion*
		Ulceration (mm diam)	Measurement module	Span length criterion, uppercase criterion, negation uncertainty inapplicability criterion, measurement criterion
		Mitotic rate	Rate module	Rate criterion
		Microsatellites		Negation uncertainty inapplicability criterion
		Level of invasion (Clark)	Level module	Uppercase criterion, level criterion
		Lymphovascular invasion		Span length criterion, uppercase criterion, negation uncertainty inapplicability criterion
		TILs		Negation uncertainty inapplicability criterion, acronym criterion*
		TILs: Distribution	TILs module	Distribution density degree criterion
		TILs: Density	TILs module	Distribution density degree criterion
		Regression	Temporality module, regress module	Temporality criterion, regress criterion, negation uncertainty inapplicability criterion
		Desmoplasia		Negation uncertainty inapplicability criterion
		Neurotropism		Negation uncertainty inapplicability criterion
		Assoc. benign naevus	Naevus module, special candidate selection module*	Span length criterion, frequency criterion*, negation uncertainty inapplicability criterion, naevus type criterion, mood degree criterion*
		Cell growth	Special candidate selection module	
		Subtype		Frequency criterion*, negation uncertainty inapplicability criterion, acronym criterion*, diagnosis criterion*, type criterion*

Note: Module or criterion marked with * means it was added after the first round error analysis.

II.2 Colorectal Cancer Corpus

Template section		**Template item**	**Post-processing module**	**Ranking criterion**
Diagnostic Summary		Summary		
		Comment		
Supporting Information	CLINICAL	Site	Tumour Site Processor	Measurement criterion, tumour site criterion
		Other sites of disease		
		Medical history		
	MACROSCOPIC	Specimen type	Subheading module	Summary criterion, tumour site criterion, medical category criterion, procedure criterion, abbreviation criterion*, frequency criterion*
		Tissue banking		Negation uncertainty inapplicability criterion
		Specimen images		Negation uncertainty inapplicability criterion
		Specimen length	Dimension Processor, Specimen Length Processor*	Maximum measurement criterion, specimen length criterion, specimen distance criterion, specimen id criterion
		Tumour site	Tumour Site Processor, subheading module	Measurement criterion, summary criterion, tumour site criterion, uppercase criterion*, specimen id criterion*
		Peritoneal reflection	Descriptor Convertor, subheading module	Summary criterion, location criterion
		Mesorectal integrity	Descriptor Convertor	Integrity criterion
		Tumour size	Dimension Processor, subheading	Summary criterion, maximum measurement criterion, size criterion, tumour distance criterion, maximum dimension criterion

			module	
		Extramuscular spread	Measurement module, tumour boarder status module	Measurement criterion, tumour boarder status criterion
		Tumour description	Subheading module, tumour description module	Summary criterion, tumour description criterion
		Overlying serosa	Subheading module*	Summary criterion*
		Perforation	Subheading module	Summary criterion, negation uncertainty inapplicability criterion
		Margins:Proximal	Measurement module, subheading module, Clear Processor*	Measurement criterion, summary criterion, sub-classification criterion
		Margins:Distal	Measurement module, subheading module, Clear Processor*	Measurement criterion, summary criterion, sub-classification criterion
		Margins:Radial	Measurement module, subheading module, Clear Processor*	Measurement criterion, summary criterion, sub-classification criterion
		Lymph nodes	Node number module, subheading module	Summary criterion, node number criterion
		Metastases	Subheading module	Summary criterion, sub-classification criterion, negation uncertainty inapplicability criterion
		Blocks selected		
		Comment		
	MICROSCOPIC	Histological type (WHO)	Subheading module	Summary criterion, negation uncertainty inapplicability criterion
		Histological grade	Subheading module	Summary criterion

		Depth of invasion	Subheading module	Measurement criterion, summary criterion, T stage criterion, depth criterion
		Serosal involvement	Subheading module	Summary criterion, negation uncertainty inapplicability criterion
		Small vessel invasion	Subheading module	Summary criterion, sub-classification criterion, negation uncertainty inapplicability criterion
		Venous invasion	Subheading module	Summary criterion, sub-classification criterion, negation uncertainty inapplicability criterion
		Perineural invasion	Subheading module	Summary criterion, negation uncertainty inapplicability criterion
		TILs	TILs module, sub-classification module	Summary criterion, sub-classification criterion, distribution density degree criterion, negation uncertainty inapplicability criterion
		Margins:Proximal	Measurement module, subheading module, Clear Processor*	Measurement criterion, summary criterion, sub-classification criterion
		Margins:Distal	Measurement module, subheading module, Clear Processor*	Measurement criterion, summary criterion, sub-classification criterion
		Margins:Radial	Measurement module, subheading module, Clear Processor*	Measurement criterion, summary criterion, sub-classification criterion
		Lymph nodes	Node number module, subheading module	Summary criterion, node number criterion
		Number involved	Involvement number module, subheading module	Summary criterion, involvement number criterion
		Distant spread	Subheading module	Summary criterion, sub-classification criterion, negation uncertainty inapplicability criterion
		Response to Rx		Regression grade criterion
		Comment		
	ANCILLARY STUDIES	Description		
	SYNTHESIS	TNM stage:T	Subheading module*	Revision criterion*

		TNM stage:N	Subheading module*	Revision criterion*
		TNM stage:M	Subheading module*	Revision criterion*
		Stage group	Subheading module	Revision criterion*
		Residual tumour (R)	Subheading module	Summary criterion, R status criterion
		Comment		

Note: Module or criterion marked with * means it was added after the first round error analysis.

II.3 Lymphoma Corpus

Template section		Template item	Post-processing module	Ranking criterion
Diagnostic Summary		Summary		
		Comment		
Supporting Information	CLINICAL	Site and laterality		Noun phrase criterion, frequency criterion, temporality criterion, laterality criterion
		Presentation	Temporality module	
		Indication for biopsy	Descriptor Convertor	
		Clinical impression		Negation uncertainty inapplicability criterion
		Disease extent		
		Other sites of disease		Negation uncertainty inapplicability criterion
		Const. symptoms		Negation uncertainty inapplicability criterion
		Medical history		Malignancy criterion*
		Predisposing factors		Negation uncertainty inapplicability criterion
	SPECIMEN	Specimen type	Descriptor Convertor	Frequency criterion*, medical category criterion, procedure criterion, specific criterion
		Size	Dimension Processor	Position criterion, size criterion, total criterion, addition criterion*
		Received in		
		Triage	Id validation module*	
		Description		
	MICROSCOPIC	Pattern of infiltration	Descriptor Convertor	Negation uncertainty inapplicability criterion, mood degree criterion, cell size criterion, architecture criterion, pattern criterion
		Cell size	Descriptor	Position criterion, tissue reaction criterion*

			Convertor	
		Cytomorphology	Descriptor Convertor	Negation uncertainty inapplicability criterion
		Tissue reactions	Mood degree module	Negation uncertainty inapplicability criterion, POS criterion*
		Grade		Negation uncertainty inapplicability criterion, grade criterion, classification criterion*
		Description		
	IMMUNOPHENOTYPING	Immunohistochemistry: Positive for	Id validation module	
		Immunohistochemistry: Negative for	Id validation module	
		Immunohistochemistry: Equivocal for	Id validation module	
		Immunohistochemistry: Comment	Id validation module	
		Flow cytometry: Positive for	Id validation module	
		Flow cytometry: Negative for	Id validation module	
		Flow cytometry: Comment	Id validation module	
	CYTOGENETICS	FISH	Id validation module	
		Cytogenetics: Comment	Id validation module	
	MOLECULAR	PCR: IgH	Id validation module	
		PCR: TCRgamma	Id validation module	

		PCR: Comment	Id validation module	
	SYNTHESIS	Lineage	Descriptor Convertor	
		Clonality		
		Diagnosis (WHO)		Negation uncertainty inapplicability criterion
		SNOMED RT Codes		
		Stage		
		Comment		

Note: Module or criterion marked with * means it was added after the first round error analysis.

Appendix III Output Examples for Some Structured Fields in Each Corpus

III.1 Melanoma Corpus

Template section		Template item	Input	Output
Diagnostic Summary		Summary		
		Comment		
Supporting Information	CLINICAL	Description		
		Site and laterality	(1) L Arm (2) (R) face	(1) left arm (2) right face
		Clinical diagnosis	(1) ? ["Li:Modality"] lentigo maligna ["Sy:Subtype"] (2) exclude["Li:Lexical Polarity Negative"] malignant melanoma ["Sy:Diagnosis"]	(1) possible lentigo maligna (2) no malignant melanoma
		Specimen type		
		Prev. Rx / Trauma	previous ["Li:Temporality"] surgical procedure ["De:Cosmetic Changes"]	surgical procedure (previous)
		Previous melanoma	Past history ["Li:Temporality"] of malignant melanoma ["Sy:Diagnosis"]	present
		Distant metastasis	metastasis	present
		Other medical history	Skin lesion ["En:Primary Lesion"] from right thigh ["De:Site and Laterality"] (small ["Li:Mood and Comment Adjuncts"] recent ["Li:Temporality"] increase in size ["De:Cosmetic Changes"])	lesion; increase in size (recent)
	MACROSCOPIC	Description		
		Size of specimen	(1) 4mm in diameter and 2mm in depth (2) 9.5mm dia (3) 7x6mm (4) 5 x 6 mm to a depth of 3 mm (5) 65mm in length, width of 30mm and maximum thickness of 13mm (6) 15mm in maximal dimension	(1) 4mm diameter and 2mm depth (4mm dia x 2mm) (2) 9.5mm diameter (3) 7mm x 6mm (4) 5mm x 6mm x 3mm (5) 65mm x 30mm x 13mm

				(6) 15mm
		Other lesions	A second, separate area ["En:Lesion (other)"] of purple discolouration ["De:Shape"] is present ["Li:Lexical Polarity Positive"]	present
	MICROSCOPIC	Description		
		Diagnosis	(1) no ["Li:Lexical Polarity Negative"] evidence ["Li:Lexical Polarity Positive"] of malignancy ["Sy:Diagnosis"] (2) sections ["En:Specimen Identifier"] show ["Li:Lexical Polarity Positive"] a malignant melanoma ["Sy:Diagnosis"]	(1) no malignancy (2) malignant melanoma
		Tumour thickness	(1) thickness of 5.0mm (2) maximum depth of 1.0mm	(1) 5.0mm (2) 1.0mm
		Excision margins: Invasive	(1) Excision appears complete ["Ma:Excision Clear"]; Clearance values are 2.1mm and 1.1mm laterally ["Ma:Excision Invasive"] and 0.7mm to the deep surface ["Ma:Excision Deep"]. (2) nearest margin of excision is 2.4mm	(1) clear - 2.1mm and 1.1mm (2) 2.4mm
		Excision margins: In-situ	(1) clear of the resection margins ["Ma:Excision Clear"] with a minimum measured deep clearance (from the invasive component) of 4.3mm ["Ma:Excision Invasive"] and a minimum measured lateral clearance (from the intraepidermal in-situ component) of 2.3mm ["Ma:Excision In Situ"]. (2) close to one lateral border, within 0.2mm	(1) clear - 2.3mm (2) 0.2mm
		Excision margins: Deep	(1) Excision appears complete ["Ma:Excision Clear"]; Clearance values are 2.1mm and 1.1mm laterally ["Ma:Excision Invasive"] and 0.7mm to the deep surface ["Ma:Excision Deep"]. (2) deep margin is 1.4mm	(1) clear - 0.7mm (2) 1.4mm
		Ulceration (mm diam)	(1) THE SURFACE ULCERATION MEASURES 4.5MM (2) no ["Li:Lexical Polarity Negative"] ulceration ["De:Ulceration"]	(1) present (4.5mm) (2) absent
		Mitotic rate	(1) four to eight mitoses per high power field (2) Mitoses are less than ten per high powered fields (3) average 3-4 per mm square (4) mitotic activity of up to 2 mitoses/mm2 (5) average 1 per 5 high power fields	(1) 4-8/hpf (2) less than 10/hpf (3) 3-4/mm2 (4) 2/mm2 (5) 1/5 hpf
		Microsatellites	? ["Li:Modality"] SATELLITE FOCUS ["En:Satellites"]	possible
		Level of invasion (Clark)	(1) Clark level IV (2) Clark level 3-4	(1) IV (2) III-IV

		Lymphovascular invasion	no ["Li:Lexical Polarity Negative"] vascular invasion ["In:Vascular/Lymphatic"] identified ["Li:Lexical Polarity Positive"]	absent
		TILs	Scanty ["Li:Mood and Comment Adjuncts"] tumour infiltrates of lymphocytes ["Re: TILs"] are noted ["Li:Lexical Polarity Positive"].	present
		TILs: Distribution	(1) heavy ["Li:Mood and Comment Adjuncts"] band-like lymphocytic infiltrate ["Re:TILs"] (2) sparse ["Li:Mood and Comment Adjuncts"] patchy ["Li:Mood and Comment Adjuncts"] lymphoid infiltrate ["Re:TILs"]	(1) band-like (2) patchy
		TILs: Density	(1) heavy ["Li:Mood and Comment Adjuncts"] band-like lymphocytic infiltrate ["Re:TILs"] (2) sparse ["Li:Mood and Comment Adjuncts"] patchy ["Li:Mood and Comment Adjuncts"] lymphoid infiltrate ["Re:TILs"]	(1) heavy (2) sparse
		Regression	(1) possible ["Li:Modality"] partial regression ["Sy:Regression"] (2) in keeping with ["Li:Lexical polarity Positive"] active regression ["Sy:Regression"] (3) consistent with ["Li:Lexical polarity Positive"] early ["Li:Temporality"] regression ["Sy:Regression"]	(1) possibly (partial) (2) present (active) (3) present (early)
		Desmoplasia	no ["Li:Lexical Polarity Negative"] obvious ["Li:Mood and Comment Adjuncts"] desmoplasia ["Re:Desmoplasia"]	absent
		Neurotropism	No ["Li:Lexical Polarity Negative"] perineural invasion ["In:Neurotropism"] is seen	absent
		Assoc. benign naevus	(1) arising from a dysplastic naevus (2) no ["Li:Lexical Polarity Negative"] associated naevus ["En:Associated naevus (type)"]	(1) dysplastic naevus (2) absent
		Cell growth	an asymmetrical poorly circumscribed proliferation ["De:Cell Growth Pattern"] of atypical melanocytes ["De:Cell Type"] arranged in confluent units and nests ["De:Cell Growth Pattern"]	asymmetrical poorly circumscribed proliferation, confluent units and nests
		Subtype	(1) malignant melanoma ["Sy:Diagnosis"] of superficial spreading type ["Sy:Subtype"] (2) Sections ["En:Specimen Identifier"] show ["Li:Lexical Polarity Positive"] a nodular ["Sy:Subtype"] malignant melanoma ["Sy:Diagnosis"]	(1) superficial spreading (2) nodular

Note: Multiple samples are separated by semicolon ";"; sample text without specification suggests the whole text is an entity of the associated medical entity type.

III.2 Colorectal Cancer Corpus

Template section		Template item	Input	Output
Diagnostic Summary		Summary		
		Comment		
Supporting Information	CLINICAL	Site	(1) Ca.rectum (2) Ca.R.colon (3) 4 cm from anal wrge	(1) rectum (2) right colon (3) 4 cm from anal verge
		Other sites of disease		
		Medical history		
	MACROSCOPIC	Specimen type	(1) Specimen type: Extended right hemicolectomy (2) AP resection	(1) extended right hemicolectomy (2) abdominoperineal resection
		Tissue banking	Tissue Banking - not done	no
		Specimen images	Macroscopic Photos - not taken	no
		Specimen length	(1) A length of large bowel 200mm ["De:Specimen Size"] with attached mesocolon 70mm wide ["De:Specimen Size"] (2) An anterior resection ["De:Specimen Type"] specimen comprising recto-sigmoid colon measuring 150mm ["De:Specimen Size"] with attached peri-colic fat up to 60mm ["De:Specimen Size"] (3) A right hemicolectomy ["De:Specimen Type"] consisting of terminal ileum (45mm in length and 35mm in circumference) ["De:Specimen Size"], caecum and ascending colon (120mm in length and 80mm in circumference) ["De:Specimen Size"], mesenteric fat (115mm in width) ["De:Specimen Size"]	(1) 200mm (2) 150mm (3) N/A
		Tumour site	(1) Located within the transverse colon ["De:Tumour Site"] (165mm distal to ileocaecal valve ["De:Tumour Site"], and 55mm proximal to the distal resection margin ["Ma:Proximal or Distal Margin"])	(1) transverse colon, 165mm distal to ileocaecal valve (2) caecum, just above the

			(2) Site of tumour : Caecum, just above the ileocaecal valve	ileocaecal valve
		Peritoneal reflection	(1) 10mm above the anterior peritoneal reflection (2) below the level of the peritoneal reflection (3) straddling the line of peritoneal reflection	(1) above (2) below (3) astride
		Mesorectal integrity	(1) distal half of the mesorectal excision is incomplete (2) status of mesorectal excision: intact (3) intactness of mesorectum: complete	(1) incomplete (2) complete (3) complete
		Tumour size	(1) 22 x 18mm (2) 28mm in axial length and 25mm in transverse dimension (3) 32mm in axial length and 35mm in width (4) 4.0x4.0 cm	(1) 22mm (2) 28mm (3) 35mm (4) 40mm
		Extramuscular spread	(1) just beyond it to a depth of 0.8 mm (2) tumour edge – infiltrative (3) broad pushing front	(1) 0.8mm (2) infiltrative (3) pushing
		Tumour description	(1) annular constricting tumour (2) brown ulcerated tumour with rounded raised border (3) bulky, ulcerated (4) central area of ulceration (5) bulk of the tumour is exophytic	(1) annular constricting (2) brown ulcerated with rounded raised border (3) bulky, ulcerated (4) central area of ulceration (5) exophytic
		Overlying serosa	(1) serosa is smooth (2) Overlying serosa: Puckered	(1) serosa is smooth (2) puckered
		Perforation	(1) perforated area near to the tumour (2) Perforation: Present, 12 x 8 mm in area, 30 mm proximal to tumour	(1) present (2) present
		Margins:Proximal	(1) distal and proximal resection margins are clear of tumour ["Ma:Clear"] (at least 20mm ["Ma:Proximal or Distal Margin"]) (2) Distance from proximal margin: 3.5cm (measured from the tumour in sigmoid colon) ["Ma:Proximal or Distal Margin"]; Proximal resection margin: No involvement, as confirmed histologically" ["Ma:Clear"] (3) 75mm from the proximal ileal resection margin ["Ma:Proximal or Distal Margin"]	(1) clear (20mm) (2) clear (3.5cm) (3) 75mm
		Margins:Distal	(1) 110mm from the distal resection margin ["Ma:Proximal or Distal Margin"]; not seen on the free margin of the specimen ["Ma:Clear"] (2) Distance from distal margin: 2.5cm (measured from the tumour in rectum)	(1) clear (110mm) (2) clear (2.5cm) (3) 130mm

			["Ma:Proximal or Distal Margin"]; Distal resection margin: No involvement, as confirmed histologically ["Ma:Clear"] (3) 130mm from the distal resection margin ["Ma:Proximal or Distal Margin"]	
		Margins:Radial	(1) 0.1cm from the nearest radial margin ["Ma:Circumferential Margin"] (2) 1.1mm from the serosal margin ["Ma:Circumferential Margin"] (3) tumour extends to within 0.5mm of the visceral peritoneum ["Ma:Circumferential Margin"]; resection margin show unremarkable small and large bowel ["Ma:Clear"]	(1) 0.1cm (2) 1.1mm (3) clear (0.5mm)
		Lymph nodes	(1) 12 lymph nodes are identified, 2 to 6mm in greatest dimension (2) Approximately twenty lymph nodes found (3) No nodes are identified (4) No apical node is identified; Up to thirty-five mesenteric nodes have been submitted for histological assessment (5) 20 pericolonic lymph nodes were identified ranging from 3mm to 5mm in diameter; five high tie lymph nodes ranging from 4mm to 7mm in diameter were identified	(1) 12 (2) 20 (3) 0 (4) 35 (5) 25
		Metastases	(1) 3 separate tumour deposits,4-5mm in greatest dimension, are also present (2) No metastases	(1) present (2) absent
		Blocks selected		
		Comment		
	MICROSCOPIC	Histological type (WHO)	(1) mucinous adenocarcinoma (2) HISTOLOGICAL TUMOUR TYPE: Adenocarcinoma, NOS	(1) mucinous adenocarcinoma (2) Adenocarcinoma, NOS
		Histological grade	(1) moderately differentiated (2) Degree of differentiation: Moderately differentiated	(1) moderately differentiated (2) Moderately differentiated
		Depth of invasion	(1) Tumour invades through the mucosa into the inner layers of the muscularis propria (2) Local invasion: Beyond muscularis propria (pT3) (3) Tumour extends to lie approximately 0.2mm from the closest serosal surface	(1) Tumour invades through the mucosa into the inner layers of the muscularis propria (2) Beyond muscularis propria (pT3) (3) Tumour extends to lie approximately 0.2mm from the closest serosal surface
		Serosal involvement	(1) serosa is clear (2) invades the serosa	(1) absent (2) present

