CHINESE-ENGLISH CROSS-LINGUAL INFORMATION RETRIEVAL IN BIOMEDICINE USING ONTOLOGY-BASED QUERY EXPANSION

A thesis submitted to the University of Manchester
for the degree of Doctor of Philosophy
in the Faculty of Engineering and Physical Sciences

2011

By
Xinkai Wang
School of Computer Science

Contents

Word Count: 42,734

List of Tables

List of Figures

Abstract

In this thesis, we propose a new approach to Chinese-English Biomedical cross-lingual information retrieval (CLIR) using query expansion based on the eCMeSH Tree, a Chinese-English ontology extended from the Chinese Medical Subject Headings (CMeSH) Tree.

The CMeSH Tree is not designed for information retrieval (IR), since it only includes heading terms and has no term weighting scheme for these terms. Therefore, we design an algorithm, which employs a rule-based parsing technique combined with the C-value term extraction algorithm and a filtering technique based on mutual information, to extract Chinese synonyms for the corresponding heading terms. We also develop a term-weighting mechanism. Following the hierarchical structure of CMeSH, we extend the CMeSH Tree to the eCMeSH Tree with synonymous terms and their weights.

We propose an algorithm to implement CLIR using the eCMeSH Tree terms to expand queries. In order to evaluate the retrieval improvements obtained from our approach, the results of the query expansion based on the eCMeSH Tree are individually compared with the results of the experiments of query expansion using the CMeSH Tree terms, query expansion using pseudo-relevance feedback, and document translation. We also evaluate the combinations of these three approaches.

This study also investigates the factors which affect the CLIR performance, including a stemming algorithm, retrieval models, and word segmentation.

Declaration

No portion of the work referred to in this thesis has been submitted in support of an application for another degree or qualification of this or any other university or other institute of learning.

Copyright

Acknowledgements

I would like to express my deep appreciation to Prof. Sophia Ananiadou for her constructive comments and advice of this thesis and for her four-year supervision.

I would also like to thank Mr. Paul Thompson and Dr. Scott Songlin Piao for their assistance and help during my PhD study.

Personal thanks must also go to my friends at the University of Manchester, Mr. Geng Li, Mr. Roufei He, Dr. Naikuo Yang, Dr. Hongfen Zhou, and Dr. Jian Yu, to whom I am greatly indebted for giving me ideas and perspectives.

Lastly, I would like to acknowledge and thank my wife and my parents for their support and encouragement.

Chapter 1

Introduction

1.1 Research Motivations

No one would doubt that the online search services like “Google” and “Yahoo”, have become a popular and effective way to access information and to gain knowledge. Compared to the traditional media, like books and broadcast programmes, search engines can return a collection of relevant documents to the user within seconds; and the contents of these documents are always up to date. Such features of the online search engines make them particularly suitable for information access in specific domains, such as biomedicine, which are undergoing a knowledge explosion.

However, precise acquisition of biomedical information via search services is not an easy task, because the biomedical concepts or terms are complicated and sophisticated, while the precision of retrieval significantly depends on the words or terms used in the queries. Different users have different goals. For example, general users with no biomedical background are hardly able to understand terms in this domain, let alone form effective queries using them.

Despite the wide usage of the online search engines in daily life, popular engines like Google are not designed to meet the requirements of acquiring precise biomedical information. They are general-purpose search engines; and their retrieval parameters are optimised using the training data composed of news reports and articles.

For the Chinese language, there are extra limitations which restrict biomedical information access:

(1) The language barrier hinders the effective access of biomedical information.

In general, up-to-date biomedical information and knowledge are usually conveyed in English; however, most of the people in China do not understand the English used in biomedicine.

(2) The existing search engines are “monolingual” information retrieval systems. For instance, “PubMed” [1], which is a famous biomedical search engine and is designed for retrieval of English domain documents, responds to the information need by “no relevant documents found”, if the keyword “癌” (meaning “cancer”). Since it is difficult for Chinese general users to understand English biomedical documents, the current strategy to spread the latest biomedical information is to translate it into Chinese before usage. However, the translated knowledge may fall behind the latest achievements. Moreover, the translation may be misrepresented or may contain errors. An improved strategy is to make use of online search engines to provide the user with the Chinese documents as well as the relevant English ones. But the existing search engines are “monolingual” information retrieval systems, where queries and the documents are expressed in the same language. Cross-lingual information retrieval (CLIR), which allows the query and documents to use different languages, is designed to satisfy such information requests.

(3) There are only two Chinese search services in the biomedical domain. The China National Knowledge Infrastructure (CNKI) [2], which is one of the biggest databases of Chinese journals and academic publications, is only accessible to subscribers. Another bibliographic literature database available is the China TCM Patent Database (CTCMPD), but its performance is not reliable.

(4) On the other hand, the international databases include only limited Chinese medical literate data. According to [FTP$^+$08], for example, only 10 of the traditional Chinese medical journals, out of 149, are indexed by MEDLINE [3].

In a word, the existing search engines are not designed for general users to access biomedical information expressed in Chinese. They are aimed at either

[1] http://www.ncbi.nlm.nih.gov/pubmed/

[2] http://www.global.cnki.net/

[3] http://www.nlm.nih.gov/databases/databases_medline.html

searching general articles like news reports, or providing a service only to professional users. The public thirst for biomedical knowledge and information is ignored, so it is necessary to study new techniques to conquer these shortcomings and make the professional knowledge open to general users.

In the past decades, many researchers have investigated cross-lingual information retrieval. Some [PS08, LT07, AF01] use bilingual dictionaries or lexicons to translate queries before retrieving; but they have been proven to give a poor performance due to the lack of contextual information. Some [LCC02, BPPM93] apply parallel or comparable domain document collections (corpora) to translate queries; however, collecting and processing them are time-consuming tasks. And others [CG04, Oar98] translate documents instead of translating queries; although this approach has the best results, its quality depends on the machine translation software, and is not suitable for the rapidly growing document collections.

Therefore, in this thesis, we propose a new approach to biomedical cross-lingual information retrieval. We use an improved bilingual biomedical concept hierarchy to expand Chinese queries, and then translate them into English using a bilingual dictionary. The novelty of this method is that it overcomes the disadvantages of a dictionary-based approach and does not require a large volume of resources as corpora-based methods do.

1.2 Research Methodologies

In knowledge acquisition, ontologies "reflect the structure of the domain and constrain the potential interpretations of terms." [SAMK05, p.239]. They provide a vocabulary which represents the concepts and their properties, and their relationships. Chinese Medical Subject Headings (CMeSH), which is the Chinese translation of Medical Subject Headings (MeSH), is a biomedical structured terminology.

In this thesis, we propose a new approach to Chinese-English cross-lingual information retrieval in the area of biomedicine, which depends on bilingual ontologies. This approach is based on the hypothesis that *a biomedical CLIR can perform better using weighted bilingual ontologies to expand queries than using classic dictionary translation approaches.* In this sense, we present our approach to creating such ontologies and use them to improve Chinese-English biomedical CLIR.

The original MeSH and CMeSH have limitations when applied in information retrieval: 1) Many synonyms or variants for each concept are ignored in MeSH-like resources, because they are originally designed for indexing or cataloguing biomedical articles in libraries. 2) The original MeSH and CMeSH do not provide any weighting to indicate the importance of each concept.

In order to solve these problems, we propose the following techniques:

(1) We propose an automatic approach to extract concepts/terms using the Google search service. The key techniques involved are both of the C-value term extraction algorithm [FAM00] and mutual information. The extracted Chinese terms are treated as synonyms of the original terms and are finally applied to expand queries.

(2) We also propose an algorithm to calculate a weight for each extracted Chinese concept. Such weight plays an important role in retrieval performance.

After the extension using the above-mentioned algorithms, the original CMeSH becomes the extended CMeSH, which we call *eCMeSH*. eCMeSH not only retains the hierarchy of MeSH, but also includes many variant terms with term weights. Chapter 3 presents the detail of this extension.

The focus of our research is the improvement of Chinese-English CLIR in biomedical literature. We propose the following approach to retrieval improvement using eCMeSH:

(1) We construct a prototype Chinese-English cross-lingual information retrieval application using a bilingual domain dictionary. Queries are Chinese sentences; and the document set consists of English full-text articles. This prototype CLIR is the baseline of this study, obviously. We will describe the baseline CLIR in Chapter 4.

(2) We improve the prototype CLIR by query expansion using eCMeSH. The eCMeSH terms are added to original queries according to two strategies: a) A query is expanded using the synonyms of each original query term. This expansion policy is aimed to determine the contribution of the vocabulary provided by eCMeSH to the CLIR. b) A query is expanded using the terms which share the “is-a” relationship with the original query term.

This strategy evaluates the CLIR retrieval performance obtained using the relations among terms in eCMeSH.

In order to evaluate the performance of our ontology-based query expansion, we compare our approach with the query expansion using the original CMeSH terms, the query expansion using pseudo-relevance feedback and the document translation approach. Pseudo-relevance feedback is a retrieval improvement technique, which automatically takes the results that are initially returned from a given query and then uses information about whether or not those results are relevant to perform a new query. Chapter 4 and 5 not only present the details of these experiments but also evaluate and analyse the results of experiments.

(3) We study the factors which affect the performance of our Chinese-English biomedical CLIR, such as retrieval module parameters and Porter stemming algorithm (a technique which removes affixes from terms) before indexing. Some studies, such as Bennett et al. [BSU08], found that optimal smoothing parameters were dependent on the collection and the query set. The results of Salton [Sal68] indicate that the effect of stemming depends on the nature of the vocabulary used. We investigate module parameters and the effect of stemming in Chapter 4.

1.3 Research Aims

The main objective of the research presented in this thesis is the development of a new approach using ontology based query expansion, in order to retrieve information from biomedical literature expressed in English using Chinese queries.

In order to achieve this objective, we adopt the following methodology:

- We present and discuss the probabilistic models and language models for monolingual information retrieval, reviewing the existing approaches to improvement of retrieval performance.

- We review the main approaches to cross-lingual information retrieval such as non-translation- and translation-based approaches, discussing their advantages and disadvantages.

- We investigate the existing approaches to biomedical information retrieval for both monolingual and cross-lingual IR.
- We discuss the approaches to term/concept extraction, concentrating on the C-value algorithm.
- We analyse the features of MeSH-like resources, discussing their limitations and the possible solutions.
- We summarise the existing approaches to the construction and evaluation of ontologies, in order to design algorithms to extend ontologies and calculate term weights.
- We evaluate the effect of model parameters on the performance of Chinese-English cross-lingual information retrieval, optimising the parameters for a probabilistic model and language model, respectively.
- We determine the performance that our new approach achieves, comparing it with document translation and pseudo-relevance feedback.

1.4 Research Contributions

The main contribution of our research is Chinese-English CLIR for biomedical literature.

Although information retrieval has been investigated over past decades, and now has successfully been applied in business, no study on cross-lingual information retrieval for the Chinese biomedical domain had been carried out before our work. Our study, where we try to search English biomedical papers and articles using Chinese queries, reports the algorithms, resources, and techniques required to the biomedical Chinese-English CLIR to give good results. It shares our experiences with other researchers in the community.

The second innovative aspect of this research lies in the extension of the original CMeSH terms. Research of biomedical information retrieval related to the Chinese language is rarely reported. The reasons are: 1) No Chinese document collection and corresponding gold standards in biomedicine are available. Usually the gold standard, which records the relevant documents in a document collection for each topic, is provided within the document collection. For Chinese, however, there is no biomedical document collection designed for information retrieval.

2) Essential linguistic resources in Chinese, such as parallel or comparable corpora, bilingual domain dictionaries, and ontologies are required to perform CLIR. Despite several Chinese/English dictionaries available to the biomedical domain, they are inadequate for information retrieval, since the relations among terms are not presented in them. The parallel or comparable corpora and other linguistic resources like ontologies can play a more important role to improve retrieval performance. 3) There are fewer Chinese syntactic parsers designed for the biomedical domain. Although some named entity extraction tools specially designed for Chinese language in this domain have been reported in recent years, utilities able to identify the constituents of the Chinese sentences from biomedical literature may provide more useful elements; in this case, the structure of sentences and the meanings of words and phrases may forge a new method to improve CLIR.

Our study does not attempt to provide solutions to all these issues. We have constructed new bilingual ontologies for the biomedical domain, which we call "eCMeSH". Unlike the original CMeSH, the new resource includes variant terms for concepts, as we discovered that the poor vocabulary coverage of the CMeSH limits its usage in CLIR. In addition, each Chinese term in eCMeSH has been assigned a term weight, since we believe the term weight can improve CLIR performance. In our study, the newly designed ontologies have proven to improve the performance of cross-lingual information retrieval. Chapter 5 evaluates the improvements of retrieval performance after using eCMeSH. We expect that they can be helpful in other tasks of natural language processing.

The third contribution is that the C-value algorithm re-developed is improved for extracting biomedical concepts from Chinese texts. The original algorithm was developed for English. We have redesigned it for Chinese.

Finally, this study has its economic value. The techniques involved in the study can be used to quickly develop a business application to retrieve English biomedical documents using Chinese queries, and with minor modification, it would be easy to create an English-Chinese information retrieval application in the area of biomedicine. In addition, the technique is not limited to the biomedical domain, and could be expanded to the newswire domain, which is more popular and closer to daily life.

1.5 Thesis Overview

This thesis is structured as follow:

Chapter 1 introduces the research motivations, methodologies, objectives, and aims. It also summarises the innovations and contributions of our study.

Chapter 2 gives background knowledge about information retrieval and cross-lingual information retrieval, which is necessary to understand our research. It discusses the state of the art of biomedical information retrieval and Chinese information retrieval. Section 2.1 introduces the background knowledge about information retrieval, and reviews previous studies on improving retrieval performance. Section 2.2 focuses on the techniques used to improve cross-lingual information retrieval, discussing their advantages and disadvantages. Section 2.3 investigates the studies on biomedical information retrieval with their achievements and limitations. Section 2.4 discusses the state of biomedical information retrieval on the Chinese language, summarising the additional difficulties in Chinese language processing. Section 2.5 is a summary of the chapter.

Chapter 3 describes the algorithm of extending the original CMeSH. Section 3.1 reviews some related research works, such as EuroWordNet, and summarises the advantages and disadvantages of the C-value algorithm. Section 3.2 introduces the background of MeSH and CMeSH, discussing their limitations. Section 3.3 illustrates the algorithm to extract candidate terms and calculate their weights. Section 3.4 discusses the evaluation of the extended CMeSH concepts. Section 3.5 summarises the entire chapter.

Chapter 4 discusses the process of conducting the following experiments: the baseline experiments, the experiments to optimise model parameters, and the experiments to evaluate query expansion using eCMeSH. Section 4.1 reviews the measures applied to evaluate information retrieval. Section 4.2 explains the common settings for the entire experiments, such as the toolkit exploited to establish experiments and pre-processing of document collection and query sets. Section 4.3 is the description of the baseline experiment, including the experimental steps and results and analysis. Section 4.4 is concentrated on the experiments to optimise the model parameters. Section 4.5 describes the experiments to evaluate query expansion using the eCMeSH Tree, which is compared with query expansion using the CMeSH Tree terms, query expansion using pseudo-relevance feedback, and document translation. Section 4.6 summarises the above experiments.

Chapter 5 puts the emphasis on improving retrieval performance using hybrid approaches. Section 5.1 attempts to combine query expansion using the eCMeSH Tree with that using pseudo-relevance feedback. Section 5.2 describes the experiments which firstly translate document collections into Chinese and then apply the eCMeSH Tree to expand queries. Based on the first two experiments, Section 5.3 focuses on the experiments using the eCMeSH Tree and pseudo-relevance feedback on the translated Chinese documents. Section 5.4 compares and analyses the experiments of hybrid approaches. Section 5.5 summarises these experiments.

Chapter 6 concludes our study. Section 6.1 reviews the whole research. Section 6.2 summarises the achievements in the study. Section 6.3 discusses the limitations of this research, and proposes future work.

Chapter 2

IR and CLIR Methodologies

Since the first idea of using computers to search for relevant information [Bus45], information retrieval (IR) has become a mature technology to discover relevance among documents, not only in the newswire domain but also in special domains. In this thesis, IR is limited to text retrieval. This chapter starts with the retrieval models and the techniques used to improve monolingual information retrieval; then it reviews approaches to cross-language environments in biomedicine; and finally it discusses the information retrieval methods applied in Chinese.

2.1 Information Retrieval

The term "information retrieval" was first coined by Mooers [Moo50]. After many early studies, such as [Luh59, Sal75, SYY75, SWY75], IR came to maturity in the mid 1990s. In this section, IR refers to "monolingual information retrieval", where queries and documents are presented in the same language.

2.1.1 The Definition

According to Manning et al. [MRS08, p.1], "information retrieval" refers to the technology of "finding material (usually documents) of an unstructured nature (usually text) that satisfies an information need from within large collections (usually stored on computers)".

The workflow of information retrieval is illustrated as Figure 2.1, which can be separated as three sections: the first focuses on techniques to prepare documents for retrieval; the second presents algorithms used to parse users' queries and then improve these queries; and the third describes the retrieval engine itself.

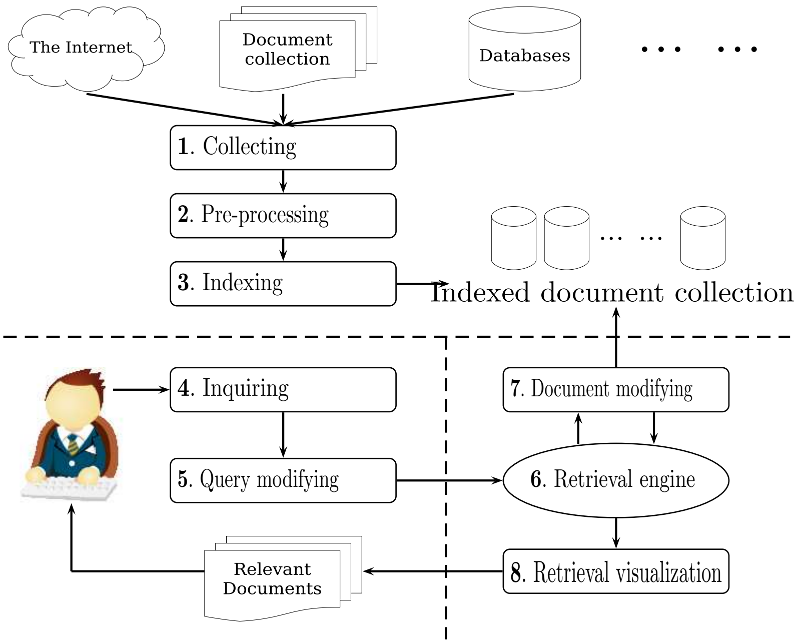


Figure 2.1: The workflow of information retrieval

The first step is collecting documents from multiple sources, such as online documents, databases, etc.

Before indexing the documents, several pre-processes are required:

(1) In general, the words which have too high or too low frequency are removed from documents at this stage, because they scarcely contribute to the performance of retrieval. This process is known as "stopping". However, stopping is not mandatory. Researchers, such as Jiang and Zhai [JZ07] and Trieschnigg [THdJK10], draw different conclusions on stopping.

(2) Generally, punctuation is ignored, except for special requirements. For example, hyphens, dots, and the percentage mark may be kept as part of index terms when the document collection is on biochemistry.

(3) For documents in alphabetic writing, upper-case words may need to be converted to lower case or vice versa, which is described as "case folding".

(4) Stemming, which remove affixes (usually suffixes) from words, may be applied, in order to reduce the size of dictionary used for indexing.

(5) *N-gram* may be required. Although in literature N-gram can include the notion of any co-occurring set of characters in the stream (e.g., an N-gram can consist of the first and last characters of a word), in this thesis N-gram is a contiguous N-character slice of a text stream. Here is an example of a tri-gram. The word "NAME" is composed of the N-grams: "_NA", "NAM", "AME", "ME_", and "E__", where "_" represents blanks. The N-gram model has two functions: i) For some languages like Chinese, where it is difficult to establish an appropriate dictionary for the purpose of indexing, N-grams are treated as index terms. ii) Under the circumstances that all information needs to be preserved, an N-gram is able to retain all symbols in documents.

The core technique at the document collection stage is indexing. The idea of indexing is to represent each document using a set of words/terms. The location of each term in each document is recorded; and the importance of each term to the document is calculated by term weights.

Queries need to be split into index terms before retrieval. The techniques applied to pre-process documents are used to handle queries. Since the original query terms are usually ambiguous, they need to be improved before use. The extended terms may come from dictionary-like resources, corpora, or the relevant contexts of the initial query feedback. In the case of cross-lingual information retrieval (CLIR), "query translation", may be required.

Document modification is an optional module, which includes i) document expansion, which employs the related corpora to find relevant contents and add these contents to documents; and ii) document translation, an approach to CLIR instead of translation of the query.

2.1.2 Terminology and Information Retrieval

According to Ananiadou [Ana88, pp.4–5], the definition of terminology adopted by the International Association for Terminology (TERMIA) is that "Terminology is concerned with the study and use of the systems of symbolic and linguistic signs employed for human communication in specialised areas of knowledge and activities. It is primarily a linguistic discipline — linguistics being interpreted here in its widest possible sense — with emphasis on semantics (systems of meanings and concepts) and pragmatics. It is interdisciplinary in the sense that it also borrows concepts and methods from semiotics, epistemology, classification, etc. It is closely linked to the subject fields whose lexicon it describes and for which it seeks to provide assistance in the ordering and use of designations. Although Terminology has became in the past mostly concerned with the lexical aspects of specialised languages, its scope extends to other levels, namely syntagmatics, syntax, etc. In its applied aspect Terminology is related to Lexicography and uses techniques of Information Science and Technology". In summary, terminology is the study of *terms* and their use.

"Terms" as used in terminology is different from the index term which is applied to information retrieval. The former is the linguistic representation of a concept in a particular subject field, and the latter is a word that captures the essence of the topic of a document. Index terms compose a controlled vocabulary, which are used as keywords to retrieve documents in an information system such as a catalogue or search engine. The automatic term recognition technique, which is the (semi)automatic aid designed for discovering potential terms, keeping track of the life-cycle of terms, etc. in the amount of specialised texts [Ana94], used here is *C-value* algorithm [FAM00].

Although terminology is not connected to information retrieval in any way, it is obviously believed that terms and their relations in specific domains may be used to improve the retrieval performance of IR. Mayr and Petras [MP08], for example, report their evaluation of the impact of terminology mappings on recall and precision in IR. They utilise "cross-concordances" to translate query terms into other terminologies to facilitate the search across different databases and terminologies. The cross-concordances correct ambiguities and imprecision in query formulation. The results of their experiments show that both recall and precision benefit from the term mapping.

In this thesis, the meaning of a term depends on its context: A term used to describe the procedure of IR is considered as the index term; a term related to MeSH, CMeSH, eCMeSH, and domain dictionaries refers to the term in terminology.

2.1.3 Retrieval Models

We review four retrieval models: the Boolean model; vector space model; probabilistic models; and language models and the early research on syntactic indexing. It is worth describing probabilistic and language models in some detail, because we carry out our experiments using these models.

2.1.3.1 The Boolean Model

The *Boolean retrieval model* is a model for information retrieval where queries are presented in a Boolean expression of terms.

The Boolean operators include *AND*, *OR*, and *NOT*, which connect terms to form a query. The operators *AND* and *OR* affect performance in opposite ways. The more *OR* operators that are used in a query, the more extraneous items are retrieved, which reduces the retrieval precision. On the other hand, the *AND* operator tends to increase retrieval precision, while recall declines.

The advantage of the Boolean model is the high precision — a document either matches the query or it does not. Therefore, it is the preference of professional users, such as the users of legal documents. By 2007, the majority of law librarians still recommended terms and connectors for high recall searches.

However, this model has its own problems: i) Boolean queries are difficult to formulate. Fox [Fox86] illustrated several operations needed to formulate a Boolean query: removal of high-frequency terms, additional synonyms and alternate spellings; moreover, it is hard to insert extra terms that are not originally included; ii) most applications of the Boolean model do not provide the assignment of term weights, on which the query-document relevance measurement depends; iii) the retrieved documents are usually presented in a random order, that is, with no ranking, because the Boolean model does not provide an estimate of the query-document relevance; iv) the size of the subset of documents to be returned is difficult to control; and v) it is difficult or impossible to find a satisfactory middle ground between *AND* and *OR*. Salton [Sal86] proposed a compromise by the use of a query formulation that is neither too broad nor too narrow.

Several studies [Sal82, SFW83] have extended the base Boolean model to add term weighting and output ranking features.

2.1.3.2 The Vector Space Model

The *vector space model* (VSM) [Sal71] uses a ranking algorithm that tries to rank documents according to the overlap between the query terms and document terms [Boo82].

In this model, all queries and documents are represented as vectors in $|V|$-dimensional space, where V is the set of all distinct terms in the document. The vector space model requires the following calculations, where the model for term weight is called *term frequency - inverse document frequency* ($tf - idf$) model: i) the weight of each index term within a given document, which points out how important the term is within a single document. This weight is usually calculated using *term frequency* (tf); ii) the weight based on the *document frequency* (df), i.e. the number of documents a term appears in. In practice, this is usually taken to be the *inverse document frequency* (idf) for scaling purpose. The effect is to boost the weight of a term that occurs in fewer documents over a term that occurs in many, as it is more discriminating; iii) the similarity measure of the query vector and the document vectors, which indicates which document comes closest to the query, and ranks the others by the closeness of the fit. *Cosine similarity* is frequently used to calculate this similarity.

Compared with the Boolean retrieval model, the vector space model has a couple of advantages: i) it is a simple model based on linear algebra; ii) term weights are not binary; iii) it provides for computing a continuous degree of similarity between queries and documents; iv) ranking documents is performed according to the similarity measure; and v) it is possible to only match a part of a document.

However, there are a few limitations to the vector space model: i) terms are assumed to be independent of each other; ii) long documents are poorly represented due to poor similarity values; iii) query terms must precisely match document terms, otherwise substrings of terms could result in a false match; and iv) it is difficult to take into account the order of terms appearing in a document.

The vector space model is applied not only to document or text retrieval but also to other information retrieval related applications, such as topic tracking [CFK$^+$04], text categorization [Joa97, Hul94], and collaborative filtering [SN00].

2.1.3.3 The Probabilistic Retrieval Model

Probabilistic retrieval models are used to estimate the probability of documents being relevant to a query [RSJ76, RvRP81]; this assumes that the terms are distributed differently in relevant and non-relevant documents.

Probabilistic models are based on the "probability ranking principle" (PRP) [vR79, pp. 113–114], which proposes that all documents can be simply ranked in decreasing order of the probability of relevance with respect to the information need. Ripley [Rip96] proves that the PRP is optimal, but it requires that all probabilities are known correctly. In practice, this is impossible.

In order to estimate the probability of relevance of the document to a query, the *binary independence model* (BIM) is introduced. “Binary” means that documents and queries are both represented as binary term vectors. “Independence” indicates that terms occurring in documents independently, that is the presence or absence of a term in a document, is independent of the presence or absence of any other term. The similarity of the document relevant to a query is calculated as Equation 2.1:

$$\begin{aligned} sim(d,q) &= \sum_{t \in q} \left(\log \frac{(s+0.5)/(S-s+0.5)}{(df_t - s + 0.5)\,/\,(N - df_t + 0.5)} \right) \\ &= \sum_{t \in q} \left(\log \frac{(s+0.5)\,(N - df_t + 0.5)}{(df_t - s + 0.5)\,(S - s + 0.5)} \right) \end{aligned} \tag{2.1}$$

where d a document

q a query

t a term

s the number of documents which contain the term t and are relevant to the query q

S the total number of documents relevant to the query q

N the number of all documents in the collection

df_t the document frequency, the number of documents in the collection that contain the term t

0.5 the value used to overcome the problem of zero probability

The problem of BIM is that it was originally designed for short catalogue records and abstracts of fairly consistent length, and does not consider the term frequency and document length carefully.

The *2-Poisson model* proposed by Bookstein and Swanson [BS75] assumes that a term plays two different roles in documents: in documents with a low average number of term occurrences, the term should not be used as an index term; in documents with a high average number of term occurrences, the term is a good index term. Robertson and Walker [RW94] present an IR model approximating the 2-Poisson model, known as the *Okapi weighting scheme*. Among the

Okapi BM series, *Okapi BM25* [RWHB$^+$92] is the optimal result. Okapi BM25 is estimated by trial and error. Its name is derived from "BM", which means "Best Match", and a version number of the last trial, 25, which was a combination of BM11 and BM15 [RWJ$^+$94]. In this thesis, we follow the explanations expressed by Spärck Jones et al. [SJWR00] and Robertson et al. [RvRP81]. Okapi BM25 described in this section ignores the relevance feedback information.

Equation 2.2 is the representation of Okapi BM25, considering the term frequency, the document length, and duplicate query terms.

$$\begin{aligned} \mathit{BM}25(d,q) = \sum_{t\in q}\Bigg(& \log\frac{N-\mathit{df}_t+0.5}{\mathit{df}_t+0.5}\times\frac{(k_1+1)\cdot \mathit{tf}_{d,t}}{k_1\times\left((1-b)+b\cdot\frac{|d|\times N}{\sum_{i\in C}|d_i|}\right)+\mathit{tf}_{d,t}} \\ & \times\frac{(k_3+1)\;\mathit{tf}_{q,t}}{k_3+\mathit{tf}_{q,t}}\Bigg) \end{aligned} \tag{2.2}$$

where	d	a document
	$\lvert d\rvert$	the length of the document d, calculated according to term
	q	a query
	t	a term
	C	the document collection
	N	the number of all documents in the collection
	df_t	the document frequency, the number of documents in the collection that contain the term t
	$\mathit{tf}_{d,t}$	the document term frequency, the number of occurrences of the term t within the document d
	$\mathit{tf}_{q,t}$	the query term frequency, the number of occurrences of the term t within the query q
	b, k_1, k_3	the Okapi BM25 tuning parameters

The first item in Equation 2.2 is derived from Equation 2.1; the second item uses the term frequency and the document length to adjust the similarity; the last item is related to the duplicate query term.

As for the parameters, Robertson and Walker [RW99] recommend the value for k_1 is 1.2, which is the lower end of the range suggested by Spärck Jones et al. [SJWR00]. Robertson and Walker [RW99] and Spärck Jones et al. [SJWR00] suggest the optimal value for b is 0.75. k_3 controls the importance of the query term. If it is set to 0, then only one instance of each term contributes to the similarity; when it is equal to a very large value, query terms contribute as much as they occur. In most cases, this factor has little impact on the final similarity, as most queries are short and do not contain duplicate terms.

Logistic Regression [CCG94] and *Pircs* [Kwo96] are two well known probabilistic models. Although they perform well, studies, for example, Luk and Kwok [LK02], show that the 2-Poisson model with Okapi BM25 weighting scheme exceeds other probabilistic models.

2.1.3.4 The Language Model

The language model [PC98, Hei98, BL99, MLS99] is based on the idea that a document is a good match to a query if the document model is likely to generate the query, which will in turn happen if the document contains the query words a number of times.

In practice, the language model for IR is based on the unigram model, because the *unigram model* is sufficient to judge the topic of a text. In addition, the unigram model is more efficient to estimate and apply than higher-order models.

The language model has many variant realisations. Lafferty and Zhai [LZ01] lay out three ways to establish language models. Figure 2.2 illustrates these approaches, where, (1) is the query likelihood language model, which uses documents to generate a query; (2) is the document likelihood language model, where the query model is used to estimate documents; (3) is the model comparison approach.

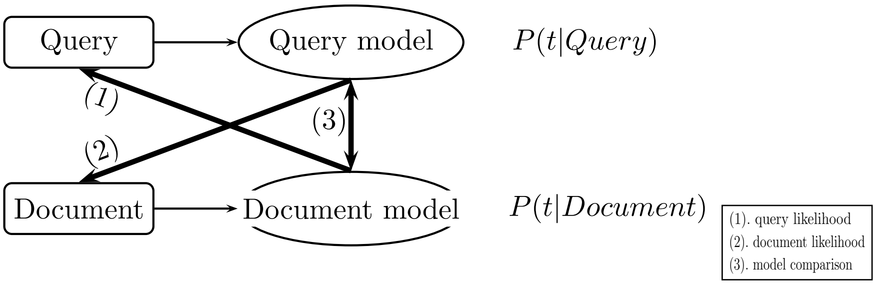


Figure 2.2: Three approaches to the implementation of language models

In the *query likelihood model*, a language model M_d constructed from each document d in the collection is applied to model the query generation process. The probability $P(d|q)$, where the probability of a document is interpreted as the likelihood that it is relevant to the query, is used to rank relevant documents.

Equation 2.3 illustrates the calculation of this language model.

$$
\begin{aligned}
\hat{P}(q|M_d) =& \prod_{i=1}^{n} \hat{P}(t_i|M_d) \\
=& \prod_{i=1}^{n} \Big((1-\lambda)\hat{P}(t_i|d) + \lambda \hat{P}(t_i|C) \Big) \\
=& \prod_{i=1}^{n} \left((1-\lambda) \frac{tf_{d,t}}{|d|} + \lambda \frac{cf_t}{T} \right)
\end{aligned}
\tag{2.3}
$$

where	d	a document
	$\|d\|$	the length of the document d, calculated according to term
	q	a query
	t	a term
	C	the document collection
	M_d	the language model of the document d
	$tf_{d,t}$	the document term frequency, the number of occurrences of the term t within the document d
	cf_t	the number of the term t in the collection
	T	the number of tokens in the entire collection
	λ	the smoothing parameter of *Jelinek-Mercer* smoothing

Correctly setting λ is important for good performance. Zhai and Lafferty [ZL02] report that a wide range of values, typically between 0.2 and 0.8, usually around 0.7, achieves good performance.

An alternative language model is the *document likelihood model*. The problem of this approach is that there is much less text available to estimate a language model based on the query text. Queries are in common very short. For example, Jansen et al. [JSP05] report that 20% of web queries in 2002 only contained a single term. The sparseness of query texts causes the models derived from queries to be unreliable. Lavrenko [Lav04, pp. 69] has reported that document likelihood models perform poorly.

The third approach to the language model is the *model comparison*. Lafferty and Zhai [LZ01] use the *Kullback-Leibler (KL) divergence* between the document language model and the query likelihood model to model the risk of returning a document d as relevant to a query q. Equation 2.4 defines the model comparison.

$$R(d;q) = KL(M_d \| M_q) = \sum_{t \in V} P(t|M_q) \log \frac{P(t|M_d)}{P(t|M_q)} \tag{2.4}$$

where	d	a document
	q	a query
	t	a term
	M_d	the language model of the document d
	M_q	the language model of the query q
	V	the set of distinct terms in the vocabulary

Lafferty and Zhai [LZ01] demonstrate that the model comparison approach outperforms both query likelihood models and document likelihood models. Manning and Schütze [MS99] point out that KL divergence is not symmetric and does not satisfy the triangle inequality and thus is not a metric. Therefore, the problem of using KL divergence as a ranking function is that scores are not comparable across queries. Kraaij and Spitters [KS03a] suggest that the similarity can be modelled as a normalised log-likelihood ratio.

All language models are faced with the zero-frequency problem, in which the frequency of a term is zero because the term does not occur in the document collection; thus the probability involving this term is zero. Smoothing is a solution to this issue. In general, all smoothing techniques are attempting to discount the probabilities of the terms appearing in the documents, and then to assign the extra probabilities to the unseen terms. Considering the efficiency of computations over a large collection of documents, there are three smoothing techniques widely applied in language models: *Jelinek-Mercer smoothing*, *Dirichlet smoothing*, and *two-stage smoothing*. Chen and Goodman [CG96] and Zhai and Lafferty [ZL04] review these smoothing methods. Equation 2.3 is the result after using the Jelinek-Mercer smoothing method, where λ controls how many probabilities are assigned to the terms that do not occur in a document. An alternative smoothing technique is *Dirichlet smoothing*. After applying Dirichlet smoothing, the query likelihood language model can be computed as Equation 2.5.

$$
\begin{aligned}
\hat{P}(q|M_d) =& \prod_{i=1}^{n} \hat{P}(t_i|M_d) \\
=& \prod_{i=1}^{n} \left(\frac{\mathit{tf}_{d,t_i} + \mu \hat{P}\left(t_i|C\right)}{|d| + \mu} \right) \\
=& \prod_{i=1}^{n} \frac{\mathit{tf}_{d,t_i}}{|d| + \mu} + \prod_{i=1}^{n} \left(\frac{\mu}{|d| + \mu} \cdot \frac{\mathit{cf}_{t_i}}{T} \right)
\end{aligned}
\tag{2.5}
$$

where	d	a document
	$\lvert d\rvert$	the length of the document d, calculated according to term
	q	a query
	t	a term
	C	the document collection
	M_d	the language model of the document d
	$\mathit{tf}_{d,t}$	the document term frequency, the number of occurrences of the term t within the document d
	cf_t	the number of the term t in the collection
	T	the number of tokens in the entire collection
	μ	the smoothing parameter of *Dirichlet* smoothing

Zhai and Lafferty [ZL04] investigated the value of μ, and proposed that the optimal μ varies from collection to collection; the recommendation in their research is around 2,000.

The *two-stage smoothing* method is an improvement of Dirichlet smoothing, which uses a two-stage strategy. In the first stage, a query language model is smoothed using Dirichlet smoothing; and in the second stage, it is smoothed using Jelinek-Mercer. It has the advantage over Dirichlet smoothing that no tuning is needed and it performs well.

2.1.3.5 Efforts of Syntactic Indexing

The above-mentioned approaches to indexing are featured because of the neglect of linguistic characteristics. Another way to index documents is to parse texts using linguistic knowledge and to then form index terms.

Some researchers, for example Jones [Jon83] and Piternick [Pit84], believe that certain positions in a text, or certain constructions of word strings, are helpful for identifying potential index terms. Baxendale tried to extract potential index terms from the "topic sentence" [Bax58], the most important sentence in a paragraph, or from prepositional phrases [Bax61]. Clarke and Wall [CW65] use a restricted dictionary in a process to provide a surface parse of sentences in order to extract noun phrases. Salton [SM83]launched experiments on syntactic analysis for automatic indexing in the *SMART* system. Hillman [Hil68] developed the *LEADER* system, which concentrated on identifying referential noun phrases from complete documents. Maeda et al. [MMS80] were concerned with semantic analysis of abstracts, in order to extract significant phrases for indexing purposes. Dillon and McDonald [DM83] developed the *FASIT* system, which identifies content bearing textual units without a full parse, and, without using semantic criteria, groups these units into quasi-synonymous sets.

However, as Salton [SM83, p.131] has pointed out, "The early test results obtained with the SMART system showed that some complicated linguistic methodologies that were believed essential to attain reasonable retrieval effectiveness were in fact not useful in raising performance. In particular, the use of syntactic analysis procedures to construct syntactic content phrases [...] could not be proved effective under any circumstances." Linguistic analysis approaches to construction of index terms have not been successful. Spärck Jones and Kay [SJK73, p.119] explain one set of reasons: i) "Retrieval needs are not properly understood"; ii) "the value of the syntactic component of an index description is affected by other system components: it may either be that the correct relationships between different components have not been established, or that other components are defective"; and iii) "essentially inadequate or inappropriate methods of handling syntax have been adopted".

However, Researchers do not stop investigating the effect of syntactic indexing. Fagan [Fag87] compared the document retrieval performance using syntactic and non-syntactic (statistical) approaches to automatic phrase indexing. Experimental results show that the effect of non-syntactic phrase indexing is inconsistent. It is concluded that syntax-based indexing has certain benefits not available with the non-syntactic approach. Salton and Smith [SS89] summarised various linguistic approaches proposed for document analysis in IR, which include syntactic analysis in IR and use of syntax for index term generation.

Smeaton et al. [SOK94] reported their approach to IR based on a syntactic analysis of both document texts and queries. They used *tree structures* (TSAs), which are constructed on the clause level and thus each document and each query can yield many TSAs, to encode and capture language ambiguities. The degree of overlap between the TSAs from documents and those from queries are computed and aggregated to yield a score for each document, which is used in ranking the collection. However, their method achieved poor performance in terms of recall and precision. The possible reasons for the poor performance lie in i) documents are relevant to queries besides sharing structural syntactic relationships; ii) language analyser (ENGCG) may be of poor quality; iii) the type of language used in TREC topic description may affect retrieval performance.

Evans and Zhai [EZ96] described a hybrid approach to the extraction of meaningful subcompounds from complex noun phrases using both corpus statistics and linguistic heuristics. Experimental results show that indexing based on such extracted subcompounds improves both recall and precision in information retrieval.

Pohlmann and Kraaij [PK97] did the research on the effect of syntactic phrase indexing on retrieval performance for Dutch texts. They compared different choices for combining terms to form head-modifier pairs and studied the effect of adding none, one, or all constituent parts of the pair as a separate index term. The results of their experiments show that using head-modifier pairs as index terms can improve both recall and precision significantly (up to 25%) but only if all constituent parts are also added separately. They concluded that using both Noun-Adjective and Noun-Noun head-modifier pairs produced the best results. Kraaij and Pohlmann [KP98] described the results of experiments contrasting syntactic phrase indexing with statistical phrase indexing for Dutch texts. Experimental results showed that syntactic phrases are slightly superior to statistical phrases when used in indexing, particularly at hight precision and that at higher recall level syntactic and statistical phrases are equally effective.

In summary, although some studies have proven that applying syntactic analysis during indexing increases the performance of IR, researchers found that syntactic indexing cannot improve retrieval performance, such as [SJK73] and [SM83], that syntactic analysis reduces retrieval performance, like [SOK94], and that syntactic and statistical phrase indexing are equally effective [KP98]. Moreover, syntactic indexing approaches are heavily dependent on the quality of language analysers and dictionaries.

2.1.4 Improving Information Retrieval

Information retrieval can be treated as the match between query terms and index terms. In practice, the set of index terms hardly covers query terms, so improvements are necessary. The improvements take effect either on the query, known as "*query reformulation*", or on documents, called "*document expansion*".

2.1.4.1 Query Reformulation

Query reformulation is an attempt to improve poor queries by: adding terms that aid retrieval, subtracting terms that degrade retrieval performance, or re-weighting the existing or new query terms. Typically, query terms will not be removed, because it is hard to determine the irrelevance of terms. So techniques used to improve information retrieval via modifying queries are of two types: adding new terms with or without term weights; or re-weighting the existing terms without adding new terms. The former is called "query expansion"; and the latter is "relevance feedback".

1. Relevance Feedback

Relevance feedback (or more precisely, *interactive relevance feedback*) is a query improvement technique which involves the human user's judgement of the relevance or non-relevance of documents to queries. The *Rocchio algorithm* [Roc71] is the classic algorithm for implementing relevance feedback, which uses the vector space model to incorporate relevance feedback information. *Ide dec-hi* and *Ide regular* [Ide71] are similar to the Rocchio method. The probabilistic approach to relevance feedback [RSJ76, RvRP81] is based on the ratio between the probability of an item x (a term or a document) being relevant or not relevant. Magennis and van Rijsbergen [MvR97] have reported that interactive relevant feedback can increase effectiveness significantly. However, the main issue of interactive relevance feedback is that it requires users to determine the relevance of a document in an iterative process.

2. Query Expansion

Query expansion can be performed interactively, know as "*interactive query* *expansion*". During the retrieval session, a user chooses the expansion terms from a list of candidate terms. An important aspect of this technique is the determination of a relatively small set of query terms. Harman [Har88, Har92] and Magennis and van Rijsbergen [MvR97] have investigated interactive query expansion and reported a significant improvement in retrieval effectiveness.

The main disadvantage of interactive query expansion is that users generally dislike providing the relevance information [DMB98]. Moreover, as they lack control over the search process, more and more studies [BSAS94, RW99, RW00] are shifting to automatic query expansion approaches. There are two groups of approaches to automatic query expansion: expansion based on "knowledge structures" and expansion using an initial set of search results.

(1). Query Expansion using Knowledge Structures

These structures may be collection-independent resources, such as dictionaries and manually constructed thesauri. Query expansion based on these structures is also known as *external techniques*, because they do not make use of statistics in document collection. Voorhees [Voo94] used WordNet to expand queries and found that queries that do not describe the information need well can be improved significantly.

The knowledge structures can also be collection-dependent, like automatically constructed thesauri [QF93, JC94]. Expansion based on these structures is called global analysis techniques, which employ the term statistics in the entire document collection. The methods to construct these structures are: term co-occurrence [SvR83, SC99], term clustering [MWZ72], and latent semantic indexing [DFLD88, Dum94, DDFL90]. Crouch [Cro88] reports that automatically constructed thesauri can work better than manually constructed resources. Mandala et al. [MTT99] conclude that a combination of different kinds of thesauri is more useful for query expansion than any single kind.

(2). Query Expansion using Search Results

The issue of expansion based on knowledge structure is that the retrieval precision is decreased as the expansion terms may be too ambiguous to help differentiate relevance. One method to overcome this problem is *local analysis* [CH79]. It starts with an initial set of results using the original query; then a certain number of terms is selected from all terms that occur in the top documents;

finally, the terms with the highest score are added to the query. Experiments [MTT99, MTT00] show that query expansions based on local analysis perform better than those based on external knowledge structures. Xu and Croft [XC00] report that query expansion using local analysis excels those using thesauri and global analysis. Although local analysis generally performs better than other query expansion techniques, the local analysis is much less efficient than global analysis, since terms occurring in the local set of documents need to be assessed. Moreover, local analysis faces the increased risk of query drift [MSB98], as the top ranked documents are assumed to be relevant, but they may not be [CH79].

2.1.4.2 Document Expansion

Another approach to retrieval improvement is document expansion. Rather than expanding a query from initially retrieved documents, document expansion modifies documents by adding potential query terms that appear in similar documents. Singhal and Pereira [SP99] propose document expansion in the context of speech retrieval using a side corpus to provide related terms. Li and Meng [LM03] use document expansion to retrieve spoken documents. Lester and Williams [LW02] and Levow and Oard [LO02] apply document expansion for topic tracking. Tseng and Juang [TJ03] propose the document-self expansion used for text categorisation. The main issue of document expansion is that the expansion process is reasonably costly. Billerbeck [BZ05] reported on document retrieval experiments and concluded that corpus-based document expansion is not a promising approach and that other document expansion methods based on extracting terms from external resources give limited improvements in some circumstances.

2.2 Cross-Lingual Information Retrieval

The information retrieval discussed in Section 2.1 is generally known as "monolingual" information retrieval, where the documents in foreign languages are treated as unwanted noise [ATO05]. In cross-lingual information retrieval (CLIR), queries and documents are expressed in different languages. CLIR uses the techniques successful in monolingual IR: it uses the same indexing algorithms and retrieval models as classic IR and also employs various more sophisticated methods used in monolingual IR to improve retrieval performance. The basic idea and technique in performing cross-lingual information retrieval is translation [RB09], translating query or document manually or automatically; however, translation is not the only approach to CLIR.

2.2.1 Non-Translation Approaches

Cross-lingual information retrieval can be implemented using non-translation approaches, such as *cross-language latent semantic indexing* [DLLL97, LL90], *cognates matching* [BMWC00], and *cross-lingual relevance model* [LCC02].

The basic idea of using latent semantic indexing (LSI) in CLIR is that term-term inter-relationships are able to be automatically modelled and used to improve retrieval. LSI attempts to examine the similarity of the contexts where words appear, and to create a reduced-dimension feature space where words appearing in similar contexts are near each other. Singular value decomposition [DDFL90, FDD$^+$88], a method derived from linear algebra, is used to discover the associative relationships. Thus, it is unnecessary to exploit any external dictionaries or thesauri to determine word associations because they are derived from analysis of existing texts. In order to adapt LSI to CLIR, an initial sample of documents is translated by human experts or by machine, to create a set of bilingual training documents. The major problem of cross-language latent semantic indexing is that it is difficult to determine the best initial set of sample documents for large document collections. Moreover, the training texts depend on translation.

Buckley et al. [BMWC00] report their attempt at cross-lingual information retrieval with cognates matching. They assume that source and target languages share many cognates, which are “words that have a common etymological origin”, as in French and English. The query terms in the source language are treated as potentially misspelled target language words. Instead of using bilingual dictionaries, the source query is expanded by adding target words from the collection that are lexicographically nearby. It is obvious that this method is not suitable for the language pair where one language is distinct from the other, such as English and Chinese.

In the cross-lingual relevance model, the probabilities of each word in the target vocabulary to a set of target documents that are relevant to a source query is calculated from either a parallel corpus or a bilingual lexicon. Lavrenko et al. [LCC02]'s experiments show that the size of parallel corpus has a significant effect on the performance of cross-lingual retrieval models. Where a high-quality parallel corpus or bilingual lexicon is not available, cross-lingual relevance models lead to a drop in performance.

