Building a semantically annotated corpus of clinical texts

Angus Roberts *, Robert Gaizauskas, Mark Hepple,

George Demetriou, Yikun Guo, Ian Roberts, Andrea Setzer

Department of Computer Science, University of Sheffield, Regent Court, 211 Portobello, Sheffield S1 4DP, United Kingdom

Abstract

In this paper we describe the construction of a semantically annotated corpus of clinical texts for use in the development and evaluation of systems for automatically extracting clinically significant information from the textual component of patient records. The paper details the sampling of textual material from a collection of 20,000 cancer patient records, the development of a semantic annotation scheme, the annotation methodology, the distribution of annotations in the final corpus, and the use of the corpus for development of an adaptive information extraction system. The resulting corpus is the most richly semantically annotated resource for clinical text processing built to date, whose value has been demonstrated through its use in developing an effective information extraction system. The detailed presentation of our corpus construction and annotation methodology will be of value to others seeking to build high-quality semantically annotated corpora in biomedical domains.

Key words: Corpora; Semantic annotation; Clinical text; Natural language processing; Gold standards; Evaluation; Information Extraction; Text mining; Temporal annotation; Annotation guidelines

1 Introduction

We describe the creation of a semantically annotated corpus of clinical texts. The documents of this corpus are drawn from the free text component of patient records, and the annotations capture clinically significant information communicated by these texts. The corpus is intended for use in developing and evaluating systems that can *automatically* extract this kind of clinically significant information from the textual component of patient records. The corpus has been created within the context of the CLinical E-Science Framework (CLEF) project [1]: a multi-site research project that has been developing the technology and techniques required for a high quality repository of electronic patient records. Such a repository must meet high standards of security and interoperability, and should enable ethical and user-friendly access to patient information, so as to facilitate both clinical care and biomedical research. CLEF has chosen to work in the area of cancer informatics, as one of the project partners – the Royal Marsden Hospital (RMH) – is a large specialist oncology centre.

Although much of the patient information needed to populate such a repository exists as structured data, e.g. database records of drug prescriptions and clinic appointments, free text material still forms an important component of electronic patient records, and contains information that is potentially significant both for day-to-day care and clinical research. For example, letters

* Corresponding author. Fax: +44 114 222 1810
Email address: a.roberts@dcs.shef.ac.uk (Angus Roberts).

written from the secondary to the primary care physician (e.g. from specialist consultant to patient GP) form a major component of any UK medical record, and free text plays a key role in the reporting of imaging and pathology findings. Clinical narratives may record, for instance, why drugs were given or discontinued, the results of physical examination, and issues considered important when discussing patient care but which are not coded for audit. Such information, when combined with that from the structured record, and suitably presented, could contribute to individual patient care, e.g. providing a consultant with a concise summary of their patient's clinical history, or access to concise histories for patients with similar conditions elsewhere. Aggregation of information across all the records in a large repository could bring benefits for clinical research. For example, being able to get answers to questions such as "*How many patients with stage 2 adenocarcinoma who were treated with tamoxifen were symptom-free after 5 years?*" could assist a researcher in formulating hypotheses that could be later explored in clinical trials.

The need to make the information that exists in clinical texts available for integration with the structured record, for subsequent use in clinical care and research, has been addressed within CLEF through the use of *information extraction* (IE) technology [2,3]. Although some IE research has focused on unsupervised methods of developing systems, as in the earlier work of Riloff [4], most practical modern IE work requires data that have been manually annotated with the events, entities and relationships that are considered to express key content for the given domain. These data serve three purposes. Firstly, the analysis of data that is required to create the annotation scheme serves to focus and clarify the information requirements of the task and domain. Secondly, the annotated data provide a *gold standard* against which to assess the

performance of systems designed to automatically identify this information in texts. Thirdly, it serves as a resource for system development: extraction rules may be created either automatically or by hand, and statistical models of the text may be built by machine learning algorithms.

This paper reports on the work done within CLEF to create an annotated corpus, to aid the development and evaluation of the CLEF IE system. To the best of our knowledge, no one else has explored the problem of producing a corpus annotated for clinical IE to the depth and extent reported here, and the resulting corpus is the most richly semantically annotated resource for clinical text processing built to date. Our annotation exercise draws its texts from a large background corpus of clinical narratives, covers multiple text types, and involves over 20 annotators. Results are encouraging, and suggest that a rich corpus to support IE in the medical domain can be created.

We reported the early development of the CLEF corpus in [5]. The current paper elaborates quantitative results from this development process, giving a much greater level of detail. Quantitative results have also previously been given, for the partially complete corpus, in [6]. The results in the current paper are final, reflecting the finished corpus. In addition, the current paper provides results and descriptions not previously published, including: annotation with UMLS CUIs; annotation of temporal expressions; the summary results of an annotator difference analysis; a discussion of time taken to annotate; detailed descriptions of the annotation guidelines, their development and application; and greater detail of our annotation methodology. We also summarise work on the corpus in use, to train and evaluate a working IE system. We believe that this detailed account of our methodology, corpus, and its use will be of benefit to other groups contemplating similar exercises.

The paper is organised as follows. In the next section, we summarise previous efforts to create annotated corpora in biomedical domains. Section 3 describes how material was selected for inclusion in our corpus, and then in Section 4, we describe the semantic annotation schema, the annotation methodology, the development of the annotation guidelines, as well as the measures for assessing the consistency of human annotations. Section 5 presents an analysis of aspects of the annotation process and Section 6 presents inter annotator agreement scores for the finished corpus, and figures on the distribution of entity and relation types by document type across the corpus. The next section describes work carried out subsequent to the initial corpus construction work, to add a layer of temporal annotation. Finally, in Section 8, we mention on-going use of the corpus for training and evaluation of our supervised machine learning IE system.

2 Annotated Corpora for Biomedical Research

Annotated corpora, or text collections, are now recognized as resources of central importance in biomedical language processing research. They may be taxonomized in various ways. For example, they can be grouped by domain (e.g protein-protein interactions, oncology), document type or genre (e.g. research article, clinical narrative, radiology report), type of annotation (e.g. semantic – entities, relations and/or syntactic – part-of-speech, parse structure), intended language processing application (e.g. information extraction, text classification), intended mode of use (e.g. for training adaptive systems, for specific system evaluation, for community wide shared task evaluation), or availability (e.g. publicly available or not publicly available). It is not our

intention to attempt a complete characterisation and review of all annotated corpus resources that have been used in biomedical language processing research. Instead we focus on a few that enable us to show where the CLEF corpus fits in the context of prior research and what novel contribution it makes.

The CLEF corpus may be characterised as a semantically annotated corpus of clinical documents of mixed type (clinic letters, radiology and histopathology reports) which is designed to support both automated training and evaluation of information extraction systems. While it is not publicly available at time of writing we are working towards its release (see below) and reusability has been an important consideration informing its design.

There are now a significant number of publicly available semantically annotated corpora designed to support information extraction research comprising texts drawn from the biomedical research literature. For example, the GENIA corpus is a collection of ∼200 Medline abstracts in the area of molecular biology that has had mentions of specific biological entities and events annotated within it [7,8]. The PennBioIE corpus [9] consists of ∼2300 Medline abstracts, in the domains of molecular genetics of oncology and inhibition of enzymes of the CYP450 class and is annotated for biomedical entity types (it is also annotated syntactically for parts-of-speech amd some portion of it has been annotated for Penn Treebank style syntactic structure). The Yapex corpus contains 200 Medline abstracts annotated for protein names [10]. The BioText project has made several semantically annotated corpora available, including one for disease-treatment relation classification consisting of ∼3500 sentences drawn from Medline abstracts labelled for DISEASE and TREATMENT and seven types of relation holding between them [11], and one for

protein-protein interaction classification consisting of ∼800 sentences drawn from full-text journal papers, where each sentence contains mentions of an interacting protein pair [12]. The ITI TXM corpus [13] has annotated tissue expressions in 238 full-text documents drawn from PubMed and protein-protein interactions in 217 documents obtained from PubMedCentral and PubMed.

While these corpora have been developed in the contexts of specific research projects they have been developed with a view to reusability and have been released to the wider research community. Other semantically annotated corpora drawn from the biomedical research literature have been developed specifically for the purpose of shared task evaluations of information extraction systems. These evaluations include the Biocreative challenge, which utilized the GENETAG corpus containing 20,000 sentences with gene/protein names annotated [14]), the LLL05 challenge task, which supplied training and test data for the task of identifying protein/gene interactions in sentences from Medline abstracts [15], and the TREC Genomics Track, which, while focussed on information retrieval rather than information extraction, did yield some datasets which could be viewed as semantically annotated, e.g. the TREC 2007 task for which human relevance judgements include lists of domain-specific entities associated with relevant passages [16].

The corpora mentioned so far consist of texts drawn from the research literature. Corpora consisting of clinical texts, e.g clinic letters, radiology and histopatholgy reports, are much rarer – getting access to clinical text for research purposes is difficult due to issues of patient confidentiality and getting permission to release them to the wider research community is even more challenging. To our knowledge the only annotated corpora intended to support research in clinical information retrieval and extraction that have been

released to the wider research community are those developed in the context of several recent shared task challenges. For example, the corpus prepared and released for the Computational Medicine Challenge [17] consists of 1954 (978 training, 976 test) radiology reports annotated with ICD-9-CM codes, where the challenge is to automatically code the unseen test data. The ImageCLEFmed 2005 and 2006 image test collections consist of ∼50,000 images with associated textual annotations (case descriptions, imaging reports) and in some cases metadata (e.g. DICOM labels), together with query topics and relevance judgements [18,19]. While intended to support medical image retrieval research, the textual component of this resource could have purely language processing applications. Finally, the I2B2 challenges, have provided training and evaluation data for de-identification of discharge summaries, the identification of smoking status from discharge summaries, and the identification of obesity and co-morbidities from discharge summaries [20].

These are the only publicly released semantically annotated clinical corpora of which we are aware. However, various research projects have developed and published descriptions of clinical corpora used for training and/or evaluation within their project which may be viewed as "semantically annotated" in some sense. Ogren et al. [21], for example, describe work on annotating disorders within clinic notes with a view to training and testing a named entity recognition system. Meystre and Haug [22] describe the development of corpus of 160 clinical documents of mixed type (diagnostic procedure reports, radiology reports, history and physicals, etc.) in which medical problems are identified manually for use in evaluating their system which attempts to extract a patient "problem list" from a clinical document. However it appears that specific mentions of these problems are not annotated where they occur

in the text, but rather that problems are associated with a text at document level, reducing the utility of the corpus for supervised learning. Denny et al. [23] construct a "gold standard" corpus of medical school lecture documents in which biomedical concepts have been manually identified for use in evaluating their KnowledgeMap tool which aims to automatically identify such concepts. Again it appears that in the gold standard the concepts are associated with the text at document level, rather than at the mention level within the running text. Assessing the ability to correctly identify the negations of clinical concepts in clinical texts is the focus of a study by Elkin et al. [24] who have manually verified whether the clinical concepts in a set of 41 clinical documents are negated or not, yielding an annotated evaluation resource for concept negation in clinical texts. Of course the long history of interest in constructing clinical information extraction systems has left a correspondingly long series of gradually maturing evaluations of these systems many of which produced evaluation resources that can be viewed as semantically annotated corpora. Friedman and Hripcsak [25] present an extensive review of work on evaluating natural language processing systems in the clincal domain, especially information extraction systems, prior to 1998, including discussion of any evaluation resources these evaluations have produced.