		Small vessel invasion	(1) focal infiltration of lymphatic vessels (2) Focal intra-lymphatic tumour permeation is highly suspicious	(1) present (2) probable
		Venous invasion	(1) Extramural vein invasion Not identified (2) Venous invasion is identified	(1) absent (2) absent
		Perineural invasion	(1) focal perineural invasion (2) Perineural invasion: No	(1) present (2) absent
		TILs	(1) Tumour infiltrating lymphocytes are inconspicuous (2) Tumour infiltrating lymphocytes and/or Crohn's-like inflammation: No (3) no significant peritumoral diffuse or nodular lymphocytic inflammatory response	(1) TILs: present (minimal). (2) TILs: absent. Crohn's like: absent. (3) Peritumoural: absent.
		Margins:Proximal	(1) 160mm from one surgical margin ["Ma:Proximal or Distal Margin"] and 140mm from the opposite surgical margin ["Ma:Proximal or Distal Margin"] (2) proximal ileal and distal colonic resection margins are clear of the tumour ["Ma:Clear"] (3) all resection margins are well clear of the tumour ["Ma:Clear"]	(1) 160mm/140mm (2) clear (3) clear
		Margins:Distal	(1) 160mm from one surgical margin ["Ma:Proximal or Distal Margin"] and 140mm from the opposite surgical margin ["Ma:Proximal or Distal Margin"] (2) proximal ileal and distal colonic resection margins are clear of the tumour ["Ma:Clear"] (3) all resection margins are well clear of the tumour ["Ma:Clear"]	(1) 160mm/140mm (2) clear (3) clear
		Margins:Radial	(1) nearest soft tissue resection margin to tumour appears to be that around dome of bladder, which is 10mm away ["Ma:Circumferential Margin"]; nearest mesenteric resection margin is 30mm away ["Ma:Circumferential Margin"] (2) "clear of the radical resection margin" ["Ma:Clear"] (3) all resection margins are well clear of the tumour ["Ma:Clear"]	(1) 10mm/30mm (2) clear (3) clear
		Lymph nodes	(1) All 8 lymph nodes dissected from the mesentery are small and reactive, including the apical ["En:Lymph Nodes"] (2) Ten separate regional mesenteric lymph nodes have been examined ["En:Lymph Nodes"] (3) One (1) of sixteen (16) lymph nodes show a small deposit of metastatic carcinoma ["Ex:Lymph Node Involvement"] (4) Node summary 1/57 ["Ex:Lymph Node Involvement"] (5) three lymph nodes were seen in fat associated with sections of the wall ["En:Lymph Nodes"]; Eight lymph nodes were isolated from pericolic fat	(1) 8 (2) 10 (3) 16 (4) 57 (5) 27 (6) 5

			["En:Lymph Nodes"]; Two lymph nodes are in identified pericolic fat associated with the wall of the bowel ["En:Lymph Nodes"]; Fourteen lymph nodes were isolated from pericolic fat ["En:Lymph Nodes"] (6) Three of the local lymph nodes shows metastatic mucinous tumour including the large proximal node ["Ex:Lymph Node Involvement"] ; "Two proximal ileocaecal nodes are clear of tumour" ["Ex:Lymph Node Involvement"]	
		Number involved	(1) 2 are infiltrated by malignant cells (2) Eleven benign lymph nodes identified (3) Three of thirteen lymph nodes are involved by metastatic carcinoma (4) 4 out of 16 lymph nodes show metastatic carcinoma (5) 1/26 (6) Three of the local lymph nodes shows metastatic mucinous tumour including the large proximal node; Two proximal ileocaecal nodes are clear of tumour (7) 10 are positive for tumour; apical lymph node is replaced by mucinous tumour with some associated scarring (8) seven of which contain metastatic tumour; six of which contain metastatic tumour	(1) 2 (2) 0 (3) 3 (4) 4 (5) 1 (6) 3 (7) 11 (8) 13
		Distant spread	(1) consistent with metastatic deposit in the bladder (2) Mesenteric deposits: Nil	(1) present (2) absent
		Response to Rx		
		Comment		
	ANCILLARY STUDIES	Description		
	SYNTHESIS	TNM stage:T		
		TNM stage:N		
		TNM stage:M		
		Stage group		
		Residual tumour (R)	(1) Residual tumour - none notified (2) Residual tumour (R) 0 (3) residual tumour	(1) none notified (2) R0 (3) residual tumour
		Comment		

Note: Multiple samples are separated by semicolon ";"; sample text without specification suggests the whole text is an entity of the associated medical entity type.

III.3 Lymphoma Corpus

Template section		Template item	Input	Output
Diagnostic Summary		Summary		
		Comment		
Supporting Information	CLINICAL	Site and laterality	(1) (L) ["De:Laterality"] recurrent ["Sy:Indication for Biopsy"] parotid ["De:Anatomical Structure"] tumour (2) cervical LN ["De:Topography"]	(1) left parotid (2) cervical lymph node
		Presentation	(1) Swelling ["Sy:Presentation"] Lt distal femur ["De:Topography"] & knee ["De:Anatomical Structure"] since 2mths ["Li:Temporality"] (2) Generalised ["Ex:Disease Extent"] lymphadenopathy ["Sy:Presentation"]	(1) swelling (since 2 months) (2) lymphadenopathy
		Indication for biopsy	(1) ? ["Li:Lexical Modality"] NHL recurrence ["Sy:Indication for Biopsy"] (2) Core bx ["De:Specimen Type"] inconclusive ["Sy:Indication for Biopsy"]	(1) relapse (2) core biopsy inconclusive
		Clinical impression	(1) ? ["Li:Lexical Modality"] lymphoma ["Sy:Clinical Impression"] (2) suspicions for ["Li:Lexical Modality"] Hodgkin's Lymphoma ["Sy:Clinical Impression"]	(1) possible lymphoma (2) possible Hodgkin's lymphoma
		Disease extent		
		Other sites of disease	(1) ? ["Li:Lexical Modality"] maybe ["Li:Lexical Modality"] spreading to kidney ["Ex:Other Sites of Disease"] (2) involving mediastinum	(1) possible (2) present
		Const. symptoms	(1) Night sweats ["Sy:Constitutional Symptoms"], weight loss ["Sy:Constitutional Symptoms"] (2) He initially experienced flu-like symptoms ["Sy:Constitutional Symptoms"] 6 weeks ago ["Li:Temporality"]	(1) night sweats, weight loss (2) flu-like symptoms
		Medical history		
		Predisposing factors	(1) Hep C (2) Post CTx	(1) hepatitis c (2) post chemotherapy

	SPECIMEN	Specimen type	(1) Excision Bx (2) core biopsies	(1) excision biopsy (2) core biopsy
		Size	(1) 8 and 3mm across (2) 40 x 20 x 15mm (3) 4 to 13mm across (4) 20mm in length and diameter 3mm	(1) 8 and 3mm (2) 40x20x15mm (3) 4-13mm (4) 20x3mm
		Received in		
		Triage		
		Description		
	MICROSCOPIC	Pattern of infiltration	(1) Sections show ["Li:Lexical Polarity Positive"] a diffuse proliferation ["De:Architecture"] of atypical lymphoid cells (2) though focal lymphoepithelial lesions ["De:Architecture"] are seen (3) scattered ["Li:Mood and Comment Adjuncts"] lymphoid follicles/nodules ["De:Architecture"]	(1) diffuse (2) lymphoepithelial lesions (3) follicular, nodular
		Cell size	(1) small lymphoid cells (2) small to intermediate lymphoid cells (3) variable size	(1) small (2) small to medium (3) mixed
		Cytomorphology	(1) Occasional binucleate Reed-Sternberg cells ["De:Cytomorphology"] are identified ["Li:Lexical Polarity Positive"]. (2) Most ["Li:Mood and Comment Adjuncts"] of the cells in the follicles ["De:Architecture"] are small ["De:Cell Size"] irregular cleaved centrocytes ["De:Cytomorphology"] (3) non-cleaved centroblasts ["De:Cytomorphology"]	(1) Reed-Sternberg (2) centrocytic (3) centroblastic
		Tissue reactions	(1) an ulcerated lesion with necrosis ["Re:Tissue Reaction"] (2) one of which consists mostly ["Li:Mood and Comment Adjuncts"] of tumourous tissue within sclerotic fibrous stroma ["Re:Tissue Reaction"]	(1) necrosis (2) sclerotic fibrous stroma
		Grade	(1) The features are of follicular lymphoma ["Sy:Diagnosis"], WHO grade 2 ["Sy:WHO Grade"]. (2) frequent mitoses suggests ["Li:Lexical Polarity Positive"] it is high grade ["Sy:WHO Grade"]	(1) WHO grade 2 (2) high grade
		Description		
	IMMUNOPHENOTYPING	Immunohistochemistry:	(1) Positive ["An:Immunohistochemistry-Positive"] - CD30	(1) CD30, CD15, CD20

		Positive for	["An:Biomarker"] +++ ["An:Immunohistochemistry-Positive"], CD15 ["An:Biomarker"] ++ ["An:Immunohistochemistry-Positive"], CD20 ["An:Biomarker"] ++ ["An:Immunohistochemistry-Positive"] (2) On immunohistochemical stains the cells show ["Li:Lexical Polarity Positive"] diffuse strong membranous staining ["An:Immunohistochemistry-Positive"] for CD20 ["An:Biomarker"], CD79a ["An:Biomarker"] and CD10 ["An:Biomarker"] with moderate widespread membrane staining ["An:Immunohistochemistry-Positive"] for CD30 ["An:Biomarker"]. (3) Immunohistochemical stains show ["Li:Lexical Polarity Positive"] that the abnormal lymphoid cells are positive ["An:Immunohistochemistry-Positive"] for CD10 ["An:Biomarker"], CD20 ["An:Biomarker"] and bcl-2 ["An:Biomarker"].	(2) CD20, CD79a, CD10, CD30 (3) CD10, CD20, bcl-2
		Immunohistochemistry: Negative for	(1) Negative ["An:Immunohistochemistry-Negative"] : CD10 ["An:Biomarker"], Cyclin D1 ["An:Biomarker"] (2) Stains for CD3 ["An:Biomarker"], CD5 ["An:Biomarker"] and cyclin D1 ["An:Biomarker"] are negative ["An:Immunohistochemistry-Negative"]. (3) Immunohistochemistry shows ["Li:Lexical Polarity Positive"] that the neoplastic cells are CD79a ["An:Biomarker"] + ["An:Immunohistochemistry-Positive"], CD20 ["An:Biomarker"] + ["An:Immunohistochemistry-Positive"], CD10 ["An:Biomarker"] + ["An:Immunohistochemistry-Positive"], CD23 ["An:Biomarker"] + ["An:Immunohistochemistry-Positive"], CD5 ["An:Biomarker"] - ["An:Immunohistochemistry-Negative"], cyclin D1 ["An:Biomarker"] - ["An:Immunohistochemistry-Negative"] and CD30 ["An:Biomarker"] -["An:Immunohistochemistry-Negative"].	(1) CD10, Cyclin D1 (2) CD3, CD5, cyclin D1 (3) CD5, cyclin D1, CD30
		Immunohistochemistry: Equivocal for	(1) Equivocal ["An:Immunohistochemistry-Equivocal"] - CD5 ["An:Biomarker"] (weak staining of atypical cells), bcl-6 ["An:Biomarker"] (some weak nuclear staining present). (2) Occasional cells show very weak and equivocal staining	(1) CD5, bcl-6 (2) CD30

			["An:Immunohistochemistry-Equivocal"] for CD30 ["An:Biomarker"].	
		Immunohistochemistry: Comment		
		Flow cytometry: Positive for	Positive ["An:Flow Cytometry-Positive"] for: Kappa ["An:Biomarker"], CD19 ["An:Biomarker"], CD10 ["An:Biomarker"], CD45 ["An:Biomarker"], CD38 ["An:Biomarker"]	Kappa, CD19, CD10, CD45, CD38
		Flow cytometry: Negative for	Negative ["An:Flow Cytometry-Negative"]: CD23 ["An:Biomarker"], kappa ["An:Biomarker"], CD10 ["An:Biomarker"]	CD23, kappa, CD10
		Flow cytometry: Comment		
	CYTOGENETICS	FISH		
		Cytogenetics: Comment		
	MOLECULAR	PCR: IgH		
		PCR: TCRgamma		
		PCR: Comment		
	SYNTHESIS	Lineage	(1) Malignant lymphoma, diffuse and follicular ["Sy:Diagnosis"], large ["De:Cell Size"] B cells ["De:Lineage"] predominating (2) HODGKINS LYMPHOMA ["Sy:Diagnosis"] (3) Precursor T lymphoblastic lymphoma ["Sy:Diagnosis"]	(1) B-cell (2) Hodgkin-like (3) T-cell
		Clonality		
		Diagnosis (WHO)	(1) WHO GRADE 1 ["Sy:WHO Grade"] FOLLICULAR LYMPHOMA ["Sy:Diagnosis"] (2) Diffuse large B cell lymphoma ["Sy:Diagnosis"] (up to 30%) and follicular lymphoma ["Sy:Diagnosis"] (3) LOW GRADE ["Sy:WHO Grade"] EXTRANODAL MARGINAL ZONE B-CELL LYMPHOMA OF MALT TYPE ["Sy:Diagnosis"] (WHO 2001 Classification) (4) No ["Li:Lexical Polarity Negative"] significant ["Li:Mood and Comment Adjuncts"] abnormality ["Sy:Diagnosis"]	(1) FOLLICULAR LYMPHOMA, WHO GRADE 1 (2) Diffuse large B cell lymphoma, follicular lymphoma (3) EXTRANODAL MARGINAL ZONE B-CELL LYMPHOMA OF mucosa-associated lymphoid tissue TYPE, LOW GRADE

				(4) no abnormality
		SNOMED RT Codes		
		Stage		
		Comment		

Note: Sample text without specification suggests the whole text is an entity of the associated medical entity type.

Appendix IV Screenshots from the Structured Reporting Web Page

Melanoma Report

Your Melanoma Report:

Report Analysis

Note: Only instances used to populate the template are presented below.

SURGICAL PATHOLOGY
CLINICAL NOTES
See 972284O. Wider excision *acral lentiginous malignant melanoma*
right plantar foot .

MACROSCOPIC EXAMINATION
"Excision base of right foot". A roughly oval piece of skin 42 x 32 x 6 mm *. Centrally there is an irregular mottled brown lesion 20 x 12 mm. Directly adjacent to this is a linear change 13 mm in length. Towards one end there is a tentative surgical incision 20 mm in length. (BP/BJB)*

MICROSCOPIC EXAMINATION
Sections are of glabrous skin from the foot, which includes the underlying skeletal muscle and subcutaneous tissue. The skin centrally shows scarring and chronic inflammation in the dermis and the subcutaneous tissue from the previous surgical procedure *. The skin adjoining and overlying this focus of scarring shows an* acral lentiginous melanoma *,* tumour thickness 1.4mm *,* Clark level III *.* Single *malignant melanocytes are noted in the adventitial tissue in occasional sweat ducts. Some of these sweat ducts with the single atypical melanocyte in the adventitia are located in the superficial subcutaneous tissue and are 2.15 mm deep. The* closest deep surgical margin is 2.2 mm clear *. The melanocytes in the dermic show no* mitoses *. There is no* angiolymphatic invasion *identified. There are no* tumour infiltrating lymphocytes *noted. There are no* satellite deposits *present. There is no* ulceration *and no* perineural invasion *identified.*

The radial growth phase *of the acral lentiginous melanoma* extends to one lateral surgical margin *in block 10 in one section.*

COMMENT
The slides have been reviewed by Dr X and he agrees with the above findings.

DIAGNOSIS
Skin right foot (plantar aspect) (wider excision of melanoma)
acral lentiginous melanoma, tumour thickness 1.4 mm, Clark level III, excision incomplete at one lateral surgical margin - please consult above report and comments.

List of Tags

- Site and laterality
- Specimen type
- Prev. Rx / Trauma
- Previous melanoma
- Distant metastasis
- Size of specimen
- Other lesions
- Ulceration(mm diam)
- Mitotic rate
- Tumour thickness
- Level of invasion (Clark)
- Lymphovascular invasion
- Neurotropism
- Microsatellites
- Assoc. benign naevus
- Excision margins: In-situ
- Excision margins: Invasive
- Excision margins: Deep
- Desmoplasia
- Diagnosis
- Subtype
- TILs
- TILs: Distribution/Density
- Excision margins: Clear
- Regression

View 1. Annotations of a single specimen document

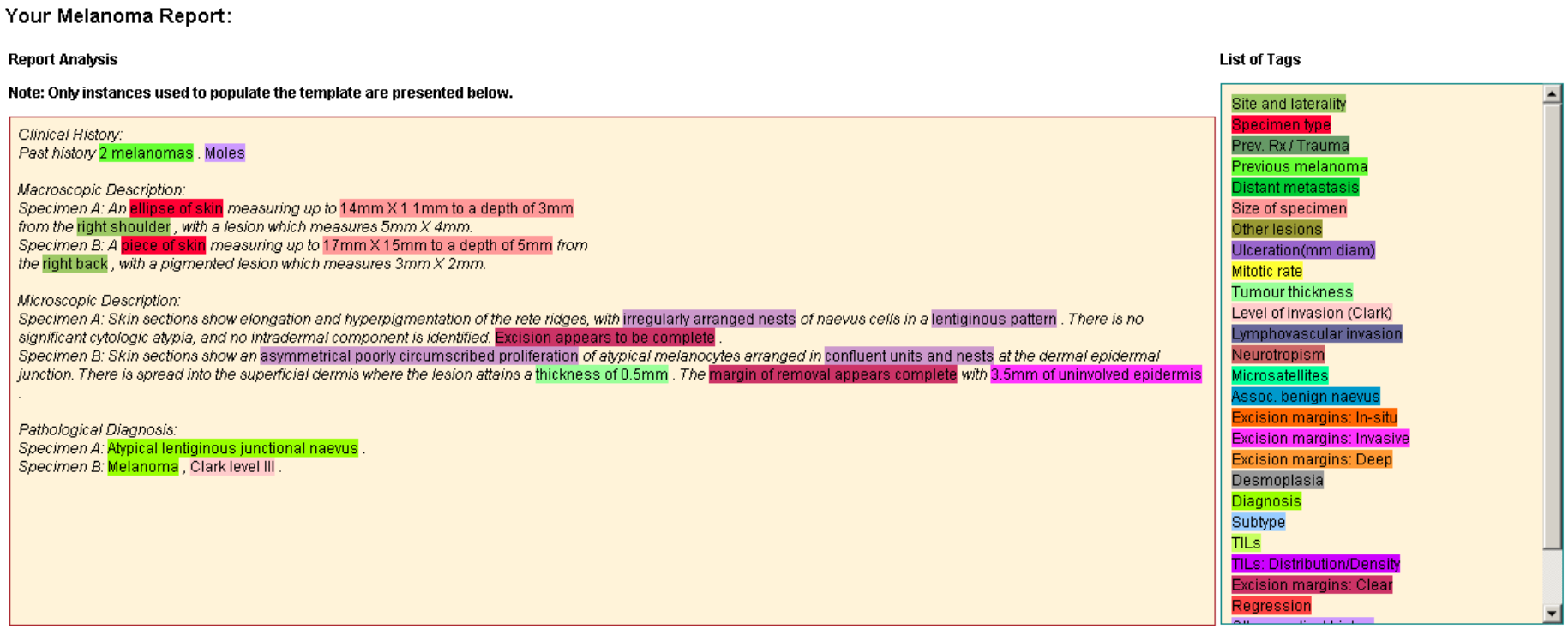
Structured Output Report

Section	Title	Description
Diagnostic Summary	Summary	Skin right foot (plantar aspect) (wider excision of melanoma) acral lentiginous melanoma, tumour thickness 1.4 mm, Clark level III, excision incomplete at one lateral surgical
	Comment	The slides have been reviewed by Dr X and he agrees with the above findings.
Supporting Information	Clinical Description	See 972284O. Wider excision acral lentiginous malignant melanoma right plantar foot.
	Macroscopic Description	"Excision base of right foot". A roughly oval piece of skin 42 x 32 x 6 mm. Centrally there is an irregular mottled brown lesion 20 x 12 mm. Directly adjacent to this is a linear change 13 mm in length. Towards one end there is a tentative surgical incision 20 mm in length. (BP/BJB)
	Microscopic Description	Sections are of glabrous skin from the foot, which includes the underlying skeletal muscle and subcutaneous tissue. The skin centrally shows scarring and chronic inflammation in the dermis and the subcutaneous tissue from the previous surgical procedure. The skin adjoining and overlying this focus of scarring shows an acral

Items in Supporting Information

Sub-section	Specimen id	Item	Value
CLINICAL	1	Site and laterality	right plantar foot
	1	Specimen type	wider excision
	1	Prev. Rx / Trauma	surgical procedure(previous)
	1	Clinical diagnosis	acral lentiginous malignant melanoma
	1	Previous melanoma	N/A
	1	Distant metastasis	N/A
	1	Other medical history	None relevant
MACROSCOPIC	1	Other lesions	N/A
	1	Size of specimen	42mm x 32mm x 6mm
MICROSCOPIC	1	Ulceration(mm diam)	absent
	1	Mitotic rate	N/A
	1	Tumour thickness	1.4mm
	1	Level of invasion (Clark)	III
	1	Lymphovascular invasion	absent
	1	Neurotropism	absent
	1	Microsatellites	absent
	1	Assoc. benign naevus	N/A
	1	Excision margins: In-situ	N/A
	1	Excision margins: Invasive	N/A
	1	Excision margins: Deep	clear - 2.2mm
	1	Desmoplasia	N/A
	1	Diagnosis	melanoma
	1	Subtype	acral lentiginous
	1	Cell growth	single, radial growth phase
	1	TILs	absent
	1	TILs: Distribution	N/A
	1	TILs: Density	N/A
	1	Regression	N/A

View 2. Structured reporting on a single specimen document



View 3. Annotations of a multiple specimen document

Structured Output Report

Section	Title	Description
Diagnostic Summary	Summary	Specimen A: Atypical lentiginous junctional naevus. Specimen B: Melanoma, Clark level III.
	Comment	N/A
Supporting Information	Clinical Description	Past history 2 melanomas. Moles
	Macroscopic Description	Specimen A: An ellipse of skin measuring up to 14mm X 1 1mm to a depth of 3mm from the right shoulder, with a lesion which measures 5mm X 4mm.
	Microscopic Description	Specimen A: Skin sections show elongation and hyperpigmentation of the rete ridges, with irregularly arranged nests of naevus cells in a lentiginous pattern. There is no significant cytologic atypia, and no intradermal component is identified. Excision appears to be complete.