Although the above-mentioned methods do not translate queries or documents, they either depend on extra resources, or are restricted by difficult circumstances. Non-translation approaches are not in the mainstream of cross-lingual information retrieval.

2.2.2 Translation-Based Approaches

Translation-based approaches to CLIR make use of dictionaries, lexicons, parallel or comparable corpora, or machine translation software to translate queries, or documents, or both of them. Xu and Weischedel [XW05] investigate the impact of lexical resources on CLIR performance. They review several resources including bilingual term lists, parallel corpora, machine translation, and stemmers on Chinese, Spanish, and Arabic CLIR and conclude that a bilingual term list and parallel corpora lead to the best CLIR performance; it can rival monolingual performance, and in the case of no parallel corpus, pseudo-parallel texts generated by machine translation can partially overcome the lack of parallel text.

2.2.2.1 Query Translation

Machine readable dictionaries (MRD) are the most common resources used to translate queries. This approach is faster and simpler than translating documents [McC99]. However, query translation based on MRD suffers from incorrect word inflection, wrong compounds and phrases translation, and inadequate spelling variants and domain terms.

Another approach to query translation is the use of parallel or comparable corpora. Parallel corpora contain the same documents in more than one language, while comparable corpora cover the same domain and contain an equivalent vocabulary. These corpora commonly are aligned by some unit of language, such as the sentence.

Queries can also be translated using machine translation (MT) software. The advantage of MT lies in its high effectiveness for translating large texts. However, queries are usually short and thus provide little context for word disambiguation. Moreover, it is difficult for machine translation to handle the grammar of queries

[AF01]. So, CLIR is difficult if the translation is only based on MT.

There are several models to implement query translation. Lavrenko et al. [LCC02] proposed the approach to translate queries based on pseudo-relevance feedback. The query to translate is first used to search a parallel or comparable collection to obtain a set of relevant documents; and the translation terms in the nearest neighbouring documents are used as the translated query. Brown et al. [BPPM93] proposed five models for determining statistical translations based on a bilingual collection of sentences. The key point of these models is the estimation of an alignment of the sentences in the language pair. Terms in the sentences in source languages are connected to terms in the translated sentences in the target language by this alignment. These translation models can also be applied in CLIR tasks. Trieschnigg et al. [THdJK10] report their attempt to use IBM Model 1 to translate concepts and terms. Bian and Chen [BC98] and Gao et al. [GNZ06] apply pointwise mutual information, which indicates the association of two events based on their joint distribution in comparison to their individual probabilities, to filter ambiguous translations in their CLIR setting. However, Manning and Schütze [MS99] found that pointwise mutual information is not ideal for measuring the association between terms, because it is biased towards low-frequency words.

2.2.2.2 Document Translation

Instead of translating queries, another approach to CLIR is the translation of documents from the source language to the target language. Usually, this is done using MT software.

Studies have shown that document translation-based CLIR is typically better than query translation-based CLIR. Oard [Oar98] compared several query translation approaches with the document translation technique and concluded that document translation may result in further improvements in retrieval effectiveness under some conditions. Chen and Gey [CG04] report their monolingual, bilingual, and multilingual retrieval experiments using the CLEF 2003 test collection. They compared query translation-based multilingual retrieval with document translation-based multilingual retrieval, where documents are translated into the query language using MT systems or statistical translation lexicons derived from parallel texts. Their results show that document translation-based retrieval is slightly better than the query translation-based retrieval. Moreover, they suggest that combining both query translation and document translation in multilingual retrieval achieves the best performance.

However, document translation based on MT has limitations. The major problem is that MT is computationally expensive and sometimes impractical. For the large collections, machine translation is a time-consuming task. Therefore the document translation approach is not suitable for cross-lingual information retrieval where documents are added or removed frequently, or the content of documents varies rapidly. Other problems of document translation are the cost of the machine translation system and the lack of language pairs.

2.2.3 Challenges in CLIR

Each query translation and document translation encounter the problem of translation ambiguity, which is often rooted in *homonymy* and *polysemy* [MS99]. Homonymy refers to a word that has at least two entirely different meanings. Polysemy refers to a word that has two or more distinct but related meanings. It is difficult to determine the most appropriate translation from several choices in the dictionary.

The second problem that CLIR tasks have to face is inflection, especially in Western languages. This can be solved by stemming and lemmatization. Stemming is the technique where different grammatical forms of a word are reduced to a stem, which is the common part and usually shorter than these forms, by removing the word endings. Lemmatization is a technique that simplifies every word to its uninflected form or lemma.

The out-of-vocabulary (OOV) word refers to a word or a phrase that cannot be found in a dictionary. Cross-lingual information retrieval tasks are significantly affected by OOV words/terms. These unknown words degrade the performance of CLIR based on the dictionary-based translation, even with the best dictionaries. Generally, OOV terms are proper names or newly created words, including compound words, proper nouns and technical terms. Their translation is crucial for a well-performed CLIR. Although additional linguistic resources can improve translation, the common and simplest strategy used to handle untranslatable query terms is to include them in the new query represented by the target language. If these terms do not exist in the target language, the query will be less likely to retrieve the relevant documents. Correct phrase translation is also becoming one of the problems in CLIR. A phrase cannot be translated word by word [LXG07].

Correct recognition of named entities (NEs) plays an important role in improving the performance of CLIR. Bilingual dictionaries often have few entities for organisation, person and location names. When NEs are wrongly segmented as ordinary words and translated with a bilingual dictionary, the results are often poor.

2.2.4 Current Approaches

The usual method to improve CLIR is to exploit more linguistic resources. Wikipedia [1] has become an important resource in CLIR. Lin et al. [LWY$^+$09] have developed a Japanese-Chinese IR system based on the query translation approach. The system employs a more conventional bilingual Japanese-Chinese dictionary and Wikipedia for translating query terms. They investigate the effects of using Wikipedia and conclude that Wikipedia can be used as a bilingual named entities dictionary. They use an iterative approach to weight-tuning and term disambiguation, which is based on the PageRank algorithm. Nguyen et al. [NOH$^+$09] report that query translations for CLIR can be implemented using only Wikipedia. An advantage of using Wikipedia is that it allows translating phrases and proper nouns. It is also scalable since it is easy to use the latest version of Wikipedia, which makes it able to handle actual terms. They map the queries to Wikipedia concepts and the corresponding translations of these concepts in the target language are used to create the final query. Their CLIR system, named as WikiTranslate, is evaluated by searching the topics in Dutch, French, and Spanish within an English data collection.

A bilingual ontology is another useful resource to translate queries. Pourmahmoud and Shamsfard [PS08] report their research on Persian-English CLIR using dictionary-based query translation. They use bilingual ontologies to annotate the documents and queries and to expand the query with related terms in pre- and post-translation query expansion and combine phrase reorganisation, pattern-based phrase translation to improve the cross-lingual information retrieval performance. Wang and Ananiadou [WA10] extend biomedical ontologies, “Chinese Medical Subject Headings” (CMeSH), and apply it to implement query expansion. Their results show that ontology-based query expansion achieves prospective improvement in biomedical Chinese-English CLIR.

[1] `http://en.wikipedia.org/wiki/Main_Page`

Yuan and Yu [YY07] present a new method for query translation which only needs a bilingual dictionary and a monolingual corpus. They use co-occurrences between pairs of terms as a statistical measure of the quality of translation. The relationships between target terms are represented as a graph. By adding all the weights of a k-complete subgraph, the best combination of terms and the probability distribution of translation are computed. Compared with other work, it is a simple method. Their experiment shows that their new approach performs well.

Some researchers attempt to apply new techniques to discover relevant queries. Query suggestion [GNN$^+$07] aims to suggest relevant queries for a given request, which helps users to specify their information needs better. It is closely related to query expansion but query suggestion suggests full queries that have been formulated by users in another language. Gao et al. [GNN$^+$07] propose query suggestion by mining relevant queries in different languages from up-to-date query logs, as it is expected that for most user queries, common formulations of these topics in the query log in the target language can be found. Cross-lingual query suggestion also plays a role in adapting the original query formulation to the common formulations of similar topics in the target language. When query suggestion is used as an alternative to query translation, this approach demonstrates higher effectiveness than traditional query translation methods using either bilingual dictionary or machine translation tools.

Lilleng and Tomassen [LT07] propose a new approach to implement query translation in cross-lingual information retrieval using feature vectors. They employ ontologies to define concepts in a particular domain (oil and gas industry domain). Their idea is to associate every concept of the ontology with a feature vector to tailor these concepts to the specific terminology used in the document collection. Synonyms, conjugations and related terms that tend to be used in connection with the concept and to provide a contextual definition of it are the elements of a feature vector. Since a feature vector includes only those terms found highly related to a concept, it can be automatically translated. A correct translation is found and verified by finding an equal semantic relation between the set of translated candidate terms and the original terms of a feature vector. Those candidate terms found to have a similar semantic relation to the original feature are selected. The result is a new translated feature vector with equally semantically related terms as the original feature vector. This feature vector-based query translation approach is able to expand a query not only to translate it. However, one problem of this method is that the characteristic of a feature vector is dependent on the quality of both the ontology and the document collection being used.

Wu and Lu [WL08] introduced a novel model called domain alignment translation to implement cross-lingual document clustering and term translation simultaneously; in the end the multi-lingual documents with similar topics can be clustered together. Their method, which uses only a bilingual dictionary, can achieve comparable performance with the machine translation approach using the Google translation tool. Although their experiments only consider words, ignoring the base phrase, the clustering in the source language and the clustering in the target language are strongly related and the clustering quality can be emphasised for future research.

2.3 Biomedical Information Retrieval

According to Trieschnigg [THdJK10], biomedical information retrieval can be defined as "the structure, analysis, organization, storage, searching, and retrieval of biomedical information". In this section, we focus on the techniques and approaches to improvements in the retrieval of biomedical information.

2.3.1 History of Biomedical IR

Biomedical information retrieval starts with the accessibility of biomedical literature. The technical focus is how to index biomedical literature effectively. The two early controlled vocabulary indices, the *Index-Catalogue* and *Index Medicus*, were created more than a century ago [GG09]. The Index-Catalogue ceased in 1950; the Index Medicus was renewed in 1960 using a "freshly revised and expanded list of standardised subject headings" [CB01] called Medical Subject Headings (MeSH) [2].

The first biomedical bibliographic retrieval systems, Medical Literature Analysis and Retrieval System (MEDLARS), was established in 1964 [Lan69]. This system provided an online inquiry service in 1971, MEDLARS ONLINE, abbreviated to MEDLINE. In the middle of the 1990s, MEDLINE [3] became accessible

[2] http://www.nlm.nih.gov/mesh/

[3] http://www.nlm.nih.gov/databases/databases_medline.html

on the Internet and became a part of PubMed [4].

Nowadays, researchers are likely to treat biomedical information retrieval as one of the first steps in knowledge acquisition [Her09], because it is able to provide other applications such as information extraction and text mining with a collection of literature with a small but condensed volume.

2.3.2 Terminological Challenges of Biomedical IR

The major challenge for information retrieval in the biomedical domain is its complex and inconsistent terminology [KN04]. In order to overcome the lexical ambiguity, several terminological resources are available: UMLS, SNOMED CT, MeSH, BioLexicon and biological databases.

The Unified Medical Language System (UMLS) is "to facilitate interoperable computer programs processing biomedical texts by integrating and distributing key terminology, classification, and coding standards" [MM98]. The main component of the UMLS is the Metathesaurus, which contains multi-lingual biomedical vocabulary and several resources that have biomedical and health-related concepts in a uniform format.

Systematized Nomenclature of Medicine-clinical Terms (SNOMED CT), which is a part of the Metathesaurus, is a multi-lingual controlled vocabulary focused on medical terminology. It is designed for providing "a consistent way of indexing, storing, retrieving and aggregating clinical data from structured, computerised clinical records".

The *Medical Subject Headings* (MeSH), which is also included in the Metathesaurus, is used for indexing, cataloguing, and searching for biomedical information and documents. MeSH consists of descriptors, also named as main headings, Supplementary Concept Records, which are lists of chemicals and drugs, and topic classifiers or subheadings, which are optionally used with main headings. Biological bases collect and link the acquired knowledge, and some of them can be used as terminological resources.

BioLexicon [VMS$^+$09] is a large-scale lexical resource for the biomedical domain, providing information about predicate-argument structure that has been bootstrapped from a biomedical corpus on the subject of E. Coli.

UniProt [5], which is aimed to provide the community with a comprehensive,

[4] `http://www.pubmed.gov/`

[5] `http://www.uniprot.org/`

high-quality and freely accessible resource of protein sequence and functional information, consists of entries from Swiss-Prot and TrEMBL. This resource can be applied to looking up gene/protein synonyms.

HUGO Gene Nomenclature Committee (HGNC) [6] records human gene names and symbols and their aliases.

Entrez Gene [MOPT07] is concentrated on genomes that have been completely sequenced and is usually used as a source of gene nomenclature.

ACROMINE [OA06, OAT10] is a database of abbreviations of biomedical terms automatically extracted from the entire MEDLINE as of April 2009. ACROMINE identifies abbreviation definitions by assuming a word sequence cooccurring frequently with a parenthetical expression to be a potential expanded form. It contains 68,007 abbreviation candidates.

2.3.3 Techniques for Improving Biomedical IR

Biomedical information retrieval suffers from low recall and precision, because of its complex and inconsistent terminology. Biomedical terms usually have many synonyms, aliases, abbreviations, acronyms and variants. In order to overcome this issue, the obvious idea is to make use of external knowledge resources. Abdou and Savoy [AS07] show that by including MeSH terms retrieval performance can be improved greatly. Hersh et al. [HCRR07], for example, summarising the studies in TREC 2007 genomics track, have employed UMLS, the Gene Ontology, the Entrez Gene database, and MeSH terms to develop biomedical IR. However, some studies show that these external resources may not function as well as expected. For instance, Huang et al. [HSHRA07] report that automatically using UMLS or the Entrez Gene database for query expansion makes a negative contribution to retrieval performance.

Because of the special properties of biomedical texts, tokenisation, which converts a stream of characters into a stream of "tokens", requires consideration. Jiang and Zhai [JZ07] investigate the tokenisation strategies used in biomedical information retrieval. They study the effect of stemming and stop-word removal for biomedical IR and evaluate a set of tokenisation strategies, including a non-functional character-removal step, a break-point normalisation step with three possible normalisation methods and three possible sets of break points, and a

[6] http://www.hugo-international.org/

Greek alphabet normalisation step. They conclude that: i) "Non-functional characters", which are words or characters that do not have lexical meaning, such as punctuation and some symbols, "should be removed from the text using a set of heuristic rules." ii) "For different types of queries, different tokenization heuristics should be applied. For queries that contain only gene symbols, removing brackets, hyphens, slashes and underlines in the tokens and replacing Greek letters with their Latin equivalents are useful. For queries that contain only full gene names and for verbose queries that also contain English words, replacing brackets, hyphens, slashes and underlines with spaces should be used. Numerical characters should not be separated from alphabetical characters." iii) Stemming can improve the retrieval performance for verbose queries. iv) Stop-word removal "either does not improve the performance, or only slightly improves the performance". However, Trieschnigg [THdJK10] reports that stop-word removal for the original queries can significantly improve retrieval effectiveness and in the worst cases slightly hurt the performance. Carpenter [Car04] compared phrase-based searching and word-based searching, and noticed that the latter performed better than the former. Huang et al. [HHR06] and Büttcher et al. [BCC04] studied the techniques to process numbers, hyphens and parentheses in biomedical texts. Zhou and Yu [ZTS06] did not apply stemming in cases where the word looked like a gene name. Urbain et al. [UGF06] only used stemming when the word was not an acronym.

Some researchers employ relevance feedback to develop high-performance biomedical information retrieval. Lin [Lin08] proposes the application of PageRank and HITS to biomedical text retrieval. He assumes that the networks formed by MEDLINE citations can be exploited for retrieval, in the same manner as hyperlink graphs on the Web. The experiments demonstrate that PageRank scores help to improve retrieval performance. Yin et al. [YHL09] present a context-sensitive approach for re-ranking retrieved documents. They train a two-dimensional context for each topic by the top N and the last N documents in the initial retrieval ranked list. The two-dimensional context contains a lexical-level context, which is constructed based on pseudo-relevance feedback with keywords, and a conceptual context, which uses MeSH terms. The results of their experiments on TREC genomics tracks show that this method yields a better retrieval performance. Nevertheless, Smucker [Smu06] reports performance degradation when using query-biased pseudo-relevance feedback. Huang et al. [HSHRA07] attempt to improve passage retrieval performance in the biomedical domain. They address the issue by constructing different indexes. After comparing the experiments of word-based and sentence-based indexes, the word-based indexing scheme is believed to be more effective than the sentence-based index.

Others have concentrated on enhancing retrieval models by adjudging parameters or integrating additional processing. Abdou and Savoy [AS06] evaluate both the Okapi BM25 model and the InB2 probabilistic model derived from the Divergence from Randomness paradigm, and conclude that the latter model performs better than the Okapi model. Moreover, a 5-gram indexing approach is compared with word-based indexing schemes, and the performance decreases slightly when n-gram indexing is used. Recently, Trieschnigg et al. [THdJK10] used a cross-lingual IR perspective on a monolingual biomedical information retrieval, to view the mismatch between terms used in a query and terms used in relevant documents in the monolingual IR as a cross-lingual matching problem. They distinguished between a concept and word-based representation language and hypothesised that the integration of a concept-based representation in biomedical IR could benefit from methods and techniques used in CLIR. The technique establishing CLIR is translation. They experimented with three types of translation model: a comparable corpus of documents in both a text and concept-based representation; term-by-term translation models trained on a comparable corpus; and a thesaurus upon a baseline retrieval model and the improved retrieval model combining translation models. They concluded that translation based on pseudo feedback using a comparable corpus in both a word- and concept-based representation performed best, that a combination of translation models could improve retrieval effectiveness, and that MeSH terms enhanced recall while an extended version of UMLS improved precision.

2.4 Information Retrieval in the Chinese Language

Although Chinese language processing is a tough task, the techniques used in English IR are proven to be effective in Chinese IR. Query translation is the mainstream approach to CLIR related to the Chinese language. Many studies focus on resolutions of unknown words and translation ambiguity. Biomedical IR in Chinese has not been comprehensively investigated.

2.4.1 Difficulties of Information Retrieval in Chinese

Chinese is hard to process, not only because of its sophisticated glyphs, but also for the reason that it features special syntactic properties. According to Zhu [Zhu85], Chinese has two grammatical features:

(1) Chinese lacks morphological signs and morphological changes. Part of speech (POS) has no sign to indicate its grammatical category. On the other hand, there is no morphological change in words when they become sentence constituents.

(2) As long as the context allows, sentence constituents, including the important function words, can be omitted.

These two basic features lead to the following characteristics:

(3) One POS can be mapped with many sentence constituents. For example, adjectives can be predicates in Chinese. Also, verbs can be sentence subjects.

(4) Rules of construction of sentences are basically the same ones that construct phrases.

(5) A grammatical relation can imply a large volume of meanings and complex semantic relations without any morphological sign.

These features make it difficult to segment and tag words in Chinese, which are fundamental to other language processing tasks. The difficulties existing for biomedical IR in the Chinese are discussed in Section 2.4.4.

2.4.2 Monolingual IR in Chinese

Many researchers report the indexing strategies used when indexing Chinese documents. Indexing techniques using model-based signatures [Chi94], superimposed coding signatures [LWW01], variable bit-block compression signatures [CWL98], and PAT-trees [Chi97] generally affect only retrieval efficiency (i.e., speed and storage). Chen et al. [CHX^{+}97] implemented several statistical and dictionary-based word segmentation methods to study the effect on retrieval effectiveness of different segmentation methods. Their results show that bigram indexing and purely statistical word segmentation perform better than the dictionary-based maximum matching method. Nie et al. [NBR96] describe several reasons why the character-based, or unigram-based, approach is not suitable for Chinese text retrieval. Jin and Wong [JW02] propose a method to construct a statistic-based automatic dictionary for indexing. Their experiments show that the information retrieval based on the dictionary outperforms static dictionary results and performs as well as the bigram indexing approach.

Some researchers have investigated the effect of use of multiple types of terms in Chinese IR; Kwok [Kwo96], for example, uses short words with single characters as terms in the Pircs retrieval system. Others apply merging the retrieval lists from different indexed terms [LZ97, Kwo99] and the hybrid index [TLW99, CLWK01, LKFK01]. Luk and Kwok [LK02] compare the retrieval performance of various indexing strategies, i.e., character, word, short-word, bigram, and Pircs indexing, and conclude that bigram indexing appears to be the best indexing strategy and that the character indexing strategy performs worst.

Various retrieval models have been studied. Chow et al. [CLWK01] used Boolean and vector space models to retrieve documents. Huang and Robertson [HR97] applied the Okapi weighting scheme to Chinese IR. McNamee et al. [MMP00] used the BM25 Okapi weight, combined with the cosine measure of the vector space model. Chen et al. [CGJ01] and He et al. [HXC$^+$96] applied the logistic regression model to Chinese IR. Kwok [Kwo96, Kwo01] employed Pircs to implement Chinese IR. Luk and Kwok [LK02] investigated several retrieval models: the vector space model, 2-Poisson model, logistic regression model, and Pircs model. Their experimental results show that the 2-Poisson model has the shortest retrieval time and the best retrieval effectiveness.

2.4.3 CLIR and Chinese

The most popular approach to cross-lingual information retrieval and Chinese is query translation, although the accuracy of translation is limited by two factors: the presence of out-of-vocabulary (OOV) words and translation ambiguity.

The existing techniques tackle the OOV problem in several ways:

(1) The simplest way is to ignore OOV words when translating them. Some systems such as BabelFish use this policy.

(2) Where OOV words are caused by transliteration, an orthographic representation such as Pinyin or the International Phonetic Alphabet are applied.

When "read aloud" by a native speaker of the language, it sounds as it would when spoken by a speaker of the foreign language. See, for example, the work of Lin and Chen [LC02].

(3) Some researchers such as Chen et al. [CJG00] attempt to involve manual intervention.

(4) Web pages are used to search for appropriate translations. For instance, Zhang et al. [ZVZ05] propose an approach exploiting juxtaposition of English text and Chinese text on the web to identify OOV terms. Lu et al. [LCL02] use the web pages written in different languages that have hyperlinks pointing to the same page to resolve OOV word problems. Kwok et al. [KCDD04] propose a web-based translation from English to Chinese, focusing on entity names and terminology.

(5) Parallel corpora are important sources of translations. Yang and Li [YL02] successfully mined parallel Chinese-English documents from the Web to find the appropriate translations for OOV words. Chen and Nie [CJG00] applied aligned English-Chinese documents from the Web to overcome the OOV problem.

The following approaches have been used to resolve translation ambiguity in Chinese CLIR:

(1) Gao et al. [GZN$^+$02] applied an improved co-occurrence approach to disambiguate dictionary-based translation.

(2) Zhang et al. [ZVZ05] used a hidden Markov model (HMM) with distance factor and window size to provide disambiguation.

(3) Zhang et al. [ZSDS00] used a mutual information value matrix to select English translation, instead of looking up a Chinese-English dictionary.

Most studies use the *Linguistic Data Consortium* [7] and the *CEDICT Chinese-English dictionary* [8] to translate Chinese queries into English.

The approach to query translation using machine translation (MT) software has also been evaluated. Xu and Weischedel [XC00] explored the relationship

[7] http://www.ldc.upenn.edu/

[8] http://www.mdbg.net/chindict/chindict.php?page=cedict

between the performance of CLIR and the size of the bilingual dictionary. They observed that the performance was not improved once the lexicon exceeded 20,000 terms. Zhu and Wang [ZW06] investigated the effect of translation quality in MT-based CLIR and concluded that "[...] it is more effective to develop a larger dictionary than to develop more rules".

Xu and Weischedel [XC00] viewed information retrieval as a query generation process and employed an HMM to extend the query generation for CLIR.

2.4.4 Biomedical IR in Chinese

Qin and Feng [QF99] applied CMeSH terms to improve the indexing quality of Chinese abstracts from 1977 concerning family planning and gynaecology. Li et al. [LHZ+01] developed an information retrieval system with the help of Chinese Medical Subject Headings terms. The following points explain the reason why biomedical information retrieval in Chinese is rarely reported:

(1) There are only two Chinese search services in the biomedical domain. The "China National Knowledge Infrastructure (CNKI) [9]", which is one of the biggest databases of Chinese journals and academic publications, is only accessible to subscribers. Another bibliographic database available is the China TCM patent database (CTCMPD), but its performance is unreliable.

(2) On the other hand, the international databases include only very limited Chinese medical bibliographic data. According to Fan [FTP+08], for example, only 10 of the traditional Chinese medical journals, out of 149, are indexed by MEDLINE.

(3) Chinese document collections and gold standards in biomedicine are unavailable. Usually the gold standard, which records the relevant documents in a document collection for each topic, is provided within the document collection. For Chinese, however, there is no biomedical document collection designed for information retrieval.

(4) Essential linguistic resources in Chinese, such as parallel or comparable corpora, domain bilingual dictionaries, and ontologies are required. Despite several Chinese/English dictionaries available for the biomedical domain,

[9] http://www.global.cnki.net/

they are inadequate for information retrieval. The parallel or comparable corpora and other linguistic resources like ontologies can play a more important role to improve retrieval performance.

(5) Fewer Chinese syntactic parsers have been designed for the biomedical domain. Although some named entity extraction tools specially designed for the Chinese language in this domain have been reported in recent years, the utilities to identify the constituents of Chinese sentences from biomedical literature may provide more useful elements; in this case, the structure of sentences and the meanings of words and phrases may forge a new method to improve CLIR.

2.5 Summary

Information retrieval can be treated as the process where query terms are matched with index terms. Various models are applied to model and compute such matches. Among them, Okapi BM25 and the query likelihood language model perform best. Query expansion is the most commonly used approach to retrieval improvement.

Cross-lingual information retrieval, in which the language of queries is different from that of the documents, makes use of the techniques of monolingual information retrieval. The performance of CLIR falls behind that of monolingual IR, from 60% to 80%. The mainstream method of CLIR is query translation.

Biomedical information retrieval is an application of IR in a domain. One of the key techniques in biomedical IR is to use domain knowledge to overcome the complication of biomedical terms.

In this study, we investigate the retrieval performance of Chinese-English CLIR in biomedicine. In order to overcome the problems caused by the complex biomedical terms, Chinese-English bilingual ontologies are used to expand queries before translating them. The results of experiments show that the non-expert words plus domain terms supplied by these ontologies improve the retrieval precision.

In the next chapter, the detailed construction of such ontologies is described.

Chapter 3

Extension of CMeSH

In Chapter 2, we reviewed biomedical information retrieval. The main challenge in biomedical IR is the complex and inconsistent terminology [KN04]. The current approach to overcoming this problem is to apply domain knowledge resources. For example, Abdou and Savoy [AS07] show that by including MeSH terms, IR retrieval performance can be improved from 2.4% to 13.5%. Hersh et al. [HCRR07], summarising the results of TREC 2007 Genomics Track, show that employing UMLS, the Gene Ontology, Entrez Gene database and MeSH terms to develop biomedical IR improves retrieval performance. However, some studies show that these external resources may not function as well as expected. Huang et al. [HSHRA07] report that using UMLS or the Entrez Gene database for automatic query expansion decreases retrieval performance. In this study, instead of directly using MeSH heading terms for query expansion, we first extend the original Chinese MeSH Tree to the eCMeSH Tree and then use the terms of the eCMeSH Tree to expand the query. This chapter focuses on the extension of the original Chinese Medical Subject Headings (CMeSH), reviews related studies, discusses the limitations of MeSH-like resources, describes the details of the extension algorithm and provides an evaluation of the eCMeSH Tree.

3.1 Related Work on Ontology Extension

The original MeSH includes an ontology, the MeSH Tree. Our study aims to extend the MeSH Tree with more features. Previous research on ontology extension has focused on two aspects: extending ontologies to multiple languages and extending general domain ontologies to special domains.

3.1.1 Extending Ontologies to Multiple Languages

"EuroWordNet" [Vos98] and "MultiWordNet" [PBG02] are lexical databases for multiple languages using the structures and methodologies of "WordNet" [Mil95].

WordNet, or the Princeton WordNet, is a lexical database for the English language. It groups English nouns, verbs, adjectives, and adverbs into various sets of synonyms, known as *synsets*, in terms of word meanings, and records the semantic relations such as hypernymy/hyponymy, meronymy/holonymy, entailment, and troponymy between these sets. It also includes short descriptions for concepts. WordNet has the following characteristics: i) to resemble a traditional dictionary; ii) to perform as a thesaurus, which indicates the given concepts mapped to appropriate words; iii) to support automatic text analysis and artificial intelligence with semantic relations. WordNet can be considered as an ontology. The verbs and nouns are hierarchically organised via the hypernymy/hyponymy relations.

EuroWordNet is a multilingual lexical database for several European languages (Dutch, Italian, Spanish, German, French, Czech and Estonian) sharing the same idea as WordNet. Like the original WordNet, EuroWordNet is structured using synsets, which are connected by the semantic relations proposed by the Princeton WordNet. EuroWordNet makes use of a common framework to build the individual word nets and integrate them in a single database. The inter-lingual-index (ILI), which is mainly taken from the Princeton WordNet, connects the synsets that are equivalent in the different languages.

Vossen [Vos02, Vos96] reviews the methodology used for the development of EuroWordNet: the EuroWordNet database was built from existing available machine readable dictionaries and lexical databases, such as the van Dale database with the bilingual Dutch-English dictionary, Dutch wordnet, Princeton WordNet 1.5, Italian wordnet, English wordnet additions, and Spanish wordnet, with semantic information developed in various projects. After the specification of a fragment of the vocabulary, two possible approaches were applied to encode the semantic relations: the "merge model" and "expand model". In the merge model, the synsets and their language-internal relations are firstly developed separately for each individual language and then the equivalent relations to WordNet are generated, resulting in a word net that is independent of WordNet. In the expand model, the WordNet synsets are translated into equivalent synsets in the other languages using bilingual dictionaries and relations, which are adapted to EuroWordNet, leading to a word net that is very close to WordNet. Since the expand model makes the multilingual system biased by the Princeton WordNet: It will not only contain all the mistakes and gaps that are presented in WordNet but it will also be structured by the (American)-English lexicalisation of Western concepts, the EuroWordNet follows the merge model. Vossen [Vos96] points out that the most serious potential drawback of the expand model is the excessive dependency on the lexical and conceptual structure of the languages involved.

Vossen et al. [Vos98] describe the issues they encountered when constructing EuroWordNet:

(1) Determining the appropriate sense distinction: distinguishing between "over-differentiation of senses", where several definitions refer to the same meaning, and "under-differentiation of senses" problem, in which different senses are collapsed in a single definition. The strategy to resolve the former problem is: when multiple sources classify the same concept differently it may be possible to merge multiple senses. A solution to the latter issue is to split the sense into separate senses.

(2) Deriving comprehensive and consistent patterns of relations for word meanings: the general way of overcoming the problem of completeness is to combine information from different resources. That is, to treat the definitions in different monolingual dictionaries as a corpus and to collect those definitions that have relevant co-occurrences of words. It is also possible to apply specific strategies for extracting more comprehensive lists of word meanings related in a specific way. Another possibility is to look for words that have the same translations and/or occur as translations for the same words in bilingual dictionaries.

(3) Overlapping relations: Vossen et al. found that some "roles or involvements of first-order-entities (concrete things) indicating arguments 'incorporated', or word meanings strongly implied, within the meaning of high-order entities (events)" [Vos98, p.169] are undifferentiated. The solution is to use the under-specified relation ROLE to broaden the interpretations of these indiscriminate relations. They also noticed that some incompatible relations overlap in interpretation. "This is the case for two classes of relations: hyponymy/synonymy versus meronymy/subevent, and agent/instrument roles versus CAUSES." [Vos98, p.171] Their methodology for overcoming this problem is to apply additional existing relations to them according to

actual situations. For example, because a couple of reasons can cause the differences in hyperonyms or hyponyms, to "indicate a less precise matching these synsets should always be linked with an EQ_NEAR_SYNONYM relation" [Vos98, p.182]. The "EQ_NEAR_SYNONYM" (equivalent near synonym), which is used "when a meaning matches multiple ILI-records simultaneously, or when multiple synsets match with the same ILI-record" [Vos99, p.5], is one of the most important *complex equivalence relations* in EuroWordNet.

(4) Specification of equivalence relations: the inter-lingual-indexes (ILIs) are linked to WordNet using automatic techniques and manual approaches. The criteria of automatic techniques include: i) "Monosemous translations of synsets with a single sense are directly taken over as translations." [Vos98, p.174] ii) "Polysemous translations are disambiguated by measuring the conceptual distance in WordNet between the senses of multiple translations [AR96]." [Vos98, p.174] Manual approaches to the construction of equivalence relations are used to resolve problems such as lexical gaps, differences in sense-differentiation, and fuzzy matching.

MultiWordNet (MWN) [PBG02] produces an Italian WordNet strictly aligned with the Princeton WordNet. It adopts a methodological framework distinct from EuroWordNet. The information contained in MWN can be browsed through the MWN browser, which allows for the access to the Spanish, Portuguese, Hebrew, Romanian and Latin WordNets. The model adopted within MultiWordNet consists of "building language specific word nets keeping as much as possible of the semantic relations available in the Princeton WordNet" [PBG02, p.293]. In order to avoid the risk that two word nets built independently for two different languages show "differences which depend only partially on divergences between the languages" [PBG02, p.293], the MultiWordNet model applies strict adherence to the Princeton WordNet criteria and subjective choices to minimise this problem. Another difference between MultiWordNet and EuroWordNet is that the former introduces automatic procedures, such as the assign procedure and the lexical gaps procedure, to speed up both the construction of synsets and the detection of divergence between the Princeton WordNet and the word net being built.

3.1.2 Extending Ontologies to the Biomedical Domain

Another research direction of ontology extension is to construct domain ontologies based on general ontologies. “MedicalWordNet” [FHS05] and “BioWordNet” [PBH08] are such examples.

Bodenreider et al. [BBM03] observed that WordNet contains many common terms for single gene diseases and high-level terms from the Gene Ontology. So they concluded that WordNet is likely to be a useful source of lay knowledge in the framework of a consumer health information system on genetic diseases. However, the direct usability of the original WordNet for biomedical NLP is severely hampered by the lack of coverage of the life sciences domain in the general English WordNet [BB01]. From the perspective of the biomedical domain, WordNet has several drawbacks:

(1) WordNet is not constructed by domain experts, thus entries with technical meanings are not always reliable. For instance, some medical terms are obsolete, such as “unction” and “ichor”. Moreover, the medical entries in WordNet “tend to be shallow, lacking intermediate nodes expressing meanings intelligible to, and salient for, medical experts.” [FHS05, p.323]

(2) Expert and non-expert terms share the same synset. For example, in the phrase “upper jaw, maxilla” and “hay fever, pollinosis”, the first part is commonly used by a layperson; and the second is a domain term. This structure causes a problem when some people use a term in its “technical, medical” sense and others apply the same term under the mistaken assumption that the same disorder or symptom is being referred to.

(3) Potentially important medical information may not be provided in WordNet. The relations among WordNet’s entities are represented as “being necessarily true and there is no room for probability, optionality, or conditionality.” [FHS05, p.323] For example, “blister” is given as a kind of body part; to a WordNet user, this implies that every body has blisters. In fact, “blister” is associated with an injury. This fact is not represented in WordNet.

Bodenreider and Burgun conducted a series of studies to overcome these drawbacks. They [BB02] investigated the difference in definitions of anatomical concepts in a specialised medical dictionary (Dorland's) and a layman's terminological resource (WordNet). They found that there are plenty of genus-differentia definitions in both general and specialised resources and that hierarchical relations are the principal type of relation found between the definiendum (the word that is given a definition) and the noun phrase head of the definiens (the word or phrase used to define other words). Burgun and Bodenreider [BB01] also report their approach to the problem of terminological overlap between WordNet and the domain vocabularies of the Metathesaurus of the UMLS. By using two semantic classes: "Animal", a general class, and "Health Disorder", a class in the medical domain, as an example, they found that 2% of the domain-specific concepts from the UMLS were present in WordNet, while 83% of the domain-specific concepts from WordNet were found in the UMLS. Terms from WordNet absent in UMLS are usually found to be lay terminology. Bodenreider et al. [BBM03] also evaluated the coverage of WordNet for terminology from molecular biology and genetic diseases. They found that the coverage for highly specialised terms is low, from 0% (for gene products) to 2.8% (for cellular components). Removing specialised markers such as hyphens, numbers, and capitals from these terms and using synonyms significantly increased the coverage of genetic disease terms, ranging from 27.4 to 31.4%.

Fellbaum et al. [FHS05] discussed how to create an entirely new kind of information resource for public health, MedicalWordNet. Instead of being conceived merely as a lexical extension of the original WordNet to medical terminology, it was proposed as a new type of repository, consisting of three large collections: i) words relevant to medicine, structured as in the existing Princeton WordNet; ii) medical facts, which are medically validated propositions; and iii) propositions which reflect the layman's medical beliefs. They built a database of sentences relevant to the medical domain. These sentences, which are generated from WordNet via its relations and from online medical statements broken down into elementary propositions, are organised in two sub-corpora: MedicalBeliefNet and MedicalFactNet. In their study, human intervention is introduced: MedicalBeliefNet is rated as understandable by layman participants; MedicalFactNet is rated for correctness by domain experts.

Poprat et al. [PBH08] reported on building the BioWordNet using lexical data, the data format and the infrastructure of WordNet. They encountered two types of issue existing in the data format underlying the WordNet lexicon and the software that helps building a WordNet, and concluded that the out-of-date format and structure of WordNet, which also caused WordNet software to fail or to give limited support in case of building and debugging a new WordNet-like resource, limited the extension of the original WordNet in the biomedical domain.

3.2 MeSH and CMeSH

MeSH is a biomedical ontology, designed for indexing or cataloguing biomedical articles in libraries. CMeSH is the translation of the English MeSH into Chinese. MeSH-like resources are widely used in NLP applications, but the limitations of these resources, such as the poor coverage of biomedical vocabulary, make them unsuitable for information retrieval.

3.2.1 Introduction to MeSH and CMeSH

In this section, we introduce the components and structure of MeSH and CMeSH. However, we must point out that hierarchical relations in them are not labelled. Therefore, relations describing that an item is "narrower than" or "broader than" the other, such as "kind_of" or "part_of" relations, are included in MeSH and CMeSH trees.

3.2.1.1 MeSH

The Medical Subject Headings (MeSH) thesaurus is a controlled vocabulary produced by the National Library of Medicine (NLM) and used for indexing, cataloguing, and searching for biomedical and health-related information and documents. The first edition of MeSH was published in 1960, as a revision of the 1954 Subject Heading Authority List; MeSH reduced the number of subheadings and re-structured the subheadings.

1. Record Types

There are five types of MeSH record: Descriptors, Publication Characteristics, Geographics, Qualifiers, and Supplementary Concept Records. All of them are searchable in PubMed. Among them, Descriptors, Qualifiers, and Supplementary Concept Records are basic types. The following list describes these five record types.

Descriptors Descriptors, also known as *main headings*, are used to index citations in the MEDLINE database, for cataloguing of publications, and in other databases. Most descriptors indicate the subject of an indexed item, such as Dementia or Carcinoma in Situ.

Publication Characteristics Publication characteristics are special descriptors, also named *publication types*. Unlike MeSH descriptors, publication characteristics describe the genre of the indexed item, rather than its content, such as historical article, and the publication components such as charts, the publication formats such as editorial, and study characteristics such as clinical trial.

Geographics Geographics are descriptors which include continents, regions, countries, states, and other geographic subdivisions. They are not used to characterise subject content.

Qualifiers Qualifiers, also called *subheadings*, are used for indexing and cataloguing in conjunction with descriptors. Qualifiers provide additional means of the aspect of a subject the citation which are concerned with. For example, *Liver/drug effects* indicates that the article or book is about the effect of drugs on the liver.

Supplementary Concept Records Supplementary concept records (SCR), formerly named of *supplementary chemical records*, are used to index chemicals, drugs, and other concepts for MEDLINE. Unlike descriptors, they have no tree numbers; however, each SCR is linked to one or more descriptors.

2. Entry Vocabulary

The entry vocabulary of MeSH can be grouped into two types: *entry terms* and *other cross-references*.

Entry terms are synonyms, alternate forms, and other closely related terms in a given MeSH record. Entry terms are generally used as aliases of a heading term, thus increasing the access points to MeSH-indexed data. Entry terms are not always strictly synonymous with the preferred term in the record or with each other.

Other cross-references suggest other descriptors in MeSH that relate to the subject and that may be useful in indexing, cataloguing, or searching a particular topic. They include: i) “See related references”, also known as “associative relationships”, used for a variety of relationships between descriptors, in which one descriptor is associated with another descriptor which may be more appropriate for a particular purpose. ii) “Consider also” refers to the descriptors which have related linguistic roots. This reference defines groups of descriptors beginning with a common stem rather than to a single descriptor. iii) “Entry combination” is certain descriptor/qualifier combinations, which are prohibited by a special MeSH data element.

3. Definitions related to MeSH Tree

MeSH Tree *node* and *node number* are defined in this section. Other definitions related to MeSH Tree can be found in Section 3.4.1.

Definition 1 (MeSH Tree node)**.** *A MeSH Tree node is an item which is subject to the following:*

(1) it does not include any other items;

(2) it is a component of the MeSH Tree structure.

In the MeSH Tree, each tree node corresponds to one MeSH term, but each MeSH term can occupy multiple tree nodes. Tree nodes are usually connected into networks through the relations existing in the terms.

Definition 2 (MeSH Tree node number)**.** *A MeSH Tree node number, also known as a tree number, is a unique string used to identify a tree node.*

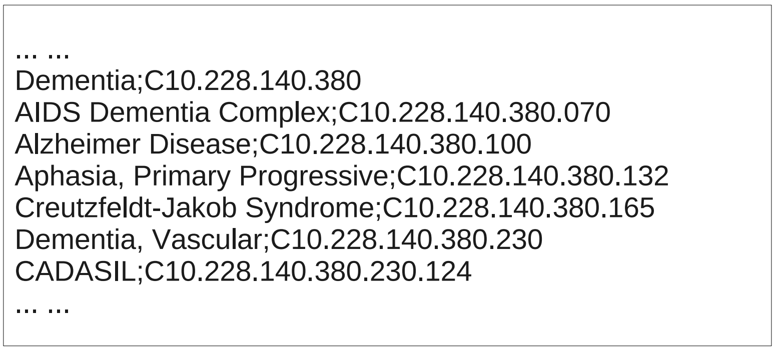
The MeSH Tree assigns each tree node a string starting with an uppercase letter followed by digits and separated by dots, which indicates the hierarchical information of the node. Here is an example, which is one of the tree node numbers that refer to the MeSH term “Alzheimer disease”.

C10.228.140.380.100

In this tree number, the root node is marked as “C10”; the parent and grandfather nodes are the nodes referred to as “C10.228.140.380” and “C10.228.140”, respectively.

4. MeSH Tree Structures

The MeSH entity vocabulary and the relationships between terms are organised as hierarchies: each term is treated as a node of a tree; and relations are represented using the tree node numbers. This tree structure is an ontology. However, hierarchical relations are not labelled. Figure 3.1 shows a segment of the MeSH tree structure from MeSH 2008.



```
... ...
Dementia;C10.228.140.380
AIDS Dementia Complex;C10.228.140.380.070
Alzheimer Disease;C10.228.140.380.100
Aphasia, Primary Progressive;C10.228.140.380.132
Creutzfeldt-Jakob Syndrome;C10.228.140.380.165
Dementia, Vascular;C10.228.140.380.230
CADASIL;C10.228.140.380.230.124
... ...
```

Figure 3.1: An example of the MeSH Tree from MeSH 2008

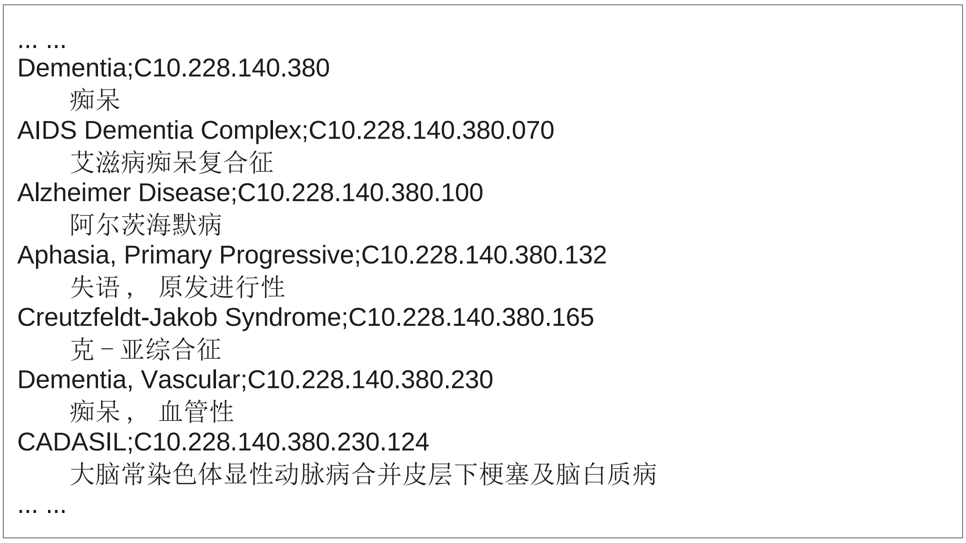
On each line, the text before the semicolon constitutes a MeSH term. After the semicolon, the string starting with a Latin letter and followed by digits and dots represents a tree number, which encodes the term’s position within the tree. The relations between terms are expressed by tree number. For example, the tree number “C10.228.140.380.100” contains the number “C10.228.140.380”. This indicates that “Alzheimer disease” is **narrower than** (here, a kind of) “Dementia”.

The MeSH tree used in this study is the 2008 MeSH tree, which has 24,763 unique terms and 48,442 tree nodes. The MeSH Tree 2008 contains 16 top categories, ranging from anatomy, organisms, diseases, chemicals and drugs to humanities, named groups, publication characteristics and geographics.

3.2.1.2 CMeSH

The Chinese Medical Subject Headings (CMeSH) is published by the Institute of Medical Information of the Chinese Academy of Medical Sciences, consisting of two different versions, i.e., a paper version and an electronic version. The basic idea of the design of CMeSH is that the terms and the hierarchical structures of MeSH should be maintained and that the traditional Chinese medical terms should be organised as MeSH does. Thus, the official CMeSH contains three parts: i) a Chinese translation of MeSH terms; ii) traditional Chinese medical subject headings, which are the special heading terms designed for the traditional Chinese medical documents; and iii) Special Classification for Medicine of China Library Classification, which is applied to classify the drugs used in traditional Chinese medicine.

The usual application of CMeSH is indexing and cataloguing biomedical literature in a library, or providing standard keywords to describe journal articles and conference papers. In this study, we use the CMeSH Tree term to expand queries. Figure 3.2 illustrates a segment of the CMeSH Tree, which is the Chinese counterpart of Figure 3.1.



```
... ...
Dementia;C10.228.140.380
    痴呆
AIDS Dementia Complex;C10.228.140.380.070
    艾滋病痴呆复合征
Alzheimer Disease;C10.228.140.380.100
    阿尔茨海默病
Aphasia, Primary Progressive;C10.228.140.380.132
    失语， 原发进行性
Creutzfeldt-Jakob Syndrome;C10.228.140.380.165
    克－亚综合征
Dementia, Vascular;C10.228.140.380.230
    痴呆， 血管性
CADASIL;C10.228.140.380.230.124
    大脑常染色体显性动脉病合并皮层下梗塞及脑白质病
... ...
```

Figure 3.2: An example of the CMeSH Tree

3.2.2 IR Using MeSH or CMeSH

Although the MeSH thesaurus is widely accepted as the standard vocabulary used for indexing, cataloguing, and searching for biomedical and health-related information and documents in libraries, researchers have used MeSH terms in IR tasks.

Lowe and Barnett [LB94] report how they use MeSH to index medical literature. They review the structure and use of MeSH, directed toward the non-expert, and outline how MeSH may help resolve a number of common difficulties encountered when searching MEDLINE. Cooper and Miller [CM98] compare lexical and statistical methods used to extract a list of suggested MeSH terms from the narrative part of the electronic patient medical records.

