THE UNIVERSITY OF
SYDNEY

A RULE-BASED METHODOLOGY AND FEATURE-BASED METHODOLOGY FOR EFFECT RELATION EXTRACTION IN CHINESE UNSTRUCTURED TEXT

JINGCHENG WANG
SID: 308212029

Supervisor: Josiah Poon

A thesis submitted in fulfilment of the
requirements for the degree of
Master of Philosophy

School of Information Technologies at
The University of Sydney

June 2015

THE UNIVERSITY OF SYDNEY

STUDENT PLAGIARISM: COURSE WORK - POLICY AND PROCEDURE

COMPLIANCE STATEMENT

INDIVIDUAL / COLLABORATIVE WORK

I/We certify that:

(1) I/We have read and understood the *University of Sydney Student Plagiarism: Coursework Policy and Procedure*;

(2) I/We understand that failure to comply with the *Student Plagiarism: Coursework Policy and Procedure* can lead to the University commencing proceedings against me/us for potential student misconduct under Chapter 8 of the *University of Sydney By-Law 1999* (as amended);

(3) this Work is substantially my/our own, and to the extent that any part of this Work is not my/our own I/we have indicated that it is not my/our own by Acknowledging the Source of that part or those parts of the Work.

Name(s): Jingcheng Wang

Signature(s): 

Date: 30/06/2015

Abstract

As the development of technology continues to accelerate, the need for better storage and easier access of documents grows. As a result, more and more text documents are being converted to or generated in digital form. However a large proportion of these digital documents are unstructured free text and the retrieval of useful information requires human access. Relying of human access to retrieve useful information from such a vast number of documents is infeasible; there is great need for systemic methodologies to make use of these unstructured texts in an automated fashion.

The Chinese language differs significantly from English, both in lexical representation and grammatical structure. These differences lead to problems in the Chinese NLP, such as word segmentation and flexible syntactic structure. Many conventional methods and approaches in Natural Language Processing (NLP) based on English text are shown to be ineffective when attending to these language specific problems in late-started Chinese NLP.

Relation Extraction is an area under NLP, looking to identify semantic relationships between entities in the text. The term “Effect Relation” is introduced in this research to refer to a specific content type of relationship between two entities, where one entity has a certain “effect” on the other entity. In this research project, a case study on Chinese text from Traditional Chinese Medicine (TCM) journal publications is built, to closely examine the forms of Effect Relation in this text domain. The relationships expected to be identified in this case study are the effect of a prescription or herb, in treatment of a disease, symptom or body part.

A rule-based methodology is introduced in this thesis. It utilises predetermined rules and templates, derived from the characteristics and pattern observed in the dataset. This methodology achieves the F-score of 0.85 in its Named Entity Recognition (NER) module; 0.79 in its Semantic Relationship Extraction (SRE) module; and the overall performance of 0.46. A second methodology taking a feature-based approach is also introduced in this thesis. It views the RE task as a classification problem and utilises mathematical classification model and features consisting of contextual information and rules. It achieves the F-scores of: 0.73 (NER), 0.88 (SRE) and overall performance of 0.41. The role of functional words in the contemporary Chinese language and in relation to the ERs in this research is explored. Functional words have been incorporated into the two ER extraction methodologies. They are found to be significantly more effective in detecting the complex structure ER entities as rules in the rule-based methodology than as features in the feature-based methodology. A TCM dataset of more than 2000 sentences and more than 1400 ER annotations have been constructed for the experimentation and evaluation of the two ER extraction methodologies. Over 500 ER related dictionary words have been collected and categorized from the dataset to support the dictionary related methods in the two methodologies.

Keywords: Relation Extraction, Medical Text Mining, Chinese text, Rule-based, Feature-based, Functional Words

Acknowledgements

I would never have been able to finish this thesis without the guidance of my supervisors, help from friends, and support from my family.

I would like to express my deepest gratitude to my supervisor, Dr. Josiah Poon, for his excellent guidance and patience in the course of my research. I would like to thank him for the numerous times his advices shed new light when I felt I've reached a dead-end in my research, and the careful proofreading of my thesis.

I would like to thank my auxiliary supervisor, Dr. Simon Poon and the members of the LEMMA reading group for their sharing of research ideas and advices with me. The regular presentations have broadened my perspective towards my field of research.

I would also like to thank my parents, my younger brother and my friends for their support and encouragement to my research. Their caring words have brought great comfort during the hours I spent struggling to write and edit the numerous drafts of this thesis.

Definitions

- Chinese character – A Chinese character is the most basic unit of measure in a Chinese text. Each character has a meaning of its own, and when combined to form a Chinese word (see "Chinese word" below), the word may take a similar meaning, or very different meaning to the characters it consists of. For example, in this piece of text: "猫会游泳" (cat can swim); there are 4 Chinese characters: "猫", "会", "游", "泳"; and 3 Chinese words: "猫", "会", "游泳". The character "游" and "泳", though both have meaning of "swim", are usually used together as one word.
- Chinese word – A Chinese word consists of one or more Chinese characters (see "Chinese characters" above).
- Effect Relation (ER) – a type of relation is Relation Extraction, defined by this paper. (Further details in Chapter 3. Effect Relation)
- Effect Relation entity (ER entity) – text tokens that are used to form Effect Relations.
- Effect Relation entity classes – classes used to mark text tokens for their involvement in an Effect Relation. These classes are: "SOURCE", "TARGET", "EFFECT" (text tokens of these 3 classes form Effect Relation), and "None" (text tokens of this class is not directly involved with Effect Relation, they may be used for context support, but otherwise disregarded).
- *n*-char word – Chinese words consists of one or more (represented by *n*) characters, for example, 1-char word refers to words such as "猫"; and 2-char word refers to words such as "游泳".
- Performance – the performance of approaches / methods in this thesis are measured by several factors commonly used for information extraction research: precision, recall and F-score. When no factors are mentioned, performance refers to F-score.
- TCM ingredients – the generic term referring to the ingredients in Traditional Chinese Medicine prescriptions, such as 枸杞 *wolfberry*. These ingredients are also often referred to as 草药 (*herb*), but there are exceptions such as 穿山甲 (*armadillo*) and 鹿茸 (*velvet antler*). In this thesis, the term "TCM ingredients" refers to all types of ingredients used in TCM prescription.

Abbreviations

- ACE – Automatic Content Extraction
- CCG – Combinatory Categorial Grammar
- DARPA – Defense Advanced Research Projects Agency
- DP – Dependency Parsing
- ER – Effect Relation
- ET – EFFECT TARGET
- FWF – Functional Word Features
- FWR – Functional Word Rules
- GATE – General Architecture for Text Engineering
- IE – Information Extraction
- LTP – Language Technology Platform
- MUC – Message Understanding Conference
- NIST – National Institute of Standards and Technology
- NLP – Natural Language Processing
- NER – Named Entity Recognition
- POS – Part-Of-Speech
- RE – Relation Extraction
- SRE – Semantic Relationship Extraction
- SRL – Semantic Role Labelling
- TAC – Text Analytics Conference
- TCM – Traditional Chinese Medicine

Table of Contents

List of Figures

List of Tables

Chapter 1. Introduction

This chapter provides a brief overview on the domains related to this research. It also provides insights to the purpose and motivation of this research, the goals set out to achieve in this project and the key contributions to the field of this research.

As the development of technology continues to accelerate, the need for better storage and easier access of documents grows. As a result, more and more text documents are being converted to or generated in digital form. A large proportion of these digital documents are unstructured free text and the retrieval of useful information requires human access. Relying of human access to retrieve useful information from such a vast number of documents is unfeasible; there is great need for systemic methodologies to make use of these unstructured texts in an automated fashion.

The Chinese language differs significantly to English, both in lexical representation and grammatical structure. These differences lead to problems in the Chinese NLP, such as word segmentation and flexible syntactic structure. Many conventional methods and approaches in Natural Language Processing (NLP) based on English text are shown to be ineffective when attending to these language specific problems in late-started Chinese NLP. This research looks into a specific type of Relation Extraction task under NLP in Chinese unstructured text.

1.1 Relation extraction

Relation Extraction (RE) is a significant topic in the task of Information Extraction (IE) under the field of Natural Language Processing (NLP). The Message Understanding Conference formally inducted this topic at its 7th conference (MUC-7) in 1998 [1]. Zelenco defines Relation Extraction as “the method of efficient detection and identification of predefined semantic relationships within a set of entities in text documents” [2]. This topic includes extracting relationships as generic as “part-to-whole” or “person-social” relations, defined in the Automatic Content Extraction (ACE) [3], which can be applied on almost all unstructured text. As it developed, Relation Extraction has also found significant applications in medical science domain, extracting specific relationships, such as “gene to disease” relationships [4] or “protein to protein” interactions [5]. This research introduces the extraction of “Effect Relation”, a relation similar to the “gene to disease” relationship.

1.2 Field of Traditional Chinese Medicine

Controversy continues about the validity of Traditional Chinese Medicine (TCM) methodology. TCM has had an extensive history in the Chinese culture and produced much literature in this field of research. However many still question the validity of TCM as a practise of medicine. As defined in the Gale Encyclopaedia of Medicine, Traditional Chinese Medicine is considered “a complete system of health care with its own unique theories of anatomy, health and treatment” [6]. TCM shows great difference to the mainstream medicine which is defined in Segen’s Medical Dictionary as “the ‘Western model’ of evidence-base practise for diagnosing and treating disease” [7]. More TCM procedures are

conducted and more TCM literatures are being produced in ways recognised by the mainstream medical field. A method to automatically extract the relationships between specific entities, such as “herb to symptom” relationships, in TCM literatures will go a long way to help gain more understanding of TCM from the “Western model” perspective.

1.3 Problem definition and motivation

This section describes the challenges anticipated in this research and my motivation to undertake this research and overcome these challenges.

1.3.1 Purpose of Effect Relation

An Effect Relation (ER) is the relation between two entities that can be described as a positive or negative effect. Its primary application in the TCM domain is to detect the effects found of an herb or prescription on a body part or symptom across a range of TCM journal publications. Similarly, the definition of Effect Relation can be revised and applied in other medical science domains to identify the relationships of certain effect between those medical entities.

1.3.2 Information extraction on Chinese text

Natural Language Processing (NLP) research on Chinese text has had a late start compared to English. Much difference exists between the two languages as English is of the Indo-European language family and Chinese is of the Sino-Tibetan language family. Many conventional methods were designed based on English text, and may not be entirely applicable to Chinese text.

One of the biggest challenges in NLP on Chinese text is word segmentation. Chinese text lacks the notion of word delimiter. Where an English sentence uses space to indicate the start and end of each word, a Chinese sentence is a string of characters and the reader needs to interpret which words are formed within. Therefore Chinese text parsers are required to perform such interpretation, known as word segmentation, which English word parsers do not need this function. This also means errors can be introduced into the dataset at this early stage.

Another challenge is the flexibility of Chinese syntax in sentence structure. Shi claims that the syntax of Chinese is often described as “more flexible than that of other languages because it is based on hypotaxis or has a ‘soft’ rule system.” [8] This flexibility often makes parsing words according to syntax rules quite difficult.

Upon examining these challenges specific to the Chinese language, I hope to develop a method using specific characteristics in the Chinese language to overcome or reduce the impact of their existence. One of these characteristics I have chosen is functional words. The Chinese functional words are very significant to interpreting the meaning of a sentence, and one of the key elements contributing to the syntactic flexibility.

1.4 Project goals

The primary aim of this project is to develop a methodology to extract “Effect Relations” (ER) from Chinese scientific documents. This methodology will cover the entire process: from the dataset preparation module, to the Named Entity Recognition (NER) module, to the Semantic Relationship Extraction (SRE) module, and finishing with the methods of evaluating this methodology. Since Effect Relation is a new concept in this research and there is no existing methodology for this research to benchmark against, it becomes difficult to evaluate the results of this research. To solve this, two methodologies will be developed. One will be set as the baseline, and the other will be compared to the baseline to evaluate the influence of the difference factors between the two methodologies.

The secondary aim of this project is to explore the role of functional words and how they can be incorporated into the ER extraction methodologies. Additional tests will be performed to evaluate the effectiveness of the incorporation of functional words.

These goals will be achieved by completing the following tasks:

- The exploration of the definition of ER to attain a concrete understanding of this type of relation and the ways it appears in text.
- The development of two methodologies to extract Effect Relation from Chinese text, one for baseline and one for comparison.
- The exploration of how functional words behaviour around or in regards to “Effect Relations” in the sentence and ways to incorporate the exploitation of these behaviours into the two methodologies.
- The construction of a dataset to conduct experiments to test the two methodologies.
- The analysis and comparison of results of the two methodologies.
- The analysis and comparison of results of the two methodologies with and without the use of functional words.

Achieving these goals will provide insight and contribution to the research of Effect Relation extractions and provide results for future research to benchmark against.

1.5 Contributions

In the course of this research project, the concept of “Effect Relation” is introduced and further refined in the environment of Chinese text from TCM journal publications. Detailed observation of the annotated dataset detected patterns, which are further used in the ER extraction methodologies; and ambiguous cases, which set down the boundaries of what is or is not considered to be an ER.

There was no existing TCM dataset, so a dataset is constructed from raw text in this project. Details of how the raw text is processed into the form in the dataset are included in this

thesis. Approximately 2000 sentences are processed and approximately 1400 Effect Relation annotations form the TCM dataset used in the experiments of this research project.

A rule-based methodology to extract Effect Relations has been developed. It consists of the pre-processing stage to perform the necessary preparation for the dataset; the NER module to arrive with potential entities and other components to form Effect Relations; the SRE module to distinguish the correct combination of ER entities to form ERs; and methods of evaluation to set the performance baseline for modules individually and the methodology as a whole.

A feature-based methodology to extract ER has in additionally been developed. It consists of similar modules as the rule-based methodology, but the methods used within the modules operate differently. By utilising the methods of evaluation, not only can we compare the difference in performance of the two methodologies, but also the difference of the modules individually.

In this research, the presence of functional words in relation to the ERs in the sentence has been examined. Some insights are obtained on how a selection of functional words can indicate the location or narrow down the range of where certain ER entities reside. These insights are useful when dealing with ER entities of a complex form, which cannot be easily detected by conventional methods. Different implementations of the functional words have been included into the two methodologies.

Finally, in the course of implementing these methodologies, a list of dictionary words has been collected. They have been very useful in the process of ER extraction on this dataset. They have also been categorised so if working with a larger or different dataset, these dictionary words can be expanded by external databases in respect to the category to enable better adaptability. For example, the category of “TCM ingredients” can be expanded by TCM databases.

In summary, the six main contributions provided to the research area of Relation Extraction on Chinese text in medical journals are:

1. Detailed definition of Effect Relation in its concept and textual form in the field of TCM literature.
2. Description of how a new TCM dataset is produced from raw text, and the construction of the TCM dataset.
3. Development of a rule-based ER extraction methodology, consisting of the Named Entity Recognition (NER) module and the Semantic Relationship Extraction (SRE) module. Extensive experiments have been carried out to understand the behaviour and characteristics of this methodology.
4. Development of a feature-based methodology, consisting of the NER module and the SRE module. Extensive experiments have been carried out to understand the

behaviour and characteristics of this methodology and its differences with the rule-based methodology.

5. Greater understanding of functional words, and their effectiveness of their implementations in rule-based and feature-based methodologies
6. Collection and categorization of dictionary words useful in ER extraction.

1.6 Thesis structure

This thesis covers the development of two methodologies to extract Effect Relation from Chinese text on Traditional Chinese Medicine journal papers. It explains the initial concept and the developed definition of the term “Effect Relation”. It further describes the rule-based methodology and feature-based methodology developed to perform the extraction. It also sets out the possible benefits by utilising functional words in these methodologies.

This chapter provides a brief overview on the domains related to this research. It also provides insights to the purpose and motivation of this research, the goals set out to achieve in this project and the key contributions to the field of this research.

Chapter 2 provides an overview of the literature and research related to the field of this research. This consists of research on the fields of information retrieval and information extraction on Chinese text, and the correct usage of functional words.

Chapter 3 describes the overall concept and definition of the term “Effect Relation”.

Chapter 4 describes how approaches were selected for the ER extraction methodologies in this research.

Chapter 5 describes the importance of functional words in the Chinese languages. It also specifies the possible ways to implement functional words on the two methodologies and improve their performance on extracting Effect Relations.

Chapter 6 explains the design of the implementation of the two methodologies, and how are functional words used in each of the approaches. It also specifies how the dataset is prepared.

Chapter 7 specifies the experimental setup used to evaluate the performance of the rule-based methodology and feature-based methodology.

Chapter 8 reports on the results found by the experiments performed.

Chapter 9 evaluates how well the development and implementation of the two methodologies and usage of the functional word performed as to achieve the goals set for this research project.

Chapter 10 concludes the work done in this research and identifies some possible directions for future work into Effect Relation extraction.

Chapter 2. Background

Chapter 2 provides an overview of the literature and research related to the field of this research. This consists of research in the fields of information retrieval and information extraction on Chinese text. This chapter also provides insight to research related to Chinese functional words. They play a significant role in the contemporary Chinese language and they may be a valuable resource to Chinese NLP.

2.1 Relation Extraction

As the development of technology continues to accelerate, the need for better storage and easier access of documents grows. As a result, more and more documents are being converted to or generated in digital form. However a large proportion of these digital documents are unstructured free text and the retrieval of useful information requires human access. Relying of human access to retrieve useful information from such a vast number of documents is unfeasible; there is great need for systemic methodologies to make use of these documents in an automated fashion. One of such example is Relation Extraction.

Relation extraction (RE) refers to the method of efficient detection and identification of predefined semantic relationships within a set of entities in text documents [9] [10].

This methodology has been a topic of the MUC held from 1987 to 1997, under supervision of DARPA1. Later it continued to take part in the Automatic Content Extraction (ACE) workshops, supervised by NIST2. This workshop is conducted as a sub-field of Text Analytics Conference (TAC) which is also currently under the supervision of NIST [11].

2.1.1 Common approaches

Approaches to the RE task will vary according to the nature of the relation intended for extraction and the domain of text the RE task is applied. In [12], Sharma et al points out the four main approaches towards RE in the past: the co-occurrence based, link based, rule-based and machine learning approaches.

The co-occurrence based approach captures the relationship between two entities which frequently co-occur in a certain field of interest. For example, [13] extracts correlated files and web pages belonging to the same task from file-access logs and web-page-access logs. The link based approach extends on the relations extracted in other approaches to construct new indirect links between two entities that both have a direct relation to another common entity [14].

The rule-based approach constructs hand-craft rules to extract specific relationships between entities, and therefore will rely heavily on both the syntactic and semantic units in the text [12]. Fundel et al. builds a set of rules on dependency parse trees of text to extract gene and protein relations [15]. Feldman et al. uses a structure-driven rule-based strategy to essentially extract relations falling under the generic template of two noun phrases connected by a verb [16]. Similarly, Sharma et al. proposes a methodology of identifying the relationship of two entities by the main verb/s in the sentence. They claim this proposed

verb-centric approach can “effectively handle complex sentence structures such as clauses and conjunctive sentences.” [12]

The machine learning approach is relatively new compare to the other approaches and shows growing interest in the field of NLP. This approach looks into learning the construct and patterns of a collection of known samples (often referred to as the training dataset), to make predictions on unknown samples (often referred to as the testing dataset). Feature-based approach and kernel based approach are two widely used approaches in the area of machine learning [11].

The feature-based approach focuses on the selection of features and vector of feature weights to perform tasks such as statistical classification and regression analysis. K Nearest Neighbours [17] is a commonly used classification algorithm of the feature-based approach. The kernel based approach calculates the similarity level between pairs of instances to study relational tasks such as clustering and statistical classification. Support Vector Machine is a learning algorithm widely used in the kernel based approach [18]. The two approaches can be described as dual, as they both employ classification models or algorithms to learn and make prediction, and are often mathematically interchanged.

2.1.2 Text parsing

As most RE methodologies utilises more information than simply the raw text, text parsing is one of the most important module, if not all, of the pre-processing stage of performing a RE task. Shallow parsers (or partial parsers) can provide word level syntactic information such as Part-Of-Speech (POS), tagging every word in a sentence of their likely grammatical function [19]. This information is significantly utilised in methodologies like [12]. Deep parsers can provide sentence structural information by fitting the sentence into a tree structuring using grammar formalisms such as the Combinatory Categorial Grammar (CCG) [20] [19]. This information is significantly utilised in methodologies like [15].

2.1.3 Named Entity Recognition

Most RE systems consist of the NER module, where entities potential to form relationship in the text are identified; and the Semantic Relation Extraction (SRE) module, where the identified entities are paired to determine the content and type of existing relationship. A well performing NER module is a key factor to a well performing RE system.

The BioAnnotator in [21] uses domain based dictionary look-up to recognise known biomedical terms and set of rules to recognise new terms. This system achieved good performance (87% precision and 94% recall on the GENIA 1.1 corpus). The method of domain based dictionary look-up is further discussed in 6.3.1.

In other examples, NER is performed in the same step as the SRE. In [22], Bundschus et al. utilises a selection of features such as dictionary feature, start window feature and key entity neighbourhood feature, which shows to be useful for NER when taking a feature-based approach. The use of these features is further discussed in 6.5.2.

2.1.4 Nature of text

In order to determine the approach for a RE task, nature of the text also needs to be considered. RE methodologies are considered domain based if they heavily rely on certain characteristics of the text. Text of different domains may differ not only in language and contents, but also how well the text is "structured" for the ease of automatic processing. In Soderland's paper [23], he gave example of two styles of text mainly dealt with in Information Extraction tasks, namely semi-structured text and unstructured free text.

Semi-structured text is "often ungrammatical and telegraphic in style, but does not follow any rigid format". Such examples may be medical records, equipment maintenance logs and rental ads [23]. Unstructured free text is text where two pieces of text from the same domain may be totally different in terms of structure. News story text is an example of this [23].

2.1.5 Relation Extraction on Chinese text

At present, the fundamental approaches of RE is still largely based on the English text. Meanwhile, the number of digital documents in Chinese or other oriental languages has grown greatly in recent decades. These documents are mostly published in Japan, Korea & Singapore, Hong Kong, Mainland China, Taiwan etc. These documents are mostly non-structured. Although the mainstream RE methodologies on Chinese text are similar to that of English text, significant differences in lexical representation and grammatical structure do exist between the two languages. Therefore, many researchers in oriental countries are exploring new techniques adaptable to their native languages [9] [24] [25] [26] [27] [28] [29].

In most impressions, Chinese and other oriental languages are generally more difficult to perform automatic text classification upon. However, Fung's [30] response to such impressions is "that Chinese and Japanese are difficult only if one seeks to fit them into an Indo-European model of linguistic theories."

2.1.5.1 Word segmentation

Linguists have argued that the smallest semantic unit is often not a single word, as defined by a string of letters delimited by spaces, but a phrase (or a term) [31] [32] [30]. However unlike English, the Chinese language does not have natural delimiters between words such as spaces. A sentence is constructed by words, which within itself is formed by characters. There are more than 60, 000 characters in the Chinese language [33], the difficulty increases as Pinchuck [31] had claimed, characters are not the unit which expresses meaning. Combination of character is quite flexible, word segmentation and extraction is difficult task for human understanding as well [33].

Therefore, performing word segmentation on Chinese text is quite difficult, especially in cases of unknown words, such as names, locations, translated terms, technical terms, abbreviations etc. [34]. Ambiguity brought by the various ways that different words can be formed from characters induces more errors to any further processes [33]. Without efficient word segmentation, many information retrieval applications, such as full-text

searching [35], document classification [36], information filtering [37] and text summary [38], cannot obtain satisfactory achievements [29].

The lexical approach and the grammatical approach are the two most common approaches for the word segmentation task on Chinese text [30]. Both Fung [30] and Chien [29] believe that the lexical approach is more suitable. For example, the Csmart system [39] "ignores the concept of words and uses character-level information to replace word-level information in the construction of IR systems" [29].

2.1.5.2 Existing methodologies

The relation intended for extraction in this research project is different from standard relations such as part-to-whole or person-social relations define in ACE08 [3] (discussed further in Chapter 3). The RE methodologies in this research cannot be a direct extension of an existing methodology, however inspirations can be drawn from certain characteristics of existing methodologies.

Li et al introduces a RE approach on Chinese text using "linguistic clues and pattern-matching" [40]. They defined a relation as: R=(X,Y,G); where X and Y are the entities of the relation R, signifying the start and end of the relation; G is the type of the relation R. They derived 8 common rules and 2 exception rules in their NER module, which are extended upon in the NER module of the rule-based methodology in this research project (discussed further in 4.1.3 and 6.3.3).

Gu et al., on the other hand, proposes a verb-centric RE methodology, similar to [12]. However their methodology is based on the dependency parse tree for both the NER module and SRE module. They used the Language Technology Platform (LTP) [41] developed by the Research Center for Social Computing and Information Retrieval at Harbin Institute of Technology (HIT-SCIR) for pre-processing the dataset, which is also used in this research (discussed further in 6.2.3).

2.2 Functional words

Studies on modern Chinese functional words are often human oriented and subjective or vague in definition [42]. Consequently they are often used merely as word delimiters in Chinese NLP methodologies. Chen [43] describes the Chinese language as "whose typology involves isolated language and root-word-based analytical language. The vast majority of Chinese words cannot themselves clearly express grammatical sense. Chinese syntactical methods mainly depend on the use of functional words and word order." There is great potential for improvement if functional words are able to play a more important role in RE methodologies in Chinese text.

In Chen's [43] work, she grouped Chinese functional words into ten categories, namely, Voice Category, Tense & Aspect Category, Mood Category, Modality Category, Positive and Negative Category, Relation Category of Sentences, Degree Category, Range Category, Object Category, Other Category.

Yu et al. proposes the idea of building the “Trinity” knowledge base of modern Chinese functional words, consisting of a dictionary of functional words, a list of their definitions and a dataset of their actual usage in sentences [44] [45].

Zhang and Zan examines in detail individual functional words. In their works, they reported on ways to identify the correct definition of the functional words, DOU [46] and JIU [47]. These are examples of the most commonly used functional words, and are part of a greater research project they conduct on the study of functional words.

In these previous works, the authors pointed out the importance of functional words to understanding the Chinese language. However these researches are mostly focused on clarifying the types and definitions of functional words and how to better extract the correct definition using the contextual information around the functional words. There has been very little work done on how to use these functional words as contextual information to improve other tasks, in this case, RE tasks. Previous works have shown the definition and usage of functional words consists of rules of high complexity, ranging from generic rules to rules dealing with exceptions. To use only a selection of these rules may not be entirely correct from the perspective of studying functional words, but it may still prove useful to recognising patterns in the text and improvement of processes in the RE task (discussed further in Chapter 5).

Chapter 3. Effect Relation

This chapter describes the background concept of the term “Effect Relation” (ER), which is the core topic discussed in this research. It explains the progress from a generic concept to a standardized definition, the manual annotation performed on dataset, patterns that arise in the annotations and many ambiguous cases that challenge the definition along the way.

3.1 Concept

Generally, the “relation” in Relation Extraction refers to any relationship between two entities, such as “A is a part of B” or “A is also called B”. The presence of a “relation” signifies a link between two entities, but the content of the relationship can vary from one relation to another. The term “Effect Relation” is introduced in this research to refer to a specific content type of relationship between two entities, where one entity has a certain “effect” on the other entity, i.e. “entity A effects (in certain way) entity B” or “entity A has the effect E on entity B”. The ER is a generic relation describing the interaction of two entities, while also focusing on the nature of the interaction. To clarify the vagueness from the many forms an interaction can take, the “effect” in an Effect Relation needs to be quantifiable (to a degree) that it can be recognised as a positive or negative effect. In this research project, a case study on Chinese text from TCM journal publications is built, to closely examine the forms of Effect Relation in this text domain. For example, in a sentence from the case study: “经验方中山药、熟地黄补益肝肾”, the text can be broken down to pieces concerned with the ERs present. (Figure 1)

经验方中	山药	、	熟地黄	补益	肝肾
In this prescription	*yam*		*rehmannia*	*benefit*	*liver and kidney*
Context	Herb1		Herb2	Effect (positive)	Body_part1

Figure 1: Example – an Effect Relation

These pieces can be organised to form a set of ERs. (Table 1)

SOURCE entity	TARGET entity	EFFECT entity	EFFECT type
山药	肝肾	补益	Positive
熟地黄	肝肾	补益	positive

Table 1: Example – Effect Relation extraction result

TCM journal publications mention a wide range of prescription names or TCM ingredients (see “TCM ingredients” in Definitions) used in treatment of a wide range of diseases or healing certain body parts. The relationships expected to be identified in the TCM case study are the effect of a prescription or herb, in treatment of a disease, symptom or body part (as given in Figure 1). The results extracted can be cross-referenced and used in statistical analysis in the understanding of interaction between TCM ingredients and symptoms. Similarly, ERs can be extended to the biomedical field where many other Relation Extraction methodologies, such as protein to protein [48] [49] or gene to disease [4] [22], have been applied.

3.2 Definition

Drawing from the concept, an ER can be defined as "the relation of an entity having an obvious positive or negative effect on another entity". It shares a similar component structure as many other relations (e.g. in [22]):

(1) $$R = < Entity_1, RIndicator, Entity_2 >$$

However ER deviates from the standard types of relations set in ACE08 [3] (Figure 2).