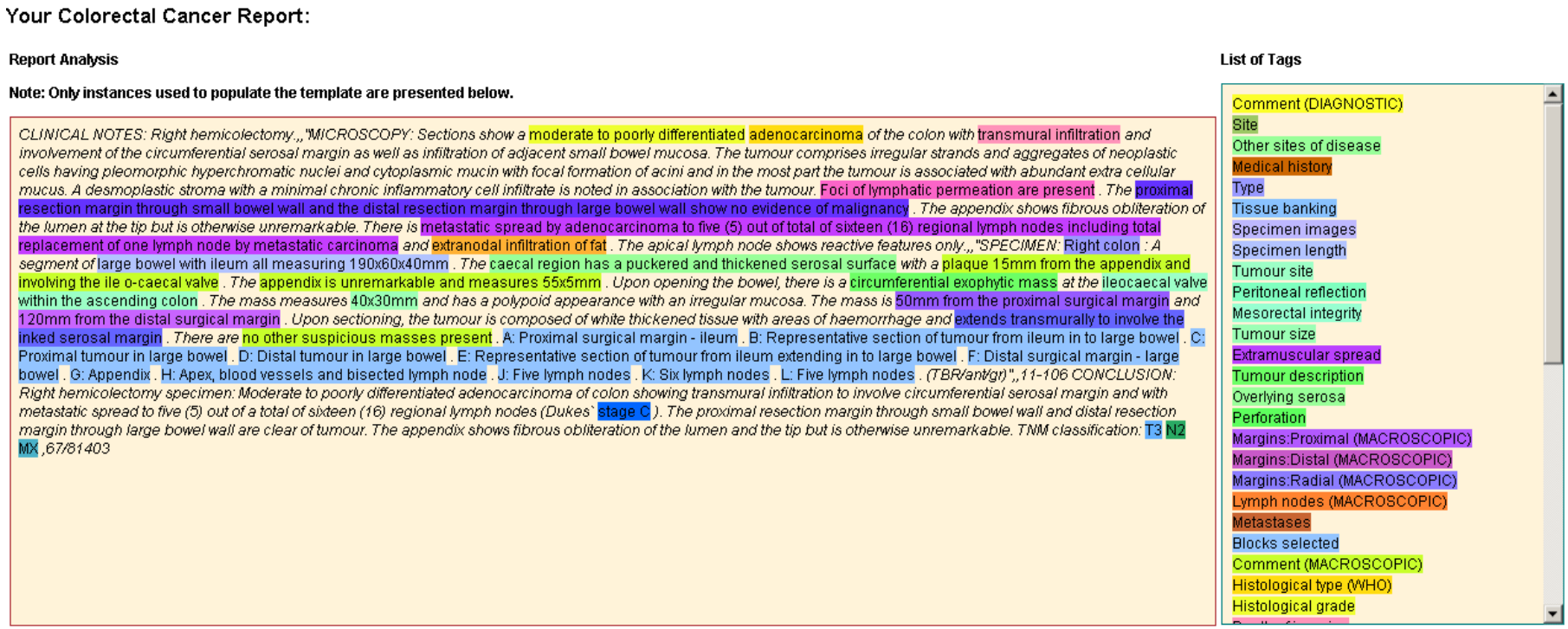
View 4. Structured reporting on a multiple specimen document – Part 1

Items in Supporting Information

Sub-section	Specimen id	Item	Value
CLINICAL	1	Site and laterality	right shoulder
	1	Specimen type	ellipse of skin
	1	Prev. Rx / Trauma	N/A
	1	Clinical diagnosis	N/A
	1	Previous melanoma	present
	1	Distant metastasis	N/A
	1	Other medical history	moles
	2	Site and laterality	right back
	2	Specimen type	piece of skin
	2	Prev. Rx / Trauma	N/A
	2	Clinical diagnosis	N/A
	2	Previous melanoma	present
	2	Distant metastasis	N/A
	2	Other medical history	moles
MACROSCOPIC	1	Other lesions	N/A
	1	Size of specimen	14mm x 11mm x 3mm
	2	Other lesions	N/A
	2	Size of specimen	17mm x 15mm x 5mm
MICROSCOPIC	1	Ulceration(mm diam)	N/A
	1	Mitotic rate	N/A
	1	Tumour thickness	N/A
	1	Level of invasion (Clark)	N/A
	1	Lymphovascular invasion	N/A
	1	Neurotropism	N/A
	1	Microsatellites	N/A
	1	Assoc. benign naevus	N/A
	1	Excision margins: In-situ	N/A
	1	Excision margins: Invasive	N/A
	1	Excision margins: Deep	N/A
	1	Desmoplasia	N/A
	1	Diagnosis	atypical lentiginous junctional naevus
	1	Subtype	N/A
	1	Cell growth	irregularly arranged nests, lentiginous pattern
	1	TILs	N/A
	1	TILs: Distribution	N/A
	1	TILs: Density	N/A
	1	Regression	N/A
	2	Ulceration(mm diam)	N/A
	2	Mitotic rate	N/A
	2	Tumour thickness	0.5mm
	2	Level of invasion (Clark)	III
	2	Lymphovascular invasion	N/A
	2	Neurotropism	N/A
	2	Microsatellites	N/A
	2	Assoc. benign naevus	N/A
	2	Excision margins: In-situ	N/A
	2	Excision margins: Invasive	clear - 3.5mm
	2	Excision margins: Deep	N/A
	2	Desmoplasia	N/A
	2	Diagnosis	melanoma
	2	Subtype	N/A
	2	Cell growth	asymmetrical poorly circumscribed proliferation, confluent units and nests
	2	TILs	N/A
	2	TILs: Distribution	N/A
	2	TILs: Density	N/A
	2	Regression	N/A

View 5. Structured reporting on a multiple specimen document – Part 2

Colorectal Cancer Report



View 6. Annotations of a single specimen document

Structured Output Report

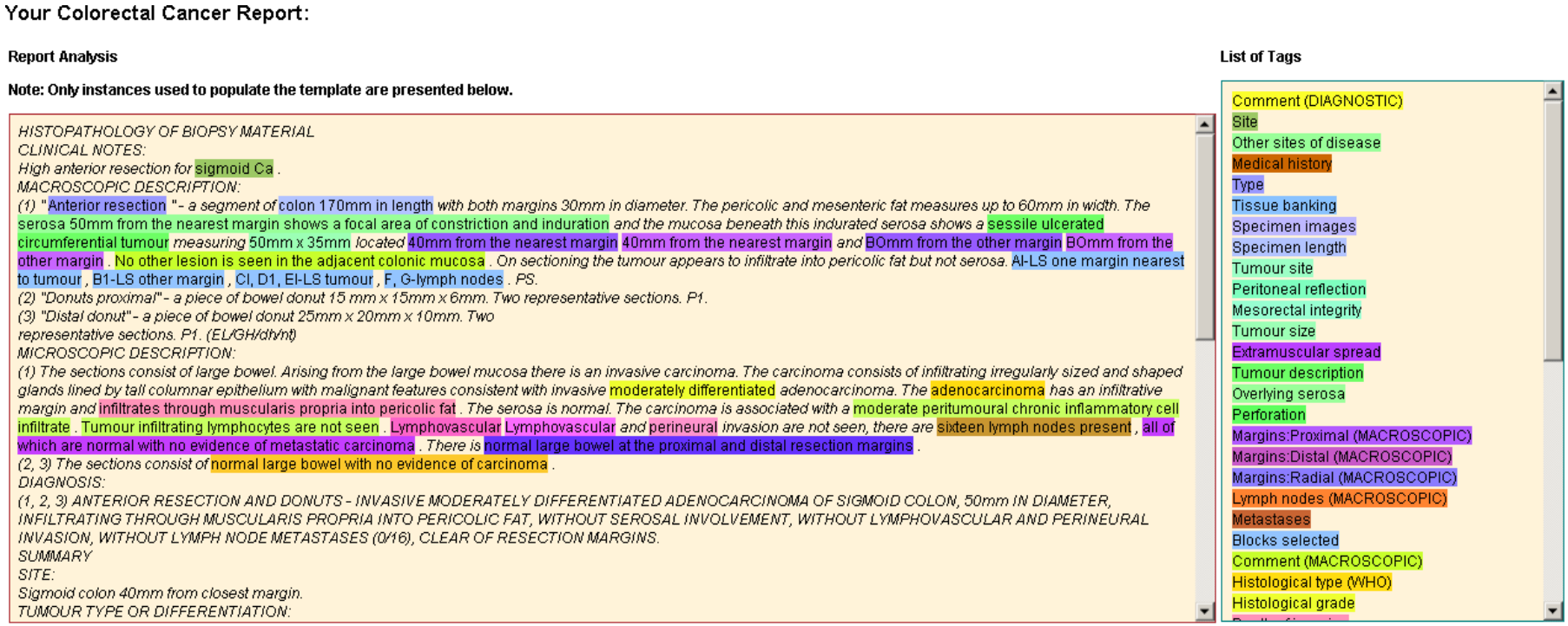
Section	Title	Description
Diagnostic Summary	Summary	Right hemicolectomy specimen: Moderate to poorly differentiated adenocarcinoma of colon showing transmural infiltration to involve circumferential serosal margin and with metastatic spread to five (5) out of a total of sixteen (16) regional lymph nodes (Dukes` stage C). The proximal resection margin through small bowel wall and distal
Supporting Information	Clinical Description	Right hemicolectomy.
	Macroscopic Description	Right colon: A segment of large bowel with ileum all measuring 190x60x40mm. The caecal region has a puckered and thickened serosal surface with a plaque 15mm from the appendix and involving the ile o-caecal valve. The appendix is unremarkable and measures 55x5mm. Upon opening the bowel, there is a circumferential
	Microscopic Description	Sections show a moderate to poorly differentiated adenocarcinoma of the colon with transmural infiltration and involvement of the circumferential serosal margin as well as infiltration of adjacent small bowel mucosa. The tumour comprises irregular strands and aggregates of neoplastic cells having pleomorphic hyperchromatic
	Ancillary Description	N/A
	Synthesis Description	N/A

View 7. Structured reporting on a single specimen document – Part 1

Items and Values

Sub-section	Item	Value
DIAGNOSTIC	Comment	N/A
CLINICAL	Site	N/A
	Other sites of disease	N/A
	Medical history	N/A
MACROSCOPIC	Type	right colon
	Tissue banking	N/A
	Specimen images	N/A
	Specimen length	190mm
	Tumour site	ileocaecal valve within the ascending colon
	Peritoneal reflection	N/A
	Mesorectal integrity	N/A
	Tumour size	40mm
	Extramuscular spread	N/A
	Tumour description	circumferential exophytic
	Overlying serosa	caecal region has a puckered and thickened serosal surface
	Perforation	N/A
	Margins:Proximal	50mm
	Margins:Distal	120mm
	Margins:Radial	N/A
	Lymph nodes	N/A
	Metastases	N/A
	Blocks selected	A: Proximal surgical margin - ileum; B: Representative section of tumour from ileum in to large bowel; C: Proximal tumour in large bowel; D: Distal tumour in large bowel; E: Representative section of tumour from ileum extending in to large bowel; F: Distal surgical margin - large bowel; G: Appendix; H: Apex, blood vessels and bisected lymph node; J: Five lymph nodes; K: Six lymph nodes; L: Five lymph nodes
	Comment	appendix is unremarkable and measures 55x5mm; no other suspicious masses present. plaque 15mm from the appendix and involving the ile o-caecal valve.
MICROSCOPIC	Histological type (WHO)	adenocarcinoma
	Histological grade	moderate to poorly differentiated
	Depth of invasion	transmural infiltration
	Serosal Involvement	N/A
	Venous invasion	N/A
	Small vessel invasion	present
	Perineural invasion	N/A
	TILs	N/A
	Margins:Proximal	clear
	Margins:Distal	clear
	Margins:Radial	N/A
	Lymph nodes	16
	Number involved	5
	Distant spread	present
	Response to Rx	N/A
	Comment	N/A
ANCILLARY	Ancillary Studies	N/A
SYNTHESIS	TNM stage:T	T3
	TNM stage:N	N2
	TNM stage:M	MX
	Stage Group	stage C
	Residual tumour (R)	N/A
	Comment	N/A

View 8. Structured reporting on a single specimen document – Part 2



View 9. Annotations of a multiple specimen document

Structured Output Report

Section	Title	Description
Diagnostic Summary	Summary	(1, 2, 3) ANTERIOR RESECTION AND DONUTS - INVASIVE MODERATELY DIFFERENTIATED ADENOCARCINOMA OF SIGMOID COLON, 50mm IN DIAMETER, INFILTRATING THROUGH MUSCULARIS PROPRIA INTO PERICOLIC FAT, WITHOUT SEROSAL INVOLVEMENT, WITHOUT LYMPHOVASCULAR AND
Supporting Information	Clinical Description	High anterior resection for sigmoid Ca.
	Macroscopic Description	(1) "Anterior resection" - a segment of colon 170mm in length with both margins 30mm in diameter. The pericolic and mesenteric fat measures up to 60mm in width. The serosa 50mm from the nearest margin shows a focal area of constriction and induration and the mucosa beneath this indurated serosa shows a sessile ulcerated
	Microscopic Description	(1) The sections consist of large bowel. Arising from the large bowel mucosa there is an invasive carcinoma. The carcinoma consists of infiltrating irregularly sized and shaped glands lined by tall columnar epithelium with malignant features consistent with invasive moderately differentiated adenocarcinoma. The adenocarcinoma has
	Ancillary Description	Immunostains for identification of microsatellite instability (MSI) have been performed on the tumour as follows: MLH1 - normal mucosa positive, tumour positive. MSH2 - normal mucosa positive, tumour positive.
	Synthesis Description	SITE: Sigmoid colon 40mm from closest margin. TUMOUR TYPE OR DIFFERENTIATION:

View

View 10. Structured reporting on a multiple specimen document – Part 1

Items and Values

Sub-section	Item	Value
DIAGNOSTIC	Comment	N/A
CLINICAL	Site	sigmoid
	Other sites of disease	N/A
	Medical history	N/A
MACROSCOPIC	Type	anterior resection
	Tissue banking	N/A
	Specimen images	N/A
	Specimen length	170mm
	Tumour site	N/A
	Peritoneal reflection	N/A
	Mesorectal integrity	N/A
	Tumour size	50mm
	Extramuscular spread	N/A
	Tumour description	sessile ulcerated circumferential
	Overlying serosa	serosa 50mm from the nearest margin shows a focal area of constriction and induration
	Perforation	N/A
	Margins:Proximal	40mm/80mm
	Margins:Distal	40mm/80mm
	Margins:Radial	N/A
	Lymph nodes	N/A
	Metastases	N/A
	Blocks selected	AI-LS one margin nearest to tumour; B1-LS other margin; CI, D1, EI-LS tumour; F, G-lymph nodes
	Comment	No other lesion is seen in the adjacent colonic mucosa.
MICROSCOPIC	Histological type (WHO)	adenocarcinoma
	Histological grade	moderately differentiated
	Depth of invasion	infiltrates through muscularis propria into pericolic fat
	Serosal Involvement	N/A
	Venous invasion	absent
	Small vessel invasion	absent
	Perineural invasion	absent
	TILs	TILs: absent. Peritumoural: present(moderate).
	Margins:Proximal	clear
	Margins:Distal	clear
	Margins:Radial	N/A
	Lymph nodes	16
	Number involved	0
	Distant spread	N/A
	Response to Rx	N/A
	Comment	normal large bowel with no evidence of carcinoma.
ANCILLARY	Ancillary Studies	Immunostains for identification of microsatellite instability (MSI) have been performed on the tumour as follows: MLH1 - normal mucosa positive, tumour positive. MSH2 - normal mucosa positive, tumour positive. MSH6 - normal mucosa positive, tumour positive. PMS2 - normal mucosa positive, tumour positive. immunostaining profile is not typical of (but does not exclude) HNPCC.
SYNTHESIS	TNM stage:T	T3
	TNM stage:N	N0
	TNM stage:M	MX
	Stage Group	ACPS - B
	Residual tumour (R)	N/A
	Comment	N/A

View 11. Structured reporting on a multiple specimen document – Part 2

Lymphoma Report

View 12. Annotations of a single specimen document

Structured Output Report

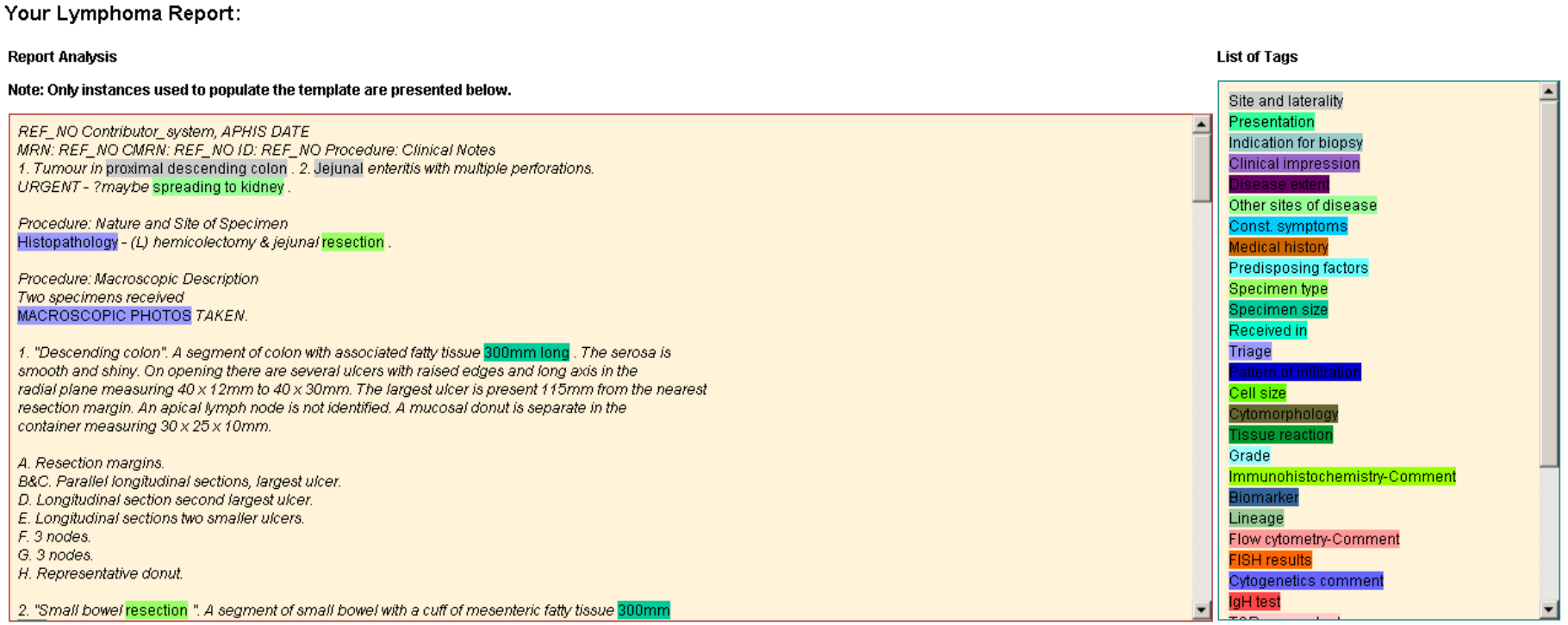
Section	Title	Description
Diagnostic Summary	Summary	Left paraaortic lymph node, core bx: MANTLE CELL LYMPHOMA. SNOMED CODES: 1 M-96733 Mantle cell lymphoma
	Comment	N/A
Supporting Information	Clinical Description	M70. B-cell lymphoma diagnosed in 2002 Noumea. Treated with 6 courses of chemo in 2002, 2003, 2004 and 2005. Blinded in (L) eye 2005 (?haemorrhage). Hemicolectomy 18.7.05 for lymphoma involvement. PET (DATE) showed mild-to-moderate glucose-avidity in the paraaortic region near L3. CT (DATE) showed enlarged paraaortic LNs to 3cm dia. Specimen: 2x 18G core biopsies (under CT guidance).
Supporting Information	Specimen Description	(DR_NAME) "CORE BX". The specimen consists of three cores of pale tissue ranging from 3 to 5mm in length. All for section. Levels x2, CD3, CD5, CD10, CD23, CD79a, bcl-2, cyclin D1, Ki67, p53.
Supporting Information	Microscopic Description	The core biopsies show a monotonous population of small lymphocytes in an apparently diffuse growth pattern with infiltration into adipose tissue. The cells have small hyperchromatic rounded nuclei. Mitoses are scant and no blastic morphology is identified. Immunohistochemical stains show the lymphoid cells have the following staining characteristics: Positive - CD20, CD79a, cyclin D1, bcl-2.