More recently, [GHG04, AS07, LKW09] have employed MeSH terms to evaluate or improve biomedical information retrieval applications. Guo et al. [GHG04] assume that the performance of biomedical retrieval can be improved using query expansion with synonyms of the original query terms. They use the UMLS Metathesaurus, which includes MeSH vocabulary, to identify query terms in topics and to determine their synonyms. In their research, MeSH terms are used to match against the MeSH fields of MEDLINE citations. Abdou and Savoy [AS07] evaluate ten different IR models, including recent developments in both probabilistic and language models and conclude that a probabilistic model developed within the Divergence Randomness framework leads to the best retrieval performance. They also report their impact evaluations on the retrieval effectiveness of manually assigned MeSH descriptors. The results of experiments show that by including these terms, retrieval performance can improve from 2.4% to 13.5%, depending on the underlying IR model. Lu et al. [LKW09] investigated the effectiveness of using MeSH in PubMed through its automatic query expansion process: automatic term mapping (ATM). They ran Boolean searches based on a collection of 55 topics and about 160,000 citations used in the 2006 and 2007 TREC Genomics Track. After automatic construction of a query by selecting keywords from a topic, they assigned different search tags to query terms. Three search tags: MeSH Terms, Text Words, and All Fields were chosen to be studied because they all make use of the MeSH field of indexed MEDLINE citations. Furthermore, they characterise the two different mechanisms by which the MeSH field is used. Their experiments suggested that “query expansion using MeSH in PubMed can generally improve retrieval performance, but the improvement may not affect end PubMed users in realistic situations" [LKW09, p.69].

In addition, MeSH terms are treated as the standard vocabulary to which terms from other resources are mapped [ECL$^+$88, Shu06]. Elkin et al. [ECL$^+$88] developed and evaluated a tool for identifying MeSH terms found in narrative texts. Their study exploits data structures (including both MeSH and entry term vocabulary) of MicroMeSH [LB87]. Experimental results showed that 90% of medical concepts identified in narrative texts can be mapped to MeSH terms. Shultz [Shu06] proposed a study to evaluate how various MEDLINE MeSH interfaces, including the PubMed MeSH database, the PubMed Automatic Term Mapping feature, the NLM Gateway Term Finder, and Ovid MEDLINE, map acronyms and initialisms to the MeSH vocabulary. Experimental results suggest that online interfaces do not always map medical acronyms and initialisms to their corresponding MeSH phrases, which may lead to inaccurate results and missed information if acronyms and initialisms are used in search strategies.

MeSH vocabulary has also been employed in the construction of Chinese medical ontologies. Zhou et al. [ZLWF07] discover novel gene networks and functional knowledge of genes using a significant bibliographic database of traditional Chinese medicine. In their research, MeSH disease headings are applied to generate the index data for gene and disease MEDLINE literature.

CMeSH is a Chinese extension of MeSH, which retains the terms and concepts of the English MeSH and their relations. Unlike the wide use of MeSH in Western language processing, only a small number of studies have so far attempted to use CMeSH to improve the performance of natural language processing (NLP) applications, such as information retrieval or information extraction systems. Qin and Feng [QF99] applied CMeSH terms to improve the indexing quality of Chinese abstracts from 1977 concerning family planning and gynaecology. Li et al. [LHZ$^+$01] developed an information retrieval system with the help of CMeSH terms.

3.2.3 Limitations of MeSH-like Resources

In contrast to research achievements using the original MeSH, the use of CMeSH is currently largely limited as a gold standard for indexing and cataloguing biomedical documents or for assigning indexing terms in IR systems. There is very little work that reports on evaluating CLIR using CMeSH. The main reason lies in the philosophy of MeSH design. As MeSH terms are intended to index, catalogue and search for biomedical literature in libraries, they must be represented succinctly, concisely, and accurately. Specifically, MeSH-like resources suffer from the following limitations:

(1) There are no term weights for MeSH terms. Term weights are essential to text mining or NLP algorithms based on vector-space, probabilistic, and statistical models. Without term weights, MeSH can thus function only as a traditional word list. Our previous study [WA10] has shown the high degree to which term weights contribute towards the improvement of retrieval performance.

As the Chinese translation of the original MeSH, CMeSH not only inherits this limitation, but also has an additional constraint:

(2) Each English MeSH heading term has one and only one Chinese translation. Many Chinese translations are ignored in the original CMeSH. It seems that CMeSH merely includes the translations of a MeSH heading term and its "scope note", which consists of several short sentences. The "entry term" in the MeSH structure is not translated at all. Like other languages, the Chinese language can express a particular concept in multiple ways. For example, "Alzheimer disease" is translated as "阿尔茨海默病" in the original CMeSH. However, it can also be written as "Alzheimer病" (meaning "Alzheimer disease"), "阿滋海默症" (meaning "Alzheimer disease"), "早老性痴呆" (meaning "dementia praesenilis"), "AD症" (meaning "AD disease"), "老年性痴呆" (meaning "senile dementia"), or "Alzheimer氏病" (meaning "Alzheimer disease"). The original CMeSH thus lacks the ability to provide synonyms for a particular term. Our results have shown that the availability of such synonyms can also increase task performance.

3.3 The Extension Algorithm

In Section 3.2, we presented the detail of MeSH and CMeSH, reviewed the related work on information retrieval using MeSH resources, and discussed the limitations of these resources. In order to counteract these limitations, we now focus on our approach to extend the CMeSH Tree to the eCMeSH Tree.

3.3.1 An Overview

Figure 3.3 illustrates the workflow of extending the CMeSH Tree to the eCMeSH Tree. Figure 3.3(a) describes the procedure of extending the original CMeSH Tree using extracted Chinese translations. Figure 3.3(b) shows the calculation of weights of Chinese terms. In both figures, the dashed grey squares represent the external modules.

Firstly, the English MeSH Tree terms, which come from MeSH Tree 2008, are aligned with the corresponding Chinese MeSH terms, since the Chinese MeSH terms are collected from an online keyword list [1], which contains 30,175 unique English-Chinese term pairs. Consequently, CMeSH may contain terms that do not appear in the original MeSH, or vice versa. Table 3.1 compares the number of terms in both versions of MeSH. The terms that do not appear in both versions of the tree are ignored in subsequent processing steps. The result of the alignment is the term list which is treated as the basis for extension. Each Chinese term in the list is considered as a seed term, which is used to search for Chinese synonyms online.

Secondly, the Google search engine is used to retrieve documents in Chinese for each seed term, which are assumed to contain candidate Chinese synonyms.

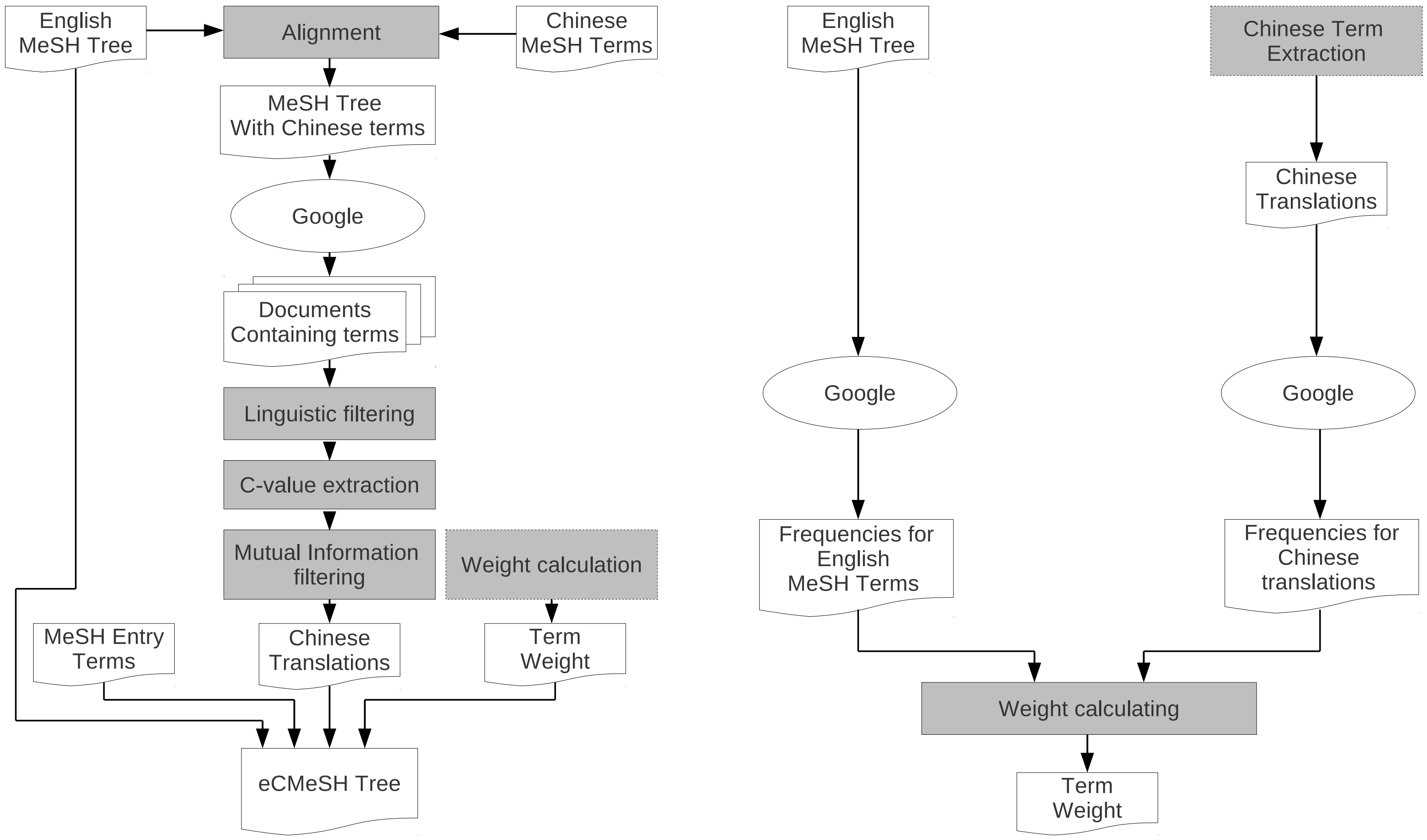
Thirdly, Chinese translations of terms are extracted from the retrieved documents using sequential application of the following: a) linguistic rules, which provide the text segments potentially containing translations; b) C-value [FAM00], which extracts candidate translations from the identified text segments; and c) mutual information filtering, which refines the candidate translations.

Fourthly, the frequencies of each English term and Chinese translation in the documents retrieved by Google are calculated; and term weights are computed using these frequencies.

Finally, the aligned term pairs, the Chinese translations, term weights, and the MeSH entry terms are merged according to the MeSH Tree hierarchy, forming the eCMeSH Tree.

Algorithm 1 gives the algorithmic description of this procedure.

[1] `http://www2.chkd.cnki.net/kns50/Dict/dict_list.aspx?firstLetter=A` (accessed on 15/07/2011)



(a) Extension of Chinese MeSH terms

(b) Calculation of term weights

Figure 3.3: The workflow of the extension algorithm

Algorithm 1: CMeSH Tree extension algorithm

```
Input  : the English MeSH Tree ($E$), the Chinese MeSH Terms ($M$)
Output: the eCMeSH Tree ($S$)

begin
    /* $A = \{< a_e, a_c > |a_e$ and $a_c$ are aligned English and
    Chinese MeSH heading terms.$\}$ */
    $M$ is aligned with the heading terms of $E \rightarrow A$
    for $a_i \in A$ do
        /* $a_{i,c}$ is the Chinese MeSH term in $a_i$. */
        $a_{i,c} \xrightarrow{\text{Google}}$ Returned documents ($R$)
        $R \xrightarrow{\text{Linguistic filtering}}$ Chinese Candidate term list ($T_l$)
    end
    $T_l \xrightarrow{\text{C-value extraction}}$ Chinese Candidate term list ($T_c$)
    $T_c \xrightarrow{\text{Mutual Information filtering}}$ Term list ($T$)
    for $t_i \in T$ do
        $t_i \xrightarrow{\text{Google}}$ Frequencies for Chinese terms ($f_c$)
        /* $e_{i,c}, e_{i,c} \in E$ is the English MeSH term which is
        expected to be associated with $t_i$. */
        $e_{i,c} \xrightarrow{\text{Google}}$ Frequencies for English terms ($f_e$)
        $f_c, f_e \xrightarrow{\text{Weight calculating}}$ Term weight of $t_i(w_{t_i})$
    end
    /* $W = \{w_i\}$ */
    $A$, MeSH entry terms ($P$), $T$, and Term weights ($W$) $\rightarrow$ the eCMeSH
    Tree ($S$)
end
```

	The online CMeSH list	**MeSH Tree 2008**
The number of the unique English terms	30,175	24,764
The number of the unique English terms in both MeSH resources	24,046	
Percentage	79.69%	97.10%

Table 3.1: The number of English terms in the online CMeSH term list and MeSH Tree 2008

3.3.2 Aligning the MeSH Tree with CMeSH Terms

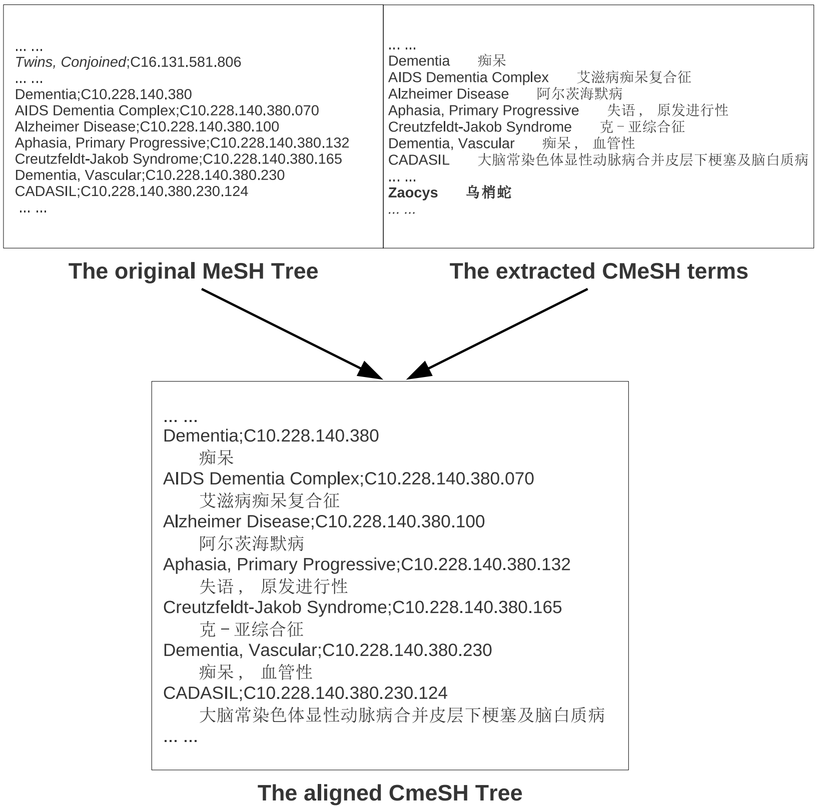
The extension of CMeSH starts with the alignment of the original English MeSH Tree with the CMeSH terms. CMeSH terms are available from an online keyword list mentioned in Section 3.3.1, which is used to index biomedical or health-related articles in the China National Knowledge Infrastructure (CNKI). The pairs of an English term and its corresponding Chinese translation are extracted from the list, illustrated in Figure 3.4. The italicised terms and the bold terms in Figure 3.4 denote the terms that cannot be found in the extracted CMeSH terms and that do not exist in the MeSH tree, respectively.

The problem with the extracted term pairs is that some terms cannot be found in the original MeSH Tree. Based on our experience, illustrated in Table 3.1, about 3% of English terms in the original MeSH Tree 2008 had no translation in the extracted terms; and approximately 20% of Chinese terms had no matching MeSH Tree terms. For example, the MeSH Tree term "Twins, Conjoined", italicised in Figure 3.4, has no counterpart in CMeSH; the term "Zaocys (乌梢蛇)", presented as a bold item in Figure 3.4, can not be found in the MeSH Tree.

Therefore, it is necessary to align the MeSH Tree with the CMeSH terms before the extension. The terms that do not occur in both the MeSH Tree and the keyword lists are ignored. The result of the alignment is the term list which is organised using the MeSH Tree structure.

3.3.3 Extracting Chinese Translations

The Chinese MeSH heading terms in the aligned list are treated as seed terms, which are then submitted to the Google search engine to obtain a set of Chinese



... ...
Twins, Conjoined;C16.131.581.806
... ...
Dementia;C10.228.140.380
AIDS Dementia Complex;C10.228.140.380.070
Alzheimer Disease;C10.228.140.380.100
Aphasia, Primary Progressive;C10.228.140.380.132
Creutzfeldt-Jakob Syndrome;C10.228.140.380.165
Dementia, Vascular;C10.228.140.380.230
CADASIL;C10.228.140.380.230.124
... ...

... ...
Dementia 痴呆
AIDS Dementia Complex 艾滋病痴呆复合征
Alzheimer Disease 阿尔茨海默病
Aphasia, Primary Progressive 失语， 原发进行性
Creutzfeldt-Jakob Syndrome 克－亚综合征
Dementia, Vascular 痴呆， 血管性
CADASIL 大脑常染色体显性动脉病合并皮层下梗塞及脑白质病
... ...
Zaocys 乌梢蛇
... ...

... ...
Dementia;C10.228.140.380
痴呆
AIDS Dementia Complex;C10.228.140.380.070
艾滋病痴呆复合征
Alzheimer Disease;C10.228.140.380.100
阿尔茨海默病
Aphasia, Primary Progressive;C10.228.140.380.132
失语， 原发进行性
Creutzfeldt-Jakob Syndrome;C10.228.140.380.165
克－亚综合征
Dementia, Vascular;C10.228.140.380.230
痴呆， 血管性
CADASIL;C10.228.140.380.230.124
大脑常染色体显性动脉病合并皮层下梗塞及脑白质病
... ...

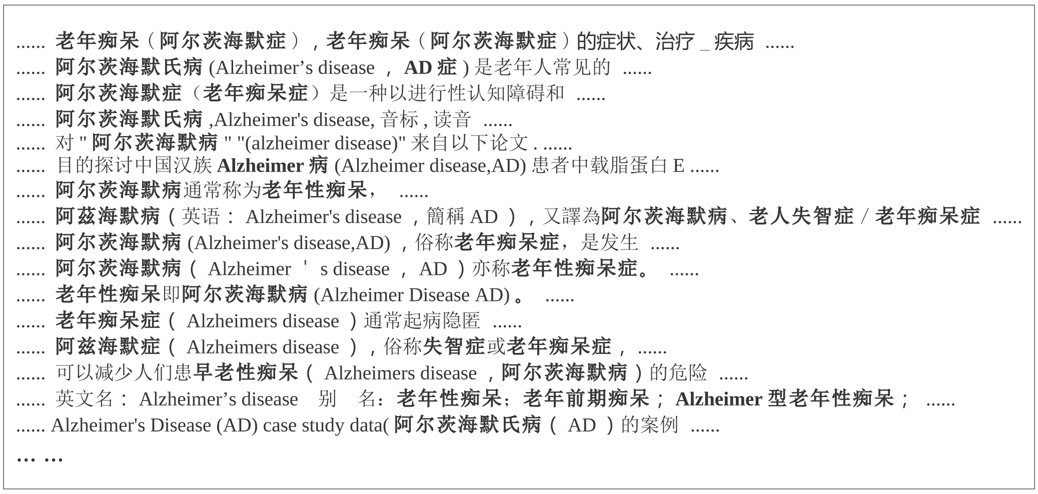
Figure 3.4: The alignment of MeSH Tree and CMeSH terms

documents which may contain synonyms of seed terms. Then the synonyms are extracted from the obtained documents using linguistic rules, C-value, and mutual information filtering.

3.3.3.1 Extracting Texts Using Linguistic Rules

The basis of this extraction is the observation that most Chinese biomedical terms tend to be accompanied by some linguistic features among the contexts. Figure 3.5 illustrates some examples about the Chinese heading term "阿尔茨海默病"("Alzheimer disease").

The biomedical terms appearing in the Chinese texts usually feature the contexts which can be used to determine both "boundaries" of the Chinese terms. In the examples of Figure 3.5, symbols like parenthesis and slash, punctuation such as period and colon, and special words like "别名" ("alias"), "亦称" ("also known as"), etc. are able to be applied to determine both boundaries of terms.



...... **老年痴呆（阿尔茨海默症），老年痴呆（阿尔茨海默症）**的症状、治疗 _ 疾病 ......

...... **阿尔茨海默氏病** (Alzheimer’s disease ， **AD 症**) 是老年人常见的 ......

...... **阿尔茨海默症**（**老年痴呆症**）是一种以进行性认知障碍和 ......

...... **阿尔茨海默氏病** ,Alzheimer's disease, 音标 , 读音 ......

...... 对 " **阿尔茨海默病** " "(alzheimer disease)" 来自以下论文 . ...

...... 目的探讨中国汉族 **Alzheimer 病** (Alzheimer disease,AD) 患者中载脂蛋白 E ......

...... **阿尔茨海默病**通常称为**老年性痴呆**， ......

...... **阿兹海默病**（英语： Alzheimer's disease ，簡稱 AD ），又譯為**阿尔茨海默病、老人失智症 / 老年痴呆症** ......

...... **阿尔茨海默病** (Alzheimer's disease,AD) ，俗称**老年痴呆症**，是发生 ......

...... **阿尔茨海默病**（ Alzheimer ' s disease ， AD ）亦称**老年性痴呆症。** ......

...... **老年性痴呆**即**阿尔茨海默病** (Alzheimer Disease AD) 。 ......

...... **老年痴呆症**（ Alzheimers disease ）通常起病隐匿 ......

...... **阿兹海默症**（ Alzheimers disease ），俗称**失智症**或**老年痴呆症**， ......

...... 可以减少人们患**早老性痴呆**（ Alzheimers disease ，**阿尔茨海默病**）的危险 ......

...... 英文名： Alzheimer’s disease 别 名：**老年性痴呆；老年前期痴呆； Alzheimer 型老年性痴呆**； ......

...... Alzheimer's Disease (AD) case study data(**阿尔茨海默氏病**（ AD ）的案例 ......

… …

Note: the bold strings are the Chinese translations of “Alzheimer disease”.

Figure 3.5: The alignment of MeSH Tree and CMeSH terms

To be specific, the Chinese biomedical terms have the following characteristics.

(1) Most Chinese biomedical terms have suffixes. For instance, “症” (meaning “disease”) in “阿滋海默症” and “病” (meaning “disease”) in “Alzheimer病” are suffixes. Other examples are “综合征” , which means “syndrome”, “复合征”, which are usually used as the end of a term referring to complex, “酶” which is the last character of an enzyme name, and “酸”, which is the indicator of an acid term.

(2) Some Chinese biomedical terms contain “inner” keywords, which can help to identify terms. For example, in “失语, 原发进行性” (meaning “Aphasia, Primary Progressive”) and “老年性痴呆症” (meaning “senile dementia”), “-性” is an important character that can indicate, when used between two adjacent verbs and nouns, that the first word describes the term after it, thus indicating a hign probability of the presence of a term. Other similar indicators are “-化-”, “-式-”, “特发”, and so forth.

(3) Biomedical terms appearing in texts are often followed by synonyms, which are often indicated using a particular set of phrases. For instance, in the sentence “. . . 阿尔茨海默病(Alzheimer disease, AD)亦称老年性痴呆症 . . . ” (meaning “. . . Alzheimer disease is also known as 老年性痴呆症 . . . ”), “亦称”, which means “also known as”, can be used to determine the beginning

of the term "老年性痴呆症". Words such as "又称" (meaning "also known as"), "俗称" (meaning "commonly known as"), "又譯為" (meaning "also translated as"), and "还叫" (meaning "also called") have a similar function.

(4) Some symbols (e.g. brackets and parentheses) and punctuation can play the role of delimiters which define the boundaries of terms. In the sentence "...可以减少人们患早老性痴呆（ Alzheimer disease, 阿尔茨海默病）的危险 ..." (meaning "...can reduce the risk of Alzheimer disease..."), the phrase between brackets contains a term. In the sentence "...阿尔茨海默病(Alzheimer disease,AD),俗称老年痴呆症,是发生 ..." (meaning "Alzheimer disease is commonly known as '俗称老年痴呆症', which occurs ..."), the second comma can decide the end of a term. However, identification only depending on such symbols may cause ambiguity. For example, the chemical term "1-(4-氟苯基)-1,3-二氢-5-异苯并呋喃腈" (citalopram) contains brackets and comma. The extracted candidates may be "4-氟苯基" ("4-fluorophenyl") or "1-(4-氟苯基)-1" ("1-(4-fluorophenyl)-1"), which are clearly incorrect and not terms.

(5) Many Chinese biomedical terms start with an English word or several Latin letters. For instance, "阿尔茨海默病" ("Alzheimer disease") can also be written as "Alzheimer症" ("Alzheimer disease") or "AD症" ("AD disease").

(6) Most Chinese biomedical terms contain a keyword to indicate their ending parts, or the ending boundaries can be determined by the words or symbols adjacent to them. The beginning of the Chinese biomedical terms cannot be simply determined by keywords, due to the lack of delimiters, which are words or linguistic characteristics such as the beginning of a sentence that are not a part of terms but lead to a term. For example, as illustrated in Figure 3.5, the term "阿尔茨海默病" ("Alzheimer disease") and its synonyms such as "老年痴呆症" ("senile dementia") and "早老性痴呆" ("dementia praesenilis") either contain a keyword or are delimited by parenthesis at the end of the term. However, the beginning of these terms are relatively hard to decide, because only a few of them can be delimited by the adjacent character before them.

According to the above-mentioned linguistic characteristics in Chinese texts, we design a set of rules, which is aimed to extract Chinese biomedical terms or the fragment of texts which may contain terms. Appendix A gives the definitions of all the rules. These 23 rules can be grouped into three layers: keyword definition rules (Level 1 and 2), rules used to determine the end of a term (Level 3) and rules applied to detect the beginning of a term (Level 4).

1. Rules defining keywords

Before extracting texts using other rules, some sets of keywords, illustrated in Appendix A.1, should be defined. These keywords define the symbols used to separate texts into sentences and words applied to determine the affixes of terms. We enumerate and define the set of rules for this purpose.

STOP defines a set of symbols which are applied to segment character streams into sentences. It includes punctuations like “。”, “？”, “?”, and “!”. However, the Western style period (“.”) and comma (“,”) are not in the set, because they are often a part of a term.

SSYM constructs a set of symbols and punctuations, which are employed to determine both boundaries of terms. Symbols like “/”, “—” “:”, “;”, “；”, “，”, and so on are listed in this set.

LSYM defines a set of symbols which are the left symbol in the pair, such as “ “”, ““”, “(”, “[”, “{”, “《”. These symbols are matched with their right counterparts. Since brackets and parentheses may cause ambiguity, they need constraining rules when applied.

RSYM This rule provides a set of symbols which are the right counterparts of symbols appearing in LSYM.

NUMB establishes a set of numbers in the Arabic, Greek, Roman, Latin, and Chinese style, e.g. “2”, “β”, “II”, “B”, and “二”.

SUFF defines a set of words which are suffixes of terms. Words such as “复合征”, “症”, “酶”, “腈” (meaning “nitrile”) and “烃” (meaning “hydrocarbon”) are examples of this set of keywords. These words are in the end part of terms (suffixes).

PREF provides a set of words which function as prefix indicators, such as “又称”, “别名”, “俗称”, etc.

INPT builds a set of special Chinese characters or words whose appearance within a phrase indicates a high probability that the phrase is a term. For example, "-性-" (meaning "-ity" or "-ness"), "-化-" (meaning "-isation" or "-ise" or "-ify"), "-式-" (meaning "type" or "style"), "-特发-" (meaning "idiopathic"), etc. are defined by this rule.

ASYM provides a set of the characters and symbols that may appear within a term, including all the Hanzi characters defined in the Unicode, numbers in NUMB, and symbols such as "-", "%", "(", and ")".

NSYM defines a set of Chinese characters and words that cannot be a part of a term, such as "这" (meaning "this"), "那些" (meaning "those"), "那里" (meaning "there"), "这里" (meaning "here"), "什么" (meaning "what"), "哪里" (meaning "where"), "但是" (meaning "only") and so on.

Appendix A.2 shows these rules. Moreover, the rule "`SUFF := ([SUFF][SUFF])`" is applied to dynamically add more elements into the set of "suff", since we observe that some keywords are composed of other keywords. For example, "蒽醌" (meaning "anthraquinone") is constituted by "蒽" (meaning "anthracene") and "醌" (meaning "quinone").

2. Rules identifying the end of terms

The end of each Chinese biomedical term is determined by this group of rules, shown in Appendix A.3, which is subdivided into two types: rules determining the end part of a term and rules identifying the end boundary (not a part) of a term.

For example, the following rule will recognise the suffix of a term and append the "EOT" (End of Term) tag at the end of the suffix. The "ˆ" indicates the beginning of a sentence; "+" means that the previous type of character(s) occur once or more than once; "1" controls the scanning direction from right to left.

```
ˆ([ASYM]+[SUFF]):1:EOT
```

The following example rule illustrates the detection of the ending boundary of a candidate term from the context. Here, keywords in the PREF set are used to identify the end of a term, but these keywords are not a part of a term.

```
ˆ([ASYM]+)[PREF]:1:EOT
```

3. Rules detecting the beginning of terms

While the previous rules are formulated for identifying the end of each candidate term, another set of rules, illustrated in Appendix A.4, is aimed at detecting the beginning of a term. This level of rules is also subdivided into two types: rules identifying the beginning of a term using its prefix indicator; or other linguistic features.

For instance, the rule illustrated as follows is used to find the beginning part of each term, only concerning the prefix indicators. It adds the additional tag "BOT" (Beginning of Term) at the end of the prefix indicator; and it scans the character stream from right to left, due to the operator "1".

```
BOT:1:[PREF]([ASYM]+)
```

The below rule aims to find the beginning of a term using information such as the beginning of a sentence or some symbols. After scanning the characters from left to right, it will add the tag "BOT" before the beginning of the term.

```
BOT:0:^|[SSYM]([ASYM]+)
```

4. Implementation of Rules

Algorithm 2 shows how we extract Chinese candidate terms for a single CMeSH term using linguistic rules. The fourth layers of rules described before are applied using the following criteria:

(1) First, we apply the rules to define keyword sets, because the rules used to determine the term's beginning and ending depend on the keyword set.

(2) Then, we apply the rules used to detect the end of terms, since most of the Chinese terms have suffixes or delimiters.

(3) The rules applied to identify the beginning of terms are employed based on the results of the third level of rules.

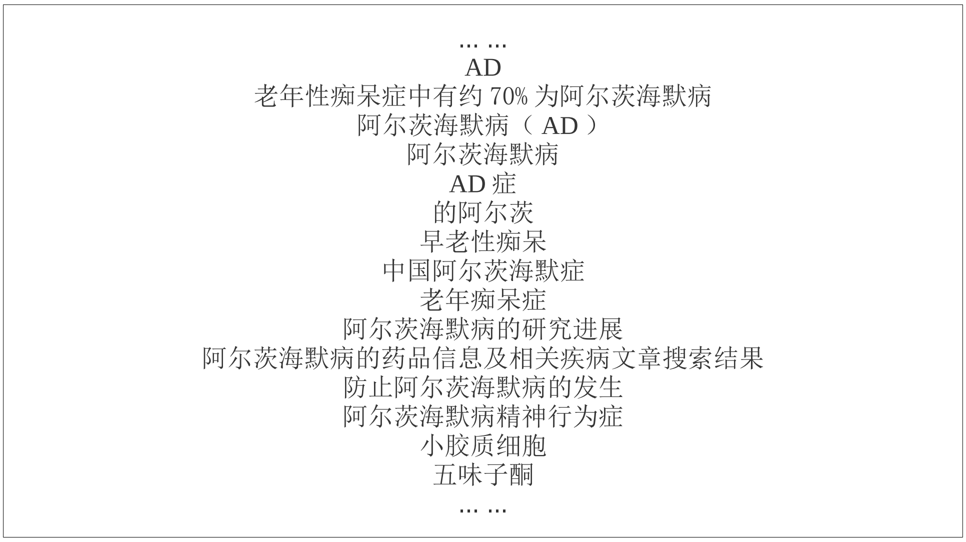
(4) The text between "BOT" and "EOT" tags is extracted as the candidate Chinese terms.

Each rule has a unique code that determines its application sequence. For rules at the same level, the implementation of a rule is determined by its code. All rules are performed for each sentence.

Algorithm 2: Extraction using linguistic rules

```
Input  : the Potentially Relevant Documents ($D$), the rules ($R$)
Output: the candidate term list ($T$)

begin
    /* $R_{1,2}$ is the first and second level rule.          */
    for $r_i \in R_1$ do
        Run $r_i$, defining keywords
    end
    /* $S = \{s\}, s$ is a sentence.                           */
    $D \xrightarrow{\text{[STOP]}} S$
    $T = \emptyset$
    for $s_i \in S$ do
        /* $R_3$ is the third level rule.                      */
        for $r_j \in R_3$ do
            $s_i \xrightarrow{r_j} T''$
        end
        /* $R_4$ is the fourth level rule.                     */
        for $r_n \in R_4$ do
            $T'' \xrightarrow{r_n} T'$
        end
        $T \leftarrow T \cup T'$
    end
end
```

Figure 3.6 gives an example of the extracted candidate terms for “阿尔茨海默病” (“Alzheimer disease”).



… …
AD
老年性痴呆症中有约 70% 为阿尔茨海默病
阿尔茨海默病（ AD ）
阿尔茨海默病
AD 症
的阿尔茨
早老性痴呆
中国阿尔茨海默症
老年痴呆症
阿尔茨海默病的研究进展
阿尔茨海默病的药品信息及相关疾病文章搜索结果
防止阿尔茨海默病的发生
阿尔茨海默病精神行为症
小胶质细胞
五味子酮
… …

Figure 3.6: The results of linguistic rule filtering

5. Issues of the Rule-Based Approach

In order to evaluate the performance of the rule-based extraction, we firstly select the results of 9 CMeSH terms at random plus the result of “阿尔茨海默病” (“Alzheimer disease”). Then, for each CMeSH term, 50% of the extracted items are randomly chosen. And finally we manually determine whether these extracted items are terms or not. Table 3.2 shows the experimental results, where the selected items are considered as the sample.

The performance of rule-based extraction has some drawbacks. According to our experiment, illustrated in Table 3.2, only 44.48% of extracted terms are correct. From our error analysis, we observe:

(1) Many terms are “nested” in other candidates, or affixed by other words. For example, in “阿尔茨海默病的研究进展” (meaning “research progress on Alzheimer disease”), the term “阿尔茨海默病” is nested; while in the example of “中国阿尔茨海默症” (meaning “China Alzheimer disease”), “中国” (China) is the prefix. We found that about 46% of extracted candidates suffer from such cascade problems.

The CMeSH term	**Extracted items**	**Selected items (sample)**	**Non-term items in sample**	**Term items in sample**	**Nested terms among sample terms**	**Irrelevant terms among sample terms**
阿尔茨海默病 (Alzheimer disease)	96	48 (50.00%)	28 (58.33%)	20 (41.67%)	12 (60.00%)	6 (30.00%)
5,7,4'-三羟基黄酮 (Apigenin)	468	234 (50.00%)	151 (64.53%)	83 (35.47%)	52 (62.65%)	29 (34.94%)
UDP葡糖4-差向异构酶 (UDPglucose 4-Epimerase)	49	24 (48.98%)	20 (83.33%)	4 (16.67%)	1 (25.00%)	0 (0.0%)
日本血吸虫 (Schistosoma japonicum)	239	119 (49.79%)	71 (59.66%)	48 (40.34%)	11 (22.92%)	21 (43.75%)
南欧斑疹热 (Boutonneuse Fever)	365	182 (49.86%)	46 (25.27%)	136 (74.73%)	49 (36.03%)	52 (38.24%)
麻风, 中间型 (Leprosy, Borderline)	726	363 (50.00%)	247 (68.04%)	116 (31.56%)	48 (41.38%)	43 (37.07%)
葡萄球菌烧灼性皮肤综合征 (Staphylococcal Scalded Skin Syndrome)	88	44 (50.00%)	17 (38.64%)	27 (61.36%)	24 (88.89%)	8 (29.63%)
腺瘤, 嗜碱性 (Adenoma, Basophil)	197	98 (49.75%)	26 (26.53%)	72 (73.47%)	37 (51.39%)	28 (38.89%)
格斯特曼综合征 (Gerstmann Syndrome)	126	63 (50.00%)	41 (65.08%)	22 (34.92%)	10 (45.45%)	8 (36.36%)
骨骼肌肌球蛋白 (Skeletal Muscle Myosins)	313	156 (49.84%)	92 (58.97%)	64 (41.03%)	29 (45.31%)	25 (39.06%)
Summary	2,667	1,331 (49.91%)	739 (55.52%)	592 (44.48%)	273 (46.12%)	220 (37.16%)

Table 3.2: The performance of the rule-based term extraction

(2) About 37% of extracted candidate terms are irrelevant items. For example, “小胶质细胞” (meaning “microglia”) and “五味子酮” (meaning “schisandra ketone”) are irrelevant to “阿尔茨海默病” (“Alzheimer disease”), although both may be terms.

(3) The others can be considered as strings without meanings, such as “的阿尔茨” (meaning “of Alz”).

Therefore, it is necessary to apply more effective methods to filter out more terms from the extracted candidate list.

3.3.3.2 Extracting Terms Using the C-value Algorithm

In order to extract Chinese biomedical terms from texts, the *C-value* term extraction technique is applied to the final results of filtering using the above-mentioned linguistic rules.

1. Introduction to C-value

C-value [FAM00] is a simple but effective tool to extract terms, especially nested terms, from free texts, using the frequencies of string occurrences to extract candidate terms. Equation 3.1 measures C-value algorithm:

$$C\text{-}value\,(a) = \begin{cases} \log_2|a| \cdot f\,(a) & a \text{ is not nested,} \\ \log_2|a| \cdot \left(f\,(a) - \dfrac{1}{P\,(T_a)} \sum_{b \in T_a} f\,(b) \right) & \text{otherwise.} \end{cases} \tag{3.1}$$

where	a	a candidate string
	b	a candidate string
	$\|a\|$	the length of the string, e.g. the number of symbols in the string a
	$f\,(\cdot)$	the frequency of occurrence of a string in the corpus
	T_a	the set of extracted candidate terms that contains the string a
	$P\,(T_a)$	the number of these candidate terms

C-value by itself is a purely statistical measure and hence is independent of language and domain knowledge; it only requires statistical information of candidate strings that occur in the corpora. According to Equation 3.1, C-value assigns a lower value to a nested term than a non-nested one, which indicates that this measure can be used to process nested names.

2. Linguistic Filtering in C-value

Generally, C-value is combined with a linguistic filter, which is often a set of simple regular expressions coupled with a stop list, used to prune the candidate terms. In our studies, the linguistic filter is not applied, because the effectiveness of linguistic rules depends on a correct and easy approach to recognise part of speech (POS); however, for Chinese texts, it is hard to determine the correct POSs. The results of linguistic rules contain plenty of texts which may contain Chinese biomedical terms.

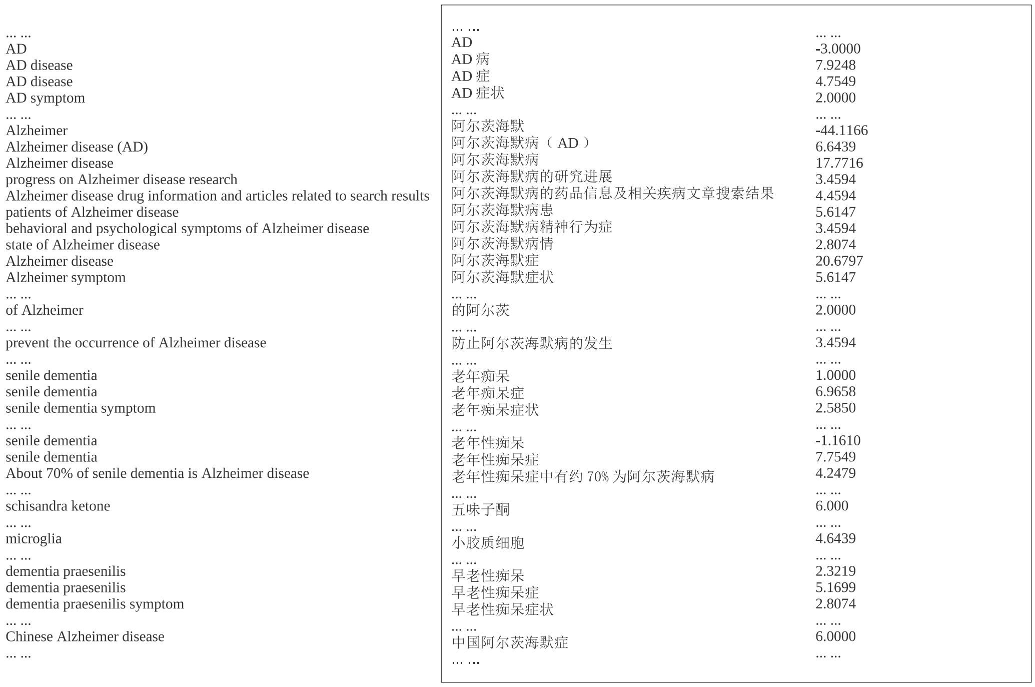
3. Threshold for C-value

The threshold used to filter out terms from the C-value results is 4.7, which means items with C-value lower than 4.7 will be ignored. This threshold is determined according to the length of biomedical terms. The average length of terms in CMeSH (the keyword list in Section 3.3.2) is 5.1 characters. We assume that a candidate is selected as a term only if it occurs twice or more in the candidate list. Suppose this candidate is not a nested term and appears twice, then according to Equation 3.1, the C-value for it is about 4.701. Figure 3.7 illustrates the results after using C-value without the threshold. The first column is English translations of the second column which is the candidate term list; the third is their c-values.

4. Issues with C-value Filtering

Although C-value filtering removes the strings which are likely not to be terms, there is a problem: some items are still nested. For example, in the term "阿尔茨海默病（AD）", either "阿尔茨海默病" or "AD" is a term. Figure 3.8 gives the filtered C-value results using the threshold 4.7.

In this example, "阿尔茨海默病（AD）", "阿尔茨海默病患", "阿尔茨海默症状", and "中国阿尔茨海默症" are the strings which contain the term "阿尔茨海默病" and "阿尔茨海默症". These items feature the prefix or suffix consisting of one or more symbols. We now use mutual information (MI) to resolve this problem.



... ...	... ...	... ...
AD	AD	-3.0000
AD disease	AD 病	7.9248
AD disease	AD 症	4.7549
AD symptom	AD 症状	2.0000
... ...	... ...	... ...
Alzheimer	阿尔茨海默	-44.1166
Alzheimer disease (AD)	阿尔茨海默病（ AD ）	6.6439
Alzheimer disease	阿尔茨海默病	17.7716
progress on Alzheimer disease research	阿尔茨海默病的研究进展	3.4594
Alzheimer disease drug information and articles related to search results	阿尔茨海默病的药品信息及相关疾病文章搜索结果	4.4594
patients of Alzheimer disease	阿尔茨海默病患	5.6147
behavioral and psychological symptoms of Alzheimer disease	阿尔茨海默病精神行为症	3.4594
state of Alzheimer disease	阿尔茨海默病情	2.8074
Alzheimer disease	阿尔茨海默症	20.6797
Alzheimer symptom	阿尔茨海默症状	5.6147
... ...	... ...	... ...
of Alzheimer	的阿尔茨	2.0000
... ...	... ...	... ...
prevent the occurrence of Alzheimer disease	防止阿尔茨海默病的发生	3.4594
... ...	... ...	... ...
senile dementia	老年痴呆	1.0000
senile dementia	老年痴呆症	6.9658
senile dementia symptom	老年痴呆症状	2.5850
... ...	... ...	... ...
senile dementia	老年性痴呆	-1.1610
senile dementia	老年性痴呆症	7.7549
About 70% of senile dementia is Alzheimer disease	老年性痴呆症中有约 70% 为阿尔茨海默病	4.2479
... ...	... ...	... ...
schisandra ketone	五味子酮	6.000
... ...	... ...	... ...
microglia	小胶质细胞	4.6439
... ...	... ...	... ...
dementia praesenilis	早老性痴呆	2.3219
dementia praesenilis	早老性痴呆症	5.1699
dementia praesenilis symptom	早老性痴呆症状	2.8074
... ...	... ...	... ...
Chinese Alzheimer disease	中国阿尔茨海默症	6.0000
... ...	... ...	... ...

Figure 3.7: The results of C-value filtering without threshold

3.3.3.3 Mutual Information

1. Definition of Mutual Information

Mutual information (MI) has been used a lot in co-occurrence analysis. Its variant, pointwise mutual information (PMI), is defined in Equation 3.2 [MS99, p.68].

$$PMI\,(x,y) = \log \frac{P_{X,Y}\,(x,y)}{P_X\,(x)\,P_Y\,(y)} \tag{3.2}$$

where x a signal

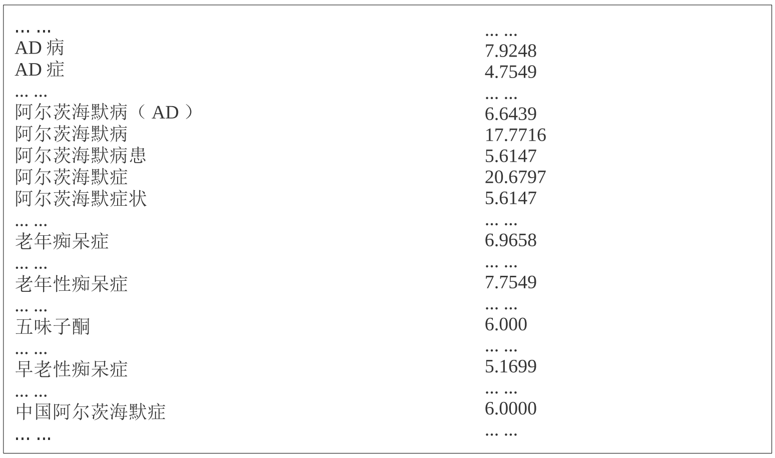
y a signal

$P_{X,Y}\,(x,y)$ the joint probability of the occurrence of the signals x and y

$P_X\,(x)$ the probability of the occurrence of the signal x

$P_Y\,(y)$ the probability of the occurrence of the signal y

Magerman and Marcus [MM90] suggested a variation of Equation 3.2, which replaces probabilities of events with the item frequency contingency table (Table 3.3). Equation 3.3 is usually used to estimate mutual information of the item x and y,



… …	… …
AD 病	7.9248
AD 症	4.7549
… …	… …
阿尔茨海默病（ AD ）	6.6439
阿尔茨海默病	17.7716
阿尔茨海默病患	5.6147
阿尔茨海默症	20.6797
阿尔茨海默症状	5.6147
… …	… …
老年痴呆症	6.9658
… …	… …
老年性痴呆症	7.7549
… …	… …
五味子酮	6.000
… …	… …
早老性痴呆症	5.1699
… …	… …
中国阿尔茨海默症	6.0000
… …	… …

Figure 3.8: The results of C-value filtering with threshold

which may refer to parts of speech, words, symbols, phrases, or even sentences.

	Y	$\overline{Y}$
X	A	B
$\overline{X}$	C	D

Table 3.3: Part of speech frequency contingency

$$MI\left(x,y\right) \approx \log_2 \frac{A \times (A+B+C+D)}{(A+C) \times (A+B)} \tag{3.3}$$

where X the event that x occurs in the corpus
Y the event that y occurs in the corpus
$\overline{X}$ the event that x does not occur in the corpus
$\overline{Y}$ the event that y does not occur in the corpus
A frequency of the co-occurrence of x and y
B frequency of x occurrence without y occurrence
C frequency of y occurrence without x occurrence
D frequency of the absence of y and x

2. Filtering Candidate Terms Using Mutual Information

The binding capacities between any two symbols in Chinese texts are different. For example, “中” (meaning “central”) and “国” (meaning “nation”) are more likely to be combined together in texts than “国” and “阿” (meaning “a big mount") are. We assume that such affinity can help determine the boundaries of the terms which are extracted using C-value.

The affinity between symbols is measured using PMI. In this study, we use an external corpus to calculate MI values. This corpus, which includes 39,654 full-text Chinese papers or articles on pharmacy and clinical and medical science, is collected from an online document publication site, "论文百事通" [2]. Table 3.4 illustrates the MI values between symbols of candidate terms listed in Figure 3.8. In this example, each candidate term is separated as a couple of overlapping bigrams; and the MI values of these bigrams are calculated using Equation 3.3.