(Relations marked with an * are symmetric relations.)

Type	Subtype
ART (artifact)	User-Owner-Inventor-Manufacturer
GEN-AFF (General affiliation)	Citizen-Resident-Religion-Ethnicity, Org-Location
METONYMY*	*None*
ORG-AFF (Org-affiliation)	Employment, Founder, Ownership, Student-Alum, Sports-Affiliation, Investor-Shareholder, Membership
PART-WHOLE (part-to-whole)	Artifact, Geographical, Subsidiary
PER-SOC* (person-social)	Business, Family, Lasting-Personal
PHYS* (physical)	Located, Near

Figure 2: ACE08 Relation Types and Subtypes [3]

ER possesses a definite asymmetry property, where when A has an effect E on B; it is definitely incorrect to say B has the same effect E on A. This is due to the fact that the two entities in an ER are definitely of two categories, and therefore not interchangeable. Furthermore, in most cases, the "effect" in an ER is a verb or action and the entities being the subject and object. A better resemblance would be as the definition given in the verb-centric approach in [12] (2):

(2) $$R = < Entity_{subject}, Verb_{main}, Entity_{object} >$$

In applying Relation Extraction on Chinese text, Gu provided a similar definition of non-taxonomic relation [50] (3),

(3) $$R = < C_{pre}, V_{rel}, C_{suc} >$$

where the Relation (R) consists of 2 Contents (C): the precursor (pre) and the successor (suc), and a Verb (V) that indicates the type of relation (rel).

(4) $$R_{Effect} = < Entity_{SOURCE}, Entity_{TARGET}, Entity_{EFFECT}, Type_{EFFECT} >$$

The ER consists of four components: a SOURCE entity, a TARGET entity and an EFFECT entity and an EFFECT type, within the scope of one sentence. The relation indicator in ER is also considered as an entity, because it is very distinctive and limited to one word in each ER. This relation (4) essentially indicates that the SOURCE entity has the effect of the EFFECT entity on the TARGET entity. The EFFECT type is directly associated with the EFFECT entity.

The components are named differently to (2), because ER is defined by its semantic meaning (e.g. “source”) rather than syntactic structure (e.g. “subject”).

The EFFECT entities are words that can be interpreted as a positive or negative effect:

- Clearly quantitative – The effect is an increase or decrease on a quantifiable scale. For example, 增加 *increase* and 减少 *decrease*.
- More descriptive – The effect is an improvement or worsening that is not directly translated to a quantifiable scale. For example, 补益 *benefit*, 调整 *adjustment* and 损伤 *damage*.

3.3 Observed characteristics of Effect Relation

From observation of the dataset, many characteristics of ER are identified. The characteristics which are further used in designing of the ER extraction methodologies are: word categories, word repeat rate and ER structure.

3.3.1 Word categories

Upon observing the words in the ERs in the dataset, it is notable that the words present in the SOURCE and TARGET entities can be classified into several categories.

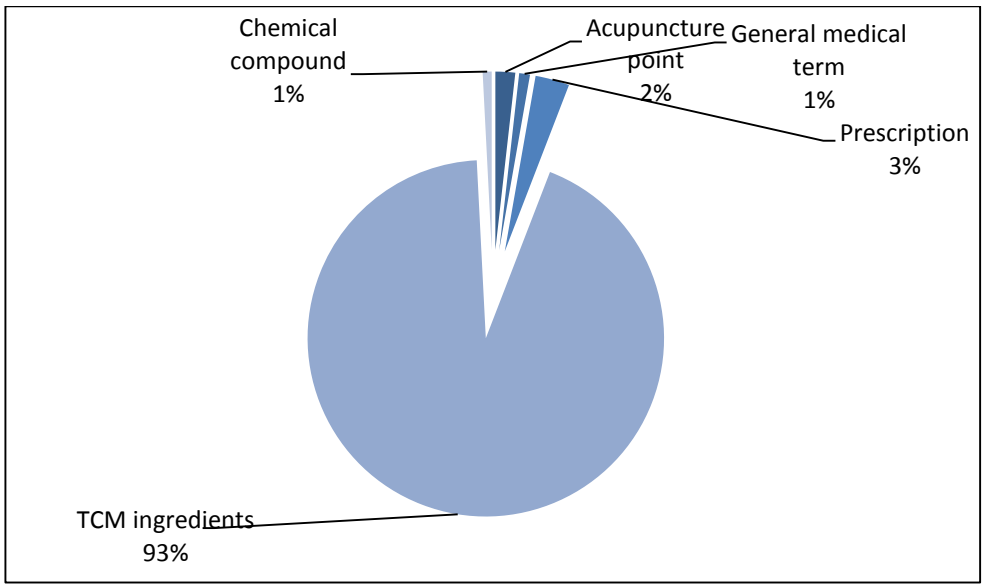

Figure 3: SOURCE word categories

The SOURCE entity always consists of one word, usually a noun. This word belongs to one of 5 categories (in order of high to low presence as Figure 3 above): Tradition Chinese Medicine ingredients (e.g. 杏仁 *almond*) – 93%, prescription (e.g. 清宁胶囊 *Qingning capsule*) – 3%, acupuncture point (e.g. 关元 *Guanyuan point*) – 2%, general medical term (e.g. 诸药 *these medicine*) – 1%, and chemical compound (e.g. 黄连素 *berberine*) – 1%.

The structure of TARGET can be defined as simple or complex. Simple structure refers to cases where the entire TARGET entity is one word, usually a noun, for example 血糖 *blood*

sugar; or as the "TARGET" inside of EFFECT TARGET (further discussed in 6.2.4), part of a verb-noun structured word, for example in the word 养血 *enrich the blood*, 养 (*enrich*) is the effect and 血 (*blood*) is the target. Complex structure refers to cases where the entire TARGET entity consists of more than one word, for example 组织对胰岛素的敏感性 *the sensitivity of tissues towards insulin*, where nouns and the context around the nouns for the entirety of this TARGET entity. Figure 4 below shows that 79.04% of the TARGET within the annotated ERs is simple structured, and 20.96% are complex structured. The complex structured TARGET entities are not further categorised.

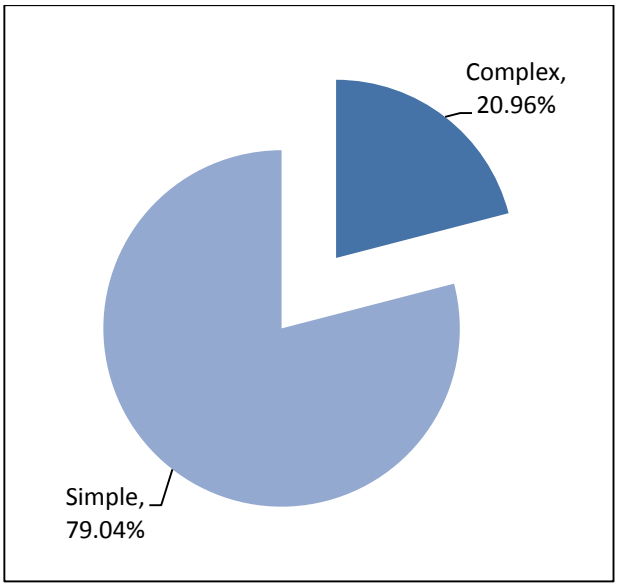

Figure 4: TARGET structure categories

Figure 5 shows that within the simple structured TARGET cases, 51% are words related to body parts (e.g. 胃 *stomach*), 45% are words related to symptoms (e.g. 寒 *cold*), and 4% does not fit into either categories above.

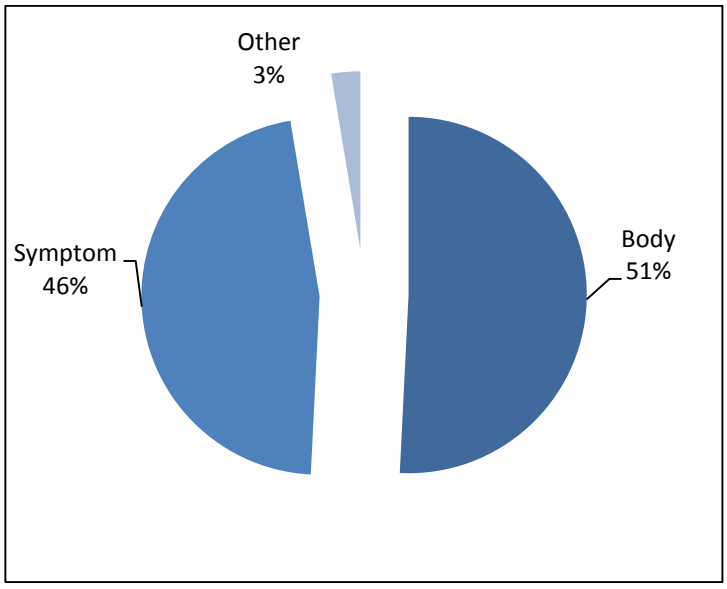

Figure 5: Simple TARGET categories

These categories of words in the ER entities can provide possible new words (beyond what has already been found in the dataset) and improve the performance of extraction methodologies when encountering the anticipated new words. This is discussed further in 6.3.1.

3.3.2 Word repeat rate

Upon observing the ERs in the dataset, certain words began to repeatedly appear. Table 2 below displays the rate for which words are repeated within its ER entity category (i.e. SOURCE, TARGET and EFFECT).

Occurrence	Repeat Rate
SOURCE	63.14%
TARGET	83.42%
EFFECT	84.72%

* $repeat\ rate = \frac{number\ of\ unique\ words/characters}{number\ of\ instances}$

Table 2: Word repeat rate

As shown in Table 2, EFFECT has the highest repeat rate: 84.72%; followed by TARGET: 83.42%. Given such repetition, it means rules based on these words will have a higher rate of usage and lower chance of encountering new words in this category. This characteristic is put to use and discussed further in 3.3.2.

3.3.3 Effect Relation structure

Upon observing the ERs in the dataset, the order of appearance of the ER entities inside each Effect Relation is quite consistent. Table 3 shows the types of order of appearance and their occurrence rate.

Order of appearance	Occurrence Rate
S, E, T	98.15%
S, T, E	1.85%

Table 3: ER order of appearance

As shown in Table 3, 98.15% of overall ERs are structured in the order of SOURCE, EFFECT then TARGET; the remaining 1.85% is in order of SOURCE, TARGET then EFFECT. ERs are only structured in the 2 orders mentioned above and in most cases in the order of "S, E, T".

Words between E and T	Occurrence Rate
No words	98.16%
1-2 words	1.34%
More than 2 words	0.50%

Table 4: Words between EFFECT and TARGET

As shown in Table 4, 98.16% of the EFFECT and TARGET in overall Effect Relations have no words in between, i.e. they are adjacent to each other; 1.34% has 1-2 words in between. From observation, they are when two TARGET entities are sharing the same EFFECT entity and are separated from each other by a punctuation mark or conjunction word; the second TARGET is not directly adjacent to the EFFECT. Only 0.50% of Effect Relations has more than

2 words in between their TARGET entity and EFFECT indicator. In most cases, the EFFECT entity and the TARGET entity are adjacent in an ER; only in rare cases are they separated.

The consistency of how the entities of an ER appear in a sentence can allow templates

3.4 Ambiguous cases

During the course of annotation, many ambiguous cases arise in the text. They seem to be different to the definition in place, but somewhat related. It is often brought to question whether they should be included into the definition. Some significant examples are given below.

3.4.1 Conditioned joining

In some cases, SOURCE entities can be joined by conjunction. The conjunction signals that these SOURCE entities are operating together to achieve the EFFECT. The decision to whether these SOURCE entities should be extracted in separate ERs or one single ER is arguable. An example is given in Figure 6.

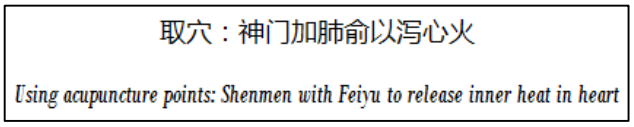

Figure 6: Example - conditioned joining

In the example in Figure 6, the SOURCE entities are in “Shenmen with Feiyu” (神门加肺俞). Given the “with” is a conjunction; this string of text can be broken into two SOURCE entities Shenmen and Feiyu. However what the example states is that the combination of these two acupuncture points has the effect to release inner heat in heart. It does not state that each independently will have the same effect. This is a common occurrence in TCM curing methodology. Often more than one acupuncture point or TCM ingredient is used to tend to a symptom. It will not be entirely true to conclude each will have the same effect independently. However if the entire combination is taken as the SOURCE entity, the purpose of the ER extraction results will be greatly reduced. There are a vast number of ways to combine curing elements, and a very low chance that the exact combination will repeat, even over vast amount of text.

In this project, the above example is considered as two independent SOURCE entities. Further research can look into implementing another element to capture the relation between these adjacent SOURCE entities, which shall make the extraction result more useful.

3.4.2 Pronoun as SOURCE

In some cases, the SOURCE entity is a pronoun, referencing another entity mentioned outside of this sentence. The scope defined for an ER is within a sentence. Therefore should this ER still be considered valid? An example is given in Figure 7.

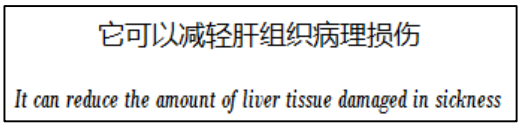

Figure 7: Example - pronoun as SOURCE

In the example in Figure 7, the SOURCE entity is “it” (它). After a correct extraction, the ER will still be meaningless, because the TARGET entity “the amount of liver tissue damaged in sickness” (肝组织病理损伤) can be effected in a “reduction” (减轻) manner, by an unknown SOURCE entity referenced by “it”.

In this project, the above example is not considered as a valid ER, because the entity referenced by the pronoun lies beyond the scope of the sentence where the pronoun resides. Detecting contents referenced by pronoun beyond the sentence level is another task in NLP, and will not be included in this project. However this task will be a promising extension to this research, as several examples of the “pronoun as SOURCE” are found in the dataset.

3.4.3 “Comparing” relation

In some cases, an EFFECT keyword is detected and the sentence does seem to be stating a relation concerning two entities. However it is actually a common attribute of the two entities being compared, resulting in a quantifiable difference expressed by the EFFECT keyword. If this type of relation is to be extracted, a new ER component will need to be defined to capture the common attribute shared by the two entities. Also the entities in this type of relation are usually referencing procedures or other entities outside of the sentence, just as discussed in 3.4.2. An example is given in Figure 8.

Figure 8: Example - "comparing" relation

In the example in Figure 8, “the model group” (模型组) is the SOURCE, “the normal control group” (正常组) is the TARGET, “the amount of TNF-A in the mouse serum” (小鼠血清 TNF-A) is the comparative and “risen” (升高) is the EFFECT of the difference from SOURCE to TARGET. However “model group” and “normal control group” are common terms in medical journal papers to reference patients treated by different procedures. Without knowing these procedures, “model group” and “normal control group” are meaningless as a result.

These “comparing” relations are quite distinctive in structure and usually contain an EFFECT keyword. They are not included in the current definition of ER. However they are considered a potential extension of the ER definition, and are discussed further in Appendix 1.

3.5 Summary

This chapter provided an insight into the overall information related to an Effect Relation; from the abstract concept, to a concrete definition and practise of annotation in text. It also discussed the patterns and ambiguous cases found during the course of the annotation.

The concept of an ER is to identify the positive or negative relationship between two entities. A definition is developed,

(1) $$ER = \ < S, T, E >$$

stating an ER must consist of a SOURCE, a TARGET and an EFFECT. Over the course of annotation, another ER entity: EFFECT TARGET is introduced, due to the manner words are segmented by the word parser.

Many patterns were found in the ER annotations, such as most words in SOURCE and TARGET belongs to a few distinctive categories and can be anticipated by using external databases. The words in the ER annotations are highly repeated, and the position of the ER entities to each other in an ER follows a strong pattern too.

Many ambiguous cases were also found over the course of annotation, such as conditional joining, pronouns used as SOURCE and relation defined in a comparison statement. These ambiguous cases made the course of annotation difficult and caused many revisions. However they also helped the definition of ER to be observed from different perspectives to arrive at the stage it is today.

Chapter 4. Relation extraction approaches

This chapter describes the examination of two approaches: rule-based and feature-based, to develop respective methods to extract Effect Relations from the dataset. It explains how methods such as dictionary lookup and template matching were selected for the rule-based approach. It explores how these methods take advantage of the characteristics of ERs and their limitations in using contextual information. It further discusses how a feature-based approach is developed to tend to such limitation.

4.1 Rule-based approach

As Sharma et al. claims in [12], there are four main approaches in use for the task of extracting relationship from unstructured text. They are namely, the co-occurrence based, the link-based, the rule-based and the machine learning approaches. The prior three approaches are closely related, as they all operate with predetermined assumptions or rules. An ER can be understood as the co-occurrence of two entities in the same sentence-level text. However, the dataset used in this research project is relatively small (as it is newly constructed); therefore ERs of the same pair of entities are not expected to reoccur often. The co-occurrence approach is not applicable in this research. For similar reasons, the linked approach is not applicable in this research. Effect Relation is not just any relation between two entities, it focuses on the positive or negative effect one entity has on the other. The extraction approach needs to be able to capture the "effect" as well as the entities, and can possibly use their presence in the same sentence to validate each other. The process, of a human trying to decide whether a selection of text is considered as an ER, shares great resemblance as the process of a machine performing the same task using a rule-based approach. As more ERs are annotated, a collection of rules and patterns are established by the annotator to keep the annotations consistent and justify decisions made on ambiguous cases. This collection of rules and patterns formed the foundation of a rule-based approach, thus this approach was chosen as the first approach towards extraction of ERs.

Some notable patterns observed in the dataset include:

(i) Each ER is constructed of a specific set of ER entities: SOURCE, TARGET and EFFECT (further discussed in 6.2.4).

(ii) Majority of the Chinese words (see "Chinese word" in Definitions) in ER entities can be categorised into a specific domain, e.g. "TCM ingredients" and "symptoms" (further discussed in 3.3.1).

(iii) Chinese words in ER entities shows to have high repetitiveness (further discussed in 3.3.2).

The rule-based approach consists of the dictionary lookup method and template matching method to draw on these patterns observed in the dataset.

4.1.1 Raw text information and text parsing information

As mentioned in 2.1.2, text parsing is an important step in the pre-processing stage of most RE methodologies. Text parsing can provide additional information, such as Chinese word segmentation, POS tags and dependency tree tags, to the previously raw dataset, i.e. only raw text. However, this step runs the risk of introducing errors into the dataset when performed by automated systems with manual correction (such errors further discussed in 6.2.6.4). The dataset in this research project is constructed from raw text and the automated text parsing system, Language Technology Platform (further discussed in 6.2.3) is used. Manual correction of the text parsing information is limited due to time constraint, therefore the presence of errors are expected in the text parsing information of the dataset. Upon acknowledgement of these errors, the rule-based approach, which usually targets a small range of information, is designed to utilize only raw text information; and the feature-based approach, which usually relies on a wider range of information as features, is designed to utilize the text parsing information, despite the possible errors.

4.1.2 Dictionary Lookup method

The dictionary lookup method is a simple rule-based method to identify entities of interest, and is used in many NER modules, such as BioAnnotator [21]. It consists of building dictionaries of word in different domains and detecting the presence of these dictionary words in the dataset. Domain refers to the categorisation of the word content, e.g. “disease” and “TCM ingredients”. Its domains of dictionary words directly correspond to the domains of which words are categorised in Pattern (ii) above. Identifying domains of dictionaries allows this method to anticipate unknown words. The dictionary words in the dictionary lookup method are initially collected from the dataset, and then extended by external sources. For example, the dictionary of the “effect” domain is first populated by “seed” words identified as “effect” in the dataset. It is then extended by performing synonym search on these “seed” words in an external Chinese dictionary (汉典 [51]). When the dataset is expanded and new “effect” words are detected, many words are already included in the “effect” domain dictionary. The repeated use of words in the ER entities identified in Pattern (iii) is an indication that the words collected in the dictionaries will repeatedly come into use in the dataset and that the dictionary lookup method is a scalable method when the dataset is expanded in the future.

The dictionary lookup method, however, is limited in terms of validation. The presence of dictionary words does not necessarily indicate ER entity. For example, most SOURCE entities are in the dictionary of “TCM ingredients” domain. However, from the perspective of the entire dataset, only a minority of the presence of TCM ingredient is a SOURCE entity (i.e. part of an ER), the majority are non-ER-related mentioning. There are many occasions for TCM journal publications to mention TCM ingredients without having to relate to an ER, such as “what ingredients are used” and “how much ingredients are used”. This limitation is worsened in cases of 1-char dictionary words (see “n-char words” in Definitions). In cases of 1-char dictionary words, such as “养” *enrich/support* and “清” *clear*, they individually have usage in non-RE-related context, and they are often used to form longer words which also have usage in non-RE-related context. Details to how the dictionary lookup method is implemented are further discussed in 6.3.1.

4.1.3 Template matching method

The template matching method is a rule-based method used often in both the NER module and SRE module, such as [40]. The template matching method consists of setting criterions to validate whether the input content is following the desired template, and setting rules to extract contents in accordance to the desired template. This method is befitting to attend to the limitation of validation in the dictionary lookup method, by remove presence of dictionary words that do not follow the template of an ER. For example, the basic requirement of an ER is to have a SOURCE, a TARGET and an EFFECT. Any presence of dictionary words that do not have enough entities in the sentence to this requirement, can be deemed as invalid and be discarded from further processing. Other criterions, such as validation according to the order and positioning of the potential ER entities, are also included in the template matching method to validate potential ER entities. The previous dictionary lookup method emphasizes on the retrieval of potential ER entities and this template matching method emphasizes on validation of these ER entities and, in turn, potential ERs. The combination of these two methods is expected to achieve a good balance of precision and recall in performance. Details to how template matching method is implemented are further discussed in 6.3.3.

4.1.4 Scalability and adaptability

A crucial weakness of the rule-based approach is its lack of scalability and adaptability. For example it has difficulty recognising new cases, i.e. entity words not included in the lookup dictionaries or ERs in different forms to the template. Reliance on the precisely targeted range of information (for example, [15] targets the dependency tree build its rules) is a double-edged sword, allowing the approach to be highly accurate when most variation are known but baffled when encountering unknown cases. In other words, the rule-based approach can perform extremely well if all ERs are uniform and strictly following the rules and criterions in the methods. However this is rarely the case in unstructured text, where the same contents can be expressed in various ways depending on factors such as the habits of the author and formality of the text document. Inevitably, new words and new forms of ER will be introduced to the growing dataset. Improvement on the scalability of rule-based approach requires higher complexity of rules. High complexity of rules means any further modification may cause possible conflicts between rules, and this will eventually lead to undesirably high amount of effort on the maintenance. A machine learning approach, on the other hand, may be able to achieve desired scalability without such high level of maintenance, and thus a feature-based approach is introduced.

4.2 Feature-based approach

The rule-based approach is very precise in its range of targeted information because high complexity of rules leads to high level of maintenance. A machine learning approach, however, can utilize a wide range of information and handle the high complexity with mathematical classification models or algorithms. Therefore the feature-based approach introduced in this research will utilise a wider range of information, i.e. contextual information.

4.2.1 Using contextual information

Contextual information refers the information around a targeted piece of text other than the text itself. For example, contextual information of a text token may be whether it is a noun or verb or information about its neighbouring text tokens; and contextual information of a sentence may be which section of the document (if the document is structured in some way) it resides in. In this research, contextual information is main focused at a text token level, and includes text parsing information such as POS and dependency tree tags. The feature-based approach utilises contextual information, forming them into features, and trains on a selection of the dataset (training dataset) using a mathematical classification algorithm to obtain the model of what an ER is in terms of features. As mentioned in 4.1.1, possible errors may exist in the text parsing information due to automated text parsing systems. This approach aims to handle the possibility of errors and unknown cases (not encountered in the training dataset), by viewing the decision over a collaboration of different features. It will still be impacted by the errors and unknown cases. However, by using a range of features to diffuse the weight of the decision, it allows the influence of a wrong decision made by a minority of wrong/unknown features to be outweighed by the influence of a correct decision made by a majority of correct/known features. Details to what features are employed in the feature-based methodology are further discussed in 6.5.2 and 6.6.

4.2.2 Building upon rule-based approach

The feature-based approach is not meant to be a completely different and separate approach to the rule-based approach. It is brought into this research to overcome the weaknesses of the rule-based approach, but also extend on its strength. Other than the text parsing information features, the feature-based approach will also include features more specific to ER extraction. The rule-based approach has its strength in RE on the known dataset and the feature-based approach will extend on this strength. Some rules in the rule-based approach will be converted to features. Most of them will be Boolean values such as whether a word is in the lookup dictionaries, or whether the surroundings of the word fit accordingly to the template. These additional features will allow the feature-based approach to be more specific for ER extraction. Details to how rules are converted into features are further discussed in 6.5.2.

4.3 Summary

The rule-based approach targets a small range of information to translate the ERs in the dataset as rules and templates. It is expected to start off leading in the race of performance, as evaluation is performed on the known dataset. However it slows down over the course of the race, i.e. the expansion of dataset and unknown cases are introduced, due to its weakness in scalability. The feature-based approach utilises a wide range of contextual information. It is expected to start off behind in the race of performance as contextual information are less indicative and there may be errors in the contextual information. However the contextual information will become more indicative as the dataset is refined and expanded. It is difficult to argue which approach will win the race, as there will always be room for both approaches to be improved, but it is fair to say the performances of rule-

based approach can provide a good baseline for comparison; and its rules can be extended or made into building blocks for other approaches to build upon.

In addition, over the course of developing both approaches, the originally conceptual semantic relationship that is ER takes more concrete syntactic forms, which in turn, consolidated the definition of ER and clarified many ambiguities.

Chapter 5. Functional words

This chapter describes the background of functional words and why they are brought into this research. It explains the distinctive features of the participation of functional words in text, and how functional words are proposed to be involved in the definition and extraction of Effect Relations.

5.1 Functional words and content words

In the Chinese grammar, words are classified into two groups: 实词 (*content words*) and 虚词 (*functional words*). Content words are concerned with the content of the text, and will differ greatly as the field of text is changed, for example, medical terms in the field of medicine, commercial terms in the field of commerce. Functional words (a.k.a. abstract words) are concerned with the structure of the sentence, and gives instructions to how the sentence should be understood, regardless of the field of content.

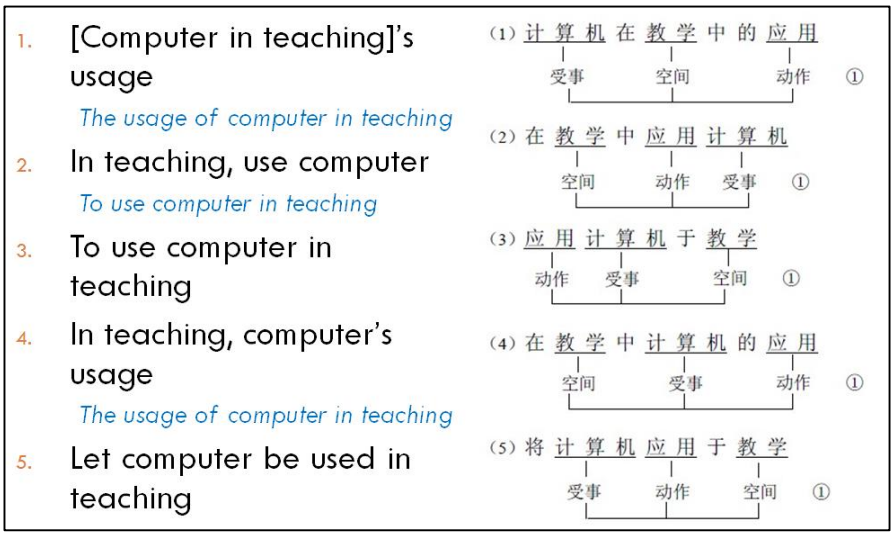

Figure 9: Example - use of functional words [52]

Figure 9 shows an example given in Dong's paper [52]. This is an example the syntactic flexibility enable by the use of functional words. The underlined words are the content words in the sentence: 计算机 *computer*, 教学 *teaching*, and 应用 *usage*. In the 5 examples, the content words remain the same, only their order has been changed. By utilising functional words, the sentences form different interpretations.

5.2 Occurrence and repetitiveness

In Liu's paper [53], he gathered information and performed statistical analysis on the number of words in the Chinese language vocabulary. In this paper, Liu mentioned 7753 1-char words: 46.7% were noun, 31.4% are verb and 12.7% were adjectives. These 3 categories cover most of what is considered in the group of 实词 (*content words*) in the Chinese grammar. What remains in the group of 实词 are 数词 (*number words*), 量词 (*counter words*), 代词 (*pronouns*) and a section of 副词 (*adverb*), the rest are considered as words in the group of functional words. As we can see from the statistics above, out of 7753 1-char words, at least 90.8% are content words, leaving less than 9.2% to be

functional words. Similarly, of the 43097 2-char words, 51.1% are noun, 36.4% are verb and 7.6% are adjective, leaving less than 4.9% to be functional words. The statistics above demonstrates that functional words occupy an extremely small portion of the Chinese vocabulary.