The CLEF corpus may be differentiated from the annotation work mentioned above in several regards. First, so far as we are aware, it is the first corpus of clinical texts to be annotated with information about clinical relations as well entities. Secondly the range of entity types for which all mentions are annotated in the running text, as opposed to merely being associated with the text at document level is much wider than in previous efforts, making the resource of significantly greater utility for supervised learning. Thirdly,

it is the first biomedical corpus to be annotated with temporal information. Taken together these features make the CLEF corpus the richest semantically annotated corpus of clinical texts yet developed. Finally, it is worth mentioning that the corpus has been designed with a view to reuse by using standards such as XML for the markup and by producing documentation for others to use, something that differentiates it from many project-specific evaluations.

3 Selection of Corpus Material

Our corpus comes from CLEFs main clinical partner, the Royal Marsden Hospital, Europe's largest specialist oncology centre. The entire corpus consists of both the structured records and free text documents from 20234 deceased patients. The free text documents are of three types: clinical narratives (with sub-types as shown in Table 1); histopathology reports; and imaging reports. Patient confidentiality is ensured through a variety of technical and organisational measures, including automatic pseudonymisation and manual inspection. Approval to use this corpus for research purposes within CLEF was sought and obtained from the Thames Valley Multi-centre Research Ethics Committee (MREC).

3.1 Document Sampling

Given the expense of human annotation, the annotated portion of the corpus – which we refer to as the gold standard corpus – has to be a relatively small subset of the whole corpus of 565000 documents. In order to avoid events that are either rare or outside of the main project requirements, the gold standard

is restricted by diagnosis, and only considers documents from those patients with a primary diagnosis code in one of the top level sub-categories of ICD-10 Chapter II (neoplasms) [26]. In addition, it only contains those sub-categories that cover more than 5% of the total number of narratives and reports in the whole corpus. The gold standard corpus consists of three portions, selected for slightly different purposes.

3.1.1 Whole patient records

Two applications in CLEF involve aggregating data across a single patient record. The CLEF chronicle builds a chronological model for a patient, integrating events from both the structured and unstructured record [27]. CLEF report generation creates aggregated graphical and textual reports from the chronicle [28]. These two applications require whole patient records for development and testing. Two whole patient records were selected for this portion of the corpus, from two of the major diagnostic categories, to give median numbers of documents, and a mix of document types and lengths. For each patient, the record comprises nine narratives, one imaging report and seven histopathology reports, plus associated structured data.

3.2 Stratified random sample

The major portion of the gold standard serves as development and evaluation material for IE. In order to ensure even training and fair evaluation across the entire corpus, the sampling of this portion is randomised and stratified, so that it reflects the population distribution along various axes. Table 1 shows the proportions of clinical narratives along two of these axes. The random sample

consists of 50 each of clinical narratives, histopathology reports, and imaging reports.

The numbers of documents chosen for annotation were based on two factors. First, preliminary experiments using documents annotated with a small number of entity types had shown that performance of an adaptive IE system plateaued with around 40 documents used for training. Second, from a purely pragmatic point of view, we only had a limited amount of annotator time. We used empirically based estimates of the time taken to annotate each document, to calculate the number of documents we could annotate in the time available. Time for annotator training was factored in.

Thirty-two documents of mixed type were also randomly chosen for use in annotator training and guideline development. These documents were annotated, but were not used as part of the final gold standard.

3.3 Development corpus

The stratified random corpus was only ever examined by annotators, and not by system developers, who remained blind to its contents throughout. This policy was implemented to avoid there being any developments of the system which were cued specifically by the characteristics of documents that might ultimately be used in scoring the system's performance, as this would contaminate the evaluation.

It is, however, essential for developers to have some documents to work with. A "mirror" corpus of the stratified random corpus was therefore created. This consisted of different documents, but with the same document types, and

stratified in the same proportions along the same axes. This corpus was never annotated. It was available to system developers as required.

4 The CLEF Annotation Schema and its Development

The CLEF gold standard is a semantically annotated corpus. We are interested in identifying the key clinical entities mentioned in the text. By entity, we mean some real-world thing or occurrence referred to in the text such as the drugs that have been administered, the tests that were carried out, etc. We are also interested in determining the relationships between entities: the condition indicated by a drug, the result of an investigation, etc.

Annotation is anchored in the text. Annotators mark spans of text with a type: drug, locus and so on. Annotators may also mark words that modify spans (such as negation), and mark relationships as links between spans. Two or more spans may refer to the same entity in the real world, in which case they co-refer. Co-referring CLEF entities are linked by the annotators. An example illustrating some aspects of annotation is shown in Figure 1. The types of annotation are described in a schema, shown in Figure 2. The CLEF entities, relations, modifiers and co-reference are also listed in Tables 2 and 3, along with descriptions and examples.

Relationships include those that are obvious from the linguistic structure of the text, and those that need some level of domain knowledge to infer. As an example of the latter, consider the example: *"FBC and U&E were requested. She was severely anaemic."* In this, knowledge is required to infer that there is a relationship `FBC has_finding anaemia`. In practice, the distinction between

linguistic and domain knowledge is blurred, and it proves difficult to decide which relationships are based on which type of knowledge. We have therefore made no attempt to differentiate between these two categories of relationship in our schema, taking the view that such a distinction could be added as a separate layer of annotation if required.

The schema is based on a set of requirements developed between clinicians and computational linguists in CLEF. The schema types are mapped to types in the UMLS semantic network, which enables us to utilize UMLS vocabularies in entity recognition. The aim of annotation was to provide general semantic types for entities, and not to map entities to any particular codified terminology. Mapping to specific terminologies was considered to be an extra layer of annotation, performed for specific applications that require it, as described in Section 4.6. For the purposes of annotation, the schema is modeled as a Protégé-Frames ontology [29]. Annotation is carried out using an adapted version of the Knowtator plugin for Protégé [30]. This was chosen for its handling of relationships, after evaluating several such tools.

4.1 The Annotation Guidelines

Consistency is critical to the quality of a gold standard. It is important that all documents are annotated to the same standard. Questions regularly arise when annotating. For example, should multi-word expressions be split? Should "myocardial infarction" be annotated as a condition only, or as a condition and a locus? To ensure consistency, a set of guidelines is provided to annotators. These describe in detail what should and should not be annotated; how to decide if two entities are related; how to deal with co-reference; and a number

of special cases. The guidelines also provide a sequence of steps, a recipe, which annotators should follow when working on a document. This recipe is designed to minimise errors of omission. The guidelines themselves were developed through a rigorous, iterative process, which is described below.

4.2 The origin of the guidelines

The guidelines originated from IE *template* definitions, in an initial CLEF IE system [3], which were themselves patterned on the set of template definitions used in the Message Understanding Conferences (see e.g. [31]). A template is a structured object representing domain-specific entities, their properties, and the relationships between them. A template represents something in the real world. The template does not, however, relate directly to a specific span of text: it is independent of the text. A template may be instantiated, even though the entity it describes is not directly mentioned in the text. For example, a text that discusses angina could lead to a `heart` template being created.

The CLEF templates modelled a large and ambitious set of nine entities with sixteen different relationships between them. Each entity also had a number of properties that were to be extracted, for example, the `course` of a `condition`, or the `goal` of an `intervention`. The entities and relationships were themselves based on an ontology that attempted to model every aspect of the patient and treatment, as described in the clinical documents.

The template definitions were drawn up in collaboration with a single medical informatician, and were tested by the same medical informatician, by manually filling the templates for a small number of documents. This set of documents

became a gold standard for system development and testing. With use, a number of problems became apparent in this gold standard. First, although there was a good formal description of how templates should be filled, there was no description of how they should be created. Should a single template be created for every mention of a patient's bladder, or should just one be created? This led to template construction that was idiosyncratic, and at odds with the requirements of information extraction. Secondly, the complexity of the ontology, the resulting templates, and the limitations of the tools used (text editors), meant that template filling was slow and painful. This in turn led to insufficient data for system development and testing. Lastly, templates are not anchored in the text. This means that when comparing a template in the gold standard to a template created by a IE system, we must first decide whether they are referring to the same thing. For example, suppose a text mentions the two distinct kidneys of a patient, and as a consequence, in the gold standard there are two `kidney` templates instantiated. If an IE system only finds a single `kidney` template, then a choice needs to be made as to which of the two gold standard templates it must be aligned with for evaluation.

Taken together, the problems we encountered meant that it was difficult to decide if evaluation scores reflected the system being evaluated, or some problem in the gold standard. The problems that we identified with our template model are in part inherent to the template representation, and in part due to the complexity of our specific template model. As originally used in the Message Understanding Conferences [31], templates are independent of the text: a product of research into full text understanding systems. Our simpler task is to extract those entities and relations explicitly mentioned in the text. This task is better served by a representation that anchors those entities and

relations directly to the text.

4.3 Developing the Guidelines

As a consequence of these difficulties, it was decided to create a new gold standard consisting of textually-anchored annotations, rather than templates. This would make evaluation easier, would simplify supervised learning using annotated text, and would also mean that one of the dedicated tools available for this style of annotation could be used. A larger number of documents would be annotated with a simplified set of entities and relations, and these would be described in explicit, methodically developed guidelines. The guidelines would be developed by a team of clinicians and computational linguists, and would be tested against a significant number of documents, before use for annotation of the final gold standard.

The starting points for the writing of the guidelines were the original ontology and template definitions. These were simplified to give an initial set of six entities and six relations, plus two modifiers (later additions changed this to the schema presented in this paper, as shown in Figure 2). The entities and relationships were agreed between a small group of computational linguists and clinicians. An initial draft set of guidelines describing the entities and relationships were then drawn up, and discussed by a larger group.

The guidelines were developed and refined using an iterative process, designed to ensure their consistency. This is shown in Figure 3. Two qualified clinicians annotated different sets of documents in 5 iterations (covering 31 documents in total). We measured the agreement between annotators according to a number

of metrics which are defined below in Section 4.5.2. Agreement for these iterations are shown in Table 5. As can be seen, agreement remains consistently high after the 5 iterations, after which very few amendments were required to the guidelines. Relation agreement does not appear so stable on iteration 5. Difference analysis showed that over half of the difference was due to a single, simple type of disagreement across a limited number of sentences in one document. One annotator had co-referred mentions with a plural or set that encompassed that mention. For example, "nail of the right thumb" has been co-referred with "all of the hand nails". Scoring without this document gave a much improved level of agreement.

During each development iteration, the clinician annotators made notes on the clarity of the guidelines, and on the relevance of the resulting annotations. At the end each iteration, a difference analysis was performed on the two sets of annotations, listing points of difference between the two annotators. The annotator notes and the difference analysis were fed into a post-iteration discussion, which informed a rewrite of the guidelines. Many of the changes consisted of either minor clarifications, or the addition of informative examples. Occasionally, major changes were made. For example, it had been intended to annotate any discussion of lymph node involvement. However, no examples were found in the development documents, and the few examples found in a larger selection of the entire CLEF corpus were difficult to interpret. In another example, it was thought that `Investigation` entities would always stand in a `has_finding` relations to an entity type of `Condition`. However, this proved false, and the schema was augmented with a new entity type of `Result`, when it was realised that not all cases could be annotated in this way.

4.4 The guidelines as a tool

The guidelines are written as a *wiki*: a set of hyperlinked web pages that can be edited and created by anyone who has access to them. Use of a wiki means that the guidelines can be edited, corrected and updated by a number of people involved in their writing. Although written in this way, the guidelines are provided to annotators as a read-only web site. Publication as a web site meant that the guidelines were dynamic, and hyperlinked. The dynamic nature of the site meant that as guidelines were updated, annotators would always be accessing the latest version. Pages of "news" were provided to publicise recent changes, and to answer common queries. Sample pages from the web site are shown in Figure 4.