The candidate term	**Bigrams and their MI values**
AD病 (AD disease)	AD D病 (2.3122)(0.0253)
AD症 (AD disease)	AD D症 (2.3122)(0.0019)
阿尔茨海默病（AD） (Alzheimer disease (AD))	阿尔 尔茨 茨海 海默 默病 病（ （A (0.4176)(0.0485) (0.0554) (0.1892) (0.2357)(0.3029)(0.0725) AD D） (2.3122)(0.0085)
阿尔茨海默病 (Alzheimer disease)	阿尔 尔茨 茨海 海默 默病 (0.4176)(0.0485) (0.0554) (0.1892) (0.2357)
阿尔茨海默病患 (patients of Alzheimer disease)	阿尔 尔茨 茨海 海默 默病 病患 (0.4176)(0.0485) (0.0554) (0.1892) (0.2357)(5.7850)
阿尔茨海默症 (Alzheimer disease)	阿尔 尔茨 茨海 海默 默症 (0.4176)(0.0485) (0.0554) (0.1892) (0.0884)
阿尔茨海默症状 (Alzheimer disease symptom)	阿尔 尔茨 茨海 海默 默症 症状 (0.4176)(0.0485) (0.0554) (0.1892) (0.0884)(4.7363)
老年痴呆症 (senile dementia)	老年 年痴 痴呆 呆症 (5.2546)(0.0509) (6.2053) (0.4498)
老年性痴呆症 (senile dementia)	老年 年性 性痴 痴呆 呆症 (5.2546)(0.0329) (0.0) (6.2053) (0.4498)
五味子酮 (schisandra ketone)	五味 味子 子酮 (0.9138)(0.0276)(0.01784)
早老性痴呆症 (dementia praesenilis)	早老 老性 性痴 痴呆 呆症 (0.2371)(0.0067) (0.0) (6.2053)(0.4498)
中国阿尔茨海默症 (China Alzheimer disease)	中国 国阿 阿尔 尔茨 茨海 海默 默症 (3.5922) (0.0) (0.4176) (0.0485) (0.0554)(0.1892)(0.0884)

Table 3.4: An example of MI values for candidate Chinese terms

[2] http://www.lwbst.com/

The MI values illustrated in Table 3.4 reflect this feature: The bigram with a high MI value is likely to be a common word, which indicates the possible boundaries of the entire term. For example, "AD", "中国" (meaning "China"), "老年" (meaning "senility"), "痴呆" (meaning "dementia"), "症状" (meaning "symptom"), and "病患" (meaning "illness") are bigrams with a high MI value. Some of them, such as "中国" and "症状", can be used to determine the boundaries of terms.

However, if the candidate term is separated according to "AD", "老年", and "痴呆", errors will be introduced. These high-MI items have these characteristics: i) they do not contain Chinese characters; or ii) they appear in the middle of a term.

Based on the above-mentioned features, we design an algorithm using MI to remove the first or last character of candidate terms, as described in Algorithm 3

Our experiment using the above-mentioned corpus shows the average MI (AMI) is 0.6920. Since the algorithm only concerns the first and last bigrams of each candidate term, it will only remove the first or last character if the bigram MI is higher than AMI.

In order to evaluate the term extraction performance of MI-based filtering approach, we use the 10 CMeSH terms illustrated in Table 3.2 and compare the number of terms filtered using MI with that obtained after using linguistic rules and C-value. Table 3.5 shows the experimental results, where "term" refers to the synonymous terms of the original CMeSH heading term; "terms@10" denotes terms appearing in the top 10 ranked terms.

According to our experiment, shown in Table 3.5, about 9.8% of terms can be recalled from the results of C-value. Figure 3.9 illustrates the results of mutual information filtering with the English translations.

3.3.4 Assigning Term Weight

After processing the potentially relevant documents using linguistic rules, C-value and mutual information, the Chinese biomedical terms are extracted and considered as the synonyms of the CMeSH heading terms. Then, we implemented an algorithm to assign term weights to these synonyms.

Equation 3.4, a variant of the one proposed by Layman et al. [LCC01], measures the term weight employing frequency of English MeSH heading term and its

Algorithm 3: Delete the first or last character of a term using MI

Input : the corpus (C), the candidate term list (T)
Output: the improved term list (T')

```
begin
    /* $S = \{s|s$ is a symbol in the corpus $C\}$ */
    $T' \leftarrow \emptyset$
    $C \rightarrow S$
    for $s_i \in S, s_j \in S$ do
        calculate $MI_{s_i,s_j}$
    end
    $\overline{MI} = \frac{\sum_{1\leq i\leq |S|}\sum_{1\leq j\leq |S|} MI_{s_i,s_j}}{|S|^2}$
    for $t_i \in T$ do
        /* $a_i$ is the $i$th symbol or character consisting of $t$.
        */
        $t_i \rightarrow \{a_{i,j}a_{i,j+1}\}, j \in [1, |t|)$
    end
    for $t_i \in T, i \in [1, |T|]$ do
        $b \leftarrow t_i$
        if $a_{i,1}$ is a Chinese character or $a_{i,|t|}$ is a Chinese character
        then
            if $a_{i,1}a_{i,2} > \overline{MI}$ and $a_{i,1}a_{i,2} > a_{i,2}a_{i,3}$ then
                $b \leftarrow \{a_{i,2}a_{i,3}\cdots a_{i,|t|}\}$
            end
            if $a_{i,(|t|-1)}a_{i,|t|} > \overline{MI}$ and $a_{i,(|t|-2)}a_{i,(|t|-1)} < a_{i,(|t|-1)}a_{i,|t|}$ then
                $b' \leftarrow \{a_{i,1}a_{i,2}\cdots a_{i,(|t|-1)}|$
            end
        end
        $T' \leftarrow \{b\} \cup T'$
    end
end
```

The CMeSH term	**Terms extracted by linguistic rules**	**Terms extracted by C-value (terms@10)**	**Terms extracted by MI (terms@10)**
阿尔茨海默病 (Alzheimer disease)	37	6	7 (16.67%)
5,7,4'-三羟基黄酮 (Apigenin)	163	4	4 (0.0%)
UDP葡糖4-差向异构酶 (UDPglucose 4-Epimerase)	6	3	3 (0.0%)
日本血吸虫 (Schistosoma japonicum)	81	4	5 (25.00%)
南欧斑疹热 (Boutonneuse Fever)	219	6	7 (16.67%)
麻风, 中间型 (Leprosy, Borderline)	233	8	8 (0.0%)
葡萄球菌烧灼性皮肤综合征 (Staphylococcal Scalded Skin Syndrome)	55	5	5 (0.0%)
腺瘤, 嗜碱性 (Adenoma, Basophil)	149	7	8 (14.29%)
格斯特曼综合征 (Gerstmann Syndrome)	45	5	5 (0.0%)
骨骼肌肌球蛋白 (Skeletal Muscle Myosins)	130	4	4 (0.0%)
Summary	1,118	51	56 (9.80%)

Table 3.5: The performance of term extraction improved by MI

counterpart Chinese synonyms' frequencies.

$$\begin{aligned} w_{ct} &= \begin{cases} w + 1.0 & \text{if } f_{ct} > f_{et} > 0, \\ w & \text{otherwise.} \end{cases} \\ w &= e^{-e^{-\frac{\log_{10}\left((f_{ct}+0.5) / (f_{et}+0.5) \right)}{2}}} \end{aligned} \tag{3.4}$$

where w_{ct} the Chinese term weight

f_{ct} the frequency of the Chinese term

f_{et} the frequency of the English MeSH heading term, which is the equivalent of that Chinese term

Equation 3.4 is designed to increase the term weight when a Chinese term is

… …	… …
AD disease	AD 病
AD disease	AD 症
… …	… …
Alzheimer disease (AD	阿尔茨海默病（AD
Alzheimer disease	阿尔茨海默病
Alzheimer disease	阿尔茨海默症
… …	… …
year dementia disease	年痴呆症
… …	… …
year dementia disease	年性痴呆症
… …	… …
schisandra ketone	五味子酮
… …	… …
dementia praesenilis	早老性痴呆症
… …	… …
nation Alzheimer disease	国阿尔茨海默症
… …	… …

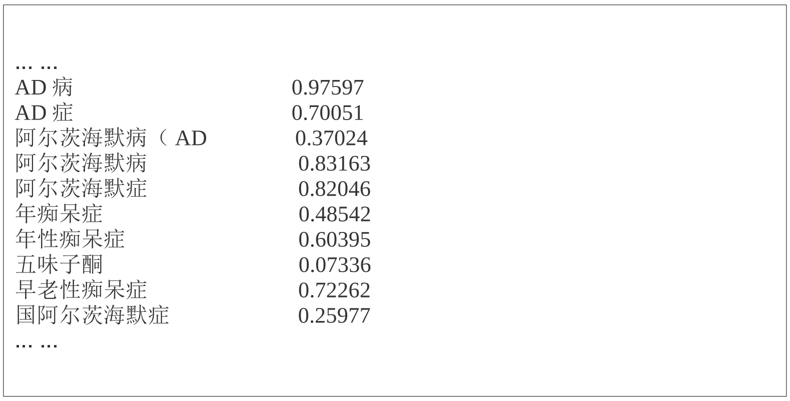
Figure 3.9: The results of mutual information filtering

more popular than its English equivalent, as indicated by the frequency of the Chinese translation as greater than that of English term. The value 0.5 used in the Equation is used to avoid the zero frequency.

Figure 3.3 illustrates the source of frequencies. The frequency of an English MeSH heading term is the number of its occurrence within the potentially relevant documents returned by the Google search engine; while the frequency of the equivalent Chinese term is the number of occurrences of this term in the documents retrieved by Google. The weight for the English term is not calculated because eCMeSH is designed to expand the Chinese query terms. Figure 3.10 is an example of term weights assigned to the terms in Figure 3.9.

3.3.5 Merging Terms and Weights

Now, we can merge the original MeSH Tree, the aligned CMeSH heading term list, the Chinese synonym list, and the weights for all Chinese terms. These resources are merged based on the infrastructure of the MeSH Tree. The result



```
... ...
AD 病                    0.97597
AD 症                    0.70051
阿尔茨海默病（AD          0.37024
阿尔茨海默病              0.83163
阿尔茨海默症              0.82046
年痴呆症                  0.48542
年性痴呆症                0.60395
五味子酮                  0.07336
早老性痴呆症              0.72262
国阿尔茨海默症            0.25977
... ...
```

Figure 3.10: An example of the Chinese weighted terms

is “the extended CMeSH Tree” (eCMeSH Tree), a bilingual biomedical ontology. Figure 3.11 is an example of the eCMeSH Tree.

In the eCMeSH Tree, the English heading terms and their Chinese equivalents are connected by tree nodes, which are represented by a string starting with the upper case Latin letters followed by dot-separated numbers. The nestedness of tree numbers indicates the taxonomy relations among the terms. All the Chinese terms including synonyms are weighted, while the English terms are not. The Chinese heading terms are extended by Chinese translations extracted from web pages which are assumed to contain Chinese synonyms. The English heading terms are extended by MeSH entry terms.

3.4 Evaluation of eCMeSH

The eCMeSH Tree constitutes an ontology, which must be evaluated. In general, ontology evaluation cannot be compared to evaluation tasks in information retrieval or classic natural language processing tasks such as part-of-speech (POS) tagging, because the notion of precision and recall cannot easily be defined. Methodologies used to evaluate ontologies generally fall under one of the following approaches:

(1) Testing the ontology in an application and evaluating the result [PM04], also called application-based evaluation;

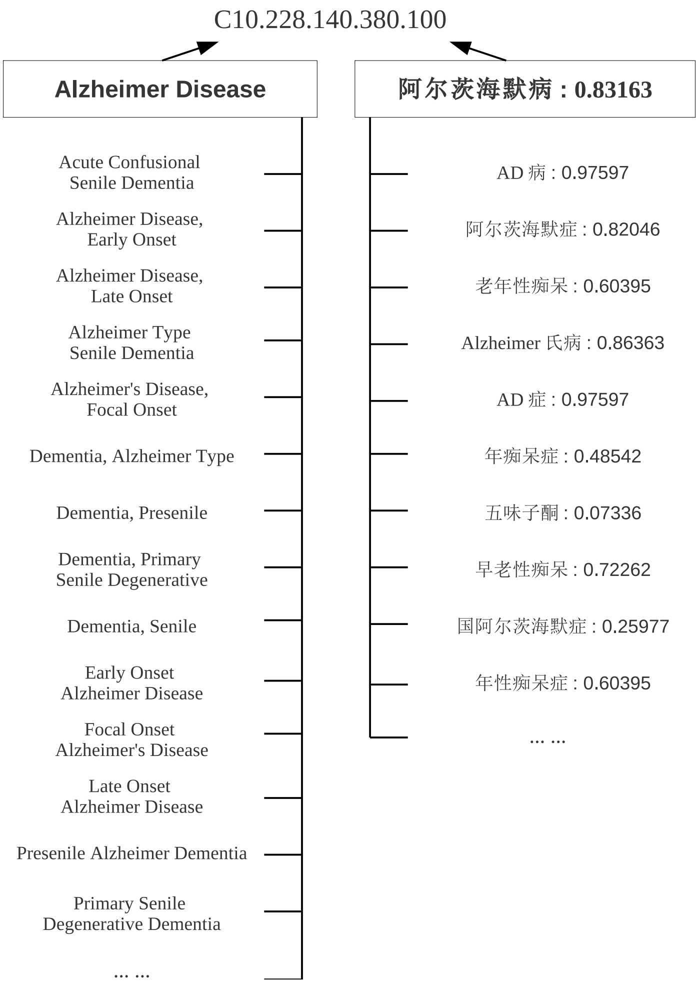

C10.228.140.380.100
Alzheimer Disease
阿尔茨海默病 : 0.83163
Acute Confusional Senile Dementia
Alzheimer Disease, Early Onset
Alzheimer Disease, Late Onset
Alzheimer Type Senile Dementia
Alzheimer's Disease, Focal Onset
Dementia, Alzheimer Type
Dementia, Presenile
Dementia, Primary Senile Degenerative
Dementia, Senile
Early Onset Alzheimer Disease
Focal Onset Alzheimer's Disease
Late Onset Alzheimer Disease
Presenile Alzheimer Dementia
Primary Senile Degenerative Dementia
... ...
AD 病 : 0.97597
阿尔茨海默症 : 0.82046
老年性痴呆 : 0.60395
Alzheimer 氏病 : 0.86363
AD 症 : 0.97597
年痴呆症 : 0.48542
五味子酮 : 0.07336
早老性痴呆 : 0.72262
国阿尔茨海默症 : 0.25977
年性痴呆症 : 0.60395
... ...

Figure 3.11: An example of the eCMeSH Tree

(2) Comparing the ontology to a "gold standard" [MS02];

(3) Human evaluation of the ontology according to a set of predefined criteria, standards, requirements, etc. [LTGP04];

(4) Comparing the ontology with a set of data (e.g., a collection of documents) from the domain to be covered by the ontology [BADW04], also called data-driven evaluation.

Evaluation of ontologies in general is carried out at three basic levels: vocabulary, taxonomy, and (non-taxonomic) semantic relations. However, in this study, we are not intending to evaluate the taxonomy and the non-taxonomic relations (semantic relations) of the eCMeSH Tree, because our extension does not add new tree nodes to the MeSH Tree. Moreover, based on the fact that the MeSH Tree, as part of the Unified Medical Language System (UMLS), has been assessed by human experts against a set of criteria [KS03b, Smi06], our evaluation of the eCMeSH Tree will serve only to evaluate the enhanced ontology vocabulary.

In order to evaluate the eCMeSH Tree terms, we design two CLIR experiments, which are compared with the baseline experiment: one aims to expand queries using the eCMeSH Tree terms just as a dictionary does; the other exploits the hierarchical structure of the eCMeSH Tree to expand queries. The experimental settings are described in Section 4.2.2. Document collection and query sets are processed using the methods discussed in Section 4.2.2.2 and Section 4.2.2.1, respectively. The results of the baseline experiment are discussed in Section 3.4.5.

3.4.1 Definitions of Concepts

Before evaluating the eCMeSH Tree terms, it is necessary to define the concepts involved in the following experiments.

Definition 3 (Depth of tree node)**.** *The depth of a tree node is the minimum number of nodes, traversing from the root node to the current node.*

Due to the special format of a MeSH Tree node number (discussed in Definition 2), the depth of a tree node is computed as follows:

$$depth\,(t) = M + 1$$

where t a tree node
M the number of “dots” existing in the tree number of node t

So according to this calculation, the depth of “C10.228.140.380.100” is 5.

Definition 4 (Parent node)**.** *The parent node of a tree node t, where $t \neq null$, is:*

(1) null, if $depth\,(t) = 1$;

(2) p, if $depth\,(t) - depth\,(p) = 1$ and after removing the last dot and substring behind the dot from t, the remaining part of t is equal to p.

According to this definition, the parent of "C10.228.140.380.100" is "C10.-228.140.380", and is not "C10.228.140", "C10.228", or "C10".

Definition 5 (Child node)**.** *A child node of a tree node t, where $t \neq null$, is:*

(1) null, if t is a leaf node, or no other tree number is formed by appending sub-string to the tree number of t;

(2) c, if $parent\,(c) = t$.

For example, the tree node "C10.228.140.380", which refers to the MeSH term "Dementia", has children, "AIDS Dementia Complex" (C10.228.140.380.-070), "Alzheimer disease" (C10.228.140.380.100), "Aphasia, Primary Progressive" (C10.228.140.380.132), "Creutzfeldt-Jakob Syndrome" (C10.228.140.380.165), and "Dementia, Vascular" (C10.228.140.380.230). But the term "CADASIL" (C10.228.140.380.230.124) is not the child of "Dementia".

Definition 6 (Sibling node)**.** *A sibling node of a tree node t, where $t \neq null$, is*

(1) null, if $|child\,(parent\,(t))| = 1$;

(2) s, if $depth\,(s) = depth\,(t)$ and $parent\,(s) = parent\,(t)$.

In the above example, "AIDS Dementia Complex", "Alzheimer disease", "Aphasia, Primary Progressive", "Creutzfeldt-Jakob Syndrome", and "Dementia, Vascular" are brothers of each other, while "CADASIL", which is a child node of "Dementia, Vascular" is not the sibling node of 'Alzheimer disease".

3.4.2 Baseline Experiment

The baseline experiment (See Section 4.3) provides a standard for other experiments to evaluate the retrieval performance. In this experiment, a bilingual dictionary, "Google and KingSoft Dictionary 2.0" [3] is applied to translate the Chinese queries into English. The details of the baseline experiment are illustrated in Section 4.3. Here, the results of the "Exp2" experiments in Table 4.6 are considered as the standard results of the baseline.

[3] `http://g.iciba.com`

3.4.3 Query Expansion Using the eCMeSH Tree Terms

This experiment is designed to evaluate the vocabulary existing in the eCMeSH Tree. We do the experiment in two ways: including term weight; and excluding term weight. Algorithm 4 shows the steps to take in this experiment.

Algorithm 4: Steps of the experiment of query expansion based on the eCMeSH Tree terms

```
Input  : the indexed document collection ($D_I$), the query set ($Q$)
Output: $MAP$, $AP$

begin
    /* $S = \{t_i | t_i \in q_j, q_j \in Q$. $t_i$ is the $i$th term of $q_j$.$\}$        */
    $Q \xrightarrow{\text{query pre-processing}} S$
    $S \xrightarrow{\text{the eCMeSH Tree terms (expanding)}} S_1$
    $S_1 \xrightarrow{\text{the eCMeSH Tree terms and the Dictionary (translating)}} S_2$
    $S_2, D_I \xrightarrow{\text{Indri (inquiring)}} R$
    Evaluate $R$ using $AP$ and $MAP$ measures
end
```

These steps are similar to those described in Section 4.5.2.1, except that when queries are expanded, the sibling terms of a term are not included. The translation and term weighting follow the methods described in Section 4.5.2.1. The experimental results are shown in Section 3.4.5. Table 3.6, presented as Lemur's "INQUERY" language, is the example of Query 161, processed using the above algorithm.

161	the original query	#combine(0.3333 阿尔茨海默病 0.3333 IDE 0.3333 作用)
	the expanded query	#combine(#syn(0.9760 AD症 0.9760 AD病 0.8636 Alzheimer氏病 0.8205 阿尔茨海默症 0.7226 早老性痴呆 0.6040 年性痴呆症 0.6040 老年性痴呆 0.2446 阿尔茨海默病 0.1862 阿滋海默症 . . .) 0.3333 IDE 0.3333 作用)
	the translated query	#combine(#syn(0.9760 #1(Alzheimer disease) 0.9760 #1(Acute Confusional Senile Dementia) 0.9760 #1(Alzheimer disease, Early Onset) 0.9760 #1(Alzheimer disease, Late Onset) 0.9760 #1(Alzheimer Type Senile Dementia) 0.9760 #1(Alzheimer disease, Focal Onset) 0.2446 #1(Dementia, Alzheimer Type) 0.2446 #1(Early Onset Alzheimer disease) . . .) 0.3333 IDE 0.3333 #1(act on) 0.3333 affect 0.3333 action 0.3333 function 0.3333 effect)

Table 3.6: An example of a query expanded based on the eCMeSH Tree

3.4.4 Query Expansion Using eCMeSH Tree Hierarchy

The hierarchical structure existing in the eCMeSH Tree can also be applied to expand queries. This experiment is aimed at evaluating the contribution of the taxonomic relations among the eCMeSH tree heading terms to CLIR retrieval performance. In this experiment, term weights are included and calculated as described in Section 4.5.2.1.

Spasić and Ananiadou [SAMK05] refer to an algorithm to compute the tree similarity (TS):

$$ts\left(C_1, C_2\right) = \frac{2 \cdot common\left(C_1, C_2\right)}{depth\left(C_1\right) + depth\left(C_2\right)} \tag{3.5}$$

where	C_1	the classes related to Term 1
	C_2	the classes related to Term 2
	$\mathit{common}\left(C_1, C_2\right)$	the number of common classes in the paths leading from the root to the given classes
	$\mathit{depth}\left(C_1, C_2\right)$	the number of classes in the path connecting the root and the given class

In this experiment, the *common* function is defined as Equation 3.6, which is subject to the following conditions: given that C_1 and C_2 denote classes of Term 1 and Term 2 respectively, if $\mathit{common}\left(C_1, C_2\right) = \mathit{depth}\left(C_2\right)$, Term 2 is the parent node of Term 1; otherwise, Term 2 is the sibling of Term 1, because they have the same depth.

$$\mathit{common}\left(C_1, C_2\right) = \begin{cases} \mathit{depth}\left(C_1\right) - 1, & \text{where } \mathit{depth}\left(C_1\right) = \mathit{depth}\left(C_2\right); \\ \mathit{depth}\left(C_2\right), & \text{otherwise.} \end{cases} \tag{3.6}$$

Using this constraint, the TS algorithm expands a Chinese query term only with its siblings and parent in the eCMeSH Tree. Algorithm 5 illustrates the steps to conduct query expansion using the hierarchical information existing in the eCMeSH Tree, which applies Equation 3.5 and Equation 3.6 to select the proper terms as the expansion of a query. The translation step is the same as that in Section 3.4.3. Table 3.7 is Query 161, which is expanded using the TS algorithm. The experimental results are shown in Section 3.4.5

Algorithm 5: Steps of the experiment of query expansion using the eCMeSH Tree Hierarchy

```
Input  : the inexed document collection ($D_I$), the query set ($Q$)
Output: $MAP$, $AP$

begin
    /* $S = \{t_i | t_i \in q_j, q_j \in Q$. $t_i$ is the $i$th term of $q_j$.}     */
    $Q \xrightarrow{\text{query pre-processing}} S$
    $S \xrightarrow{\text{the eCMeSH Tree: parent and sibling nodes (expanding)}} S_1$
    $S_1 \xrightarrow{\text{the eCMeSH Tree terms and the Dictionary (translating)}} S_2$
    $S_2, D_I \xrightarrow{\text{Indri (inquiring)}} R$
    Evaluate $R$ using $AP$ and $MAP$ measures
end
```

161	the original query	#combine(0.3333 阿尔茨海默病 0.3333 IDE 0.3333 作用)
	the expanded query	#combine(#syn(0.3439 痴呆 0.0000 #1(失语, 原发进行性) 0.0148 克-亚综合征 0.2264 #1(痴呆, 血管性) . . .) 0.3333 IDE 0.3333 作用)
	the translated query	#combine(0.3439 Dementia #syn(0.0000 #1(Aphasia, Primary Progressive) 0.0148 #1(Creutzfeldt-Jakob Syndrome) 0.2264 #1(Dementia, Vascular) . . .) 0.3333 IDE 0.3333 #1(act on) 0.3333 affect 0.3333 action 0.3333 function 0.3333 effect)

Table 3.7: An example of a query expanded using the tree selection algorithm

3.4.5 Experimental Results and Discussion

The results, measured using MAP, of the above-mentioned experiments are illustrated in Table 3.8, where the Okapi BM25 model is abbreviated as "BM25"; the query likelihood language model with Jelinek-Mercer smoothing is referred to as "LM"; "Automatic WS" denotes the automatic word segmentation; "Manual WS" stands for word segmentation performed manually; the baseline experiment is represented by "Baseline"; query expansions using the eCMeSH Tree with and without term weights are abbreviated as "eW" and "eN", respectively; and the result of experiment using the tree selection algorithm is marked as "TS".

Table 3.8 suggests that the eCMeSH Tree terms help improve the retrieval performance of Chinese-English Biomedical CLIR, which indicate the improvements of vocabulary of the eCMeSH Tree. The best retrieval performances are attained when the eCMeSH Tree terms are used to expand queries with term weights. For the TREC 2006 Track, it is 0.3055, increased by 16.51%. For the 2007 Track, the performance achieved when Okapi BM25 and manual word segmentation are applied is 0.1899, which is improved by 9.45%.

Query expansion using the eCMeSH Tree without term weights also produces improvements on the retrieval performance, but the gain is less than query expansion using the eCMeSH Tree terms with term weights. This implies the term weight enhances the retrieval performance.

The experiments based on the tree selection perform worst, in most cases, even worse than the baseline. For example, compared with the baseline, the performance of query expansion using the eCMeSH Tree hierarchies drops by

	BM25								LM							
	automatic WS				manual WS				automatic WS				manual WS			
	Baseline	eW	eN	TS	Baseline	eW	eN	TS	Baseline	eW	eN	TS	Baseline	eW	eN	TS
2006	0.2309	**0.2645**	0.2503	0.2176	0.2622	**0.3055**	0.2857	0.2689	0.2278	**0.2385**	0.2497	0.2160	0.2619	**0.2922**	0.2842	0.2605
		(14.55%)	(8.40%)	(-5.76%)		(16.51%)	(8.96%)	(2.56%)		(4.70%)	(9.61%)	(-5.18%)		(11.57%)	(8.51%)	(-0.53%)
2007	0.1353	**0.1413**	0.1435	0.1391	0.1735	**0.1899**	0.1813	0.1563	0.1330	**0.1374**	0.1341	0.1357	0.1695	**0.1897**	0.1799	0.1548
		(4.43%)	(6.06%)	(2.81%)		(9.45%)	(4.50%)	(-9.91%)		(3.31%)	(0.83%)	(2.03%)		(11.92%)	(6.14%)	(-8.67%)

Table 3.8: MAPs for the experiments evaluating the eCMeSH Tree

9.91% when manually segmented terms are applied to the Okapi BM25 model, but for the 2007 Track using the language model and automatic word segmentation, this approach to CLIR increases the final performance by 2.81%. The reason for this instability is that relations among sibling tree nodes are not closer than that among child nodes.

In summary, our experiments prove that the vocabulary of our extended CMeSH Tree, or the eCMeSH Tree, has been improved.

3.5 Summary

In this chapter, we propose a new approach to extend the biomedical ontology, CMeSH Tree. Unlike the other related ontology extensions, we not only add synonyms to the original CMeSH Tree terms, but also assign term weights to all the Chinese terms. Instead of applying text mining approaches, we use a rule-based method plus a term extraction algorithm and mutual information filtering to extract terms, which are considered as Chinese translations, from web pages, because text mining approaches require a biomedical training set in the Chinese language.

The next chapter focuses on using the eCMeSH Tree to implement cross-lingual information retrieval. It evaluates several experiments of using the eCMeSH Tree, compares our eCMeSH query expansion approach with other classic methods such as pseudo-relevance feedback and document translation, and analyses the experimental results.

Chapter 4

Implementation of CLIR Using eCMeSH Tree Terms

In this chapter, we focus on query expansion using eCMeSH Tree terms, which were described in Chapter 3. In order to evaluate the performance of our approach, we launch three "individual" experiments: i) query expansion using the original CMeSH Tree terms, ii) query expansion using pseudo-relevance feedback, and iii) the document translation approach. We then compare the results of the eCMeSH method, measured by mean average precision (MAP), with the results of these three experiments.

4.1 The Approaches to Evaluating Information Retrieval

To measure ad hoc information retrieval in the standard way, three things are required: a collection of documents, a test suite of information needs represented as queries, and a set of relevance judgements. The standard approach to information retrieval evaluation revolves around the notion of relevant and non-relevant documents. With respect to a user's information need, a document in the test collection is given a binary classification as either relevant or non-relevant. This decision is referred to as the *gold standard* or *ground truth* judgement of relevance. The size of the test document collection and suite of information needs should be reasonable. It has been found that the sufficient minimum of information needs is 50.

Relevance is assessed relative to an information need, not a query. A document is relevant if it addresses the stated information need, not because it contains all or some the words in the query.

4.1.1 Standard Test Collections

A series of standard test collections have been applied to evaluate IR performance, for example:

The *Cranfield* collection The Cranfield collection [1] was the pioneering test collection, which allows precise quantitative measures of information retrieval effectiveness. The Cranfield collection was assembled in the United Kingdom in the late 1950s. It contains 1398 abstracts of aerodynamics journal articles, a set of 225 queries, and exhaustive relevance judgements of all query-document pairs. The Cranfield collection is too small to satisfy the requirement of large experiments.

Text REtrieval Conference (TREC) Since 1992 the United States National Institute of Standards and Technology (NIST) has been building a large information retrieval evaluation framework. There have been many tracks over a range of different test collections. The best known test collections are the ones used for the TREC Ad Hoc track during the first eight TREC evaluations between 1992 and 1999. These test collections consist of 1.89 million documents (most of them are newswire documents) and relevance judgements for 450 information needs, which are called topics and specified in detailed text passages. The early TRECs comprise 50 topics and the test collections are different but overlap. TRECs 6-8 cover 150 topics and 528,000 newswire articles. TREC information retrieval tracks also include domain-specific retrieval tasks. For example, information retrieval in biomedical literature was in 2006 and 2007. The test collections used established in TREC genomics tracks are also used in our research.

GOV2 In recent years, NIST has evaluated larger document collections. GOV2 is one of them. This collection contains 25 million web pages, and is currently the largest web collection readily available for research purposes.

[1] `http://ir.dcs.gla.ac.uk/resources/test_collections/cran/`

NII Test Collection for IR System (NTCIR) The NTCIR project has built various test collections of similar sizes to the TREC collections. They concentrate on East Asian languages and cross-lingual information retrieval.

Cross Language Evaluation Forum (CLEF) This evaluation series focuses on European languages and cross-language information retrieval.

Reuters-21578 and Reuters-RCV1 The Reuters-21578 collection is usually used in text classification. It contains 21578 newswire articles. More recently, Reuters released the much larger Reuters Corpus Volume 1 (RCV1), which contains 806,791 documents.

In this study, the TREC 2006 and 2007 Genomics Track test collections have been used to evaluate query expansion using eCMeSH. The 2006 and 2007 tasks share the same document collection, containing 162,259 (about 11.9 GB) full-text biomedical publications in HTML format. The 2006 task has 28 queries, in which Query 173 and 180 have no relevant documents, so the number of the 2006 task queries used in this thesis is 26; while the 2007 task provides 36 queries.

4.1.2 Evaluation of Unranked Retrieval Sets

The two most frequent and basic measures for information retrieval effectiveness are precision and recall, which were first used by Kent et al. [KBLJP55]. The following Venn diagram (Figure 4.1) shows the measures of precision and recall.

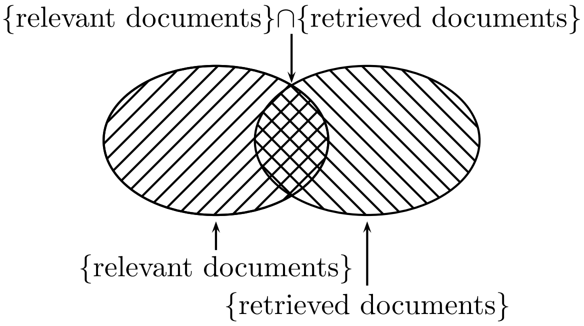


Figure 4.1: Precision and recall

Precision (P) is the percentage of retrieved documents that are relevant. It measures the exactness or fidelity of retrieval.

$$\begin{aligned} Precision &= P(\text{relevant}|\text{retrieved}) \\ &= \frac{|\{\text{relevant documents}\} \cap \{\text{retrieved documents}\}|}{|\{\text{retrieved documents}\}|} \end{aligned}$$

Recall (R) is the percentage of relevant documents that are retrieved. It evaluates the completeness of retrieval.

$$\begin{aligned} Recall &= P(\text{retrieved}|\text{relevant}) \\ &= \frac{|\{\text{relevant documents}\} \cap \{\text{retrieved documents}\}|}{|\{\text{relevant documents}\}|} \end{aligned}$$

The advantage of using precision and recall is that one is more important than the other in many cases. For example, web searches always provide users with ordered results where the first items are most likely to be relevant to particular queries (high precision), but they are not designed for returning every relevant document (high recall) to users. In contrast, some professional and special search engines like law or legislation search and desktop search engines require high recall results. However, recall is a non-decreasing function of the number of documents retrieved: users can always get a recall of 1 by retrieving all documents for all queries. On the other hand, precision usually decreases as the number of documents retrieved is increased.

The *F measure* [vR79] is usually used to trade off precision versus recall.

$$F = \frac{1}{\alpha\frac{1}{P} + (1-\alpha)\frac{1}{R}} = \frac{(\beta^2+1)PR}{\beta^2 P + R}$$

where $\beta^2 = (1-\alpha)/\alpha$, and $\alpha \in [0,1]$. Values of $\beta < 1$ emphasise precision, while values of $\beta > 1$ emphasise recall. The default *balanced F measure* weights precision and recall equally, which means $\alpha = 0.5$ or $\beta = 1$:

$$F_{\beta=1} = \frac{2PR}{P+R}$$

4.1.3 Evaluation of Ranked Retrieval Results

Ranked retrieval results are now standard with search engines. Precision, recall, and F measure are not suitable for directly measuring the ranked results. The following section describes some of the evaluation approaches.

4.1.3.1 11-Point Interpolated Average Precision

In a typical retrieval system, relevant documents, usually the top k relevant document, are returned to users as an ordered set according to their ranks. For the set, precision and recall values can be plotted to give a *precision-recall curve*, which has a distinctive saw-tooth shape. In order to remove these jags, it is standard practice to use an interpolated precision. The interpolated precision $P_{interpolated}$ at a certain recall level R is defined as the highest precision found for any recall level $R' \leq R$:

$$P_{interpolated}\,(R) = \max_{R' \leq R} P(R')$$

Examining the entire precision-recall curve is very informative. In practice, there is often a requirement to use a few numbers or a single number to represent the result. The first eight TREC Ad Hoc evaluations use *11-point interpolated average precision.* Table 4.1 is an example of 11-point interpolated average precision of Query 160 of the TREC 2006 Geonomics Track. The interpolated precision is measured at the 11 recall levels of 0.0, 0.1, 0.2, . . . , 1.0. For each recall level, the arithmetic mean of the interpolated precision at that recall level for each information need in the test collection is used as the final precision at that recall level.

Recall	**Interpolated Precision**
0.00	0.6400
0.10	0.5714
0.20	0.1175
0.30	0.0000
0.40	0.0000
0.50	0.0000
0.60	0.0000
0.70	0.0000
0.80	0.0000
0.90	0.0000
1.00	0.0000

Table 4.1: An example of 11-point interpolated average precision

4.1.3.2 Mean average precision

Mean average precision provides a single-figure measure of quality across recall levels. Among various evaluation measures, MAP has been shown to have especially good discrimination and stability. *Average precision* (AP) is the average of the precision obtained for the set of the top k documents existing after each relevant document is retrieved, and this value is then averaged over information needs. If the set of relevant documents for an information need $q_j \in Q$ is $\{d_1, d_2, \ldots, d_m\}$ and R_{jk} is the set of ranked retrieval results from the top result until document d_k appears, then:

$$MAP\,(Q) = \frac{1}{|Q|} \sum_{j=1}^{|Q|} \left(\frac{1}{m_j} \sum_{k=1}^{m_j} Precision\,(R_{jk}) \right)$$

The MAP value estimates the average area under the precision-recall curve for a set of queries. The above measure calculates all recall levels. For many applications, measuring at fixed low levels of retrieved results, such as 10 or 30 documents, is useful. This is referred to as precision at k. It has the advantage that any estimate of the size of the set of relevant documents is not required. But it is the least stable of the commonly used evaluation measures and does not average well.

In our study, we use AP and MAP to measure the results of all experiments, because MAP evaluates the performance of IR over the entire query set. The first 1,000 returned documents are concerned when calculating MAP, that is, m_j in Section 4.1.3.2 is set to 1,000.

4.1.3.3 R-precision

R-precision requires a set of known relevant documents *Rel*, from which the precision of the top *Rel* documents returned is calculated. If there are $|Rel|$ relevant documents for a query and the top $|Rel|$ results of a system are examined with the result that r of them are relevant, then the precision (referred to as R-precision) and recall are all $r/|Rel|$. R-precision describes only one point on the precision-recall curve, rather than attempting to summarize effectiveness across the curve. However, R-precision is highly correlated with MAP empirically.

4.1.3.4 Receiver operating characteristics (ROC) curve

The *Receiver operating characteristics (ROC) curve* is another concept used in evaluation of IR. An ROC curve plots the true positive rate or sensitivity against the false positive rate or (1-specificity).

4.1.3.5 Cumulative gain

Cumulative gain, and in particular *normalised discounted cumulative gain (NDCG)* are employed with machine learning approaches to ranking. NDCG is designed for situations of non-binary notions of relevance. It is evaluated over some number k of top search results. For a set of queries Q, let $R(j, d)$ be the relevance score assessors gave to document d for query j. Then,

$$NDCG\,(Q, k) = \frac{1}{|Q|} \sum_{j=1}^{|Q|} Z_{kj} \sum_{m=1}^{k} \frac{2^{R(j,m)} - 1}{\log_2 (1 + m)}$$

where Z_{kj} is a normalisation factor calculated to make it so that a perfect rank's NDCG at k for query j is 1. For queries for which $k' < k$ documents are retrieved, the last summation is done up to k'.

4.1.4 Comparison of Different Retrieval Experiments

The previous section discussed the measures used to evaluate the performance of each round of retrieval, but these measures cannot be applied to determine the better retrieval approach between two different retrieval methods. The reason for this is that some queries, such as Query 164, whose MAPs are very low, are harder than others, which introduces an inherent noise in an evaluation. The relevance of documents is judged by human, thus is affected by the behaviour of the human judge.

Statistical significance tests plays an important role in helping researchers to decide whether a retrieval approach performs truly or by chance better than do the others, given the set of queries, judgements, and documents in an evaluation. A good significance test allows researchers to determine significant improvements even if the improvements are small.

In IR, three commonly used approaches to significance test are: the Wilcoxon test [Wil45], Student's t-test [SXP01] and the sign test [WL06]. According to

Smucker et al. [SAC07], the Wilcoxon test and sign test "disagree with the other tests and each other. [...] the Wilcoxon and sign test should no longer be used by IR researchers." [SAC07, p.623] Therefore, in our study, t-test [SXP01] are applied to conduct significance test.

A significance test consists of three ingredients:

(1) A test statistic used to judge the two retrieval approaches. In our study, the test statistic is computed based on average precision of each individual query, using Equation 4.1.

(2) A *null hypothesis* H_0. Our null hypothesis states that both retrieval methods result in the similar retrieval performance.

(3) A significance level α which is the probability of mistakenly rejecting the null hypothesis. In our evaluations, α is set to 0.05.

If a retrieval approach is presented as $X = (x_1, x_2, \ldots x_n)$, where x_i is the AP of the ith query, the null hypothesis H_0 of t-test at level α can be judged by Equation 4.1.

$$t = \frac{\overline{X_0} - \overline{X_1}}{s_w \cdot \sqrt{\dfrac{1}{n_1} + \dfrac{1}{n_0}}} \leq -t_{0.05}\left(n_0 + n_1 - 2\right) \tag{4.1}$$

$$\begin{aligned} s_w^2 &= \frac{\left(n_0 - 1\right) S_0^2 + \left(n_1 - 1\right) S_1^2}{n_0 + n_1 - 2} \\ &= \frac{\sum_{i=1}^{n_1}\left(X_{1,i} - \overline{X_1}\right)^2 + \sum_{j=1}^{n_0}\left(X_{0,i} - \overline{X_0}\right)^2}{n_0 + n_1 - 2} \end{aligned}$$

where	$\overline{X_0}$	the average of all individual APs of method X_0
	$\overline{X_1}$	the average of all individual APs of method X_1
	$t_{0.05}\left(n_0 + n_1 - 2\right)$	the quantile of the t distribution when significance level is 0.05 and the degree of freedom is $n_0 + n_1 - 2$
	S_0^2	the sample variance of X_0
	S_1^2	the sample variance of X_1
	n_0	the total number of queries in X_0
	n_1	the total number of queries in X_1

In general, t_λ is available in a t-distribution quantile table. If Equation 4.1 is true, the H_0 needs to be rejected, which indicated there is a significant difference existing between two retrieval methods. The following is an example.

X_0	**Query ID**	160	161	162	163	164	165	166	167	168
	AP	0.2349	0.000	0.2792	0.5660	0.0000	0.1153	0.0008	0.1629	0.8902
	Query ID	169	170	171	172	174	175	176	177	178
	AP	0.2686	0.7566	0.0002	0.3700	0.0837	0.2592	0.2188	0.0000	0.3611
	Query ID	179	181	182	183	184	185	186	187	
	AP	0.0080	0.2327	0.4151	0.2214	0.1000	0.6005	0.0004	0.6667	
	$\overline{\mathbf{X_0}}$	0.2622								
	$\mathbf{S_0^2}$	1.6252								
X_1	**Query ID**	160	161	162	163	164	165	166	167	168
	AP	0.2206	0.0000	0.2583	0.4905	0.0000	0.1141	0.0001	0.1533	0.8327
	Query ID	169	170	171	172	174	175	176	177	178
	AP	0.2602	0.5685	0.0000	0.3449	0.0755	0.2417	0.2098	0.0000	0.3367
	Query ID	179	181	182	183	184	185	186	187	
	AP	0.0052	0.2157	0.3920	0.2007	0.0854	0.4976	0.0000	0.5000	
	$\overline{\mathbf{X_1}}$	0.2309								
	$\mathbf{S_1^2}$	1.1764								
t	0.09529									

Table 4.2: An example of t-test significance test

The retrieval performance of each individual query of 2006 is listed in Table 4.2, where "X_0" refers to the baseline experiment, described in Section 4.3, using Okapi BM25 and manual word segmentation without stemming; "X_1" stands for the experiment under the same conditions as X_0 except for automatic word segmentation.

From a t-test table, $t_{0.05}\,(50)$ is 1.6759. t is greater than -1.6759. This means that retrieval methods X_0 and X_1 do not have significant difference.

In following experiments, we evaluate the significant difference between the approach leading to the best performance with others and mark the method which has a significant difference.

4.2 Experimental Settings

In this section, we first introduce the toolkits designed for establishing IR systems; and then we discuss the necessary pre-processing for both queries and documents.

4.2.1 Experimental Frameworks and Toolkits

Software toolkits have been designed to help implement IR applications, including SMART, Okapi, Lucene, and Lemur.

SMART SMART [2] is one of the most famous and widely-used experimental systems in information retrieval. It was originally developed by Gerard Salton and his colleagues in the 1980s, and the newest release is SMART 11. The source codes of SMART can be freely downloaded from the official site, so it is convenient for researchers to make the system satisfy their requirements.

SMART applies the vector space model to implement information retrieval. It provides a complete processing for information retrieval. Users use it first to index a set of documents, then retrieve documents for their queries, and finally evaluate the results via the system. SMART has word stopping, stemming, and weighting modules, which are able to be called by users according to the requirements of tasks.

A major constraint of SMART is that the maximum size of the document collection is 500 MB. Another problem is that SMART does not provide good API documents, so users have to look at the source code before hacking. Moreover, SMART can only process documents in English.

Okapi Okapi [3] is another well-known experimental document retrieval system, which was developed at the City University, London at the end of the 1980s. It is based on a probabilistic retrieval model, and uses BM25 and its variants to calculate term weight. Okapi provides indexing and searching utilities for users. Many exciting results in TREC tracks have been obtained through the Okapi system.

The problem with the Okapi framework is that no source code is available, so it is impossible to customise the system for a special task.

Lucene Toolkit Unlike the Lemur toolkit, the Lucene toolkit [4] is designed to provide a high performance, scalable information retrieval application programming interface (API) in Java. It is not a ready-to-use and full-featured

[2] `ftp://ftp.cs.cornell.edu/pub/smart/`

[3] `http://www.soi.city.ac.uk/~andym/OKAPI-PACK/index.html`

[4] `http://lucene.apache.org/`

search application but a software library. Lucene concerns itself with text indexing and searching. Its query expressions are the Boolean operators (AND, OR, and NOT) and limited regular functions on terms and fields (a term is used in Lucene to describe the metadata of articles, like title, author, URL, and so on).

Lemur Toolkit The Lemur toolkit is one of the products of the Lemur project [5], which was started in 2000 by the Center for Intelligent Information Retrieval (CIIR) at the University of Massachusets, Amherst, and the Language Technologies Institute (LIT) at Carnegie Mellon University. The Lemur toolkit comprises an open-source Indri search engine which provides a combination of inference network and language model for retrieval, a query log toolkit to capture and analyse user interaction data, and a set of structured query operators. It also includes other linguistic tools such as a Chinese part-of-speech tagger and a stemmer for Arabic texts.

The Lemur toolkit is featured as the structured query language, *INQUERY* [CCH92], which is illustrated by the following structured Query 161 of the TREC 2006 Genomics Track (Table 4.3).

```
#weight(0.5 #combine(0.3333 #2(Alzheimer disease) 0.3333 IDE 0.3333 action) 0.5 #weight(0.7421 Alzheimer 0.7019 illness 0.6504 disease 0.5723 function 0.4710 affected 0.3241 damages 0.3200 treatment 0.2988 brain ...))
```

Table 4.3: An example of INQUERY

The decimal before a term is used to weight the importance of this term. The strings starting with "#" are operators. For example, "#weight" indicates that the terms contribute unequally to the final result according to the weights associated with each of them. "#combine" means that terms have equal influence on the final result. "#*N*" is an ordered distance operator, which controls the window of *N* words within which all occurrences of terms appear ordered. For more details, please refer to the Lemur manual.

In this study, we construct the experimental CLIR system using the Lemur toolkit, taking advantage of its INQUERY query language and the built-in retrieval models.

[5] http://www.lemurproject.org/

4.2.2 Experimental Settings and Pre-Processing

4.2.2.1 Query Pre-Processing

The original test set for the TREC Genomics Tracks was written in English. Before being used in our experiments, it needs to be translated into Chinese and separated into words according to corresponding policies.

1. Query Translation

The original 2006 and 2007 Genomics task topics are manually translated into Chinese before any experiment, since there are no Chinese topics designed for biomedical information retrieval. Researchers such as [LOR04] have applied the same translation policy to cope with the construction of queries. Appendix B lists the original topics used in the 2006 and 2007 TREC Genomics tracks and their Chinese counterparts.

2. Word Segmentation

After the translation, it is necessary to segment the translated sentences into words or phrases before further operations. The word segmentation tool is *BaseSeg* [ZHL06], which is based on the conditional random field model. Since this tool is trained using a newswire domain corpus, many biomedical terms are wrongly recognised. The incorrect query terms significantly decrease the performance of retrieval, illustrated in the following experiments.