Statistical analysis of functional words occurrence (based on 835 functional word entries in [54]) is performed on the TCM dataset constructed in this research project (further discussed in 0). The result shows, from 2023 sentences, 1902 sentences were found to contain at least one functional word (94.02%). From 59852 words, 19196 words were found to be functional words (32.07%). These results indicate that functional words play a significant role in terms of occurrence in the dataset.

Further analysis is performed to examine the distribution of the most repeat text tokens in the dataset. A sample of the top 100 most repeated text tokens in the dataset are selected and categorised in terms of their syntactic role and their usage in the ER extraction methodologies. Figure 10 below shows the distribution of these categories.

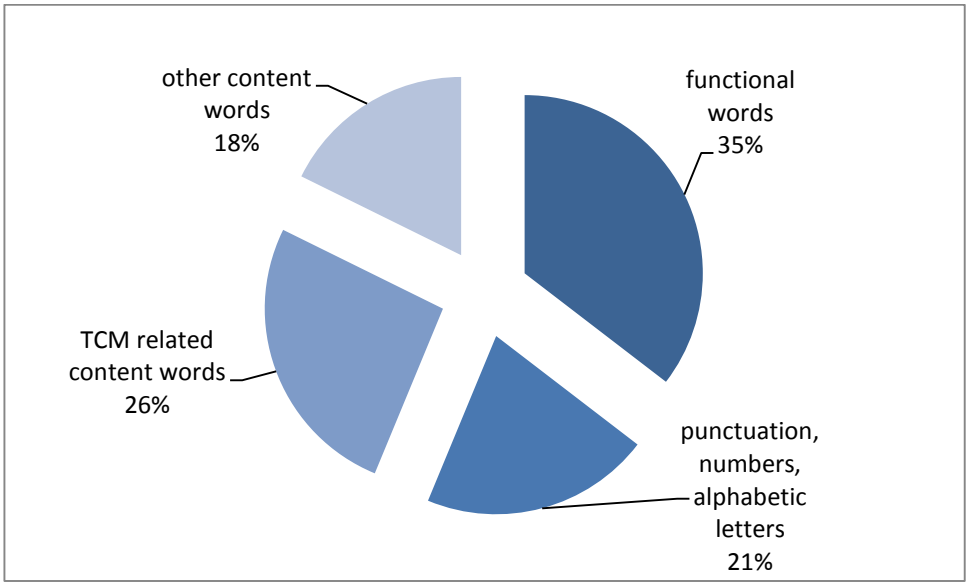

Figure 10: Top 100 most repeated text token category distribution

In the categories in Figure 10, punctuation, numbers and alphabetic letters (21%) are used as features in the feature-based methodology (further discussed in 6.5 and 6.6). TCM related content words (26%) are used in the dictionary lookup method (further discussed in 6.3.1). The remaining portions, i.e. functional words (35%) and other content words (18%), are not used in the RE extraction methodologies.

Results of the statistical analyses above show that despite the minor portion functional words takes up in the Chinese vocabulary, they play a significant role in terms of occurrence in text, in the Chinese language, in this case, the TCM dataset used in this research. In other words, compared to content words, functional words are limited in vocabulary and frequent in usage. They show great potential as resources for the RE methodologies.

5.3 Using functional words

Functional words are very crucial to the understanding of the contemporary Chinese language. However its limited vocabulary and frequent usage creates complexity in understanding how they exactly function and develops more than one definition and way of usage. For example, Zan discussed the 7 definitions and 21 ways of usage of the word JIU (“就”), in her paper [47]. It is impractical to design rules of high complexity to determine the exact definition and way of usage of functional words in the dataset, but a selection of such rules may be beneficial to the existing methodologies.

Functional words give indications of their surroundings. The range and type of indication they give depend on the functional word. A “5C categorization” system was used at an early stage of this research to categorize ER according to the type of functional word related to this ER (refer to Appendix 1 for more information). This system was eventually discarded a significant increase in numbers of ERs without relation functional word. However this system inspired detailed observations of how functional words relate to the entities in ERs, which can be useful in identifying potential ER entities from the dataset, i.e. the NER module.

Functional words like: “的” and “之” (of) usually indicate a connection between its neighbours, and that by combining with its neighbours; it will form a new noun concept (object / action / concept).

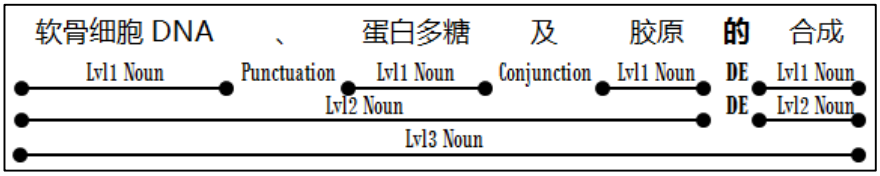

Figure 11: Example – usage of DE

For example in Figure 11, “软骨细胞 DNA”, “蛋白多糖”, “胶原” before “的” and “合成” after “的” are the basic nouns (lvl1 noun) in the clause. “软骨细胞 DNA”, “蛋白多糖” and “胶原” form a bigger noun concept (lvl2 noun) since they are joined by slight-pause mark “、” and a conjunction “及”. The two noun concepts on either side of “的” can be combined with “的” to form a new noun concept (lvl3 noun).

Meanwhile, functional words like: “能够” and “可以” usually indicate that there is a noun concept in the text tokens “shortly” before it, and a verb + noun concept lies in the text tokens coming “shortly” after it. However the exact range of “shortly” may vary according to other text tokens in the sentence.

Figure 12: Example – usage of NENGGOU

For example in Figure 12, “石韦” is the noun concept before “能够”, “养阴” and “补肾” are the two sets of verb + noun concepts “shortly” after “能够”.

Other functional words may interfere with this pattern. For example, if functional words like: “使” and “把” appears, it will switch the order of the verb and noun concept.

Figure 13: Example – usage of SHI

For example in Figure 13, when “使” is used after “能够”, the verb + noun concept is transformed to a noun + verb (and action) concept: “局部血液濡养”, “得到”, “改善”.

In addition, some content words are found to act like functional words, where they do not hold specific content by themselves, instead they help facilitate other content words. These words often work in pairs, such as “具有…的作用” (have … effect), and the content they are facilitating, i.e. the verb + noun concept, lies in the text tokens between the pair.

Figure 14: Example – usage of JUYOU

For example in Figure 14, 2 verb + noun concepts: “改善” + “肾虚证候” and “支持” + “黄体功能” lies between “具有” and “的作用”.

Above observations show certain functional words can provide strong indication on the whereabouts of the SOURCE, TARGET and EFFECT entities in relation to their own position; but due to other factors in the sentence, often they cannot pinpoint the exact location and length of these entities. Under most circumstances, functional words will not seem helpful to the existing rule-based methodology. The dictionary lookup method can pinpoint the exact location and length of the potential ER entities based on text token, and the template matching method can validate these potential ER entities against sentence level criterions. This, however, is not true for a selection of ER entities, namely, complex TARGET entities.

As discussed in 3.3.1, although most TARGET entities in the dataset follow a simple structure (consists of only one word), there is a small portion that are complex structured, e.g. “软骨细胞 DNA、蛋白多糖及胶原的合成” in Figure 11. These complex structured TARGET entities vary greatly in length and form. The current dictionary lookup and template matching methods are not capable of identifying these complex structured TARGET entities, while rules based on functional words are less affected by the various forms and length of the ER entity. Therefore by employing these functional word rules in addition to the existing methods, it improves the capability of identifying complex structured TARGET entities of the rule-based methodology. Similarly, these functional word rules can be expressed as Boolean values and added as features in the classification model

of the feature-based methodology to improve the detection of potential ER entities in the NER module.

5.4 Summary

This chapter explores the importance of functional words in understanding Chinese syntax. Statistical results show the usage of functional words in the general Chinese language and the TCM dataset used in this research. Rules in regards of functional words are mainly concerned with the NER module, as they can help identify ER entities in the dataset. Examples are given to how functional words can give indications in regards to the ER entities nearby. The benefits of using functional word rules lie mostly in identifying complex structured TARGET entities, which the rule-based methodology is initially unable to do. These rules can also be converted to features and help with the NER module of the feature-based methodology.

Chapter 6. Design

This chapter describes the design of how the rule-based methodology and feature-based methodology will be implemented respectively to obtain the final result. It explains details of the entire process, from preparing the dataset for these methodologies to perform on; to the methods and stages within these methodologies.

6.1 Overview

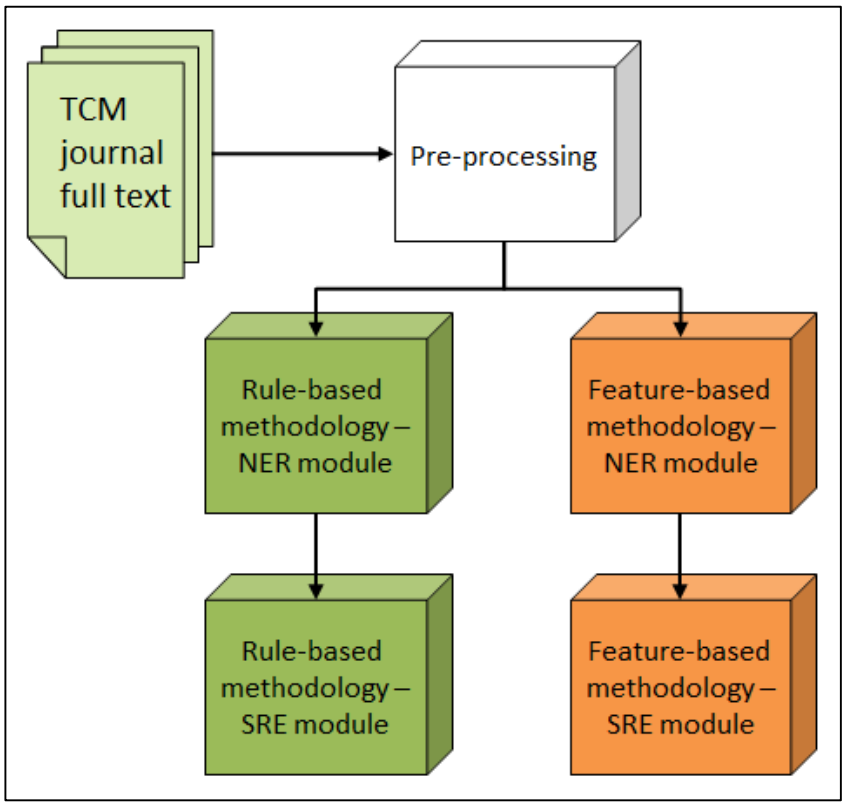

Figure 15: Architecture of the rule-based and feature-based methodologies

The design of the two ER extraction methodologies in this thesis is displayed in Figure 15. This design follows a similar sequence of processes as many other methodologies. This design have proven effective in extraction of generic relationships, such as in [55], and specific relationships, such as in [12]. It starts with raw text, in this case, full text extracted from a selected set of TCM journal publications. Pre-processing is performed (further discussed in 0), transforming raw text into a dataset. ER extraction is then performed on this common dataset using the rule-based methodology and the feature-based methodology. Both methodologies comprise two modules: the Named Entity Recognition module and the Semantic Relationship Extraction module (Figure 16).

In many RE methodologies, the NER module is performed with the help of external tools. For example [12] utilised external tools such as WordNet [56] and VerbNet [57] to identify generic named entities in the text. The entities required in this research are quite different from the generic entities identified by conventional NER tools. Many ER entities are terms specific to the TCM domain. In many cases, a noun entity is embedded as a part of what conventionally be considered a verb term (further discussed in 6.2.4). Therefore the NER

modules in the two methodologies are carried out by methods defined in this thesis. As ER is a specific relationship defined in this research, the procedures within the NER modules and SRE modules have been modified to suit the characteristics of ER. The procedures within the NER modules and SRE modules are displayed in Figure 16.

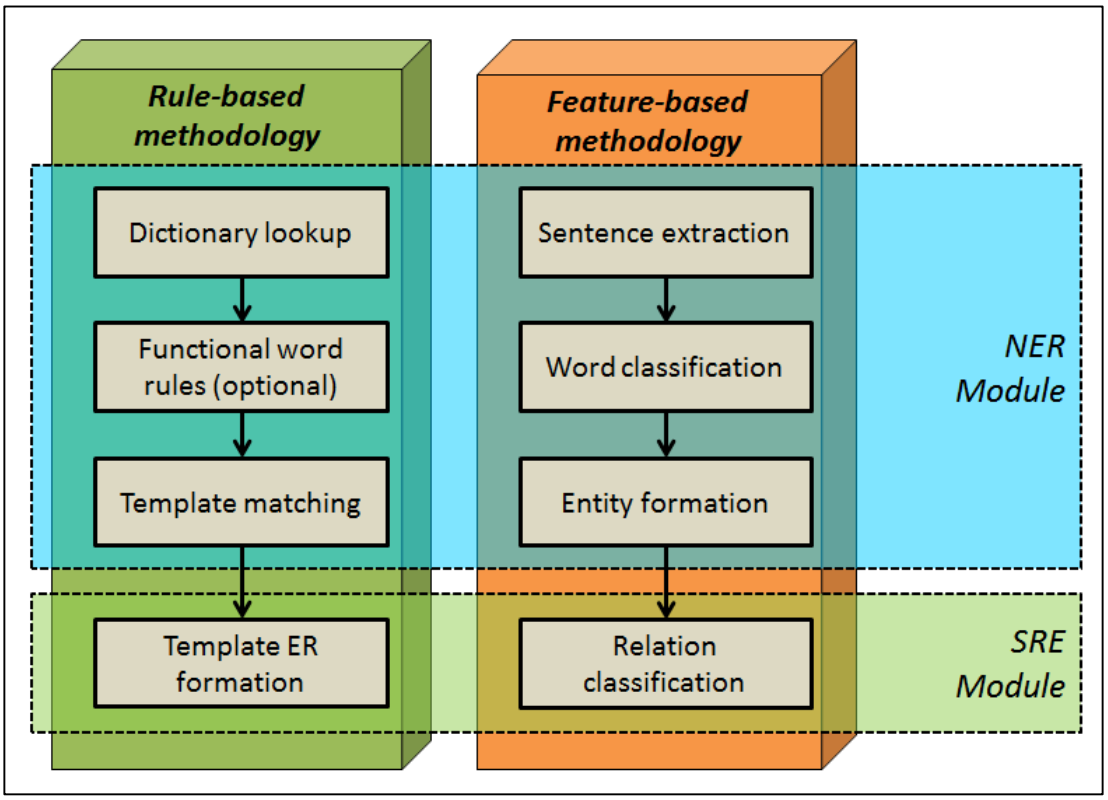

Figure 16: Methods in modules of the rule-based and feature-based methodologies

As depicted in Figure 16, the NER module and SRE module in the two methodologies are achieved with different methods.

The NER module will identify entities in the dataset that have the potential to form an ER. The NER module of the rule-based methodology aims to achieve this via three methods: dictionary lookup, functional words rules and template matching (further discussed in 6.3); and the feature-based methodology aims to achieve this via three methods: sentence extraction, word classification, entity formation (further discussed in 6.5).

The SRE module will use the results of the NER module and determine what ERs are formed by the entities identified in the NER module. The SRE module of the rule-based methodology aims to achieve this using the template ER formation method (further discussed in 6.4); and the feature-based methodology aims to achieve this using the relation classification method (further discussed in 6.6).

6.2 Pre-processing: preparing the dataset

Effect Relation is a type of relationship introduced in this research. The pre-processing stage will construct a new dataset from raw text.

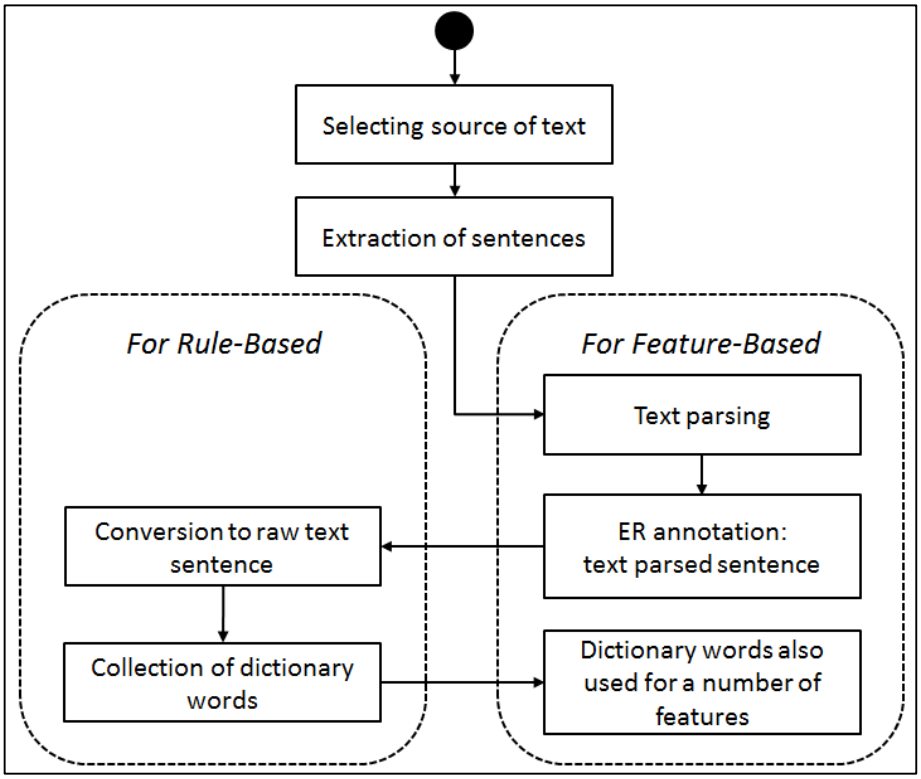

Figure 17: Pre-processing procedures

As shown in Figure 17, the pre-processing procedures consists of steps: selecting source of text, extraction of sentences, text parsing, ER annotations, collection of dictionary words. ER annotation is initially done on the text parsed sentence for the feature-based methodology, and then converted to raw text sentences for the rule-based methodology (further discussed in 6.2.4). The pre-processing stage also deals with errors in the dataset detected along the procedures.

6.2.1 Selecting source of text

For the ERs resulting from the extraction method to be meaningful, we must first assume the contexts that these ERs were extracted from are formal, factual and conclusive statements. This is due to be fact that extraction in this proposal looks mostly into the structural context and little in the semantic context. When an ER is extracted, it is assumed that the author of the text was not writing this piece of text as an assumption, prediction, question or other non-conclusive manner.

For this reason, commonly used datasets of unstructured Chinese text, such as the popular Xinhua news articles corpus, are not suitable for ER extraction. Scientific journal publication fit the requirement as they are unstructured text and written as in a formal, factual and conclusive manner. Traditional Chinese medicine (TCM) was selected as the case study as it potentially contains a large number of ERs defined between TCM ingredients or

prescriptions, and human body parts or symptoms. Therefore a new dataset is to be constructed based on TCM journal publications.

The TCM journal publications used as case study are relatively more "structured" than text type, such as newspaper articles or blogs. These journal publications consist of standard journal style section headings such as "结果" *Results* and "讨论" *Discussion* to organise the contents of the document, however these headings are not consistent over different publications and are not explicitly marked, making it relatively difficult to automatically identify the document sections in the publications. The sentences in the publications are generally full sentences and grammatically correct, however the structure of sentences and paragraphs in these journal publications are not standardised and may vary significantly from author to author. Overall the TCM journal publications used for the dataset are considered as "unstructured text".

Initially 551 pieces of TCM journal publication, covering a wide range of disease treated and a wide range of journals were collected. Due to the time constrain of this research project, not all journal publications can be annotated. Ten pieces of journal papers were randomly selected from the category of treating diabetes, because a significant portion of these journal papers was based on treating diabetes (274 out of 551). Upon completion of ER annotation (further discussed in 6.2.4), one publication was randomly selected from each of the other categories, resulting with a total of 31 pieces of TCM journal publications being used in the final version of the dataset in this research project. These publications cover a wide range of diseases, such as diabetes, rhinitis and rheumatism (details in Appendix 1). They also come from a wide range of journals, such as Clinical Journal of Guang Ming TCM, Medical Innovation of China and Beijing Journal of Traditional Chinese Medicine (details in Appendix 3).

6.2.2 Extraction of sentences

Extraction of raw text from the selected journal publication results with a long string of text tokens for each of the publications. Sentencing is then performed on these long strings of text tokens. The strings of text tokens need to be sentenced because the scope of ER is within one sentence, and methods in the rule-based and feature-based methodologies require the dataset to be organized in terms of sentences. A total of 2023 sentences are extracted from the long strings of text tokens of the TCM journal publications.

6.2.3 Text parsing

The Language Technology Platform (LTP) [41] developed by the Research Center for Social Computing and Information Retrieval at Harbin Institute of Technology (HIT-SCIR) was used in to perform word parsing on the sentenced text. LTP was chosen for its high performance and the numerous word parsing tasks it can perform, such as word segmentation, Part-Of-Speech Tagging (POS), Named Entity Recognition (NER), Dependency Parsing (DP) and Semantic Role Labelling (SRL). A total of 48231 words are obtained from text parsing of the extracted sentences.

6.2.4 ER Annotation

As ER is specifically defined in this research, there are no existing Effect Relation dataset available. The feature-based methodology requires text parsing information, so manual annotation is performed on the text parsed sentences collected from TCM journal publications. Certain rules are set for consistency of the Effect Relation annotation:

I. As mentioned in 3.2, an ER is a combination of 3 entities: SOURCE (S), TARGET (T) and EFFECT (E), (EFFECT type does not need to be explicitly extracted as it is directly dependent on the EFFECT entity).

An ER entity must consist of one or more Chinese words. If more than one word, there must not be any other words within the boundaries of this entity.

II. For consistency and applicability over large text, the text parsing task, including word segmentation is based on the output of the LTP word parser [41].

III. On principle, each ER must have exactly one SOURCE entity, one TARGET entity and one EFFECT entity.

However due to the word segmentation of the LTP parser, there are cases where the EFFECT entity and the TARGET entity are parts of the same Chinese word, e.g. "补血" *enrich the blood* is considered as one word. Therefore a new entity: EFFECT TARGET (ET) is created, to tend to these cases. The revised rules for Effect Relation annotation are:

IV. Each Effect Relation must have exactly one SOURCE entity.

V. Each Effect Relation must have either:
   i. exactly one EFFECT entity and exactly one TARGET entity, or
   ii. exactly one EFFECT TARGET entity
   iii. In other words, an ER must either be <S, E, T> or <S, ET> (not necessary in this order of appearance).

The rule-based methodology does not require text parsing information and can use raw text sentences. Raw text sentences are not affected by the word segmentation issue, so the EFFECT TARGET entities in the text parsed sentences are further annotated into the appropriate EFFECT and TARGET entities in the raw text sentences. A total of 1486 annotations are annotated in the TCM dataset.

6.2.5 Collection of dictionary words

The dictionary lookup method (further discussed in 6.3.1) in the rule-based methodology and the word classification method (further discussed in 6.5.2) in the feature-based methodology require dictionary words. Collection of dictionary words is performed after the annotation, based on the words in the ER entities in the dataset. A selection of words present in ER entities is collected and organised into categories. Some of these categorise were discussed earlier in 3.3.1.

These categories are listed below:

- Categories for SOURCE entities:
    - TCM ingredient
    - Prescription name
    - Acupuncture point
    - Chemical compound
    - General medical term
- Categories for TARGET entities:
    - Symptom
    - Body part
- Categories for EFFECT entities:
    - Positive effect
    - Negative effect

Categorization of the dictionary words allows anticipation of unknown words (as discussed in 4.1.2). A total of 483 dictionary words are collected from the TCM dataset.

6.2.6 Errors in dataset

Upon building the ER dataset from raw text extracted from TCM journal papers, some errors appeared, which could negatively affect the performance of the ER extraction methodologies, if not dealt with.

6.2.6.1 Invalid new line indicator

The scope of an Effect Relation is within a sentence, so the first level of structure of the ER dataset is in sentences. New line indicators are generally used as a new sentence indicator, along with punctuations. However in many of the TCM journal papers, text is broken into two columns of journal paper format and new line indicator is given according to the way they appeared in the document, thereby most new line indicators used are invalid. This problem is resolved by removing all new line indicators, and only using Chinese punctuations like "。" and "；" as indicators for new sentences. However some errors continue to exist. For example, titles and paragraph headings usually do not have ending punctuations, so they are often included into the next sentence. In addition, text extracted from tables and figure captions are wrongly included into sentences too. These errors have to be removed manually.

6.2.6.2 Headers and footers

Many journal papers have headers and footers to include information about the journal or the author. This information is usually wrongly included into the next sentence. To resolve this problem, the first three sentences and last three sentences are removed for each journal paper. However some traces of this error continue to exist for headers and footers not on the first and last page; and require manual removal.

6.2.6.3 Detection of wrong words

Some publications used in the dataset are scanning of image instead of text. The Chinese character recognition module of GATE [58] is used to extract text from these scanned documents. Some documents are not well scanned and the characters returned are incorrect. These incorrect characters lead to errors on further processes like text parsing and dictionary lookup. This type of errors requires manual correction.

6.2.6.4 Error in text parser

As mentioned before (6.2.3), the Language Technology Platform (LTP) [41] is used to perform text parsing on sentenced text. LTP reports have obtained 0.97 of F-score for the Chinese word segmentation task and 0.98 of precision for the POS tagging task [59]. However due to much difference in the nature of the People's Daily dataset [60] which LTP trained upon, and the TCM journal papers used in this research, the performance of LTP has been significantly lowered. From sampling a selection of LTP parsed sentences in the dataset, word segmentation appears to be performing at around 0.90 of F-score and POS tagging at around 0.87. These errors are mostly due to TCM specific vocabulary not being recognition and broken down to form other words; and many quotes of TCM literature that was written in 文言文 *classical Chinese format*, which is significantly different to the vocabulary and grammar of 白话文 *contemporary Chinese format* used in LTP. These errors are manually corrected. However only the errors that occurred in sentences containing ERs are corrected; and only the word segmentation and POS tagging label are corrected, other data labels are not corrected.

Upon observation, the above errors have been manually removed or corrected, but errors may still exist where not observed.

6.3 Rule-based NER module

After the dataset is prepared, the rule-based NER module can be performed. It aims to identify the words in each sentence that form ER entities, which then can form ERs.

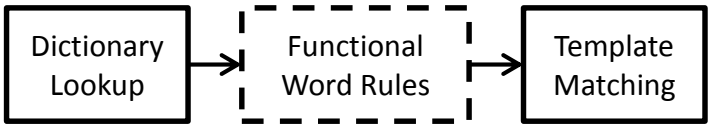

Figure 18: Rule-based NER module methods flowchart

As shown in Figure 18, this module consists of three methods: dictionary lookup, functional word rules and template matching, in the given sequence. The functional word rules method is optional; it can be implemented or set aside to see the effect of involving functional words.

6.3.1 Dictionary lookup

The dictionary lookup method is the first stage in the rule-based NER module. The aim is to identify all words that may be an ER entity or part of an ER entity in the dataset. From Figure 3 and Figure 5 above, we can see that the majority of SOURCE and TARGET words can be categorised into TCM ingredients, body parts, and symptoms. Therefore it is

foreseeable that majority of SOURCE and TARGET can be identified using lists of words from the above categories, forming “dictionaries”. As discussed in 3.3.2, statistically the repeat rate for words in EFFECT is significantly higher than SOURCE and TARGET (Table 2). An explanation for this observation is that EFFECT entities are words indicating either positive or negative change on a quantitative scale, eg. 增加 (*increase*), 降低 (*reduce*); or qualitative scale, e.g 改善 (*improve*), 阻止 (*prevent*). There are many words which share similar meaning to other words in the list, simply taking a slightly different form, e.g. 加, 添加, 增加, 添增 all mean *add*. Unlike another word category such as TCM ingredient, that new TCM ingredients may be mentioned in treating new disease or prescription, the EFFECT category have relatively low potential to include new words as more text is added to the dataset.

Comparatively, SOURCE words are mostly traditional Chinese medicine ingredients (93% from Figure 3 above). New TCM ingredients may be mentioned when treating new disease or prescription. Therefore the list of SOURCE words has a higher potential to grow, and is expected a lower recall in the dictionary lookup for SOURCE entities. However, if this list can be expanded by ingredient names external databases, that can provide a more complete list of TCM ingredient over the whole medical field, this list can be used as a cross-reference to remove more negative results, without the cost of lowering recall.

In terms of identifying TARGET entities, the dictionary lookup method will be targeting only a selection of ER annotations. As discussed in 3.3.1, TARGET consists of simple and complex structure. This step is only expected to attend to the simple structured TARGET entities, which is still a majority (87%).

It is also worth noting that this step does not rely on word segmentation, therefore it would not be affected by the errors of the LTP word parser.

6.3.2 Functional word rules

The functional word rules method is an optional stage after the dictionary lookup method. The aim is to detect additional ER entities. However as most ER entities are of simple structure and are easier to be detected by dictionary lookup, functional word rules would be redundant to detect the same entities. The key target for functional word rules are the complex TARGET entities that cannot be detected by the dictionary lookup stage.