View 13. Structured reporting on a single specimen document – Part 1

Items in Supporting Information

Sub-section	Specimen id	Item	Value
CLINICAL	1	Site and laterality	paraaortic lymph nodes
	1	Presentation	N/A
	1	Indication for biopsy	N/A
	1	Clinical impression	N/A
	1	Disease extent	N/A
	1	Other sites of disease	N/A
	1	Const. symptoms	N/A
	1	Medical history	b-cell lymphoma
	1	Predisposing factors	6 courses of chemotherapy
SPECIMEN	1	Specimen type	core biopsy
	1	Specimen size	3-5mm
	1	Received in	N/A
	1	Triage	N/A
MICROSCOPIC	1	Pattern of infiltration	diffuse
	1	Cell size	small
	1	Cytomorphology	no blastic
	1	Tissue reactions	occasional aggregates of follicular dendritic cells
	1	Grade	N/A
IMMUNOPHENOTYPING	1	Immunohistochemistry-Positive for	CD20, CD79a, cyclin D1, bcl-2
	1	Immunohistochemistry-Negative for	CD23, CD10, CD3
	1	Immunohistochemistry-Equivocal for	CD5
	1	Immunohistochemistry-Comment	stains scattered T-cells positively and shows weak blush in B-cells. Ki-67 stains approximately 20-30% of nuclei. Stains for kappa and lambda are difficult to interpret but there appears to be more cells with cytoplasmic staining for lambda.
	1	Flow cytometry-Positive for	N/A
	1	Flow cytometry-Negative for	N/A
	1	Flow cytometry-Comment	N/A
CYTOGENETICS	1	FISH	N/A
	1	Cytogenetics comment	N/A
MOLECULAR	1	IgH	N/A
	1	TCRgamma	N/A
	1	PCR comment	N/A
SYNTHESIS	1	Lineage	N/A
	1	Clonality	N/A
	1	Diagnosis (WHO)	MANTLE CELL LYMPHOMA
	1	Stage	N/A
	1	Comment	N/A
	1	SNOMED RT codes	M-96733 Mantle cell lymphoma P1-03120 Core biopsy T-C4480 Paraaortic lymph node

View 14. Structured reporting on a single specimen document – Part 2



View 15. Annotations of a multiple specimen document

Structured Output Report

Section	Title	Description
Diagnostic Summary	Summary	Left paraaortic lymph node, core bx: MANTLE CELL LYMPHOMA. SNOMED CODES: 1 M-96733 Mantle cell lymphoma
	Comment	N/A
Supporting Information	Clinical Description	M70. B-cell lymphoma diagnosed in 2002 Noumea. Treated with 6 courses of chemo in 2002, 2003, 2004 and 2005. Blinded in (L) eye 2005 (?haemorrhage). Hemicolectomy 18.7.05 for lymphoma involvement. PET (DATE) showed mild-to-moderate glucose-avidity in the paraaortic region near L3. CT (DATE) showed enlarged paraaortic LNs to 3cm dia. Specimen: 2x 18G core biopsies (under CT guidance).
Supporting Information	Specimen Description	(DR_NAME) "CORE BX". The specimen consists of three cores of pale tissue ranging from 3 to 5mm in length. All for section. Levels x2, CD3, CD5, CD10, CD23, CD79a, bcl-2, cyclin D1, Ki67, p53.
Supporting Information	Microscopic Description	The core biopsies show a monotonous population of small lymphocytes in an apparently diffuse growth pattern with infiltration into adipose tissue. The cells have small hyperchromatic rounded nuclei. Mitoses are scant and no blastic morphology is identified. Immunohistochemical stains show the lymphoid cells have the following staining characteristics: Positive - CD20, CD79a, cyclin D1, bcl-2.

View 16. Structured reporting on a multiple specimen document – Part 1

Items in Supporting Information

Sub-section	Specimen id	Item	Value
CLINICAL	1	Site and laterality	proximal descending colon
	1	Presentation	N/A
	1	Indication for biopsy	N/A
	1	Clinical impression	N/A
	1	Disease extent	N/A
	1	Other sites of disease	N/A
	1	Const. symptoms	N/A
	1	Medical history	N/A
	1	Predisposing factors	N/A
	2	Site and laterality	jejunal
	2	Presentation	N/A
	2	Indication for biopsy	N/A
	2	Clinical impression	N/A
	2	Disease extent	N/A
	2	Other sites of disease	possible
	2	Const. symptoms	N/A
	2	Medical history	N/A
	2	Predisposing factors	N/A
SPECIMEN	1	Specimen type	resection
	1	Specimen size	300mm
	1	Received in	N/A
	1	Triage	histopathology, macroscopic photos
	2	Specimen type	resection
	2	Specimen size	300mm
	2	Received in	N/A
	2	Triage	histopathology, macroscopic photos
MICROSCOPIC	1, 2	Pattern of infiltration	not follicular
	1, 2	Cell size	large
	1, 2	Cytomorphology	N/A
	1, 2	Tissue reactions	necrosis, plasma cells
	1, 2	Grade	N/A
IMMUNOPHENOTYPING	1, 2	Immunohistochemistry-Positive for	N/A
	1, 2	Immunohistochemistry-Negative for	CD5, CD10, bcl-6, cyclin D1, lambda
	1, 2	Immunohistochemistry-Equivocal for	N/A
	1, 2	Immunohistochemistry-Comment	express CD79a, CD20, CD30, CD23 and bcl-2. Approximately 50% of nuclei stain with the proliferative marker Ki-67. scattered interspersed small T cells that express CD3. Kappa appears to stain the cytoplasm of most of the cells. suggestive of light chain restriction. strong cytoplasmic staining of approximately 50% of the atypical lymphoid cells.
	1, 2	Flow cytometry-Positive for	N/A
	1, 2	Flow cytometry-Negative for	N/A
	1, 2	Flow cytometry-Comment	N/A
CYTOGENETICS	1	FISH	N/A
	1	Cytogenetics comment	N/A
	2	FISH	N/A
	2	Cytogenetics comment	N/A
MOLECULAR	1	IgH	A MONOCLONAL band, of 102 bp, was detected.
	1	TCRgamma	N/A
	1	PCR comment	This result was confirmed by repeat DNA extraction and PCR. PCR amplification of DNA with primers flanking the region of gene rearrangement. DNA analysis by high resolution gel electrophoresis with ethidium bromide staining. Band sizing shows a +/-4 bp between run variability. IgH gene rearrangement studies were performed after the method of Brisco et al (1990) Br J Haem 75: 163-167 using the LJH and FR3A primer set.
	2	IgH	Irregular bands detected. No definite evidence of a monoclonal IgH rearrangement detected.
	2	TCRgamma	N/A
	2	PCR comment	PCR amplification of DNA with primers flanking the region of gene rearrangement. DNA analysis by high resolution gel electrophoresis with ethidium bromide staining. Band sizing shows a +/-4 bp between run variability. Ig H gene rearrangement studies were performed after the method of Brisco et al (1990) Br J Haem 75: 163-167 using the LJH and FR3A primer set.
SYNTHESIS	1, 2	Lineage	B-cell
	1, 2	Clonality	N/A
	1, 2	Diagnosis (WHO)	MULTIFOCAL DIFFUSE LARGE B CELL LYMPHOMA, POST-TRANSPLANT LYMPHOPROLIFERATIVE DISORDER; perforation
	1, 2	Stage	N/A
	1, 2	Comment	please see above report.
	1, 2	SNOMED RT codes	M-15500 Transplanted organ M-39210 Perforation M-95903 Malignant lymphoma P1-03000 Excision T-58000 Small intestine T-59460 Descending colon

View 17. Structured reporting on a multiple specimen document – Part 2

Appendix V Examples of Poorly-written Reports

In most cases, the reports were written in a similar way with comparable headings for each section. However, the individual writing styles or language preferences of the pathologists did play an important role in determining how easy or difficult the automatic structured reporting was to perform on a narrative report.

Each of the following examples is representative of the poorly-written reports drawn from the test sets, with reasons following it.

Example 1 - Report #1 in the Melanoma Test Set

Specimen:
RIGHT INNER ANKLE

Macroscopic: VT/FHS
The specimen is a skin ellipse 15 x 7 mm. In the centre there is an irregular black lesion 6x7 mm and close to the nearest line of resection. (Four pieces, one block).

Microscopic:
Sections show an area of SUPERFICIAL SPREADING MALIGNANT MELANOMA with NODULAR MELANOMA invading the dermis to a depth of 0.6 mm. There is marked regressive change at the base of the lesion and the line of excision is clear laterally by less than 0.5 mm in one area; the deep line is clear by 5 mm. In the area where the lesion is very close to the line of resection this is superficial spreading malignant melanoma and it is a skip lesion separated by relatively normal epidermis from the main lesion.

Discussion:

There are several issues in this report:

1. There is no “CLINICAL HISTORY” section, which suggests that some important information such as the patient’s past history and a posited diagnosis are not available in the report.
2. It lacks of diagnosis summary, so the final diagnosis is not immediately visible.
3. The report uses over-complicated clause structure, hindering clear information expression and making it difficult to follow.
4. A number of items of vital information are omitted in the report, e.g., the presence or absence of lymphovascular invasion and neurotropism.

On the whole, this report is short and limited.

Example 2 - Report #2 in the Melanoma Test Set

HISTOPATHOLOGY

MACROSCOPIC:
SPECIMEN CONSISTS OF ONE ELLIPSE OF SKIN AND FAT MEASURING 22X13X5MM

AND BEARING A CENTRAL, DARK PAPULE MEASURING 5MM. EXCISION ? MARGINS INKED. LESION ALL FOR SECTION IN TWO PIECES. (BFQ)

MICROSCOPIC:
SECTIONS OF THE "LEFT ELBOW" LESION SHOW AN ULCERATED MALIGNANT MELANOMA, CLARK LEVEL IV, BRESLOW THICKNESS 1.65MM WITH A MINIMAL ADJACENT COMPONENT OF SUPERFICIAL SPREADING PATTERN. THE CLOSEST PERIPHERAL MARGIN FROM THE IN-SITU
MELANOMA MEASURES 1.35MM. THE CELLS ARE DEVOID OF PIGMENT AND PREDOMINANTLY
EPITHELIOID IN TYPE. SCATTERED MITOSES ARE READILY SEEN. THERE IS NO VASCULAR/LYMPHATIC INVASION PRESENT. NO NEUROTROPISM IS SEEN. FEATURES OF EARLY REGRESSION ARE NOTED AT THE EDGES OF THE LESION.

CONCLUSION:
SKIN "LEFT ELBOW"
ULCERATED, MALIGNANT MELANOMA, CLARK LEVEL IV, BRESLOW THICKNESS 1.65MM WITH AN ADJACENT COMPONENT OF SUPERFICIAL SPREADING PATTERN AND EARLY REGRESSION.
THE CLOSEST PERIPHERAL MARGIN MEASURES 1.35MM.
IMMUNOHISTOCHEMISTRY IS PROCEEDING TO CONFIRM THE DEPTH OF THE LESION.

Discussion:

Several problems are presented with this report:

1. The arbitrary use of All Caps font in the whole report increases the difficulty for processing.
2. It has the same issue discussed in the previous report, without a “CLINICAL HISTORY” section.
3. Some vital information is omitted, such as the distribution and density of tumour-infiltrating lymphocytes, any associated benign melanocytic lesion.

This report provides more details and uses more precise language than the previous one, but it is still hard for processing.

Example 3 - Report #3 in the Melanoma Test Set

SURGICAL PATHOLOGY

CLINICAL DETAILS
1. ? NMM left knee.
2. Irritated naevus right neck.

NATURE OF SPECIMEN
I) Skin biopsy.
II) Skin biopsy.

MACROSCOPIC:
Specimen I: Labelled "L knee", the specimen consists of an ellipse of skin, 15x9x7mm, bearing on its surface a brown seborrhoeic dome shaped lesion 7x6mm, Block 1A - two transverse sections; lB-lC - ends.

Specimen II: Labelled "Right neck", the specimen consists of an ellipse of skin, 14x5x3mm, bearing on its surface a keratotic grey papule 3mm in diameter. Block 2A - two transverse sections; 2B-2C - ends.

30/3 JG: nm 30/03/01

MICROSCOPIC:
Specimen I: Sections show nodular malignant melanoma. The lesion is composed of large pleomorphic spindled cells arranged in a somewhat "Spitzoid" pattern. Clark level is IV (dermal-subcutaneous interface); Breslow thickness is 2.5mm. Although, the lesion gives a low power impression of 'symmetrical' growth, high power examination of the centre of base of the lesion shows small nests and single atypical melanocytes invading lower reticular dermis and infiltrating sweat glands at the dermal-subcutaneous junction. This is confirmed on immunoperoxidase stains for Melan A and S100. **The lesion is positive in its superficial aspect for HMB45, There is a moderate amount of pagetoid invasion of the epidermis.** There is no junctional component beyond the dermal component. There is no surface ulceration. There is a mild patchy lymphocyte response. **Very occasional mitoses are present (<l/mm sg.) however one of these is an atypical mitosis.** The lesion is clear of resection margins by a minimum of 1.2mm (lateral).

CONCLUSION: Nodular malignant melanoma; Level IV, 2.5mm thick.

Specimen II: Sections show excoriated intradermal naevus. There is no evidence of malignancy. Lines of resection are clear of the lesion.

Discussion:

The main issues in this report are:

1. The representation of specimen identifiers (ids) is not consistent in the whole report: in "CLINICAL HISTORY" section, an id is denoted as an Arabic numeral followed with period "."; in "SPECIMEN" section, an id is rendered as a Roman numeral tailed with bracket ")"; in "MACROSCOPIC" and "MICROSCOPIC" sections, an id is started with the lexicon "Specimen". Such inconsistent representation of ids is complicating specimen id detection, which results in incorrect specimen context detection.
2. Delimitation of several sentences is incorrect (highlighted in bold), which hinders the sentence boundary detection on them.
3. The diagnosis summary is misplaced by being placed before the end of the "MICROSCOPIC" section, which can affect the section context detection.
4. The improper unit for mitotic rate ("/mm sg.") prevents extraction of the correct value.

In brief, this is a poorly-written multiple specimen report with abnormal grammatical structures.

Example 4 - Report #1 in the Colorectal Cancer Test Set

CLINICAL NOTES: Anterior resection for rectosigmoid cancer - another one fond in the sigmoid. 1: Distal and proximal rings. 2: Sigmoid-rectum bowel (two primary cancers).
MACROSCOPIC DESCRIPTION: (A) Sigmoid colon: A segment of large bowel measuring 210mm in length and 59mm in internal circumference at proximal resection margin, and 67mm in internal circumference at distal resection margin. Approximately 50mm from the proximal resection margin there is a fungating ulcerating pale tan tumour measuring 24 x 13mm in area and approximately 7mm above surrounding mucosa. This tumour invades the surrounding mesenteric fat and reaches within 1mm of the serosal surface without penetrating through. The serosal surface at this area is puckered and slightly roughened. 100mm distal to this tumour there is a second larger fungating ulcerated tumour measuring 30mm in diameter and up to 10mm in thickness. This larger tumour also involves the surrounding pericolic fat and reaches within 1mm of serosal surface which is slightly roughened and darker brown. Elsewhere there are two pale tan small polypoid lesions; the more distal one approximately 5mm from distal resection margin. The polyps measure up to 3mm

and 4mm and are 10mm apart. Nineteen lymph nodes identified in the surrounding mesenteric fat, the largest measuring up to 7mm. Specimen inking: serosal surface inked with silver nitrate, proximal margin inked blue, distal margin inked green. *** 10/03. Representative sections. A1-A13. A1: proximal resection margin. A2-A4: smaller more proximal tumour. A5: section from bowel mucosa between the two tumours. A6-A7: the more distal larger tumour. A8-A9: composite blocks showing the larger tumour and its distance to distal resection margin (yellow ink on adjoining edges). A10: two small polyps near distal resection margin. A11-A13: lymph nodes. Tissue Bank - A small piece of larger tumour submitted in TB1. (B) Proximal sphincter donut: Received on a spike is a bowel donut measuring 13mm in length and 20mm in external diameter. Macroscopically unremarkable. *** 10/03. No blocks submitted. (C) Distal sphincter donut: A bowel donut measuring 14mm in length and 22mm in external diameter. Macroscopically unremarkable. *** 10/03. No blocks submitted. SS/SR PREVIOUS BIOPSY/CYTOLOGY: Nil.
MICROSCOPIC DESCRIPTION: (A) The tumour 50mm from the proximal resection margin is a moderately well differentiated adenocarcinoma with puckering of the bowel wall associated with transmural infiltration of the muscularis by tumour glands with short extension into the perimuscular lamina. No extramural vascular invasion is identified. The tumour 100mm distal to this is larger with extensive ulceration, is less differentiated, shows transmural spread into the pericolic fat where extramural vascular invasion is identified. Within the submucosa, vascular invasion is also prominent. The tumour is clear of the serosal surface. The two discrete polyps separate from the tumour are metaplastic polyps with regular serrated glands gaping towards the surface and lined by hypermature mucinous epithelium. A total of twenty three lymph nodes are identified, two are largely replaced by metastatic carcinoma and one node shows two small nests of tumour within the peripheral sinus one 0.4mm, the other 0.13mm. The tumours are clear of the mucosal and radial resection margin.
ANTERIOR RESECTION OF RECTOSIGMOID: SYNCHRONOUS ADENOCARCINOMA OF THE COLON. PROXIMAL TUMOUR; MODERATELY WELL DIFFERENTIATED, TRANSMURAL SPREAD, NO EXTRAMURAL VASCULAR INVASION. pT3 DISTAL TUMOUR; POORLY DIFFERENTIATED ADENOCARCINOMA, TRANSMURAL SPREAD, EXTRAMURAL VASCULAR INVASION. pT3 LYMPH NODE METASTASES (2/20) + ISOLATED TUMOUR CELLS pN1b BOTH CLEAR OF MUCOSAL AND RADIAL RESECTION MARGINS.