Instead of attempting to fix this issue using external lexicons or tools, we manually segment all the queries and compare the results of both strategies. Table 4.4 gives two examples of the results of word segmentation using both an automatic approach and manual processing. All the experiments in our study, except evaluation of retrieval models, are launched using the queries segmented using automatic approach and manual processing, respectively. In the following experiments, we refer to these methods as ‘automatic WS’ and ‘manual WS’, respectively.

3. Word Filtering

When segmentation is finished, a filtering operation is carried out on the Chinese words or phrases, which follows these policies, for the reason that words such as prepositions, adverbs, and link and auxiliary verbs cannot give a positive

161	automatic	在 阿尔茨海 默病 中I DE 的 作用 是 什么
	manual	在 阿尔茨海默病 中 IDE 的 作用 是 什么
200	automatic	什么 血清 [蛋 白质] 改变 狼疮 中 与 高病 活性 相关 的 表达 ？
	manual	什么 血清 [蛋白质] 改变 狼疮 中 与 高病活性 相关 的 表达 ？

Table 4.4: Word segmentation results: automatic and manual approaches

improvement on retrieval performance.

- Words will be removed if they are not nouns or noun phrases, verbs (except link and auxiliary verbs), or adjectives.
- Words without Chinese characters, like "IDE", are retained in the query, since these words are highly likely to indicate terms.
- Punctuation will be erased from the query terms.
- Words including punctuation are retained as query terms.

Table 4.5 shows the filtering results of the above examples.

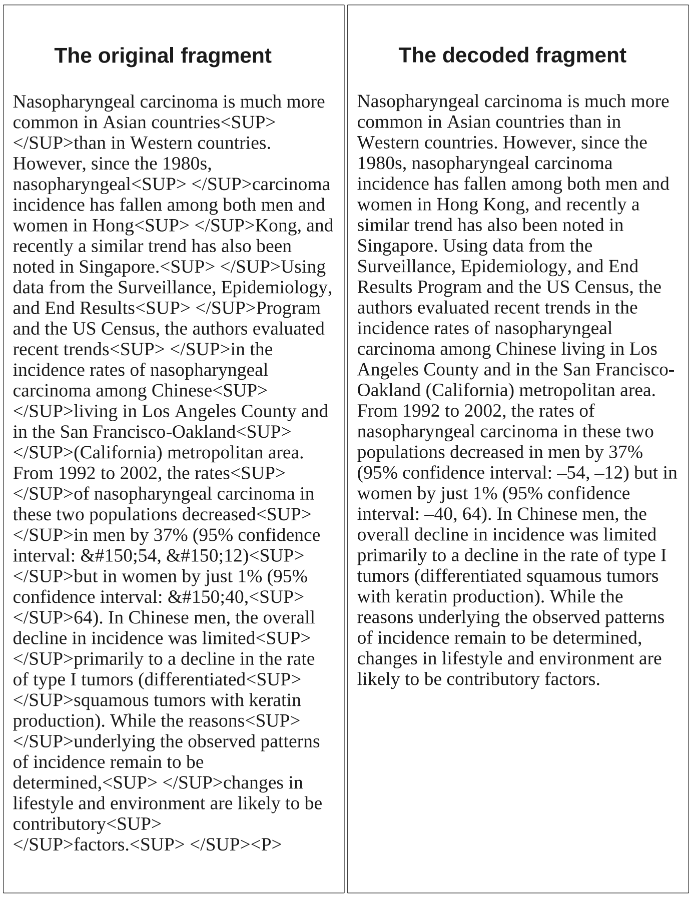
161	automatic	阿尔茨海 默病 DE 作用
	manual	阿尔茨海默病 IDE 作用
200	automatic	血清 [蛋 白质 改变 狼疮 高病 活性 表达
	manual	血清 蛋白质 改变 狼疮 高病活性 表达

Table 4.5: Filtering results: automatic and manual approaches

4.2.2.2 Document Pre-Processing

1. Document Decoding and Conversion to Plain Text

The 2006 and 2007 TREC Genomics tracks are designed for paragraph retrieval, that is, to find out the paragraphs which are relevant to each query. However, ad hoc retrieval, which finds all documents discussing a particular topic, is the only task we launch. Therefore, all metadata stored in documents are deleted, and all documents are restored as plain texts, before other operations. Figure 4.2 gives a fragment of a document numbered "16282240" and the result of decoding.



The original fragment

Nasopharyngeal carcinoma is much more common in Asian countriesthan in Western countries. However, since the 1980s, nasopharyngealcarcinoma incidence has fallen among both men and women in HongKong, and recently a similar trend has also been noted in Singapore.Using data from the Surveillance, Epidemiology, and End ResultsProgram and the US Census, the authors evaluated recent trendsin the incidence rates of nasopharyngeal carcinoma among Chineseliving in Los Angeles County and in the San Francisco-Oakland(California) metropolitan area. From 1992 to 2002, the ratesof nasopharyngeal carcinoma in these two populations decreasedin men by 37% (95% confidence interval: &#150;54, &#150;12)but in women by just 1% (95% confidence interval: &#150;40,64). In Chinese men, the overall decline in incidence was limitedprimarily to a decline in the rate of type I tumors (differentiatedsquamous tumors with keratin production). While the reasonsunderlying the observed patterns of incidence remain to be determined,changes in lifestyle and environment are likely to be contributoryfactors.<P>

The decoded fragment

Nasopharyngeal carcinoma is much more common in Asian countries than in Western countries. However, since the 1980s, nasopharyngeal carcinoma incidence has fallen among both men and women in Hong Kong, and recently a similar trend has also been noted in Singapore. Using data from the Surveillance, Epidemiology, and End Results Program and the US Census, the authors evaluated recent trends in the incidence rates of nasopharyngeal carcinoma among Chinese living in Los Angeles County and in the San Francisco-Oakland (California) metropolitan area. From 1992 to 2002, the rates of nasopharyngeal carcinoma in these two populations decreased in men by 37% (95% confidence interval: –54, –12) but in women by just 1% (95% confidence interval: –40, 64). In Chinese men, the overall decline in incidence was limited primarily to a decline in the rate of type I tumors (differentiated squamous tumors with keratin production). While the reasons underlying the observed patterns of incidence remain to be determined, changes in lifestyle and environment are likely to be contributory factors.

Figure 4.2: An example of document decoding

2. Tokenisation

After converting the documents into plain text, a tokenisation or lexical analysis process is applied to translate the characters into "words" or "tokens". Tokenisation can decrease the length of index terms, hence index efficiency may be improved by this processing. Tokenisation takes the following factors into account:

Word separation

Biomedical terms are often composed of multiple words. Nenadic et al. [NSA05] observe that in a collection of MEDLINE citations, 85% of the terms consisted of more than one word. Splitting multi-word terms into multiple individual tokens may result in nondescript index terms. However, Carpenter [Car04] reported that phrase-based retrieval performed worse than word-based retrieval. In our study, we use a word-based strategy for indexing documents, because the performance of phrase-based retrieval depends on the precise phrase recognition. *N-gramming* tokenisation has been successfully used for languages without proper word separators such as Chinese, Japanese and Korean. Research has shown that in English, word-based searching performs no worse than n-gram based searching, so in our research, n-gramming tokenisation is not applied.

Case

In English texts, the first letter of sentences, proper names, and some abbreviations always make use of uppercase letters. As for some gene and protein names, upper and lowercase letters are often mixed, for example, 'PrP Proteins' and 'PrPSc Proteins'. Jiang and Zhai [JZ07] report that since capitalisation is not consistent, the cases of letters might not affect the retrieval performance. In our study, all uppercase letters are converted to lowercase.

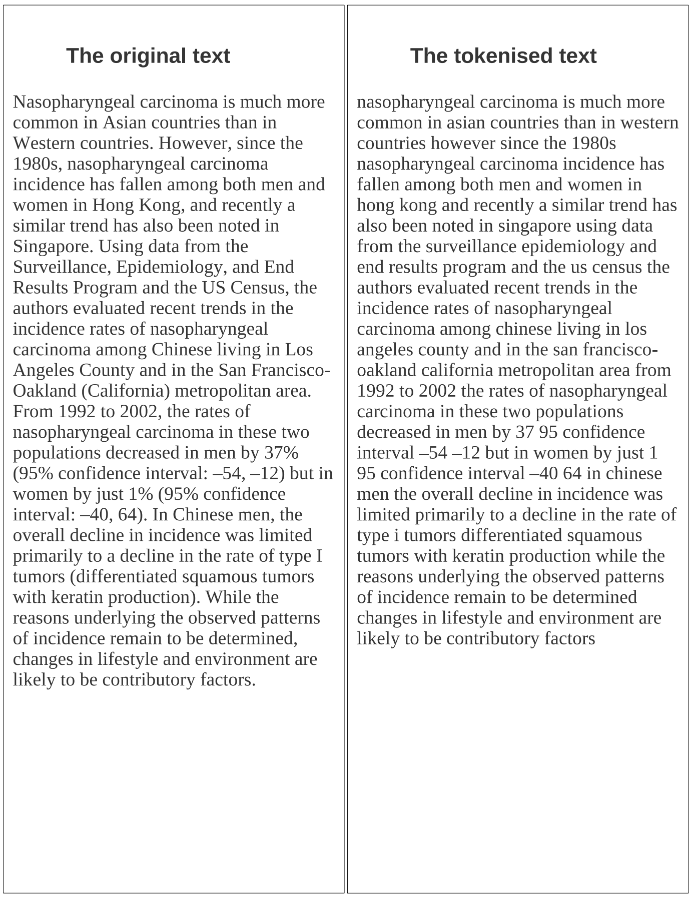
Numbers

In general domain information retrieval, numbers are usually ignored because they are typically not valuable in an index without their surrounding context. However, in biomedical IR, since numbers always indicate the sub-family, members or other variants of proteins and genes, it is necessary to keep numbers in the index vocabulary. The problem of numbering in the biomedical domain is that various numbers are used, for example, Arabic numerals ('HT29 Cells'), the Greek alphabet ('Betaherpesvirinae' or 'β herpesvirinae'), and Roman numbers ('PTLV II'). Unlike other studies [BCC04, HHR06, JZ07], in which numbers are made uniform during lexical analysis, we retain numbers' original forms. Instead of providing a proximity operator to make sure the split numbers and words appeared close together in matching documents [PL03], we apply a policy that is similar to Tomlinson [Tom03]'s strategy: if a number is connected to a term without white space, then it is treated as part of this term; otherwise, it is separated as a new term.

Punctuation

In our experiments, all punctuation marks except hyphen and parentheses are removed before indexing. Hyphens are used to attach prefixes and suffixes. Our strategy to hyphens is: i) all in-term hyphens are retained; and ii) all hyphens at the end or beginning of words are removed. Parentheses are used to provide additional or explanatory information. We remove all parentheses during lexical analysis.

Figure 4.3 illustrates the tokenisation result of the above example texts.



The original text

Nasopharyngeal carcinoma is much more common in Asian countries than in Western countries. However, since the 1980s, nasopharyngeal carcinoma incidence has fallen among both men and women in Hong Kong, and recently a similar trend has also been noted in Singapore. Using data from the Surveillance, Epidemiology, and End Results Program and the US Census, the authors evaluated recent trends in the incidence rates of nasopharyngeal carcinoma among Chinese living in Los Angeles County and in the San Francisco-Oakland (California) metropolitan area. From 1992 to 2002, the rates of nasopharyngeal carcinoma in these two populations decreased in men by 37% (95% confidence interval: –54, –12) but in women by just 1% (95% confidence interval: –40, 64). In Chinese men, the overall decline in incidence was limited primarily to a decline in the rate of type I tumors (differentiated squamous tumors with keratin production). While the reasons underlying the observed patterns of incidence remain to be determined, changes in lifestyle and environment are likely to be contributory factors.

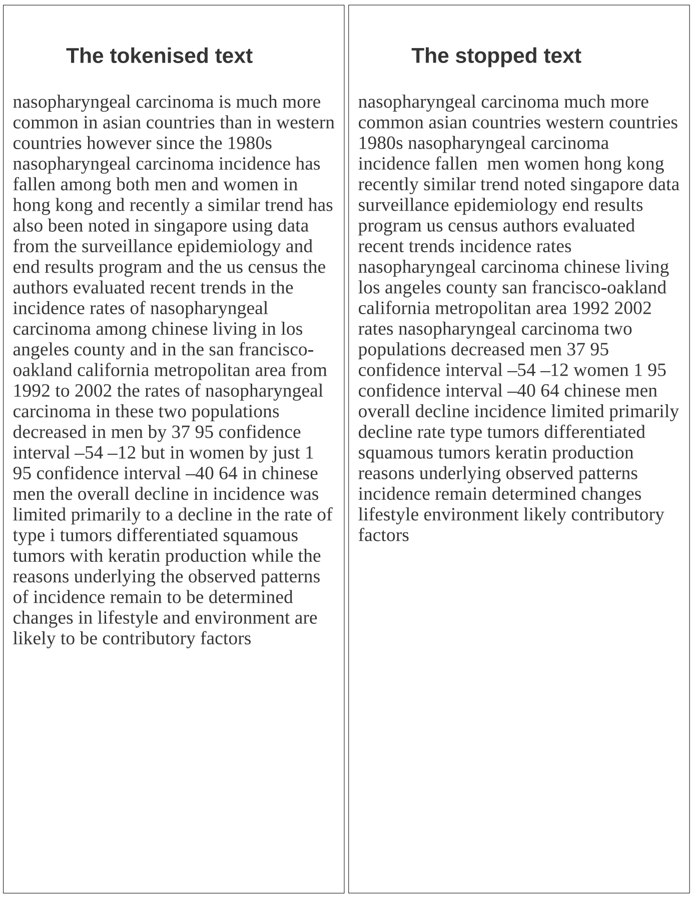
The tokenised text

nasopharyngeal carcinoma is much more common in asian countries than in western countries however since the 1980s nasopharyngeal carcinoma incidence has fallen among both men and women in hong kong and recently a similar trend has also been noted in singapore using data from the surveillance epidemiology and end results program and the us census the authors evaluated recent trends in the incidence rates of nasopharyngeal carcinoma among chinese living in los angeles county and in the san francisco-oakland california metropolitan area from 1992 to 2002 the rates of nasopharyngeal carcinoma in these two populations decreased in men by 37 95 confidence interval –54 –12 but in women by just 1 95 confidence interval –40 64 in chinese men the overall decline in incidence was limited primarily to a decline in the rate of type i tumors differentiated squamous tumors with keratin production while the reasons underlying the observed patterns of incidence remain to be determined changes in lifestyle and environment are likely to be contributory factors

Figure 4.3: Tokenisation of the text decoded in Figure 4.2

3. Stopword Removal

Frequently used and uninformative words are commonly filtered out from tokens. This operation not only reduce the number of index terms, but also removes words that are rarely used for searching. We exploit the standard PubMed stop list [6] to remove stopwords. Figure 4.4 shows the result of stopping unwanted words in the above example.



The tokenised text

nasopharyngeal carcinoma is much more common in asian countries than in western countries however since the 1980s nasopharyngeal carcinoma incidence has fallen among both men and women in hong kong and recently a similar trend has also been noted in singapore using data from the surveillance epidemiology and end results program and the us census the authors evaluated recent trends in the incidence rates of nasopharyngeal carcinoma among chinese living in los angeles county and in the san francisco-oakland california metropolitan area from 1992 to 2002 the rates of nasopharyngeal carcinoma in these two populations decreased in men by 37 95 confidence interval –54 –12 but in women by just 1 95 confidence interval –40 64 in chinese men the overall decline in incidence was limited primarily to a decline in the rate of type i tumors differentiated squamous tumors with keratin production while the reasons underlying the observed patterns of incidence remain to be determined changes in lifestyle and environment are likely to be contributory factors

The stopped text

nasopharyngeal carcinoma much more common asian countries western countries 1980s nasopharyngeal carcinoma incidence fallen men women hong kong recently similar trend noted singapore data surveillance epidemiology end results program us census authors evaluated recent trends incidence rates nasopharyngeal carcinoma chinese living los angeles county san francisco-oakland california metropolitan area 1992 2002 rates nasopharyngeal carcinoma two populations decreased men 37 95 confidence interval –54 –12 women 1 95 confidence interval –40 64 chinese men overall decline incidence limited primarily decline rate type tumors differentiated squamous tumors keratin production reasons underlying observed patterns incidence remain determined changes lifestyle environment likely contributory factors

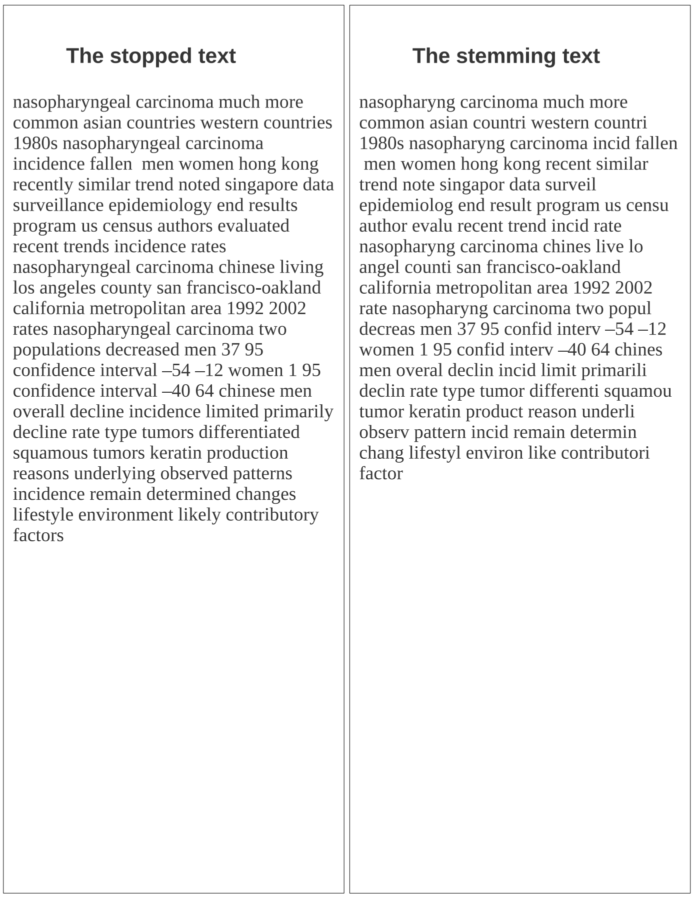
Figure 4.4: The result of stopword removal

4. Stemming

Stemming is aimed to remove word endings to obtain the root of the word. This operation can enhance retrieval recall: the same query term returns more documents containing the actual query term or a similar word. However, the document containing words with meaning unrelated to the same stem may also be returned. For example, if Porter stemming [Por80] is used, ‘universe’ and ‘university’ share the same stem ‘univers’. In order to overcome this limitation,

[6] http://www.ncbi.nlm.nih.gov/books/NBK3827/table/pubmedhelp.T43/

Zhou and Yu [ZY06] tried to prevent stemming when a word looked like a gene name. Urbain et al. [UGF06] only used stemming when the word was not an acronym. In this study, retrieval with and without stemming have been reviewed respectively at the baseline stage (discussed in Section 4.3). Based on these results, stemming has not been applied in the other experiments. Figure 4.5 illustrates the result of the given example texts processed by Porter stemming.



The stopped text

nasopharyngeal carcinoma much more common asian countries western countries 1980s nasopharyngeal carcinoma incidence fallen men women hong kong recently similar trend noted singapore data surveillance epidemiology end results program us census authors evaluated recent trends incidence rates nasopharyngeal carcinoma chinese living los angeles county san francisco-oakland california metropolitan area 1992 2002 rates nasopharyngeal carcinoma two populations decreased men 37 95 confidence interval –54 –12 women 1 95 confidence interval –40 64 chinese men overall decline incidence limited primarily decline rate type tumors differentiated squamous tumors keratin production reasons underlying observed patterns incidence remain determined changes lifestyle environment likely contributory factors

The stemming text

nasopharyng carcinoma much more common asian countri western countri 1980s nasopharyng carcinoma incid fallen men women hong kong recent similar trend note singapor data surveil epidemiolog end result program us censu author evalu recent trend incid rate nasopharyng carcinoma chines live lo angel counti san francisco-oakland california metropolitan area 1992 2002 rate nasopharyng carcinoma two popul decreas men 37 95 confid interv –54 –12 women 1 95 confid interv –40 64 chines men overal declin incid limit primarili declin rate type tumor differenti squamou tumor keratin product reason underli observ pattern incid remain determin chang lifestyl environ like contributori factor

Figure 4.5: The result of stemming

4.3 Baseline Experiment

The baseline experiment built in this section is the plain dictionary-based approach to CLIR. The purposes of this experiment are: i) to provide a basic retrieval performance for comparison; ii) to establish a baseline IR system, which can be easily extended to more complicated applications; and iii) to evaluate the retrieval performance with and without stemming.

4.3.1 Steps

In the baseline experiment, all documents in the 2006 TREC Genomics test collection are processed following the approaches described in Section 4.2.2.2; all queries of the 2006 and 2007 tracks have been processed using the methods described in Section 4.2.2.1. Because Query 173 and Query 180 have no relevant documents in the test collection, these two queries are ignored in all experiments. The dictionary used to translate Chinese words into English is “Google and Kingsoft Dictionary 2.0” (谷歌金山词霸 2.0) [7]; and in the following experiments, “the Dictionary” refers to the Google and Kingsoft Dictionary 2.0. If more than one translation is available, the first one in the candidate translation will be selected. Indexing and retrieval are carried out using the Lemur toolkit. The retrieval model used in the baseline is Okapi BM25, abbreviated to “BM25”; and the query likelihood language model with Jelinek-Mercer smoothing is abbreviated to “LM”, its parameters using the system default value. The policy for out-of-vocabulary words which contain Chinese characters is neglect, i.e., all words containing Chinese characters that cannot be translated will be ignored, but English acronyms or abbreviations will be retained in the query. The retrieval performance is measured using MAP.

We launch two experiments: one is processed using the Porter stemming algorithm; the other uses original words in the documents as index terms. The steps to carry out these experiments are described by Algorithm 6.

4.3.2 Results

Table 4.6 shows the results of the experiments with and without stemming. “Exp1” refers to the experiment using the Porter stemming algorithm, and “Exp2” denotes the experiment without stemming.

In the point view of the types of method used in the baseline experiment, the significance tests can be grouped into three aspects: significance test about word segmentation, significance test of retrieval model, and significance test for

[7] `http://g.iciba.com`

Algorithm 6: Steps of baseline experiment

```
Input  : the document collection ($D$), the query set ($Q$)
Output: $MAP$, $AP$

begin
    /* $S = \{t_i | t_i \in q_j, q_j \in Q.$ $t_i$ is the $i$th term of $q_j$.$\}$ */
    $Q \xrightarrow{\text{query pre-processing}} S$
    $D \xrightarrow{\text{document pre-processing}} D_1$
    $D \xrightarrow{\text{document pre-processing with Porter stemming algorithm}} D_2$
    $D_1, D_2 \xrightarrow{\text{Indri (indexing)}} D_3$
    $S \xrightarrow{\text{The Dictionary}} S_1$
    $S_1 \xrightarrow{\text{Porter stemming algorithm}} S_2$
    $S_1, D_3 \xrightarrow{\text{Indri (inquiring)}} R_1$
    $S_2, D_3 \xrightarrow{\text{Indri (inquiring)}} R_2$
    Evaluate $R_1$ and $R_2$ using $AP$ and $MAP$ measures
end
```

	BM25				LM			
	automatic WS		manual WS		automatic WS		manual WS	
	Exp1	Exp2	Exp1	Exp2	Exp1	Exp2	Exp1	Exp2
2006	0.2278	**0.2309**	0.2619	**0.2622**	0.2307	**0.2310**	**0.2619**	**0.2619**
2007	0.1330	**0.1352**	0.1695	**0.1735**	0.1333	**0.1350**	0.1696	**0.1733**

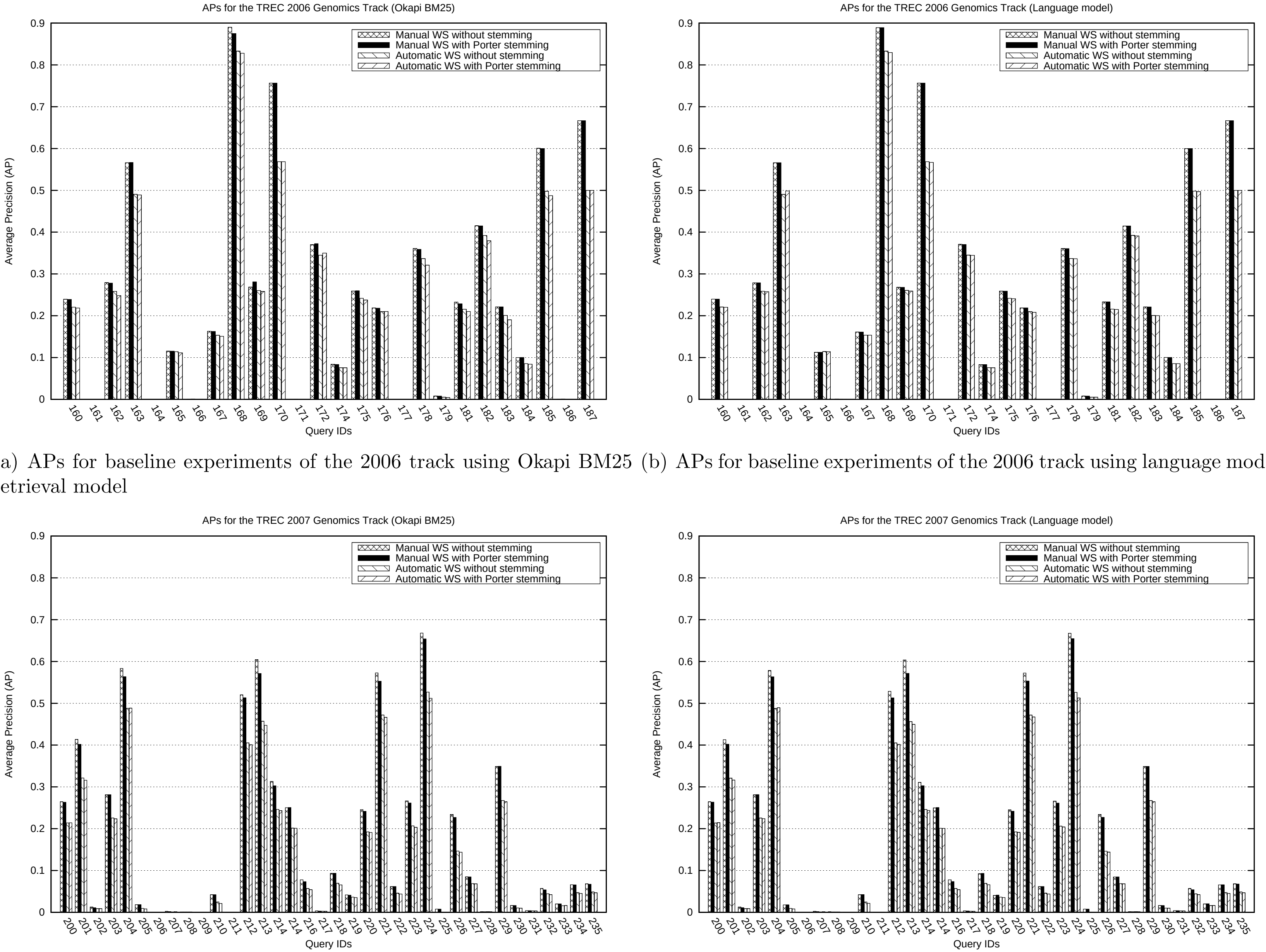
Table 4.6: MAPs for baseline experiments

stemming. For each test, the method which has the best retrieval performance is evaluated with other corresponding methods. The results show that there is no significant difference among these approaches.

The following figures illustrate the APs of each single query in both baseline experiments. Figure 4.6(a) and Figure 4.6(b) are the average precision for each query in the TREC 2006 Genomics Track using different retrieval models. Figure 4.6(c) and Figure 4.6(d) show the experimental results of the 2007 Track.

4.3.3 Analysis and Discussion

The baseline experimental results suggest that:



(a) APs for baseline experiments of the 2006 track using Okapi BM25 retrieval model

(b) APs for baseline experiments of the 2006 track using language model

(c) APs for baseline experiments of the 2007 track using Okapi BM25 retrieval model

(d) APs for baseline experiments of the 2007 track using language model

Figure 4.6: Baseline results measured by AP

(1) The strategy of word segmentation significantly affects the performance of CLIR.

(2) Stemming does not make any positive contributions to the retrieval performance, which is in agreement with the conclusion of Abdou et al. [ARS05].

(3) The Okapi BM25 retrieval model performs slightly better than the language model.

1. The effect of word segmentation

Table 4.6 shows that the best performance (expressed in bold) of each group is achieved when manual word segmentation is used. For example, the 2006 track's queries under Okapi BM25 attain the best performance of 0.2622 when using manual word segmentation, while in the same situation using automatic word segmentation, the performance declined by 13.56%, to 0.2309. The experiments using the 2007 queries also illustrate a similar decrease in retrieval performance.

It is clear that the segmentation tool, "BaseSeg", causes the performance to drop, because it was trained using newswire corpus. Biomedical terms are wrongly separated into words, since the tool lacks the essential knowledge to identify biomedical terms from context; these wrong terms cannot be found in the Dictionary and are ignored in the further steps. Manual word segmentation prevents such incorrect segmentation.

2. The effect of stemming

Almost all the groups of experiments reach the best performances when stemming is not applied. For instance, in Table 4.6, using the language model and manual word segmentation, the queries in the 2007 Track attain a performance of 0.1733 when no stemming algorithm is applied, compared to 0.1696 when Porter stemming is used. However, stemming only has a slight effect on the results. In contrast with the variation caused by word segmentation methods, the change of performance from the stemming algorithm is lower than 3%.

A possible explanation of the fall in retrieval performance is that stemming, which removes word endings to attain word roots, introduces extra errors. Different terms become equivalent because they share the same root. So in the following experiments, "baseline experiment" refers to the baseline that does not use stemming.

3. The effect of retrieval models

Table 4.6 also shows that the Okapi BM25 model performs slightly better than the query likelihood language model, but not significantly. For example, for the TREC 2006 Track, using manual word segmentation, the performances of Okapi BM25 and the language model are 0.2622 and 0.2619, respectively.

4. The retrieval performance for a single query

The retrieval performance has also been evaluated using average precision, illustrated in Figure 4.6. Not all queries are improved by using the Dictionary to translate Chinese terms into English equivalents. For example, the APs of Queries 161, 164, 171 and 206 are zero, because no relevant documents retrieved for them. In contrast, Queries 168, 170, 185, 187, 204, 213 and 224 have high AP values.

4.4 Experiments Concerning Parameter Optimisation

The two popular retrieval models, Okapi BM25 and the language model, are applied in all the experiments. For the language model, we use the query likelihood language model with Jelinek-Mercer smoothing. However, the parameters of these models in the Lemur toolkit are trained using the newswire corpora, which may not be optimised for biomedical texts. In this section, we conduct experiments to evaluate the optimal parameters of both retrieval models.

4.4.1 Steps

The queries of the 2006 and 2007 Tracks are processed using the methods described in Section 4.2.2.1 and translated into English query terms using the Dictionary. Unlike the other experiments, which index all the documents in the 2006 test collection, the experiments carried out in this section require the training set and a test set. Two thirds of the documents in the 2006 document collection (about 108,100 documents) were selected at random for the purposes of training; the remaining documents were used as the test set. Both sets of documents are processed using the document pre-processing methods (Section 4.2.2.2) and then indexed using Indri. Algorithm 7 gives the details of the experimental steps.

The training collection was used to determine the range of each parameter and the step values of parameters. For Okapi BM25 (expressed by Equation 2.2), k_1 , which calibrates the within-document term frequency scaling, ranges from 0.1 to 1.8, stepped by 0.1. k_3 , which controls the importance of query terms, ranges from 0 to 10, stepped by 1. b, which determines the scaling by document length, ranges from 0.05 to 0.95, stepped by 0.05. For the language model (represented as Equation 2.3), the range of λ is from 0.0 to 1.0; and its step value is 0.1. These ranges and step values of parameters were applied to the test collection.

In these experiments, the query set consists of all queries of the TREC 2006 and 2007 Genomics tracks as the parameters are applied to both tasks.

4.4.2 Results

Algorithm 7 attempts to find the parameters which give the best retrieval performance. Through trial and error, we obtained the optimal values of parameters. Okapi BM25 performs best, when $k_1 = 1.4$, $k_3 = 6$, and $b = 0.75$. Figure 4.7 illustrates the segment of experimental data which contains the best retrieval performance. In the case of the query likelihood language model, the optimal value for λ is 0.6, as shown in Figure 4.8. These optimal parameters are applied in all the following experiments.

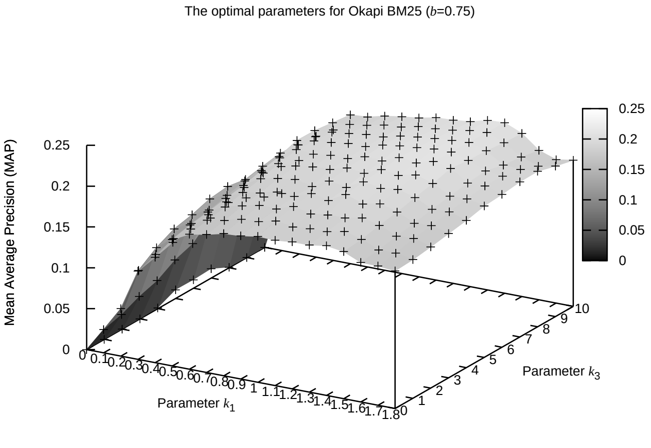


Figure 4.7: Results of parameter optimisation for Okapi Bm25

Algorithm 7: Steps of baseline experiment

```
Input  : the document collection ($(D)$), the query set ($(Q)$)
Output: the optimal $k_1, k_3, \text{and} b$; the optimal $\lambda$

begin
    /* $S = \{t_i | t_i \in q_j, q_j \in Q$. $t_i$ is the $i$th term of $q_j$.$\}$ */
    $Q \xrightarrow{\text{query pre-processing}} S$
    $D_0 \xrightarrow{\text{document pre-processing}} D'$
    $D' \xrightarrow{\text{separation}} D_0$(the training set, about 108,100 documents)
    $D' \xrightarrow{\text{separation}} D_1$(the test set, about 54,200 documents)
    /* Parameter ranges and step values for Okapi BM25 */
    Use $D_0$ to determine $b \in [0.05, 0.95]$, and the step value is 0.05
    Use $D_0$ to determine $k_1 \in [0.1, 1.8]$, and the step value is 0.1
    Use $D_0$ to determine $k_3 \in [0, 10]$, and the step value is 1
    /* Parameter ranges and step values for language model */
    Use $D_0$ to determine $\lambda \in [0.0, 1.0]$, and the step value is 0.1
    /* Optimisation of the parameters of Okapi BM25 */
    $k_{1o} = 0.0, k_{3o} = 0, b_o = 0.0, mmap = -0.1$
    for $b$ from 0.05 to 0.95 step 0.05 do
        for $k_1$ from 0.1 to 1.8 step 0.1 do
            for $k_3$ from 0 to 10 step 1 do
                $D_1, b, k_1, k_3 \xrightarrow{\text{Okapi BM25 (inquiring)}} R_0$
                Calculate the $MAP$ for $R_0, map_r$
                if $mmap < map_r$ then
                    $mmap = map_r$
                    $k_{3o} = k_3$
                    $k_{1o} = k_1$
                    $b_o = b$
                end
            end
        end
    end
    /* Optimisation of the parameter of language model */
    $\lambda_o = 0, mmap = -0.1$
    for $\lambda$ from 0 to 1.0 step 0.1 do
        $D_1, \lambda \xrightarrow{\text{language model (inquiring)}} R_1$
        Calculate the $MAP$ for $R_1, map_r$
        if $mmap < map_r$ then
            $mmap = map_r$
            $\lambda_o = \lambda$
        end
    end
end
```

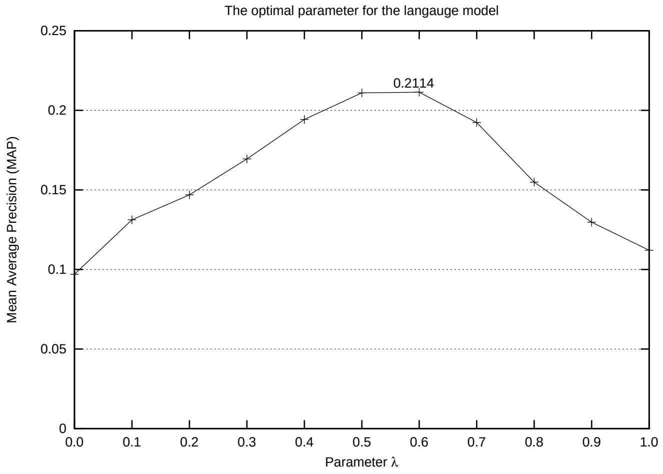


Figure 4.8: Results of parameter optimisation for language model

4.5 Experiments of Evaluating Query Expansion Using eCMeSH

The eCMeSH Tree consists of terms and the relations among them. We assume that they can be employed as a lexical resource to improve the retrieval performance of CLIR. In this section, several experiments have been carried out to evaluate the retrieval performance using the eCMeSH Tree:

Query translation using the CMeSH Tree This experiment studies the contribution of the original CMeSH terms to the retrieval performance, when a query is translated into English using CMeSH terms.

Query expansion using the eCMeSH Tree This experiment investigates the retrieval improvements attained when the eCMeSH Tree is used to expand queries, compared with the method using the CMeSH Tree.

Query expansion using pseudo-relevance feedback Pseudo-relevance feedback (PRF) is employed to expand a query. This experiment compares the performance attained via eCMeSH and PRF.

Document translation Compared with query expansion, document translation is a different approach to CLIR. This experiment is designed to show the difference between the eCMeSH method and the document translation approach.

4.5.1 Query Translation Using the CMeSH Tree

In order to evaluate the improvement attained via the eCMeSH Tree terms, it is necessary to know the retrieval performance using the original CMeSH Tree terms. This section describes the experiment's steps and performance.

4.5.1.1 Steps

In this experiment, documents in the 2006 TREC Genomics test collection are processed following the approaches described in Section 4.2.2.2 and indexed; all queries are processed using the methods described in Section 4.2.2.1. The original CMeSH Tree terms aligned in Section 3.3.2 are used to translate the Chinese query terms into the English equivalents. This translation follows these criteria:

(1) If a Chinese term is found in the CMeSH Tree, then its English MeSH heading term is used to replace this Chinese term. For example, "阿尔茨海默病" has the equivalent English heading term, "Alzheimer Disease". Then "Alzheimer Disease" is selected as the translation of "阿尔茨海默病" to form the English query.

(2) If a term cannot be found in the CMeSH Tree, then the dictionary is used to translate it. All translations listed in the dictionary will be included in the new query.

(3) If different terms have identical translations, then the translated terms are merged.

(4) All untranslatable Chinese terms will be ignored, but English acronyms or abbreviations will be retained in the query.

Algorithm 8 gives the detail of the steps to carry out this experiment.

As an example, Table 4.7 stet presented in Lemur's structured query language, *INQUERY*, illustrates the original Query 161 and the translation using the CMeSH Tree terms. For the purpose of comparing the result with other experiments, the term weight is introduced, calculated as $^1/_N$, where N is the total number of query terms. It is clear that the retrieval performance of queries weighted by this scheme is equivalent to that of queries without term weight.

Algorithm 8: Steps of the experiment using CMeSH Tree terms to translate the query

```
Input  : the indexed document collection ($D_I$), the query set ($Q$)
Output: MAP, AP

begin
    /* $S = \{t_i | t_i \in q_j, q_j \in Q.$ $t_i$ is the $i$th term of $q_j$.$\}$ */
    $Q \xrightarrow{\text{query pre-processing}} S$
    $S \xrightarrow{\text{the CMeSH Tree terms (translating)}} S_1$
    $S_1, D_I \xrightarrow{\text{Indri (inquiring)}} R$
    Evaluate $R$ using AP and MAP measures
end
```

161	the original query	#combine(0.3333 阿尔茨海默病 0.3333 IDE 0.3333 作用)
	the translated query	#combine(0.3333 #1(Alzheimer Disease) 0.3333 IDE 0.3333 #1(act on) 0.3333 affect 0.3333 action 0.3333 function 0.3333 effect)

Table 4.7: An example of a query translated using the CMeSH Tree

4.5.1.2 Results

The results of the experiment are illustrated in Table 4.8, which includes the retrieval performances, measured by *MAP*, of manual and automatic word segmentations under Okapi BM25 and the language model, respectively. The baseline experiments, abbreviated as "Baseline", are compared with these results, referred to as "CMeSH".

	BM25				LM			
	automatic WS		manual WS		automatic WS		manual WS	
	Baseline	CMeSH	Baseline	CMeSH	Baseline	CMeSH	Baseline	CMeSH
2006	**0.2309**	0.1976 (-3.33%)	**0.2622**	0.2503 (-4.54%)	**0.2278**	0.1935 (-15.06%)	**0.2619**	0.2418 (-7.67%)
2007	**0.1353**	0.0911 (-32.67%)	**0.1735**	0.1344 (-22.54%)	**0.1330**	0.0789 (-40.68%)	**0.1695**	0.1154 (-31.92%)

Table 4.8: MAPs for the experiments of query translation using the CMeSH Tree terms

Significance test in this experiment is focused on the difference between the retrieval approach causing the best performance and others. The test result implies that they have no significant difference.

Figure 4.9 compares the retrieval performance of each query in the experiments using the CMeSH Tree terms to translate queries with that of the baseline, from the point of view of average precision (AP): Figure 4.9(a) and Figure 4.9(b) are the results for the TREC 2006 Genomics Track using different retrieval models; Figure 4.9(c) and Figure 4.9(d) show the experimental results of the 2007 Track under the Okapi BM25 and language models.

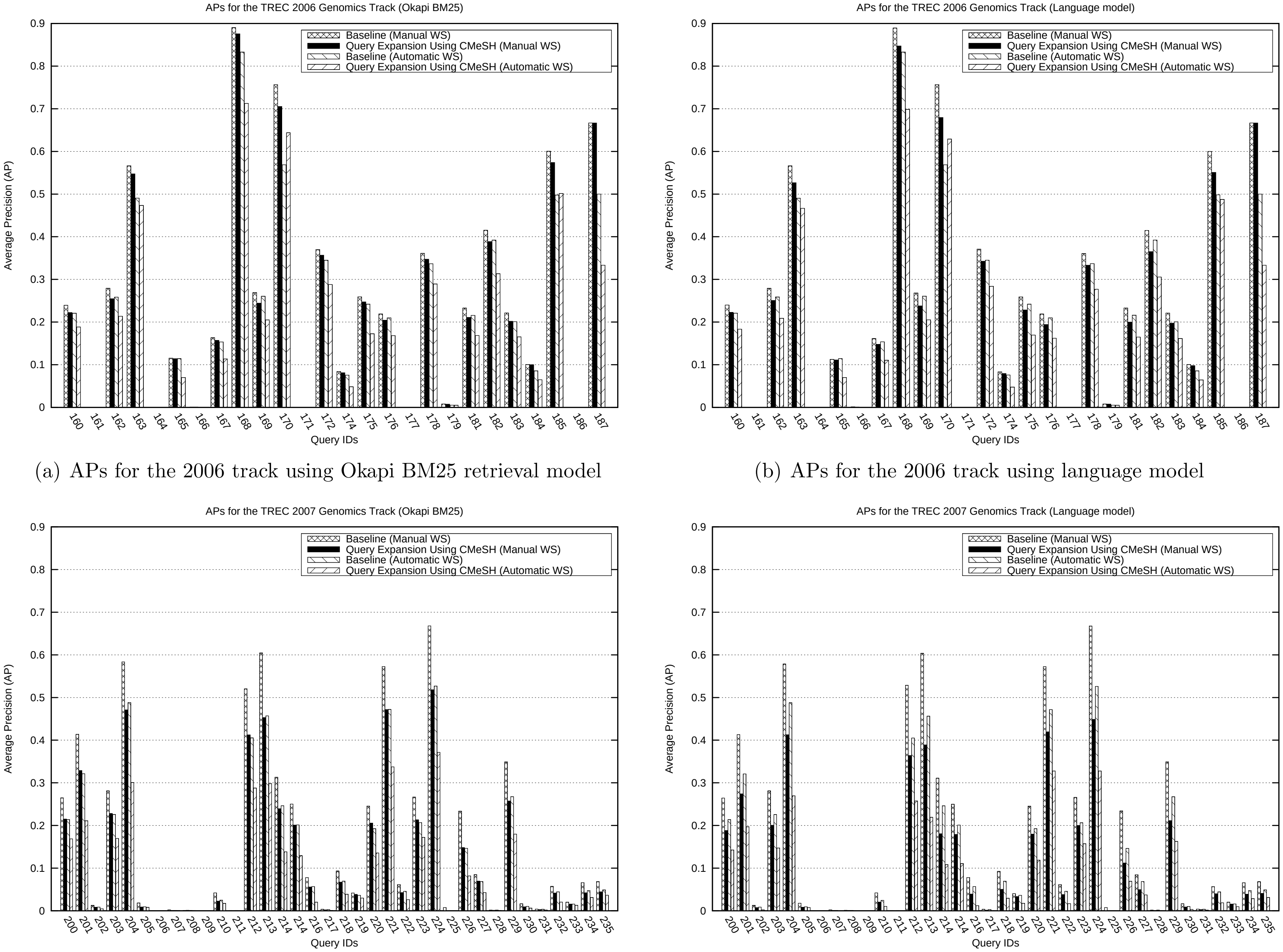
4.5.1.3 Discussion

From Table 4.8, it is clear that the experiments applying the original CMeSH Tree terms to translate queries perform worse than the baseline experiment for both retrieval models and both word segmentation strategies. Compared with the baseline experiment's results, for example, the MAP of the 2006 Track using manual word segmentation under Okapi BM25 is 0.2503, falling by 4.54%; and the retrieval performance of the corresponding 2007 task falls by 22.54%. The reason for this decrease is that the original CMeSH has a poor coverage of the biomedical terms. In the CMeSH Tree, each English heading term has only one Chinese counterpart or no Chinese term. When the CMeSH Tree terms are employed to translate the Chinese queries, the unmatched Chinese query terms are ignored, which leads to the drop in performance. Figure 4.9 shows the drop in performance from the point of view of each single query.

It is also noticed from Table 4.8 that the level of performance drop in the 2007 Track is greater than that in the 2006 Track, averaging 24.45% in 2007's queries, 7.65% in 2006's only. This difference reflects the fact that the 2006 queries are structurally different from those of 2007.

4.5.2 Experiment of Query Expansion Using the eCMeSH Tree

The experiments described in this section are designed to evaluate the retrieval improvements attained after using the eCMeSH Tree terms to expand the original queries. Experiments with and without term weights are conducted, but more attention is paid to the experiments using term weights.



(a) APs for the 2006 track using Okapi BM25 retrieval model

(b) APs for the 2006 track using language model

(c) APs for the 2007 track using Okapi BM25 retrieval model

(d) APs for the 2007 track using language model

Figure 4.9: Results of query translation using the CMeSH Tree terms measured by AP

4.5.2.1 Steps

Before any retrieval, all documents in the 2006 TREC Genomics track test collection are processed using the pre-processing mentioned in Section 4.2.2.2; and queries from the 2006 and 2007 tracks are processed following the methods described in Section 4.2.2.1. The eCMeSH Tree terms constructed from the algorithms in Section 3.3 are used to expand and translate the Chinese query terms. This expansion is performed as follows:

(1) If a Chinese query term is found in the eCMeSH Tree, any siblings (Definition 5) and child nodes (Definition 6) are sorted into a list according to their term weights. Only the top 20 terms from the list are added to the query along with their term weights. For instance, when "阿尔茨海默病" is expanded, according to eCMeSH, the extra terms added to the original terms are: "AD症" (0.97597), "AD病" (0.97597), "Alzheimer氏病" (0.86363), "阿尔茨海默症" (0.82046), "早老性痴呆" (0.72262), "年性痴呆症" (0.60395), "老年性痴呆" (0.60395), "阿滋海默症" (0.1862), etc.

(2) Terms which are not found in the eCMeSH Tree are ignored, except for those without Chinese characters (e.g. acronyms), which are retained in the query, given their likelihood of representing terms. For example, "IDE" is not listed in the eCMeSH Tree terms, but it will be retained because it is highly indicative of a term.