Two lists of functional words are collected: “starting” functional words and “ending” functional words. “Starting” functional words include words like: 包括 *include*, 从而 *thereby*, 具有 *have*. “Ending” functional words include words like: 的作用 *effect of*, 之效 *effect of,* and punctuations.

As “starting” functional words are quite common in the text, functional word rules apply only to the presence of both a “starting” functional word and a SOURCE keyword. Once both conditions are met, this method scans for an “ending” functional word after the TARGET keyword. EFFECT keywords and TARGET keywords are then extracted from the text between the “starting” and “ending” functional words. The remaining text will be matched against a template of positions and number of characters compared to the

extracted EFFECT and TARGET keywords to determine if it is a complex TARGET entity. Any EFFECT keywords form the boundary of the complex TARGET entity and are excluded. TARGET keywords are included in the complex TARGET entity, but a scan for conjunction functional words will be performed to determine if this complex TARGET entity should be broken into parts.

Figure 19: Example - using functional word rules – simple TARGET entities

In the example in Figure 19, “金银花” is a common TCM herb and is detected as a SOURCE entity, while “具有” (in bold) is detected as a “starting” functional word. Conditions are met to scan for an “ending” functional word, which in this case is present as “等功效” (in bold). The content between the “starting” and “ending” functional words are an effectively potential EFFECT TARGET entity. It is scanned for further possible breakdown. Next, “、” are recognised to serve similar purpose as conjunctions to separate EFFECT TARGET entities. At this stage, potential EFFECT TARGET entities are extracted as “抗菌”, “抗病毒” and “解热”. If characters such as “抗” and “解” are collected in the dictionary lookup list as EFFECT words, these will be further broken down as EFFECT and TARGET entity sets. If these words are not yet collected, they will be extracted as EFFECT TARGET entities, awaiting manual annotation after the results are collected. Such functional word pairs are effective for alerting the presence and rough whereabouts in a sentence. This may not seem very useful in detecting simple TARGET entity, as collected dictionary lookup words will also be effective. However, in case of complex TARGET entities, where some or none of words in the entity is recognised by the collected dictionary lookup words, functional word rules will be more effective.

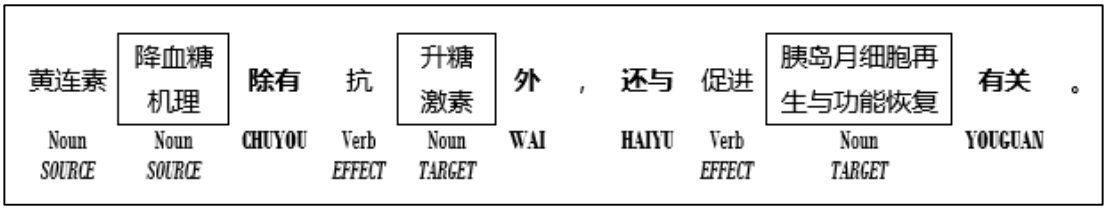

Figure 20: Example - using functional word rules – complex TARGET entities

In the example in Figure 20, the boxed texts are the complex SOURCE/TARGET entities. Complex refers that they consist of multiple words and would not have been collected in the dictionary lookup words as a whole. There are 2 pairs of functional word sets in this sentence. They are marked in bold. The first is “除有” as the “starting” functional word and “外” as the “ending” functional word. The second is “还与” as the “starting” functional word and “有关” as the “ending” functional word. With these 2 pairs of functional word sets, a basic parameter can be set for the 2 potential EFFECT TARGET entities lying within the 2 pairs. With collected dictionary lookup words, the first words “抗” and “促进” can be marked as EFFECT entities and the remaining words will be marked as TARGET entities.

Although these TARGET entities are not recognised in the dictionary lookup words, they are still detected, with the help of functional words setting a range and removal of recognised words. This is a similar case for the SOURCE entity in this sentence, which is a complex SOURCE entity. It should be “黄连素降血糖机理”. However only “黄连素”, which is a common chemical substance used in TCM, will be recognised by the dictionary lookup words. The rest will be either ignored or mistaken for another EFFECT TARGET entity. However the “starting” functional word indicated there is not a breakage prior to it in the sentence and that the recognised SOURCE entity and unknown words between the SOURCE and itself should be considered as part of the SOURCE entity. In this case, the functional word rules allow a complex SOURCE entity to be detected.

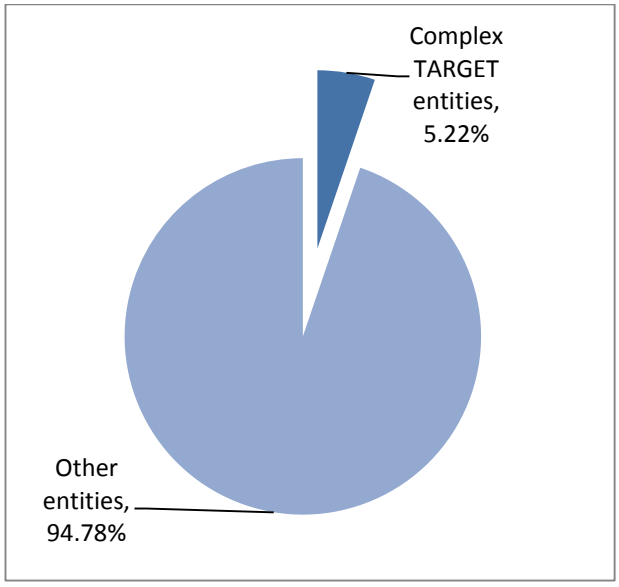

Figure 21: Distribution of complex TARGET entities in all ER entities

Complex TARGET entities make up a small portion in all the TARGET entities, 20.96% as shown in Figure 4; and even smaller portion in all the ER entities, 5.22% as shown in Figure 21. This means even if functional word rules are successful in detecting complex TARGET entities in the NER module, there will not be significant improvement in the overall performance of the module. Although the portion is small, but its number will grow as the dataset is expanded in the future, and functional word rules will come in handy if it proofs its effectiveness in detecting complex TARGET entities. Therefore, in addition to the performance of the module, the result of the rule-based NER module will also be analysed for the increase of complex TARGET entities found by using functional word rules.

6.3.3 Template matching

The template matching method is the last stage in the rule-based NER module. The aim is to validate the potential ER entities identified in the previous stage (dictionary lookup or functional word rules) with a set of sentence level criterions ER entities need to meet. The dictionary matching method is expected to achieve high recall, for its capability to cover most of the words involved, however a low precision is also anticipated. This is because dictionary matching is detecting the presence of words, but presence does not necessarily mean existence of ER. Many words used in the dictionary lookup are one character words

and may be used quite commonly under different meanings. This will result in false positive detections, lowering precision. Also, the detection of entities is scored against the entities marked by ER annotations. Therefore, words that are recognised by the dictionary lookup words, as SOURCE, TARGET or EFFECT, but are by themselves, should not be considered an entity because there are no other entities present around it to form a valid ER. In other words, the dictionary lookup method only considers the content of individual words, but does not take the context of these words into account.

To resolve the issues above and improve on the precision, a template (a set of criterions) is determined to validate the potential ER entities identified by the dictionary lookup and functional word rules. Any potential ER entities not matching the template are removed from the results.

In each sentence:

(i) There must be at least one SOURCE, one TARGET and one EFFECT entity presence.

(ii) Same entities that are next to each other will be combined into one

(iii) After combination, all EFFECT entity must have a neighbouring TARGET entity, and vice versa.

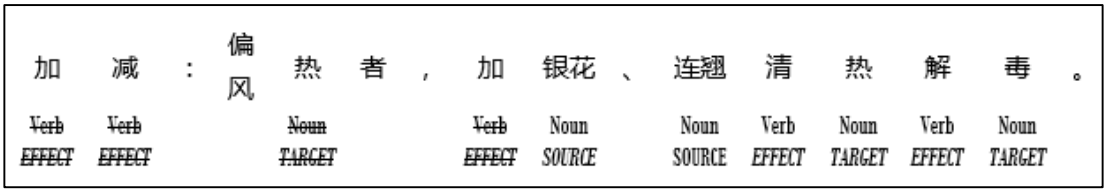

Figure 22: Example – using template matching

In the example in Figure 22, many potential entity words are recognised by the dictionary lookup, however some do not form a valid ER and are therefore incorrectly marked (crossed out). This happens when potential entity words are common in usage and can be used in various situations. In many cases, these incorrectly marked cases are removed using template rule (i), as these occurrences of potential entity word do not have other potential entity words in the sentence to form a valid ER. However, in Figure 22, rule (i) is not effective because there are valid ER's in the sentence. By applying rule (ii), "加" and "减" are recognised as one potential EFFECT entity. By applying rule (iii), "加减", "热" and "加" are removed because these EFFECT and TARGET entities are not neighbouring to each other. "清", "热", "解" and "毒" do fit under rule (iii) and remain marked.

This template is only aimed for detecting simple structure ER entities. Complex entities such as complex TARGET entities are left out for their small portion to all the ER entities; and other methods such as the functional word rules method are used for their detection.

6.4 Rule-based SRE module

The rule-based SRE module consists of one method: template ER formation. It aims to extract ERs formed from given ER entities. This will be done by using the following template:

(i) Neighbouring SOURCE, TARGET, EFFECT entities are combined to form SOURCE, TARGET, EFFECT chunks, respectively. Words that are marked as “None” (meaning not potential ER entities) are left out of the list of chunks. Neighbouring here refers to no other words in between the two entities.

(ii) Neighbouring SOURCE, TARGET, EFFECT chunks are combined to form SOURCE, TARGET, EFFECT clusters, respectively. When distance is the same and have no punctuation in between, after takes priority over prior. Neighbouring here refers to no other entities in between the two chunks.

(iii) For each of the EFFECT chunks in a sentence, it will be linked to the nearest TARGET chunk, forming an EFFECT & TARGET chunk.

(iv) For each of the EFFECT & TARGET chunk in a sentence, it will be linked to each of the SOURCE chunks in the nearest SOURCE cluster prior in the sentence.

(v) A complete ER is formed when it consists of one SOURCE chunk, one TARGET chunk and one EFFECT chunk.

For example, when given the sentence: “经验方中山药、熟地黄、杜仲补益肝肾，强壮筋骨。”

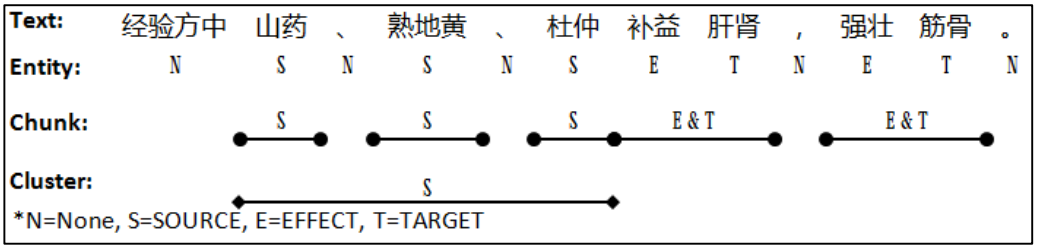

Figure 23: Example – Forming chunk and cluster from ER entities

In Figure 23, SOURCE entities: “山药”, “熟地黄”, “杜仲”; TARGET entities: “肝肾”, “筋骨”; and EFFECT entities: “补益”, ”强壮” are given. By using the template:

(i) They formed their own chunk because there are no adjacent entities of the same type.

(ii) The 3 SOURCE chunks are combined into a SOURCE cluster.

(iii) The 2 EFFECT chunks are linked to its respective nearest TARGET chunks, and are combined to form 2 EFFECT & TARGET chunks.

(iv) Each of the SOURCE chunks in the cluster is linked to the each of EFFECT & TARGET chunks.

(v) A total of 6 ERs are formed and extracted from this example (Table 5).

SOURCE entity	TARGET entity	EFFECT entity
山药	肝肾	补益
熟地黄	肝肾	补益
杜仲	肝肾	补益
山药	筋骨	强壮
熟地黄	筋骨	强壮
杜仲	筋骨	强壮

Table 5: Example ERs formed

6.5 Feature-based NER module

After the dataset is prepared, feature-based NER module can be performed.

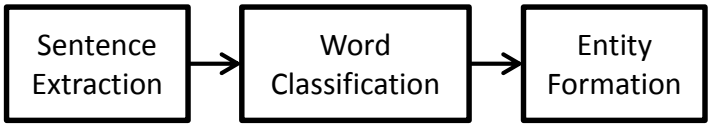

Figure 24: Feature-based NER module method flowchart

As shown in Figure 24, this module consists of 3 methods: sentence extraction, word classification and entity formation, performed in the given sequence.

6.5.1 Sentence extraction

The sentence extraction method is the first stage in the feature-based NER module. The aim is to extract sentences that hold potential to contain ERs from the dataset. This method is a supporting method to the core NER method: word classification. Given a dataset of sentenced text, sentences with occurrence of words from the EFFECT keyword list are extracted. The reason for using the EFFECT keyword list as the criteria is due to high repetition of words in EFFECT entities. There is lower chance of missing sentences due to new words. The results of this stage will be used in the next stage: word extraction (6.5.2). The purpose of this stage is to provide the first line of screening. It will rid the dataset of cases that very obviously do not contain ER. This allows further stages, such as word extraction stage, which has a much more complex procedure, to process on a smaller dataset. This stage will also reduce the effect of the high number of negative cases out balancing low number of positive cases in the sentences. Given that the results will be used further, the stage leans towards higher recall than precision when given a fair trade-off, where negative recalls can be removed by further processes, but negative precision cannot be recovered.

6.5.2 Word classification

The word classification method is the second stage in the feature-based NER module. The aim is to classify words to an ER entity class they have potential to be, or “None” class, meaning they do not hold potential to be part of an ER entity. Given a trimmed dataset from sentence extraction (6.5.1), this stage will look at each word and its neighbours to determine which class this word should belong to. The selection of features to be used will consist of information given by the LTP word parser and results of some rule-based calculations. These features are:

- Text – actual text of the word and neighbours
- POS – Part-Of-Speech tag of the word and neighbours
- Relate – Dependency tree tag of the word and neighbours
- Position – position of the word in relation to the length of the sentence
- EFFECT keywords – whether the word and neighbours contain an EFFECT keyword
- SOURCE keywords – whether the word and neighbours contain a word from categories common for SOURCE keywords
- TARGET keywords – whether the word and neighbours contain a word from categories common for TARGET keywords
- Delimiter keywords – whether the word and neighbours contain a word from categories common for words used as delimiter
- Functional word start window – whether the word and neighbours contain a functional word common to appear before a complex TARGET entity
- Functional word end window – whether the word and neighbours contain a functional word common to appear before a complex TARGET entity

The best threshold of how many neighbours to consider can be determined in the experiment. The results of this stage will be used in the next stage: chunk formation (6.5.3).

The "functional word start" and "function word end" features in the list above are the simplified feature-version of functional word rules in the rule-based methodology. However the scope and complexity of the rules have been reduced from the sentence level to word level of the examined word and its neighbours. Similarly to the functional word rules (6.3.2), in addition to the performance of the module, the result of the feature-based NER module will also be analysed for the increase of complex TARGET entities found by using functional word rules.

6.5.3 Entity formation

The entity formation method is the third stage of the feature-based NER module. The aim is to form entities using the words from the results of the word classification method (6.5.2). This method is a supporting method to the core NER method: word classification. The word classification method outputs a dataset where each word is marked as one of ER entity (SOURCE, TARGET, EFFECT or EFFECT TARGET) or "None". The entity formation method performs "adjacent join" process, where neighbouring entities of the same type are combined. This is due to the fact that any ER entity can consist of one or more words. In order to perform the next method: feature-based relation classification (6.6), complete entities are required, not individual words. Neighbouring words marked with the same ER entity type are combined into one ER entity. These ER entities can further used to form relations later. In addition, after forming the entities, the same template as the rule-based template matching (6.3.3) is applied to each sentence to remove stray words that have been marked as ER entity but not enough to form ER relations.

In each sentence:

(i) There must be at least one SOURCE, one TARGET and one EFFECT entity presence.

(ii) Same entities that are next to each other will be combined into one

(iii) After combination, all EFFECT entity must have a neighbouring TARGET entity, and vice versa.

6.6 Feature-based SRE module

This feature-based SRE module consists of one method: relation classification. It takes a list of combinations of ER entities necessary to form an ER in each sentence as dataset. This method will classify each of the combination items as “is an ER” class or “is not an ER” class. The list of item with the “is an ER” class will be further processed for evaluation against the results of the rule-based methodology. The selection of features to be used will consist of information results of some rule-based calculations.

These features are:

- ER word span – the number of words spanning between first appearing ER entity and last appearing ER entity.
- Entity word distance – the number of words spanning between the three ER entities (3 distances).
- SOURCE in between – the number of other SOURCE entities identified within the span of this potential ER.
- Comma in between – the number of commas within the span of this potential ER.
- Slight-pause mark in between – the number of 、 (slight-pause mark) within the span of this potential ER.

The numbers in the features above are further clustered into 5 categories (using k-means algorithm on numbers collected from the training dataset). The number of category is used as features.

6.7 Summary

This chapter provided details to how methodologies developed in this project are carried out in proper implementation. These details include how the dataset is constructed and prepared for experimentation. This chapter described the two modules shared by the methodologies: NER module and SRE module, and the methods within. It also listed and explained the rules in the rule-based methodology and features in the feature-based methodology.

Chapter 7. Experiments

This chapter describes the experiments that will be performed to evaluate the performance and hypothesis drawn on the methodologies developed. It also explains the background aspects of these experiments such as the source of input for those experiments requiring input, how the dataset will be used and the methods of evaluation used to evaluation the experiments.

7.1 Background

There are some common background setups to consider in carrying out the experiments in this chapter. These setups include: dataset of different ratio of positive and negative cases, source of input when requiring previous results and the methods to evaluate the experiments.

7.1.1 Ratio of cases in dataset

The ratio of positive cases and negatives cases in the dataset will affect the performance of a method, especially when it requires training and testing dataset. In this case, the reality of ERs existence in the text is that negative cases outweigh positive cases by far. For example, within the 31 TCM journal publication used to construct the TCM dataset, a total of 2023 sentences were extracted. The ratio of sentences with one or more ER to those with none is 1:10.12. The highly unbalanced positive / negative ratio in the dataset will impact the performance of methods in the NER modules. Therefore experiments of NER in both methodologies will be tested with both a balanced ratio dataset and an actual ratio dataset.

A balanced ratio dataset allows the evaluation of the performance of this method in a balanced environment. The result of using this dataset will be used as base line. Ideally, the pre-processing stage can be improved to obtain a more balanced dataset. In this research, balanced ratio dataset contains the same number of sentences with at least one ER and sentences without ER.

An actual ratio dataset allows the evaluation of the performance of this method in a realistic environment. The result of using this dataset will be used to compare with the balanced ratio dataset result to see the impact of the outweighing ratio.

7.1.2 Dataset

	Balanced ratio dataset	Actual ratio dataset
Total sentences:	400	2023
Sentences w/ >0 ER:	200	200
Total words:	7742	48231
Words belonging to ER entity:	2431	2431
Total text tokens:	13749	85734
Instances of ER:	1486	1486

Table 6: Details of balanced ratio dataset and actual ratio dataset

Details of the balanced ratio and actual ratio dataset (7.1.1) when mentioned in the experiments are reported in Table 6. The “words” in total words and words belonging to ER

entity are based on the word segmentation performed by LTP (6.2.3) in the pre-processing stage.

7.1.3 Source of input

There are particular experiments that require input, which are results from previous experiments/stages. They will be evaluated on being provided with both golden input and actual input.

Golden input is the result from the previous experiment / stage, which have been configured to be a zero error result. By using this input, it allows the evaluation of performance of this particular stage in the entire method, regardless of the error rate of previous stages.

Actual input is the actual results generated from the previous experiment / stage. By using this input, it allows the evaluation of the performance of the entire method up to this particular stage, with the accumulation of errors from this and previous experiments / stages.

7.1.4 Methods of evaluation

Although the rule-based methodology and feature-based methodology are being tested on the same dataset, there are still some differences in the output of the two methodologies which may cause problems for comparing the two results. A common platform for evaluating both methodologies is explained below.

7.1.4.1 Precision, Recall and F-score

The performance and scoring of experiments in this thesis refer to precision, recall and F-score, unless stated otherwise. Precision and recall are calculated by scores of the standard binary classification matrix:

		Gold standard	
		Positive (P)	**Negative (N)**
Test	**Positive (P')**	True Positive (TP)	False Positive (FP)
	Negative (N')	False Negative (FN)	True Negative (TN)

Table 7: Standard binary classification matrix

(i) $$Precision = \frac{\text{True Positives}}{\text{True Positives+False Positives}}$$

(ii) $$Recall = \frac{\text{True Positives}}{\text{True Positives+False Negatives}}$$

(iii) $$Fscore = 2 \cdot \frac{\text{Precision} \cdot \text{Recall}}{\text{Precision+Recall}}$$

Precision, recall and F-score are calculated according to formula (i) , (ii) and (iii) respectively. In the field of text mining, the ratio of positive cases in the dataset is usually heavily

outweighed by negative cases, but the focus of performance lies more on identifying the positive cases correctly rather than negative cases correctly. Therefore it is standard practise to use precision, recall and F-score for measure of performance, as they disregard true negatives.

7.1.4.2 Evaluation 1: NER module

For the result produced by the methods in the NER module in the two methodologies, the biggest difference lies with word segmentation. As discussed earlier in 6.2.4, there are two versions of ER annotation in the dataset:

Version 1. ER annotation is performed on text parsed sentences

Version 2. ER annotation is performed on raw text sentences.

The rule-based NER module uses Version 1 as input and generates output with minimum word segmentation. That is, it only performs word segmentation for the entities identified; all other characters are kept as individual characters. The feature-based NER module uses Version 2 as input and generates outputs in the same manner. Version 1 contain EFFECT TARGET entities (discussed in 6.2.4), which do not exist in Version 2 (further annotated as separate EFFECT and TARGET entities). In order to create a platform where both methodologies can be evaluated upon, the outputs of both methodologies are converted to characters and evaluated on the character level, i.e.:

(1) All sentences are seen as string of characters, not words

(2) Each character has an ER entity property:
    a. For rule-based: SOURCE, TARGET, EFFECT, or None
    b. For feature-based: SOURCE, TARGET, EFFECT, EFFECT TARGET, or None

(3) The converted outputs are evaluated against the respective answers, where the EFFECT TARGET entity characters are considered correctly identified, if:
    a. In rule-based: the characters are correctly marked as EFFECT or TARGET, according to ER annotation version 2
    b. In feature-based: the characters are correctly marked as EFFECT TARGET, according to ER annotation version 1

7.1.4.3 Evaluation 2: SRE module

The evaluation of the template ER formation method in the rule-bases SRE module and the relation classification method in the feature-based SRE module is scored based on the ERs correctly extracted/classified from the dataset.

		Gold standard	
		ER in the dataset (P)	**Not ER in the dataset (N)**
Test	**ER in the extraction (P')**	True Positive	False Positive
	Not ER in the extraction (N')	False Negative	True Negative (not counted)

Table 8: SRE module evaluation classification matrix

As indicated in Table 8, each extracted ER that is in the dataset, is a true positive; that is not in the dataset is a false positive. Each ER in the dataset that is not extracted is a false negative. Precision, recall and F-score are calculated from these scores (7.1.4.1).

7.1.4.4 10-fold cross validation

When implementing the methods of evaluation above for comparing the two methodologies, the ten-fold cross validation is performed in addition on the results of the feature-based methodology. The given dataset will be separated to 10 equal selections. One selection will be used as the testing set, and the rest used as the training set. The output will be a prediction result, which will generate a score after comparing with the correct result. There will be a rotation through each of the 10 selections being used as the testing set. The 10 prediction results are put together, when needed, as the prediction result of the whole dataset and used as an actual input of another experiment / stage. The 10 scores are averaged to obtain the final score of evaluation. This method of evaluation reduces the chance of over-fitting in the dataset.

7.1.4.5 Evaluation 3: Sentence extraction

The evaluation of the sentence extraction method in the feature-based NER module is scored based on sentences extracted.

		Gold standard	
		Contains >0 ER (P)	**Contains 0 ER (N)**
Test	**Extracted (P')**	True Positive	False Positive
	Not extracted (N')	False Negative	True Negative (disregarded)

Table 9: Sentence Extraction evaluation classification matrix

As indicated in Table 9, each sentence extracted, with one or more ERs, is a true positive; without any ERs is a false positive. Each sentence not extracted with one or more ERs is a false negative; without any ERs is a true negative and disregarded. Precision, recall and F-score are calculated from these scores (7.1.4.1).

7.1.4.6 Evaluation 4: Entity formation

The evaluation of the rule-based entity formation method in the feature-based NER module is scored based on the entities formed.

		Gold standard	
		Is entity (P)	**Not entity (N)**
Test	**Is entity (P')**	True Positive	False Positive
	Not entity (N')	False Negative	True Negative (not counted)

Table 10: Entity Formation evaluation classification matrix

As indicated in Table 10, each entity formed, that is the exact match of the entity in an ER is a true positive; that is not exact match is a false positive. Each ER entity in the dataset unformed is a false negative. Precision, recall and F-score are calculated from these scores (7.1.4.1).

7.2 Overview

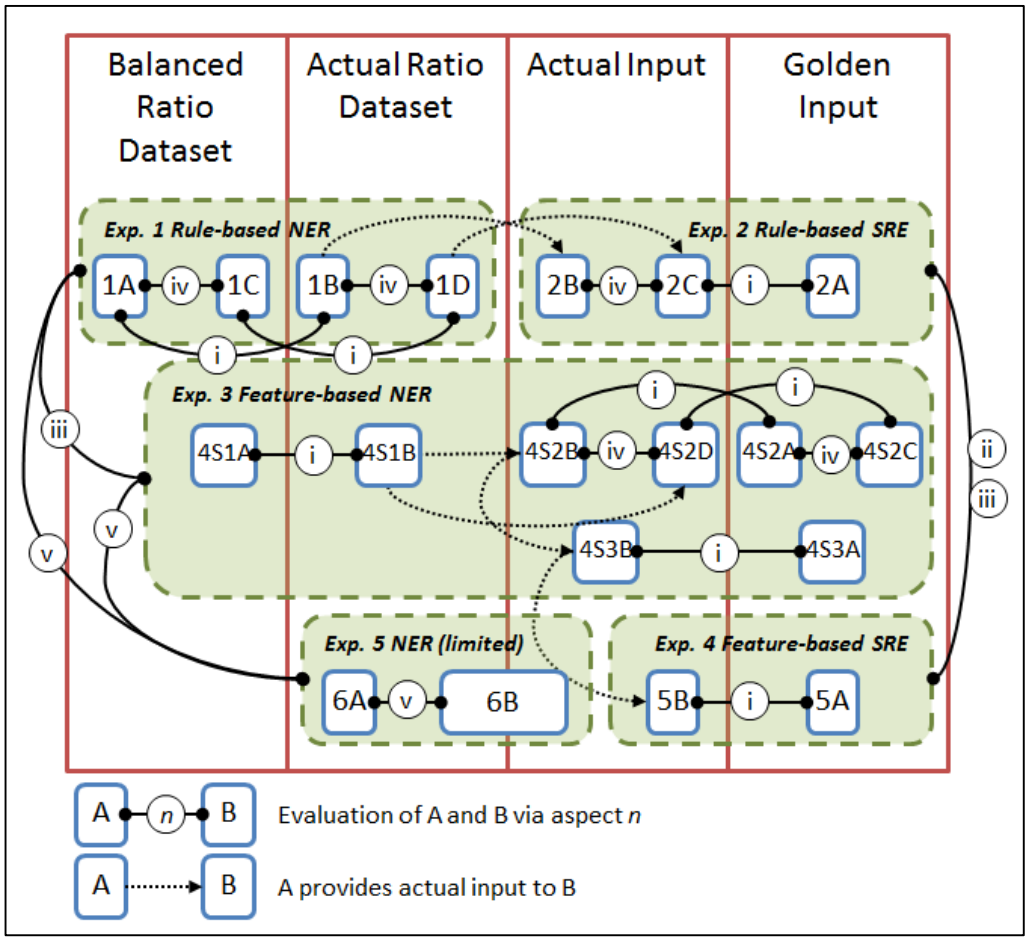

Figure 25: Overview of experiments

Figure 25 is an overview of the experiments reported in this thesis. It depicts the type of dataset each experiment uses and the aspect of evaluation performed on the results of experiments. The aspect number n in Figure 25 refers to the aspects discussed below:

(i) Evaluation of the difference in performance of a stage/module/methodology under ideal dataset and actual dataset – comparing experiments using balanced ratio dataset with actual ratio dataset, or experiments using golden input with actual input

(ii) Comparison of the overall performance of the two methodologies – using actual ratio dataset / actual inputs as dataset for all the stages in each methodology to view the operation of different stages as one continuous process and errors introduced in previous stages will be carried over to the last stage.

(iii) Comparison of the performance of a stage/module in the two methodologies – using balanced ratio dataset / golden inputs to view the operation of different stages as separate processes and errors introduced in previous stages will not affect other stages.