Discussion:

This report is problematic, as

1. The inconsistent use of specimen ids in different sections. The specimen id "(A)" in "MACROSCOPIC" section refers to id "2" in "CLINICAL HISTORY" section. The id "1" in "CLINICAL HISTORY" section is divided into ids "(B)" and "(C)" in "MACROSCOPIC" section.
2. Besides lack of several mandatory information items (e.g., the presence or absence of tumour perforation and the microscopic residual tumour status), the tumour site and specimen type is not mentioned explicitly in the "MACROSCOPIC" section.
3. The T and N stage values are scattered in the diagnosis summary, which makes their extraction more difficult.
4. In the "MICROSCOPIC" section, when describing the differentiation of the tumour, "poorly differentiated" is preferred to use rather than "less differentiated".

Generally, this report is relatively well-organized, but with imprecise language usage.

Example 5 - Report #2 in the Colorectal Cancer Test Set

Dysplastic polyp. Transverse colon - Right hemicolectomy.
One specimen container received labelled `GLASSENBURY`. The name and biopsy number on the cassettes supplied match those on the specimen request and the specimen containers. The contents are labelled `Right hemi colon`. The specimen consists of a length of large bowel 240mm long with terminal ileum measuring 25mm in length and appendix measuring 60mm in length. There is mesenteric fat measuring up to 60mm in width attached to the specimen as well as omentum measuring 180x110x up to 45mm. The serosal surface of the bowel is unremarkable. On opening there is a sessile polyp measuring 65x40mm. The polyp is located 40mm from the distal resection margin. There is a 2mm raised area of mucosa located 60mm from the ileocaecal junction. There is a diverticulum located 115mm from the ileocaecal junction. No other focal abnormalities are identified within the mucosa. Blocks: 1a - distal margin, 1b - terminal ileal margin, 1c-1u - the polyp all blocked, 1v - diverticulum, 1w - appendix, 1x - the 2mm polyp, 1y - apical lymph nodes x 2, 1z - five mesenteric lymph nodes, 1aa - five mesenteric lymph nodes, 1ab - five mesenteric lymph nodes, 1ac - five lymph nodes, 1ad - three lymph nodes. Tissue remains.
Sections confirm a large (65x 40mm) tubulo-villous adenoma with moderate cytologic atypia. In blocks 1g and 1i there is an invasive adenocarcinoma with tumour islands seen within the submucosa but not entering the muscularis propria. There is no evidence of lymphovascular space permeation although there is some retraction artefact noted in block 1i. A small amount of black dye is present within the adenoma but separate to the tumour. There is no evidence of nodal metastasis in 22 nodes. A separate tubular adenoma with focal low grade dysplasia is also present 60mm from the ileo-caecal junction. The resection margins are free of adenomatous and dysplastic change.
CACOLON Procedure Right hemicolectomy 240mm large bowel, 25 small. Tumour type : Adenocarcinoma arising in atubulovillous adenoma Tumour grade: Well differentiated (AJCC) Location : Transverse colon Size : 3mm Longitudinal and 2mm transverse Depth of invasion: T1=Into submucosa but not not muscularis. Resection margins: Clear Mesenteric deposits and other organs: Nil Perforation: Absent Lymphovascular invasion: Absent Perineural invasion: Absent Tumour Border: Infiltrative Lymph nodes: Apical node not specifically identified. 22 lymph nodes sampled. 0 lymph nodes show tumour Capsular involvement not seen. NO = No regional nodes Polyps : Tumour arose in a tubulovillous adenoma. Non-tumorous bowel: separate tubular adenoma, diverticulum present and black dye in lamina propria consistent with previous biopsy site. Staging: **Stage 1 = T1 or T2 N0 M0** Stage IIa = T3 N0 M0 Stage IIb = T4 N0 M0 Stage IIIa = T1-2, N1, M0 Stage IIIb = T3-4, N2, M0 Stage IIIc = Any T, N2, M0 Stage IV = Any T, Any N, M1 p = Pathologist, x = Dont know

Discussion:

Several issues are apparent in this report:

1. Lack of section headings, which is the main obstacle to section context detection.
2. The expression of staging information is inappropriate: only content regarding the diagnosis should be recorded (highlighted in bold) and other references of staging should be excluded.
3. In the last paragraph, “25 small” is incomplete, which should be modified to “25mm small bowel”.

Briefly, most of the sentences are in regular structures, though each major aspect if kept to one clause or sentence would have been beneficial.

Example 6 - Report #3 in the Colorectal Cancer Test Set

CLINICAL HISTORY: Ultra low anterior resection for rectal carcinoma. Recto-sigmoid resection and proximal and distal donuts.
SPECIMEN: Two pots received. 1. Rectal tumour: A segment of large bowel comprising approximately 125mm length of sigmoid and 95mm length of rectum. The segment of true mesentery is 130mm length x 80mm width, which falls away to perirectal fat. Arising from the antimesenteric side of the mucosa, at 14mm from distal surgical margin, there is a fungating, ulcerated, bleeding tumour (56x45mm), which obscures approximately 70% of the lumen. The tumour invades muscularis propria and extends into perirectal fat; the tumour is 9mm from the radial soft tissue margin. At 30mm proximal to the peritoneal reflection, the serosa is focally drawn into the tumour, but does not appear to breach the surface. The tumour is well clear of the proximal resection margin (greater than 150mm). 1A: Tumour with deepest invasion (radial soft tissue margin green). 1B and 1C: Sections of tumour with overlying puckered serosa. 1D: Distal surgical resection margin. 1E: Proximal surgical resection margin. 1F: Further representative section of tumour. 1G: Tumour and adjacent proximal mucosa. Throughout sigmoid colon there are approximately one dozen diverticula, one of which is ulcerated at the base. There are no other masses or polyps along the bowel wall. 1H and 1J: Two diverticula. A total of 27 lymph nodes are found within the mesentery, up to 16mm diameter. None of the nodes grossly appear to contain tumour. No vessels grossly contain tumour. 1K: Vascular mesenteric resection margin. 1L: Apical lymph node. 1M: Six lymph nodes. 1N: Six lymph nodes. 1P: Six lymph nodes. 1Q: Six lymph nodes. 1R: Two lymph nodes. 2. Proximal and distal donuts: Two specimens received in the pot. The first is an intact donut (20x16x12mm). The mucosa is unremarkable and there are no masses. Radial margin is inked black. 2A and 2B: Intact donut trisected. The non-intact segment is a crescent-shaped piece of intestine (20x25x11mm). There is an area of ulcerated mucosa (15x4mm), at one end. The specimen cannot be oriented, all margins are inked green. 2C to 2F: Eleven transverse slices processed. All processed. (LJ/tb/sf)
SPECIMEN: The specimen is a recto-sigmoid resection along with proximal and distal donuts. LARGE BOWEL IN GENERAL: There is hypertrophy of main muscle coat and then indication of formation of diverticulae. None of the diverticulae sectioned is inflamed and so the patient has diverticulosis with no evidence of diverticulitis in the sections. TUMOUR: The mucosa gives way to a moderately and rather poorly differentiated adenocarcinoma which is raised at its edges and ulcerated in its central portion. The appearances of the tumour are quite consistent with a primary arising in large bowel. Some foci of tumour have considerable necrosis and there is also spotty calcification. TUMOUR SPREAD: 1. Direct spread: The proximal and distal margins of excision will be considered under the heading of donuts. The tumour itself is through main muscle coat and is out into pericolic fat. 2. Vascular and perivascular spread: There are foci which indicate that there is possible lymphovascular spread. 3. Neural and perineural infiltration: This has not been identified. 4. Lymphatic spread: Twenty-five (25) genuine lymph nodes have been found and none contain metastatic tumour (0 out of 25). 1. In summary, there is an elevated and ulcerated moderately and poorly differentiated adenocarcinoma of the large bowel associated with necrosis and calcification. The tumour is through the wall and through the serosa and out into surrounding fat. There are no lymph node metastases. 2. There is evidence of diverticular disease in this material with hypertophy of main muscle coat and diverticulae forming. However no malignancy has been identified. 3. There are rather haemorrhagic fragments of large bowel including mucosa and main muscle coat. There is no evidence of malignancy in this material. The absence of malignancy in both the proximal and distal donuts indicates that the proximal and distal margins of this specimen are free of tumour. Immunoperoxidase studies will be carried out on lymph nodes to determine if micrometastases are present or not and a supplementary report will be issued when available.
MICROSCOPY:
11-2393 CONCLUSION: Recto-sigmoid resection in which there is moderately and poorly differentiated adenocarcinoma which is through to the pericolic fat. There are no lymph node metastases (Dukes` B). SYNOPTIC REPORT FOR LARGE BOWEL MALIGNANCY SUPPLEMENTARY Immunoperoxidase stains were carried out looking for micrometastases in lymph nodes and none were found.

Discussion:

The major issues in this report include:

1. Misuse of "SPECIMEN" as a microscopic examination heading (highlighted in bold), can lead to inaccurate section context detection.
2. No content is reported under "SYNOPTIC REPORT" heading.
3. TNM stages are not reported in "CONCLUSION" section.

Generally, this report is also in poor organization, without sufficient information.

Example 7 - Report #4 in the Colorectal Cancer Test Set

CLINICAL HISTORY: Low rectal cancer. Rectosigmoid colon and distal donut.
SPECIMEN: 1. Low anterior resection: Received in formalin is a segment of large bowel (170mm length x 45mm diameter), that is half sigmoid and half rectum. The attached mesentery is 90mm length x 125mm width, which folds away to peri-rectal fat. At 9mm from the distal (false) surgical margin there is an exophytic fungating tumour (25x24mm) and an adjacent ulcer that extends proximally for a total dimension of 65mm length x 26mm diameter. The exophytic area of the tumour appears to invade only muscularis propria but where the ulcer extends proximally there is invasion of muscularis propria into surrounding peri-rectal fat, to a depth of 4mm. The tumour is heterogeneous white/tan with ulceration and haemorrhage throughout. The tumour is well clear of the deep fatty resection margin (28mm). The rest of the bowel mucosa is unremarkable without any polyps or other masses. 1A: Proximal surgical margin. 1B: Distal (false) surgical margin. 1C to 1F. Sections of ulcer/mass with invasion into surrounding fat. 1G to 1J: Sections of exophytic portion of tumour. Fifty-two (52) possible lymph nodes are found, up to 14mm diameter. Many of the nodes appear grossly involved by firm white tumour and several are haemorrhagic. 1K to 1M: Three high tie vascular mesenteric resection margins. 1N: Two high tie lymph nodes. 1P: Two lymph nodes grossly involved. 1Q: Four lymph nodes. 1R: Six lymph nodes. 1S: Four lymph nodes. 1T: Four lymph nodes. 1U: Five lymph nodes. 1V: Five lymph nodes. 1W: Five lymph nodes. 1X: Five lymph nodes. 1Y: Five lymph nodes. 1Z: Five lymph nodes. 2. Distal donut: In formalin a short segment of bowel (8mm length x 20mm diameter) that is unremarkable. All processed as five transverse slices. Two blocks. (LJ/sas/gr)
MICROSCOPY:
11-1923 CONCLUSION: 1 and 2. Recto-sigmoid colon in which ulcerated adenocarcinoma of the large bowel has been identified. There are numerous lymph node metastases.
SYNOPTIC REPORT FOR LARGE BOWEL MALIGNANCY SPECIMEN. 1. The specimen is a recto-sigmoid colon with an associated distal donut. TUMOUR: The tumour is a moderately and sometimes poorly differentiated adenocarcinoma consistent with a primary arising in large bowel. The tumour tends to be exophytic in its pattern of growth on the margins and endophytic in the centre. The centre also tends to be ulcerated and a little scarred with acutely inflamed slough on the surface. The bowel outside the tumour appears normal. EXTENT OF SPREAD: The proximal and distal margins of excision are free of tumour. The tumour is through main muscle coat and is well out into surrounding fat. LYMPHOVASCULAR INVASION: Lymphovascular invasion has been identified. NEURAL AND PERINEURAL INFILTRATION: No definite neural or perineural infiltration has been identified. LYMPH NODES: Numerous lymph nodes were found and many contain metastatic disease. Fourteen (14) lymph nodes contain metastatic carcinoma associated with considerable necrosis. The number of lymph nodes found in total is forty-two (42). In summary, there is an ulcerated moderately and poorly differentiated adenocarcinoma arising in the large bowel which is through main muscle coat and into surrounding fat. There are numerous lymph node metastases (Dukes` C). 2. Sections taken from the donuts are free of tumour.

Discussion:

There are several issues in this report:

1. Misuse of "SYNOPTIC REPORT" as a microscopic examination heading (highlighted in bold), hindering the detection of section context.
2. It also lacks staging information in the "CONCLUSION" section.
3. It omits to record other vital information as well, such as tumour site in "MACROSCOPIC" section and the status of the nonperitonealised circumferential margin in the "MICROSCOPIC" section.

Likewise, this is also a poorly-organised report, with limited information.

Example 8 - Report #5 in the Colorectal Cancer Test Set

69 yo male LGIE found caecal cancer CT Abdo -> ? paracolic LN Laparoscopic R hemicolectomy One specimen container received labelled `WOOD`. The name and biopsy number on the cassettes supplied match those on the specimen request and the specimen containers. The contents are labelled `Right hemicolectomy`. The specimen consists of 100mm of the ceacum/right colon, 40mm of terminal ileum and appendix 65mm in length. There is mesenteric fat up to 110x90mm attached to the specimen. The serosal surface is deeply indented over an area measuring 15x10mm adjacent to the base of the appendix. On opening the mucosal surface underlying, this indentation corresponds to an ulcerated fungating tumour 50mm in diameter. Tumour is located at least 60mm from the distal resection margin and is adjacent to the ileocaecal junction. On sectioning through the tumour, it extends into pericolic fat. On examination of the rest of the mucosa no polyps or other focal abnormalities are identified. Blocks selected 1a - distal margin, 1b - terminal ileal margin, 1c-e - deepest part of tumour, 1f&g - tumour and appendix, 1h&i - tumour and normal mucosa, 1j - appendix, 1k - 3 apical lymph nodes, 1l- 1 mesenteric lymph node bisected, 1m - 4 mesenteric lymph nodes, 1o - 4 mesenteric lymph nodes, 1p - 3 mesenteric lymph nodes. Tissue remains. Procedure: Right hemicolectomy. Tumour type: Mucin secreting adenocarcinoma. Tumour grade: Moderately differentiated. Location: See macroscopic description. Size: 50mm. Depth of invasion: Into pericolic fat. Resection margins: Clear of tumour. Mesenteric deposits: Nil. Perforation: Nil. Lymphovascular invasion: Nil seen. Perineural invasion: Nil seen. Border: Infiltrative. Lymph nodes: 22 lymph nodes including apical lymph nodes - all clear of tumour. Section of additional lymph node to follow. Supplementary report to follow. Non-tumourous bowel: Diverticular disease.

The additional lymph node is clear of tumour.

Discussion:

The major problems in this report include:

1. Similar to Example 5, it has no section headings.
2. It seems that the contents for the microscopic examination are missing.
3. A subheading referring to "Supplementary Report" is omitted from the last paragraph.
4. No staging information is provided.
5. Frequent use of abbreviations or acronyms, such as "yo" and "R", increases the difficulty for medical entity recognition.

In summary, this report is less-organised, short and limited.

Example 9 - Report #1 in the Lymphoma Test Set

REF_NO Contributor_system, APHIS DATE
MRN: REF_NO CMRN: REF_NO ID: REF_NO

Procedure: Clinical Notes
(R) inguinal LN mass max. diam. 4cm - Splenic lesions also on CT

Procedure: Nature and Site of Specimen
18G core x 3 from (R) groin LN mass - 1. in saline - 2. in formalin - by DR_NAME.

Procedure: Macroscopic Description
Two specimens received.

1. "R lymph node core". Two pale tan cores of tissue 9 x 10mm in length received in formalin.
Almost all embedded in one block. A small portion retained in formalin in case ultrastructural studies required.

2. "(R) lymph core". A pale tan core 14mm in length received in saline. Most sent for Flow Cytometry and a small portion frozen for molecular studies if required. No blocks taken. (wc/pjt)

Procedure: Microscopic Report
1. "R lymph node core".
The core biopsies show effacement of any normal lymph node architecture. There is abundant stromal fibrosis and a heterogenous infiltrate of lymphoid cells. There are numerous atypical large lymphoid cells, some with bizarre nuclear morphology. Some of these cells have prominent nucleoli and there are some binucleate forms with the appearance of Reed-Sternberg cells. In the background there are small lymphoid cells as well as occasional plasma cells, macrophages and eosinophils.

The features favour Hodgkins lymphoma. (wac/swm)

DR_NAME comments:

There is a lymphoid infiltrate in a fibrous background. Scattered large cells are present which often have double or multiple blurred nuclei.

Immunoperoxidase stains show:

CD30 repeated +/- only occasional large cells faint cytoplasmic staining
CD20 large cells negative ; CD3 small cells only ++
CD15 some large cells positive
CD83 large cells strongly positive
Fascin -large cells strongly positive

The appearances are very suggestive of Hodgkin's disease but not entirely diagnostic.

2. "(R) lymph core".
Flow cytometry showed "NO RESULTS AS INSUFFICIENT CD45 POSITIVE VIABLE CELLS PRESENT".

Procedure: Summary
Lymph node, R groin - Hodgkin's disease
SNOMED CODES:
1 M-96500 965-966 HODGKIN'S DISEASE
T-C4000 Lymph node
T-D7000 Inguinal region, NOS

```
Procedure:  History Upload Request Detail
            Req Dr: DR_NAME - REF_NO
            Client: HOSP_NAME   MRN: REF_NO
```

Discussion:

There are two main issues in this report:

1. Specimen ids are used contradictorily in different sections. The specimen id "1" in "SPECIMEN" section corresponds to id "2" in the "MACROSCOPIC" section, while id "2" in "SPECIMEN" section is referred to id "2" in "MACROSCOPIC" section.
2. The ambiguous use of "R lymph node" or "(R) lymph" in the "MACROSCOPIC" and "MICROSCOPIC" sections, obstructing clear information about the site of the biopsy.

Except for these problems, this is a relatively well-organized report.

Example 10 - Report #2 in the Lymphoma Test Set

```
Procedure:  Clinical Notes
            (R) femoral pathological fracture ?lymphoma. Large lytic lesion mid femur
              Flow cytometry : CLL
      Procedure:  Macroscopic Description
            Three specimens received.

            1.  "Right femur bone".  A piece of bony tissue 40 x 15 x 5mm.  The periosteum
            appears ragged, irregular and shows some eburnation.  The inner surface shows some
            congestion.  The specimen kept in decalcification.

            2.  "Curettings from right thigh lesion in formalin".  Soft congested partly pale tan
            tissue                  including blood clot measuring 30 x 30mm in aggregate.  All
            embedded in one block.

            3.  "Fracture haematoma from right femur".  A friable part of semisolid altered blood
            clot 30 x 30mm.  Some embedded in one block.
            (rw/kms)

      Procedure:  Microscopic Report
            1,2,3.  Apart from evidence of recent fracture and repair (extensive haemorrhage,
            granulation tissue and newbone all three specimens show poorly defined sheets of
            small and medium sized lymphoid cells.

            Positive  :  CD20 +++, CD23 +
                      :  CD3 (++ scattered small lymphocytes)
            Negative :  CD10, CD56, CD5, MPO, CD138

            Bone Tumour Meeting DATE
            DR_NAME agreed that the features were indicative of a diffuse small cell malignant
            lymphoma.
            DR_NAME and DR_NAME could not remember a similar case with localised bone
            involvement.

      Procedure:  Summary
            L. femur - diffuse, small cell malignant lymphoma, CLL type.
          SNOMED CODES:
          1  M-96703 Malignant lymphoma, small lymphocytic
            T-12710 Femur
```

Procedure: History Upload Request Detail
Req Dr: DR_NAME - REF_NO
Client: HOSP_NO MRN: REF_NO

Procedure: Supplementary Report
Addendum

DR_NAME comments:
I agree that such extensive involvement of the bone marrow of one bone would be unusual in CLL, especially without peripheral blood involvement. There are some cells with largish vesicular nuclei and prominent nucleoli. These could be paraimmunoblasts of CLL but they might be large cells on the outskirts of a diffuse large B-cell malignant lymphoma.

Procedure: Supplementary Summary
L. femur - diffuse malignant lymphoma **?** large B cell **?** CLL type.