The hyperlinked nature of the guidelines is in contrast to the more common method of presenting annotation guidelines as a technical document. Hyperlinking meant that annotators could quickly navigate them, finding the relevant section for their work, and could easily move to related sections. For example, an annotator thinking about how to annotate the `has_location` relation, could easily jump to the section about the `Locus` entity, an argument of that relation, via hyperlinks on every mention of `Locus` on the `has_location` pages. In addition to hyperlinks within pages, each page was provided with a top level menu bar, giving access to tables summarising the guidelines, and to the top level sections. Links for the next and previous page were also provided, so that the guidelines could be read in a linear style if required.

The idea of guidelines-as-a-tool is also reflected in the writing style. Writing is in an easily digested style with short sentences, heavy use of bullet points,

tables, examples, and sub-sections. The aim is to present the information clearly, and in a quickly accessible form. Annotators work with the guidelines open in a web browser, switching back and forth from the guidelines to their annotation tool. The guidelines comprise nine main sections:

(1) **News:** a section describing recent changes to the guidelines, answers to common questions, and other annotation related news items.
(2) **Terminology:** a table giving definitions and examples of the technical terms used in annotation, such as *Entity*, *Co-reference*.
(3) **Summary tables:** of entities, modifiers, and relations, each type with a description, examples, and hyperlinks to the relevant guidelines. Tables 2 and Tables 3 are adapted from these.
(4) **A *recipe*** for annotating: a step-by-step guide of how to read a document and mark the relevant annotations. This recipe was independent of the annotation tool used.
(5) **General guidelines:** that give a high-level philosophy of what should and should not be annotated.
(6) **Entity guidelines:** specific guidelines for each entity.
(7) **Relation guidelines:** specific guidelines for each relation.
(8) **Modifier guidelines:** specific guidelines for each modifier.
(9) **Report guidelines:** guidelines specific to histopathology and imaging reports

The annotation recipe describes in detail how a document should be annotated. It was expected that a consistent annotation method would produce more consistent annotations. In reality, however, it is difficult to supervise annotation, and so it is not clear whether annotators always adopted the recipe, or opted for faster shortcut methods of annotation. The recipe is summarised

below:

(1) Read the document through in its entirety, marking no annotations, to get an understanding.
(2) Read the document a second time, adding annotations for the mentions (including pronouns) of the entities.
(3) Go through each of the conditions, loci, and interventions, checking for modifiers, qualifications, and associated text that signify further annotations.
(4) Go through each of the mentions in turn, and check to see if it co-refers with any other mention.
(5) Go through each of the mentions in turn, and decide if any have relationships with other entities.
(6) Record any questions, uncertainties, ambiguities, tool bugs and issues.

The general guidelines give a high level philosophy of what should and should not be annotated. They discuss issues such as whether to annotate overlapping terms; how and when complex terms should be broken down into their component parts; how to treat conjunctions; whether annotator domain knowledge may be applied to infer relationships, or whether they should be clearly stated in the text.

Each entity, relationship, and modifier has a single web page detailing specific guidelines for that annotation. These pages have a consistent format. For entities, the page first lists the kinds of things that should be annotated as this entity type, each with an example. This is followed by the kinds of things that should not be annotated, again with examples. The next section describes how mentions of this entity type take part in complex phrases, and how they

are modified by other words. Other sections may follow, specific to the entity type. For relations, the possible arguments are first described, in tabular form. This is followed by further sections, discussing for example: when entities do and do not take part in this relation type; the use of clinical knowledge to infer relations; whether one-to-many relations are allowed for this relation type.

4.5 Annotation Methodology

The annotation methodology follows established natural language processing standards [32]. Annotators work to agreed guidelines; documents are annotated by at least two annotators; documents are only used where there is an acceptable level of agreement between annotators; differences are resolved by a third experienced annotator. These points are discussed further below.

4.5.1 Double Annotation

A singly annotated document can reflect many problems: the idiosyncrasies of an individual annotator; one-off errors made by a single annotator; annotators who consistently under-perform. There are many alternative annotation schemes designed to overcome this, all of which involve more annotator time. Double annotation is a widely used alternative, in which each document is independently annotated by two annotators, and the sets of annotations compared for agreement.

4.5.2 Agreement Metrics

Agreement between annotators is defined in terms of *matches* and *non-matches* between the two double annotation sets created for each document, one set created per annotator. An annotation in one set matches that in the other set if they have the same type, and the same character offsets (textual span). In all other cases, the annotation is considered a non-match. For every match in the first set, there will be an equivalent match in the second set. The total number of matches is the sum of these (i.e. double the number of matches in any one set). The total number of non-matches is the sum of non-matches in each set. Agreement between double annotated documents can then be calculated as inter annotator agreement (IAA), as in Equation 1.

$$IAA = \frac{\text{matches}}{\text{matches} + \text{non-matches}} \tag{1}$$

We report IAA as a percentage. Overall figures are macro-averaged across all entity or relationship types. In addition to the "strict" version of IAA described above, in which entity spans must match exactly, we use a second "lenient" IAA, in which partial matches, i.e. overlaps, are counted as a half match. Together, these show how much disagreement is down to annotators finding similar entities, but differing in the exact spans of text marked. We used both scores in development. Results given below explicitly state the score being used.

Two variations of IAA for relations were also used. First, all relationships found were scored. This has the drawback that an annotator who failed to find a relationship because they had not found one or both the entities would be penalized. To overcome this, a Corrected IAA (referred to as CIAA) was

calculated, including only those relationships where both annotators had found the two entities involved. This allows us to isolate, to some extent, relationship scoring from entity scoring.

In the initial stages of the annotation exercise, during guideline development, IAA was calculated directly with the Knowtator plugin for Protégé [30]. During the training of annotators and "production" annotation, we wished to have a more fine-grained control over IAA calculation, giving the different types of IAA scores for different combinations of annotators and parameters, and producing hyperlinked error reports. To this end, we customised our own ANNALIST scoring tool [33]. Unless otherwise stated, scores given in this paper have been calculated using ANNALIST.

The metrics used are equivalent to others more commonly used in IE evaluations, as shown in Table 4. IAA also approximates the widely used kappa score, which is itself not appropriate in this case [34].

4.5.3 Difference Resolution

Double annotation can be used to improve the quality of annotation, and therefore the quality of statistical models trained on those annotations. This is achieved by combining double annotations to give a set closer to the "truth" (although it is generally accepted as impossible to define an "absolute truth" gold standard in an annotation task with the complexity of CLEF's). The resolution process is carried out by a third experienced annotator, the *consensus* annotator. All agreements from the original annotators are accepted into a consensus set, and the third annotator adjudicates on differences, according to a set of strict consensus guidelines. These consensus guidelines are

designed to ensure that annotations remain at least double annotated, and that the consensus annotator cannot easily overrule both of the double annotators to enforce their own single annotation. The consensus annotator cannot, for example, create new annotations that have not been previously created by one of the double annotators, and cannot delete an annotation that has been created by both double annotators. Amongst other rules, the consensus annotation guidelines rule how to deal with overlapping annotations; how to deal with annotations of the same span but different type; and how to deal with different arguments for relationship annotations.

4.6 Annotating CUIs

As described in Section 4, the CLEF entity types map to high level types in the UMLS semantic network. This gives a coarse-grained semantic typing to entities, appropriate for most CLEF use cases. For one CLEF use case, however, a more fine grained typing was required over a small number of narratives, using UMLS concept identifiers (CUIs). We therefore assigned CUIs to all entity mentions in a portion of the narratives: 35 from the stratified random sample, and 5 from a single patient of the whole patient record.

It is not easy to assign CUIs fully automatically, as a term may be ambiguous, and relate to several concepts in the UMLS. The term "cold", for example, has a CUI associating it with the temperature, and a CUI associating it with the infection. The context in which a term is mentioned is therefore required to disambiguate the possible CUIs. We therefore adopted a semi-automated approach to CUI annotation, using the GATE language processing toolkit [35,36]. A custom GATE module took each entity mention in turn

from annotated gold standard documents. The mention was queried against the UMLS Knowledge Source Server API (UMLSKS API) [37], to fetch a list of possible CUIs for that mention, together with their UMLS semantic type, and a textual definition if available. The results were presented to a single human annotator, who examined them in the light of the mention's surrounding context. Where a single CUI had been automatically assigned, the annotator could either choose or reject that assignment. Where several CUIs were possible for a mention, the annotator could choose either one or none of the CUIs. In those cases where no suitable CUI had been automatically assigned, the annotator performed a more sophisticated manual search of the the UMLS via its web interface. The most suitable CUI found via the web interface was attached to the mention.

5 Analysis of the annotation process

This section presents some qualitative and quantitative results relating to the annotation process and guideline development.

5.1 Annotator Expertise

In order to examine how easily the guidelines could be applied by other annotators with varying levels of expertise, we also gave a batch of documents to the two clinicians who assisted in guideline development 4.3, another clinician, a biologist with some linguistics background, and a computational linguist. Each was given very limited training. The resultant annotations were compared with each other, and with a consensus set created from the two development anno-

tators. The IAA matrices for this group are shown in Table 6 for entities, and Table 7 for relations. It is interesting to note that both the biologist and the computational linguist achieve closer agreement with the consensus set, than does the clinician. A difference analysis suggested that the computational linguist was finding more pronominal co-references and verbally signaled relations than the clinician, but that unsurprisingly, the clinician found more relations requiring domain knowledge to resolve. A combination of both linguistic and life science knowledge appears to be best: of the three non-development annotators, the biologist with some linguistics background achieved the closest agreement with the consensus set.

This difference reflects a major issue in the development of the guidelines: the extent to which annotators should apply domain specific knowledge to their analysis. Much of clinical text can be understood, even if laboriously and simplistically, by a non-clinician armed with a medical dictionary. The basic meaning is exposed by the linguistic constructs of the text. Some relationships between entities in the text, however, require deeper understanding. For example, the condition for which a particular drug was given may be unclear to the non-clinician. In writing the guidelines, we decided that such relationships should be annotated, although this requirement is not easy to formulate as specific rules.

5.2 Different text sub-genres

The guidelines were mainly developed against clinical narratives. We were interested to see if the same guidelines could be applied to imaging and histopathology reports. We found that the guidelines could be quickly adapted

with minimal change, to give excellent IAA after only two iterations, as is shown in Table 8. Of those entities and relationships with an IAA below 75%, the majority reflect bias due to a small sample size. The fact that report IAA is better than clinical narrative IAA may reflect the greater regularity of the reports.

5.3 Annotation: Training and Consistency

In total, around 25 annotators were involved in guideline development and annotation. They included practicing clinicians, medical informaticians, and final year medical students. Each given an initial 2.5 hours of training.

After the initial training session, annotators were given two training batches to annotate, which comprised documents originally used in the debugging exercise, and for which consensus annotations had been created. IAA scores were computed between annotators, and against the consensus set. The results are shown for one group of annotators, in Table 9 for entities, and Table 10 for relationships. These figures allowed us to identify and offer remedial training to under-performing annotators and to refine the guidelines further.

The matrices allow us to look at two factors. First, the IAA between annotators and the consensus set gives us a measure of consistency between annotators and our notion of truth. For entities, the trainee annotators clearly agree with the consensus as closely as the expert annotators do. For relations, they do not agree so closely. Second, the matrices allow us to examine the internal consistency between trainee annotators. Are they applying the guidelines consistently, even if not in agreement with the consensus? The wide

range of relation IAA scores suggests that relationship annotation is inconsistent. Again, this may reflect the difficulty in applying highly domain-specific knowledge to relationships between entities.