(3) Query terms without a weight (e.g. acronyms) will be assigned a term weight of $^1/_N$, where N is the total number of query terms (after processing and before expanding) in a certain query. In Query 161, "IDE" is such an example: it has no term weight due to its absence in the eCMeSH; but the weight 0.3333 is assigned to "IDE", because after word segmentation and term filtering Query 161 consists of three terms: "阿尔茨海默病", "IDE", and "作用".

After expansion, the queries are translated into the English equivalents using the following criteria:

(1) If a Chinese term is found in the eCMeSH Tree, then its English counterparts, including both MeSH heading terms and entry terms, are used to replace this Chinese term. Each of these English terms inherits the term weight from the Chinese term. In the above example, "早老性痴呆"

is found in the eCMeSH Tree node C10.228.140.380.100, so the heading term, “Alzheimer Disease”, and all entry terms, “Acute Confusional Senile Dementia”, “Alzheimer Disease, Early Onset”, “Alzheimer Disease, Late Onset”, “Alzheimer Type Senile Dementia”, “Alzheimer’s Disease, Focal Onset”, etc. are selected as the translation. Moreover, each of these English terms have the same term weight, 0.72262, which is derived from the original Chinese terms.

(2) If a term cannot be found in the eCMeSH Tree, then the dictionary is used to translate it. All translations listed in the dictionary will be included in the new query; each translation uses the term weight of the original term. The term “作用” is translated, by the Google and Kingsoft Dictionary, into the English equivalents, “act on”, “affect”, “action”, “function”, and “effect”; and each is assigned the term weight 0.3333.

(3) If different terms have identical translations, then the translated terms are merged. The new term weight is the maximum weight amongst the duplicates. The term “阿尔茨海默病” and “AD症” share the same translations, for example; this rule will merge the same English terms and re-assign the term weight as 0.97597, which is derived from “AD症”.

(4) All untranslatable terms will be ignored.

Algorithm 9 summarises the steps to carry out the experiments of query expansion based on the eCMeSH Tree.

Table 4.7 shows the original Query 161, the expanded query using the eCMeSH Tree, and the translated result.

The experiment using the above-mentioned rules and criteria to expand queries, except for term weight assignment, is also conducted, aimed at investigating the contribution of a term weight.

4.5.2.2 Results

The results of query expansions using the eCMeSH Tree with and without term weights are illustrated in Table 4.10,compared with the baseline experiment, abbreviated as “Baseline”. “eCMeSH-W” refers to the experiments which include term weights; and “eCMeSH-N” represents the experiments that do not apply term weighting.

Algorithm 9: Steps of the experiment using eCMeSH Tree terms to expand and translate queries

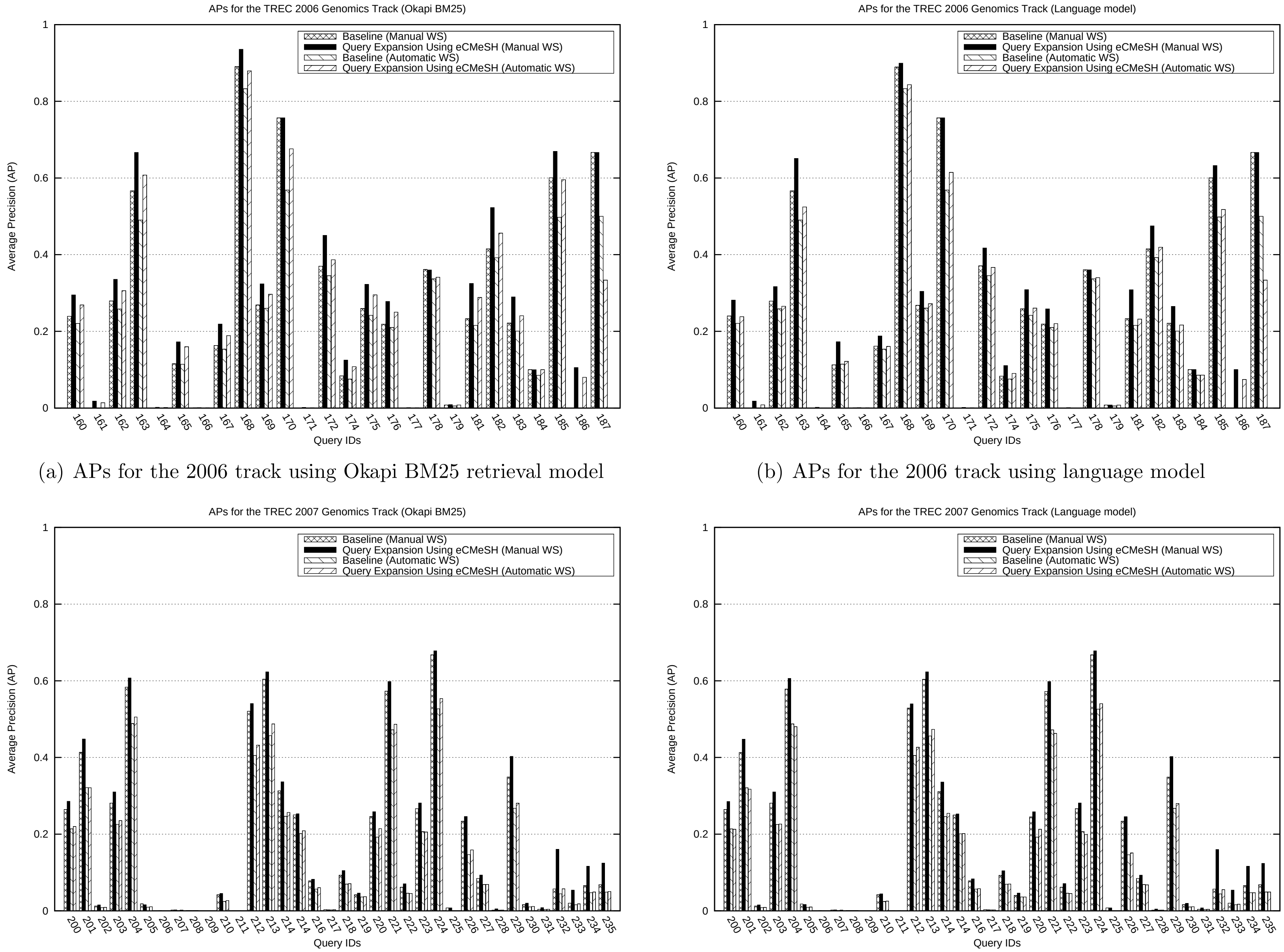
```
Input  : the indexed document collection ($D_I$), the query set ($Q$)
Output: $MAP$, $AP$

begin
    /* $S = \{t_i | t_i \in q_j, q_j \in Q.$ $t_i$ is the $i$th term of $q_j.\}$ */
    $Q \xrightarrow{\text{query pre-processing}} S$
    $S \xrightarrow{\text{the eCMeSH Tree terms (expanding)}} S_1$
    $S \xrightarrow{\text{the eCMeSH Tree terms and the Dictionary (translating)}} S_2$
    $S_2, D_I \xrightarrow{\text{Indri (inquiring)}} R$
    Evaluate $R$ using $AP$ and $MAP$ measures
end
```

Significance tests for these experiments are carried out between any two retrieval methods using the same query set. Calculations shows that there is no significant difference among these methods.

The AP measure of each single query is illustrated in Figure 4.10: Figure 4.10(a) focuses on the retrieval performance of the 2006 Track using the Okapi BM25 model; and Figure 4.10(b) shows the average precisions of queries in the 2006 Track under the language model. Figure 4.10(c) and Figure 4.10(d) compare the experimental results of the 2007 Track using different retrieval models. However, the results of the experiment that applies no term weighting scheme are not represented in this figure.

4.5.2.3 Discussion

The experimental results prove that the eCMeSH Tree terms can effectively improve the retrieval performance. Table 4.10 compares the retrieval performance of query expansion with the baseline. The best performance of query expansion based on the eCMeSH Tree using Okapi BM25 achieves 0.3058 for the 2006 Track and 0.1901 for the 2007 Track, respectively, when automatic and manual word segmentation are applied; compared with the baseline, they are increased by 16.63% and 9.57%, respectively. For the language model, the performance drops slightly. For example, query expansion based on the eCMeSH Tree terms with term weights using the language model and manual word segmentation is 0.2925, while using Okapi BM25, the result is 0.3058. The other experiments show the



(a) APs for the 2006 track using Okapi BM25 retrieval model

(b) APs for the 2006 track using language model

(c) APs for the 2007 track using Okapi BM25 retrieval model

(d) APs for the 2007 track using language model

Figure 4.10: Results of query expansion using the eCMeSH Tree terms measured by AP

161	the original query	#combine(0.3333 阿尔茨海默病 0.3333 IDE 0.3333 作用)
	the expanded query	#combine(#syn(0.9760 AD症 0.9760 AD病 0.8636 Alzheimer氏病 0.8205 阿尔茨海默症 0.7226 早老性痴呆 0.6040 年性痴呆症 0.6040 老年性痴呆 0.2446 阿尔茨海默病 0.1862 阿滋海默症 ... 0.0000 #1(失语, 原发进行性) 0.0148 克-亚综合征 0.2264 #1(痴呆, 血管性) ...) 0.3333 IDE 0.3333 作用)
	the translated query	#combine(#syn(0.9760 #1(Alzheimer Disease) 0.9760 #1(Acute Confusional Senile Dementia) 0.9760 #1(Alzheimer Disease, Early Onset) 0.9760 #1(Alzheimer Disease, Late Onset) 0.9760 #1(Alzheimer Type Senile Dementia) 0.9760 #1(Alzheimer's Disease, Focal Onset) 0.2446 #1(Dementia, Alzheimer Type) 0.2446 #1(Early Onset Alzheimer Disease) ... 0.0000 #1(Aphasia, Primary Progressive) 0.0148 #1(Creutzfeldt-Jakob Syndrome) 0.2264 #1(Dementia, Vascular) ...) 0.3333 IDE 0.3333 #1(act on) 0.3333 affect 0.3333 action 0.3333 function 0.3333 effect)

Table 4.9: An example of a query expanded using the eCMeSH Tree

same trend. Figure 4.10 reflects how the eCMeSH Tree terms produce effects on each single query: Query 181's AP attains 0.3248 after applying manual word segmentation, the Okapi BM25 model and term weight; while for the baseline, it is only 0.2327. The improved AP is increased by 39.58%. Therefore, in the following experiments, when it comes to the experimental results related to the eCMeSH Tree, the results are presented by the experiments using term weights.

It is clear that the term weight improves the retrieval performance to some degree. For instance, the 2007 Track achieves 0.1813 when Okapi BM25 and manual word segmentation are used without term weights; while the result reaches 0.1901, increased by 4.85%, when term weights are included. The other experiments also illustrate performance improvements when term weights are used.

As in the experiments of query translation using the CMeSH Tree terms (described in Section 4.5.1.3), the degree of performance increment for the 2007 Track is slightly lower than that of the 2006 Track.

	BM25						LM					
	automatic WS			manual WS			automatic WS			manual WS		
	Baseline	eCMeSH-W	eCMeSH-N	Baseline	eCMeSH-W	eCMeSH-N	Baseline	eCMeSH-W	eCMeSH-N	Baseline	eCMeSH-W	eCMeSH-N
2006	0.2309	**0.2647**	0.2503	0.2622	**0.3058**	0.2857	0.2278	0.2390	**0.2497**	0.2619	**0.2925**	0.2842
		(14.64%)	(8.40%)		(16.63%)	(8.96%)		(4.92%)	(9.61%)		(11.68%)	(8.51%)
2007	0.1353	0.1415	**0.1435**	0.1735	**0.1901**	0.1813	0.1330	**0.1375**	0.1341	0.1695	**0.1899**	0.1799
		(4.58%)	(6.06%)		(9.57%)	(4.50%)		(3.38%)	(0.83%)		(12.04%)	(6.14%)

Table 4.10: MAPs for the experiments of query expansion using the eCMeSH Tree

4.5.3 Experiment of Query Expansion Using Pseudo-Relevance Feedback

Pseudo-relevance feedback (PRF), which is described in Section 2.1.4.1, provides the automatic approach to analysing the relevant documents returned from the initial search. Queries can be expanded using PRF, which requires no external linguistic resources such as bilingual dictionaries, lexicons or ontologies. In this section, we describe the steps and results of the experiments and discuss the results.

4.5.3.1 Steps

Before the experiments, we process all documents in the 2006 TREC Genomics test collection using the methods described in Section 4.2.2.2; all the queries of the 2006 and 2007 tracks are processed following the methods discussed in Section 4.2.2.1. Then the pseudo-relevant feedback built in to Indri is applied to expand the translated English query terms. This expansion is carried out as follows:

(1) The Dictionary is used to translate the Chinese query terms into English using a term-by-term translation policy. If a term has more than one translation, only the first translation is selected. For example, “阿尔茨海默病” has two translations: “alzheimer disease” and “AD”; and only “alzheimer disease” is selected as the translation. Terms not found in the Dictionary are always omitted, but terms without Chinese characters are retained. “IDE”, for instance, will be retained in the query.

(2) The initial term weight is assigned as $^1/_N$, where N is the total number of query terms (after processing and before expanding) in a certain query. Following this rule, the terms “阿尔茨海默病”, “作用” and “IDE” in Query 161 are all assigned the weight 0.3333.

(3) The initial queries are supplied to the Indri search engine; additional query terms are obtained using Indri’s pseudo-relevance feedback method.

(4) New query terms obtained in step 3 are merged with the initial query to construct an expanded query; this expanded query is submitted to the system again to search for relevant documents.

The steps to carry out the experiments are described in Algorithm 10.

Algorithm 10: Steps of the experiment using eCMeSH Tree terms to translate query

```
Input  : the indexed document collection ($D_I$), the query set ($Q$)
Output: $MAP$, $AP$

begin
    /* $S = \{t_i | t_i \in q_j, q_j \in Q$. $t_i$ is the $i$th term of $q_j$.$\}$    */
    $Q \xrightarrow{\text{query pre-processing}} S$
    $S \xrightarrow{\text{the Dictionary (translating)}} S_1$
    $S_2 = \emptyset$
    for $s_i \in S_1$ do
        $s_i \xrightarrow{\text{Indri (inquiring)}} R_1$
        Select the top 25 ranked terms $T$ from $R_1$
        $S_2 = S_2 \cup T$
    end
    $S_2, D_I \xrightarrow{\text{Indri (inquiring)}} R$
    Evaluate $R$ using $AP$ and $MAP$ measures
end
```

We select the top 50 documents returned by Indri at the initial retrieval as those which are most likely to be relevant to the original query, and the top 25 ranked terms extracted from these documents as those which are most likely to correspond to relevant new query terms to be added to the original query. The weights used to adjust original query terms and the terms resulting from the application of the relevance feedback method are both 0.5 and 0.5. Other parameters of Indri's pseudo-relevance feedback are configured using their default values. Table 4.11 gives a pseudo-relevance feedback example for Query 161.

4.5.3.2 Results

Table 4.12 illustrates the results of query expansion using pseudo-relevance feedback, abbreviated to "PRF", which are compared with the results of the baseline experiment, represented as "Baseline".

In this experiment, the retrieval approach to the best performance, i.e. query expansion using PRF with manual word segmentation under Okapi BM25, is set as reference; other methods are compared with it. The results of significance tests show that these methods have no significant difference.

161	the Chinese query	#combine(0.3333 阿尔茨海默病 0.3333 IDE 0.3333 作用)
	the translated query	#combine(0.3333 #1(alzheimer's disease) 0.3333 IDE 0.3333 action)
	the expanded query	#weight(0.5 #combine(0.3333 #1(alzheimer's disease) 0.3333 IDE 0.3333 action) 0.5 #weight(0.7421 Alzheimer 0.7019 illness 0.6504 disease 0.5723 function 0.4710 affected 0.3241 damages 0.3200 treatment 0.2988 brain ...))

Table 4.11: An example of a query expanded using pseudo-relevance feedback

	BM25				LM			
	automatic WS		manual WS		automatic WS		manual WS	
	Baseline	PRF	Baseline	PRF	Baseline	PRF	Baseline	PRF
2006	0.2309	**0.2737**	0.2622	**0.3009**	0.2278	**0.2765**	0.2619	**0.3178**
		(15.03%)		(13.04%)		(21.38%)		(21.34%)
2007	0.1353	**0.1654**	0.1735	**0.2154**	0.1330	**0.1591**	0.1695	**0.2123**
		(22.25%)		(24.15%)		(19.62%)		(12.04%)

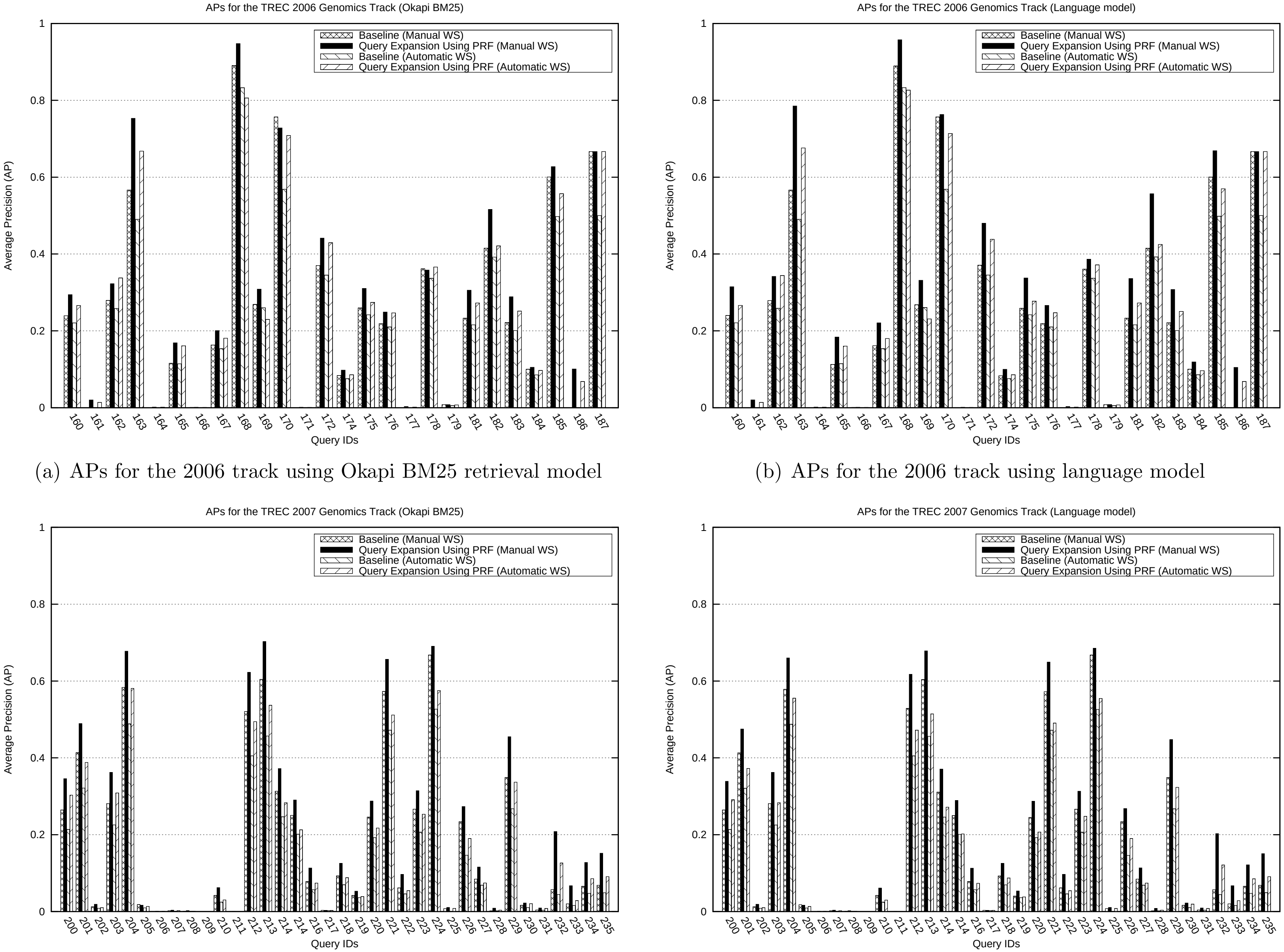
Table 4.12: MAPs for the experiments of query expansion using pseudo-relevance feedback

Figure 4.11 measures the retrieval performance for each query using average precision. The difference existing in the 2006 Track between Okapi BM25 and the language model is illustrated in Figure 4.11(a) and Figure 4.11(b). Figure 4.11(c) and Figure 4.11(d) compare the experimental results of the 2007 Track using both retrieval models.

4.5.3.3 Discussion

There is no doubt that the query expansion using pseudo-relevance feedback is able to improve the retrieval performance greatly. In Table 4.12, for example, the best retrieval for the 2006 Track is 0.3178, when the language model and manual word segmentation are applied. Compared with the baseline result, this result is increased by 21.34%. The 2007 Track achieves the best performance of 0.2154 under Okapi BM25 using manually separated queries. The detail of retrieval performance for each query is shown in Figure 4.11, where pseudo-relevance feedback improves the retrieval performance for most of the queries.

Unlike the experiments using the CMeSH Tree terms (described in Section 4.5.1)



(a) APs for the 2006 track using Okapi BM25 retrieval model

(b) APs for the 2006 track using language model

(c) APs for the 2007 track using Okapi BM25 retrieval model

(d) APs for the 2007 track using language model

Figure 4.11: Results of query expansion using pseudo-relevance feedback measured by AP

and applying the eCMeSH Tree terms (discussed in Section 4.5.2), where the best performance for both models and both years are attained under Okapi BM25, the 2006 Track reaches its best performance when the language model is employed; and the 2007 Track achieves the optimal result under the language model. A possible reason for such results is that the optimal parameters for Okapi BM25 described in Section 4.4 are not optimal for pseudo-relevance feedback.

4.5.4 Experiment on Document Translation

Document translation (investigated in Section 2.2.2.2) is another approach to cross-lingual information retrieval. This method is based on the idea that the document collection is translated into the target language before searching for content; it converts the cross-language context into the monolingual environment. In our experiments, the Google translation service [8] is applied to translate the TREC 2006 Genomics Track document collection into Chinese before retrieval.

4.5.4.1 Steps

The first step in the experiments is to process the queries of the 2006 and 2007 Tracks using the methods described in Section 4.2.2.1. Then all the documents in the 2006 Track test collection are translated into Chinese equivalents. Table 4.13 illustrates an example: a fragment of an English document, coming from Document 11574508, is translated into Chinese using the Google translation service. After translation, retrieval is performed using segmented Chinese queries. During the experiments, no dictionaries or ontologies are used to expand the Chinese queries; moreover, no term weight is assigned to terms. Algorithm 11 shows the steps in conducting these experiments.

The translated Chinese documents are indexed using the bigram procedure, because this indexing strategy avoids the incorrectness introduced by word segmentation tools and retains all the information in a document. Queries, therefore, are separated using the same bigram algorithm. Table 4.14 is the example of Query 161, which is segmented as bigrams.

[8] `http://translate.google.com/`

All biopsies were performed under local anaesthesia with a pre-operative injection of 0.5 mg alfentanilum (Rapifen®; Janssen-Cilag, Birkerod, Denmark) and 2.5 mg midazolam (Dormicum®; Roche, Basel, Switzerland). The scrotal area was washed with 0.5% chlorhexidine solution followed by physiological saline and then draped. Seven millilitres of prilocaine hydrochloride (Citanest®10 mg/ml, Astra; Södertälje, Sweden) was injected around the vas deferens as described by Li et al. (1992), using a 22 gauge, 5 cm needle (Microlance 3®, 22G 0.7x50 mm; Becton Dickinson, Dublin, Ireland). Additionally, 1 ml prilocaine hydrochloride was injected into the scrotal skin. The testis was grasped between the thumb and forefinger of the non-dominant hand and rotated ventrally to prevent epididymal injury. The needle was inserted into the cranial pole towards the centre of the testicle. Two biopsies were taken close to each other, first with a 14 gauge (n = 45) and then with a 16 gauge (n = 44) needle (Bard MAGNUM Biopsy Instrument, C.R.Bard Inc., Covington, GA, USA), both with a 19 mm notch. Three quarters of the testicular material was fixed and plastic-embedded for histopathological assessment, and one quarter reserved for direct microscopy. Evaluation of testicular biopsy specimens was as described by Rosenlund et al. (Rosenlund et al., 1998Go).
所有的活组织切片检查是在局部麻醉下进行手术前注射了0.5毫克alfentanilum（Rapifen®;扬森 - Cilag，Birkerod，丹麦）和2.5毫克咪达唑仑（Dormicum®;，巴塞尔，瑞士罗氏公司）。阴囊面积洗净，用0.5%洗必泰，生理盐水解决方案，然后搭着。七毫升丙胺卡因盐酸盐（Citanest®10毫克/毫升，阿斯特拉南泰利耶，瑞典）是围绕注入输精管，输精管，李等。（1992年），使用22号，5厘米的针（Microlance3®，22G0.7x50毫米; Becton Dickinson公司，都柏林，爱尔兰）。此外，1毫升丙胺卡因盐酸注入阴囊皮肤。睾丸掌握非惯用手的拇指和食指之间和旋转的腹部，以防止附睾损伤。针插入颅极对睾丸中心。两个活检接近对方，先用14计（N = 45），然后用16号（N= 44）针（巴德MAGNUM活检仪器，CRBard公司，科文顿，GA，美国），既一个19毫米的缺口。四分之三的睾丸材料是固定的，塑料包埋病理组织学评估，一季度直接镜检保留。睾丸活检标本的评价是Rosenlund等。（Rosenlund等。1998Go）。

Table 4.13: An example of document translation

4.5.4.2 Results

The results of the document translation experiment, abbreviated to "DT", are compared with the baseline results in Table 4.15.

In this experiment, the retrieval methods of automatic word segmentation in the baseline experiment, which are marked with "*", are significatively different from the best retrieval approaches of document translation. This implies that document translation performs much better than the baseline approach.

Figure 4.12 gives the details of the APs of each single query in both baseline experiment and the document translation approach. The average precision for the TREC 2006 Track's queries are illustrated in Figure 4.11(a) and Figure 4.11(b); the 2007's performance for a single query is compared in Figure 4.11(c) and

Algorithm 11: The steps of the experiment using eCMeSH Tree terms to translate query

```
Input  : the document collection ($D$), the query set ($Q$)
Output: $MAP$, $AP$

begin
    /* $S = \{t_i | t_i \in q_j, q_j \in Q$. $t_i$ is the $i$th term of $q_j$.$\}$ */
    $Q \xrightarrow{\text{query pre-processing}} S$
    $D \xrightarrow{\text{Google tanslation (translating into Chinese)}} D_1$
    $D_1 \xrightarrow{\text{bigram (splitting)}} D_2$
    $D_2 \xrightarrow{\text{Indri (indexing)}} D_3$
    $S \xrightarrow{\text{bigram (splitting)}} S_1$
    $S_1, D_3 \xrightarrow{\text{Indri (inquiring)}} R$
    Evaluate $R$ using $AP$ and $MAP$ measures
end
```

161	the original query	#combine(0.3333 阿尔茨海默病 0.3333 IDE 0.3333 作用)
	the query as bigrams	#combine(0.3333 #3(阿尔 尔茨 茨海 海默 默病) 0.3333 #3(ID DE) 0.3333 作用)

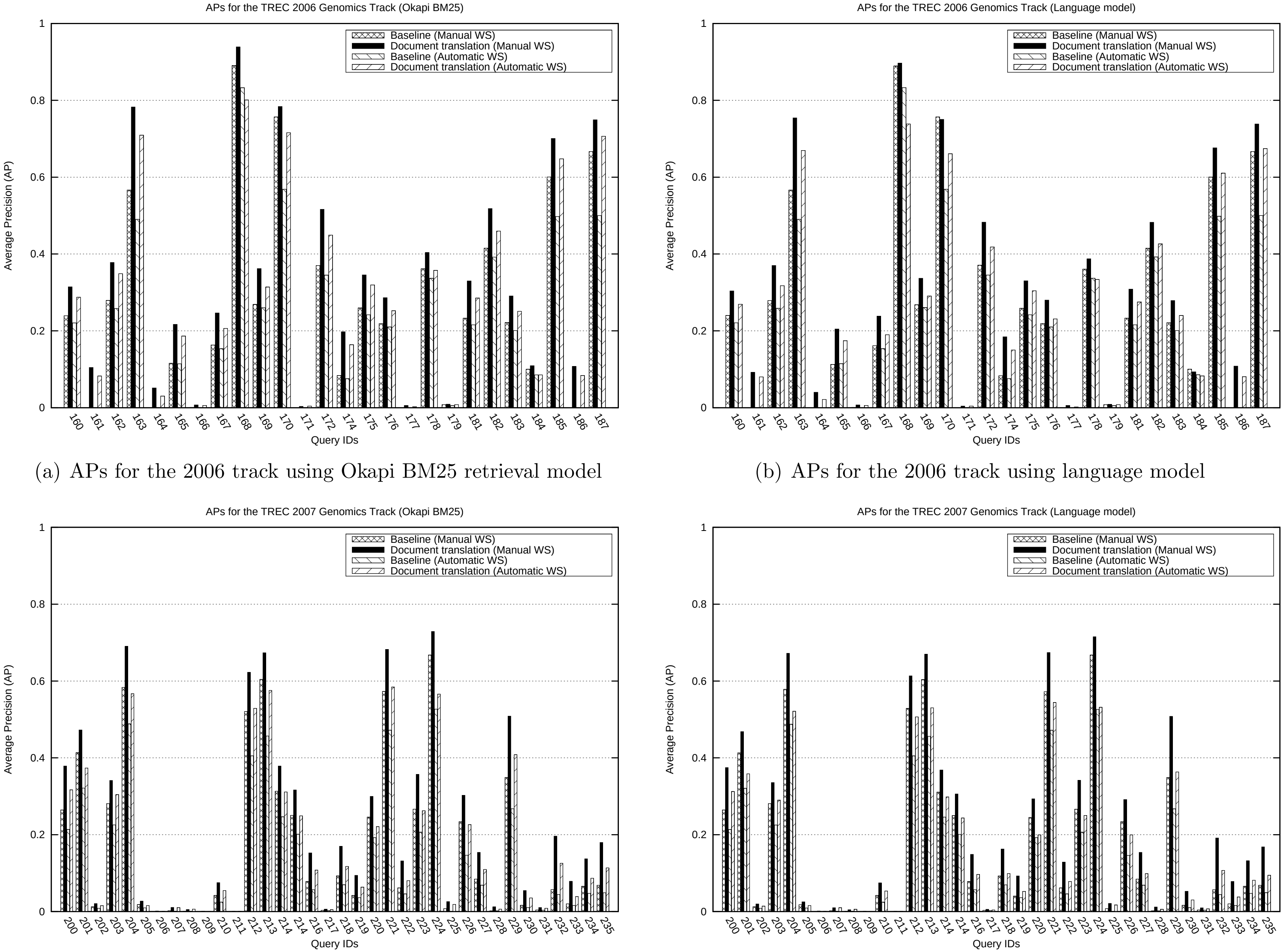
Table 4.14: An example of a query for document translation

Figure 4.11(d).

4.5.4.3 Discussion

The results illustrated in Table 4.15 suggest that the document translation approach to CLIR greatly improves the retrieval performance. The average degree of improvement is 28.58%. Of the results, the best retrieval of the 2006 and 2007 tracks are 0.3368 and 0.2305, attained when Okapi BM25 is applied with the manually separated queries. Compared to the baseline results, the best performances are increased by 28.54% and 32.85%, respectively. Unlike dictionary-based translation methods, which suffer from out-of-vocabulary terms, document translation can at least produce a translation for each term, although it may be inappropriate. This improves retrieval performance.

It is also noticed that the degree of increment in retrieval performance for the 2007 Track exceeds that of the 2006 Track. For example, when the automatic



(a) APs for the 2006 track using Okapi BM25 retrieval model

(b) APs for the 2006 track using language model

(c) APs for the 2007 track using Okapi BM25 retrieval model

(d) APs for the 2007 track using language model

Figure 4.12: Results of document translation measured by AP

	BM25				LM			
	automatic WS		manual WS		automatic WS		manual WS	
	Baseline	DT	Baseline	DT	Baseline	DT	Baseline	DT
2006	0.2309*	**0.2985**	0.2622	**0.3368**	0.2278*	**0.2791**	0.2619	**0.3216**
		(29.27%)		(28.45%)		(22.52%)		(22.79%)
2007	0.1353*	**0.1800**	0.1735	**0.2305**	0.1330*	**0.1683**	0.1695	**0.2257**
		(33.04%)		(32.85%)		(26.54%)		(33.16%)

Table 4.15: MAPs for the experiments in document translation

word segmentation and the language model are applied on queries, the 2007 Track achieves 0.1683, increased by 26.54%, while the 2006's result is 0.2791, increased by 22.52%. Compared with the experiments based on the eCMeSH Tree (described in Section 4.5.2) and the experiments using pseudo-relevance feedback (discussed in Section 4.5.3), where the improvement in the 2007 Track is less than that of 2006, document translation is a more effective approach to information needs for the 2007 queries than the dictionary-based and relevance feedback methods.

4.5.5 Comparison and Analysis

The retrieval performance of the above experiments has been compared with that of the baseline, but a comparison between each individual experiment was not carried out. In this section, we compare the experimental results of the four individual experiments.

Table 4.16 illustrates the retrieval performance of the experiments on query expansion using the CMeSH Tree, represented as "CMeSH", the experiments on query expansion using the eCMeSH, abbreviated to "eCMeSH", query expansion using pseudo-relevance feedback, marked "PRF", and document translation method, referred to as "DT". The tags "D1" and "D2" are used in the discussion described in Section 5.4.

Considering the approaches to query expansion based on a dictionary, i.e. the expansions using the eCMeSH Tree and the CMeSH Tree, the CLIR performance decreases significantly from eCMeSH's 0.3058 to CMeSH's 0.2503, for the TREC 2006 Track; and, for the 2007 Track, from 0.1901 to 0.1344. When document translation is used, the CLIR reaches the best performances of 0.33368 and 0.2305 for the 2006 and 2007 tracks, respectively. Compared with the experiments which exploit the eCMeSH Tree to expand queries, the method that applies pseudo-relevance feedback to expand queries achieves performances of 0.3178, for the 2006 Track, and 0.2154, for the 2007 Track, which are increases of 8.65% and 13.31%. The document translation method provides more improvements on retrieval performance than both the eCMeSH Tree and pseudo-relevance feedback approaches. For instance, when Okapi BM25 and manual word segmentation are employed, the retrieval performance of the 2007 Track is increased by 21.25% over the eCMeSH Tree, and 7.01% over pseudo-relevance feedback.

The experimental results suggest the following conclusions:

(1) For the query expansion approaches based on dictionaries, our query expansion using the eCMeSH Tree can greatly improve CLIR performance. eCMeSH provides more synonyms for an expanded query, which leads to the improvement.

(2) Query expansion using pseudo-relevance feedback gives more improvements on CLIR performance than query expansion using the eCMeSH Tree does. Pseudo-relevance feedback analyses the relevant documents returned initially, and adds the top-ranked terms to the original query; this is an effective way to form a new query.

(3) The document translation approach exceeds all the query expansion methods. The machine translation software applied to translate the document collection into Chinese gives English biomedical terms an equivalent Chinese translation, which, to some degree, resolves the out-of-vocabulary issue.

However, pseudo-relevance feedback and document translation have their own problems. Although pseudo-relevance feedback is independent of language, when applied to expand queries in CLIR, it indeed requires the translation of either queries or the document collection. In our experiment, queries have been translated into Chinese before the use of pseudo-relevance feedback.

The experiment (PRF) described in Section 4.5.3 utilises the Google and Kingsoft Dictionary to translate the Chinese queries into English equivalents. We also conduct two experiments which employ the eCMeSH Tree terms and the CMeSH Tree terms to translate queries. The translation is carried out using a term-by-term translation policy. If a Chinese term is found in eCMeSH or CMeSH, then its corresponding English heading term is treated as the translation. The untranslatable Chinese terms are ignored, but words containing Latin symbols are retained in the translated query. Other experimental settings are the same as in Section 4.5.3. Table 4.17 illustrates the results. "PRF-D" refers to the experiment using the Dictionary to translate the Chinese queries; "PRF-e" denotes the experiment where the eCMeSH Tree terms are applied to translate queries; and "PRF-C" is the abbreviation for the experiment in which the translation is finished by the CMeSH Tree terms.

From Table 4.17, it is clear that the retrieval performance of query expansion using pseudo-relevance feedback depends on the quality of the linguistic resources that are used to translate queries. For instance, for the 2006 Track, the best performance of translation using the Dictionary is 0.3009. However, when the eCMeSH Tree terms are applied to translate Chinese queries into English, the retrieval performance becomes 0.2771, decreased by 7.91%. The retrieval performance drops by 15.39% when the Dictionary is substituted by the CMeSH tree.

The major problem of the document translation approach lies in that it is computationally expensive. In our experiment, the translation of the TREC 2006 Genomics document collection takes about five months on a desktop computer with a 1.44 GHz Intel Dual Core CPU, 3 GB memory, and 250 GB hard disk. Moreover, the translated documents have to be re-indexed before any search. Therefore the document translation approach is not suitable for cross-lingual information retrieval where documents are added or removed frequently, or the contents of documents vary rapidly.

4.6 Summary

In this chapter, we evaluate the effectiveness of the eCMeSH Tree when it is used to expand queries in Chinese-English biomedical cross-lingual information retrieval. Compared to other dictionary-based approaches, the results show that the query expansion based on the eCMeSH Tree is an effective approach to CLIR. Although the query expansion using pseudo-relevance feedback method and document translation approach, which are different mechanisms to implement CLIR, lead to better retrieval performance, they suffer from drawbacks: i) Pseudo-relevance feedback is independent on language, but when it is used to implement

CLIR, the retrieval performance depends on the quality of the query translation. ii) Document translation depends on machine translation software, which is a computationally expensive approach, and is not suitable for the cases where documents are added or removed frequently, or documents' contents change rapidly.

We compare Okapi BM25, a probabilistic retrieval model, with the query likelihood language model using Jelinek-Mercer smoothing. The experimental results illustrate that Okapi BM25 performs better than the language model.

We also optimise the parameters for both retrieval models. For Okapi BM25, the optimal parameters are $k_1 = 1.4$, $k_3 = 6$, and $b = 0.75$; for the language model, the optimal parameter is $\lambda = 0.6$.

In the next chapter, the experiments combining query expansion based on the eCMeSH Tree, query expansion using pseudo-relevance feedback, and document translation method are conducted to evaluate the retrieval performance improvements attained by a hybrid approach.

	BM25								LM							
	automatic WS				manual WS				automatic WS				manual WS			
	eCMeSH	CMeSH	PRF	DT	eCMeSH	CMeSH	PRF	DT	eCMeSH	CMeSH	PRF	DT	eCMeSH	CMeSH	PRF	DT
2006	0.2647	0.1976	0.2737	**0.2985**	0.3058	0.2503	0.3009	**0.3368**	0.2390	0.1935	0.2765	**0.2791**	0.2925	0.2418	0.3178	**0.3216**
		(-25.35%)	(3.40%)	(12.77%)$^{\mathrm{D2}}$		(-18.15%)	(-1.60%)	(10.14%)$^{\mathrm{D2}}$		(-19.04%)	(15.69%)	(16.78%)$^{\mathrm{D2}}$		(-17.33%)	(8.65%)	(9.95%)$^{\mathrm{D2}}$
				(9.06%)$^{\mathrm{D1}}$				(11.93%)$^{\mathrm{D1}}$				(0.94%)$^{\mathrm{D1}}$				(1.20%)$^{\mathrm{D1}}$
2007	0.1415	0.0911	0.1654	**0.1800**	0.1901	0.1344	0.2154	**0.2305**	0.1375	0.0789	0.1591	**0.1683**	0.1899	0.1154	0.2123	**0.2257**
		(-35.62%)	(16.47%)	(27.21%)$^{\mathrm{D2}}$		(-23.30%)	(13.31%)	(21.25%)$^{\mathrm{D2}}$		(-42.62%)	(15.71%)	(22.40%)$^{\mathrm{D2}}$		(-39.23%)	(11.80%)	(18.85%)$^{\mathrm{D2}}$
				(8.83%)$^{\mathrm{D1}}$				(7.01%)$^{\mathrm{D1}}$				(5.78%)$^{\mathrm{D1}}$				(6.31%)$^{\mathrm{D1}}$

Table 4.16: Comparisons of experiments using the CMeSH Tree, the eCMeSH Tree, pseudo-relevance feedback and document translation

	BM25						LM					
	automatic WS			manual WS			automatic WS			manual WS		
	PRF-D	PRF-e	PRF-C	PRF-D	PRF-e	PRF-C	PRF-D	PRF-e	PRF-C	PRF-D	PRF-e	PRF-C
2006	**0.2737**	0.2390 (-12.68%)	0.2205 (-19.44%)	**0.3009**	0.2771 (-7.91%)	0.2546 (-15.39%)	**0.2765**	0.2379 (-13.96%)	0.2193 (-20.69%)	**0.3178**	0.2763 (-13.06%)	0.2539 (-20.11%)
2007	**0.1654**	0.1275 (-22.91%)	0.1085 (-34.40%)	**0.2154**	0.1699 (-22.52%)	0.1483 (-31.15%)	**0.1591**	0.1268 (-20.30%)	0.1079 (-32.18%)	**0.2123**	0.1691 (-20.35%)	0.1477 (-30.43%)

Table 4.17: Effects of resource quality on retrieval performance of query expansion using pseudo-relevance feedback

Chapter 5

Improvement to CLIR Using Hybrid Approaches

In Chapter 4, the method of expanding queries using the eCMeSH Tree terms is compared with three individual approaches: query translation using the CMeSH Tree, query expansion based on pseudo-relevance feedback (PRF), and document translation (DT). The results show that all three approaches can improve retrieval performance effectively. However, their combination was not evaluated. In this chapter, we attempt to improve CLIR performance using hybrid approaches based on these methods. The experiments are:

Query expansion using the eCMeSH Tree and PRF This experiment evaluates the retrieval improvements attained after combining query expansion using the eCMeSH Tree with pseudo-relevance feedback.

Query expansion using the eCMeSH Tree with DT This experiment is designed to determine the contributions of the eCMeSH Tree terms and document translation to the performance of CLIR.

Query expansion using the eCMeSH Tree and PRF with DT The retrieval performance in experiment is enhanced by the eCMeSH Tree terms, pseudo-relevance feedback, and document translation.

5.1 Query Expansion Using the eCMeSH Tree and Pseudo-Relevance Feedback

Pseudo-relevance feedback and the eCMeSH Tree have both proved to be effective approaches to CLIR. However, the effect of combining the two methods has not been evaluated. In this section, an experiment is carried out to determine the contribution of such combination to the retrieval performance of CLIR.

5.1.1 Steps

Before query expansion, the TREC 2006 Genomics Track document collection is processed using the methods described in Section 4.2.2.2; and the queries of both 2006 and 2007 tracks are handled using the methods discussed in Section 4.2.2.1. Then a two-stage query expansion is carried out: i) the eCMeSH Tree is first applied to expand queries, as described in Section 4.5.2.1 and the expanded queries are translated using the methods mentioned in Section 4.5.2.1; and ii) pseudo-relevance feedback is then performed as described in Section 4.5.3.1. Algorithm 12 illustrates the steps to carry out these experiments.

Table 5.1 shows the example of Query 161 after expansion using the eCMeSH Tree and pseudo-relevance feedback.

5.1.2 Results

In Table 5.2, the results, referred to as “eCMeSH + PRF”, are compared with the results of the baseline experiment, or “Baseline”; the query expansion based only on the eCMeSH Tree, “eCMeSH”; and the query expansion using pseudo-relevance feedback, “PRF”.

The retrieval method leading to the best performance in this experiment, that is, query expansion using the eCMeSH Tree and pseudo-relevance feedback under Okapi BM25 with manual word segmentation, are compared with other related methods. Results of significance tests indicate that there is no significant difference among these methods.

Figure 5.1 illustrates the APs of each single query in both baseline experiment and the query expansion using the eCMeSH Tree and pseudo-relevance feedback. Figure 5.1(a) and Figure 5.1(b) are the average precisions for each query in the

Algorithm 12: Steps in the experiment using the eCMeSH Tree terms and pseudo-relevance feedback to expand the query

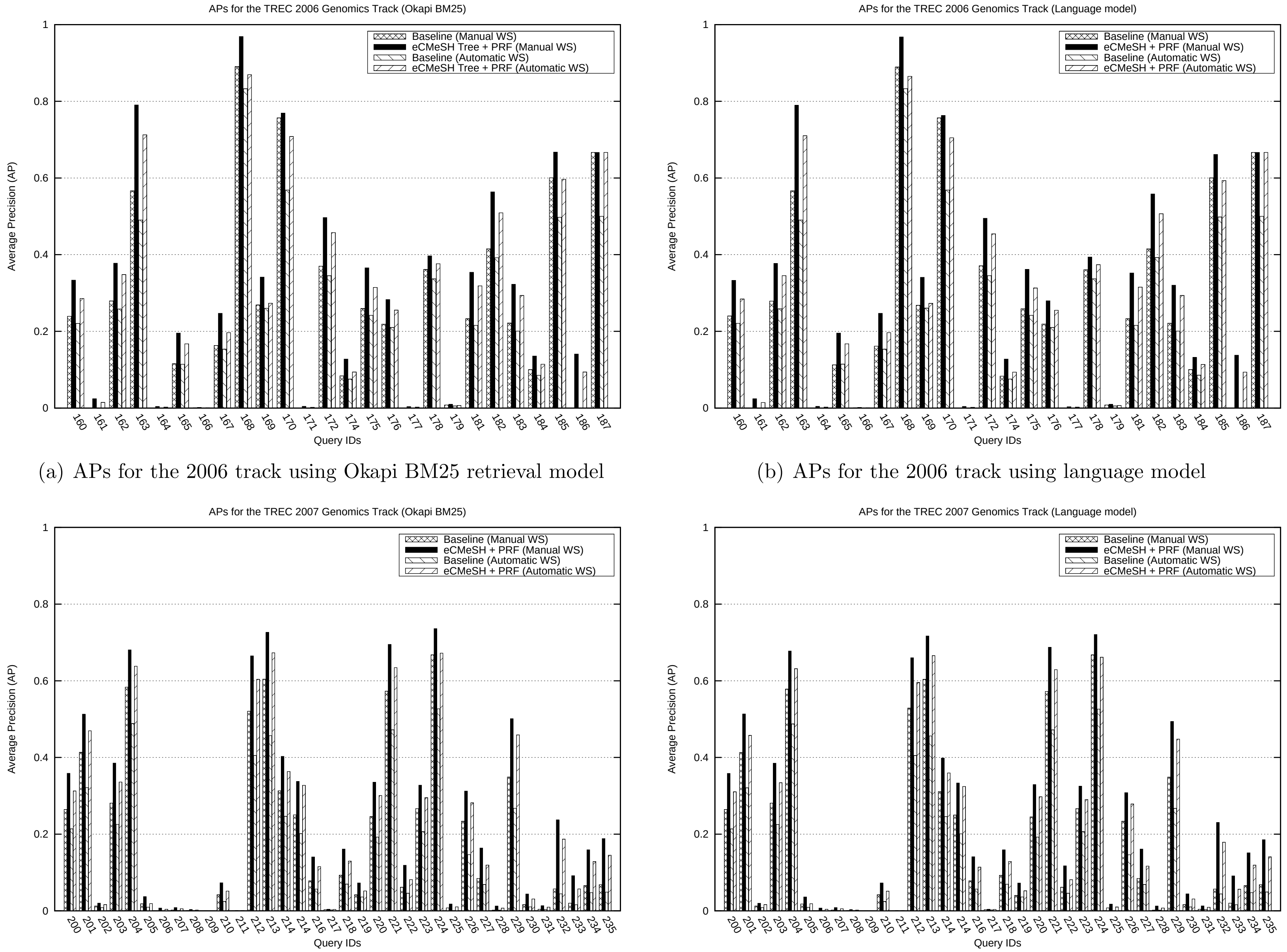
```
Input  : the indexed document collection ($D_I$), the query set ($Q$)
Output: MAP, AP

begin
    /* $S = \{t_i | t_i \in q_j, q_j \in Q.$ $t_i$ is the $i$th term of $q_j$.$\}$ */
    $Q \xrightarrow{\text{query pre-processings}} S$
    $S \xrightarrow{\text{the eCMeSH Tree (expanding)}} S_1$
    $S_1 \xrightarrow{\text{the Dictionary and the eCMeSH Tree (translating)}} S_2$
    $S_3 = \emptyset$
    for $s_i \in S_2$ do
        $s_i, D_I \xrightarrow{\text{Indri (inquiring)}} R_1$
        Select the top 25 ranked terms $T$ from $R_1$
        $S_3 = S_3 \cup T$
    end
    $S_2, D_I \xrightarrow{\text{Indri (inquiring)}} R$
    Evaluate $R$ using AP and MAP measures
end
```

TREC 2006 Genomics Track using different retrieval models. Figure 5.1(c) and Figure 5.1(d) shows the results of the 2007 Track.