Procedure: History Upload Request Detail
Req Dr: DR_NAME - REF_NO
Client: HOSP_NAME MRN: REF_NO

Discussion:

Major issues in this report include:

1. The laterality of biopsy in the diagnosis summary is inconsistent with those indicated in other sections ("left" *vs.* "right"), making it difficult to follow the report precisely.
2. No specimen id is provided in "SUPPLEMENTARY REPORT" section, increasing the difficulty for specimen context detection.
3. The hedging expressions for the diagnosis are not appropriate in the diagnosis summary (highlighted in bold).

On the whole, this report is difficult to follow with contrary language usage.

Example 11 - Report #3 in the Lymphoma Test Set

REF_NO
MRN: REF_NO
Procedure: Referred MRN
HOSP_NAME MR REF_NO

Procedure: Clinical Notes
3 lesions in (R) lobe liver fund in liver incidentally on CT scan.

Procedure: Nature and Site of Specimen
Right hemihepatectomy / cholecystectomy.

Procedure: Macroscopic Description
Two specimens were received.

1. "Liver in formalin". A portion of the right lobe of liver that appears to include segments VI and VII as well as V and VIII. The falciform ligament is not present on the specimen. It measures 210 x 160 x 110mm and weighs 1280g. The capsular surface is smooth and shows no focal abnormalities. Serial slicing reveals two ill-

defined pale grey lesions. The first measures up to 20mm in diameter and appears to be the lower portion of segment VIII. It is more than 30mm from the surgical margin. The second lesion is up to 15mm across and abuts the capsule in the lower portion of what appears to be segment VII. It is greater than 70mm from the surgical margin. On the anterior surface of segment V there is a small inconspicous pale grey lesion causing a very slight puckering of the overlying capsule. It measures up to 5mm across (lesion 3). In addition, in the lateral portion of segment VII, there is an ill-defined haemorrhagic blush in the subcapsular location up to 40mm across (?surgical artifact). No other focal abnormalities are identified. The uninvolved liver parenchyma appears normal and is not cirrhotic.
A&B. Lesion 1.
C&D. Lesion 2.
E&F. Lesion 3.
G. Ill-defined haemorrhagic appearing area.
H. Representative normal appearing liver.

2. "Gallbladder". A gallbladder 90mm in length and up to 75mm in open circumference. The serosal surface is unremarkable. The wall thickness is up to 2mm. The mucosal surface is velvety and green and shows no focal abnormalities. No stones are present with the specimen.

A. Cystic duct at surgical margin and neck of galbladder.
B. Body and fundus.
(wac/mcm)

Procedure: Microscopic Report

1. "Liver in formalin".
All three of the liver lesions have a similar appearance and are characterised by a localised atypical lymphoid infiltrate causing marked expansion and confluence of portal tracts. The infiltrate is composed predominantly of small lymphoid cells with a very thin rim of cytoplasm, together with occasional centroblast-like cells and scattered plasma cells. Occasional small groups of lymphoid cells appear to infiltrate into bile duct epithelium suggestive of lymphoepithelial lesions. In addition, there are a number of small follicle centres throughout the infiltrate.
Immunohistochemical stains show the majority of cells are CD20, CD79a positive B cells with focal staining for CD5. The cells also show expression of CD43.
The tumour cells appear to be negative for CD10, CD23 and cyclin D1.
There are numerous CD3 positive T cells mostly at the periphery of the lymphoid infiltrate. Very ocasional CD138 positive plasma cells are present.
CD21 highlights residual small follicle centres.

The uninvolved liver has a normal architecture and shows moderate panlobular macro and microvesicular steatosis. By contrast, the liver lobules between the involved portal tracts lack significant amounts of steatosis. No Mallory's hyaline is identified. The portal tracts and lobules are otherwise unremarkable and lack significant inflammation or fibrosis.

The features are of a B cell non-Hodgkins lymphoma. The morphological and immunohistochemical findings suggest either a mantle cell lymphoma or an extranodal marginal zone B cell lymphoma.
Further immunoperoxidase stains are in progress in an attempt to further classify the tumour. An addendum report will be issued.
The resection appears are well clear of the tumours (at least 30mm clearance).

2. "Gallbladder"
Sections of gallbladder showing a few foci of perivascular lymphoid infiltration in the subserosa.
There is no mucosal inflammation. No other significant histological abnormalities are identified.

Procedure: Summary
Liver:
- Low to intermediate grade B-cell lymphoma.
- Steatosis.
- Further immunoperoxidase stains pending. An addendum report will be issued.
/
Gallbladder - No significant abnormality.
SNOMED CODES:
1 M-50080 Fatty degeneration
M-95903 Malignant lymphoma
P3-44202IPX
T-62000 Liver
T-E0000 Cell, NOS
2 M-00100 Normal tissue
T-63000 Gall bladder

Procedure: Pathologist Notes - Not for Publication
Also seen by DR_NAME and DR_NAME, who agree with the above.

Procedure: History Upload Request Detail
Req Dr: DR_NAME - REF_NO
Client: HOSP_NAME MRN: REF_NO

Procedure: Supplementary Report
Further immunoperoxidase stains performed revealed the following characteristics:
Positive: CD20, CD79a, CD43, CD5
Equivocal: cyclin D1 (Repeated also by HOSP_NAME.)
Negative: CD10, CD23
CD138 positive plasma cells are present.
CD21 highlights residual small follicle centres.

The slides were also reviewed by DR_NAME and DR_NAME. The consensus view is that the tumour should be regarded as a Mantle cell lymphoma (W.H.O. classification). Common sites of involvement by this lymphoma are lymph nodes, spleen and bone marrow. Extranodal sites include GI tract and Waldeyer's ring. Liver involvement is uncommon but has been previously reported.

Procedure: Supplementary Summary
Liver:
- Mantle cell lymphoma (an intermediate grade B-cell lymphoma).
- Steatosis.
- Further immunoperoxidase stains pending. An addendum report will be issued.
/
Gallbladder - No significant abnormality.

Procedure: History Upload Request Detail
Req Dr: DR_NAME - REF_NO
Client: HOSP_NAME MRN: REF_NO

Discussion:

Despite the report's length, there is no mention of specimen handling or triage and the definite WHO grade (though it is implicitly referred to as "intermediate grade"). It is notable that there is no specimen id used in the diagnosis summary and "SUPPLEMENTARY REPORT" section, which is a disadvantage for specimen context detection. The test results and interpretation for ancillary studies are not clear. For instance, it is difficult to tell from "The cells also show expression of CD43"

whether it is a positive result. All of these features lead to a long-winded report which is not immediately informative.

Example 12 - Report #4 in the Lymphoma Test Set

REF_NO
MRN: REF_NO CMRN: REF_NO ID: REF_NO Procedure: CLINICAL DETAILS
Liver R lobe laterally. Core needles: 3x 18G passes.

Procedure: MACROSCOPIC DESCRIPTION
(DR_NAME)

"LIVER BIOPSY 18G X 3". Five pale tan core biopsies 12, 8, 7, 6 and 4mm in length and small tiny white fragments up to 2mm across. Specimen entirely embedded in blocks A - B.

Procedure: MICROSCOPIC REPORT
The core biopsies show the hepatic mass lesion is malignant lymphoma, diffuse large cell type.
The cells are large, pleomorphic, and show foci of single cell necrosis.
It has a very high Ki67 labelling.
Immunochemically the tumour cells stain for CD45, CD20, CD79a, CD10 and bcl-2.
The tumour is negative for CD30, ALK-1, CD138, MPO, TdT and CK.
Only scanty reactive T-cells are admixed.
The adjacent liver tissue shows moderate macrovesicular steatosis.

Procedure: COMMENT
See CS-08-932 for the complete report.

Procedure: SUMMARY
Liver / core bx - MALIGNANT LYMPHOMA.
Diffuse large cell type, B-cell, lambda.
.
SNOMED CODES:
1 M-95903 Malignant lymphoma, NOS
M-96803 Malignant lymphoma, large cell, diffuse
P1-03120Core biopsy
T-62000 Liver, NOS

Discussion:

This report omits some vital information, such as the fluid delivering the specimen, specimen handling or triage. It also contains imprecise language. For example, “the tumour cells stain for CD45, CD20, CD79a, CD10 and bcl-2” is too ambiguous to report a positive result for ancillary studies. In summary, this is a less informative report with imprecise language.

Example 13 - Report #5 in the Lymphoma Test Set

REF_NO
MRN: REF_NO CMRN: REF_NO ID: REF_NO Procedure: CLINICAL DETAILS
Tru-cut biopsies (L) glenoid - ? Met ?Lymphoma / myeloma. Histopath.

Procedure: MACROSCOPIC DESCRIPTION
(DR_NAME)

"BIOPSY LEFT GLENOID". Five fragments of grey tissue from 2 to 6mm across. All embedded in one block. Levels and spares ordered.

Procedure: MICROSCOPIC REPORT
(DR_NAME/DR_NAME)

"BIOPSY LEFT GLENOID". The core biopsies consist of fibrous and adipose tissue diffusely infiltrated by discohesive malignant cells. Tumour cells are round to oval with irregular nuclear contours and a variably prominent central nucleoli. Large amounts of eosinophilic cytoplasm are present. The mitotic activity is brisk (up to 8 per 10 high power fields) with abnormal forms seen.
Apoptotic debris is scattered in the background. A few admixed osteoclasts are present and a small amount of woven bone is also seen.

Immunohistochemical stains show the tumour cells staining positively with CD20, CD79a and negatively with CD138, CD3 and CD30. Ki 67 stains approximately 40 % of the tumour cells.
The appearances are consistent with a diffuse large B cell lymphoma.

Procedure: SUMMARY
Left glenoid, biopsy - Diffuse large B cell lymphoma.
SNOMED CODES:
1 M-95903 Lymphoma, NOS
T-E0000 THE CELL

Procedure: COMMENT
Bone Tumour Meeting DATE.
Left shoulder pain 4/12.
MRI - large destructive lesion with permeative changes involving the glenoid.
DR_NAME, DR_NAME and DR_NAME agreed with diffuse large B cell lymphoma.
NAME DATE.

Discussion:

This report does not provide sufficient information, e.g., the fluid delivering the specimen; cell size is implicitly indicated in the diagnosis “diffuse large B cell lymphoma”; a vague expression of specimen type: “biopsy” is used in “MACROSCOPIC” section and the diagnosis summary. It is preferable for the clinical history heading to begin on a new line. In brief, this report is subject to insufficient or underreported information.

Summary

Recommendations for writing a melanoma pathology report have been provided in a previous work on the project (Patrick and Scolyer, 2008). It can be seen from the above examples that most of the issues addressed should be taken into consideration when writing pathology reports of other cancer diseases.

1. Appropriate Use of Section Headings

A report should have at least four major section headings to delimitate the contents in the associated sections of clinical history, macroscopic examination, microscopic examination, and diagnosis summary.

2. Proper Use of Specimen Identifiers

In a multiple specimen document, it is critical to use specimen identifiers properly in every section, and the representation of them should remain consistent in the whole document.

3. Simple Grammatical Structure

Pathologists should try to use simple clauses and sentence structures to describe their findings. For example, each major aspect is kept to one clause or sentence (e.g., "No extramural vascular invasion is identified."). Short sentences are preferable, as they are easier to read and comprehend, which clearly delineates where description of a feature starts and ends.

4. Careful Use of "ALL-CAPS" Font

ALL-CAPS font should only be used in the diagnosis summary section, to highlight the key features of the cancer disease.

5. Precise Language

Precise language makes the report easier to follow. Ambiguous expressions of the findings or diagnoses, which can complicate the reading, should be avoided.

Pathologists should carefully consider these suggestions when writing a cancer pathology report. A precise and easy-to-read report will ultimately lead to better automatic structured reporting of it.

Bibliography

Abacha, A.M. and Zweigenbaum, P. (2011). Medical entity recognition: a comparison of semantic and statistical methods. *In Proceedings of the 2011 Workshop on Biomedical Natural Language Processing*, Portland, USA, 56-64, Association for Computational Linguistics.

Alpaydin, E. (2004). *Introduction to machine learning*. The MIT Press, Cambridge, MA.

Aronow, D.B., Feng, F.F. and Croft, W.B. (1999). Ad hoc classification of radiology reports. *J Am Med Inform Assn*. **6**(5): 393-411.

Aronson, A.R. (2001). Effective mapping of biomedical text to the UMLS Metathesaurus: the MetaMap program. *Proc AMIA Symp*: 17-21.

Balch, C.M. (2002). Surgical margins for melanoma: is 2 cm too much? *Anz J Surg*. **72**(4): 251-252.

Beattie, G.C., McAdam, T.K., Elliott, S., Sloan, J.M. and Irwin, S.T. (2003). Improvement in quality of colorectal cancer pathology reporting with a standardized proforma--a comparative study. *Colorectal Dis*. **5**(6): 558-562.

Ben-Gal, I. (2007). Bayesian networks. *In Encyclopedia of Statistics in Quality and Reliability*, John Wiley & Sons.

Bishop, C.M. (2007). *Pattern recognition and machine learning*. Springer, New York, NY.

Bosman, F.T., Carneiro, F., Hruban, R.H. and Theise, N.D. (2010). *WHO classification of tumours of the digestive system*. IARC Press, Lyon, France.

Bundschus, M., Dejori, M., Stetter, M., Tresp, V. and Kriegel, H.P. (2008). Extraction of semantic biomedical relations from text using conditional random fields. *Bmc Bioinformatics*. **9**.

Burges, C.J.C. (1998). A tutorial on support vector machines for pattern recognition. *Data Min Knowl Disc*. **2**(2): 121-167.

Cancer protocols. College of American Pathologists (CAP). Available from http://www.cap.org/. Accessed June 13 2013.

Chan, N.G., Duggal, A., Weir, M.M. and Driman, D.K. (2008). Pathological reporting of colorectal cancer specimens: a retrospective survey in an academic Canadian pathology department. *Can J Surg*. **51**(4): 284-288.

Chang, C.C. and Lin, C.J. (2011). LIBSVM: A library for support vector machines. *Acm T Intel Syst Tec*. **2**(3).

Chapman, B.E., Lee, S., Kang, H.P. and Chapman, W.W. (2011a). Document-level classification of CT pulmonary angiography reports based on an extension of the ConText algorithm. *J Biomed Inform*. **44**(5): 728-737.

Chapman, W.W., Bridewell, W., Hanbury, P., Cooper, G.F. and Buchanan, B.G. (2001a). Evaluation of negation phrases in narrative clinical reports. *J Am Med Inform Assn*: 105-109.

Chapman, W.W., Bridewell, W., Hanbury, P., Cooper, G.F. and Buchanan, B.G. (2001b). A simple algorithm for identifying negated findings and diseases in discharge summaries. *J Biomed Inform*. **34**(5): 301-310.

Chapman, W.W., Chu, D. and Dowling, J.N. (2007a). ConText: an algorithm for identifying contextual features from clinical text. *In Proceedings of the Workshop on Biological, Translational, and Clinical Language Processing (BioNLP 2007)*, Prague, Czech Republic, 81-88, Association for Computational Linguistics.

Chapman, W.W., Chu, D. and Dowling, J.N. 2007b. *ConText: An algorithm for identifying contextual features from clinical text*. Association for Computational Linguistics, Prague, Czech Republic.

Chapman, W.W., Dowling, J.N. and Hripcsak, G. (2008). Evaluation of training with an annotation schema for manual annotation of clinical conditions from emergency department reports. *Int J Med Inform*. **77**(2): 107-113.

Chapman, W.W., Nadkarni, P.M., Hirschman, L., D'Avolio, L.W., Savova, G.K. and Uzuner, O. (2011b). Overcoming barriers to NLP for clinical text: the role of shared tasks and the need for additional creative solutions. *J Am Med Inform Assn*. **18**(5): 540-543.

Chapuis, P.H., Chan, C., Lin, B.P.C., Armstrong, K., Armstrong, B., Spigelman, A.D., O'Connell, D., Leong, D. and Dent, O.F. (2007). Pathology reporting of resected colorectal cancers in New South Wales in 2000. *Anz J Surg*. **77**(11): 963-969.

Chen, C.-H., Ping, X.-O., Wang, Z.-J. and Hsieh, S.-L. (2010). The keyword-based and semantic-driven data matching approach for assisting structuralizing the textual clinical documents. *In the 3rd International Conference on Biomedical Engineering and Informatics (BMEI 2010)*, Yantai, China, 6: 2532-2535, IEEE.

MUC-7 information extraction task definition. NIST. Available from http://www.itl.nist.gov/iaui/894.02/related_projects/muc/proceedings/ie_task.html. Accessed June 13 2013.

Christensen, L., Haug, P.J. and Fiszman, M. (2002). MPLUS: a probabilistic medical language understanding system. *In Proceedings of the 2002 Workshop on Natural Language Processing in the Biomedical Domain*, Philadelphia, USA, 29-36, Association for Computational Linguistics.

Chun, H.-W., Tsuruoka, Y., Kim, J.-D., Shiba, R., Nagata, N., Hishiki, T. and Tsujii, J. (2006). Extraction of gene-disease relations from medline using domain dictionaries and machine learning. *In Proceedings of the Pacific Symposium on Biocomputing (PSB 2006)*, Maui, USA, 4-15, PSB.

Clark, C., Aberdeen, J., Coarr, M., Tresner-Kirsch, D., Wellner, B., Yeh, A. and Hirschman, L. (2011). MITRE system for clinical assertion status classification. *J Am Med Inform Assn*. **18**(5): 563-567.

Coden, A., Savova, G., Sominsky, I., Tanenblatt, M., Masanz, J., Schuler, K., Cooper, J., Guan, W. and de Groen, P.C. (2009). Automatically extracting cancer disease characteristics from pathology reports into a Disease Knowledge Representation Model. *J Biomed Inform*. **42**(5): 937-949.

Cohen, K.B., Fox, L., Ogren, P.V. and Hunter, L. (2005). Corpus design for biomedical natural language processing. *In Proceedings of the 2005 ACL-ISMB Workshop on Linking Biological Literature, Ontologies and Databases*, Detroit, MI, 38-45, Association for Computational Linguistics.

Collins, M. (2002). Ranking algorithms for named-entity extraction: boosting and the voted perceptron. *In Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics*, Philadelphia, USA, 489-496, Association for Computational Linguistics

Cortes, C. and Vapnik, V. (1995). Support-vector networks. *Mach Learn*. **20**(3): 273-297.

Cristianini, N. and Shawe-Taylor, J. (2000). *An introduction to Support Vector Machines and other kernel-based learning methods*. Cambridge University Press, Cambridge, UK.

Cross, S.S., Bull, A.D. and Smith, J.H.F. (1989). Is there any justification for the routine examination of bowel resection margins in colorectal adenocarcinoma. *J Clin Pathol*. **42**(10): 1040-1042.

D'Avolio, L.W., Litwin, M.S., Rogers, S.O., Jr. and Bui, A.A. (2007). Automatic identification and classification of surgical margin status from pathology reports following prostate cancer surgery. *AMIA Annu Symp Proc*: 160-164.

de Bruijn, B., Cherry, C., Kiritchenko, S., Martin, J. and Zhu, X.D. (2011). Machine-learned solutions for three stages of clinical information extraction: the state of the art at i2b2 2010. *J Am Med Inform Assn*. **18**(5): 557-562.

de Marneffe, M.-C., MacCartney, B. and Manning, C.D. (2006). Generating Typed Dependency Parses from Phrase Structure Parses. *In LREC 2006: International Conference on Language Resources and Evaluation*, Genoa, Italy, 449-454, LREC 2006 Committees.

Deleger, L., Grouin, C. and Zweigenbaum, P. (2010). Extracting medical information from narrative patient records: the case of medication-related information. *J Am Med Inform Assn*. **17**(5): 555-558.

Denny, J.C., Miller, R.A., Johnson, K.B. and Spickard, A., 3rd. (2008). Development and evaluation of a clinical note section header terminology. *AMIA Annu Symp Proc*: 156-160.

Denny, J.C., Spickard, A., Johnson, K.B., Peterson, N.B., Peterson, J.F. and Miller, R.A. (2009). Evaluation of a method to identify and categorize section headers in clinical documents. *J Am Med Inform Assn*. **16**(6): 806-815.

Dogan, R.I., Neveol, A. and Lu, Z.Y. (2011). A context-blocks model for identifying clinical relationships in patient records. *Bmc Bioinformatics*. **12**.

Dolin, R.H., Alschuler, L., Beebe, C., Biron, P.V., Boyer, S.L., Essin, D., Kimber, E., Lincoln, T. and Mattison, J.E. (2001). The HL7 clinical document architecture. *J Am Med Inform Assn*. **8**(6): 552-569.