5.4 Annotator difference analysis

During the initial guideline development process, we exhaustively examined differences between double annotators, and used the results of these analyses to both inform guideline writing, and to provide feedback to annotators. During the annotation of the final gold standard, a full analysis of all differences between the double annotations over the entire gold standard would be prohibitively time consuming, and so has not been carried out. Where documents showed poor agreement between the annotators, ad-hoc difference analysis was carried out to provide feedback and information for the consensus annotator. Most differences fell into a small number of categories. Some of these are described below, with examples from narratives given in Table 11.

(1) **Occurrence** A straightforward difference in which one annotator marked a span of text or a relation, and the other did not. Such an error could be due to a disagreement, or due to one annotator unintentionally missing something: reasons are not always clear.

(2) **Textual extent** The two annotators marked overlapping spans with the same entity type. They agreed that an annotation occurred, but disagreed on exactly what text should be marked.

(3) **Typing** The annotators agreed on annotating a specific extent of text, but assigned different entity types to that extent. Most commonly, there were confusions between `Intervention` and `Investigation`, and also

between `Condition` and `Result`.

(4) **Term decomposition** One annotator marked a span as a multi-word term, with a single annotation. The other annotator decomposed the term. This was most common with `Condition` and `Locus`. For example, should "lung cancer" be marked as a single `Condition`, or a `Condition` and `Locus`? Despite rigid guidelines on how to decompose terms (based on occurrence in a standard dictionary), differences still arose.

(5) **Granularity** Usually where one annotator marked a high level `Investigation` name and the other marked a nearby component part of that `Investigation`.

(6) **Term ambiguity** One annotator marked a span of text, but it was being used in a different sense to that implied by the annotation entity type.

(7) **Locus modification** `Locus` may be modified by both `Sub-location` and `Laterality` (e.g. "Right lobe of the lower pole of the thyroid"). This sometimes led to differences when annotating a complex anatomy expression.

(8) **Multiple compounding differences** Some examples show multiple differences that compound each other. Differences in the way in which a `Locus` and its modifiers are annotated can lead to differences in relationships, and so on.

5.5 Time taken to annotate

During the initial guideline development process, we timed the annotation of five narratives by a single annotator, in order to provide data for planning the main annotation process. The time to annotate these narratives had a range of 15 to 70 minutes, with a mean of 34 minutes. The wide range of times was

not a simple function of document length: the annotators have reported that some of the shortest documents have been some of the hardest to annotate, and vice versa. Although we did not measure time to annotate documents in the main annotation exercise, the mean time of our small sample was born out by anecdote, with annotators reporting around half an hour per narrative throughout the full annotation exercise.

It should also be remembered that each document was double annotated, and followed by a consensus annotation (15 minutes for this last step, by anecdote). Together with the time taken to process annotations, check IAA scores and so on, each document probably took around 1.5 hours to fully annotate. This excludes time taken for training, guideline and schema development, CUI annotation and time annotation.

6 Constructing the final corpus

Once guideline development and annotator training had been completed, annotators proceeded to double annotate the "production" corpus, consisting of the stratified random corpus and the whole patient corpus. Documents were annotated in batches of 5. On completion of a batch by two annotators, IAA was calculated for that batch. If IAA was not acceptable, then the batch was re-annotated by a further annotator. If IAA was acceptable, then the batch was put forward for consensus annotation. In the initial stages of the annotation exercise, an acceptable IAA was considered to be one that passed an arbitrary threshold of at least 65% lenient entity IAA, and at least 50% relation CIAA. As the annotation progressed, however, it became apparent that IAA could be skewed below these thresholds for one of two reasons. Firstly,

there were occasional "outlier" batches with very few relations, in which a small absolute number of disagreements could lead to poor IAA. Secondly, a single simple, obvious, and repeated, mistake on the part of one annotator, could also skew the IAA below the threshold. For example, one annotator completely omitted to annotate an obvious `Intervention` mentioned multiple times in one document, whereas the other annotator marked it. Given the expense of repeating annotation, it was therefore decided that low agreement on a particular double annotation batch should not mean that the batch was rejected, if these systematic errors could be corrected in the consensus annotation stage. Consensus annotation of batches with IAA below the threshold was therefore allowed where IAA had suffered in one of the above ways, and if the consensus annotator was confident of being able to correct the mistake.

Once consensus annotation had been completed, the consensus annotations were processed into two forms for use throughout the CLEF project, and beyond CLEF if we are able to make the corpus publicly available. First, the annotations were processed into XML files conforming to an XML schema embodying Figure 2, and incorporating attributes for character offsets, text of the mentions, and CUIs where appropriate. Second, the annotations were processed into GATE datastores, for use in training and evaluation of the CLEF IE system.

The final stratified random portion of the corpus is described in Tables 12 (narratives), 13 (histopathology reports), and 14 (imaging reports). Each table shows distribution of entities and relations across that document type. The tables also show the IAA between the double annotators, for each entity and relation type. Note that the final gold standard consists of a consensus of the double annotation, created by a third annotator. Systems trained and

evaluated with the gold standard use this consensus. The IAAs between double annotators that are given do not therefore provide an upper bound on system performance, but an indication of how hard a recognition task is.

The results illustrate that despite training and the use of extensive guidelines, clinically trained annotators are well below perfect agreement on single annotation tasks, such as finding all of the `Investigations` in a document. The results also illustrate that relation annotation is highly dependent on entity annotation, as would be expected. CIAA, corrected for entity recognition, is significantly higher than uncorrected IAA. It is apparent that the overall annotation of a document is hard. Annotators are asked to look for multiple, coarsely defined entities and complex relationships between them. Documents vary in their type, from simple letters to complex reports; they vary in the style of writing; in size; and in the pathophysiology being discussed.

7 Temporal Annotation

If the course of a patient's illness and treatment is to be modelled then the clinical entities and relationships found within text must be located in time so that they can be integrated with time-stamped information from the structured component of the patient record to construct a coherent history. To support this modelling the annotation scheme for clinical entities and relations specified above has been augmented to capture aspects of temporal information. In this section we describe the temporal annotation schema, the process of temporal annotation and the distribution of temporal annotations found in the portion of the corpus annotated so far.

Only a subset of the clinical entities identified above are 'event-like' and hence temporally situated. These are the CLEF investigations, interventions and conditions, which we refer to in the following as TLCs (Temporally Located CLEF entities). It is interesting to note that the clinical events that we wish to temporally locate are mostly expressed in clinical text by nouns and noun phrases, which contrasts with the predominant use of verbs to express events elsewhere. We observe that most occurrences of CLEF entities in these three categories correspond to events that we would hope to temporally anchor, the exceptions being a small proportion of uses that are generic and hence not temporally situated. The exclusion of other CLEF entity types, such as drugs and results, from the TLC class is not meant to imply that time considerations do not arise for the other CLEF entity types. For example, a drug might be prescribed or discontinued at a particular time, and a result produced by an investigation that is done at a particular time. But here the temporal involvement of the drug or result is a secondary consequence of its relation to the event which is temporally locatable. Directly anchoring a drug to a date, for example, has no clear meaning without also characterising the event, i.e. was the drug prescribed or discontinued on that day? We take such considerations to be a matter of broader temporal analysis, and instead here restrict our attention to just the CLEF entity types that can be directly temporally located.

The aim of the CLEF temporal gold standard is to capture temporal relations between TLCs and time expressions. Time expressions include dates and times (both absolute and relative), as well as durations, as specified in the TimeML

TIMEX3 standard [38]. Temporal relations are encoded as CTlink annotations which identify the TLCs and time expression related as well as specifying the relation type. Relation types include, for example, `before`, `after`, `overlap`, `includes`. For a full list see Table 15 or Figure 6. Our scheme requires annotation of only those temporal relations holding between TLCs and the date of the letter (Task A), and between TLCs and temporal expressions appearing in the same sentence (Task B). These tasks are similar to, but not identical with, those addressed by the TempEval challenge within SemEval 2007 [39]. The scheme is graphically depicted in Figure 6.

7.2 Annotation of Temporal Information

The temporal annotation scheme described in the previous section, which is still under development, has to date been used to annotate ten patient letters (narrative data) from the clinically-annotated corpus described above in Section 3. In time we intend to annotate all of the gold standard corpus.

Temporal annotation is done through a combination of manual and automatic methods. TLCs can be immediately identified from the clinical entity annotations already present in the letters. Temporal expressions are annotated and normalized to ISO dates by the GUTime tagger [40], which annotates in accordance with the TIMEX3 standard. This annotation is manually checked and corrected as necessary. After these automatic steps, we manually annotate the temporal relations holding between TLCs and the date of the letter (Task A), and between TLCs and temporal expressions appearing in the same sentence (Task B).

The distribution of annotations for the different subtypes of CTLinks, TLCs and time expressions for the ten development documents annotated so far are shown in Tables 15 and 16. Note that some TLCs are marked as hypothetical. For example in *no palliative chemotherapy or radiotherapy would be appropriate* the terms *chemotherapy* and *radiotherapy* are marked as TLCs but clearly have no 'occurrence' that can be located in time and hence will not participate in any CTLinks.

8 Using the Corpus: the CLEF IE system

The CLEF corpus has been created to enable the training and evaluation of the CLEF IE system, which can be applied to previously unseen clinical texts, to automatically extract the entities, modifiers and relationships that the annotation schema describes. This system has been built using the GATE NLP toolkit [35,36], which allows language processing applications to be constructed as a pipeline of processing components. Documents are passed down the pipeline being analysed by each component in turn, with the results of this analysis being available to later components. The CLEF IE pipeline is outlined in Figure 5, with separate pipelines being shown for training and application of the system (although the two pipelines substantially overlap). In either case, the pipeline has three main parts:

Linguistic preprocessing: Firstly, the text of each document is split into tokens (such as words, numbers and punctuation) and sentences, and then

part of speech (POS) information is added.

Dictionary-based term look-up: Next, medically significant terms are identified, using a dictionary-based look-up approach. This is done using Termino: a large-scale terminological resource designed specifically for text processing [41]. Termino consists of two parts. The first is a database constructed from existing terminology resources. Termino provides uniform access to these resources, and links from recognised terms back to resource entries. The second part consists of finite state recognisers compiled from terms in the database. Our principle terminology source in CLEF is the Unified Medical Language System (UMLS) [42], which is the largest source of medical vocabulary, and which links terms to other information, such as semantic types.

Statistical recognition of entities and relations: we treat the recognition of both entities and relations as classification tasks, using Support Vector Machines (SVMs) as trainable classifiers, as they have proven to be effective for a range of NLP tasks. We use an SVM implementation provided as part of the GATE toolkit. We will discuss the recognition of entities and relations separately in turn.

8.1 CLEF entity recognition

SVMs are binary classifiers, and so separate classifiers must be trained to recognise the different entity types. Furthermore, our classifiers apply to individual tokens, and so multi-token entities are recognised using a BE (Begin/End) style of boundary learning. This is handled by the GATE Learning API [43]. A pair of binary classifiers are trained for each entity type: one for

the begin (B) token, and one for the end (E) token. For our five entity types, ten binary classifiers are therefore built, and each is applied independently of the others. A post-processing step is required to combine pairs of B and E tokens, to find the boundaries of candidate entities, and to adjudicate between conflicting (i.e. overlapping) candidates.

The features used to classify each token are based on the token itself, and the token on either side of it. Features include the morphological root and affix (for words), a generalisation of the POS, token type (e.g. word, number) and orthographic type (e.g. upper/lower case). So that dictionary look up can contribute to entity recognition, a further feature indicates whether the token is part of term recognised by Termino, taking the term's type as it value if it is, and the value `null` otherwise.