5.1.3 Discussion

The results show that the combination of eCMeSH Tree terms and pseudo-relevance feedback to expand queries greatly improves the CLIR retrieval performance. In Table 5.2, the best retrievals are 0.3304 for the 2006 Track and 0.2375 for the 2007 Track, when manually segmented queries are used under Okapi BM25. Compared with the corresponding experiments: the baseline, the query expansion based only on the eCMeSH Tree, and the query expansion using only pseudo-relevance feedback, this hybrid approach gives improvements of 26.01%, 8.04%, and 9.80% for the 2006 Track, and 36.89%, 24.93%, and 10.26% for the 2007 Track, respectively. Figure 5.1 shows the great improvements attained in the experiments using the eCMeSH Tree and pseudo-relevance feedback to expand queries, compared with the baseline experiment.

Moreover, the contribution of the combination of the eCMeSH Tree terms



(a) APs for the 2006 track using Okapi BM25 retrieval model

(b) APs for the 2006 track using language model

(c) APs for the 2007 track using Okapi BM25 retrieval model

(d) APs for the 2007 track using language model

Figure 5.1: Results of query expansion using the eCMeSH Tree and pseudo-relevance feedback measured by AP

161	the original query	#combine(0.3333 阿尔茨海默病 0.3333 IDE 0.3333 作用)
	the expanded query (eCMeSH)	#combine(#syn(0.9760 AD症 0.9760 AD病 0.8636 Alzheimer氏病 0.8205 阿尔茨海默症 0.7226 早老性痴呆 0.6040 年性痴呆症 0.6040 老年性痴呆 0.2446 阿尔茨海默病 0.1862 阿滋海默症 . . .) 0.3333 IDE 0.3333 作用)
	the translated query	#combine(#syn(0.9760 #1(Alzheimer Disease) 0.9760 #1(Acute Confusional Senile Dementia) 0.9760 #1(Alzheimer Disease, Early Onset) 0.9760 #1(Alzheimer Disease, Late Onset) 0.9760 #1(Alzheimer Type Senile Dementia) 0.9760 #1(Alzheimer's Disease, Focal Onset) 0.2446 #1(Dementia, Alzheimer Type) 0.2446 #1(Early Onset Alzheimer Disease) . . .) 0.3333 IDE 0.3333 #1(act on) 0.3333 affect 0.3333 action 0.3333 function 0.3333 effect)
	the expanded query (PRF)	#weight(0.5 #combine(#syn(0.2446 #1(Acute Confusional Senile Dementia) 0.3796 #1(Alzheimer's Disease) 0.2446 #1(Dementia, Alzheimer Type) 0.2446 #1(Early Onset Alzheimer Disease) . . .) 0.3333 IDE 0.3333 #1(act on) 0.3333 affect 0.3333 action 0.3333 function 0.3333 effect) 0.5 #weight(0.7421 Alzheimer 0.7019 illness 0.6504 disease 0.5723 function 0.4710 affected 0.3241 damages 0.3200 treatment 0.2988 brain . . .))

Table 5.1: An example of a query expanded using the eCMeSH Tree and pseudo-relevance feedback

and pseudo-relevance feedback to retrieval performance is greater than the simple accumulation of each individual expansion approach. For instance, for the 2007 Track, the improvement achieved by this hybrid method, compared to the baseline, is 36.89%; while compared with the query expansion based on the eCMeSH Tree and the query expansion using pseudo-relevance feedback, the improvements are 24.93% and 10.26%, respectively. The sum of these two individual improvements is 35.19%, which is less than 36.89%. This indicates that this hybrid approach enhances both individual approaches.

It is also noticed that the improvement in retrieval for the 2007 Track is much higher than that for the 2006 Track. The average increment for the 2006 Track is 26.68% with a maximum of 27.89%, while the 2007 Track has an average increment of 45.07, maximised at 54.96%. This suggests that the combination approach is more effective for the 2007 queries than for the 2006 information need.

	BM25						LM					
	Automatic WS			Manual WS			Automatic WS			Manual WS		
2006	Baseline	eCMeSH	PRF	Baseline	eCMeSH	PRF	Baseline	eCMeSH	PRF	Baseline	eCMeSH	PRF
	0.2309	0.2647	0.2737	0.2622	0.3058	0.3009	0.2310	0.2390	0.2765	0.2619	0.2925	0.3178
	eCMeSH + PRF			eCMeSH + PRF			eCMeSH + PRF			eCMeSH + PRF		
	0.2953			**0.3304**			**0.2941**			**0.3287**		
	(27.89%)	(11.56%)	(7.89%)	(26.01%)	(8.04%)	(9.80%)	(27.32%)	(10.50%)	(6.37%)	(25.51%)	(12.38%)	(3.43%)
2007	Baseline	eCMeSH	PRF	Baseline	eCMeSH	PRF	Baseline	eCMeSH	PRF	Baseline	eCMeSH	PRF
	0.1352	0.1415	0.1654	0.1735	0.1901	0.2154	0.1350	0.1287	0.1591	0.1733	0.1899	0.2123
	eCMeSH + PRF			eCMeSH + PRF			eCMeSH + PRF			eCMeSH + PRF		
	0.2095			**0.2375**			**0.2064**			**0.2349**		
	(54.96%)	(48.06%)	(26.66%)	(36.89%)	(24.93%)	(10.26%)	(52.89%)	(60.37%)	(29.73%)	(35.55%)	(23.70%)	(10.65%)

Table 5.2: MAPs for the experiments of query expansion using the eCMeSH Tree and pseudo-relevance feedback

5.2 Query Expansion Using the eCMeSH Tree with Document Translation

In Section 4.5.4, the document translation approach can make a great improvement to the retrieval performance. This section is designed to evaluate the retrieval improvements attained when query expansion based on the eCMeSH Tree and the document translation approach are combined.

5.2.1 Steps

The first step in this experiment is to process the 2006 and 2007 queries using the methods described in Section 4.2.2.1. Then all the documents in the TREC 2006 Genomics Track document collection are translated into Chinese and indexed using bigrams, which is the same processing as discussed in Section 4.5.4.1. The third step is to apply the eCMeSH Tree to expand the Chinese queries. Algorithm 13 illustrates the steps in this experiment; expanded queries do not need to be translated into English, as the document collection has been translated from English into Chinese.

Algorithm 13: Steps in the experiment using the eCMeSH Tree terms to expand the query with document translation

```
Input  : the document collection ($D$), the query set ($Q$)
Output: $MAP$, $AP$

begin
    /* $S = \{t_i | t_i \in q_j, q_j \in Q.$ $t_i$ is the $i$th term of $q_j$.} */
    $Q \xrightarrow{\text{query pre-processings}} S$
    /* document translation */
    $D \xrightarrow{\text{Google tanslation (translating into Chinese)}} D_1$
    $D_1 \xrightarrow{\text{bigram (splitting)}} D_2$
    $D_2 \xrightarrow{\text{Indri (indexing)}} D_3$
    $S \xrightarrow{\text{the eCMeSH Tree (expanding)}} S_1$
    $S_1 \xrightarrow{\text{bigram (splitting)}} S_2$
    $S_2, D_3 \xrightarrow{\text{Indri (inquiring)}} R$
    Evaluate $R$ using $AP$ and $MAP$ measures
end
```

Table 5.3 shows the example of Query 161 after expansion using the eCMeSH Tree and document translation.

161	the original query	#combine(0.3333 阿尔茨海默病 0.3333 IDE 0.3333 作用)
	the expanded query (eCMeSH)	#combine(#syn(0.9760 AD症 0.9760 AD病 0.8636 Alzheimer氏病 0.8205 阿尔茨海默症 0.7226 早老性痴呆 0.6040 年性痴呆症 0.6040 老年性痴呆 0.2446 阿尔茨海默病 0.1862 阿滋海默症 . . .) 0.3333 IDE 0.3333 作用)
	the split query (bigram)	#combine(#syn(0.9760 #3(AD D症) 0.9760 #3(AD D病) 0.8636 #3(Al lz zh he ei im me er r氏 氏病) 0.8205 #3(阿尔 尔茨 茨海 海默 默症) 0.7226 #3(早老 老性 性痴 痴呆) . . .) 0.3333 #3(ID DE) 0.3333 作用)

Table 5.3: An example of a query expanded using the eCMeSH Tree and document translation

5.2.2 Results

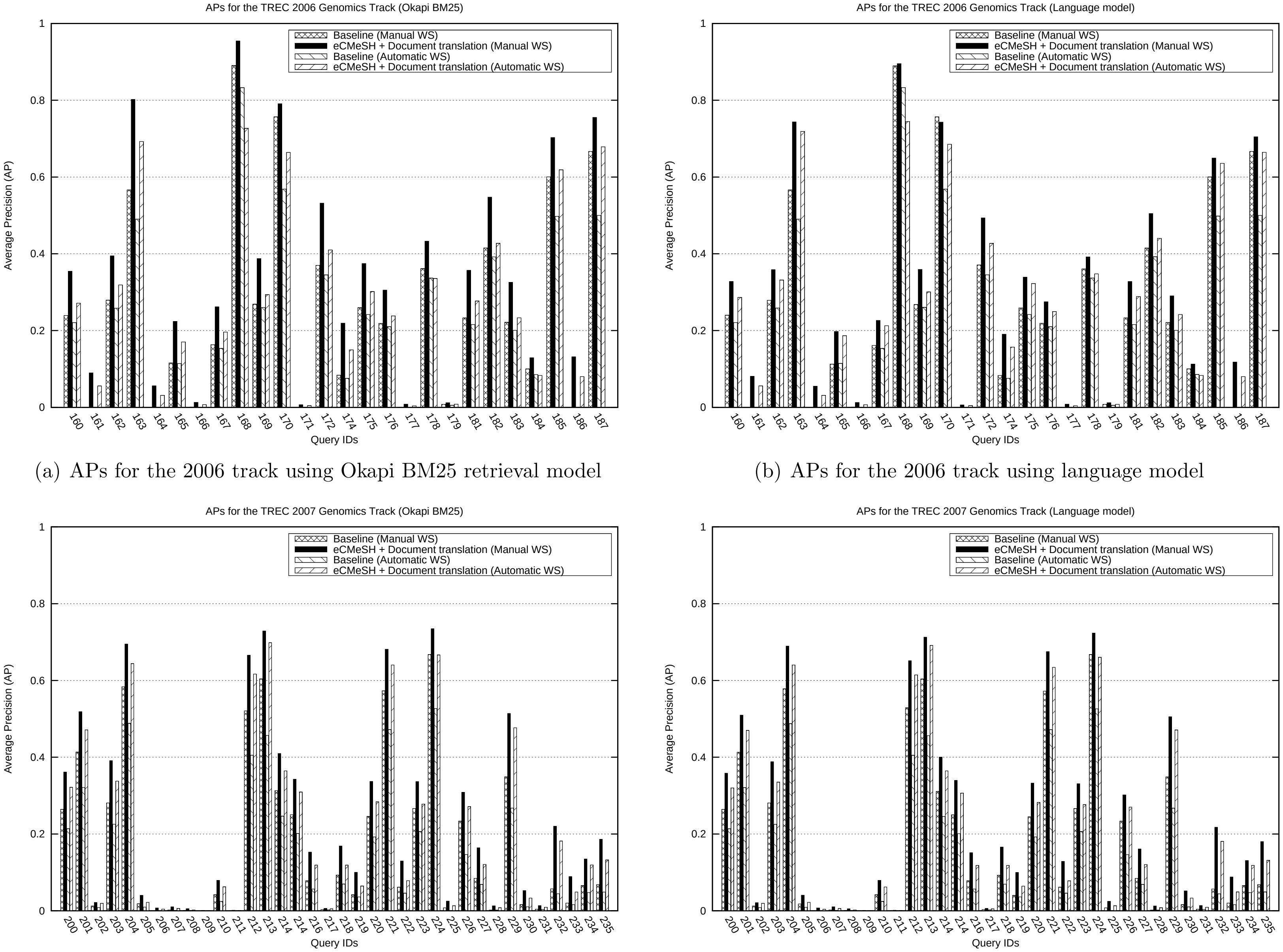
The results of applying the queries expanded by the eCMeSH Tree terms to the translated document collection, abbreviated as "eCMeSH + DT", are illustrated in Table 5.4, which compares them with Baseline, eCMeSH and DT.

Retrieval methods which have significant difference with the approach leading to the best performance are marked by "*" in Table 5.4. This indicates that the retrieval method of query expansion using the eCMeSH Tree with document translation performs better than the baseline method for both query sets, and than the eCMeSH only method for 2007 query set with automatic WS.

Figure 5.2 illustrates the APs of each single query in both the baseline experiment and the experiment of query expansion using the eCMeSH Tree on the translated documents. The average precision for each query in the TREC 2006 Genomics Track using different retrieval models is shown in Figure 5.2(a) and Figure 5.2(b). The corresponding experimental results for the 2007 Track are demonstrated in Figure 5.2(c) and Figure 5.2(d).

	BM25						LM					
	Automatic WS			Manual WS			Automatic WS			Manual WS		
2006	Baseline	eCMeSH	DT	Baseline	eCMeSH	DT	Baseline	eCMeSH	DT	Baseline	eCMeSH	DT
	0.2309*	0.2647	**0.2985**	0.2622	0.3058	0.3368	0.2310*	0.2390*	0.2791	0.2619	0.2925	0.3216
	eCMeSH + DT			eCMeSH + DT			eCMeSH + DT			eCMeSH + DT		
	0.2799			**0.3526**			**0.2890**			**0.3239**		
	(21.22%)	(5.74%)	(-6.23%)	(34.48%)	(15.30%)	(4.69%)	(25.11%)	(20.92%)	(3.55%)	(23.67%)	(10.74%)	(0.72%)
2007	Baseline	eCMeSH	DT	Baseline	eCMeSH	DT	Baseline	eCMeSH	DT	Baseline	eCMeSH	DT
	0.1352*	0.1415*	0.1800	0.1735*	0.1901	0.2305	0.1350*	0.1287*	0.1683*	0.1733*	0.1899	0.2257
	eCMeSH + DT			eCMeSH + DT			eCMeSH + DT			eCMeSH + DT		
	0.2100			**0.2401**			**0.2083**			**0.2366**		
	(55.33%)	(48.41%	(16.67%)	(38.39%)	(26.30%)	(4.16%)	(54.30%)	(61.85%)	(23.77%)	(36.53%)	(24.59%)	(4.83%)

Table 5.4: MAPs for the experiments of query expansion using the eCMeSH Tree with document translation



(a) APs for the 2006 track using Okapi BM25 retrieval model

(b) APs for the 2006 track using language model

(c) APs for the 2007 track using Okapi BM25 retrieval model

(d) APs for the 2007 track using language model

Figure 5.2: Results of query expansion using the eCMeSH Tree with document translation measured by AP

5.2.3 Discussion

The results in Table 5.4 show that the best retrieval performances for the 2006 and 2007 tracks are 0.3526 and 0.2401 respectively when manual word segmentation and Okapi BM25 are applied. Compared to the baseline, the query expansion using only the eCMeSH Tree, and the document translation method, the hybrid approach to CLIR have been increased respectively by 34.48% (for the 2006 Track) and 38.39% (for the 2007 Track), 15.30% (for the 2006 Track) and 26.30% (for the 2007 Track), and 4.69% (for the 2006 Track) and 4.16% (for the 2007 Track). Figure 5.2 shows that the combination of the eCMeSH Tree terms with the translated documents produces a great improvement on the retrieval performance for most of the queries.

As in the experiments conducted in Section 5.1, the contribution of query expansion using the eCMeSH Tree terms with document translation to retrieval performance is greater than the simple addition of each individual expansion approach. For example, the improvement achieved by this hybrid method on the 2006 query set, compared to the baseline, is 34.48%; while compared with the method of the query expansion based on the eCMeSH Tree and the document translation approach, the improvements are 15.30% and 4.69%, respectively. The sum of these two improvements is 19.99%, which is less than 34.48%.

Table 5.4 proves that this hybrid approach to CLIR is more effective for information needs like the 2007 queries than for the 2006 query set. For the 2007 Track, compared to the baseline, the best performance under Okapi BM25 and the language model are 0.2401 and 0.2366, improved by 38.39% and 36.53%, respectively. In contrast, the corresponding improvements for the 2006 query set are only 34.48% and 23.67%.

It is noticed that the approach to CLIR combining query expansion using eCMeSH Tree terms and document translation performs worse than the method of applying only document translation when automatically segmented queries of the 2006 Track are applied on the Okapi BM25 model. The reason for this drop may be the poor performance of the word segmentation tool, which introduces extra incorrect terms into the queries.

5.3 Query Expansion using the eCMeSH Tree and Pseudo-Relevance Feedback with Document Translation

The previous two sections evaluated the hybrid approaches to CLIR, where the query expansion using the eCMeSH Tree is combined separately with the query expansion using pseudo-relevance feedback and document translation. In this section, an experiment is designed to apply all three methods to implement CLIR.

5.3.1 Steps

Algorithm 14 illustrates the steps in carrying out the experiments which apply query expansion using the eCMeSH Tree and pseudo-relevance feedback to the translated documents.

The queries of the 2006 and 2007 tracks are first processed using the methods discussed in Section 4.2.2.1. All documents in the 2006 Track collection are translated into the Chinese equivalents using the Google translation service, and indexed by Indri. Then the segmented queries are expanded employing the eCMeSH Tree terms, as described in Section 4.5.2.1. The expanded queries are separated into bigrams and applied to retrieve relevant documents from the translated document collection. For each query, the top ranked 25 terms, which do not appear in the original query expanded using the eCMeSH Tree, selected from the top 50 relevant documents, are added to the original queries. The algorithms applied to extract candidate terms from the first 50 relevant documents are our approach to extending the CMeSH Tree, as described in Section 3.3.3. The term weight, or the algorithm used to rank candidate terms, shown in Equation 5.1, is *tf-idf*, where R is local set, here consisting of 50 top documents; tf_t is term frequency, i.e. the presence of term t among R; N is the number of all documents in the collection; and df_t is document frequency, which means the number of documents in the collection that contain term t.

$$w_i = \log\left(1 + \frac{\sum_{d \in R} tf_t}{|R|}\right) \cdot \log\left(\frac{N}{df_t}\right) \tag{5.1}$$

Finally, new queries are separated to evaluate the retrieval performance. Table 5.5 shows how Query 161 is processed according to these steps.

Algorithm 14: Steps in the experiment using the eCMeSH Tree terms and pseudo-relevance feedback to expand queries with document translation

```
Input   : the document collection ($D$), the query set ($Q$)
Output: $MAP$, $AP$

begin
    /* $S = \{t_i | t_i \in q_j, q_j \in Q$. $t_i$ is the $i$th term of $q_j$.$\}$    */
    $Q \xrightarrow{\text{query pre-processings}} S$
    /* document translation    */
    $D \xrightarrow{\text{Google tanslation (translating into Chinese)}} D_1$
    $D_1 \xrightarrow{\text{bigram (splitting)}} D_2$
    $D_2 \xrightarrow{\text{Indri (indexing)}} D_3$
    /* query expansion using the eCMeSH Tree    */
    $S \xrightarrow{\text{the eCMeSH Tree (expanding)}} S_1$
    $S_1 \xrightarrow{\text{bigram (splitting)}} S_2$
    /* query expansion using pseudo-relevance feedback    */
    $S_3 = \emptyset$
    for $s_i \in S_2$ do
        $s_i, D_3 \xrightarrow{\text{Indri (inquiring)}} R_1$
        /* $L$ is the ranked candidate term list    */
        The first 50 documents in $R_1 \xrightarrow{CMeSHextensionalgorithms} L$
        Select the top 25 ranked terms $T$ from $L$
        $S_3 = S_3 \cup T$
    end
    /* query term separation    */
    $S_1 \xrightarrow{\text{bigram (splitting)}} S_2$
    $S_3, D_3 \xrightarrow{\text{Indri (inquiring)}} R$
    Evaluate $R$ using $AP$ and $MAP$ measures
end
```

161	the original query	#combine(0.3333 阿尔茨海默病 0.3333 IDE 0.3333 作用)
	the expanded query (eCMeSH)	#combine(#syn(0.9760 AD症 0.9760 AD病 0.8636 Alzheimer氏病 0.8205 阿尔茨海默症 0.7226 早老性痴呆 0.6040 年性痴呆症 0.6040 老年性痴呆 0.2446 阿尔茨海默病 0.1862 阿滋海默症 . . .) 0.3333 IDE 0.3333 作用)
	the split query (bigram)	#combine(#syn(0.9760 #3(AD D症) 0.9760 #3(AD D病) 0.8636 #3(Al lz zh he ei im me er r氏 氏病) 0.8205 #3(阿尔 尔茨 茨海 海默 默症) 0.7226 #3(早老 老性 性痴 痴呆) . . .) 0.3333 #3(ID DE) 0.3333 作用)
	the expanded query (PRF)	#weight(0.5 #combine(#syn(0.9760 #3(AD D症) 0.9760 #3(AD D病) 0.8636 #3(Al lz zh he ei im me er r氏 氏病) 0.8205 #3(阿尔 尔茨 茨海 海默 默症) 0.7226 #3(早老 老性 性痴 痴呆) . . .) 0.3333 #3(ID DE) 0.3333 作用)) 0.5 #weight(0.8392 基因 0.8010 #3(血清 清D DH HT) 0.7958 年龄 0.6873 PE 0.5997 #3(认识 识能 能力 力退 退化) 0.5567 机制 0.5441 PC 0.4975 #3(前列 列腺) 0.4896 #3(双氢 氢睾 睾酮) . . . 0.3644 #3(神经 经退 退行 行性 性疾 疾病)))

Table 5.5: An example of a query expanded using the eCMeSH Tree and pseudo-relevance feedback

5.3.2 Results

Table 5.6 compares the results of this hybrid approach, abbreviated to "eCMeSH + PRF + DT" with Baseline, eCMeSH, PRF and DT.

Significance tests of this experiment are conducted by comparisons between retrieval method using the eCMeSH Tree terms and pseudo-relevance feedback to expand queries with document translation under Okapi BM25 with manually split queries and other related retrieval approaches. Table 5.6 represents these methods using "*". The significant difference implies that this hybrid approach to CLIR performs much better than methods used in the baseline and eCMeSH-only experiments, and that methods using automatic word segmentation perform worse than those using manually-separated queries.

Figure 5.3 reflects the retrieval improvements, compared with the baseline, attained using the hybrid approach from the point of view of each single query. Figure 5.3(a) and Figure 5.3(b) are the average precisions for each query in the TREC 2006 Genomics Track using different retrieval models. Figure 5.3(c) and Figure 5.3(d) illustrate the experimental results of the 2007 Track under the

	BM25								LM							
	Automatic WS				Manual WS				Automatic WS				Manual WS			
2006	Baseline	eCMeSH	PRF	DT	Baseline	eCMeSH	PRF	DT	Baseline	eCMeSH	PRF	DT	Baseline	eCMeSH	PRF	DT
	0.2309*	0.2647*	0.2737	0.2985	0.2622*	0.3058	0.3009	0.3368	0.2310*	0.2390*	0.2765	0.2791	0.2619*	0.2925	0.3178	0.3216
	eCMeSH + PRF + DT				eCMeSH + PRF + DT				eCMeSH + PRF + DT				eCMeSH + PRF + DT			
	0.3018				**0.3782**				**0.2973**				**0.3779**			
	(30.71%)	(14.02%)	(10.27%)	(1.11%)	(44.24%)	(23.68%)	(25.69%)	(12.29%)	(28.70%)	(24.39%)	(7.52%)	(6.52%)	(44.29%)	(29.20%)	(18.91%)	(17.51%)
2007	Baseline	eCMeSH	PRF	DT	Baseline	eCMeSH	PRF	DT	Baseline	eCMeSH	PRF	DT	Baseline	eCMeSH	PRF	DT
	0.1352*	0.1415*	0.1654*	0.1800	0.1735*	0.1901	0.2154	0.2305	0.1350*	0.1287*	0.1591*	0.1683*	0.1733*	0.1899	0.2123	0.2257
	eCMeSH + PRF + DT				eCMeSH + PRF + DT				eCMeSH + PRF + DT				eCMeSH + PRF + DT			
	0.2172				**0.2514**				**0.2103**				**0.2497**			
	(60.65%)	(53.50%)	(31.32%)	(20.67%)	(44.90%)	(32.25%)	(16.71%)	(9.07%)	(55.78%)	(63.40%)	(32.18%)	(24.96%)	(44.09%)	(31.49%)	(17.62%)	(10.63%)

Table 5.6: MAPs for the experiments combining the eCMeSH Tree, pseudo-relevance feedback, and document translation

corresponding retrieval models.

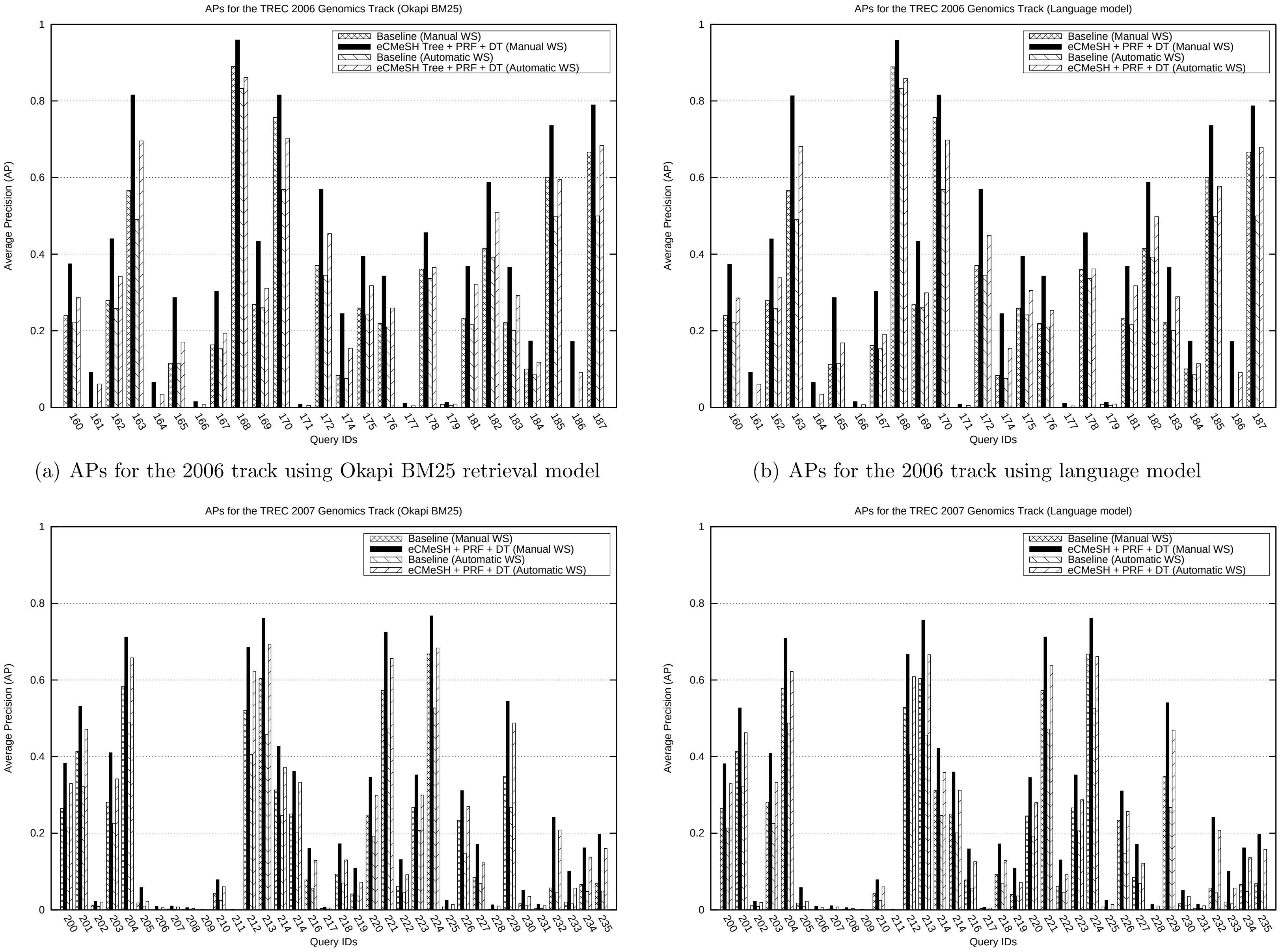
5.3.3 Discussion

The hybrid approach to CLIR, which applies the eCMeSH Tree and pseudo-relevance feedback to expand queries to retrieve the Chinese document collection translated by machine translation software, can vastly improve the retrieval performance. The best performance of the 2006 Track attained when Okapi BM25 and manual word segmentation are used is 0.3782, which is increased by 44.24% compared with the baseline, by 23.68% compared with only the eCMeSH Tree method, by 25.69% compared to the pseudo-relevance feedback, and by 12.29% compared with the document translation approach. The 2007 Track also achieves the best performance under the same conditions. Compared to the baseline, the eCMeSH-only approach, the pseudo-relevance feedback-only method, and the document translation approach, the 2007 Track is improved by 44.90%, 32.25%, 16.71%, and 9.97%, respectively. Figure 5.3 gives the detail of the improvements measured by AP for each single query.

In this round of experiments, we do not observe the increase that happens in Section 5.1 and Section 5.2.1. The sum of incremental degrees of query expansion using the eCMeSH Tree, and query expansion using pseudo-relevance feedback, and document translation is greater than that of the hybrid approach against the baseline experiment for both 2006 and 2007 tracks. A possible reason is that the pseudo-relevance feedback applied in this experiment is different from that used in Section 4.5.3. Because the translated Chinese documents are separated and indexed using bigrams, the Indri built-in pseudo-relevance feedback, which returns bigrams as terms, cannot be used. We design the pseudo-relevance feedback algorithm, whose term weight is calculated using Equation 5.1.

5.4 Comparison and Analysis

In this section, we compare of the above hybrid experiments. Table 5.7 illustrates the results of query expansion using the eCMeSH Tree and pseudo-relevance feedback, marked “e+PRF”, query expansion using the eCMeSH Tree with document translation, referred to as “e+DT”, and query expansion using the eCMeSH Tree and pseudo-relevance feedback with document translation, abbreviated to “e+PRF+DT”.



(a) APs for the 2006 track using Okapi BM25 retrieval model

(b) APs for the 2006 track using language model

(c) APs for the 2007 track using Okapi BM25 retrieval model

(d) APs for the 2007 track using language model

Figure 5.3: Results of query expansion using the eCMeSH Tree and pseudo-relevance feedback with document translation measured by AP

The best CLIR performance is attained when the queries expanded using the eCMeSH Tree and pseudo-relevance feedback are applied in the translated Chinese documents. For the TREC 2006 Track, it is 0.3782; and for the 2007 Track, it achieves 0.2524.

The eCMeSH Tree smooths the difference of the query expansion using pseudo-relevance feedback and the document translation approach. In this discussion, the result of the 2006 Track using Okapi BM25 and manual word segmentation is considered as the example. Table 4.16 shows that the difference (represented as "D1") between the document translation method and pseudo-relevance feedback is 11.93%, while after combining the eCMeSH Tree terms, the difference (marked "D2") between "e+DT" and "e+PRF" is 6.72%. The decrease of D2 against D1 is 43.67%, which is the contribution of the eCMeSH Tree terms to minimising the difference between various approaches to CLIR. Moreover, the introduced eCMeSH Tree terms are too strong to reduce the performance of the query expansion experiment using the eCMeSH Tree against document translation. For example, the 2006 Track under the language model using manual word segmentation achieves 0.3239, compared with the corresponding experiment combining the eCMeSH Tree with pseudo-relevance feedback, which is decreased by 1.19%. We avoid discussing such smoothing using the "e+PRF+DT" data as an example, because the pseudo-relevance feedback described in Section 5.3.1, in the "e+PRF+DT" experiment, is different from that used in the "e+PRF" experiment.

5.5 Summary

In this section, we extend the individual experiments conducted in Section 4.5 to evaluate the retrieval improvements attained using hybrid approaches. The results suggest that the approaches combining individual methods produce great improvements. The best retrieval performances for the 2006 and 2007 sets are attained when the eCMeSH Tree and pseudo-relevance feedback are applied to expand the queries in a translated Chinese document collection.

In the next, final, chapter, we will draw conclusions from this study.

	BM25						LM					
	automatic WS			manual WS			automatic WS			manual WS		
	e+PRF	e+DT	e+PRF+DT	e+PRF	e+DT	e+PRF+DT	e+PRF	e+DT	e+PRF+DT	e+PRF	e+DT	e+PRF+DT
2006	0.2953	0.2799	**0.3018**	0.3304	0.3526	**0.3782**	0.2941	0.2890	**0.2973**	0.3278	0.3239	**0.3779**
		(-5.22%)	(2.20%)		(6.72%)	(14.47%)		(-1.73%)	(1.09%)		(-1.19%)	(15.28%)
			(7.82%)			(7.26%)			(2.87%)			(16.67%)
2007	0.2095	0.2100	**0.2172**	0.2375	0.2401	**0.2514**	0.2064	0.2083	**0.2103**	0.2349	0.2366	**0.2497**
		(0.24%)	(3.68%)		(1.10%)	(5.85%)		(0.92%)	(1.89%)		(0.72%)	(6.30%)
			(3.43%)			(4.71%)			(0.96%)			(5.54%)

Table 5.7: Comparisons of hybrid approaches

Chapter 6

Conclusions

6.1 Research Review

In this thesis, we were concerned with the problem of improving Chinese-English biomedical cross-lingual information retrieval. In particular, we were concerned with the problem of making use of ontologies to expand queries.

Utilising online resources, we proposed an algorithm to construct bilingual ontologies, namely the Extended Chinese Medical Subject Headings (eCMeSH) Tree, suitable for query expansion. The basic idea is to extend the original Chinese Medical Subject Headings (CMeSH) with synonymous Chinese terms and their weights.

The problems related to the CMeSH Tree lie in: i) each English MeSH heading term has one and only one Chinese translation; and ii) there are no term weights for either English or Chinese heading terms. These limitations make the original CMeSH Tree terms unsuitable for query expansion, as evaluated in Section 4.5.1. However, the CMeSH Tree provides a prototype for bilingual ontologies.

In order to overcome the missing-term problem, we first used Google to retrieve Chinese documents containing the candidate terms, and then employed a three-level parsing and filtering technique, which combines a rule-based parsing approach with the C-value term extraction algorithm and a filter based on mutual information, to extract the Chinese candidate terms. The details of the extraction algorithm were described in Section 3.3.3.

We proposed a term weighting scheme, which resolved the lack of term weights. Our approach exploited the frequency of English heading terms and the frequency of each corresponding Chinese term, using the Google search engine to compute the term weight for each Chinese term; the English heading terms have no calculated weights, because they can inherit term weights from their Chinese equivalents. Section 3.3.4 discussed this term weight scheme.

We investigated the relations between the eCMeSH Tree terms and CLIR retrieval performance, using serial experiments. The baseline experiment described in Section 4.3 provides the comparison standard for further experiments. The experiment of query expansion using the CMeSH Tree, discussed in Section 4.5.1, showed that the CMeSH Tree terms decreased retrieval performance. The experiment of query expansion using the eCMeSH Tree, described in Section 4.5.2, proved that the eCMeSH Tree can improve CLIR retrieval performances over the baseline's 0.2622 (for the TREC 2006 Track) and 0.1735 (for the 2007 Track) to 0.3058 and 0.1901, respectively.

We also compared the approach to CLIR using the eCMeSH Tree to expand queries with other methods, such as query expansion using pseudo-relevance feedback, discussed in Section 4.5.3, and document translation, described in Section 4.5.4. Our query expansion using the eCMeSH Tree performed slightly worse than the query expansion using pseudo-relevance feedback and document translation.

We finally evaluated the improvements in retrieval performance gained by combining the individual approaches to CLIR: the eCMeSH Tree, pseudo-relevance feedback, and document translation. We conducted three hybrid experiments: query expansion using the eCMeSH Tree and pseudo-relevance feedback (Section 5.1), query expansion using the eCMeSH Tree with document translation (Section 5.2), and query expansion using the eCMeSH Tree and pseudo-relevance feedback with document translation (Section 5.3). The results show that, compared with individual approaches, all the hybrid approaches greatly improved the retrieval performance, and that the best performances for the 2006 and 2007 tracks were 0.3782 and 0.2514, attained when eCMeSH and pseudo-relevance feedback were used to expand queries for the translated Chinese document collection.

The problem we faced in the evaluation of CLIR using different approaches was that the retrieval performance may be affected by several factors: stemming, term segmentation, and retrieval models, for example. The experiments described in Section 4.3 compared the retrieval performances with and without the Porter stemming algorithm showing that the stemming algorithm did not improve the

CLIR. In all experiments, we illustrated the results using both manual and automatic word segmentation schemes, because no appropriate word segmentation tool has been designed for Chinese biomedical texts. The results suggest that the manual word segmentation led to much better retrieval performance than did the word segmentation using tools. We also compared the retrieval performance for each experiment under the Okapi BM25 model, a probabilistic retrieval model, with that attained when the query likelihood language model was applied. In general, Okapi BM25 performed slightly better than the language model. We also designed an experiment which optimised the parameters in Okapi BM25 and the language model.

6.2 Objectives Attained

The principal objective attained in this work, as shown by our methodology and the results of our experiments, was the approach to Chinese-English biomedical cross-lingual information retrieval using the eCMeSH Tree to expand queries.

The study presented in this thesis has made the following contributions:

- We developed an algorithm which utilises online resources and an automatic term extraction technique to expand the CMeSH Tree as the eCMeSH Tree.

- We investigated existing approaches, based on translation, to CLIR, such as dictionary-based query expansion using the eCMeSH Tree, non-dictionary based query expansion using pseudo-relevance feedback, and document translation, and combined them. We identified and evaluated the strengths and weaknesses of each method.

- We evaluated the retrieval performance attained under the Okapi BM25 model, a representative of the probabilistic model, and the query likelihood language model, respectively, and optimised the parameters of these two models.

- We discussed the effects to the CLIR retrieval performance resulting from the presence and absence of the stemming algorithm.

- We compared the CLIR retrieval performance when the manual and automatic word segmentation strategies were applied.

6.3 Limitations and Future Work

The main focus of the work presented in this thesis was the investigation of our hypothesis for Chinese-English cross-lingual information retrieval on biomedical literature, namely that a biomedical CLIR can perform better using weighted bilingual ontologies to expand queries than using classic dictionary translation approaches. For this reason, a number of issues related to the study were not undertaken, since they were beyond the scope of our study.

For example, the good-performance word segmentation tool is essential. In this study, we used “BaseSeg” [ZHL06], a newswire-doma segmentation tool, to separate the Chinese queries into terms. Results have proven that poor retrieval performance was attained using this tool. Gu et al. [GPD08] propose a biomedical named entity recogniser for Chinese. According to their study, the *F-score* is 65.60%, which does not satisfy our requirement.

Document collection is another issue. There is no gold standard or document set in the Chinese language for the biomedical domain, which implies that English-Chinese cross-lingual information retrieval cannot be conducted.

The study itself raised other issues. One of them is the term weighting scheme used in our work:

(1) During the extension of the eCMeSH Tree, the term weight computed using Equation 3.4 is not monotonic to f_{ct}, the frequency of a Chinese term, which means that some high-frequency terms may be assigned a low-term weight. We need to find a monotonic function to calculate term weights.

(2) The strategy of making use of the eCMeSH Tree term weights is debatable. The English terms in eCMeSH have no calculated term weights; and their weights are inherited from the corresponding Chinese terms. Perhaps, the term weight for an English term should be calculated using the same function as the weight calculation of the Chinese terms. Moreover, the current way to determine such inherited weights is to select the maximum weight among all the possible weights, as several different Chinese terms can correspond to the same English term. A better solution might be to replace the maximum weight with the average value of all the possible weights of an English term.

The language model evaluated in this thesis may be a problem. It is the query likelihood language model, which was discussed in Section 2.1.3.4. We did not evaluate the *model comparison* approach to the language model, because previous researchers [LZ01, MS99] had come to different conclusions about the retrieval performance using the model comparison.

The pseudo-relevance feedback performed in the experiment using the eCMeSH Tree and pseudo-relevance feedback to expand queries over a set of translated Chinese documents, described in Section 5.3.1, is another issue. In this experiment, the weight, presented in Equation 5.1, used to rank candidate terms, is computed using the $tf \cdot idf$ measure. There are other choices, such as *term selection value (TSV)*[RW99] and *Kullback-Leibler divergence (KLD)*[Cro00, p.154]. In our study, the candidate term weighting function is equal to the function used to rank the term, but Robertson and Spärck Jones [RSJ76] recommend a term weighting function, which is applied after ranking.

Apart from the above-mentioned problems which should be resolved in future work, there is also a series of research aspects related to biomedical CLIR requiring further investigation, such as the following:

- English-Chinese cross-lingual information retrieval on biomedical literature: the English-Chinese CLIR in the biomedical domain enables English users to access biomedical documents written in Chinese, especially the huge volume of documents on traditional Chinese medicine. The eCMeSH Tree, consisting of a number of Chinese biomedical terms organised by taxonomy, is a valuable bilingual resource from which to start such CLIR.

- Improved retrieval model including domain knowledge: the principal reason why a general domain search engine produces poor performance on biomedical information need is that the applied retrieval models do not include specialised knowledge. Although some measures have been tried, like using biomedical terms as indexing terms, these attempts have not been successful. The difficulty lies in what level or kind of domain knowledge should be used and where to use it, and how to compute its relevance.

Bibliography

[AF01] Mohammed Aljlayl and Ophir Frieder. Effective Arabic-English Cross-Language Information Retrieval via Machine-Readable Dictionaries and Machine Translation. In *Proceedings of the 10th International Conference on Information and Knowledge Management*, CIKM '01, pages 295–302. ACM, 2001.

[Ana88] Sophia Ananiadou. *Towards a Methodology for Automatic Term Recognition*. PhD thesis, University of Manchester, 1988.

[Ana94] Sophia Ananiadou. A Methodology for Automatic Term Recognition. In *Proceedings of the 15th Conference on Computational Linguistics - Volume 2*, COLING '94, pages 1034–1038, Stroudsburg, PA, USA, 1994. Association for Computational Linguistics.

[AR96] Eneko Agirre and German Rigau. Word Sense Disambiguation using Conceptual Density. In *Proceedings of the 16th International Conference on Computational Linguistics (COLING'96)*, pages 16–22, Copenhagen, Denmark, 1996.

[ARS05] Samir Abdou, Patrick Ruch, and Jacques Savoy. Evaluation of Stemming, Query Expansion and Manual Indexing Approaches for the Genomic Task. In *The 14th Text REtrieval Conference Proceedings (TREC 2005)*, pages 863–871, 2005.

[AS06] Samir Abdou and Jacques Savoy. Report on the TREC 2006 Genomics Experiment. In *Proceedings of the 15th Text REtrieval Conference (2006)*, Gaithersburg, Maryland, USA, 2006. National Institute of Standards and Technology (NIST).

[AS07] Samir Abdou and Jacques Savoy. Searching in MEDLINE: Query

Expansion and Manual Indexing Evaluation. *Information Processing and Management*, 44(2):781–789, 2007.

[ATO05] Mustafa Abusalah, John Tait, and Michael Oakes. Literature Review of Cross Language Information Retrieval. *World Academy of Science, Engineering and Technology*, 4:175–177, 2005.

[BADW04] Christopher Brewster, Harith Alani, Srinandan Dasmahapatra, and Yorick Wilks. Data Driven Ontology Evaluation. In *Proceedings of International Conference on Language Resources and Evaluation*, pages 24–30, Lisbon, 2004.

[Bax58] Phyllis B. Baxendale. Machine-made Index for Technical Literature: An Experiment. *IBM Journal of Research and Development*, 2(4):354–361, 1958.

[Bax61] Phyllis B. Baxendale. An Empirical Model for Machine Indexing. In: Machine Indexing; Pro#ress and Probleme. In *Third Institute for Information Storage and Retrieval*, pages 207–218, Center for Technology and Administration, School of Government and Public Administration, The American University, Washington, D.C., 1961.

[BB01] Anita Burgun and Olivier Bodenreider. Comparing Terms, Concepts and Semantic Classes in WordNet and the Unified Medical Language System. In *Proceedings of the NAACL 2001 Workshop 'WordNet and Other Lexical Resources: Applications, Extensions and Customizations'*, pages 77–82, Pittsburgh, PA, USA, 2001. New Brunswick, NJ: Association for Computational Linguistics.

[BB02] Olivier Bodenreider and Anita Burgun. Characterizing the Definitions of Anatomical Concepts in WordNet and Specialized Sources. In *Proceedings of the 1st International Conference of the Global WordNet Association*, pages 223–239, Mysore, India, 2002.

[BBM03] Olivier Bodenreider, Anita Burgun, and Joyce A. Mitchell. Evaluation of WordNet as a Source of Lay Knowledge for Molecular Biology and Genetic Diseases: A Feasibility Study. *Studies in Health Technology and Informatics*, 95:379–384, 2003.

[BC98] Guo-Wei Bian and Hsin-Hsi Chen. Integrating Query Translation and Document Translation in a Cross-Language Information Retrieval System. In *Proceedings of the Third Conference of the Association for Machine Translation in the Americas on Machine Translation and the Information Soup*, AMTA '98, pages 250–265, 1998.

[BCC04] Stefan Büttcher, Charles L. A. Clarke, and Gordon V. Cormack. Domain-Specific Synonym Expansion and Validation for Biomedical Information Retrieval. In *Proceedings of the 13th Text REtrieval Conference, TREC-2004*, Gaithersburg, Maryland, USA, 2004. National Institute of Standards and Technology (NIST).

[BL99] Adam Berger and John Lafferty. Information Retrieval as Statistical Translation. In *Proceedings of the 22nd Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR '99, pages 222–229, Berkeley, California, United States, 1999.

[BMWC00] Chris Buckley, Mandar Mitra, Janet A. Walz, and Claire Cardie. Using Clustering and SuperConcepts within SMART: TREC 6. *Information Processing and Management*, 36:109–131, 2000.

[Boo82] Abraham Bookstein. Explanation and Generalization of Vector Models in Information Retrieval. In *Proceedings of the 5th Annual ACM Conference on Research and Development in Information Retrieval*, pages 118–132, West Berlin, Germany, 1982. Springer-Verlag New York, Inc.

[BPPM93] Peter F. Brown, Vincent J. Della Pietra, Stephen A. Della Pietra, and Robert L. Mercer. The Mathematics of Statistical Machine Translation: Parameter Estimation. *Computational Linguistics*, 19(2):263–311, 1993.

[BS75] Abraham Bookstein and Donald R. Swanson. A Decision Theoretic Foundation for Indexing. *Journal of the American Society for Information Science*, 26(1):45–50, 1975.

[BSAS94] Chris Buckley, Gerard Salton, James Allan, and Amit Singhal. Automatic Query Expansion Using SMART: TREC 3. In D. K.

Harman, editor, *Proceedings of Text REtrieval Conference (TREC)*, pages 69–80, Gaithersburg, MD, 1994. National Institute of Standards and Technology Special Publication.

[BSU08] Graham Bennett, Falk Scholer, and Alexandra Uitdenbogerd. A Comparative Study of Probabilistic and Language Models for Information Retrieval. In *Nineteenth Australasian Database Conference (ADC)*, pages 65–74. Vol. 75 of CRPIT, ACS, 2008.

[Bus45] Vannevar Bush. As We May Think. The Atlantic Monthly, July 1945.

[BZ05] Bodo Billerberk and Justin Zobel. Document Expansion versus Query Expansion for Ad-hoc Retrieval. In A. Turpin and R. Wilkinson, editors, *Proceedings of the Tenth Australasian Document Computing Symposium*, pages 34–41, 2005.

[Car04] Bob Carpenter. Phrasal Queries with LingPipe and Lucene: Ad Hoc Genomics Text Retrieval. In *Proceedings of the 13th Text REtrieval Conference, TREC 2004*, Gaithersburg, Maryland, USA, 2004. National Institute of Standards and Technology (NIST).

[CB01] Margaret H. Coletti and Howard L. Bleich. Medical Subject Headings Used to Search the Biomedical Literature. *Journal of the American Medical Information Association*, 8(4):317–323, 2001.