(iv) Evaluation of how effective is functional words in improving the two methodologies – by comparing experiment results involving and not involving functional words.

(v) Evaluation of how well the NER module of the two methodologies can handle unknown words – by comparing experiment results of performing NER module in the two methodologies with limited dictionary words.

7.3 Rule-based methodology

The experiments for the rule-based methodology will test the methods described in rule-based NER module and SRE module.

7.3.1 Experiment 1: Rule-based NER module

The aim of this experiment is to:

- Evaluate the performance of different combinations of methods: dictionary lookup, template matching and function word rules, used in the rule-based NER module
- Evaluate the effectiveness of the functional word rules method in this module
- Identity all text tokens as either one of the ER entity types (SOURCE, TARGET and EFFECT) or as “None” (not an ER entity). This result is to be used as actual input in Experiment 2 (7.3.2)

This experiment consists of 3 methods:

(M1) Dictionary lookup

(M2) Template matching

(M3) Functional word rules

In (M1), the detection of longer words takes priority over shorter words, and earlier presence takes priority over later presence.

This experiment will be carried out in 2 combinations of the methods:

C1. (M1) -> (M2) – to see the performance of dictionary lookup and template matching on its own

C2. (M1) -> (M3) -> (M2) – to see the effect of involving functional words

There will be 2 experiments to test the methods in C1, using different datasets: balanced ratio dataset and actual ratio dataset (7.1.1). The setups of these experiments are given in Table 11 and Table 12 below.

Test Parameter	Value
Method:	Rule-based NER module
Processes:	M1 → M2
Dataset:	Balanced ratio
Test Measured:	Precision, recall, F-score
Evaluation:	Evaluation 1 (7.1.4.1)

Table 11: Experiment 1A Setup

Test Parameter	Value
Method:	Rule-based NER module
Processes:	M1 → M2
Dataset:	Actual ratio
Test Measured:	Precision, recall, F-score
Evaluation:	Evaluation 1 (7.1.4.1)

Table 12: Experiment 1B Setup

The setups of Experiment 1A and 1B will test the performance of the implementation of two rule-based methods: dictionary lookup and template matching. Their results will set the baseline for using the balanced ratio dataset and actual ratio dataset. The difference between their results will also show how much the rule-based NER module is impacted by an unbalanced dataset. These results will be compared against results of Experiment 1C and 1D and the corresponding part of the results of Experiment 4 (7.4.1).

There will be 2 experiments to test the methods in C2, using different datasets: balanced ratio dataset and actual ratio dataset (7.1.1). The setups of these experiments are given in Table 13 and Table 14 below.

Test Parameter	Value
Method:	Rule-based NER module
Processes:	M1 → M3 → M2
Dataset:	Balanced ratio
Test Measured:	Precision, recall, F-score
Evaluation:	Evaluation 1 (7.1.4.1)

Table 13: Experiment 1C Setup

Test Parameter	Value
Method:	Rule-based NER module
Processes:	M1 → M3 → M2
Dataset:	Actual ratio
Test Measured:	Precision, recall, F-score
Evaluation:	Evaluation 1 (7.1.4.1)

Table 14: Experiment 1D Setup

The setups of Experiment 1C and 1D will test the performance of the implementation of two aforementioned rule-based methods with the addition of the functional word rules method. Their results will be compared with the baseline set in Experiment 1A and 1B.

The hypothesis for these two experiments is similar to that of Experiment 1A and 1B, in their comparison to Experiment 3, difference in precision and recall, and difference between the results of Experiment 1C and 1D. In addition, it is also expected that by using functional word rules, there will be an increase in recall, depending on the number of complex TARGET entities in the dataset, which as a result will also raise the F-score.

The results of Experiment 1A, 1B, 1C and 1D will be evaluated on the character level, as mentioned in 7.1.4.1 and compared with the corresponding results in Experiment 3 (7.4.1).

7.3.2 Experiment 2: Rule-based SRE module

The aim of this experiment is to:

- Evaluate the performance of the template used in the rule-based template ER formation method

There will be 3 experiments to test the template ER formation method, using different datasets: golden input, actual input without functional word rules and actual input with functional word rules (FWR) (7.1.2). The setups of these experiments are given in Table 15, Table 16, and Table 17 below.

Test Parameter	Value
Method:	Template ER formation
Dataset:	Golden input
Test Measured:	Precision, recall, F-score
Evaluation:	Evaluation 2 (7.1.4.3)

Table 15: Experiment 2A Setup

Test Parameter	Value
Method:	Template ER formation
Dataset:	Actual input (without FWR)
Test Measured:	Precision, recall, F-score
Evaluation:	Evaluation 2 (7.1.4.3)

Table 16: Experiment 2B Setup

Test Parameter	Value
Method:	Template ER formation
Dataset:	Actual input (with FWR)
Test Measured:	Precision, recall, F-score
Evaluation:	Evaluation 2 (7.1.4.3)

Table 17: Experiment 2C Setup

The setup of Experiment 2A, 2B and 2C will test the performance of template ER formation method with different inputs. Their results will set the baseline for using golden input and actual input.

The hypothesis for these three experiments is that there will be a decrease of performance in the result of using golden input (Experiment 2A) to using actual input (Experiment 2B and 2C), due to any errors carried over from Experiment 1 (7.3.1). It is expected that there will be some improvement from Experiment 2B to Experiment 2C, depending on the magnitude of improvement by using functional word rules in Experiment 1.

The results of these three experiments will be evaluated by the SRE module evaluation method (7.1.4.3) and compared against the corresponding results of Experiment 4 (7.4.2).

7.4 Feature-based methodology

The experiments for the feature-based methodology will test the implementation of methods described in feature-based NER module (6.5) and SRE module (6.6).

7.4.1 Experiment 3: Feature-based NER module

The methods in the feature-based NER module will be separately tested, to evaluate their performance as an individual method and an additional method to a sequence.

7.4.1.1 Stage 1: Sentence extraction

The aim of this stage is to:

- Evaluate the performance of the rules used in the sentence extraction method
- Extract sentences that hold potential to contain ERs, to be used as actual input in Stage 2 (7.4.1.2)

There will be 1 experiment to test the sentence extraction method, using different datasets: balanced ratio dataset and actual ratio dataset (7.1.1). The setups of these experiments are given in Table 18 and Table 19 below.

Test Parameter	Value
Method:	Sentence extraction
Dataset:	balanced ratio dataset
Rules:	dictionary lookup on EFFECT words
Test Measured:	Precision, Recall, F-score
Evaluation:	Evaluation 3 (7.1.4.5)

Table 18: Experiment 3S1 Setup

Test Parameter	Value
Method:	Sentence extraction
Dataset:	actual ratio dataset
Rules:	dictionary lookup on EFFECT words
Test Measured:	Precision, Recall, F-score
Evaluation:	Evaluation 3 (7.1.4.5)

Table 19: Experiment 3S1 Setup

The hypothesis for this experiment is that it will produce a result of high recall but low precision. The list of EFFECT keywords will be adjusted to produce the best F-score.

7.4.1.2 Stage 2: Word classification

The aim of this stage is to:

- Evaluate the performance of the word classification method, using different features and different dataset
- Evaluate the effectiveness of the functional word features in this method
- Classify words to an ER entity class or "None" class, to be used as actual input in Stage 3 (7.4.1.3)

Features used in this stage are listed below:

- Standard Features
  - Text
  - POS
  - Relate
  - Position
  - EFFECT keywords
  - SOURCE keywords
  - TARGET keywords
  - Delimiter keywords

- Functional word features
    - Functional word start window
    - Functional word end window

(Details of these features are explained in 6.5.2)

There will be 2 experiments to test the word classification method with only standard features, using different datasets: golden input and actual input. The setups of these experiments are given in Table 20 and Table 21 below.

Test Parameter	Value
Method:	Word classification
Dataset:	Golden input
Features:	*Standard features listed above*
Test Measured:	Precision, Recall, F-score
Evaluation:	Evaluation 1 + 10-fold (7.1.4.4)

Table 20: Experiment 3S2A Setup

Test Parameter	Value
Method:	Word classification
Dataset:	Actual input
Features:	*Standard features listed above*
Test Measured:	Precision, Recall, F-score
Evaluation:	Evaluation 1 + 10-fold (7.1.4.4)

Table 21: Experiment 3S2B Setup

The setups of Experiment 3S2A and 3S2B will test the performance of using the standard features in the word classification method, for NER module. Their results will set the baseline for using golden input and actual input. These results will be compared against results of Experiment 3S2C and 3S2D.

It is expected that the result of Experiment 3S2A will be better than result of Experiment 3S2B, due to more negative cases in the actual input, from which we can see the extent of how much the negative ratio impacts the performance of this stage.

There will be 2 experiments to test the word classification method with both standard and functional word features (FWF). The setups of these experiments in this stage are given in Table 22 and Table 23 below.

Test Parameter	Value
Method:	Word classification with FWF
Dataset:	Golden input
Features:	*Standard and functional features listed above*
Test Measured:	Precision, Recall, F-score
Evaluation:	Evaluation 1 + 10-fold (7.1.4.4)

Table 22: Experiment 3S2C Setup

Test Parameter	Value
Method:	Word classification with FWF
Dataset:	Actual input
Features:	*Standard and functional features listed above*
Test Measured:	Precision, Recall, F-score
Evaluation:	Evaluation 1 + 10-fold (7.1.4.4)

Table 23: Experiment 3S2D Setup

The setups of Experiment 3S2C and 3S2D will test the performance of using the standard and functional word features in the word classification method, for the NER module. Their results will show the effect of using functional word features by comparing with the results of Experiment 3S2A and 3S2B.

It is expected that the result of Experiment 3S2A will be better than result of Experiment 3S2B, due to more negative cases in the actual input, from which we can see the extent of how much the negative ratio impacts the performance of this stage.

7.4.1.3 Stage 3: Entity formation

The aim of this stage is to:

- Evaluate the performance of rules used in the entity formation method
- Form entities, to be used as actual input in Experiment 4 (7.4.2)

There will be 2 experiments to test the entity formation method, using different dataset: golden input and actual input. The setups of these experiments are given in Table 24 and Table 25 below.

Test Parameter	Value
Method:	Entity formation
Dataset:	Golden input
Rules:	Adjacent join
Test Measured:	Precision, Recall, F-score
Evaluation:	Evaluation 4 (0)

Table 24: Experiment 3S3A Setup

Test Parameter	Value
Method:	Entity formation
Dataset:	Actual input
Rules:	Adjacent join
Test Measured:	Precision, Recall, F-score
Evaluation:	Evaluation 4 (0)

Table 25: Experiment 3S3B Setup

The setups of Experiment 3S3A and 3S3B will test the performance of forming entities from individual words identified as ER entities.

Result from using actual input will be considered as the final score for performance of the feature-based NER module, and will be used to compare with the corresponding results from Experiment 1 (7.3.1).

7.4.2 Experiment 4: Feature-based Semantic Relation Extraction module

The aim of this experiment is to:

- Evaluate the performance of the relation classification method in the feature-based SRE module

The features used are listed below:

- ER word span
- Entity word distance
- SOURCE in between
- Comma in between
- Slight-pause mark in between

(Details of these features explained in 6.6)

There will be 2 experiments to test the relation classification method, using different dataset: golden input and actual input. The setups of these experiments are given in Table 26 and Table 27 below.

Test Parameter	Value
Method:	Relation classification
Dataset:	Golden input
Features:	*Listed above*
Test Measured:	Precision, Recall, F-score
Evaluation:	Evaluation 2 + 10-fold (7.1.4.4)

Table 26: Experiment 4A Setup

Test Parameter	Value
Method:	Relation classification
Dataset:	Actual input
Features:	*Listed above*
Test Measured:	Precision, Recall, F-score
Evaluation:	Evaluation 2 + 10-fold (7.1.4.4)

Table 27: Experiment 4B Setup

Result from using actual input will be considered as the final score for performance of the feature-based relation classification methodology, and will be used to compare with the corresponding results from Experiment 2 (7.3.2).

7.5 Experiment 5: NER module (limited dictionary words)

The aim of this experiment is to:

- Evaluate the performance of the rule-based NER module, when using a limited keyword list for the dictionary lookup method
- Evaluate the performance of the feature-based NER module, when using a limited keyword list for the SOURCE, TARGET and EFFECT keyword features.

Dealing with unknown words is one of the weaknesses of the rule-based methodology, which the feature-based methodology aims to overcome. However in Experiment 1 and Experiment 4 above, both experiments are using the full list of keywords collected from the dataset, which do not demonstrate their performance in dealing with unknown keywords.

In this experiment, 50% of dictionary words is randomly selected and applied in the respective methods of the two methodologies:

- Rule-based NER module – dictionary lookup method
- Feature-based NER module – word classification method features: SOURCE keywords, TARGET keywords, EFFECT keywords

7.5.1 Rule-based NER module

This experiment follows the same processes as Experiment 1, and consists of 3 methods:

(M1) Dictionary lookup

(M2) Template matching

(M3) Functional word rules

There will be 1 experiment to test the methods in this module, using the actual ratio dataset (7.1.1). The setup of this experiment is given in Table 28 below.

Test Parameter	Value
Method:	Rule-based NER module
Processes:	M1 → M3 → M2
Dataset:	Actual ratio
Test Measured:	Precision, recall, F-score
Evaluation:	Evaluation 1 (7.1.4.1)

Table 28: Experiment 6A Setup

The setup of Experiment 6A will test the combined performance of M1, M3 and M2, while using the limited dictionary words for M1. Its result will be compared against result of Experiment 6A (0).

7.5.2 Feature-based NER module

This experiment follows the same processes as Experiment 3, and consists of 3 methods:

(M1) Sentence extraction

(M2) Word classification

(M3) Entity formation

Features used in M2 are listed below:

- Standard Features
  - Text
  - POS
  - Relate
  - Position
  - EFFECT keywords
  - SOURCE keywords
  - TARGET keywords
  - Delimiter keywords

(Details of these features are explained in 6.5.2)

There will be 1 experiment to test the methods in this module with only standard features, using the actual ratio dataset for M1 and actual input for M2 and M3. The setup of this experiment is given in Table 29 below.

Test Parameter	Value
Method:	M1 → M2 → M3
Dataset:	Actual ratio dataset + actual input
Features:	*Standard features listed above*
Test Measured:	Precision, Recall, F-score
Evaluation:	Evaluation 1 + 10-fold (7.1.4.4)

Table 29: Experiment 6B Setup

The setup of Experiment 6B will test the combined performance of M1, M2 and M3, while using the limited keyword list for M2. This result will be compared against result of Experiment 6A (7.5.1).

The hypothesis for these two experiments is that the result of Experiment 6B should be better than result of Experiment 6B, as the feature-based methodology is expected to perform better than the rule-based methodology in dealing with unknown words.

7.6 Summary

This chapter outlined the setups of experiments performed in this research and the rationales in choosing these setups. The common background aspects for performing these experiments are clarified in detail. In each experiment, the aim of the experiment and

further use of the results are stated. A hypothesis to the likely outcome is also explained for some of the experiments.

Experiment 1 and 2 are related to the modules in the rule-based methodology.

Experiment 1 aims to evaluate the performance of different combinations of methods used in the rule-based NER module. The differently combined methods will also evaluate the effectiveness of the functional word rules method in this module. The result will be compared with the result of corresponding NER module in the feature-based methodology (Experiment 4). The result is also used as actual input for Experiment 2.

Experiment 2 aims to evaluate the performance of the template used in the rule-based template ER formation method. The result will be compared with result of the feature-based SRE module (Experiment 4).

Experiment 3 and 4 are related to the modules in the feature-based methodology.

Experiment 3 aims to evaluate the performance of the methods used in the feature-based NER module. The performance will be evaluated in both the aspect of method performance without errors from previous results, and the aspect of accumulated performance with errors from previous results. The result will be compared with the result of corresponding NER module in the rule-based methodology (Experiment 1). The result is also used as actual input for Experiment 4.

Experiment 4 aims to evaluate the performance of the methods used in the feature-based SRE module. The performance will be evaluated in both the aspect of method performance without errors from previous results, and the aspect of accumulated performance with errors from previous results. The result will be compared with the result of corresponding NER module in the rule-based methodology (Experiment 2).

Experiment 5 to compare the performance of the NER module of both methodologies and evaluate how well each performs in dealing with unknown words. Experiment 5 is carried out with the setups of the relevant parts of Experiment 1 and Experiment 3.

Chapter 8. Results

This chapter reports the results of the experiments performed using the rule-based methodology and feature-based methodology, on different source of input, different dataset, and different rules and features of the respective methodology.

8.1 Rule-based methodology

8.1.1 Experiment 1: Rule-based NER module

The precision, recall and F-score of the rule-based NER module on the balanced ratio dataset (Experiment 1A) is reported in Table 30.

	Exp. 1A Balanced ratio dataset		
	Precision	Recall	F-score
Dictionary	0.62	0.87	0.72
Template	0.88	0.81	0.84

Table 30: Results – Rule-based NER module, balanced ratio dataset (Exp. 1A)

In the result of Experiment 1A above, there is a significant difference between the precision and recall of the dictionary lookup method, 0.62 and 0.87 respectively. This is expected as dictionary lookup ensures the retrieval of all entities that contains any dictionary words, but cannot differentiate between the actual ER entities and other presence of dictionary words. Also as mentioned in Figure 4, around 13% of TARGET entities are complex structured, this portion is not supported for detection by dictionary lookup. Until more rules are introduced to capture complex structured TARGET entities, it would be very difficult to raise recall of the dictionary lookup (0.87) any higher.

In the result of the template matching method following the dictionary lookup method, recall is slightly lowered, from 0.87 to 0.81, but a worthy trade-off that raised precision from 0.62 to 0.88, consequently, raising the F-score from 0.72 to 0.84. It shows that the template matching method is very effective in validating the instances of dictionary words for if they belong to an ER entity or not. However, this method cannot tend to complex TARGET entities because it relies on the dictionary lookup method for identifying potential ER entity candidates.

The precision, recall and F-score of the rule-based NER module on the balanced ratio dataset (Experiment 1B) are reported in Table 31. Experiment 1A and 1B are represented in the graph in Figure 26.

	Exp. 1B Actual ratio dataset		
	Precision	Recall	F-score
Dictionary	0.23	0.86	0.36
Template	0.59	0.79	0.68

Table 31: Results – Rule-based NER module, actual ratio dataset (1B)

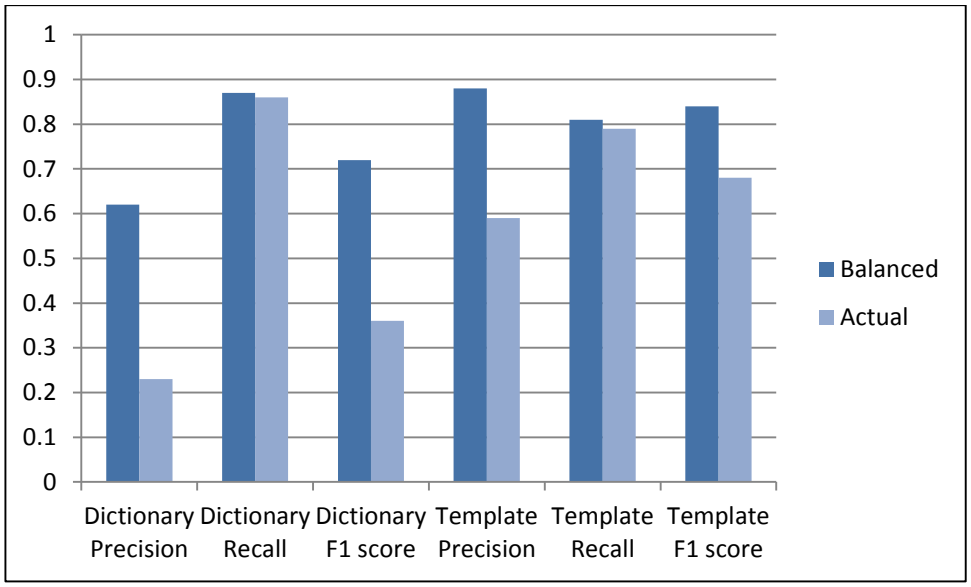

Figure 26: Result Graph – Rule-based NER module, balanced & actual ratio dataset (Exp. 1A & 1B)

In the result of Experiment 1B above, similar patterns exist as mentioned in the results of Experiment 1A. What's notable is the significant drop in the precision of the dictionary lookup method, from 0.62 to 0.23, resulting in a similarly significant drop in F-score, from 0.72 to 0.36. This is most likely due to a high increase of instances of dictionary words that do not form ER entity in the actual ratio dataset, causing this approach to mistaken them for ER entity candidates.

The template matching method also experiences a drop, but not as severe as dictionary lookup, from 0.88 to 0.59. This shows the template matching method is less impacted than the dictionary lookup method in dealing with more realistic dataset.

The precision, recall and F-score of the rule-based NER module, with additional functional word rules (FW), on the balanced ratio dataset (Experiment 1C) is reported in Table 32 and represented in the graph in Figure 27.

	Exp. 1C Balanced ratio dataset with FW		
	Precision	Recall	F-score
Dictionary	0.62	0.87	0.72
Functional	0.62	0.90	0.85
Template	0.86	0.85	0.85

Table 32: Results – Rule-based NER module, balanced ratio dataset with FW (Exp. 1C)

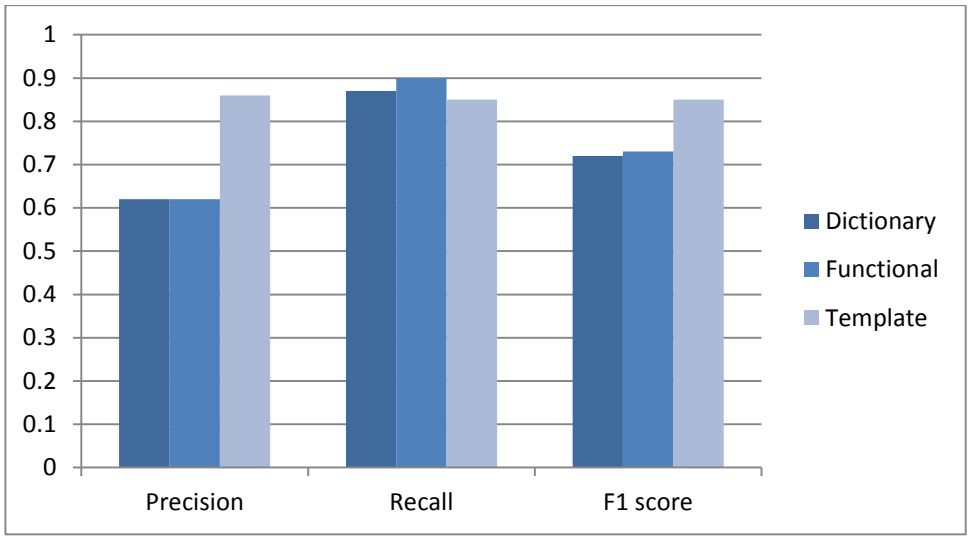

Figure 27: Result Graph – Rule-based NER module, balanced ratio dataset with FW (Exp. 1C)

In the result of Experiment 1C above, the functional word rules method is able to slightly raise the recall, from 0.87 to 0.90. Other results are quite similar to those of Experiment 1A.

The precision, recall and F-score of the rule-based NER module, with additional functional word rules, on the actual ratio dataset (Experiment 1D) are reported in Table 33.

	Exp. 1D Actual ratio dataset with FW		
	Precision	Recall	F-score
Dictionary	0.23	0.86	0.36
Functional	0.23	0.90	0.37
Template	0.57	0.84	0.68

Table 33: Results – Rule-based NER module, actual ratio dataset with FW (Exp. 1D)

In the result of Experiment 1D above, the functional word rules method is similarly able to slightly raise the recall, from 0.86 to 0.90.

Experiment 1C and 1D are represented in the graph in Figure 28.

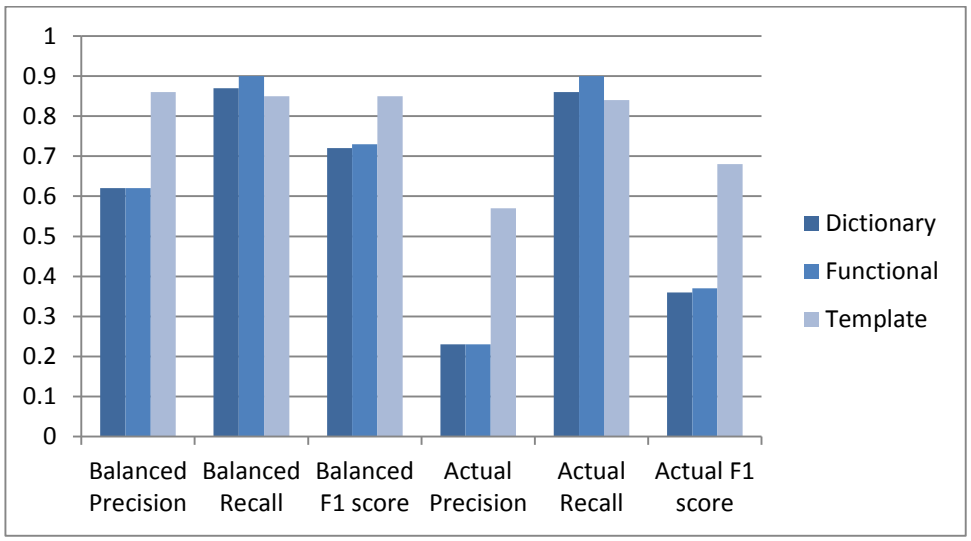

Figure 28: Result Graph – Rule-based NER module, balanced & actual ratio dataset with FW (Exp. 1C & 1D)

In the results of Experiment 1C and 1D above, there is a significant drop of precision from dictionary lookup and functional word rules processes on balanced dataset to actual dataset, from 0.62 to 0.23. The template matching process is less impacted by the actual ratio dataset, from 0.86 to 0.57. Other results are quite similar to those of Experiment 1B.

	Complex TARGET entities found
Without FW (exact match)	0.0%
Without FW (fuzzy matching)	8.1%
With FW (exact match)	60.9%
With FW (fuzzy matching)	91.2%

Table 34: Detection of complex TARGET entities in rule-based NER module

Furthermore, the rate of complex TARGET entities detected in term of all complex TARGET entities in the dataset, with and without using functional word rules are reported in Table 34. This is measured by exact match and fuzzy match (over 50% accuracy of correct words). Fuzzy match signifies that the presence of the complex TARGET entity is detected; and exact match signifies that the exact location and scope of the TARGET entity is detected. Table 34 shows that no complex TARGET entities of exact matched are found without using functional word rules. 8.1% of fuzzy matched is found, likely from the keywords within the complex TARGET entities. When using functional word rules, 60.9% are found as exact matched, and 91.2% are found as fuzzy matched. The extra found in fuzzy matched is likely from complex TARGET entities that are found but included extra words.

8.1.2 Experiment 2: Rule-based SRE module

The precision, recall and F-score of the rule-based SRE module, on golden input (Experiment 2A), and on actual input with (Experiment 2C) and without (Experiment 2B) functional word rules from the previous NER module are reported in Table 35 and represented in the graph in Figure 29.

	Precision	Recall	F-score
Exp. 2A Golden input	0.9	0.71	0.79
Exp. 2B Actual input	0.55	0.39	0.46
Exp. 2C Actual with FW	0.52	0.42	0.46

Table 35: Results – Rule-based SRE module (Exp. 2A, 2B & 2C)

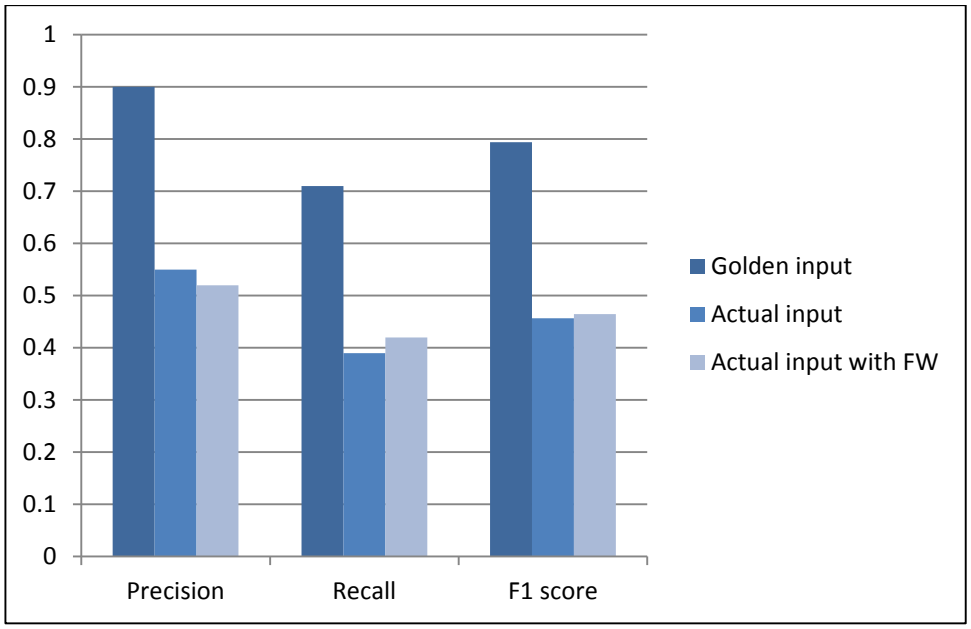

Figure 29: Result Graph – Rule-based SRE module (Exp. 2A, 2B & 2C)

In the result of Experiment 2A above, there is a significant decrease between precision and recall, from 0.9 to 0.7. This shows that the majority of ERs formed from identified ER entities using the template in this method is valid, but there is still a significant portion of ERs that are not formed by the template.