Douthat, A. (1998). The Message Understanding Conference Scoring Software User's Manual. *In In Proceedings of the Seventh Message Understanding Conference (MUC-7)*, Fairfax, Virginia, Appendix G, Morgan Kaufmann.

Dworak, O. (1992). Synoptic surgical pathology reporting. *Hum Pathol*. **23**(1): 85-85.

Dzeroski, S. and Zenko, B. (2004). Is combining classifiers with stacking better than selecting the best one? *Mach Learn*. **54**(3): 255-273.

Eckstein, R., Ackland, S., Brown, I. and Ellis, D. (2010). *Colorectal cancer structured reporting protocol*. Royal College of Pathologists of Australasia. Available from http://www.rcpa.edu.au/Publications/StructuredReporting/CancerProtocols.htm. Accessed 13 July 2012.

Edge, S.E., Byrd, D.R., Carducci, M.A. and Compton, C.A. (2010). *AJCC cancer staging manual*. Springer, New York, NY.

Elkin, P.L., Brown, S.H., Lincoln, M.J., Hogarth, M. and Rector, A. (2003). A formal representation for messages containing compositional expressions. *Int J Med Inform*. **71**(2-3): 89-102.

Fagan, M.J., Griffith, R.A., Obbard, L. and O'Connor, C.J. (2003). Improving the physical diagnosis skills of third-year medical students - A controlled trial of a literature-based curriculum. *J Gen Intern Med*. **18**(8): 652-655.

Fellbaum, C. (1998). *WordNet: an electronic lexical database*. MIT press, Cambridge, MA.

Fiszman, M., Chapman, W.W., Aronsky, D., Evans, R.S. and Haug, P.J. (2000). Automatic detection of acute bacterial pneumonia from chest x-ray reports. *J Am Med Inform Assn*. **7**(6): 593-604.

Friedman, C. (2000). A broad-coverage natural language processing system. *J Am Med Inform Assn*: 270-274.

Friedman, C., Alderson, P.O., Austin, J.H.M., Cimino, J.J. and Johnson, S.B. (1994). A general natural-language text processor for clinical radiology. *J Am Med Inform Assn*. **1**(2): 161-174.

Friedman, C., Shagina, L., Lussier, Y. and Hripcsak, G. (2004). Automated encoding of clinical documents based on natural language processing. *J Am Med Inform Assn*. **11**(5): 392-402.

Fritz, A.G. (2000). *International Classification of Diseases for Oncology: ICD-O*. World Health Organization (WHO), Geneva, Switzerland.

Frunza, O. and Inkpen, D. (2011). Extracting relations between diseases, treatments, and tests from clinical data. *In Advances in Artificial Intelligence*, Springer Berlin Heidelberg.

Fuller, M. and Zobel, J. (1998). Conflation-based comparison of stemming algorithms. *In Proceedings of the Third Australian Document Computing Symposium*, Sydney, Australia, 8-13, University of Sydney.

Fundel, K., Kuffner, R. and Zimmer, R. (2007). RelEx - Relation extraction using dependency parse trees. *Bioinformatics*. **23**(3): 365-371.

Garla, V., Lo Re, V., Dorey-Stein, Z., Kidwai, F., Scotch, M., Womack, J., Justice, A. and Brandt, C. (2011). The Yale cTAKES extensions for document classification: architecture and application. *J Am Med Inform Assn*. **18**(5): 614-620.

Gill, A.J., Johns, A.L., Eckstein, R., Samra, J.S., Kaufman, A., Chang, D.K., Merrett, N.D., Cosman, P.H., Smith, R.C., Biankin, A.V. and Kench, J.G. (2009). Synoptic reporting improves histopathological assessment of pancreatic resection specimens. *Pathology*. **41**(2): 161-167.

Giuliano, C., Lavelli, A. and Romano, L. (2006). Exploiting shallow linguistic information for relation extraction from biomedical literature. *In Proceedings of the 11th Conference of the European Chapter of the Association for Computational Linguistics (EACL 2006)*, Trento, Italy, 401-408, Association for Computational Linguistics.

Goldberger, A.L., Amaral, L.A.N., Glass, L., Hausdorff, J.M., Ivanov, P.C., Mark, R.G., Mietus, J.E., Moody, G.B., Peng, C.K. and Stanley, H.E. (2000). PhysioBank, PhysioToolkit, and PhysioNet - Components of a new research resource for complex physiologic signals. *Circulation*. **101**(23): E215-E220.

Goryachev, S., Sordo, M., Zeng, Q. and Ngo, L. (2006). Implementation and evaluation of four different methods of negation detection. *Technical report*, DSG.

Graham, D.M. and Appelman, H.D. (1990). Crohn's-like lymphoid reaction and colorectal carcinoma: a potential histologic prognosticator. *Mod Pathol*. **3**(3): 332-335.

Griffantibartoli, F., Arnone, G.B., Ceppa, P., Ravera, G., Carrabetta, S. and Civalleri, D. (1994). Malignant-tumors in the head of the pancreas and the periampullary region - diagnostic and prognostic aspects. *Anticancer Res*. **14**(2B): 657-666.

Grishman, R. and Sundheim, B. (1996). Message understanding conference-6: A brief history. *In Proceedings of the 16th International Conference on Computational Linguistics (COLING)*, Copenhagen, DK, 1: 466–471, Association for Computational Linguistics.

Guerry, D.t., Synnestvedt, M., Elder, D.E. and Schultz, D. (1993). Lessons from tumor progression: the invasive radial growth phase of melanoma is common, incapable of metastasis, and indolent. *J Invest Dermatol*. **100**(3): 342S-345S.

Guiasu, S. and Shenitzer, A. (1985). The principle of Maximum-Entropy. *Math Intell*. **7**(1): 42-48.

Hahn, U., Romacker, M. and Schulz, S. (2002). MEDSYNDIKATE - a natural language system for the extraction of medical information from findings reports. *Int J Med Inform*. **67**(1-3): 63-74.

Halgrim, S.R., Xia, F., Solti, I., Cadag, E. and Uzuner, O. (2011). A cascade of classifiers for extracting medication information from discharge summaries. *J Biomed Semantics*. **2 Suppl 3**: S2.

Hammond, E.H. and Henson, D.E. (1995). The role of pathologists in cancer-patient staging. *Am J Clin Pathol*. **103**(6): 679-680.

Harkema, H., Chapman, W.W., Saul, M., Dellon, E.S., Schoen, R.E. and Mehrotra, A. (2011). Developing a natural language processing application for measuring the quality of colonoscopy procedures. *J Am Med Inform Assn*. **18**: I150-I156.

Harkema, H., Dowling, J.N., Thornblade, T. and Chapman, W.W. (2009). ConText: an algorithm for determining negation, experiencer, and temporal status from clinical reports. *J Biomed Inform*. **42**(5): 839-851.

Harvey, J.M., Sterrett, G.F., McEvoy, S., Fritschi, L., Jamrozik, K., Ingram, D., Joseph, D., Dewar, J. and Byrne, M.J. (2005). Pathology reporting of breast cancer: trends in 1989-1999, following the introduction of mammographic screening in Western Australia. *Pathology*. **37**(5): 341-346.

Hastie, T., Tibshirani, R. and Friedman, J.H. (2001). *The elements of statistical learning*. Springer, New York, NY.

Haug, P., Koehler, D., Lau, L.M., Wang, P., Rocha, R. and Huff, S. (1994). A natural-language understanding system combining syntactic and semantic techniques. *J Am Med Inform Assn*: 247-251.

Haydu, L.E., Holt, P.E., Karim, R.Z., Madronio, C.M., Thompson, J.F., Armstrong, B.K. and Scolyer, R.A. (2010). Quality of histopathological reporting on melanoma and influence of use of a synoptic template. *Histopathology*. **56**(6): 768-774.

Hina, S., Atwell, E. and Johnson, O. (2010). Semantic tagging of medical narratives with top level concepts from SNOMED CT healthcare data standard. *International Journal of Intelligent Computing Research*. **1**(3): 204-210.

Hofmann, T., Scholkopf, B. and Smola, A.J. (2008). Kernel methods in machine learning. *Ann Stat*. **36**(3): 1171-1220.

Hsu, C.W., Chang, C.C. and Lin, C.J. (2010). A practical guide to support vector classification. *Technical report*, National Taiwan University.

Huang, Y. and Lowe, H.J. (2007). A novel hybrid approach to automated negation detection in clinical radiology reports. *J Am Med Inform Assn*. **14**(3): 304-311.

Huang, Y., Lowe, H.J. and Hersh, W.R. (2003). A pilot study of contextual UMLS indexing to improve the precision of concept-based representation in XML-structured clinical radiology reports. *J Am Med Inform Assn*. **10**(6): 580-587.

Huang, Y., Lowe, H.J., Klein, D. and Cucina, R.J. (2005). Improved identification of noun phrases in clinical radiology reports using a high-performance statistical natural language parser augmented with the UMLS Specialist Lexicon. *J Am Med Inform Assn*. **12**(3): 275-285.

Assertion annotation guidelines. Informatics for Integrating Biology and the Bedside (i2b2). Available from https://www.i2b2.org/NLP/Relations/Documentation.php. Accessed July 20 2013.

Concept annotation guidelines. Informatics for Integrating Biology and the Bedside (i2b2). Available from https://www.i2b2.org/NLP/Relations/Documentation.php. Accessed July 20 2013.

Relation annotation guidelines. Informatics for Integrating Biology and the Bedside (i2b2). Available from https://www.i2b2.org/NLP/Relations/Documentation.php. Accessed July 20 2013.

Systematized Nomenclature of Medicine - Clinical Terms (SNOMED-CT). The International Health Terminology Standards Development Organization (IHTSDO). Available from http://www.ihtsdo.org/snomed-ct. Accessed June 13 2013.

Jiang, J. and Zhai, C. (2007). A systematic exploration of the feature space for relation extraction. *In Proceedings of the Annual Conference of the North American Chapter of the Association for Computational Linguistics (NAACL HLT 2007)*, Rochester, USA, 113-120, Association for Computational Linguistics.

Johnson, S.B. (1999). A semantic lexicon for medical language processing. *J Am Med Inform Assn*. **6**(3): 205-218.

Jonnalagadda, S., Cohen, T., Wu, S. and Gonzalez, G. (2012). Enhancing clinical concept extraction with distributional semantics. *J Biomed Inform*. **45**(1): 129-140.

Kang, N., Singh, B., Afzal, Z., van Mulligen, E.M. and Kors, J.A. (2013). Using rule-based natural language processing to improve disease normalization in biomedical text. *J Am Med Inform Assn*. **20**(5): 876-881.

Karim, R.Z., van den Berg, K.S., Colman, M.H., McCarthy, S.W., Thompson, J.F. and Scolyer, R.A. (2008). The advantage of using a synoptic pathology report format for cutaneous melanoma. *Histopathology*. **52**(2): 130-138.

Kierszenbaum, A.L. and Tres, L. (2011). *Histology and cell biology: an introduction to pathology*. Elsevier Health Sciences, Philadelphia, PA.

Kim, J.D., Ohta, T., Tateisi, Y. and Tsujii, J. (2003). GENIA corpus-a semantically annotated corpus for bio-textmining. *Bioinformatics*. **19**: i180-i182.

Kipper-Schuler, k., Kaggal, V., Masanz, J., Ogren, P. and Savova, G. (2008). System evaluation on a named entity corpus from clinical notes. *In Proceedings of the International Conference on Language Resources and Evaluation (LREC 2008)*, Marrakech, Morocco, 3007-3011, European Language Resources Association.

Klein, D. and Manning, C.D. (2003). Accurate unlexicalized parsing. *In Proceedings of the 41st Meeting of the Association for Computational Linguistics*, 1: 423-430.

Koelzer, V.H. and Lugli, A. (2014). The tumor border configuration of colorectal cancer as a histomorphological prognostic indicator. *Front Oncol*. **4**: 29.

Korenius, T., Laurikkala, J., Järvelin, K. and Juhola, M. (2004). Stemming and lemmatization in the clustering of finnish text documents. *In CIKM 04:Proceedings of the thirteenth ACM international conference on Information and knowledge management*, Washington, D.C, USA, 625-633, ACM.

Kripke, S. (1980). *Naming and Necessity*. Harvard University Press, Cambridge, MA.

Lafferty, J., McCallum, A. and Pereira, F. (2001). Conditional random fields: probabilistic models for segmenting and labeling sequence data. *In Proceedings of the 18th International Conference on Machine Learning (ICML 2001)*, Williamstown, USA, 282-289, Morgan Kaufmann.

Leech, G. (1993). Corpus annotation schemes. *Literary and linguistic computing*. **8**(4): 275-281.

Levenshtein, V.I. (1966). Binary codes capable of correcting deletions, insertions and reversals. *Soviet Physics Doklady*. **10**(8): 707-710.

Li, D., Kipper-Schuler, K. and Savova, G. (2008). Conditional random fields and support vector machines for disorder named entity recognition in clinical texts. *In Proceedings of the Workshop on Current Trends in Biomedical Natural Language Processing (BioNLP 2008)*, Columbus, USA, 94-95, Association for Computational Linguistics.

Li, Y. and Martinez, D. (2010). Information extraction of multiple categories from pathology reports. *In Proceedings of the Australasian Language Technology Association Workshop 2010*, Melbourne, Australia, 8: 41-48, Australasian Language Technology Association

Liu, H., Komandur, R. and Verspoor, K. (2011). From graphs to events: a subgraph matching approach for information extraction from biomedical text. *In Proceedings of BioNLP Shared Task 2011 Workshop*, Portland, USA, 164-172, Association for Computational Linguistics.

Liu, H.F., Hu, Z.Z., Zhang, J. and Wu, C. (2006). BioThesaurus: a web-based thesaurus of protein and gene names. *Bioinformatics*. **22**(1): 103-105.

Logical Observation Identifiers Names and Codes (LOINC). Regenstrief Institute, Inc. Available from https://loinc.org. Accessed July 20 2013.

Long, W. (2005). Extracting diagnoses from discharge summaries. *AMIA Annu Symp Proc*: 470-474.

Lussier, Y., Borlawsky, T., Rappaport, D., Liu, Y. and Friedman, C. (2006). PhenoGO: assigning phenotypic context to gene ontology annotations with natural language processing. *Pac Symp Biocomput*: 64-75.

Malouf, R. (2002). A comparison of algorithms for maximum entropy parameter estimation. *In COLING02: proceedings of the 6th conference on Natural language learning*, Taipei, Taiwan, 20: 1-7, Association for Computational Linguistics.

Marcus, M.P., Santorini, B. and Marcinkiewicz, M.A. (1994). Building a large annotated corpus of English: the Penn Treebank. *Computational linguistics*. **19**(2): 313–330.

Martinez, D. and Li, Y. (2011). Information extraction from pathology reports in a hospital setting. *In CIKM 2011: Proceedings of the 20th ACM international conference on Information and knowledge management*, Glasgow, UK, 1877-1882, ACM.

McCowan, I.A., Moore, D.C., Nguyen, A.N., Bowman, R.V., Clarke, B.E., Duhig, E.E. and Fry, M.J. (2007). Collection of cancer stage data by classifying free-text medical reports. *J Am Med Inform Assn*. **14**(6): 736-745.

McCray, A.T. (2003). An upper-level ontology for the biomedical domain. *Comp Funct Genom*. **4**(1): 80-84.

Meystre, S. and Haug, P.J. (2005). Automation of a problem list using natural language processing. *BMC Medical Informatics and Decision Making*. **5**: 30.

Meystre, S.M., Thibault, J., Shen, S.Y., Hurdle, J.F. and South, B.R. (2010). Textractor: a hybrid system for medications and reason for their prescription extraction from clinical text documents. *J Am Med Inform Assn*. **17**(5): 559-562.

Michael-Robinson, J.M., Biemer-Huttmann, A., Purdie, D.M., Walsh, M.D., Simms, L.A., Biden, K.G., Young, J.P., Leggett, B.A., Jass, J.R. and Radford-Smith, G.L. (2001). Tumour infiltrating lymphocytes and apoptosis are independent features in colorectal cancer stratified according to microsatellite instability status. *Gut*. **48**(3): 360-366.

Miller, W., Ota, D., Giacco, G., Guinee, V., Irimura, T., Nicolson, G. and Cleary, K. (1985). Absence of a relationship of size of primary colon-carcinoma with metastasis and survival. *Clin Exp Metastas*. **3**(3): 189-196.

Minard, A.-L., Ligozat, A.-L. and Grau, B. (2011). Multi-class SVM for relation extraction from clinical reports. *In Proceedings of Recent Advances in Natural Language Processing (RANLP 2011)*, Hissar, Bulgaria, 604-609, RANLP 2011 Organising Committee.

Minka, T.P. (2003). Algorithms for maximum-likelihood logistic regression. *Technical Report*, Carnegie Mellon University.

Mitchell, K.J., Becich, M.J., Berman, J.J., Chapman, W.W., Gilbertson, J., Gupta, D., Harrison, J., Legowski, E. and Crowley, R.S. (2004). Implementation and evaluation of a negation tagger in a pipeline-based system for information extract from pathology reports. *Stud Health Technol Inform*. **107**(Pt 1): 663-667.

Mitsumori, T., Murata, M., Fukuda, Y., Doi, K. and Doi, H. (2006). Extracting protein-protein interaction information from biomedical text with SVM. *Ieice T Inf Syst*. **E89d**(8): 2464-2466.

Mohanty, S.K., Piccoli, A.L., Devine, L.J., Patel, A.A., William, G.C., Winters, S.B., Becich, M.J. and Parwani, A.V. (2007). Synoptic tool for reporting of hematological and lymphoid neoplasms based on World Health Organization classification and College of American Pathologists checklist. *Bmc Cancer*. **7**.

Moody, G.B. and Lehman, L.H. (2009). Predicting Acute Hypotensive Episodes: The 10th Annual PhysioNet/Computers in Cardiology Challenge. *Comput Cardiol*: 541-544.

Mutalik, P.G., Deshpande, A. and Nadkarni, P.M. (2001). Use of general-purpose negation detection to augment concept indexing of medical documents: A quantitative study using the UMLS. *J Am Med Inform Assn*. **8**(6): 598-609.

Nagakawa, T., Mori, K., Nakano, T., Kadoya, M., Kobayashi, H., Akiyama, T., Kayahara, M., Ohta, T., Ueno, K., Higashino, Y., Konishi, I. and Miyazaki, I. (1993). Perineural invasion of carcinoma of the pancreas and biliary-tract. *Brit J Surg*. **80**(5): 619-621.

NCI dictionary of cancer terms. National Cancer Institute (NCI). Available from http://www.cancer.gov/dictionary. Accessed July 12 2013.

Nguyen, A., Lawley, M., Hansen, D. and Colquist, S. (2012). Structured pathology reporting for cancer from free text: lung cancer case study. *electronic Journal of Health Informatics*. **7**(1): e8.

Nguyen, A.N., Lawley, M.J., Hansen, D.P., Bowman, R.V., Clarke, B.E., Duhig, E.E. and Colquist, S. (2010). Symbolic rule-based classification of lung cancer stages from free-text pathology reports. *J Am Med Inform Assn*. **17**(4): 440-445.

Nieman, L.Z., Cheng, L., Hormann, M., Farnie, M.A., Molony, D.A. and Butler, P. (2006). The impact of preclinical preceptorships on learning the fundamentals of clinical medicine and physical diagnosis skills. *Acad Med*. **81**(4): 342-346.

Nikolova, I. and Angelova, G. (2011). Identifying relations between medical concepts by parsing UMLS definitions. *In Conceptual Structures for Discovering Knowledge*, Springer Berlin Heidelberg.

Nivre, J., Hall, J., Nilsson, J., Chanev, A., Eryigit, G., Kübler, S., Marinov, S. and Marsi, E. (2007). MaltParser: a language-independent system for data-driven dependency parsing. *Natural Language Engineering*. **13**(2): 95-135.

Unified Medical Language System (UMLS). National Library of Medicine (NLM). Available from http://www.nlm.nih.gov/research/umls/. Accessed June 13 2013.

MetaMap Transfer (MMTx). National Library of Medicine (NLM). Available from http://mmtx.nlm.nih.gov/MMTx/. Accessed June 13 2013.

Nocedal, J. and Wright, S.J. (1999). *Numerical optimization*. Springer New York, NY.

Nochomovitz, L.E. (1998). Application of synoptic reports. *Arch Pathol Lab Med*. **122**(6): 493-494.