The recognition performance of this system is shown by the results in Table 17, which were computed over the 77 clinical narrative documents of the CLEF corpus, using ten-fold cross-validation. Scores are provided for the standard metrics of Precision (P), Recall (R) and F-measure (F1), with scores macro-averaged across the ten folds. As an indicator of the difficulty of each entity recognition task, the table also provides Inter Annotator Agreement (IAA) scores for the two independent annotators (but note that the system is trained on a third *consensus* annotation). Observe that the overall F1 performance of this system falls only 3% behind that of the overall averaged IAA.

The use of Termino dictionary lookup as a feature in a supervised statistical entity recognition system is an attempt to address two major challenges in entity recognition. Firstly, pure dictionary lookup can give poor precision, due to term ambiguity with general language ("I", for example, is both a pronoun

and an abbreviation for Iodine). Secondly, supervised statistical techniques are restricted to a model based only on those entities found in the training data. Although we have not performed a proper error analysis of our results, inspection reveals that both types of errors still occur, even if at a reduced rate. In addition, we cannot rule out errors due to e.g. incorrect POS tagging and morphological analysis. A more detailed account of our entity recognition approach has been published [44].

8.2 CLEF relation recognition

Relation extraction is treated as a classification task by taking a set of entity pairs that *might* be related and requiring the system to assign to each one of the relationship types, or the type `null` to indicate that no relation holds. The set of candidate pairs to be considered is restricted firstly by allowing only pairs whose types can be linked by some relation (e.g. no CLEF relation can link `Drug-or-device` and `Result` entities, so no such pairs are created), and secondly by only pairing entities that are no more than n sentences apart (we here allow only pairs for entities in the same or adjacent sentences). For classifier training, this set of candidate pairs is computed, and those for which a relation is asserted in the gold standard are assigned that relation type as class, and all others the class `null`. These pairs constitute the instances for which the classifier model is built. In classifier application, the corresponding set of entity pairs are computed for an unseen text (after entity extraction has been done) and the model applied to determine which pairs are related and how. As with entity recognition, we use an SVM implementation available in GATE, and use the GATE Learning API to handle the task of recasting

this multi-class classification task as a combination of binary classifiers, with a post-processing step to reconcile conflicts.

We have explored using a range of different features sets with these classifiers, including features such as the surface string, morphological root and POS of the tokens of the two entities and of the n tokens appearing to either side of the entities. Other features include the types of the two entities, their linear order (i.e. which appears first), and the distance between them (measured as number of sentence boundaries). This feature exploration and the resultant optimally performing feature set are fully described in [45]. We used the optimally performing feature set with the system to produce the relation extraction results shown in Table 18, which were again computed over the 77 clinical narrative documents of the CLEF corpus, using ten-fold cross-validation, with macro-averaging of scores across the ten folds. Note that the entities provided as input to relation extraction are those of the gold standard corpus, rather than the result of automatic entity recognition, so that we can see the performance of relation extraction in isolation from the damaging effects of errorful input. To give an indication of the difficulty of relation extraction, the table includes scores for agreement between the two independent annotators analysing texts, but these are *corrected* IAA, i.e. they compare only the relationships for which both of the related entities have been found by *both* annotators. Observe that the overall system F1 is 70%, compared to a CIAA of 75%. A more detailed account of our relation extraction approach has been published [45].

9 Discussion and Conclusions

We have described the CLEF corpus: a semantically annotated corpus designed to support the training and evaluation of information extraction systems developed to extract information of clinical significance from free text clinic notes, imaging reports and histopathology reports. We have described the design of the annotated corpus, including the number of texts it contains, the principles by which they were selected from a large body of unannotated texts and the annotation schema according to which clinical and temporal entities and relations of significance have been annotated in the texts. We also described the annotation process that was undertaken with a view to ensuring, as far as is possible given constraints of time and money, the quality and consistency of the annotation, and we have reported results of inter-annotator agreement, which show that promising levels of inter-annotator agreement can be achieved. We have examined the applicability of annotation guidelines to several clinical text types, and our results suggest that guidelines developed for one type may be fruitfully applied to others. We have also reported the distribution of entity and relation types, both clinical and temporal, across the corpus, giving a sense of how well represented each entity and relation type is in the corpus.

We believe the CLEF corpus makes a significant contribution to research on clinical language processing both in terms of the resource produced and the methodology adopted to develop this resource. Nonetheless there are limitations both to the resulting resource and to the methodology.

Regarding the resulting resource, we must consider the size of the resource, and

the quality of annotation. The size of the corpus is a straightforward function of the available annotator time. Quality of annotation will reflect both the consistency and completeness of the guidelines, and the correct application of those guidelines by annotators. The former could be improved by investing more time in iterative development and debugging of the guidelines. The latter could be improved by additional annotation steps. As with any annotated corpus, annotation quality will to some extent reflect the overriding expense of annotator time. Anything that reduces the burden on annotators, may be expected to improve both quality and the size of the final corpus. Techniques that might reduce this burden are discussed below.

Regarding the corpus development methodology, the most obvious limitation is that such efforts require a lot of annotator labour and that annotators find the work hard. Since the annotation requires specialist medical knowledge the pool of possible annotators is relatively small. Furthermore we found the recruitment, training and co-ordination of annotators at different sites working on sensitive data to be logistically complex, also requiring significant effort. Because the work was difficult a number of annotators resigned after a limited contribution forcing us into an iterative cycle of recruitment and training.

Various steps could be taken to address these difficulties in future annotation exercises. To attempt to utilize annotator effort most effectively, so-called active learning or mixed initiative approaches could be explored ([46,47]). In these approaches annotation and system learning stages are interleaved so that at any point an annotator is correcting and augmenting annotations that the system has added to a document rather than annotating a document from scratch. As the system learns, the amount of human annotator input per annotated document should go down and human effort should be concentrated on

difficult cases, i.e. ones the system has missed or annotated incorrectly. Thus more annotated text should result from equivalent annotator effort when using active learning as compared with not using it.

To address the difficulty of the task, one approach is simply to reduce the scope of the annotation scheme and to focus on fewer entities or relations. This may or may not be possible depending on the intended application. Another approach, and one which could also help with the logistical difficulties, is to move to a distributed, collaborative annotation framework in which the grain size of annotation instances is reduced to a snippet, e.g., a single sentence. A number of such collaborative annotation tools are emerging – see, e.g. [48,49]. Such an approach has numerous advantages: the annotation effort can be distributed globally, drawing on interested parties anywhere; smaller annotation grain size reduces the unit of useful annotation meaning smaller levels of effort can be exploited, reduce the difficulty for annotators by focusing effort on single decision types over small snippets of text; annotation of individual instances can be repeated until a satisfactory level of agreement is reached, or the instance is eliminated as problematic; rogue or poor quality annotators can be identified and their annotations removed. There are, however, non-trivial obstacles to using such a methodology in our domain, including the need to protect patient confidentiality, and the fact that some of the inter-sentential relations annotated in our corpus would be excluded if only snippets of text were presented to annotators.

These considerations all point to ways in which the difficulties we have encountered in our annotation effort could be mitigated in future annotation projects. Nonetheless, despite these difficulties, the annotated CLEF corpus is the richest resource of semantically marked up clinical text yet created,

one which we hope will be of wide-ranging interest and utility to the clinical language processing research community.

Availability

The current availability of all of the resources in this paper is described on the project web site [50], together with links to each available resource. Most of the software, including the ANNALIST scoring tool, is available for download, as is the final version of the guidelines.

At the time of publication, there is some limited availability of the CLEF gold standard. We are able to share small samples of data from the gold standard, which may include short extracts of documents. In order to ensure anonymity, such releases go through a triple manual inspection, by an ethicist, a clinician, and a confidentiality expert. Full release of the whole gold standard will be made on the project web site [50], after approval by a UK Multi-centre Research Ethics Committee.

Acknowledgements

This research was supported by UK Medical Research Council grant number RB106367, "CLEF Services". We would like to thank the Royal Marsden Hospital for providing the corpus; our annotators at the University of Manchester and University College London; and members of CLEF Services who have helped with clinical expertise and logistics, particularly Jay Kola, Bill Wheeldin, James Cunningham, and Colin Puleston (all at the University of

Manchester); and Dipak Kalra, Archana Tapuria, and Nathan Lea (all at University College London).

References

[1] Rector A, Rogers J, Taweel A, Ingram D, Kalra D, Milan J, et al. CLEF — joining up healthcare with clinical and post-genomic research. In: Proceedings of UK e-Science All Hands Meeting 2003. Nottingham, UK; 2003. p. 264–267.

[2] Grishman R. Information Extraction. In: Mitkov R, editor. The Oxford Handbook of Computational Linguistics; 2003. Chapter 30.

[3] Harkema H, Roberts I, Gaizauskas R, Hepple M. Information Extraction from Clinical Records. In: Cox SJ, Walker DW, editors. Proceedings of the UK e-Science All Hands Meeting 2005. Nottingham, UK; 2005. p. 254–258.

[4] Riloff E. Automatically Generating Extraction Patterns from Untagged Text. In: AAAI/IAAI, Vol. 2; 1996. p. 1044–1049.

[5] Roberts A, Gaizauskas R, Hepple M, Davis N, Demetriou G, Guo Y, et al. The CLEF Corpus: Semantic Annotation of Clinical Text. In: Proc AMIA Symp. Chicago, IL, USA; 2007. p. 625–629.

[6] Roberts A, Gaizauskas R, Hepple M, Demetriou G, Guo Y, Setzer A, et al. Semantic Annotation of Clinical Text: The CLEF Corpus. In: Proceedings of Building and evaluating resources for biomedical text mining: workshop at Sixth International Conference on Language Resources and Evaluation, LREC 2008. Marrakech, Morocco: ELRA; 2008. .

[7] Kim JD, Ohta T, Tateisi Y, Tsujii J. GENIA corpus — a semantically annotated corpus for bio-textmining. Bioinformatics. 2003;19(1):i180–i182.

[8] Kim JD, Ohta T, Tsujii J. Corpus annotation for mining biomedical events from literature. BMC Bioinformatics. 2008;9(1).

[9] Kulick S, Bies A, Liberman M, Mandel M, McDonald R, Palmer M, et al. Integrated Annotation for Biomedical Information Extraction. In: Hirschman L, Pustejovsky J, editors. HLT-NAACL 2004 Workshop: BioLINK 2004, Linking Biological Literature, Ontologies and Databases. Boston, Massachusetts, USA: Association for Computational Linguistics; 2004. p. 61–68.

[10] Franzén K, Gunnar, Eriksson, Olsson F, Asker L, Lidén P, et al. Protein names and how to find them. Int J Med Inform. 2002;67(1–3):49–61.

[11] Rosario B, Hearst MA. Classifying semantic relations in bioscience texts. In: Proceedings of the 42nd Annual Meeting on Association for Computational Linguistics. Morristown, NJ, USA: Association for Computational Linguistics; 2004. p. 430.

[12] Rosario B, Hearst MA. Multi-way relation classification: application to protein-protein interactions. In: Proceedings of the conference on Human Language Technology and Empirical Methods in Natural Language Processing. Morristown, NJ, USA: Association for Computational Linguistics; 2005. p. 732–739.

[13] Alex B, Grover C, Haddow B, Kabadjov M, Klein E, Matthews M, et al. The ITI TXM Corpora: Tissue Expressions and Protein-Protein Interactions. In: Proceedings of Building and evaluating resources for biomedical text mining: Workshop at Sixth International Conference on Language Resources and Evaluation, LREC 2008. Marrakech, Morocco; 2008. p. 11–18. In press.

[14] Tanabe L, Xie N, Thom LH, Matten W, Wilbur WJ. GENETAG: a tagged corpus for gene/protein named entity recognition. BMC Bioinformatics. 2005;6(Suppl 1)(S3).