In the results of Experiment 2B and 2C, a slight increase in F-score between actual input and actual input with functional word rules, following the trend of the slight increase in the NER results (8.1.1).

There is significant decrease in precision (dropped by ≈0.38), recall (≈0.32) and F-score (≈0.33) between scores of golden input and actual input (both with and without functional word rules). The significant difference shows the errors in the results of the previous NER module has a heavy impact on the performance of this SRE module.

8.1.3 Rule-based methodology summary

The collection of the precision, recall and F-score of the best performance of rule-based methodology modules (Experiment 1 and 2) are represented in Figure 30 and Figure 31 below.

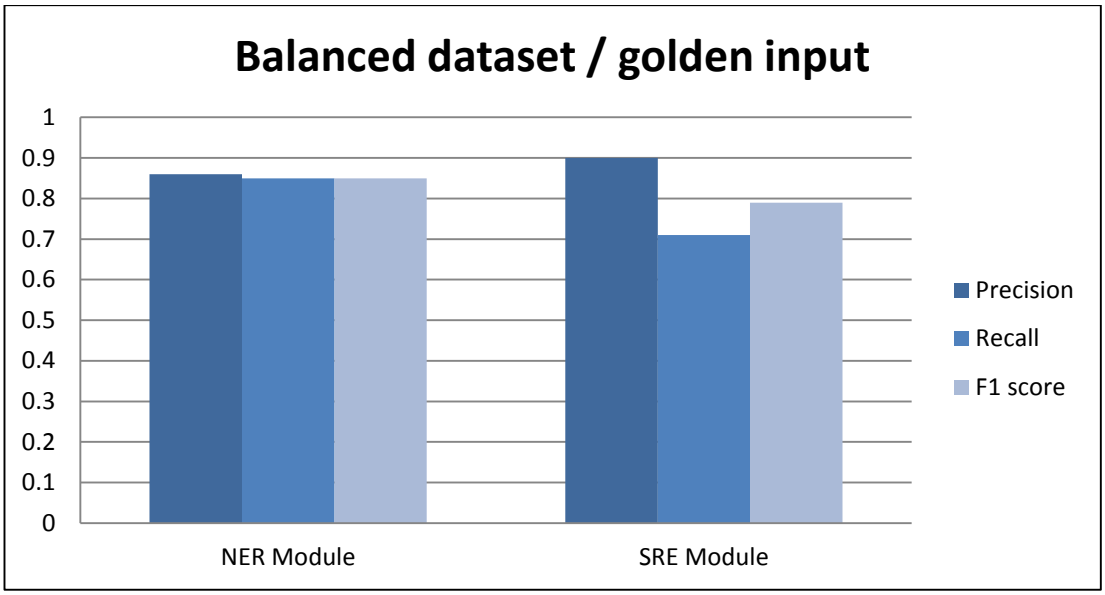

Figure 30: Result Graph – Rule-based methodology modules performance, balanced ratio dataset / golden input (Exp. 1 & 2)

Figure 30 displays the performance of the modules using balanced ratio dataset or golden input, which reflect on the effectiveness of each module independently.

The individual performance of the NER module has a relatively high and very balanced precision and recall, 0.86 and 0.85 respectively. In using the result of this NER module for the SRE module, a low amount of incorrect potential ER entities are carried over and a low number of ER entities are missed out in the dataset.

The individual performance of the SRE module has relatively high precision (0.9) but a moderate recall (0.71).

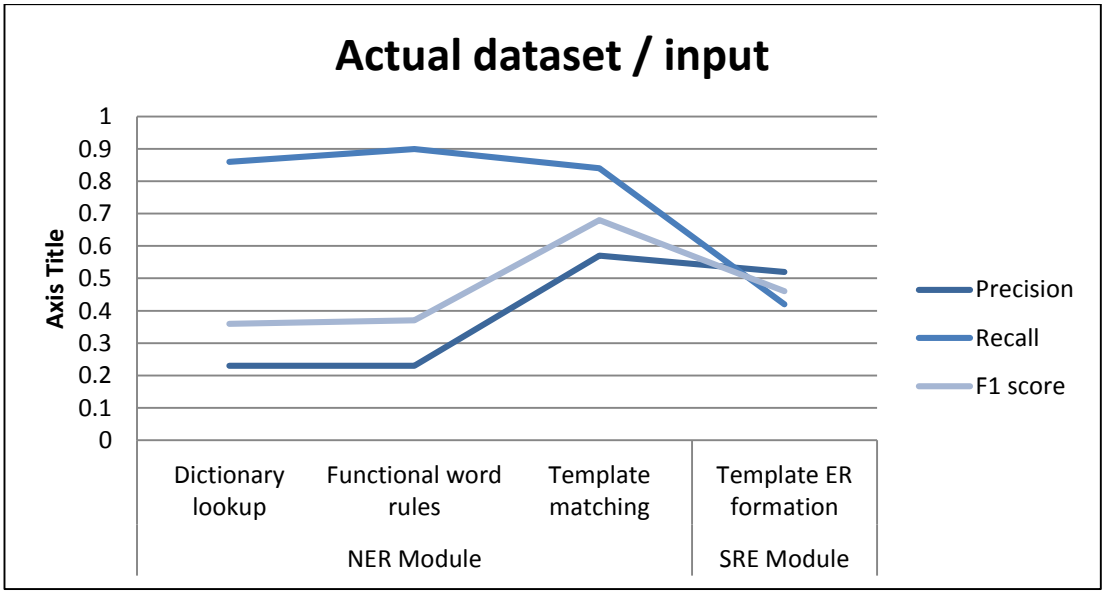

Figure 31: Result Graph – Rule-based methodology performance, actual ratio dataset / input (Exp. 1 & 2)

Figure 31 displays the performance of the methods using actual ratio dataset or actual input, which reflects the effectiveness of all methods performed through the entire process.

The dictionary lookup method has the performance with a relatively high recall (0.86) and very low precision (0.23). The low precision is likely due to the high number of negative cases in the actual ratio dataset. The ER entities detected in this method do not include any complex TARGET entities.

The functional word rules method did not have any improvements on the precision, as its main purpose is to identify new potential ER entities, not the validation of existing potential ER entities. It was able to increase the recall slightly, to 0.9. This is expected, as this method was developed to detect complex TARGET entities. Only a small portion of TARGET entities are complex TARGET entities, and even a smaller proportion among all ER entities. Some complex TARGET entities are detected amongst the ER entities with this method.

The template matching method was very effective in bringing balance to the previously outbalanced precision and recall. Recall was slightly decreased, to 0.84, and precision was significantly increased to 0.57, consequently bring the F-score for the NER module on actual ratio dataset to a moderate score, 0.68.

Given the results of the NER as input, the SRE module suffers loss of performance from carrying over the errors. The precision of the template ER formation method was slightly decreased to 0.52, but recall was more significantly decreased, to 0.42, consequently bringing the F-score to 0.46.

The difference in F-score of the rule-based NER module using balanced ratio dataset (0.85) and actual ratio dataset (0.68) shows the magnitude of impact of the high number of negative cases in the actual dataset. The difference in F-score of the rule-based SRE module using golden input (0.79) and actual input (0.46) shows the importance of the NER module to the SRE module.

8.2 Feature-based methodology

8.2.1 Experiment 3: Feature-based NER module

8.2.1.1 Sentence extraction

The precision, recall and F-score of the sentence extraction method, using balanced ratio dataset (Experiment 3S1A) and actual ratio dataset (Experiment 3S1B) are reported in Table 36, and represented in the graph in Figure 32.

	Precision	Recall	F-score
Exp. 3S1A Balanced ratio dataset	0.68	0.98	0.80
Exp. 3S1B Actual ratio dataset	0.52	0.98	0.68

Table 36: Results – Sentence Extraction method (Exp. 3S1A & 3S1B)

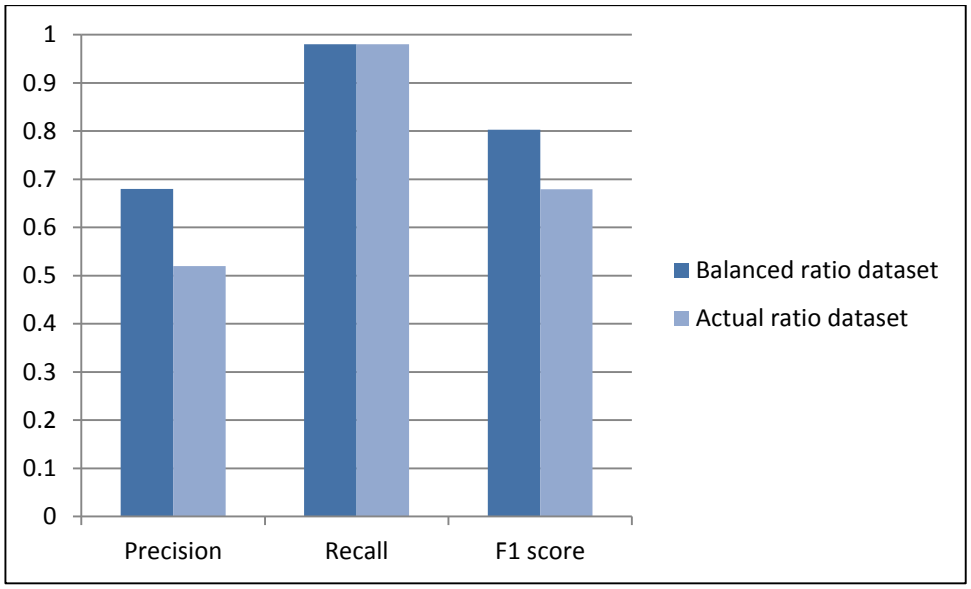

Figure 32: Result Graph – Sentence Extraction method (Exp. 3S1A & 3S1B)

In the results of Experiment 3S1A and 3S1B above, there is a significant decrease in precision in balanced ratio dataset and actual ratio dataset, 0.68 and 0.52 respectively, but no difference in recall, both at 0.98. This is in line with the fact that the actual ratio dataset contains the same number of positive cases but much more negative cases. Considering the number of negative cases increased by over 10 times, this decrease shows that this method is not too heavily affect by increased negative cases. There is an expected significant difference between precision and recall, 0.68 (precision-balanced) and 0.52 (precision-actual) to 0.98 (recall-both) respectively. This shows that the EFFECT dictionary words used in this method is ranged enough to identify most sentences with one or more RE, but there are still a significant number of occurrences of these EFFECT dictionary words that do not lead to an ER.

8.2.1.2 Word classification

The precision, recall and F-score of the word classification method using the Decision Tree classifier (selected in Appendix 4) using golden and actual input, with and without functional word features (Experiment 3S2A, 3S2B, 3S2C and 3S2D) are reported in Table 37, and the F-scores are represented in the graph in Figure 33.

	Precision	Recall	F-Score
Exp. 3S2A Golden input	0.69	0.75	0.72
Exp. 3S2B Actual input	0.35	0.65	0.45
Exp. 3S2C Golden input with FW	0.69	0.77	0.73
Exp. 3S2D Actual input with FW	0.34	0.67	0.45

Table 37: Results – Word Classification method (Exp. 3S2A, 3S2B, 3S2C & 3S2D)

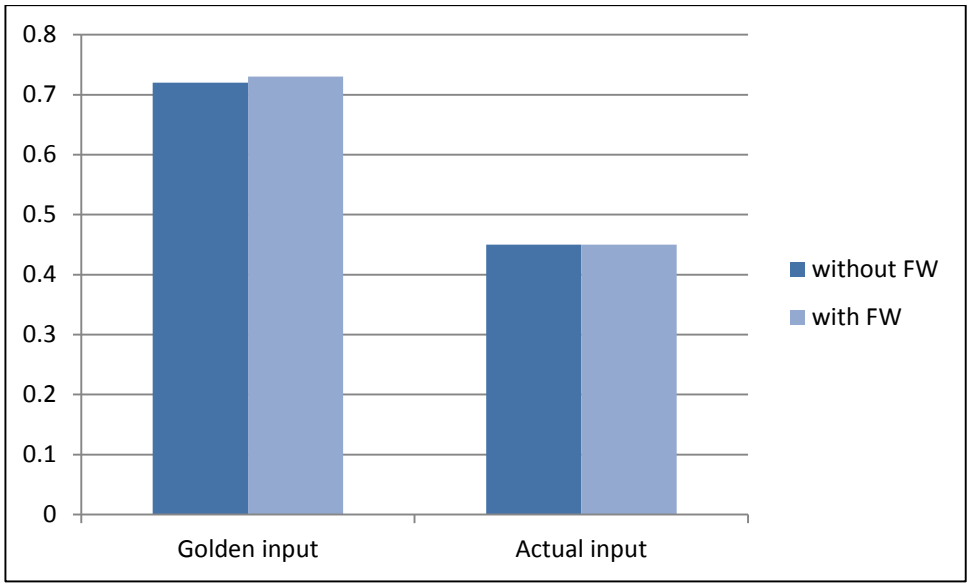

Figure 33: Result Graph – Word Classification method (Exp. 3S2A, 3S2B, 3S2C & 3S2D)

In the results of Experiment 4 above, there is a slight increase in F-score when using functional word features (≈0.01), showing that the functional word features are not very effective in the feature-based NER module. There is a significant decrease in F-score from using golden input to actual input (≈0.28). The significant difference shows the errors in the results of the previous sentence extraction method has a heavy impact on the performance of this word classification method.

	Complex TARGET entities found
Without FW (exact match)	0.0%
Without FW (fuzzy matching)	10.3%
With FW (exact match)	0.0%
With FW (fuzzy matching)	13.2%

Table 38: Detection of complex TARGET entities in feature-based NER module

Furthermore, the rate of complex TARGET entities detected in term of all complex TARGET entities in the dataset, with and without using functional word features are reported in Table 38. This is measured by exact match and fuzzy match (over 50% accuracy of correct words). Table 38 shows that when functional words features are not, no complex TARGET entities of exact matched are found; 10.3% of fuzzy matched is found. When using functional word features, none are found as exact matched, and 13.2% are found as fuzzy matched. There is no increase of exact matches by using functional word features.

8.2.1.3 Entity formation

The precision, recall and F-score of the entity formation method (Experiment 3S3A and 3S3B) are reported in Table 39, and represented in the graph in Figure 34.

	Precision	Recall	F-score
Exp. 3S3A Golden input	1.00	1.00	1.00
Exp. 3S3B Actual input	0.56	0.62	0.59

Table 39: Results – Entity Formation method (Exp. 3S3A & 3S3B)

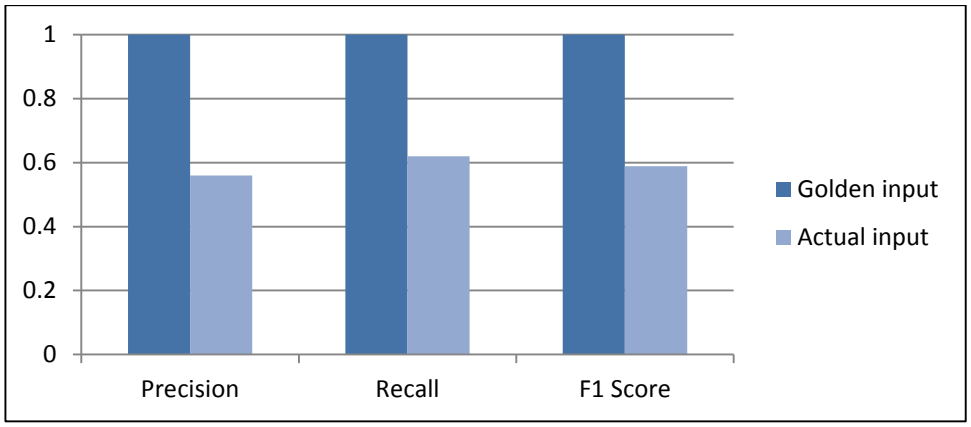

Figure 34: Result Graph – Entity Formation method (Exp. 4S3A & 4S3B)

In the results of Experiment 3S3A, the precision, recall and F-score are all 1.00. This shows that there are no pairs of separate ER entities of the same type that are adjacent to each other in the dataset, and that the entity formation method is able to define entities with zero error, if given an input of zero error.

In the results of Experiment 3S3B, the precision, recall and F-score have decreased significantly (≈0.41). This decrease shows that the errors of the actual input of the previous stages have a heavy impact on this method.

8.2.2 Experiment 4: Feature-based SRE module

The precision, recall and F-score of the relation classification method in the feature-based SRE module, using the Decision Tree classifier (selected in Appendix 4) using golden and actual input (Experiment 4A and 4B) are reported in Table 40, and the F-scores are represented in the graph in Figure 33.

	Precision	Recall	F-score
Exp. 4A Golden input	0.89	0.88	0.88
Exp. 4B Actual input	0.40	0.43	0.41

Table 40: Results – Feature-based SRE module Relation Classification method (Exp. 4A & 4B)

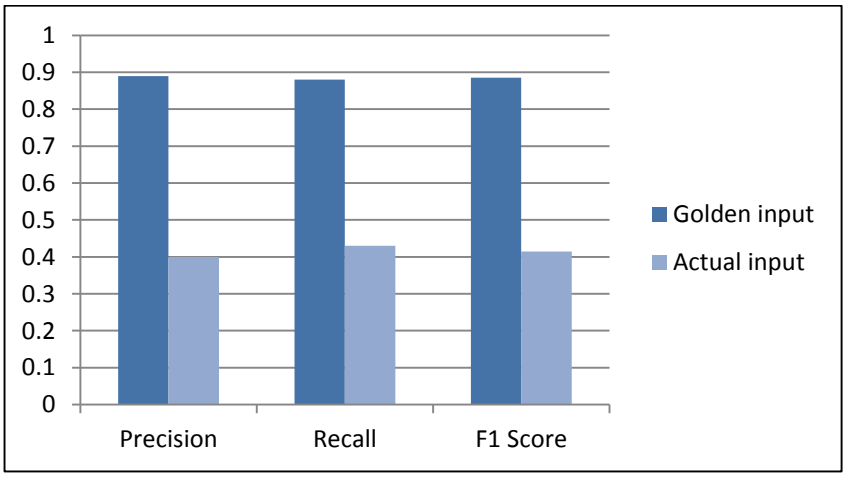

Figure 35: Result Graph – Feature-based SRE module Relation Classification method (Exp. 5A & 5B)

In the results of Experiment 5A, the precision and recall are quite even, 0.89 and 0.88 respectively. This shows that the features used in the relation classification method do not show significant sign of weakness or strength in precision or recall.

In the results of Experiment 5B, the precision, recall and F-score have decreased significantly (≈0.47). This decrease shows that the errors of the actual input from previous experiments / stages have a heavy impact on the performance of this method.

8.2.3 Feature-based methodology summary

The collection of the precision, recall and F-score of the best performance of feature-based methodology modules (Experiment 3 and 4) are represented in Figure 36 and Figure 37 below.

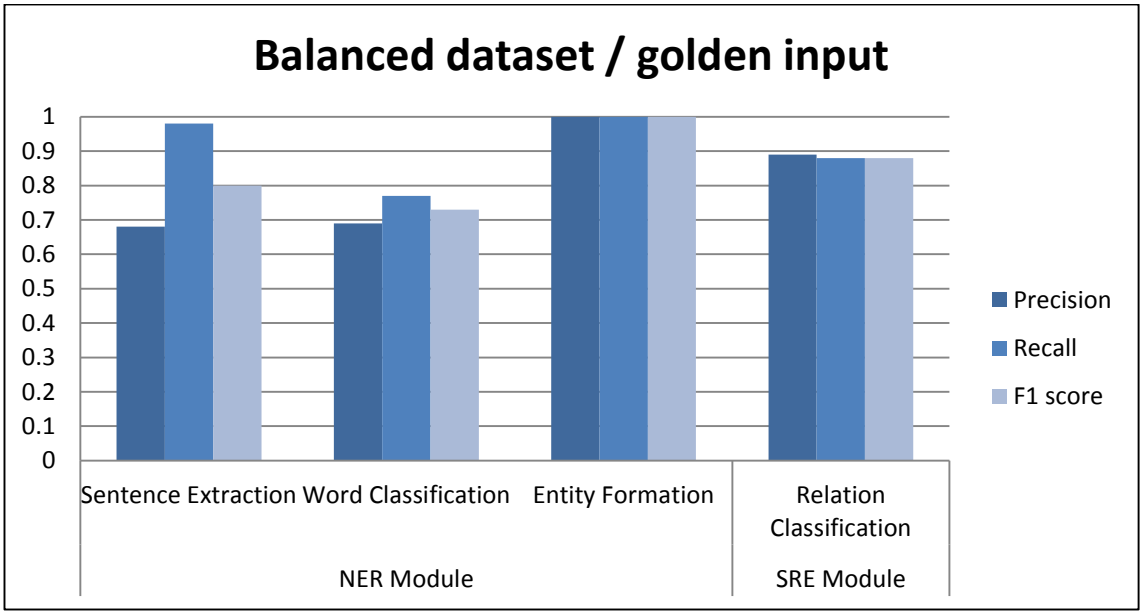

Figure 36: Result Graph - Feature-based methodology, balanced ratio dataset / golden input (Exp. 3 & 4)

Figure 36 displays the performance of the methods using balanced ratio dataset or golden input, which reflect on the effectiveness of each method independently.

The sentence extraction method has a very high recall (0.98), but a moderate precision (0.68). On its own, it will be more effective to trade off some recall for precision to achieve a more balanced F-score. However, as it is the first stage for the methods in the NER module, it is still favourable to have a higher recall, which is difficult to increase afterwards, and a lower precision, which is easier to regain later.

The word classification method has the lowest overall performance (F-score 0.73) amongst the other methods. It will likely lower the overall performance of methods using actual input following it. It also has a higher recall than precision, leaving the restoration of the precision / recall balance to further methods.

The entity formation method has a perfect performance (all scores 1.0) when performed without errors from input. The relation classification method also has a relatively high

performance and it has a very even balanced precision and recall (0.89 and 0.88 respectively).

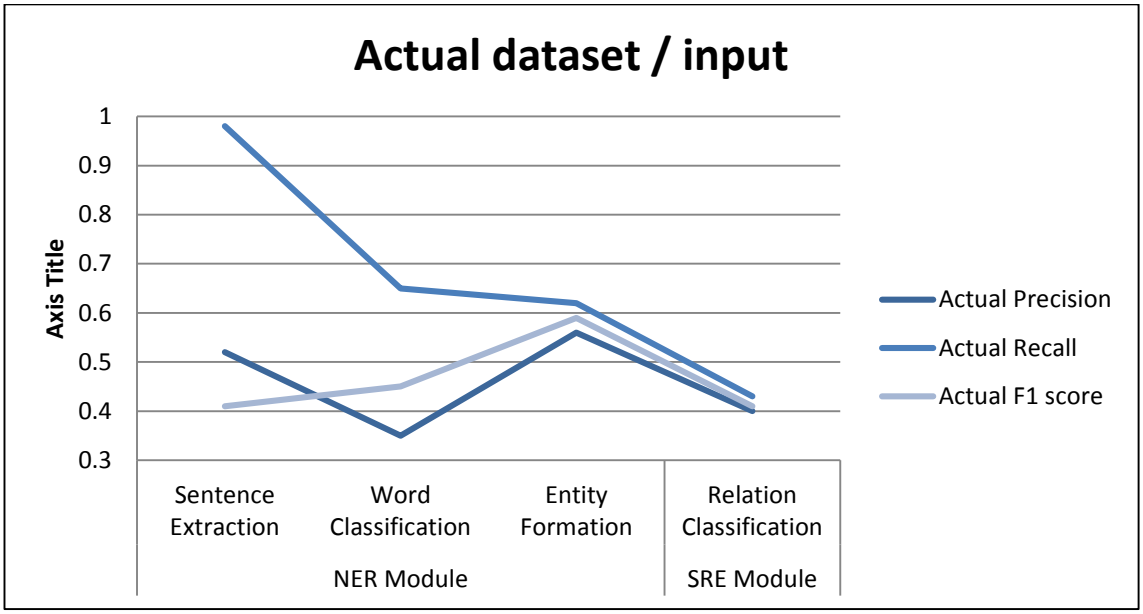

Figure 37: Result Graph - Feature-based methodology, actual ratio dataset / actual input (Exp. 3 & 4)

Figure 37 displays the performance of the methods using actual ratio dataset or actual input, which reflects the effectiveness of all methods performed through the entire process.

The sentence extraction method has the performance of a very high recall (0.98) and relatively low precision (0.52). The low precision is likely due to the high number of negative cases in the actual ratio dataset.

As expected from observation of the performance using golden input, the word classification method is the weakest link in the whole process. Recall decreases significantly in this method (0.65). Precision in word classification is the lowest value among other processes in the overall process (0.35).

Recall continues to decrease at the entity formation method (0.62), but precision increase to its highest point in the whole process (0.56), bringing balance to the precision and recall. The decrease in recall is likely from words wrongly classified as ER entity are next to correctly classified ones of the same type. When the entity formation method is implemented, it caused these correctly classified words to form ER entities with the incorrect number of words, consequently lowering recall. The increase in precision is likely from this method removing words wrongly classified as ER entity that do not have other ER entities around it to form an ER.

Finally recall and precision lowered in the last method: relation classification, scoring 0.4 and 0.43 respectively, maintaining the precision and recall balance.

The difference in F-score of the feature-based NER module word classification method using golden input (0.73) and actual input (0.45) shows the magnitude of impact of the high number of negative cases in the actual dataset after the sentence extraction method. The

difference in F-score of the feature-based SRE module using golden input (0.88) and actual input (0.41) shows the importance of the NER module to the SRE module.

8.3 Experiment 5: NER Module (limited dictionary words)

The precision, recall and F-score of the NER modules of both methodologies using limited dictionary words are reported in Table 41 and they are represented with the results of NER modules of both methodologies (Experiment 1 and Experiment 3) with full dictionary words in Figure 38.

	Precision	Recall	F-score
Exp. 5A Rule-based NER module	0.52	0.29	0.37
Exp. 5B Feature-based NER module	0.50	0.52	0.51

Table 41: Results - NER Modules, limited dictionary words (Exp. 5)

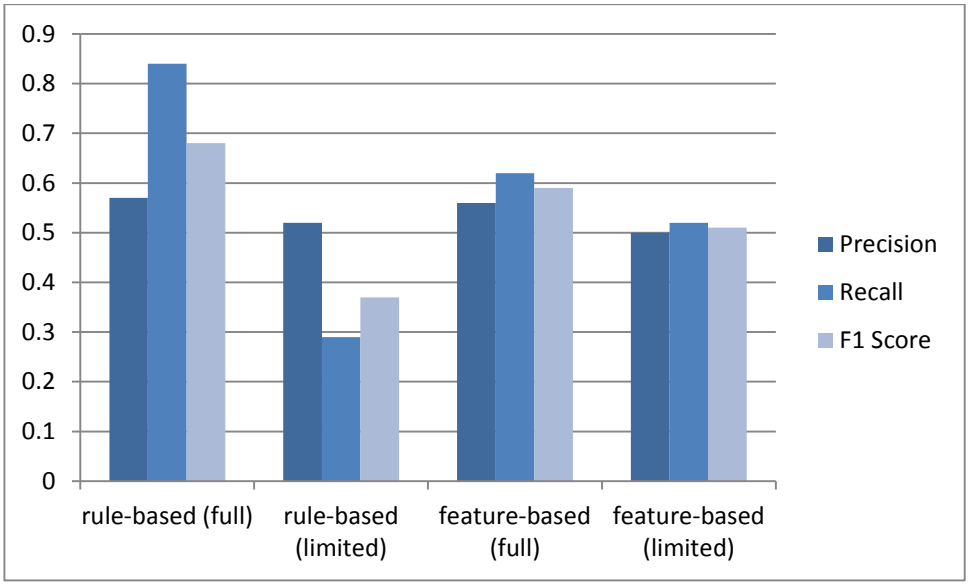

Figure 38: Result Graph – NER modules, full / limited dictionary words (Exp. 5)

In Figure 38, the results of rule-based NER module and feature-based NER module using limited dictionary words (Experiment 5A and 5B) are compared with the results of rule-based NER module and feature-based NER module using full dictionary words (Experiment 1D and 3S3B). In perspective of the rule-based NER module, there is a slight decrease in precision and sharp decrease in recall after dictionary words are limited. In perspective of the feature-based NER module, there is slight decrease for both precision and recall after dictionary words are limited. This comparison shows that although the rule-based NER module excels the feature-based NER module in F-score when using full set of dictionary words, it is possible for a sharp drop in performance for the rule-based NER module when applied to a dataset with many unknown words, whereas the impact of unknown words on feature-based NER module is relatively lower.