Norris, D., Ellis, D., Green, M. and Joske, D. (2010). *Tumours of haematopoietic and lymphoid tissue structured reporting protocol*. Royal College of Pathologists of Australasia. Available from http://www.rcpa.edu.au/Publications/StructuredReporting/CancerProtocols.htm. Accessed 13 June 2012.

Ogino, S., Nosho, K., Irahara, N., Meyerhardt, J.A., Baba, Y., Shima, K., Glickman, J.N., Ferrone, C.R., Mino-Kenudson, M., Tanaka, N., Dranoff, G., Giovannucci, E.L. and Fuchs, C.S. (2009). Lymphocytic reaction to colorectal cancer is associated with longer survival, independent of lymph node count, microsatellite instability, and CPG island methylator phenotype. *Clin Cancer Res*. **15**(20): 6412-6420.

Oliveira Filho, R.S., Ferreira, L.M., Biasi, L.J., Enokihara, M.M., Paiva, G.R. and Wagner, J. (2003). Vertical growth phase and positive sentinel node in thin melanoma. *Braz J Med Biol Res*. **36**(3): 347-350.

Palmer, D.D. and Hearst, M.A. (1994). Adaptive Sentence Boundary Disambiguation. *In ANLC 94: Proceedings of the fourth conference on Applied natural language processing*, Stuttgart, Germany, 78-83 Association for Computational Linguistics

Patrick, J. and Li, M. (2010). High accuracy information extraction of medication information from clinical notes: 2009 i2b2 medication extraction challenge. *J Am Med Inform Assn*. **17**(5): 524-527.

Patrick, J. and Sabbagh, M. (2011). An active learning process for extraction and standardisation of medical measurements by a trainable FSA. *In Proceedings of the 12th International Conference on*

Computational Linguistics and Intelligent Text Processing, Tokyo, Japan, Part II: 151-162, Springer-Verlag.

Patrick, J., Sabbagh, M., Jain, S. and Zheng, H. (2010). Spelling correction in clinical notes with emphasis on first suggestion accuracy. *In the 2nd Workshop on Building and Evaluating Resources for Biomedical Text Mining*, Valletta, Malta, 2-8, The National Centre for Text Mining.

Patrick, J. and Scolyer, R.A. (2008). Information extraction from narrative pathology reports on melanoma. *Quality Use of Pathology Program Project Report*, The University of Sydney, Royal Prince Alfred Hospital.

Patrick, J., Wang, Y. and Budd, P. (2007a). An automated system for conversion of clinical notes into SNOMED clinical terminology. *In Proceedings of the Fifth Australasian Symposium on ACSW Frontiers*, Ballarat, Australia, 68: 219-226, Australian Computer Society, Inc.

Patrick, J., Wang, Y. and Budd, P. (2007b). An automated system for conversion of clinical notes into SNOMED clinical terminology. *In Proceedings of the fifth Australasian symposium on ACSW frontiers*, Ballarat, Australia, 68: 219-226, Australian Computer Society, Inc.

Patrick, J.D., Nguyen, D.H.M., Wang, Y.F. and Li, M. (2011). A knowledge discovery and reuse pipeline for information extraction in clinical notes. *J Am Med Inform Assn*. **18**(5): 574-579.

Porter, M.F. (2006). An algorithm for suffix stripping. *Program-Electron Lib*. **40**(3): 211-218.

Pudil, P., Novovicova, J. and Kittler, J. (1994). Floating search methods in feature-selection. *Pattern Recogn Lett*. **15**(11): 1119-1125.

Qu, Z.H., Ninan, S., Almosa, A., Chang, K.G., Kuruvilla, S. and Nguyen, N. (2007). Synoptic reporting in tumor pathology - Advantages of a web-based system. *Am J Clin Pathol*. **127**(6): 898-903.

Cancer reporting protocols. Available from http://www.rcpa.edu.au/Publications/StructuredReporting/CancerProtocols.htm. Accessed 13 July 2013.

Structured pathology reporting of cancer. Available from http://www.rcpa.edu.au/publications/structuredreporting.htm. Accessed 13 July 2013.

Macroscopic Cut-Up Manual for Colorectal tumour. RCPA - Royal College of Pathologists of Australasia. Available from http://www.rcpa.edu.au/Library/Practising-Pathology/Macroscopic-Cut-Up/Specimen/Gastrointestinal/Colorectal/Colorectal-tumour. Accessed July 12 2013.

Reintgen, D. (2001). Establishing a standard of care for the patient with melanoma. *Ann Surg Oncol*. **8**(2): 91-91.

RelEx dependency relationship extractor. launchpad. Available from https://launchpad.net/relex. Accessed June 13 2013.

Reynar, J.C. and Ratnaparkhi, A. (1997). A maximum entropy approach to identifying sentence boundaries. *In ANLC 97: Proceedings of the fifth conference on Applied natural language processing*, Washington, D.C, USA, 16-19 Association for Computational Linguistics.

Richesson, R.L. and Krischer, J. (2007). Data standards in clinical research: gaps, overlaps, challenges and future directions. *J Am Med Inform Assn*. **14**(6): 687-696.

Rink, B., Harabagiu, S. and Roberts, K. (2011). Automatic extraction of relations between medical concepts in clinical texts. *J Am Med Inform Assn*. **18**(5): 594-600.

Roberts, A., Gaizauskas, R., Hepple, M., Davis, N., Demetriou, G., Guo, Y., Kola, J., Roberts, I., Setzer, A., Tapuria, A. and Wheeldin, B. (2007). The CLEF corpus: semantic annotation of clinical text. *AMIA Annu Symp Proc*: 625-629.

Roberts, A., Gaizauskas, R., Hepple, M., Demetriou, G., Guo, Y., Roberts, I. and Setzer, A. (2009). Building a semantically annotated corpus of clinical texts. *J Biomed Inform*. **42**(5): 950-966.

Roberts, A., Gaizauskas, R., Hepple, M. and Guo, Y. (2008a). Combining terminology resources and statistical methods for entity recognition: an evaluation. *In Proceedings of the International Conference on Language Resources and Evaluation (LREC 2008)*, Marrakech, Morocco, 2974-2979, European Language Resources Association.

Roberts, A., Gaizauskas, R., Hepple, M. and Guo, Y.K. (2008b). Mining clinical relationships from patient narratives. *Bmc Bioinformatics*. **9**.

Rosario, B. and Hearst, M.A. (2004). Classifying semantic relations in bioscience texts. *In Proceedings of the 42nd Annual Meeting on Association for Computational Linguistics (ACL 2004)*, Barcelona, Spain, 431-438, Association for Computational Linguistics.

Sætre, R., Sagae, K. and Tsujii, J. (2007). Syntactic features for protein-protein interaction extraction. *In Proceedings of LBM 2007: the 2nd International Symposium on Languages in Biology and Medicine*, Biopolis, Singapore, 9, BMC Bioinformatics.

Sager, N., Lyman, M., Bucknall, C., Nhan, N. and Tick, L.J. (1994). Natural-language processing and the representation of clinical-data. *J Am Med Inform Assn*. **1**(2): 142-160.

Sahiner, B., Chan, H.P., Petrick, N., Wagner, R.F. and Hadjiiski, L. (2000). Feature selection and classifier performance in computer-aided diagnosis: The effect of finite sample size. *Med Phys*. **27**(7): 1509-1522.

Saltz, L.B. (2002). *Colorectal cancer: multimodality management*. Humana Press, Totowa, NJ.

Sang, E. and Erik, F. (2002). Introduction to the CoNLL-2002 shared task: Language-independent named entity recognition. *In COLING02: proceedings of the 6th conference on Natural language learning*, Taipei, Taiwan, 155–158, Association for Computational Linguistics.

Savova, G.K., Chapman, W.W., Zheng, J.P. and Crowley, R.S. (2011). Anaphoric relations in the clinical narrative: corpus creation. *J Am Med Inform Assn*. **18**(4): 459-465.

Savova, G.K., Masanz, J.J., Ogren, P.V., Zheng, J.P., Sohn, S., Kipper-Schuler, K.C. and Chute, C.G. (2010). Mayo clinical Text Analysis and Knowledge Extraction System (cTAKES): architecture, component evaluation and applications. *J Am Med Inform Assn*. **17**(5): 507-513.

Schadow, G. and McDonald, C.J. (2003). Extracting structured information from free text pathology reports. *AMIA Annu Symp Proc*: 584-588.

Schuemie, M.J., Jelier, R. and Kors, J.A. (2007). Peregrine: lightweight gene name normalization by dictionary lookup. *In Proceedings of the BioCreAtIvE II Workshop*, Madrid, Spain, 131-133, CNIO.

Scolyer, R., Ellis, D., Heenan, P. and James, C. (2010). *Primary cutaneous melanoma structured reporting protocol*. Royal College of Pathologists of Australasia. Available from http://www.rcpa.edu.au/Publications/StructuredReporting/CancerProtocols.htm. Accessed 13 June 2013.

Scolyer, R.A., Shaw, H.M., Thompson, J.F., Li, L.X., Colman, M.H., Lo, S.K., McCarthy, S.W., Palmer, A.A., Nicoll, K.D., Dutta, B., Slobedman, E., Watson, G.F. and Stretch, J.R. (2003). Interobserver reproducibility of histopathologic prognostic variables in primary cutaneous melanomas. *Am J Surg Pathol*. **27**(12): 1571-1576.

Scolyer, R.A., Thompson, J.F., Stretch, J.R., Sharma, R. and McCarthy, S.W. (2004). Pathology of melanocytic lesions: New, controversial, and clinically important issues. *J Surg Oncol*. **86**(4): 200-211.

Sha, F. and Pereira, F. (2003). Shallow parsing with conditional random fields. *In Proceedings of HLT-NAACL 2003*, Edmonton, Canada, 134-141, Association for Computational Linguistics.

Shen, H. and Sarkar, K. (2005). Voting between multiple data representations for text chunking. *In Advances in Artificial Intelligence: 18th Conference of the Canadian Society for Computational Studies of Intelligence*, Victoria, Canada, 389-400, Springer Berlin Heidelberg.

Shewchuk, J.R. (1994). An introduction to the conjugate gradient method without the agonizing pain. *Technical Report*, Carnegie Mellon University.

Sleator, D. and Temperley, D. (1991). Parsing English with a link grammar. *Technical Report*, Carnegie Mellon University.

Sobin, L., Gospodarowicz, M., Wittekind, C. and International Union against Cancer. (2009). *TNM classification of malignant tumours*. Wiley-Blackwell, Chichester, UK and Hoboken, New Jersey.

Sohn, S., Wu, S. and Chute, C.G. (2012). Dependency parser-based negation detection in clinical narratives. *AMIA Summits Transl Sci Proc*. **2012**: 1-8.

Sondergaard, K. and Hou-Jensen, K. (1985). Partial regression in thin primary cutaneous malignant melanomas clinical stage I. A study of 486 cases. *Virchows Arch A Pathol Anat Histopathol*. **408**(2-3): 241-247.

Spackman, K.A., Campbell, K.E. and Cote, R.A. (1997). SNOMED RT: a reference terminology for health care. *J Am Med Inform Assn*: 640-644.

Srigley, J.R., McGowan, T., Maclean, A., Raby, M., Ross, J., Kramer, S. and Sawka, C. (2009). Standardized synoptic cancer pathology reporting: a population-based approach. *J Surg Oncol*. **99**(8): 517-524.

Stamp, M. (2004). *A revealing introduction to hidden Markov models*. Department of Computer Science, San Jose State University. Available from http://www.cs.sjsu.edu/faculty/stamp/RUA/HMM.pdf. Accessed 13 June 2013.

Stone, M. (1974). Cross-validatory choice and assessment of statistical predictions. *J R Stat Soc B*. **36**(2): 111-147.

Stone, P.J., Dunphy, D.C., Smith, M.S. and Ogilvie, D.M. (1966). *The General Inquirer: a computer approach to content analysis*. MIT press, Cambridge, MA.

Sutton, C., McCallum, A. and Rohanimanesh, K. (2007). Dynamic conditional random fields: factorized probabilistic models for labeling and segmenting sequence data. *J Mach Learn Res*. **8**: 693-723.

Taira, R.K., Soderland, S.G. and Jakobovits, R.M. (2001). Automatic structuring of radiology free-text reports. *Radiographics*. **21**(1): 237-245.

Tang, B., Cao, H., Wu, Y., Jiang, M. and Xu, H. (2012). Clinical entity recognition using structural support vector machines with rich features. *In Proceedings of the ACM Sixth International Workshop On Data And Text Mining In Biomedical Informatics (DTMBIO 2012)*, Maui, USA, 13-20, ACM.

Tannapfel, A., Wittekind, C. and Hunefeld, G. (1992). Ductal adenocarcinoma of the pancreas - histopathological features and prognosis. *Int J Pancreatol*. **12**(2): 145-152.

Tanushi, H., Dalianis, H., Duneld, M., Kvist, M., Skeppstedt, M. and Velupillai, S. (2013). Negation scope delimitation in clinical text using three approaches: NegEx, PyConTextNLP and SynNeg. *In Proceedings of the 19th Nordic Conference of Computational Linguistics (NODALIDA 2013)*, Oslo, Norway, 387-398, Linköping University Electronic Press.

Tateisi, Y. and Tsujii, J. (2004). Part-of-speech annotation of biology research abstracts. *In Proceedings of 4th International Conference on Language Resource and Evaluation (LREC2004)*, Lisbon, Portugal, IV: 1267-1270, European Language Resources Association.

Taylor, R.C., Patel, A., Panageas, K.S., Busam, K.J. and Brady, M.S. (2007). Tumor-infiltrating lymphocytes predict sentinel lymph node positivity in patients with cutaneous melanoma. *J Clin Oncol*. **25**(7): 869-875.

Thompson, C.A., Califf, M.E. and Mooney, R.J. (1999). Active Learning for Natural Language Parsing and Information Extraction. *In the Sixteenth International Conference on Machine Learning (ICML1999)*, Bled, Slovenia, 406-414, Morgan Kaufmann.

Tikk, D. and Solt, I. (2010). Improving textual medication extraction using combined conditional random fields and rule-based systems. *J Am Med Inform Assn*. **17**(5): 540-544.

Torii, M., Hu, Z.Z., Wu, C.H. and Liu, H.F. (2009). BioTagger-GM: a gene/protein name recognition system. *J Am Med Inform Assn*. **16**(2): 247-255.

Tsai, R.T.H., Wu, S.H., Chou, W.C., Lin, Y.C., He, D., Hsiang, J., Sung, T.Y. and Hsu, W.L. (2006). Various criteria in the evaluation of biomedical named entity recognition. *Bmc Bioinformatics*. **7**: 1-8.

Tsochantaridis, I., Joachims, T., Hofmann, T. and Altun, Y. (2005). Large margin methods for structured and interdependent output variables. *J Mach Learn Res*. **6**: 1453-1484.

Tsuruoka, Y., Tateishi, Y., Kim, J.D., Ohta, T., McNaught, J., Ananiadou, S. and Tsujii, J. (2005). Developing a robust part-of-speech tagger for biomedical text. *Lect Notes Comput Sc*. **3746**: 382-392.

Turchin, A., Kolatkar, N.S., Grant, R.W., Makhni, E.C., Pendergrass, M.L. and Einbinder, J.S. (2006). Using regular expressions to abstract blood pressure and treatment intensification information from the text of physician notes. *J Am Med Inform Assn*. **13**(6): 691-695.

Uzuner, O., Bodnari, A., Shen, S.Y., Forbush, T., Pestian, J. and South, B.R. (2012). Evaluating the state of the art in coreference resolution for electronic medical records. *J Am Med Inform Assn*. **19**(5): 786-791.

Uzuner, O., Solti, I. and Cadag, E. (2010). Extracting medication information from clinical text. *J Am Med Inform Assn*. **17**(5): 514-518.

Uzuner, O., South, B.R., Shen, S.Y. and DuVall, S.L. (2011). 2010 i2b2/VA challenge on concepts, assertions, and relations in clinical text. *J Am Med Inform Assn*. **18**(5): 552-556.

Uzuner, O., Zhang, X.R. and Sibanda, T. (2009). Machine learning and rule-based approaches to assertion classification. *J Am Med Inform Assn*. **16**(1): 109-115.

Van Cutsem, E., Nordlinger, B., Cervantes, A. and Grp, E.G.W. (2010). Advanced colorectal cancer: ESMO clinical practice guidelines for treatment. *Ann Oncol*. **21**: v93-v97.

Verbeke, C.S., Leitch, D., Menon, K.V., McMahon, M.J., Guillou, P.J. and Anthoney, A. (2006). Redefining the R1 resection in pancreatic cancer. *Brit J Surg*. **93**(10): 1232-1237.

Vincze, V., Szarvas, G., Farkas, R., Mora, G. and Csirik, J. (2008). The BioScope corpus: biomedical texts annotated for uncertainty, negation and their scopes. *Bmc Bioinformatics*. **9**.

Wang, H., Zhao, T., Tan, H. and Zhang, S. (2008). Biomedical named entity recognition based on classifiers ensemble. *International Journal of Computer Science and Applications*. **5**(2): 1-11.

Wang, X.Y., Chase, H., Markatou, M., Hripcsak, G. and Friedman, C. (2010). Selecting information in electronic health records for knowledge acquisition. *J Biomed Inform*. **43**(4): 595-601.

Wang, Y. (2009). Annotating and recognising named entities in clinical notes. *In Proceedings of the ACL-IJCNLP 2009 Student Research Workshop*, Suntec, Singapore, 18-26, ACL and AFNLP.

Wang, Y. and Patrick, J. (2008). Mapping Clinical Notes to Medical Terminology at Point of Care. *In Proceedings of Current Trends in Biomedical Natural Language Processing (BioNLP 2008)*, Columbus, USA, 102-103, Association for Computational Linguistics.

Wang, Y. and Patrick, J. (2009). Cascading classifiers for named entity recognition in clinical notes. *In Workshop Biomedical Information Extraction 2009*, Borovets, Bulgaria, 42-49, Association for Computational Linguistics.

Grady Ward's Moby. The Institute for Language Speech and Hearing, The University of Sheffield. Available from http://icon.shef.ac.uk/Moby/. Accessed 13 July 2013.

Way, T.W., Sahiner, B., Hadjiiski, L.M. and Chan, H.P. (2010). Effect of finite sample size on feature selection and classification: A simulation study. *Med Phys*. **37**(2): 907-920.

Whitney, A.W. (1971). A direct method of nonparametric measurement selection. *IEEE Transactions on Computers*. **20**(9): 1100-1103

ICD-O-3: International Classification of Diseases for Oncology, 3rd Edition. WHO - World Health Organization. Available from http://www.who.int/classifications/icd/adaptations/oncology/en/. Accessed July 12 2013.

International Classification of Diseases (ICD). World Health Organization (WHO). Available from http://www.cdc.gov/nchs/icd.htm. Accessed June 13 2013.

Wikipedia. Wikimedia Foundation, Inc. Available from http://www.wikipedia.org. Accessed June 13 2013.

Wright, F.C., Law, C.H.L., Last, L.D., Ritacco, R., Kumar, D., Hsieh, E., Khalifa, M. and Smith, A.J. (2004). Barriers to optimal assessment of lymph nodes in colorectal cancer specimens. *Am J Clin Pathol*. **121**(5): 663-670.

Xu, H., Stenner, S.P., Doan, S., Johnson, K.B., Waitman, L.R. and Denny, J.C. (2010). MedEx: a medication information extraction system for clinical narratives. *J Am Med Inform Assn*. **17**(1): 19-24.

Xu, Y., Tsujii, J. and Chang, E.I.C. (2012). Named entity recognition of follow-up and time information in 20 000 radiology reports. *J Am Med Inform Assn*. **19**(5): 792-799.

Xuan, W., Watson, S.J. and Meng, F. (2007). Tagging Sentence Boundaries in Biomedical Literature. *In CICLing 2007: Eighth International Conference on Intelligent Text Processing and Computational Linguistics*, Mexico City, Mexico, 186 – 195, Springer-Verlag Berlin Heidelberg.

Yang, H. (2010). Automatic extraction of medication information from medical discharge summaries. *J Am Med Inform Assn*. **17**(5): 545-548.

Zhou, G.D. and Su, J. (2004). Exploring deep knowledge resources in biomedical name recognition. *In Proceedings of the 2004 Joint Workshop on Natural Language Processing in Biomedicine and its Applications (JNLPBA 2004)*, Geneva, Switzerland, 171-178, COLING.

Zweigenbaum, P., Bachimont, B., Bouaud, J., Charlet, J. and Boisvieux, J.F. (1995). A multi-lingual architecture for building a normalised conceptual representation from medical language. *Proc Annu Symp Comput Appl Med Care*: 357-361.