[15] Nédellec C. Learning Language in Logic - Genic Interaction Extraction Challenge. In: Proceedings of the ICML05 Workshop on Learning Language in Logic. Bonn, Germany; 2005. p. 31–37.

[16] TREC Genomics Track. [cited 6 June 2008]; Available from http://ir.ohsu.edu/genomics;.

[17] Pestian JP, Brew C, Matykiewicz P, Hovermale D, Johnson N, Cohen KB, et al. A shared task involving multi-label classification of clinical free text. In: Biological, translational, and clinical language processing. Prague, Czech Republic: Association for Computational Linguistics; 2007. p. 97–104.

[18] Hersh WR, Muller H, Jensen JR, Yang J, Gorman PN, Ruch P. Advancing Biomedical Image Retrieval: Development and Analysis of a Test Collection. J Am Med Inform Assoc. 2006;13(5):488–496.

[19] Mller H, Deselaers T, Lehmann TM, Clough PD, Hersh W. Overview of the ImageCLEFmed 2006 medical retrieval and annotation tasks. In: Cross Language Evaluation Forum (CLEF) Workshop 2006. vol. 4730. Alicante, Spain: Springer; 2007. p. 595–608.

[20] i2b2 NLP shared task. [cited 6 June 2008]; Available from http://ir.ohsu.edu/genomics/;.

[21] Ogren PV, Savova G, Buntrock JD, Chute CG. Building and Evaluating Annotated Corpora for Medical NLP Systems. In: Proc AMIA Symp; 2006. p. 1050.

[22] Meystre S, Haug PJ. Natural language processing to extract medical problems from electronic clinical documents: Performance evaluation. Journal of Biomedical Informatics. 2006;39(6):589–599.

[23] Denny JC, Smithers JD, Miller RA, Spickard A. "Understanding" Medical School Curriculum Content Using KnowledgeMap. Journal of the American

Medical Informatics Association. 2003;10(4):351–362.

[24] Elkin PL, Brown SH, Bauer BA, Husser CS, Carruth W, Bergstrom LR, et al. A controlled trial of automated classification of negation from clinical notes. BMC Medical Informatics and Decision Making. 2005;5(13).

[25] Friedman C, Hripcsak G. Evaluating natural language processors in the clinical domain. Methods of Information in Medicine. 1998;37(4-5):334–44.

[26] International Classification of Diseases (ICD). [cited 6 June 2008]; Available from http://www.who.int/classifications/icd;.

[27] Rogers J, Puleston C, Rector A. The CLEF Chronicle: Patient Histories Derived from Electronic Health Records. Data Engineering Workshops, 2006 Proceedings 22nd International Conference on. 2006;p. x109–x109.

[28] Hallett C, Power R, Scott D. Summarisation and Visualisation of e-Health Data Repositories. In: Proceedings of the UK e-Science All Hands Meeting. Nottingham, UK; 2006. p. 69–77.

[29] Gennari JH, Musen MA, Fergerson RW, Grosso WE, Crubézy M, Eriksson H, et al. The evolution of Protégé: an environment for knowledge-based systems development. International Journal Human-Computer Studies. 2003;58(1):89–123.

[30] Ogren PV. Knowtator: a Protégé plug-in for annotated corpus construction. In: Proceedings of the 2006 Conference of the North American Chapter of the Association for Computational Linguistics on Human Language Technology. Morristown, NJ, USA: Association for Computational Linguistics; 2006. p. 273–275.

[31] Defense Advanced Research Projects Agency. Proceedings of the Seventh Message Understanding Conference (MUC-7); 1998. Available at http://www.itl.nist.gov/iaui/894.02/related_projects/muc/.

[32] Boisen S, Crystal MR, Schwartz R, Stone R, Weischedel R. Annotating resources for information extraction. In: Proceedings of the Second Language Resources and Evaluation, LREC 2000; 2000. p. 1211–1214.

[33] Demetriou G, Gaizauskas R, Sun H, Roberts A. ANNALIST – ANNotation ALIgnment and Scoring Tool. In: Proceedings of the Sixth International Conference on Language Resources and Evaluation, LREC 2008. Marrakech, Morocco: ELRA; 2008. In press.

[34] Hripcsak G, Rothschild A. Agreement, F-measure and reliability in information retrieval. J Am Med Inform Assoc. 2005 May-June;12(3):296–298.

[35] Cunningham H, Maynard D, Bontcheva K, Tablan V. GATE: A Framework and Graphical Development Environment for Robust NLP Tools and Applications. In: Proceedings of the 40th Anniversary Meeting of the Association for Computational Linguistics. Philadelphia, PA, USA; 2002. p. 168–175.

[36] GATE – General Architecture for Text Engineering. [cited 6 June 2008]; Available from http://gate.ac.uk;.

[37] UMLS Knowledge Sources, 2007AB; 2007.

[38] Pustejovsky J, no JC, Ingria R, Saurí R, Gaizauskas R, Setzer A, et al. TimeML: Robust Specification of Event and Temporal Expressions in Text. In: Proceedings of the Fifth International Workshop on Computational Semantics (IWCS-5). Tilburg; 2003. .

[39] Verhagen M, Gaizauskas R, Schilder F, Hepple M, Katz G, Pustejovsky J. SemEval-2007 Task 15: TempEval Temporal Relation Identification. In: Proceedings of the 4th International Workshop on Semantic Evaluations. Prague; 2007. p. 75–80.

[40] Mani I, Wilson G. Robust temporal processing of news. In: Proceedings of the 38th Annual Meeting of the Association for Computational Linguistics (ACL

2000). New Brunswick, New Jersey; 2000. p. 69–76.

[41] Harkema H, Gaizauskas R, Hepple M, Davis N, Guo Y, Roberts A, et al. A Large-Scale Resource for Storing and Recognizing Technical Terminology. In: Proceedings of 4th International Conference on Language Resources and Evaluation. Lisbon, Portugal; 2004. p. 83–86.

[42] Lindberg D, Humphreys B, McCray A. The Unified Medical Language System. Methods Inf Med. 1993;32(4):281–291.

[43] Li Y, Bontcheva K, Cunningham H. SVM Based Learning System for Information Extraction. In: Deterministic and statistical methods in machine learning: first international workshop. No. 3635 in Lecture Notes in Computer Science. Springer; 2005. p. 319–339.

[44] Roberts A, Gaizauskas R, Hepple M, Guo Y. Combining terminology resources and statistical methods for entity recognition: an evaluation. In: Proceedings of the Sixth International Conference on Language Resources and Evaluation, LREC 2008. Marrakech, Morocco; 2008. .

[45] Roberts A, Gaizauskas R, Hepple M. Extracting Clinical Relationships from Patient Narratives. In: Proceedings of the Workshop on BioNLP 2008. Columbus, OH, USA: Association for Computational Linguistics; 2008. .

[46] Thompson CA, Califf ME, Mooney RJ. Active learning for natural language parsing and information extraction. In: Proc. 16th International Conf. on Machine Learning. Morgan Kaufmann, San Francisco, CA; 1999. p. 406–414.

[47] Ghani R, Jones R, Mitchell T, Riloff E. Active Learning For Information Extraction With Multiple View Feature Sets. In: Proceedings of the 20th International Conference on Machine Learning (ICML 2003) Workshop on Adaptive Text Extraction and Mining; 2003. .

[48] SAFE, the Semantic Annotation Factory Environment. [cited 2 October 2008]; Available from http://gate.ac.uk/safe/;.

[49] BioNotate. [cited 2 October 2008]; Available from http://sourceforge.net/projects/bionotate/;.

[50] Clinical E-Science Framework: Sheffield NLP. [cited 2 October 2008]; Available from http://nlp.shef.ac.uk/clef/;.

Document		Diagnosis						Total
Type	**Subtype**	**Digest**	**Breast**	**Haemat.**	**Resp.**	**Female genital**	**Male genital**	
Narrative	**To GP**	9.41	12.36	11.59	5.63	4.64	4.91	48.56
	Discharge	7.08	2.74	1.75	2.27	2.63	0.52	16.98
	Case note	4.25	2.95	2.07	1.96	2.41	1.07	14.72
	Other letter	1.92	1.57	1.30	0.76	0.83	0.50	6.88
	To consultant	1.31	2.04	0.75	0.80	0.61	0.25	5.77
	To referer	1.50	0.40	0.32	0.65	0.37	0.32	3.56
	To patient	0.57	0.95	0.21	0.25	0.33	0.30	2.60
	Report	0.15	0.20	0.14	0.11	0.11	0.02	0.72
	Audit	0.01	0.18	0.00	0.01	0.00	0.00	0.21
Narratives total		26.21	23.38	18.13	12.45	11.94	7.89	100.00
Imaging	**CT scan**	10.00	3.58	3.99	3.45	4.84	1.64	27.51
	Mammogram	0.02	1.03	0.03	0.02	0.02	0.00	1.11
	MRI	0.51	0.82	0.45	0.32	0.16	0.62	2.88
	Ultrasound	1.81	3.76	1.28	0.60	1.30	0.48	9.24
	X-ray	11.64	13.35	15.30	9.82	5.38	3.78	59.27
Imaging total		23.98	22.54	21.04	14.22	11.70	6.51	100.00
Histopathology (all)		22.74	18.48	28.94	6.49	15.9	7.44	100.00

Table 1
Percentage of all CLEF documents by diagnosis and document sub-type

Entity type	Description	Examples
Condition	Symptom, diagnosis, complication, conditions, problems, functions and processes, injury.	• This patient has had a lymph node biopsy which shows melanoma in his right groin. • It is clearly secondaries from the melanoma on his right second toe.
Intervention	Action performed by doctor or other clinician targeted at a patient, **Locus**, or **Condition** with the objective of changing (the properties) of, or treating, a **Condition**.	• Although his PET scan is normal he does need a groin dissection. • We agreed to treat with DTIC, and then consider radiotherapy.
Investigation	Interaction between doctor and patient or **Locus** aimed at measuring or studying, but not changing, some aspect of a **Condition**. **Investigations** have findings or interpretations, whereas **Interventions** usually do not.	• This patient has had a lymph node biopsy . . . • Although his PET scan is normal he does need a groin dissection. • We will perform a CT scan to look at the left pelvic side wall . . .
Result	The numeric or qualitative finding of an **Investigation**, excluding **Condition**.	• Although his PET scan is normal . . . • Other examples include the numeric values of tests, such as "80mg".
Drug or device	Usually a drug. Occasionally, medical devices such as suture material and drains will also be mentioned in texts.	• This pain was initially relieved by co-codamol.
Locus	Anatomical structure or location, body substance, or physiologic function, typically the locus of a **Condition**.	• This patient has had a lymph node biopsy which shows melanoma in his right groin . . . • It is clearly secondaries from the melanoma on his right second toe. • Although his PET scan is normal he does need a groin dissection. • We will perform a CT scan to look at the left pelvic side wall.

Table 2
CLEF entities. In the examples, mentions of the entity type are underlined. Adapted from the CLEF Annotation Guidelines (see Availability).