8.4 Comparison of methodologies

The results of the two methodologies are compared below to see the difference in performance of the modules in the two methodologies.

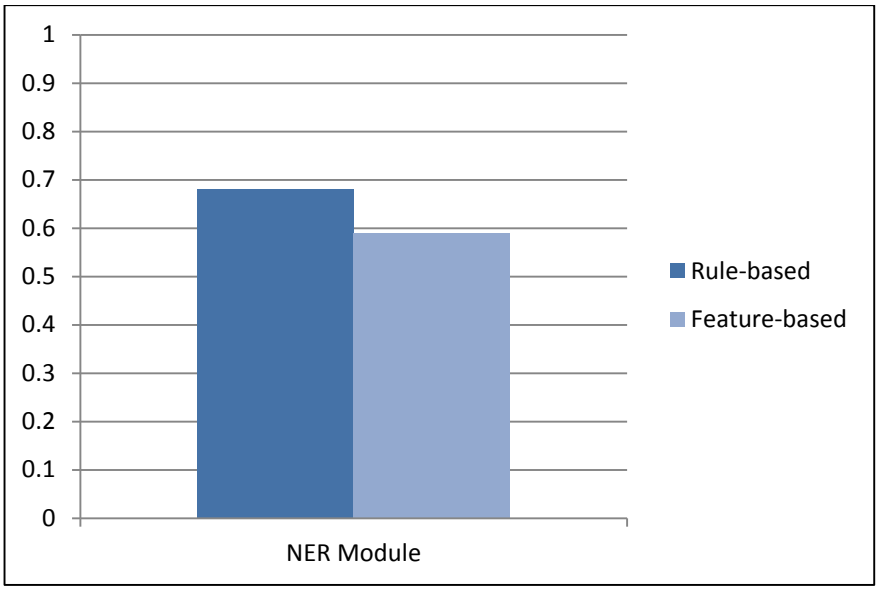

Figure 39: Result Graph – NER modules, actual dataset / input (Exp. 1 & 3)

Figure 39 compares the performance of the NER modules of the two methodologies from their results on actual ratio dataset. The rule-based NER module was evaluated as a combination of methods, and the feature-based NER module was evaluated by each individual method. In the NER module, a factor was underestimated in the designing of the feature-based methodology, i.e. contextual information on the sentence level.

Looking into more details of the NER modules of both methodologies, the dictionary lookup method in rule-based NER module and the word classification method in feature-based NER module are of corresponding roles and both focus on individual words. The dictionary lookup method looks only at the text of the word. The word classification method considers additional contextual information, e.g. POS tag, and neighbouring words. When compared alone, the latter method performs better than the prior method, 0.45 (Table 37) and 0.36 (Table 33) respectively in F-score. The word classification method also had a better balance between precision and recall. With this balance, the word classification method is less impacted from using balanced ratio dataset to actual ratio dataset, 0.73 to 0.45 (Table 37); compared to the dictionary lookup method, 0.72 (Table 32) to 0.36 (Table 33). However the respectively methods following these two methods changed the outcome of the NER module comparison.

The template matching method in the rule-based NER module and the entity formation method in the NER module are of corresponding roles and focus on information on the sentence level. The dictionary lookup method relies heavily on the template matching method, to remove a large amount of invalid potential ER entities based on sentence level information, in effect, preserving much of the high recall and improve the precision significantly. The two methods are an effective pair. The entity formation method in the

feature-based NER module essentially performs in the same way as the template matching method. However the word classification method is more balanced by itself and less reliant on other methods. The entity formation method has less potential ER entities to work on, and its function to remove invalid retrievals partially overlaps with functions within the word classification, in effect, it was not able to improve the precision by much and the recall of word classification is already lower than that of dictionary lookup. As a result, the overall feature-based NER module (0.59) performs worse than the rule-based NER module (0.68). In order for the feature-based NER module in the feature-based methodology to improve, another sentence level method that can work more effectively with the word classification method needs to be implemented.

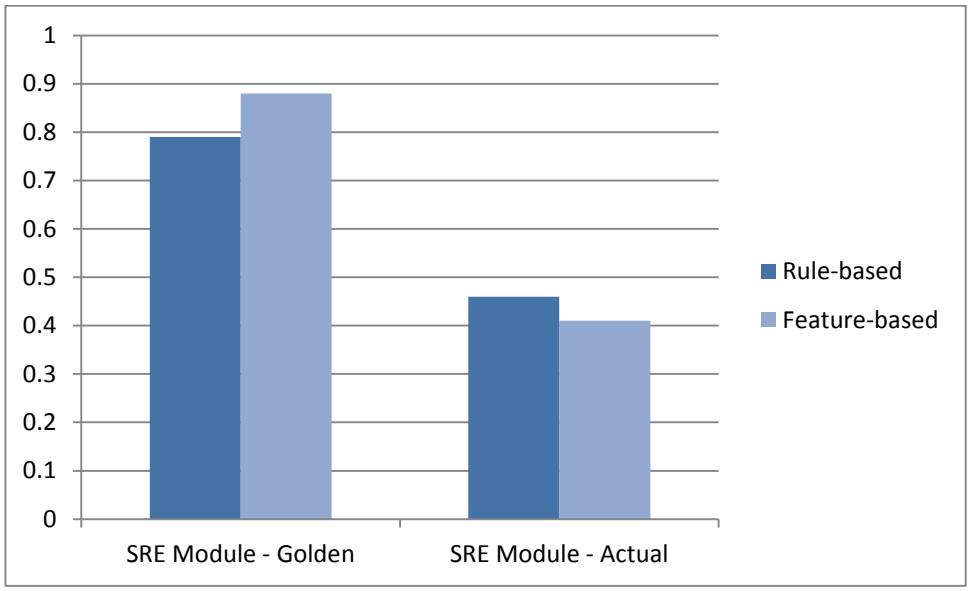

Figure 40: Result Graph – SRE modules, golden & actual input (Exp. 2 & 4)

Figure 40 compares the performance of the SRE modules of the two methodologies from their results on both golden input and actual input. It shows that the feature-based SRE module (0.88) performs better than the rule-based SRE module (0.79) when given golden input. This implies that as a method / module alone, the relation classification method performs better than the corresponding template ER formation method, and achieves a more balanced precision and recall. However, due to the errors carried over in the actual input of the NER module, the feature-based SRE module performs worse than the rule-based SRE module when using actual input. Consequently, the overall performance of the rule-based methodology performs better than the feature-based methodology.

8.5 Summary

The results of the experiments in this chapter can be summarised as below:

- In the NER module, rule-based methodology performs better than feature-based methodology when using full set of dictionary words (8.1.1 and 8.2.1). However the feature-based NER module performs better than rule-based NER module when using limited dictionary words (8.3).

- In the SRE module, feature-based approach performs better than rule-based approach (8.1.3 and 8.2.2).
- The performance of both the rule-based and feature-based NER module decrease significantly when switching from balanced ratio dataset to actual ratio dataset, indicating magnitude of impact of the high ratio of negative cases in the actual dataset (8.1.3 and 8.2.3).
- The performance of both the rule-based and feature-based SRE module decrease significantly when switching from golden input to actual input, indicating the magnitude of importance the NER module is to the SRE module (8.1.3 and 8.2.3).
- Over the course of the two modules, rule-based approach performs better than feature-based approach (8.4).
- Functional words are very effective as rules in the rule-based approach to detect complex TARGET entities. However due to the small portion complex TARGET entities take in all ER entities, this improvement is not significantly shown in the overall performance of the rule-based NER module (8.1.1).
- In the feature-based NER module, functional words as features shows no improvement in the detection of complex TARGET entities and shows no significant improvement in the performance of the module (8.2.1.2).

Chapter 9. Discussion

Chapter 9 discusses the evolution of the two methodologies used in this research project and evaluates how well they have performed as to achieve the goals set for this research project. It also discusses the role of functional words in these two methodologies and their performance.

9.1 Rule-based methodology

The rule-based methodology offers a simple start into the RE task, especially in the case of this research, where a new type of relation: Effect Relation is introduced. Detailed observation of the dataset and desired results allowed the detection of characteristics in the text, which were eventually summarised and translated into rules for the rule-based approach. These straightforward rules allow the approach to extract what is wanted or remove what is unwanted in the observed dataset. However these rules often lead to an unbalanced emphasis on precision and recall. For example, the dictionary lookup method emphasizes significantly on recall (as shown in the results in Table 30), as it can only ensure the inclusion of wanted text. On the other hand, the template matching method places the emphasis on precision, as it can only ensure the exclusion of unwanted text and requires previous input. By applying these methods together, more balance between precision and recall is achieved (as shown in results in Table 30).

Unstructured text, such as the TCM journal publications used in this research, is very flexible in its syntactic structure. This flexibility, when translated into rules, brings high complexity and possibly need for exception cases. As a result, the construction of rules is limited by the level of detail of the human observation on the ever-growing dataset. As the complexity of the rules increases, the effort to resolve possible conflicts between existing rules and new rules also increases, and sometimes requires existing rules to be completely reshaped. For example, towards the start of annotating the dataset, most annotated ERs had a functional word residing within the sentence of the ER. This observation proved useful when these functional words acted as an additional criterion to retrieve ER potential sentences. It significantly improved the precision of the sentences retrieved. This observation was further explored to categorise ERs by different groups of functional words (as discussed in Appendix 1). However as the annotated dataset grew larger and the definition of ER adapted to more text, the number of ERs without functional words outgrew those with functional words, so this additional criterion to identify the presence of ER by functional words was eventually removed. Although this example of rule becoming obsolete as dataset is scaled up, occurred in the process of annotation, similar situation can happen in the evolution of rule-based approaches.

Overall, the rule-based methodology has a close resemblance to the process of the annotation of ERs, and is therefore set a baseline for ER extraction performance. The rule-based methodology in this research project is still at a relatively simple stage (using 3 methods: dictionary lookup, template matching and template ER formation) and open for further extension. However from a long term perspective, the effort to maintain rule-based methodology at a high complexity level is undesirable.

9.2 Feature-based methodology

The feature-based methodology is explored in this research project aiming to overcome the rigidness of the aforementioned rule-based methodology in ER extraction. It is considered as a feature-based approach because its core methods are feature-based (word classification and relation classification), but it also consists of some rule-based methods as support methods (sentence extraction and entity formation). It builds on the rule-based approach by converting the dictionary lookup method into features in the word classification method, and using some parts of the template matching method in the entity formation method. It views the ER extraction as a classification problem, which is decided by a collaboration of features. These features consist of features converted from the dictionary lookup method, and different contextual information around the classification problems, i.e. the two modules: NER and SRE. Many features were experimented to evaluate their effectiveness.

In the comparison of results in Chapter 8, the performance of the feature-based NER module was lower than the rule-based NER module. This is explained in the results (8.4) that the methods in the feature-based NER module did not work effectively together as those in the rule-based module, especially using sentence level information.

Another cause for the low performance of the feature-based approach is the incorrect contextual information generated by the LTP parser. This parser not only provides word segmentation for the text, but also provides the contextual information of each word, such as POS tag and dependency tag. Some errors are found and corrected in the dataset, but more errors may remain undetected (6.2.6.4). The feature-based approach is heavily reliant on the collaboration of contextual information for make the correct classification, but if the contextual information is incorrect, the performance will be greatly limited. By contrast, the rule-based approach uses very limited contextual information; consequently it is not as heavily influenced by incorrect contextual information in the dataset. In the simulation of encountering unknown words by using limited dictionary words for the NER modules of both methodologies (8.2.2), the feature-based NER module outperforms the rule-based NER modules, demonstrating it can handle unknown words more effectively.

Building on this observation, this rule-based methodology will be much more restrictive to the domain of text it is originally based on than the feature-based methodology. When moved to another domain, the rule-based methodology will require large amount of modification or faces the awkward situation of performance dropping drastically. This is due to the fact that upon changing domain of text, much of the lexical representations and grammatical structure that the rule-based methodology relies on in the original domain of text also changes. An experiment for performance of rule-based methodology on the People's Daily corpus [60], widely used for Chinese NLP, was proposed near the start of the research. However it was later cancelled because the ERs found in this corpus are not very meaningful. Unlike the ERs in the TCM journal dataset, the ERs found in the People's Daily are not in conclusive formal statements. There was no other processed Chinese text corpus of similar nature of the TCM journal dataset available, so the evaluation of applying this rule-based approach on another domain of text was not performed. However it is anticipated that the dictionary words for SOURCE and TARGET used in the dictionary lookup

method will greatly differ if moved to a different domain of medical journals. Similar situation will apply for the templates used in template matching and template ER formation methods if moved to a different domain. The feature-based methodology, on the other hand, views the ER extraction task through the perspective of contextual information features, which are similar in most domain of text, and handles high complexity of features with mathematical classification models. This allows shifting to a larger dataset or dataset in different domain of text more efficiently than the rule-based methodology. However as some features are converted from the rule-based methodology, those features will need to undergo similar modifications as mentioned for the rule-based methodology.

The SRE module in the feature-based methodology performs better than that in the rule-based methodology. This shows that in dealing with the various forms ERs can take, the collaboration of features has its advantage in adapting than rules. It is when errors are carried over from the actual input, i.e. the NER module, does the relation classification method perform worse than the template ER formation method.

Overall, the two modules of the feature-based methodology (NER module and SRE module) have been compared with the baseline set by the rule-based methodology; and the two methodologies show to have advantage in different modules of ER extraction.

9.3 Use of functional words

The functional word rules method in the rule-based methodology performed significantly well in detecting the presence of complex TARGET entities and relatively well in exact extraction (finding the exact location and scope) of complex TARGET entities (Table 34). In contrast, the functional word features in the feature-based methodology performed poorly in both detection of presence and exact extraction. This is likely due to the following reasons:

- Loss of complexity – the functional word rules method is applied after the dictionary lookup method, so the input information provided to the functional word rules method is on the sentence level, i.e. it knows the potential entities in the sentence. By contrast, the functional word features are included in the word classification method and performed as one process. Its input information is on the word level, i.e. the current word and its limited neighbours.
- Limited search space – the number of neighbours for examining the functional word features are limited, whereas the functional word rules method can search the entire sentence when the conditions are met.
- Small portion of complex TARGET entities – the portion of complex TARGET entities compared to other entities in the dataset is very small. It is very difficult for the functional word features to influence the collaborated decision of all the features to detect the rare cases of complex TARGET entities; whereas rule-based methods

can detect cases that fit the conditions no matter how rare they may be. This is one of the weaknesses of feature-based approach in dealing with exceptional cases.

Due to the above reasons, functional word features are not as effective in detecting complex TARGET entities as the functional word rules method.

9.4 Summary

In summary, the rule-based methodology takes an effective approach to both the definition and extraction of ER. It has set a baseline for other methodologies to compare against. The rules used in this methodology can be converted to elements used for other methodologies.

The feature-based methodology takes a different perspective of viewing the ER extraction as a series of classification problems. It extends on the rule-based methodology by utilising certain rules as features and certain rule-based methods as support methods. Its NER module performs lower than that of rule-based methodology in this dataset, however experiments has also found the feature-based NER module performs better when dealing with unknown words. The feature-based SRE module performs better than rule-based; and overall the feature-based methodology performs lower than the rule-based methodology (Figure 39).

In terms of using functional words, the functional word rules method can significantly improve the detection of complex TARGET entities; whereas the functional word features, and in turn reflects that feature-based methods are not suited for detection of minority cases, e.g. complex TARGET entities.

Chapter 10. Conclusion

Chapter 10 concludes the work done in this research project and identifies some possible directions for future work into ER extraction.

Relation Extraction has found outstanding application in the domain of medical science. It is a valuable mean to tasks such as constructing relationship networks amongst genes and diseases, or studying the interactions between proteins. Effect Relation is a type of relationship describing the positive or negative effect of one entity to another. In this research, the extraction of ER has been applied to the text domain of TCM journal publications as case study. This case study explores the characteristics of ERs, such as the effect of a TCM ingredient on a symptom, in the domain of TCM (refer to 1.5 Contributions 1). Furthermore this case study is built on Chinese unstructured text, and takes into consideration the common NLP problems faced in this field of text in the development of ER extraction methodology.

Two methodologies are developed in this research, taking different approaches toward this RE task and their results to be compared for evaluation. The two methodologies implement the standard RE system architecture, encompassing the Named Entity Recognition (NER) module and the Semantic Relationship Extraction (SRE) module, with different methods within these modules modified for ER extraction.

One ER extraction methodology takes a rule-based approach (6.3 and 6.4). It utilises predetermined rules and templates, derived from the characteristics and pattern observed in the dataset. The performances of the individual modules in the rule-based methodology are reported as F-scores: 0.85 (NER) and 0.79 (SRE); while the overall performance of this methodology is reported as F-score: 0.46 (refer to 1.5 Contributions 3). There are no other existing works to benchmark this overall performance. The overall performance can be considered relatively high as it suffers loss of performance from errors being introduced as early as the pre-processing stage and carried through the NER module and SRE module.

The second methodology takes a feature-based approach (6.5 and 6.6). It views the RE task as a classification problem and utilises mathematical classification model and features consisting of contextual information and rules. The performances of the individual modules in the feature-based methodology are reported as F-scores: 0.73 (NER) and 0.88 (SRE); while the overall performance of this methodology is reported as F-score: 0.41 (refer to 1.5 Contributions 4).

ER defined in this research is a semantic relationship. It does not follow a specific set of rules in its textual form. The rule-based methodology performs better than feature-based methodology because the textual forms of ER in the dataset of this research project have been carefully studied. Experiments in this thesis have shown that the feature-based methodology excels the rule-based methodology when dealing with words unknown to the dictionaries in both methodologies. In terms of scalability and adaptability, predetermined rules and template will perform poorly against new unknown forms of ER in a larger dataset or dataset in a different domain of text. The addition and revision of rules will create complexity and result in high maintenance, limiting extent of improvements can be made

on the rule-based methodology. The feature-based methodology, however, can handle the increase in complexity with mathematical classification models. It will respond to new forms of ER more effectively as the given training dataset grows. Therefore the feature-based methodology will be more ideal when exploring ER extraction in other domains of text.

The actual ratio dataset used in this research contains more than 10 times of negative cases than that of the balanced ratio dataset (7.1.1). The sharp increase in negative cases is also reflected as significant decrease in performance of both methodologies (8.1.3 and 8.2.3). This is a common reality in most RE tasks as the size of the desired information is often minimal to the size of the dataset. This stresses the importance of the pre-processing stage to reduce negative cases and provide the RE methodologies with more balanced dataset.

This thesis explored the important role functional words play in the contemporary Chinese language and the potential benefits to incorporate functional words into the ER extraction methodologies (Chapter 5) (refer to 1.5 Contributions 5). Experiments have found the functional words to be a valuable resource as additional rules in the rule-based NER module. It can significantly improve the detection of complex TARGET entities (8.1.1). However functional words are not as effective when used as features in the feature-based NER module (8.2.1.2). This also reflects that the feature-based methodology is weaker in the detection of minority cases, e.g. complex TARGET entities (9.3).

Over the course of this research project, a TCM dataset consisting of 1486 instances of ER (6.2.4), 2023 sentences (6.2.2) and 48231 words (6.2.3) is constructed for the full text of 31 TCM journal publications (7.1.2) (refer to 1.5 Contributions 2). A total of 483 dictionary words are collected and categorized from the dataset (refer to 1.5 Contributions 6). These dictionary words have shown to be valuable to both ER extraction methodologies.

10.1 Future work

Several areas for improvements are identified but not able to be pursuit due to the time constraint of this research project. They can be applied to future research for improvements on both ER extraction methodologies. These improvements include:

- Cleaner dataset – more thorough cleaning of the dataset to remove or correct any hidden errors.
- Improved word parser – to provide more accurate contextual information. This is similar to the improvement above, but requires a word parser that is more suited to the domain of text in the dataset.
- Reference to external database – as shown in the experiments, dictionary words play a significant role in both methodologies used in this research project. The current list of dictionary words is collected mainly from the used dataset, but they have been categorised and future research can expand on them using external database to the respective category. Significant time and effort have been taken to

prepare the current list of dictionary words, and it is expected to be effective for articles of close vocabulary resemblance to the current dataset. If the dataset is to be expanded, or these methodologies are to be performed on other datasets, the use of existing dictionaries and other external database is recommended to reduce time and effort of dictionary words preparation.

Several areas of future work are suggested in the directions of:

- Exploring a kernel-based approach in the ER extraction task
- Application of the ER concept on another medical domain
- Exploration of the forms of complex ER entities and more effective ways for identification.

Reference

[1] N. A. Chinchor, “Overview of MUC-7/MET-2,” in *the Seventh Message Understanding Conference (MUC-7)*, Virginia, 1998.

[2] D. Zelenko, C. Aone and A. Richardella, “Kernel Methods for Relation Extraction,” *Journal of Machine Learning Research,* vol. 3, pp. 1083-1106, 2003.

[3] ACE08, “Automatic Content Extraction 2008 Evaluation Plan (ACE08) Assessment of Detection and Recognition of Entities and Relations Within and Across Documents,” 2008.

[4] H.-W. Chun, Y. Tsuruoka, J.-D. Kim, R. Shiba, N. Nagata, T. Hishiki and J.-i. Tsujii, “Extraction of Gene-Disease Relations from Medline Using Domain Dictionaries and Machine Learning,” *Pacific Symposium on Biocomputing,* vol. 11, pp. 4-15, 2006.

[5] Wang, Jian; Ji, Ming-Hui; Lin, Hong-Fei; Yang, Zhi-Hao;, “Protein-protein interaction extraction based on contextual and syntactic features,” *Journal of Computer Applications,* vol. 32, no. 4, pp. 1074-1077, 2012.

[6] L. J. Fundukian and Gale Group., Gale Encyclopedia of Medicine, Detroit, 2006.

[7] J. C. Segen, Segen's Medical Dictionary, Huntingdon Valley, PA, United States: Farlex, Inc., 2011.

[8] D. Shi, “The flexibility of Chinese syntax and the theory of syntax,” *Contemporary linguistics,* vol. 2, no. 1, pp. 18-26, 2000.

[9] Z. Wu and G. Tseng, “Chinese text segmentation for text retrieval: Achievements and problems,” *Journal of the American Society for Information Science,* vol. 44, no. 9, p. 532–542, 1993.

[10] M. Zhang, G. Zhou and A. Aw, “Exploring syntactic structured features over parse trees for relation extraction using kernel methods,” *Information processing & management,* vol. 44, no. 2, pp. 687-701, 2008.

[11] H. Jung, S.-K. Song, S. Lee and S.-P. Choi, Survey on Kernel-based Relation Extraction,

INTECH Open Access Publisher, 2012.

[12] A. Sharma, S. R. Swaminathan and H. Yang, “A Verb-Centric Approach for Relationship Extraction in Biomedical Text,” in *Semantic Computing (ICSC), 2010 IEEE Fourth International Conference*, Pittsburgh, PA, USA, 2010.

[13] Q. Song, Y. Watanabe and H. Yokota, “Relationship extraction methods based on co-occurrence in web pages and files,” in *Proceedings of the 13th International Conference on Information Integration and Web-based Applications and Services*, Ho Chi Minh City, Vietnam, 5-7 Dec. 2011.

[14] M. Weeber, H. Klein, L. T. d. J.-v. d. Berg and R. Vos, “Using concepts in literature-based discovery: Simulating Swanson's Raynaud–fish oil and migraine–magnesium discoveries,” *Journal of the American Society for Information Science and Technology,* vol. 52, no. 7, pp. 548-557, 2001.

[15] K. Fundel, R. Küffner and R. Zimmer, “RelEx—Relation extraction using dependency parse trees,” *Bioinformatics,* vol. 23, no. 3, pp. 365-371, 2007.

[16] R. Feldman, Y. Regev, M. Finkelstein-Landau, E. Hurvitz and B. Kogan, “Mining biomedical literature using information extraction,” *Current Drug Discovery,* vol. 2, no. 10, pp. 19-23, 2002.

[17] A. Panchenko, S. Adeykin, A. Romanov and P. Romanov, “Extraction of Semantic Relations between Concepts with KNN Algorithms on Wikipedia,” in *Proceedings of Concept Discovery in Unstructured Data Workshop (CDUD) of International Conference On Formal Concept Analysis*, Leuven, Belgium, May 2012.

[18] S. Zhao and R. Grishman, “Extracting relations with integrated information using kernel methods,” in *Proceedings of the 43rd Annual Meeting on Association for Computational Linguistics*, Ann Arbor, Michigan, US, 2005.

[19] C. Manning and H. Schütze, Foundations of Statistical Natural Language Processing, Cambridge, MA: MIT Press, May 1999.

[20] J. C. Park, H. S. Kim and J. J. Kim, “Bidirectional incremental parsing for automatic pathway identification with combinatory categorial grammar,” *Pacific Symposium on Biocomputing,*

vol. 6, pp. 396-407, 2001.

[21] L. V. Subramaniam, S. Mukherjea, P. Kankar, B. Srivastava, V. S. Batra, P. V. Kamesam and R. Kothari, “Information Extraction from Biomedical Literature: Methodology, Evaluation and an Application,” in *Proceedings of the twelfth international conference on Information and knowledge management*, New Orleans, LA, USA, Nov. 2003.

[22] M. Bundschus, M. Dejori, M. Stetter, V. Tresp and H.-P. Kriegel, “Extraction of semantic biomedical relations from text using conditional random fields,” *BMC Bioinformatics,* vol. 9, no. 207, 2008.

[23] S. Soderland, “Learning Information Extraction Rule for Semi-Structured and Free Text,” *Machine Learning,* vol. 34, no. 1-3, pp. 233-272, 1999.

[24] Y. Ogawa and M. Iwasaki, “A New Character-based Indexing Organization Using Frequency Data for Japanese Documents,” in *Proceeding SIGIR '95 Proceedings of the 18th annual international ACM SIGIR conference on Research and development in information retrieval*, New York, NY, USA, 1995.

[25] L.-F. Chien, “Fast and Quasi-Natural Language Search for Gigabytes of Chinese Texts,” in *Proceedings of the 18th annual international ACM SIGIR conference on Research and development in information retrieval*, New York, NY, USA, 1995.

[26] J.-Y. Nie, M. Brisebois and X. Ren, “On Chinese text retrieval,” in *SIGIR '96 Proceedings of the 19th annual international ACM SIGIR conference on Research and development in information retrieval*, New York, USA, 1996.

[27] T. Liang, S. Lee and W. Yang, “Optimal Weight Assignment for a Chinese Signature File,” *Information Processing and Management,* vol. 32, no. 2, pp. 227-237, 1996.

[28] J. H. Lee, J. H. Shin and J. S. Ahn, “An Effective Indexing Method for Korean Text Retrieval,” in *Proceedings of the Workshop on Information Retrieval with Oriental Languages*, Korea, 1996.

[29] L.-F. Chien, “PAT-tree-based keyword extraction for Chinese information retrieval,” in *Proceeding SIGIR '97 Proceedings of the 20th annual international ACM SIGIR conference on Research and development in information retrieval*, New York, NY, USA, 1997.

[30] P. Fung, “Extracting Key Terms from Chinese and Japanese texts,” *Computer Processing of Oriental Languages,* vol. 12, no. 1, pp. 99-121, 1998.

[31] I. Pinchuck, Scientific and technical translation, Boulder, Colorado: Westview Press, 1977.

[32] J. C. Sager, A practical course in terminology processing, Amsterdam, The Netherlands: John Benjamins Publishing, 1990.

[33] Y. Ma, G. Su, J. Li and S. Li, “A novel text subject extraction method,” in *Natural Language Processing and Knowledge Engineering, Proceedings, 2003 International Conference*, Beijing, China, 26-29 Oct. 2003.

[34] K.-J. Chen and S.-l. Liu , “Word Identification for Mandarin Chinese Sentences,” in *Proceedings of the 14th conference on Computational linguistics-Volume 1. Association for Computational Linguistics*, Stroudsburg, PA, USA, 1992.

[35] C. Faloutsos, “Access Methods of Text,” *ACM Computing Surveys (CSUR),* vol. 17, no. 1, pp. 49-74, 1985.

[36] W. B. Croft, “Clustering Large Files of Documents Using the Single-Link Method,” *Journal of the American Society for Information Science,* vol. 28, no. 6, pp. 341-344, 1977.

[37] N. J. Belkin and W. B. Croft, “Information Filtering and Information Retrieval: Two Sides of the Same Coin?,” *Communications of the ACM,* vol. 35, no. 12, pp. 29-38, 1992.

[38] D. D. Lewis and K. S. Jones, “Natural Language Processing for Information Retrieval,” *Communications of the ACM,* vol. 39, no. 1, pp. 92-101, 1996.

[39] L.-F. Chien, “Csmart - A High Performance Chinese Document Retrieval System,” in *Proceedings of the 1995 International Conference of Computer Processing of Oriental Languages (ICCPOL)*, Honolulu, Hawaii, 1995.

[40] T.-Y. Li, L. Liu, D.-W. Zhao and Y. Cao, “Eliciting Relations from Requirements Text Based on Dependency Analysis,” *Jisuanji Xuebao(Chinese Journal of Computers),* vol. 36, no. 1, pp. 54-62, 2013.