Relation type	First argument type	Second argument type	Description	Examples
has_target	**Investigation Intervention**	**Locus**	Relates an intervention or an investigation to the bodily locus at which it is targetted.	• This patient has had a [arg2] lymph node [arg1] biopsy • . . . he does need a [arg2] groin [arg1] dissection
has_finding	**Investigation**	**Condition Result**	Relates a condition to an investigation that demonstrated its presence, or a result to the investigation that produced that result.	• This patient has had a lymph node [arg1] biopsy which shows [arg2] melanoma • Although his [arg1] PET scan is [arg2] normal . . .
has_indication	**Drug or device Investigation Intervention**	**Condition**	Relates a condition to a drug, intervention, or investigation that is targetted at that condition.	• Her facial [arg2] pain was initially relieved by [arg1] co-codamol
has_location	**Condition**	**Locus**	Relationship between a condition and a locus: describes the bodily location of a specific condition. May also describe the location of malignant disease in lymph nodes, relating an involvement to a locus.	• . . . a biopsy which shows [arg1] melanoma in his right [arg2] groin • It is clearly secondaries from the [arg1] melanoma on his right [arg2] second toe • Her[arg2] facial [arg1] pain was initially relieved by co-codamol
Modifies	**Negation signal**	**Condition**	Relates a condition to its negation or uncertainty about it.	• There was [arg1] no evidence of extra pelvic [arg2] secondaries
Modifies	**Laterality signal**	**Locus Intervention**	Relates a bodily locus or intervention to its sidedness: *right, left, bilateral.*	• . . . on his [arg1] right [arg2] second toe • [arg1] right [arg2] thoracotomy
Modifies	**Sub-location signal**	**Locus**	Relates a bodily locus to other information about the location: *upper, lower, extra*, etc.	• [arg1] extra [arg2] pelvic
Co-refers	**Any**	**Any**	Relates two spans of text where they refer to the same entity in the real world. Includes both lexical co-reference and co-reference that requires domain knowledge, as in the examples.	• [arg1] Haemoglobin 7.5g/dl. Given this [arg1] Hb, treatment was postponed. • He has a [arg1] melanoma. The [arg1] tumour is in his 2nd toe.

Table 3
CLEF relations, modifiers, and co-reference. Each example shows a single relation of the given type. Arguments are underlined and preceded by their argument number. Adapted from the CLEF Annotation Guidelines (see Availability).

Agreement metric	**IE evaluation metric**
Match	2 × correct
Non-match	Spurious + missing
IAA	F1 measure

Table 4

Equivalence of annotator agreement metrics and standard IE metrics

		Debug iteration				
		1	**2**	**3**	**4**	**5**
Entities	**Matches**	244	244	308	462	276
	Partial matches	2	6	22	6	1
	Non-matches	45	32	93	51	22
	IAA	**84**	**87**	**74**	**89**	**92**
Relations	**Matches**	170	78	116	412	170
	Partial matches	3	5	14	6	1
	Non-matches	31	60	89	131	103
	IAA	**84**	**56**	**56**	**75**	**62**

Table 5

Lenient inter annotator agreement (IAA, %) for each guideline development iteration of five documents. During development, IAAs were calculated using the Knowtator annotation tool.

D2	77 *(72)*				
C	67 *(60)*	68 *(62)*			
B	76 *(70)*	80 *(74)*	69 *(64)*		
L	67 *(62)*	73 *(66)*	60 *(53)*	69 *(62)*	
Consensus	**85** ***(82)***	**89** ***(86)***	**68** ***(61)***	**78** ***(72)***	**73** ***(68)***
	D1	D2	C	B	L

Table 6

Entity agreement by annotators by expertise, over five documents. Lenient IAA, with strict IAA in italics and parentheses, both as %. D1 and D2: development annotators; C: clinician; B: biologist with linguistics background; L: computational linguist.

D2	63 *(45)*				
C	51 *(35)*	57 *(37)*			
B	56 *(41)*	57 *(43)*	63 *(40)*		
L	57 *(36)*	62 *(42)*	49 *(27)*	51 *(33)*	
Consensus	**87** ***(74)***	**74** ***(66)***	**50** ***(34)***	**55** ***(40)***	**56** ***(36)***
	D1	D2	C	B	L

Table 7

Relation agreement by annotators by expertise, over five documents. Corrected IAA, with uncorrected IAA in italics and parentheses, both as %. D1 and D2: development annotators; C: clinician; B: biologist with linguistics background; L: computational linguist.

		Narratives	Imaging	Histopath.
Iterations		5	2	2
Entities	**Condition**	91	100	92
	Intervention	82	100	n/a
	Investigation	97	75	95
	Result	100	20	80
	Drug or device	83	100	n/a
	Locus	94	97	92
	Negation signal	100	93	64
	Laterality signal	100	83	100
	Sub-location signal	100	67	50
	All	**92**	**90**	**88**
Relations	**has_target**	83	96	70
	has_finding	86	0	63
	has_indication	44	0	0
	has_location	66	90	81
	modifies (Negation)	100	100	91
	modifies (Laterality)	100	82	95
	modifies (Sub-location)	100	75	100
	corefers	52	92	67
	All	**62**	**84**	**70**

Table 8

Lenient IAA (entities) and corrected IAA (relations), both as %, on different document types. IAA was measured after the given number of guideline development iterations, with each iteration consisting of five documents. n/a means that there were no entities or relations for that type

D2	77 *(73)*								
1	76 *(70)*	79 *(71)*							
2	76 *(73)*	81 *(76)*	79 *(73)*						
3	76 *(72)*	83 *(78)*	89 *(86)*	82 *(77)*					
4	75 *(70)*	84 *(79)*	83 *(78)*	81 *(80)*	85 *(82)*				
5	76 *(62)*	84 *(79)*	71 *(62)*	88 *(66)*	80 *(53)*	78 *(62)*			
6	78 *(75)*	84 *(77)*	89 *(86)*	84 *(81)*	95 *(94)*	87 *(84)*	82 *(78)*		
7	79 *(75)*	81 *(75)*	81 *(75)*	83 *(79)*	86 *(83)*	82 *(79)*	82 *(79)*	88 *(84)*	
C	**85** ***(82)***	**89** ***(86)***	**84** ***(80)***	**84** ***(80)***	**88** ***(86)***	**85** ***(81)***	**83** ***(80)***	**91** ***(87)***	**87** ***(85)***
	D1	D2	1	2	3	4	5	6	7

Table 9

Lenient IAA (strict IAA in italics and parentheses)(%) for entities in five documents, between 7 trainee annotators, two expert development annotators (D1 and D2) and a consensus C created from D1 and D2.

D2	63 *(45)*								
1	54 *(42)*	44 *(36)*							
2	55 *(39)*	44 *(35)*	41 *(32)*						
3	65 *(48)*	59 *(48)*	60 *(53)*	49 *(39)*					
4	74 *(58)*	64 *(54)*	54 *(45)*	59 *(44)*	62 *(53)*				
5	66 *(41)*	48 *(37)*	43 *(31)*	47 *(40)*	54 *(41)*	54 *(35)*			
6	56 *(41)*	51 *(44)*	50 *(46)*	54 *(44)*	66 *(62)*	56 *(49)*	46 *(35)*		
7	69 *(52)*	54 *(43)*	52 *(43)*	52 *(41)*	59 *(52)*	61 *(48)*	64 *(50)*	57 *(50)*	
C	**87** ***(74)***	**74** ***(66)***	**52** ***(46)***	**52** ***(42)***	**61** ***(54)***	**68** ***(59)***	**57** ***(44)***	**61** ***(56)***	**71** ***(61)***
	D1	D2	1	2	3	4	5	6	7

Table 10

Corrected IAA (uncorrected IAA in italics and parentheses)(%) for relations in five documents, between 7 trainee annotators, two expert development annotators (D1 and D2) and a consensus C created from D1 and D2.

Text	Annotator 1 response	Annotator 2 response	Type of difference
no evidence of disseminated disease	disease[condition]	disseminated disease[condition]	Textual extent
tumour markers demonstrate CA125 306	CA125[investigation] has_result 306[result]	tumour markers[investigation] has_result CA125 306[result]	Textual extent; granularity
emergency admission with acute renal failure	acute renal failure[condition]	acute[condition] and failure[condition], both has_location renal	Term decomposition (Annotator 2 may have meant an acute failure has_location kidney)
I will continue to liase with the Renal team	–	renal[locus]	Occurrence; term ambiguity (Renal is an elision of "renal medicine", and not a reference to a patient's anatomical locus)
CT scan shows a partial response in the left lung lesion	CT scan[investigation] has_finding partial response[result]	(1) CT scan[investigation] (2) response[condition] has_location lung[locus]	Typing; occurrence (relation). (Annotator 2 gave no [result]).
no change in the right apical mass	no[negation] modifies change[condition]	no change[negation] modifies mass[condition]	Textual extent
After discussion at the meeting today	discussion[intervention]	–	Occurrence (entity)
an infusional Morphine pump	(1) infusional[intervention] (2) morphine[drug or device]	morphine pump[drug or device]	Occurrence (entity); textual extent
widespread metastatic disease to bone	(1) metastatic[condition] (2) bone[locus]	metastatic disease[condition] has_location bone[locus]	Textual extent; occurrence (relation)
thoraco lumber bony tenderness	tenderness[condition] with three has_location: thoraco[locus]; lumber[locus]; bony[locus]	(1) tenderness[condition] has_loation bony[locus] (2) thoraco lumber[sub-location] modifies bony[locus]	Locus modification
Blood tests were performed	tests[investigation] has_location blood[locus]	blood tests[investigation]	Term decomposition
chest: dullness to percussion in the right hemi-thorax	(1) chest[locus] (2) hemi-thorax[locus] modified by left[laterality] (3) percussion[investigation] has_finding dullness[result] (4) percussion[investigation] has_target hemi-thorax[locus]	(1) dullness[condition] has_location chest[locus] (2) percussion[investigation] has_finding dullness[result] (3) percussion[investigation] has_target chest[locus] (4) thorax[locus] modified by left[laterality] (5) thorax[locus] modified by hemi[sub-location]	Compounding of multiple differences in a single small example

Table 11
Examples of annotator difference, for narratives. In the annotator responses, annotated text is underscored, followed by an entity type in square brackets and teletype. Relation types are also in teletype, with modifiers simplified to a single `modifies` relation and its reverse, `modified by`. Text in a normal font with no underlining are comments. Where an annotator created several entities and relations, these may be numbered. A dash – means that no annotation was given by that annotator. The types of difference listed are described in Section 5.4.

Entity	**Number**	**Strict IAA**	**Lenient IAA**
Condition	429	81	84
Drug or device	172	84	85
Intervention	191	64	66
Investigation	220	77	82
Locus	284	78	81
Result	125	69	74
Laterality	76	95	95
Negation	55	67	76
Sub-location	49	63	64
Overall	**1601**	**77**	**80**
Relation	**Number**	**IAA**	**CIAA**
has_finding	233	48	76
has_indication	168	35	51
has_location	205	59	80
has_target	95	45	64
Modifies (Laterality)	73	70	93
Modifies (Negation)	67	63	90
Modifies (Sub-location)	43	52	98
Overall	**884**	**52**	**75**

Table 12

Distribution and IAA (%) of entities and relations in the 50 narrative documents in the CLEF stratified random corpus.

Entity	**Number**	**Strict IAA**	**Lenient IAA**
Condition	357	67	73
Drug or device	12	59	59
Intervention	53	57	62
Investigation	145	56	58
Locus	357	71	75
Result	96	29	33
Laterality	14	88	88
Negation	50	71	78
Sub-location	77	29	36
Overall	**1161**	**62**	**67**
Relation	**Number**	**IAA**	**CIAA**
has_finding	263	26	69
has_indication	47	15	30
has_location	270	44	70
has_target	86	20	47
Modifies (Laterality)	14	70	89
Modifies (Negation)	54	67	100
Modifies (Sub-location)	79	29	100
Overall	**813**	**36**	**72**

Table 13

Distribution and IAA (%) of entities and relations in the 50 histopathology reports in the CLEF stratified random corpus.