[41] “Language Technology Platform Cloud,” Research Center for Social Computing and Information Retrieval at Harbin Harbin Institute of Technology, [Online]. Available:

http://www.ltp-cloud.com/. [Accessed 2013].

[42] H. Zan and X. Zhu, “NLP oriented studies on Chinese functional words and the construction of their generalized knowledge base,” *Contemporary Linguistics,* vol. 2, pp. 124-135, 2009.

[43] Q. Chen, “Study on the Machine Tractable Thesaurus Dictionary of Contemporary Chinese Functional Words for Information Processing and Design Information Terms for Dictionary Entries,” *Computational Linguistics and Chinese Language Processing,* vol. 10, no. 4, pp. 459-472, 2005.

[44] S. Yu, X. Zhu and Y. Liu, “The Development of Knowledge-base of Generalized Functional Words of Contemporary Chinese,” *Journal of Chinese Language and Computing,* vol. 13, no. 1, pp. 89-98, 2003.

[45] S. Yu, X. Zhu and Y. Liu, “NLP oriented studies on Chinese functional words (面向自然语言理解的汉语虚词研究),” *民族语言文字信息技术研究——第十一届全国民族语言文字信息学术研讨会论文集,* vol. 277, p. 270, 2007.

[46] J. Zhang and H. Zan, “Automatic Recognition Research on Chinese Adverb DOU’s Usages,” *Acta Scientiarum Naturalium Universitatis Pekinensis,* vol. 49, no. 1, 2013.

[47] H.-Y. Zan, J.-H. Zhang, X.-F. Zhu and S.-W. Yu, “Research on Usages of Chinese Adverb JIU and Its Automatic Identification,” *Journal of Chinese Information Processing,* vol. 24, no. 5, pp. 10-16, 2010.

[48] A. K. Ramani, R. C. Bunescu, R. J. Mooney and E. M. Marcotte, “Consolidating the set of known human protein-protein interactions in preparation for large-scale mapping of the human interactome,” *Genome Biology,* vol. 6, no. 5, 2005.

[49] R. Bunescu, R. Ge, R. J. Kate, E. M. Marcotte, R. J. Mooney, A. K. Ramani and Y. W. Wong, “Comparative Experiments on Learning Information Extractors for Proteins and their Interactions,” *Artificial Intelligence in Medicine,* vol. 33, no. 2, pp. 139-155, 2005.

[50] L.-L. Gu and S.-Y. Sun, “Approach to Chinese ontology non-taxonomic relation extraction based on semantic dependency,” *Computer Engineering and Design,* vol. 33, no. 4, pp. 1676-1681, 2012.

[51] “汉典,” [Online]. Available: http://www.zdic.net/. [Accessed 2013].

[52] Z. Dong, “Logic Semantics and its application on Machine Translation,” *Chinese Machine Translation,* pp. 25-45, 1984.

[53] Y. Liu, “A Review of Chinese Vocabulary Statistics Studies,” *Chinese Academic Journal Electronic Publishing House,* no. 1, 2009.

[54] Z. Wang, Dictionary of Functional Words in Contemporary Chinese (现代汉语虚词词典), Shanghai: Shanghai Lexicographical Publishing House, 1998.

[55] Y. Chen, D.-Q. Zheng and T.-J. Zhao, “Chinese Relation Extraction Based on Deep Belief Nets,” *Journal of Software,* vol. 23, no. 10, pp. 2572-2585, 2012.

[56] “Princeton University "About WordNet." WordNet.,” Princeton University., 2010. [Online]. Available: http://wordnet.princeton.edu/. [Accessed 2014].

[57] K. K. Schuler, “VerbNet: A broad coverage, comprehensive,” Department of Linguistics, University of Colorado Boulder , [Online]. Available: http://verbs.colorado.edu/~mpalmer/projects/verbnet.html. [Accessed 2014].

[58] “GATE - General Architecture for Text Engineering,” The University of Sheffield, [Online]. Available: https://gate.ac.uk/. [Accessed 2013].

[59] “A brief introduction of LTP-cloud and the service it provided,” Research Center for Social Computing and Information Retrieval at Harbin Harbin Institute of Technology, [Online]. Available: http://www.ltp-cloud.com/intro/en/. [Accessed 2013].

[60] “People's Daily Processed Text Corpus,” University of Beijing, [Online]. Available: http://www.icl.pku.edu.cn. [Accessed 2014].

[61] F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot and É. Duchesnay, “Scikit-learn: Machine Learning in Python,” *Journal of Machine Learning Research,* vol. 12, p. 2825−2830, 2011.

[62] “sklearn.svm.SVC,” scikit learn, [Online]. Available: http://scikit-

learn.org/stable/modules/generated/sklearn.svm.SVC.html. [Accessed 2015].

[63] “sklearn.svm.LinearSVC,” scikit learn, [Online]. Available: http://scikit-learn.org/stable/modules/generated/sklearn.svm.LinearSVC.html. [Accessed 2015].

[64] “sklearn.naive_bayes.BernoulliNB,” scikit learn, [Online]. Available: http://scikit-learn.org/stable/modules/generated/sklearn.naive_bayes.BernoulliNB.html. [Accessed 2015].

[65] “sklearn.naive_bayes.GaussianNB,” scikit learn, [Online]. Available: http://scikit-learn.org/stable/modules/generated/sklearn.naive_bayes.GaussianNB.html. [Accessed 2015].

[66] “sklearn.linear_model.SGDClassifier,” scikit learn, [Online]. Available: http://scikit-learn.org/stable/modules/generated/sklearn.linear_model.SGDClassifier.html. [Accessed 2015].

[67] “sklearn.linear_model.RidgeClassifier,” scikit learn, [Online]. Available: http://scikit-learn.org/stable/modules/generated/sklearn.linear_model.RidgeClassifier.html. [Accessed 2015].

[68] “sklearn.linear_model.Perceptron,” scikit learn, [Online]. Available: http://scikit-learn.org/stable/modules/generated/sklearn.linear_model.Perceptron.html. [Accessed 2015].

[69] “sklearn.linear_model.PassiveAggressiveClassifier,” scikit learn, [Online]. Available: http://scikit-learn.org/stable/modules/generated/sklearn.linear_model.PassiveAggressiveClassifier.html. [Accessed 2015].

[70] “sklearn.neighbors.KNeighborsClassifier,” scikit learn, [Online]. Available: http://scikit-learn.org/stable/modules/generated/sklearn.neighbors.KNeighborsClassifier.html. [Accessed 2015].

[71] “sklearn.neighbors.NearestCentroid,” scikit learn, [Online]. Available: http://scikit-learn.org/stable/modules/generated/sklearn.neighbors.NearestCentroid.html. [Accessed 2015].

[72] “sklearn.tree.DecisionTreeClassifier,” scikit learn, [Online]. Available: http://scikit-learn.org/0.15/modules/generated/sklearn.tree.DecisionTreeClassifier.html. [Accessed 2015].

[73] “3.2.3.3.1. sklearn.ensemble.RandomForestClassifier,” scikit learn, [Online]. Available: http://scikit-learn.org/0.15/modules/generated/sklearn.ensemble.RandomForestClassifier.html. [Accessed 2015].

[74] “sklearn.ensemble.AdaBoostClassifier,” scikit learn, [Online]. Available: http://scikit-learn.org/0.15/modules/generated/sklearn.ensemble.AdaBoostClassifier.html. [Accessed 2015].

[75] “sklearn.lda.LDA,” scikit learn, [Online]. Available: http://scikit-learn.org/stable/modules/generated/sklearn.lda.LDA.html. [Accessed 2015].

[76] “sklearn.qda.QDA,” scikit learn, [Online]. Available: http://scikit-learn.org/stable/modules/generated/sklearn.qda.QDA.html. [Accessed 2015].

[77] “NLPIR,” Huaping, Zhang, [Online]. Available: http://ictclas.nlpir.org/. [Accessed 2014].

[78] F. Sebastiani and C. N. D. Ricerche, “Machine Learning in Automated Text Categorization,” *ACM Computing Surveys,* vol. 34, no. 1, p. 1–47, 2002.

[79] G. Salton and C. Buckley, “Term-weighting approaches in automatic text retrieval,” *Information Processing and Management,* 1988.

[80] H. P. Luhn, “A statistical approach to the mechanized encoding and searching of literary information,” *IBM Journal of Research and Development,* vol. 1, no. 4, pp. 309-317, 1957.

[81] J. Graupmann and R. Schenkel, ““The Light-Weight Semantic Web: Integrating Information Extraction and Information Retrieval for Heterogeneous Environments,” *SIGIR 2005 Workshop on Heterogeneous and Distributed Information Retrieval (HDIR),* 2005.

[82] R. Gaizauskas and A. Robertson, “Coupling information retrieval and information extraction: a new text technology for gathering information from the Web,” in *Computer-Assisted Information Searching on Internet Conference(RIAO 1997)*, Montreal, Canada, 1997.

[83] J. Cowie and Y. Wilks, “Information Extraction,” *Communications of the ACM,* vol. 39, no. 1, pp. 80-91, 1996.

[84] D. Freitag, “Multistrategy learning for information extraction,” in *Proceedings of the Fifteenth International, Machine Learning Conference*, 1998.

[85] S. P. Choi, C. H. Jeong, Y. S. Choi and S. H. Myaeng, “Relation Extraction based on Extended Composite Kernel using Flat Lexical Features,” *Journal of KIISE : Software and Applications,* vol. 36, no. 8, 2009.

[86] A. Culotta and J. Sorensen, “Dependency Tree Kernels for Relation Extraction,” in *Proceedings of the 42nd Annual Meeting on Association for Computational Linguistics*, 2004.

[87] X. Tong, M. Cui and G. Song, “Research on Chinese Text Automatic Categorization Based on VSM,” in *Wireless Communications, Networking and Mobile Computing, 2007. WiCom 2007, International Conference*, 2007.

[88] M. Jiang, L. Wang, Y. Lu and S. Liao, “A RBF Network for Chinese Text Classification Based on Concept Feature Extraction,” *Neural Information Processing, Lecture Notes in Computer Science,* vol. 4234, pp. 285-294, 2006.

[89] B. Rink, S. Harabagiu and K. Roberts, “Automatic extraction of relations between medical concepts in clinical texts,” *Journal of the American Medical Informatics Association,* 2011.

[90] “sklearn.naive_bayes.MultinomialNB,” scikit learn, [Online]. Available: http://scikit-learn.org/stable/modules/generated/sklearn.naive_bayes.MultinomialNB.html. [Accessed 2015].

[91] A. Skusa, A. Rüegg and J. Köhler, “Extraction of biological interaction network from scientific literature,” *Briefings in Bioinformatics,* vol. 6, no. 3, pp. 263-276, 2005.

[92] H. W. Chun, Y. Tsuruoka, J. Kim, R. Shiba, N. Nagata, T. Hishiki and J. Tsujii, “Extraction of gene-disease relations from Medline using domain dictionaries and machine learning,” *Pac Symp Biocomput,* pp. 4-15, 2006.

[93] R.-q. Lin, J.-x. Chen, X.-f. Yang and H.-l. Xu, “Research on Multi-information Fusion Chinese Relation Extraction Technology,” *Journal of Xiamen University(Natural Science),* 2011.

[94] V. Garcia, E. Debreuve and M. Barlaud, “Fast k Nearest Neighbor Search using GPU,” *Computer Vision and Pattern Recognition Workshops, 2008. CVPRW '08. IEEE Computer Society Conference,* pp. 1-6, 2008.

[95] T.-S. Chua and J. Liu, “Learning Pattern Rules for Chinese Named Entity Extraction,” *AAAI/IAAI,* pp. 411-418, 2002.

[96] N. Bach and S. Badaskar, “A Survey on Relation Extraction,” *Literature review for Language and Statistics II,* 2007.

[97] R. C. Bunescu and R. J. Mooney, “A Shortest Path Dependency Kernel for Relation Extraction,” in *Proceedings of the Human Language Technology Conference and Conference on Empirical Methods in Natural Language Processing*, Stroudsburg, PA, USA, 2005.

[98] M. E. Califf and R. J. Mooney, “Relational learning of pattern-match rules for information extraction,” in *Proceeding AAAI '99/IAAI '99 Proceedings of the sixteenth national conference on Artificial intelligence and the eleventh Innovative applications of artificial intelligence conference innovative applications of artificial intelligence*, Menlo Park, CA, USA, 1999.

[99] C. Li, M. Liakata and D. Rebholz-Schuhmann, “Biological network extraction from scientific literature: state of the art and challenges,” *Briefings in Bioinformatics,* vol. 15, no. 5, pp. 856-877, 2014.

Appendix

Appendix 1. Functional words in definition of ER

During the early stages of manual ER annotation, it is found that most ERs have a functional word nearby that is crucial to the interpretation of the words in the ER. Building upon this observation, these crucial functional words were added to the definition of ER, that is, a new ER entity type: KEY FUNCTIONAL (KF) is included as a necessary component in an ER. KEY is used to distinguish amongst the possibly many functional words in the sentence, which functional word is facilitating this particular ER. This new ER component was very helpful to the sentence extraction method (6.5.1). Many sentences amongst those retrieved by the presence of EFFECT words did not contain any ERs. Functional words were used as a second level of filter, and most sentences retrieved contain at least one valid ER. It is observed that ERs with certain KF entities share similar meaning, and this led to ERs being categorised into "5C" classes.

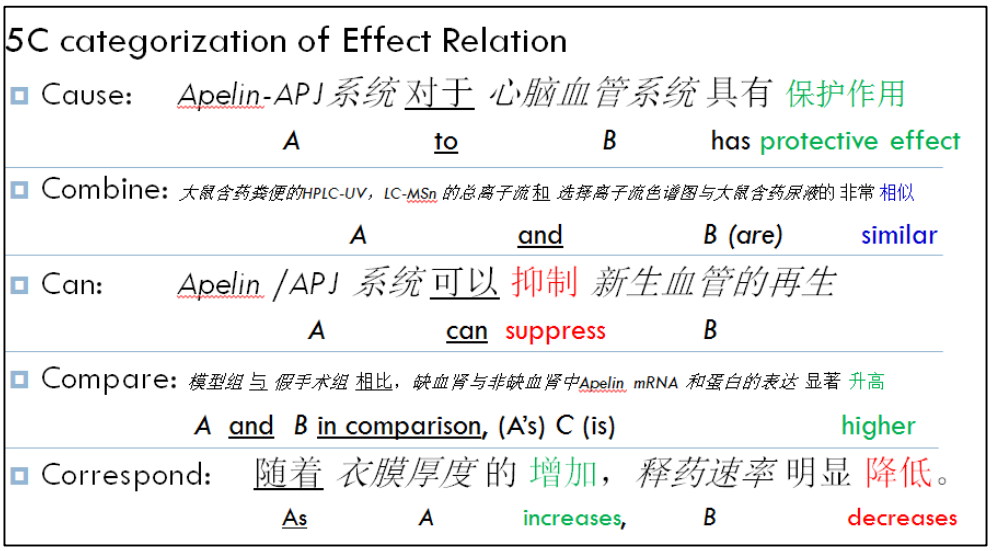

Figure 41: 5C ER categorization

As the KF entities underlined in Figure 41, there are 5 categories:

- Cause: A's effect on B. KF examples: "对于", "对".
- Combine: A and B share a similar (neutral) relation. KF examples: "联合", "一起".
- Can: A can have an effect on B. KF examples: "可以", "能够".
- Compare: The relation of A and B, in terms of the difference in their common property C. KF example: "比较", "相比".
- Correspond: The effect on A, as B is affected. KF examples: "随着", "越"

However as the TCM dataset grew, more examples were found to hold the concept of ER but not facilitated by any functional words. The definition of ER is revised to necessarily require functional words and the 5C categorization system was discarded. Manual ER annotation was performed again to retrieve possible ERs missed out due to the necessity of

functional words in the previous round of annotation. The current ER annotations in use do not necessarily contain a functional word, and resemble closest to the examples of 5C categories: “Cause” and “Can”.

Appendix 2. Dataset TCM journal papers – by disease categories

Disease Category	Number of journal paper
糖尿病 diabetes	10
膀胱炎 cystitis	1
腹泻 diarrhoea	1
肾炎 nephritis	1
肺癌 lung cancer	1
肺炎 pneumonia	1
风湿病 rheumatism	1
淋巴结炎 lymphadenitis	1
关节炎 arthritis	1
高血压 high blood pressure	1
盆腔炎 pelvic inflammatory disease	1
前列腺炎 prostatitis	1
脑炎 encephalitis	1
乳腺癌 breast cancer	1
冠心病 coronary heart disease	1
咽炎 pharyngitis	1
扁桃体炎 tonsillitis	1
肝炎 hepatitis	1
肠胃炎 gastroenteritis	1
肠癌 colorectal cancer	1

Appendix 3. Dataset TCM journal papers – by journal name

Journal name	Number of journal paper
江苏中医药 Jiangsu Journal of Traditional Chinese Medicine	2
中医临床研究 Clinical Journal of Chinese Medicine	2
中医药导报 Guiding Journal of Traditional Chinese Medicine and Pharmacy	2
云南中医中药杂志 Yunnan Journal of Traditional Chinese Medicine and Materia Medica	1
内蒙古中医药 Nei Mongol Journal of Traditional Chinese Medicine	1
中国中医药科技 Chinese Journal of Traditional Medical Science and Technology	1
中国中医药现代远程教育 Chinese Medicine Modern Distance Education of China	1
中国药学报 Chinese Medicine and Pharmacology	1
甘肃中医 Gansu Journal of TCM	2
光明中医 Clinical Journal of Guang Ming Traditional Chinese Medicine	2
中外医学研究 Chinese and Foreign Medical Research	1
中国现代实用医学杂志 Chinese Journal of Current Practical Medicine	1
吉林中医药 Jilin Journal of Traditional Chinese Medicine	1
中成药 Chinese Traditional Patent Medicine	1
中国中西医结合杂志 Chinese Journal of Integrative Medicine	1
安徽中医临床杂志 Clinical Journal of Anhui Traditional Chinese Medicine	1
上海医学 Shanghai Medicine Journal	1
中国现代医药杂志 Modern Medicine Journal of China	1
实用中医药杂志 Journal of Practical Traditional Chinese Medicine	1
浙江中西医结合杂志 Zhejiang Journal of Integrated Traditional Chinese and Western Medicine	1
北京中医药 Beijing Journal of Traditional Chinese Medicine	1
医学信息 Medical Information	1
中医中药 China Journal of Chinese Materia Medica	1
中国医学创新 Medical Innovation of China	2

Appendix 4. Experiment 6: Evaluation of classification models

Experiment

The aim of this experiment is to:

- Evaluate the performance of the feature-based word classification method and feature-based relation classification method over different classification models.
- Compare and determine the classifier to use for Experiment 3 (7.4.1.2) and Experiment 4 (7.4.2).

A list of the commonly used classifiers is collected from Scikit learn [61]:

- SVM (Support Vector Machine) [62]
- Linear SVM [63]
- Bernoulli Naive Bayes [64]
- Gaussian Naïve Bayes [65]
- SGD (Stochastic Gradient Descent) Classifier [66]
- Ridge Classifier [67]
- Perceptron [68]
- Passive Aggressive Classifier [69]
- K Nearest Neighbours Classifier [70]
- Nearest Centroid [71]
- Decision Tree Classifier [72]
- Random Forest Classifier [73]
- Ada Boost Classifier [74]
- LDA (Linear Discriminant Analysis) [75]
- QDA (Quadratic Discriminant Analysis) [76]

There will be 2 experiments to test the word classification method, using different datasets: golden input and actual input (7.1.2). The setups of these experiments are given in Table 42 below.

Test Parameter	Value
Method:	Word classification
Dataset:	Golden input
Feature/s:	Text, POS tag, EFFECT keywords
Classifier/s:	*Listed above*
Test Measured:	F-score
Evaluation:	Evaluation 1 + 10-fold (7.1.4.4)

Table 42: Experiment 6A Setup

The setups of Experiment 6A will test the performance of different classifiers, given golden input (only sentences containing at least one ER). A simplified set of features: text, POS and EFFECT keyword (explained in 6.5.2), is used to reduce run time.

There will be 2 experiments to test the relation classification method, using different datasets: balanced ratio dataset and actual ratio dataset (7.1.1). Both the balanced ratio and actual ratio dataset contain golden NER results (7.1.2). The setups of these experiments are given in Table 43 and Table 44 below.

Test Parameter	Value
Method:	Relation classification
Dataset:	Balanced ratio dataset
Feature/s:	Word distance
Classifier/s:	*Listed above*
Test Measured:	F-score
Evaluation:	Evaluation 2 + 10-fold (7.1.4.4)

Table 43: Experiment 6B Setup

Test Parameter	Value
Method:	Relation classification
Dataset:	Actual ratio dataset
Feature/s:	Word distance
Classifier/s:	*Listed above*
Test Measured:	F-score
Evaluation:	Evaluation 2 + 10-fold (7.1.4.4)

Table 44: Experiment 6C Setup

The setups of Experiment 6B and 6C will test the performance of different classifiers in relation classification, given balanced ratio dataset and actual ratio dataset. Only the feature: word distance (explained in 6.6), is used to reduce run time. Experiment 3C will also show which classifiers are more impacted by the unbalanced cases in the actual ratio dataset (7.1.1).

Results

The F-score of the feature-based NER module with different classifiers, using golden input is reported in Table 45 and represented in the graph in Figure 42.

	6A: Word Classification
	Golden (F-score)
SVM	0.00
Linear SVM	0.69
Bernoulli Naive Bayes	0.68
Gaussian Naïve Bayes	0.22
SGD Classifier	0.00
Ridge Classifier	0.14
Perceptron	0.00
Passive Aggressive	0.00
K Nearest Neighbours	0.30
Nearest Centroid	0.40
Decision Tree	0.71
Random Forest	0.61
Ada Boost	0.67
LDA	0.69
QDA	0.01

Table 45: Results – Word classification classifier comparison (Exp. 6A)

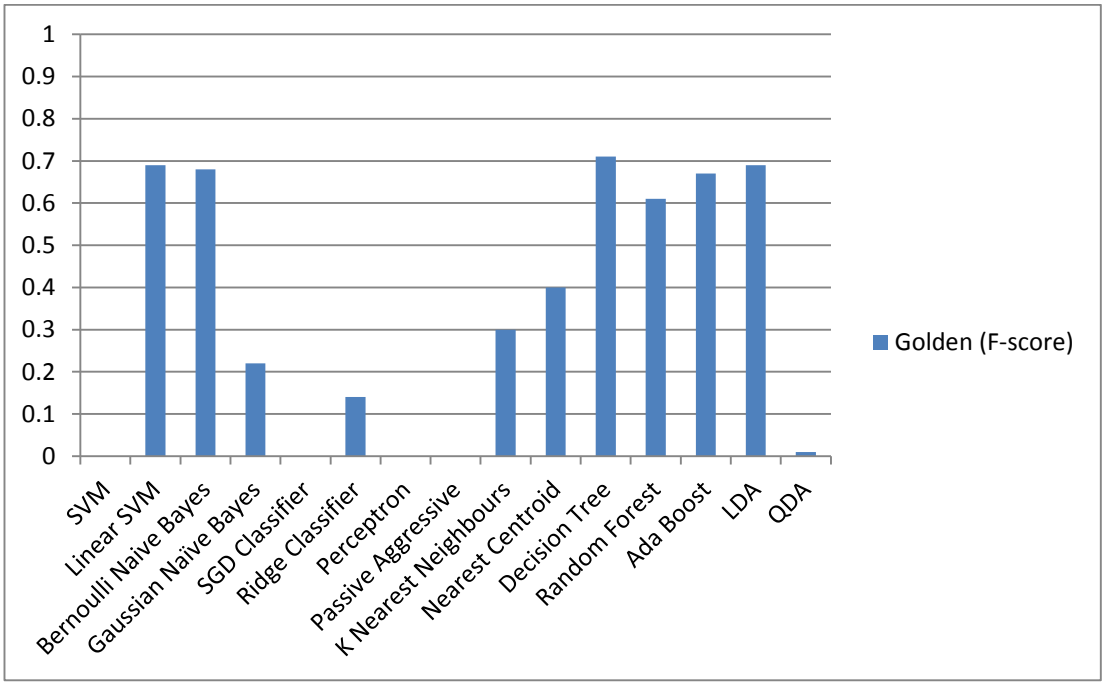

Figure 42: Result Graph – Word classification classifier comparison (Exp. 6A)

In the result of Experiment 3A above, Linear SVM, Decision Tree and LDA have the highest score, 0.69, 0.71 and 0.72 respectively. Bernoulli Naïve Bayes, Random Forest and Ada Boost scored slightly lower than the aforementioned classifiers, ranged within 0.56 and 0.68. Gaussian Naïve Bayes, Ridge Classifier, K Nearest Neighbours and Nearest Centroid scored poorly, ranged within 0.13 and 0.4. SVM, SGD Classifier, Perceptron, Passive Aggressive and QDA scored 0 or close to 0. They are not suitable for classifying this classification problem in the NER module.

The F-score of the feature-based relation classification approach amongst multiple classifiers on the balanced and actual ratio datasets is reported in Table 46 and represented in the graph in Figure 43.

	6C & 6D: Relation Classification	
	Balanced (F-score)	Actual (F-score)
SVM	0.80	0.79
Linear SVM	0.74	0.74
Bernoulli Naive Bayes	0.76	0.69
Gaussian Naïve Bayes	0.71	0.69
SGD Classifier	0.72	0.39
Ridge Classifier	0.64	0.67
Perceptron	0.62	0.56
Passive Aggressive	0.62	0.88
K Nearest Neighbours	0.80	0.79
Nearest Centroid	0.69	0.63
Decision Tree	0.82	0.74
Random Forest	0.81	0.79
Ada Boost	0.78	0.73
LDA	0.71	0.64
QDA	0.68	0.67

Table 46: Results – Relation Classification classifier comparison (Exp. 6B & 6C)

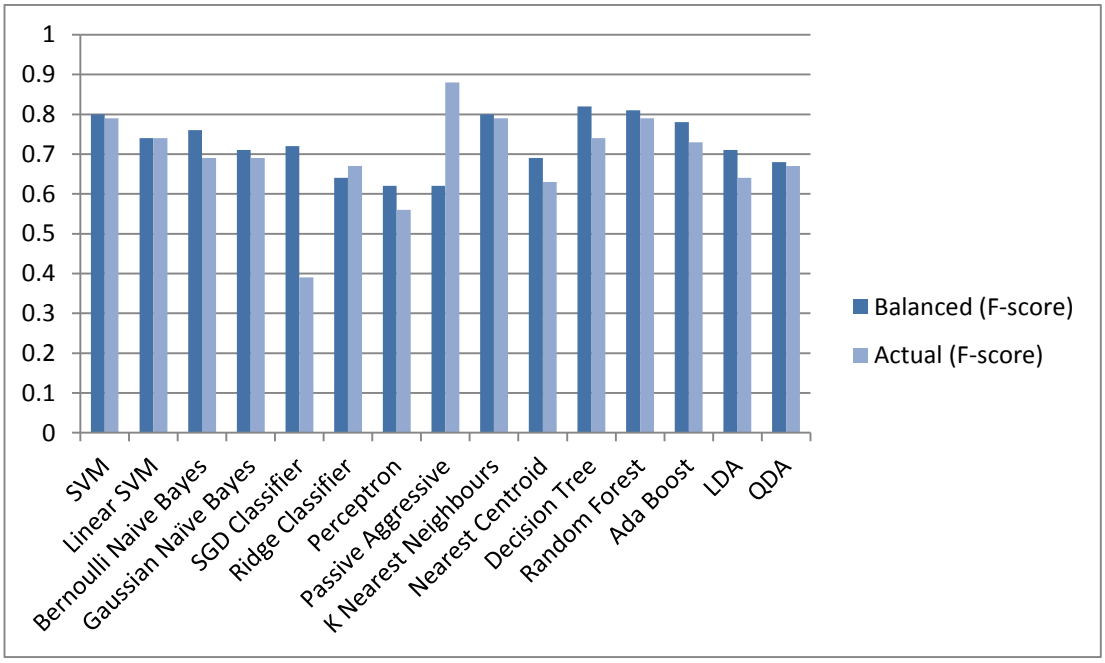

Figure 43: Result Graph – Relation Classification classifier comparison (Exp. 6B & 6C)

In the results of Experiment 3C and 3D above, SVM, K Nearest Neighbours, Decision Tree and Random Forest have the highest score for balanced ratio dataset (0.8, 0.8, 0.82 and 0.81 respectively). However Decision Tree shows a significantly drop in its actual ratio dataset score (from 0.82 to 0.74), while the other 3 shows little difference (≈0.02).

SGD Classifier shows to have a significant weakness in dealing with actual ratio dataset (decrease from 0.72 to 0.39). Ridge Classifier and Passive Aggressive shows to be strengthened in dealing with actual ratio dataset. Passive Aggressive particularly has a significant increase (from 0.62 to 0.88).

The other classifiers have a lower balanced ratio dataset score and similar drop in actual ratio dataset score.

Overall, the Decision Tree classifier shows high performance in the results of this experiment. It is selected as the classification model for the word classification method in Experiment 3 (7.4.1) and the relation classification method in Experiment 4 (7.4.2).