Entity	**Number**	**Strict IAA**	**Lenient IAA**
Condition	270	77	81
Drug or device	13	32	42
Intervention	10	43	43
Investigation	66	70	74
Locus	373	75	81
Result	71	48	52
Laterality	85	91	92
Negation	53	65	76
Sub-location	125	36	46
Overall	**1066**	**69**	**75**
Relation	**Number**	**IAA**	**CIAA**
has_finding	156	33	55
has_indication	12	14	22
has_location	268	45	77
has_target	51	67	81
Modifies (Laterality)	82	55	80
Modifies (Negation)	59	51	94
Modifies (Sub-location)	125	32	93
Overall	**753**	**43**	**76**

Table 14

Distribution and IAA (%) of entities and relations in the 50 imaging reports in the CLEF stratified random corpus.

CTLink	**Task A**	**Task B**
After	5	18
Ended_by	3	0
Begun_by	4	0
Overlap	7	26
Before	5	135
None	4	8
Is_included	31	67
Unknown	6	14
Includes	13	137
Total	**78**	**405**

Table 15

Distribution of CTLinks by type for tasks A and B, over 10 development documents.

TLCs	Not hypothetical	243
	Hypothetical	16
	Total	**259**
Time Expression	Duration	3
	Date	52
	Total	**55**

Table 16

Distribution of TLCs and temporal expressions, over 10 development documents.

Entity type	**Metric**			**IAA**
	P	R	F1	
Condition	0.819	0.654	0.724	0.751
Drug-or-device	0.83	0.592	0.684	0.781
Intervention	0.75	0.616	0.665	0.554
Investigation	0.831	0.659	0.73	0.745
Locus	0.8	0.616	0.694	0.793
Overall	0.807	0.631	0.707	0.737

Table 17

Entity recognition scores for the CLEF IE System.

Relation	**Metric**			**CIAA**
	P	R	F1	
has_finding	0.63	0.82	0.71	0.80
has_indication	0.44	0.47	0.41	0.50
has_location	0.73	0.83	0.76	0.80
has_target	0.59	0.68	0.62	0.63
laterality_modifies	0.86	0.89	0.85	0.94
negation_modifies	0.81	0.93	0.85	0.93
sub_location_modifies	0.87	0.95	0.90	0.96
Overall	0.64	0.76	0.70	0.75

Table 18

Relation extraction scores for the CLEF IE System.

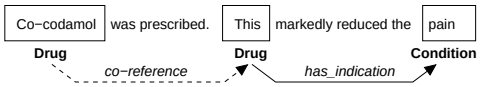

Fig. 1. Annotations, co-reference, relationships.

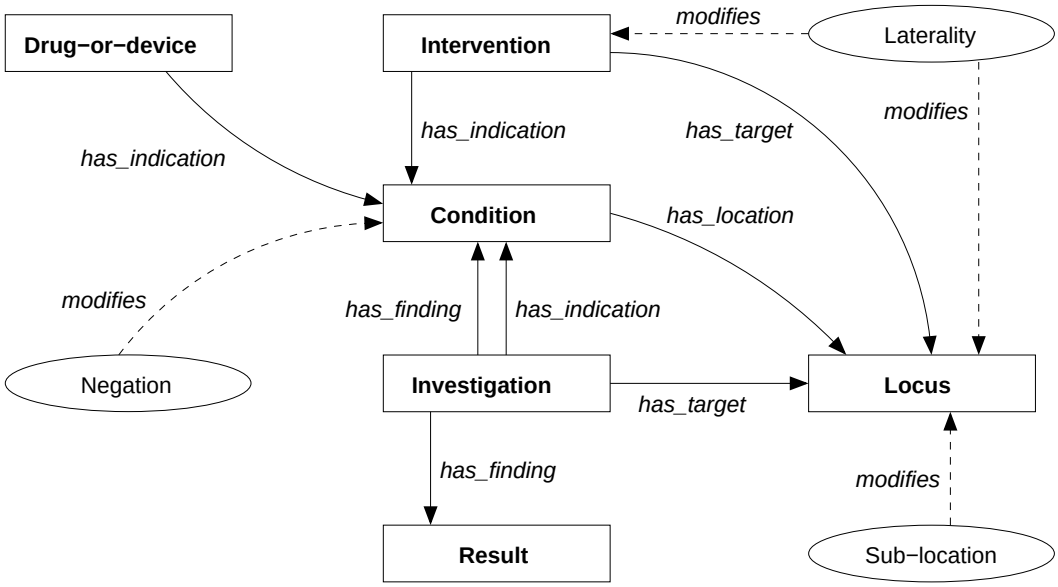

Fig. 2. The CLEF annotation schema. Rectangles: entities; ovals: modifiers; solid lines: relationships; dotted lines: modifier relationships.

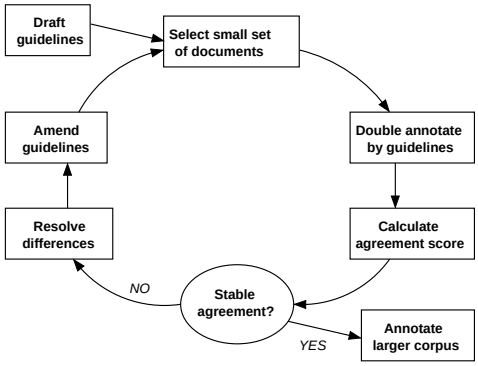

Fig. 3. Iterative development of guidelines.

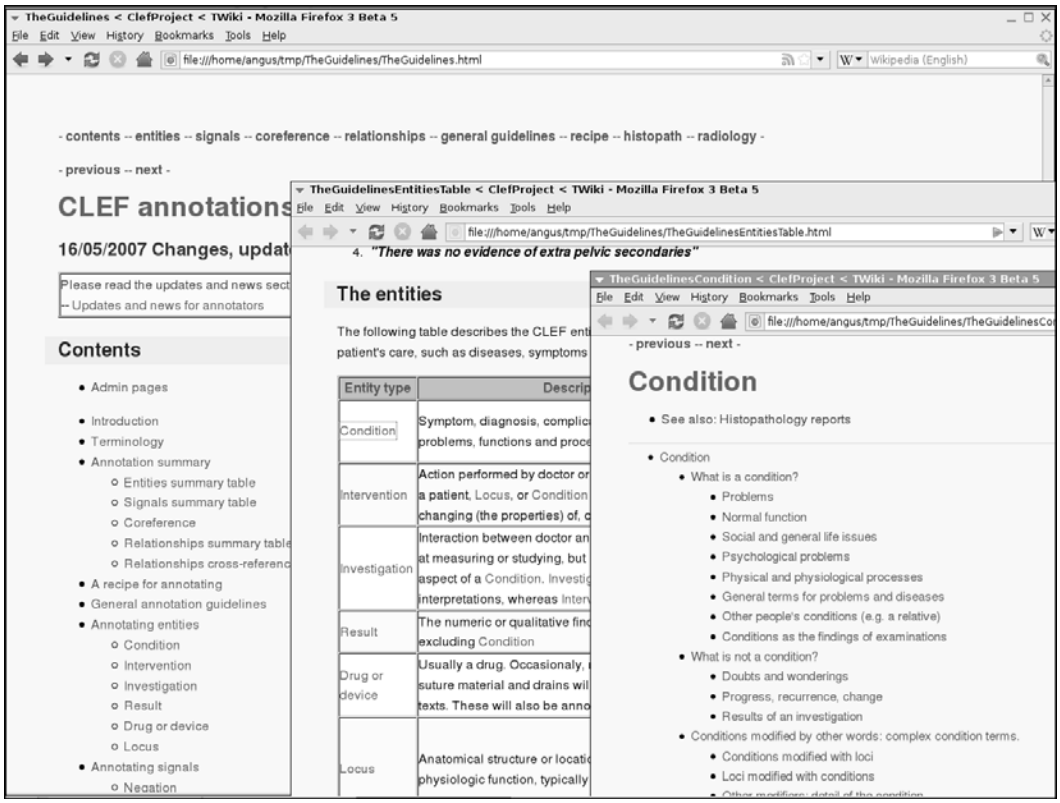

Fig. 4. The CLEF Annotation Guidelines web site. From a window showing the menus and contents, the user has opened a table of all entities, and from this window has opened the `Condition` guidelines.

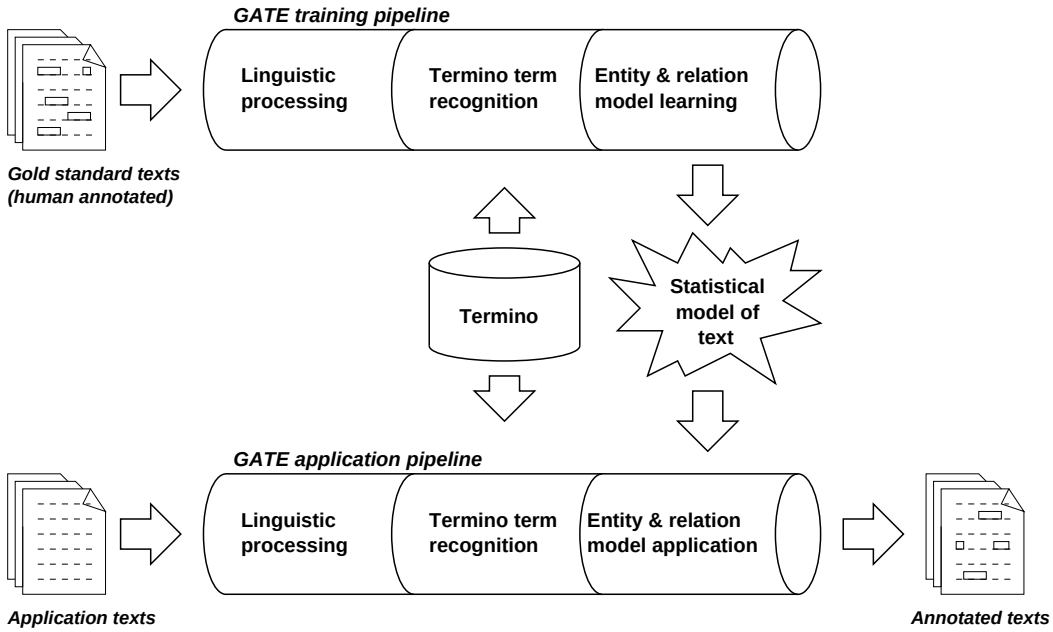

Fig. 5. The CLEF Information Extraction system.

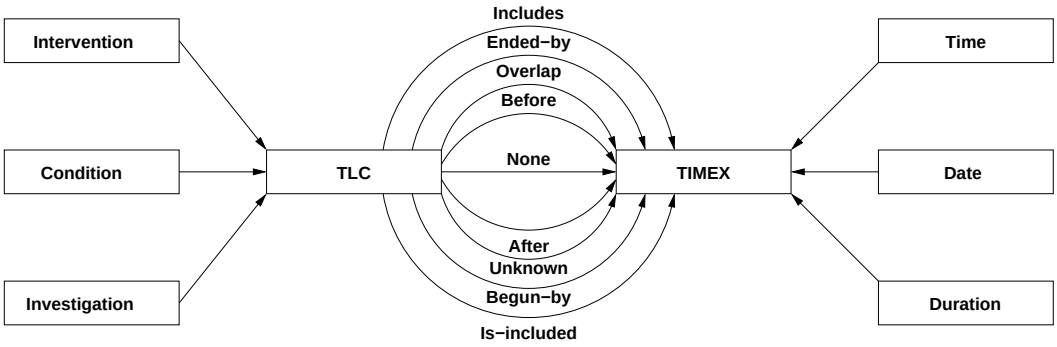

Fig. 6. The Temporal Annotation Schema.