[CCG94] William S. Cooper, Aitao Chen, and Fredric C. Gey. Full Text Retrieval Based on Probabilistic Equations with Coefficients Fitted by Logistic Regression. In *Proceedings of the Second Text REtrieval Conference (TREC-2)*, pages 57–66, Gaithersburg, MD, 1994.

[CCH92] James P. Callan, William Bruce Croft, and Stephen M. Harding. The INQUERY Retrieval System. In *Proceedings of the Third International Conference on Database and Expert Systems Applications (DEXA '1992)*, pages 78–83, Valencia, Spain, 1992. Springer-Verlag.

[CFK$^+$04] Margaret Connell, Ao Feng, Giridhar Kumaran, Hema Raghavan, Chirag Shah, and James Allan. UMass at TDT 2004. In *Topic Detection and Tracking Workshop (TDT)*, Gaithersburg, MD, 2004. National Institute of Standards and Technology.

[CG96] Stanley F. Chen and Joshua Goodman. An Empirical Study of Smoothing Techniques for Language Modeling. In *Proceedings of the 34th Annual Meeting on Association for Computational Linguistics*, ACL '96, pages 310–318, Santa Cruz, California, 1996. Association for Computational Linguistics.

[CG04] Aitao Chen and Fredric C. Gey. Combining Query Translation and Document Translation in Cross-Language Retrieval. In *Comparative Evaluation of Multilingual Information Access Systems*, volume 3237, pages 121–124, Berlin, Heidelberg, 2004. Springer-Verlag.

[CGJ01] Aitao Chen, Fredric C. Gey, and Hailing Jiang. Berkeley at NTCIR-2: Chinese, Japanese, and English IR Experiments. In *Proceedings of the 2nd NTCIR Workshop on Research in Chinese & Japanese Text Retrieval and Text Summarization*, pages 137–145, National Institute of Informatics, Tokyo, Japan, 2001.

[CH79] William Bruce Croft and David J. Harper. Using Probabilistic Models of Document Retrieval without Relevance Information. *Journal of Documentation*, 35(4):285–295, 1979.

[Chi94] Lee-Feng Chien. A Model-based Signature File Approach for Full-text Retrieval of Chinese Document Database. *Computer Processing of Chinese and Oriental Languages*, 8(Supplement):59–76, 1994.

[Chi97] Lee-Feng Chien. PAT-tree-based Keyword Extraction for Chinese Information Retrieval. In *Proceedings of the 20th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval)*, pages 50–58, Philadelphia, PA, 1997.

[CHX$^+$97] Aitao Chen, Jianzhang He, Liangjie Xu, Fredric C. Gey, and Jason Meggs. Chinese Text Retrieval without Using a Dictionary. In *Proceedings of the 20th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR '97, pages 42–49, 1997.

[CJG00] Aitao Chen, Hailing Jiang, and Fredric C. Gey. Combining Multiple Sources for Short Query Translation in Chinese-English Cross-Language Information Retrieval. In *Proceedings of the 5th International Workshop Information Retrieval with Asian Languages*, IRAL '00, pages 17–23, Hong Kong, China, 2000. ACM Press, New York.

[CLWK01] Ken C. W. Chow, Robert W. P. Luk, Kam Fai Wong, and Kui Lam Kwok. Hybrid Term Indexing for Weighted Boolean and Vector Space Models. *Proceedings of International Journal of Computer Processing of Oriental Languages*, 14(2):133–151, 2001.

[CM98] Gregory F. Cooper and Randolph F. Miller. An Experiment Comparing Lexical and a Statistical Methods for Extracting MeSH Terms from Clinical Free Text. *Journal of the American Medical Informatics Association*, 5(1):62–75, 1998.

[Cro88] Carlyn J. Crouch. A Cluster-Based Approach to Thesaurus Construction. In *Proceedings of the 11th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR '88, pages 309–320, Grenoble, France, 1988. ACM.

[Cro00] William Bruce Croft. *Advances in Information Retrieval.* Kluwer Academic Publisher, Norwell, MA, 2000.

[CW65] Dan C. Clarke and R. E. Wall. An Economical Program for Limited Parsing of English. In *Proceedings of the November 30–December 1, 1965, Fall Joint Computer Conference, Part I*, AFIPS '65 (Fall, part I), pages 307–316, New York, NY, USA, 1965. ACM.

[CWL98] S. K. Chan, C. Y. Wong, and Robert W. P. Luk. Variable Bit-Block Compression Signature for English-Chinese Information Retrieval. In *Proceedings of the Conference on Information Retrieval with Asian Langauges 98*, IRAL '98, pages 61–66, KRDL, National University of Singapore, 1998.

[DDFL90] Scott Deerwester, Susan T. Dumais, George W. Furnas, and Thomas K. Landauer. Indexing by Latent Semantic Analysis. *Journal of the American Society for Information System*, 41(6):391–407, 1990.

[DFLD88] Susan T. Dumais, George W. Furnas, Thomas K. Landauer, and Scott Deerwester. Using Latent Semantic Analysis to Improve Information Retrieval. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, pages 281–285, New York, 1988. ACM Press.

[DLLL97] Susan T. Dumais, Todd A. Letsche, Michael L. Littman, and Thomas K. Landauer. Automatic Cross-Language Retrieval Using Latent Semantic Indexing. In *AAAI Spring Symposium on Cross-Language Text and Speech Retrieval*, pages 18–24, USA, 1997. Stanford University.

[DM83] Martin Dillon and Laura K. McDonald. Fully Automatic Book Indexing. *Journal of Documentation*, 39(3):135–154, 1983.

[DMB98] Simon Dennis, Robert McArthur, and Peter D. Bruza. Searching the World Wide Web Made Easy? The Cognitive Load Imposed by Query Refinement Mechanisms. In J. Kay and M. Milosavljevic, editors, *Proceedings of Australian Document Computing Conference*, pages 65–71, Sydney, Australia, 1998.

[Dum94] Susan T. Dumais. Latent Semantic Indexing (LSI): TREC-3 Report. In D. K. Harman, editor, *Proceedings of Text REtrieval Conference (TREC)*, pages 219–230, New York, 1994. National Institute of Standards and Technology Special Publication 500-225, ACM Press.

[ECL$^+$88] Peter L. Elkin, James J. Cimino, Henry J. Lowe, David B. Aronow, Tom H. Payne, Pierre S. Pincetl, and G. Octo Barnett. Mapping to MeSH: The Art of Trapping MeSH Equivalence from within Narrative Text. In *Proceedings of the Twelfth Annual Symposium on Computer Applications in Medical Care*, pages 185–190. IEEE Comput. Soc. Press, 1988.

[EZ96] David A. Evans and Chengxiang Zhai. Noun-Phrase Analysis in Unrestricted Text for Information Retrieval. In *Proceedings of the 34th annual meeting on Association for Computational Linguistics*, ACL '96, pages 17–24, Stroudsburg, PA, USA, 1996. Association for Computational Linguistics.

[Fag87] Joel L. Fagan. *Experiments in Automatic Phrase Indexing for Document Retrieval: A Comparison of Syntactic and non-Syntactic Methods*. PhD thesis, Cornell University, 1987.

[FAM00] Katerina Frantzi, Sophia Ananiadou, and Hideki Mima. Automatic Recognition of Multi-word Terms: the C-value/NC-value Method. *International Journal of Digital Library*, 3(2):117–132, 2000.

[FDD+88] George W. Furnas, Scott Deerwester, Susan T. Dumais, Thomas K. Landauer, Richard A. Harshman, Lynn A. Streeter, and Karen E. Lochbaum. Information Retrieval Using a Singular Value Decomposition Model of Latent Semantic Structure. In *Proceedings of the 11th ACM International Conference on Research and Development in Information Retrieval*, SIGIR '88, pages 465–480, 1988.

[FHS05] Christiane Fellbaum, Udo Hahn, and Barry Smith. Towards New Information Resources for Public Health — From WordNet to MedicalWordNet. *Journal of Biomedical Informatics*, 39:321–332, 2005.

[Fox86] Edward A. Fox. Information Retrieval: Research into New Capabilities. In S. Lambert and S. Ropiequet, editors, *CD-ROM: The New Papyrus*, pages 143–174, Redmond, Washington, 1986. Microsoft Press.

[FTP+08] Wei-Yu Fan, Yuan-Yuan Tong, Yan-Li Pan, Wen-Ling Shang, Jia-Yi Shen, Wei Li, and Li-Jun Li. Traditional Chinese Medical Journals Currently Published in Mainland China. *The Journal of Alternative and Complementary Medicine*, 14(5):595–609, 2008.

[GG09] Stephen J. Greenberg and Patricia E. Gallagher. The Great Contribution: Index Medicus, Index-Catalogue, and IndexCat. *The Journal of the Medical Library Association*, 97(2):108–113, 2009.

[GHG04] Yikun Guo, Henk Harkema, and Rob Gaizauskas. Sheffield University and the TREC 2004 Genomics Track: Query Expansion Using Synonymous Terms. In E. M. Voorhees and L. P. Buckland, editors, *Proceedings of Text REtrieval Conference (TREC)*, Gaithersburg, MD, 2004. National Institute of Standards and Technology Special Publication 500-261.

[GNN$^+$07] Wei Gao, Cheng Niu, Jian-Yun Nie, Ming Zhou, Jian Hu, Kam-Fai Wong, and Hsiao-Wuen Hon. Cross-Lingual Query Suggestion Using Query Logs of Different Languages. In *Proceedings of the 30th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR '07, pages 463–470, 2007.

[GNZ06] Jianfeng Gao, Jian-Yun Nie, and Ming Zhou. Statistical Query Translation Models for Cross-Language Information Retrieval. *ACM Transactions on Asian Language Information Processing (TALIP)*, 5(4):323–359, 2006.

[GPD08] Baohua Gu, Fred Popowich, and Veronica Dahl. Recognizing Biomedical Named Entities in Chinese Research Abstracts. In *Proceedings of the Canadian Society for computational studies of intelligence, 21st conference on Advances in artificial intelligence*, Canadian AI '08, pages 114–126, Windsor, Ontario, 2008.

[GZN$^+$02] Jianfeng Gao, Ming Zhou, Jian-Yun Nie, Hongzhao He, and Weijun Chen. Resolving Query Translation Ambiguity Using a Decaying Co-Occurrence Model and Syntactic Dependence Relations. In *Proceedings of the 25th Annual International ACMSIGIR Conference on Research and Development in Information Retrieval*, SIGIR '02, pages 183–190, Tampere, Finland, 2002. ACM Press, New York.

[Har88] Donna Harman. Towards Interactive Query Expansion. In *Proceedings of the 11th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR '88, pages 321–331, Grenoble, France, 1988.

[Har92] Donna Harman. Relevance Feedback Revisited. In *Proceedings of the 15th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR '92, pages 1–10, Copenhagen, Denmark, 1992.

[HCRR07] William Hersh, Aaron Cohen, Lynn Ruslen, and Phoebe Roberts. TREC 2007 Genomics Track. In *Proceedings of the 16th Text REtrieval Conference (TREC-16)*, Gaithersburg, Maryland, USA, 2007. National Institute of Standards and Technology (NIST).

[Hei98] Djoerd Heimstra. A Linguistically Motivated Probabilistic Model of Information Retrieval. In *Proceedings of the Second European Conference on Research and Advanced Technology for Digital Libraries*, ECDL '98, pages 569–584, 1998.

[Her09] William R. Hersh. *Information Retrieval: A Health and Biomedical Perspective*. Springer-Verlan, New York, USA, 3rd edition edition, 2009.

[HHR06] Xiangji Huang, Bin Hu, and Hashmatullah Rohian. York University at TREC 2006: Genomics Track. In *Proceedings of the 15th Text REtrieval Conference*, TREC-15, Gaithersburg, Maryland, USA, 2006. National Institute of Standards and Technology (NIST).

[Hil68] Donald J. Hillman. Negotiation of Inquiries in an On-line Retrieval System. *Information Storage and Retrieval*, 4(2):219–238, 1968.

[HR97] Xiangji Huang and S. E. Robertson. Okapi Chinese Text Retrieval Experiments at TREC-6. In *Proceedings of theTREC-6 Conference*, pages 137–142, Gaithersburg, MD, 1997.

[HSHRA07] Xiangji Huang, Damon Sotoudeh-Hosseini, Hashmat Rohian, and Xiangdong An. York University at TREC 2007: Genomics Track. In *Proceedings of the 16th Text REtrieval Conference*, TREC-16, Gaithersburg, Maryland, USA, 2007. National Institute of Standards and Technology (NIST).

[Hul94] David Hull. Improving Text Retrieval for the Routing Problem Using Latent Semantic Indexing. In *Proceedings of the 17th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR '94, pages 282–291, Dublin, Ireland, 1994.

[HXC$^+$96] Jianzhang He, Jack Xu, Aitao Chen, Jason Meggs, and Fredric C. Gey. Berkeley Chinese Information Retrieval at TREC-5: Technical Report. In *Proceedings of the TREC-5 Conference*, pages 181–186, Gaithersburg, MD, 1996.

[Ide71] E. Ide. New Experiments in Relevance Feedback. In Gerard Salto, editor, *The SMART System Experiments in Automatic Document Processing*, pages 337–354, Englewood Cliffs, NY, USA, 1971. Prentice Hall.

[JC94] Yufeng Jing and William Bruce Croft. An Association Thesaurus for Information Retrieval. In *Proceedings of th 4th International Conference "Recherche d'Information Assistee par Ordinateur"*, RIAO '94, pages 146–160, 1994.

[Joa97] Thorsten Joachims. A Probabilistic Analysis of the Rocchio Algorithm with TFIDF for Text Categorization. In *Proceedings of the Fourteenth International Conference on Machine Learning*, ICML '97, pages 143–151, 1997.

[Jon83] Kevin P. Jones. How do we Index? A Report of some ASLIB Informatics Group Activities. *Journal of Documentation*, 39(1):1–23, 1983.

[JSP05] Bernard J. Jansen, Amanda Spink, and Jan Pedersen. A Temporal Comparison of Altavista Web Searching. *Journal of the American Society for Information Science and Technology*, 56(6):559–570, 2005.

[JW02] Honglan Jin and Kam-Fai Wong. A Chinese Dictionary Construction Algorithm for Information Retrieval. *ACM Transactions on Asian Language Information Processing (TALIP)*, 1(4):281–296, 2002.

[JZ07] Jing Jiang and ChengXiang Zhai. An Empirical Study of Tokenization Strategies for Biomedical Information Retrieval. *Information Retrieval*, 10(4-5):341–363, 2007.

[KBLJP55] Allen Kent, Madeline M. Berry, Fred U. Luehrs Jr., and J. W. Perry. Machine Literature Searching VIII. Operation Criteria for Designing Information Retrieval Systems. *American Documentation*, 6(2):93–101, April 1955.

[KCDD04] Kui-Lam Kwok, Sora Choi, Norbert Dinstl, and Peter Deng. NTCIR-4 Chinese, English, Korean Cross Language Retrieval Experiments using PIRCS. In *Working Notes of the Fourth NTCIR Workshop Meeting*, NTCIR4, pages 186–192. National Institute of Informatics, Tokyo, Japan, 2004.

[KN04] Michael Krauthammer and Goran Nenadic. Term Identification in the Biomedical Literature. *Journal of Biomedical Informatics*, 37(6):512–526, 2004.

[KP98] Wessel Kraaij and Renée Pohlmann. Comparing the Effect of Syntactic vs. Statistical Phrase Indexing Strategies for Dutch. In *Proceedings of the Second European Conference on Research and Advanced Technology for Digital Libraries*, ECDL '98, pages 605–617, London, UK, 1998. Springer-Verlag.

[KS03a] Wessel Kraaij and Martijn Spitters. Language Models for Topic Tracking. In W. B. Croft and J. Lafferty, editors, *Language Modeling for Information Retrieval*, pages 95–124, 2003.

[KS03b] Anand Kumar and Barry Smith. The Unified Medical Language System and the Gene Ontology: Some Critical Reflections. In *KI2003: Advances in AI (2003)*, pages 135–148, 2003.

[Kwo96] Kui-Lam Kwok. A Network Approach to Probabilistic Information Retrieval. *ACM Transactions on Information Systems (TOIS)*, 13(3):324–353, 1996.

[Kwo99] Kui-Lam Kwok. Employing Multiple Representations for Chinese Information Retrieval. *Journal of the American Society for Information Science and Technology*, 50(8):709–723, 1999.

[Kwo01] Kui-Lam Kwok. NTCIR-2 Chinese, Cross Language Retrieval Experiments Using Pircs. In *Proceedings of the 2nd NTCIR Workshop on Research in Chinese & Japanese Text Retrieval and Text Summarization*, pages 111–118, 2001.

[Lan69] F. Wilfrid Lancaster. MEDLARS: Report on the Evaluation of Its Operating Efficiency. *American Documentation*, 20(2):119–142, 1969.

[Lav04] Victor Lavrenko. *A Generative Theory of Relevance*. PhD thesis, University of Massachusetts, 2004.

[LB87] Henry J. Lowe and G. Octo Barnett. MicroMeSH: A Microcomputer System for Searching and Exploring the National Library of Medicine's Medical Subject Headings (MeSH) Vocabulary. In *Proc Annual Symp Comput Appl Med Care*, pages 717–720, 1987.

[LB94] Henry J. Lowe and G. Octo Barnett. Understanding and Using the Medical Subject Headings (MeSH) Vocabulary to Perform Literature Searches. *Journal of the American Medical Association*, 271(14):1–18, 1994.

[LC02] Wei-Hao Lin and Hsin-Hsi Chen. Backward Machine Transliteration by Learning Phonetic Similarity. In *proceedings of the 6th conference on Natural language learning - Volume 20*, COLING-02, pages 1–7, Taipei, Taiwan, 2002.

[LCC01] T. R. Lynam, C. L. A. Clarke, and G. V. Cormack. Information Extraction with Term Frequencies. In *Proceedings of the First International Conference on Human Language Technology Research*, HLT '01, pages 1–4, 2001.

[LCC02] Victor Lavrenko, Martin Choquette, and William Bruce Croft. Cross-Lingual Relevance Models. In *Proceedings of the 25th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR '02, pages 175–182, New York, NY, USA, 2002. ACM.

[LCL02] Wen-Hsiang Lu, Lee-Feng Chien, and Hsi-Jian Lee. Translation of Web Queries Using Anchor Text Mining. *ACM Transactions on Asian Language Information Processing (TALIP)*, 2(1):159–172, 2002.

[LHZ$^+$01] Danya Li, Tiejun Hu, Wenyan Zhu, Qing Qian, Huiling Ren, Junlian Li, and Bin Yang. Retrieval System for the Chinese Medical Subject Headings (in Chinese). *Chinese Journal of Medical Library*, 4, 2001.

[Lin08] Jimmy Lin. Pagerank without Hyperlinks: Reranking with PubMed Related Article Networks for Biomedical Text Retrieval. *BMC Bioinformatics*, 9(270), 2008.

[LK02] Robert W. P. Luk and Kui-Lam Kwok. A Comparison of Chinese Document Indexing Strategies and Retrieval Models. *ACM Transactions on Asian Language Information Processing (TALIP)*, 1(3):225–268, 2002.

[LKFK01] Robert W. P. Luk, Wong. Kam-Fai, and Kui-Lam Kwok. Hybrid Term Indexing: an Evaluation. In *Proceedings of the 2nd NTCIR Workshop on Research in Chinese & Japanese Text Retrieval and Text Summarization*, pages 130–136, 2001.

[LKW09] Zhiyong Lu, Won Kim, and W. John Wilbur. Evaluation of Query Expansion Using MeSH in PubMed. *Information Retrieval*, 12(1):69–80, 2009.

[LL90] Thomas K. Landauer and Michael L. Littman. Fully Automatic Cross-Language Document Retrieval Using Latent Semantic Indexing. In *Proceedings of the Sixth Annual Conference of the UW Centre for the New Oxford English Dictionary and Text Research*, pages 31–38, Waterloo, Ontario, 1990.

[LM03] Yuk-Chi Li and Helen M. Meng. Document Expansion Using a Side Collection for Monolingual and Cross-language Spoken Document Retrieval. In *ISCA Workshop on Multilingual Spoken Document Retrieval*, pages 85–90, Hong Kong, 2003.

[LO02] Gina-Anne Levow and Douglas W. Oard. *Signal Boosting for Translingual Topic Tracking: Document Expansion and N-best Translation*, pages 175–195. Kluwer Academic Publishers, Norwell, MA, 2002.

[LOR04] Gina-Anne Levow, Douglas W. Oard, and Philip Resnik. Dictionary-Based Techniques for Cross-Language Information Retrieval. *Information Processing & Management*, 41(3):523–547, 2004.

[LT07] Jeanine Lilleng and Stein L. Tomassen. Cross-Lingual Information Retrieval by Feature Vectors. In *Lecture Notes in Computer Science*, volume 4592/2007, pages 229–239, 2007.

[LTGP04] Adolfo Lozano-Tello and Asunción Gómez-Pérez. ONTOMETRIC: A Method to Choose the Appropriate Ontology. *Journal of Database Management*, 15(2):1–18, 2004.

[Luh59] Hans Peter Luhn. Auto-Encoding of Documents for Information Retrieval Systems. In *Modern Trends in Documentation*, pages 45–58, 1959.

[LW02] Nicholas Lester and Hugh E. Williams. Topic Tracking at RMIT University. In *Topic Detection and Tracking Workshop (TDT)*, Gaithersburg, MD, 2002. National Institute of Standards and Technology.

[LWW01] Wai Lam, Kam-Fai Wong, and Chi-Yin Wong. Chinese Document Indexing Based on a New Partitioned Signature File: Model and Evaluation. *Journal of the American Society for Information Science and Technology*, 52(7):584–597, 2001.

[LWY$^+$09] Chu-Cheng Lin, Yu-Chun Wang, Chih-Hao Yeh, Wei-Chi Tsai, and Richard Tzong-Han Tsai. Learning Weights for Translation Candidates in Japanese-Chinese Information Retrieval. *Expert Systems with Application*, 36(4):7695–7699, 2009.

[LXG07] Chengye Lu, Yue Xu, and Shlomo Geva. Translation Disambiguation in Web-based Translation Extraction for English-Chinese CLIR. In *Proceedings of the 2007 ACM Symposium on Applied Computing*, SAC '07, pages 819–823, 2007.

[LZ97] Mun-Kew Leong and Hong Zhou. Preliminary Qualitative Analysis of Segmented vs Bigram Indexing in Chinese. In *Proceedings of the 6th Text Retrieval Conference*, TREC-6, pages 551–557, 1997.

[LZ01] John Lafferty and Chengxiang Zhai. Document Language Models, Query Models and Risk Minimization for Information Retrieval. In

Proceedings of the 24th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval, SIGIR '01, pages 111–119, New Orleans, Louisiana, United States, 2001.

[McC99] J. S. McCarley. Should We Translate the Documents or the Queries in Cross-Language Information Retrieval? In *Proceedings of the 37th Meeting of the Association for Computational Linguistics on Computational Linguistics*, 1999.

[Mil95] George A. Miller. WordNet: A Lexical Database for English. *Communications of the ACM*, 38(11):39–41, 1995.

[MLS99] David R. H. Miller, Tim Leek, and Richard M. Schwartz. A Hidden Markov Model Information Retrieval System. In *Proceedings of the 22nd Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR '99, pages 214–221, Berkeley, California, United States, 1999.

[MM90] David M. Magerman and Mitchell P. Marcus. Parsing a Natural Language Using Mutual Information Statistics. In *Proceedings of AAAI-90, English National Conference on Artificial Intelligence*, pages 984–989, 1990.

[MM98] A. T. McCray and R. A. Miller. Making the Conceptual Connections: the Unified Medical Language System (UMLS) after a Decade of Research and Development. *Journal of the American Medical Informatics Association*, 5(1):129–130, 1998.

[MMP00] Paul McNamee, James Mayfield, and Christine Piatko. The Haircut System at TREC-9. In *Proceedings of the Ninth Text REtrieval Conference (TREC-9)*, pages 273–294, Gaithersburg, MD, 2000.

[MMS80] Takashi Maeda, Yoshio Momouchi, and Hajime Sawamura. An Automatic Method for Extracting Significant Phrases in Scientific or Technical Documents. *Information Processing & Management*, 16(3):119–127, 1980.

[Moo50] Calvin Mooers. The Theory of Digital Handling of Non-numerical

Information and Its Implications to Machine Economics. In *Proceedings of the Meetings of the Association for Computing Machinery*, Rutgers University, March 1950.

[MOPT07] Donna R. Maglott, James Ostell, Kim D. Pruitt, and Tatiana A. Tatusova. Entrez Gene: Gene-Centered Information at NCBI. *Nucleic Acids Research*, 35(Database issue):26–31, 2007.

[MP08] Philipp Mayr and Vivien Petras. Cross-Concordances: Terminology Mapping and Its Effectiveness for Information Retrieval. In *World Library and Information Congress: 74th IFLA General Conference and Council*, Québec, Canada, 2008.

[MRS08] Christopher D. Manning, Prabhakar Raghavan, and Hinrich Schütze. *Introduction to Information Retrieval.* Cambridge University Press, 2008.

[MS99] Christopher D. Manning and Hinrich Schütze. *Foundations of Statistical Natural Language Processing.* MIT Press, Cambridge, Massachusetts, USA, 1999.

[MS02] Alexander Maedche and Steffen Staab. Measuring Similarity between Ontologies. In *Proceedings of the 13th International Conference on Knowledge Engineering and Knowledge Management. Ontologies and the Semantic Web*, EKAW '02, pages 251–263, 2002.

[MSB98] Mandar Mitra, Amit Singhal, and Chris Buckley. Improving Automatic Query Expansion. In *Proceedings of the 21st Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR '98, pages 206–214, Melbourne, Australia, 1998.

[MTT99] Rila Mandala, Takenobu Tokunaga, and Hozumi Tanaka. Combining Multiple Evidence from Different Types of Thesaurus for Query Expansion. In *Proceedings of the 22nd Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR '99, pages 191–197, Berkeley, California, United States, 1999.

[MTT00] Rila. Mandala, Takenobu Tokunaga, and Hozumi Tanaka. The Exploration and Analysis of Using Multiple Thesaurus Types for Query Expansion in Information Retrieval. *Journal of Natural Language Processing*, 7(2):117–140, 2000.

[MvR97] Mark Magennis and Cornelis J. van Rijsbergen. The Potential and Actual Effectiveness of Interactive Query Expansion. *SIGIR Forum*, 21(SI):324–332, 1997.

[MWZ72] Jack Minker, Gerald A. Wilson, and Barbara H. Zimmerman. An Evaluation of Query Expansion by the Addition of Clustered Terms for a Document Retrieval System. *Information Storage and Retrieval*, 8(6):329–348, 1972.

[NBR96] Jian-Yun Nie, Marting Brisebois, and Xiaobo Ren. On Chinese Text Retrieval. In *Proceedings of the 19th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR '96, pages 225–233, 1996.

[NOH$^+$09] D. Nguyen, A. Overwijk, C. Hauff, R. B. Trieschnigg, D. Hiemstra, and F. M. G. de Jong. WikiTranslate: Query Translation for Cross-Lingual Information Retrieval Using Only Wikipedia. In *Evaluating Systems for Multilingual and Multimodal Information Access*, pages 58–65, 2009.

[NSA05] Goran Nenadic, Irena Spasic, and Sophia Ananiadou. Mining Biomedical Abstracts: What's in a Term? In *Natural Language Processing IJCNLP 2004*, pages 797–806, 2005.

[OA06] Naoaki Okazaki and Sophia Ananiadou. Building an Abbreviation Dictionary Using a Term Recognition Approach. *Bioinformatics*, 22(24):3089–3095, 2006.

[Oar98] Douglas W. Oard. A Comparative Study of Query and Document Translation for Cross-Language Information Retrieval. In *Proceedings of the Third Conference of the Association for Machine Translation in the Americas on Machine Translation and the Information Soup*, AMTA '98, pages 472–483. Springer-Verlag, 1998.

[OAT10] Naoaki Okazaki, Sophia Ananiadou, and Jun'ichi Tsujii. Building a High Quality Sense Inventory for Improved Abbreviation Disambiguation. *Bioinformatics*, 26(9):1246–1253, 2010.

[PBG02] Emanuele Pianta, Luisa Bentivogli, and Christian Girardi. MultiWordNet: Developing an Aligned Multilingual Database. In *Proceedings of the First International Conference on Global WordNet*, pages 293–302, Mysore, India, 2002.

[PBH08] Michael Poprat, Elena Beisswanger, and Udo Hahn. Building a Biowordnet by Using WordNet's Data Formats and WordNet's Software Infrastructure — a Failure Story. In *Software Engineering, Testing, and Quality Assurance for Natural Language Processing*, SETQA-NLP '08, pages 31–39, Stroudsburg, PA, USA, 2008. Association for Computational Linguistics.

[PC98] Jay M. Ponte and William Bruce Croft. A Language Modelling Approach to Information Retrieval. In *Proceedings of the 21st Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR '98, pages 275–281, Melbourne, Australia, 1998.

[Pit84] Anne B. Piternick. Searching Vocabularies: A Developing Category of Online Search Tools. *Online Review*, 8(5):441–449, 1984.

[PK97] Renée Pohlmann and Wessel Kraaij. The Effect of Syntactic Phrase Indexing on Retrieval Performance for Dutch Texts. In *Proceedings of RIAO '97*, pages 176–187, 1997.

[PL03] Ari Pirkola and Erkka Leppänen. TREC 2003 Genomics Track Experiments at UTA: Query Expansion with Predefined High Frequency Terms. In *Proceedings of the 12th Text REtrieval Conference, TREC 2003*, pages 796–805, Gaithersburg, Maryland, USA, 2003. National Institute of Standards and Technology (NIST).

[PM04] Robert Porzel and Rainer Malaka. A Task-based Approach for Ontology Evaluation. In *Proceedings of ECAI 2004 Workshop on Ontology Learning and Population*, 2004.

[Por80] Martin F. Porter. An Algorithm for Suffix Stripping. *Program*, 14(3):130–137, 1980.

[PS08] Solmaz Pourmahmoud and Shamsfard Shamsfard. Semantic Cross-Lingual Information Retrieval. In *the 23rd International Symposium on Computer and Information Sciences,*, ISCIS'08, pages 1–4, 2008.

[QF93] Yonggang Qiu and H. P. Frei. Concept Based Query Expansion. In *Proceedings of the 16th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR '93, pages 160–169, Pittsburgh, Pennsylvania, United States, 1993.

[QF99] Yayun Qin and Changqi Feng. 中文医学主题词表(机读版)在文献标引中的应用 (Literature Citations Using Chinese Medical Subject Headings (Machine-Readable Version)) (in Chinese). *Journal of Medical Intelligence*, 5, 1999.

[RB09] Fuji Ren and David B. Bracewell. Advanced Information Retrieval. *Electronic Notes in Theoretical Computer Science*, 225:303–317, 2009.

[Rip96] Brian D. Ripley. *Pattern Recognition and Neural Networks.* Cambridge University Press, 1996.

[Roc71] Joseph J. Rocchio. Relevance Feedback in Information Retrieval. In Gerard Salton, editor, *The SMART Retrieval System — Experiments in Automatic Document Processing*, pages 313–323, Englewood Cliffs, NH, USA, 1971. Prentice Hall.

[RSJ76] Stephen E. Robertson and Karen Spärck Jones. Relevance Weighting of Search Terms. *Journal of American Society for Information Science*, 27(3):129–146, 1976.

[RvRP81] Stephen E. Robertson, Cornelis J. van Rijsbergen, and Martin F. Porter. Probabilistic Models of Indexing and Searching. In *Proceedings of the Third Annual ACM Conference on Research and Development in Information Retrieval*, pages 35–56, Cambridge, England, 1981.

[RW94] Stephen E. Robertson and Steve Walker. Some Simple Effective Approximations to the 2-Poisson Model for Probabilistic Weighted Retrieval. In *Proceedings of the 17th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR '94, pages 232–241, Dublin Ireland, 1994.

[RW99] Stephen E. Robertson and Steve Walker. Okapi/Keenbow at TREC-8. In E. M. Voorhees and D. K. Harman, editors, *Proceedings of Text REtrieval Conference (TREC)*, pages 151–161, Gaithersburg, MD, 1999. National Institute of Standards and Technology Special Publication 500-246.

[RW00] Stephen E. Robertson and Steve Walker. Microsoft Cambridge at TREC-9: Filtering Track. In E. M. Voorhees and D. K. Harman, editors, *Proceedings of Text REtrieval Conference (TREC)*, pages 361–368, Gaithersburg, MD, 2000. National Institute of Standards and Technology Special Publication 500-249.

[RWHB$^+$92] Stephen E. Robertson, Steve Walker, Micheline Hancock-Beaulieu, Aarron Gull, and Marianna Lau. Okapi at TREC. In D. K. Harman, editor, *Proceedings of Text REtrieval Conference (TREC)*, pages 21–30, Gaithersburg, MD, 1992. National Institute of Standards and Technology Special Publication 500-207.

[RWJ$^+$94] Stephen E. Robertson, Steve Walker, S. Jones, M. M. Hancock-Beaulieu, and M. Gatford. Okapi at TREC-3. In D. K. Harman, editor, *Proceedings of Text REtrieval Conference (TREC)*, pages 109–126, Gaithersburg, MD, 1994. National Institute of Standards and Technology Special Publication 500-225.

[SAC07] Mark D. Smucker, James Allan, and Ben Carterette. A Comparison of Statistical Significance Tests for Information Retrieval. In *Proceedings of the Sixteenth ACM Conference on Conference on Information and Knowledge Management*, CIKM '07, pages 623–632, 2007.

[Sal68] Gerard Salton. *Automatic Information Organization and Retrieval.* McGraw-Hill, New York, 1968.

[Sal71] Gerard Salton. *The SMART Retrieval System: Experiments in Automatic Document Processing.* Prentice Hall, 1971.

[Sal75] Gerard Salton. *A Theory of Indexing.* Society for Industrial and Applied Mathematics, Philadelphia, PA, USA, 1975.

[Sal82] Gerard Salton. A Blueprint for Automatic Boolean Query Processing. *ACM SIGIR Forum*, 17(2):6–25, 1982.

[Sal86] Gerard Salton. Another Look at Automatic Text-Retrieval Systems. *Communications of ACM*, 29(7):648–656, 1986.

[SAMK05] Irena Spasić, Sophia Ananiadou, John McNaught, and Anand Kumar. Text Mining and Ontologies in Biomedicine: Making Sense of Raw Text. *Briefings in Bioinformatics*, 6(3):239–251, 2005.

[SC99] Mark Sanderson and William Bruce Croft. Deriving Concept Hierarchies from Text. In *Proceedings of the 22nd Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR '99, pages 206–213, Berkeley, California, United States, 1999.

[SFW83] Gerard Salton, Edward A. Fox, and Harry Wu. Extended Boolean Information Retrieval. *Communications of ACM*, 26(11):1022–1036, 1983.

[Shu06] Mary Shultz. Mapping of Medical Acronyms and Initialisms to Medical Subject Headings (MeSH) across Selected Systems. *Journal of the Medical Library Association*, 94(4):410–414, 2006.

[SJK73] Karen Spärck Jones and Martin Kay. *Linguistics and Information Science.* Academic Press, New York, NY, 1973.

[SJWR00] Karen Spärck Jones, Steve Walker, and Stephen E. Robertson. A Probabilistic Model of Information Retrieval: Development and Comparative Experiments, Parts 1&2. *Information Processing and Management*, 36(6):779–840, 2000.

[SM83] Gerard Salton and Michael J. McGill. *Introduction to Modern Information Retrieval.* McGraw Hill Book Co., New York, 1983.

[Smi06] Barry Smith. From Concepts to Clinical Reality: An Essay on the Benchmarking of Biomedical Terminologies. *Journal of Biomedical Informatics*, 39(3):288–298, 2006.

[Smu06] Mark D. Smucker. UMass Genomics 2006: Query-Biased Pseudo Relevance Feedback. In *Proceedings of the 15th Text REtrieval Conference (TREC 2006)*, Gaithersburg, Maryland, USA, 2006. National Institute of Standards and Technology (NIST).

[SN00] Ian Soboroff and Charles Nicholas. Collaborative Filtering and the Generalized Vector Space Model. In *Proceedings of the 23rd Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR '00, pages 351–353, Athens, Greece, 2000.

[SOK94] Alan F. Smeaton, Ruairi O'Donnell, and Fergus Kelledy. Indexing Structures Derived from Syntax in TREC-3: System Description. In *The Third Text REtrieval Conference (TREC-3)*, TREC '94, pages 55–67, 1994.

[SP99] Amit Singhal and Fernando Pereira. Document Expansion for Speech Retrieval. In *Proceedings of the 22nd Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR '99, pages 34–41, Berkeley, California, United States, 1999.

[SS89] Gerard Salton and Maria Smith. On the Application of Syntactic Methodologies in Automatic Text Analysis. In *Proceedings of the 12th annual international ACM SIGIR conference on Research and development in information retrieval*, SIGIR '89, pages 137–150, New York, NY, USA, 1989. ACM.

[SvR83] Alan F. Smeaton and Cornelis J. van Rijsbergen. The Retrieval Effects of Query Expansion on a Feedback Document Retrieval System. *Computer Journal*, 26(3):239–246, 1983.

[SWY75] Gerard Salton, Andrew K. C. Wong, and Chung-Shu Yang. A Vector Space Model for Automatic Indexing. *Communications of the ACM*, 18(11):613–620, 1975.

[SXP01] Ju Sheng, Shiqian Xie, and Chengyi Pan. *概率论与数理统计(第四版) (Probability and Mathematical Statistics, the 4th Edition) (in Chinese)*. Higher Education Press, Beijing, P. R. China, 2001.

[SYY75] Gerard Salton, Chung-Shu. Yang, and Clement Tak Yu. A Theory of Term Importance in Automatic Text Analysis. *Journal of the American Society for Information Science*, 26:33–44, 1975.

[THdJK10] Dolf Trieschnigg, Djoerd Hiemstra, Franciska de Jong, and Wessel Kraaij. A Cross-Lingual Framework for Monolingual Biomedical Information Retrieval. In *Proceedings of the 19th ACM International Conference on Information and Knowledge Management*, pages 169–178, New York, NY, USA, 2010. ACM.

[TJ03] Yuen-Hsien Tseng and Da-Wei Juang. Document-Self Expansion for Text Categorization. In *Proceedings of the 26th Annual International ACM SIGIR Conference on Research and Development in Informaion Retrieval*, SIGIR '03, pages 399–400, New York, Toronto, Canada, 2003.

[TLW99] T. F. Tsang, R. W. P. Luk, and K. F. Wong. Hybrid Term Indexing Using Words and Bigrams. In *Proceedings of the Information Retrieval with Asian Languages 1999 Conference*, pages 112–117, 1999.

[Tom03] Stephen Tomlinson. Robust, Web and Genomic Retrieval with Hummingbird SearchServer at TREC 2003. In *Proceedings of the 12th Text REtrieval Conference (TREC 2003)*, pages 254–267, Gaithersburg, Maryland, USA, 2003. National Institute of Standards and Technology (NIST).

[UGF06] Jay Urbain, Nazli Goharian, and Ophir Frieder. IIT TREC 2006: Genomics Track. In *Proceedings of the 15th Text REtrieval Conference TREC 2006*, Gaithersburg, Maryland, USA, 2006. National Institute of Standards and Technology (NIST).

[VMS$^+$09] Giulia Venturi, Simonetta Montemagni, Marchi Simone, Yutaka Sasaki, Paul Thompson, John McNaught, and Sophia Ananiadou.

Bootstrapping a Verb Lexicon for Biomedical Information Extraction. In A. Gelbukh, editor, *Proceedings of the 10th International Conference on Intelligent Text Processing and Computational Linguistics (CICLing 2009)*, pages 134–148. Springer, 2009.

[Voo94] Ellen M. Voorhees. Query Expansion Using Lexical-Semantic Relations. In *Proceedings of the 17th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR '94, pages 61–69, Dublin, Ireland, 1994.

[Vos96] Piek Vossen. Right or Wrong: Combining Lexical Resources in the EuroWordNet Project. In *Proceedings of Euralex-96 International Congress*, pages 715–728, 1996.

[Vos98] Piek Vossen. *EuroWordNet: a Multilingual Database with Lexical Semantic Networks.* Kluwer Academic Publishers, Norwell, MA, USA, 1998.

[Vos99] Piek Vossen. EuroWordNet as a Multilingual Database. In Wolfgang Teubert, editor, *TWC*. Amsterdam: Vrije Universiteit, 1999.

[Vos02] Piek Vossen. WordNet, EuroWordNet and Global WordNet. *Revue Française de Linguistique Appliquee/RFLA*, VII(1), 2002.

[vR79] C. J. van Rijsbergen. *Information Retrieval.* Butterworths, 2nd edition, 1979.

[WA10] Xinkai Wang and Sophia Ananiadou. A Task-Oriented Extension of Chinese MeSH Concepts Hierarchy. In *Proceedings of the 2nd Workshop on Building and Evaluating Resources for Biomedical Text Mining*, pages 23–30, Malta, 2010.

[Wil45] Frank Wilcoxon. Individual Comparisons by Ranking Methods. *Biometrics Bulletin*, 1(6):80–83, 1945.

[WL06] Jinglong Wang and Xiaojun Liang. 非参数统计分析 *(in Chinese).* Higher Education Press, Beijing, P. R. China, 2006.

[WL08] Ke Wu and Bao-Liang Lu. A Refinement Framework for Cross-Language Text Categorization. In *Proceedings of the 4th Asian*

Information Retrieval Conference on Information Retrieval Technology, AIRS '08, pages 401–411, 2008.

[XC00] Jinxi Xu and William Bruce Croft. Improving the Effectiveness of Information Retrieval with Local Context Analysis. *ACM Transactions on Information Systems*, 18(1):79–112, 2000.

[XW05] Jinxi Xu and Ralph Weischedel. Empirical Studies on the Impact of Lexical Resources on CLIR Performance. *Information Processing & Management*, 41(3):475–487, 2005.

[YHL09] Xiaoshi Yin, Xiangji Huang, and Zhoujun Li. Towards a Better Ranking for Biomedical Information Retrieval Using Context. In *IEEE International Conference on Bioinformatics & Biomedicine*, BIBM '09, pages 344–349, Los Alamitos, CA, USA, 2009. IEEE Computer Society.

[YL02] Christopher C. Yang and Kar Wing Li. Mining English/Chinese Parallel Documents from the World Wide Web. In *Proceedings of the 11th International World Wide Web Conference*, pages 188–192, Honolulu, Hawaii, 2002. ACM Press, New York.

[YY07] Song An Yuan and Song Nian Yu. A New Method for Cross-Language Information Retrieval by Summing Weights of Graphs. In *Proceedings of the Fourth International Conference on Fuzzy Systems Knowledge Discovery — Volume 02*, volume 02 of *FSKD '07*, pages 326–330, 2007.

[ZHL06] Hai Zhao, Chang-Ning Huang, and Mu Li. An Improved Chinese Word Segmentation System with Conditional Random Field. In *Proceedings of the 15th SIGHAN Workshop on Chinese Language Processing (SIGHAN-5)*, pages 162–165, Sydney, Australia, 2006.

[Zhu85] Dexi Zhu. *The Questions and Answers on Grammar (in Chinese)*. The Commercial Press, Beijing, 1985.

[ZL02] ChengXiang Zhai and John D. Lafferty. Two-stage Language Models for Information Retrieval. In *Proceedings of the 25th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR '02, pages 49–56, 2002.

[ZL04] ChengXiang Zhai and John D. Lafferty. A Study of Smoothing Methods for Language Models Applied to Information Retrieval. *ACM Transactions on Information System (TOIS)*, 22(2):179–214, 2004.

[ZLWF07] Xuezhong Zhou, Baoyan Liu, Zhaohui Wu, and Yi Feng. Itegrative Mining of Traditional Chinese Medicine Literature and MEDLINE for Functional Gene Networks. *Artificial Intelligence in Medicine*, 41(2):87–104, 2007.

[ZSDS00] Yibo Zhang, Le Sun, Lin Du, and Yufang Sun. Query Translation in Chinese-English Cross-Language Information Retrieval. In *Proceedings of the 2000 Joint SIGDAT conference on Empirical methods in natural language processing and very large corpora: held in conjunction with the 38th Annual Meeting of the Association for Computational Linguistics - Volume 13*, EMNLP '00, pages 104–109. Association for Computational Linguistics, 2000.

[ZTS06] Wei Zhou, Vetle I. Torvik, and Neil R. Smalheiser. ADAM: Another Database of Abbreviations in MEDLINE. *Bioinformatics*, 22(22):2813–2818, 2006.

[ZVZ05] Ying Zhang, Phil Vines, and Justin Zobel. Chinese OOV Translation and Post-translation Query Expansion in Chinese-English Cross-lingual Information Retrieval. *ACM Transactions on Asian Language Information Processing (TALIP)*, 4(2):57–77, 2005.

[ZW06] Jiang Zhu and Haifeng Wang. The Effect of Translation Quality in MT-Based Cross-Language Information Retrieval. In *ACL-44 Proceedings of the 21st International Conference on Computational Linguistics and the 44th annual meeting of the Association for Computational Linguistics*, ACL-44, pages 593–600. Association for Computational Linguistics, 2006.

[ZY06] Wei Zhou and Clement Tak Yu. TREC Genomics Track at UIC. In *Proceedings of the 15th Text REtrieval Conference TREC 2006*, Gaithersburg, Maryland, USA, 2006. National Institute of Standards and Technology (NIST).

Appendix A

Linguistic Rules to Extract Candidate Chinese Terms

A.1 Keyword Sets

Set name	Keywords
stop	。, ？, ！, ?, !, . . . , …
ssym	/, —, :, ;, ，, 、, ：, ；, $, %, @
lsym	“, ‘, (, [, {, 《, 『, （, 〈, 「, 【, 〔
rsym	”, ’,),], }, 》, 』, ）, 〉, 」, 】, 〕
asym	\p{Han}, NUMB, -, %, (,), &, ·
inpt	性, 化, 式, 样, 特发, 型, 形, 质, 伴, 杂, 代, 基, 类, 特异, 脱, 去, 寡, 多, 亚, 半, 抗, 异构

Table A.1: Definitions of keyword sets (Part I)

Note: \p{Han} in the “asym” set is a regular expression, which matches the Hanzi characters defined in Unicode script. The range of Hanzi characters is as follows: from 0x4E00 to 0x9FFF, from 0x3400 to 0x4DBF, from 0x20000 to 0x2A6DF, from 0x2A700 to 0x2B73F, from 0x2B840 to 0x2B81F, and from 0xF900 to 0xFAFF.

Set name	Keywords
nsym	这, 那些, 哪里, 那里, 这里, 这里, 但是, 可是, 然而, 而且, 导致, 可能, 或许, 应该, 不能, 禁止, 严禁, 可以, 不可, 症状, 病情, 病况, 病人, 科学, 科目, 科研, 种属
pref	又称, 俗称, 别称, 还称, 又叫, 还叫, 也叫, 也称, 称为, 就是, 成为, 亦称, 亦即, 别名, 即, 简称, 略称, 所谓, 所谓的, 记为, 译作, 译为, 记作, 简记, 写成, 写为, 同于, 属于

Table A.2: Definitions of keyword sets (Part II)

Set name	Keywords
numb	甲, 乙, 丙, 丁, 戊, 己, 庚, 辛, 壬, 癸, 零, 一, 二, 三, 四, 五, 六, 七, 八, 九, 十, 〇, 壹, 贰, 叁, 肆, 伍, 陆, 柒, 捌, 玖, 拾, 百, 千, 万, 亿, 佰, 0, 1, 2, 3, 4, 5, 6, 7, 8, 9, ０, １, ２, ３, ４, ５, ６, ７, ８, ９, Ⅰ, Ⅱ, Ⅲ, Ⅳ, Ⅴ, Ⅵ, Ⅶ, Ⅷ, Ⅸ, Ⅹ, Ⅺ, Ⅻ, I, II, III, IV, V, VI, VII, VIII, IX, X, XI, XII, ⅰ, ⅱ, ⅳ, ⅳ, ⅴ, ⅵ, ⅶ, ⅷ, ⅸ, ⅹ, ⅺ, ⅻ, i, ii, iii, iv, v, vi, vii, viii, ix, x, xi, xii, a, b, c, d, e, f, g, A, B, C, D, E, F, G, α, β, γ, δ, ϵ, ζ, η, θ, ι, κ, λ, μ, ν, ξ, π, ρ, σ, τ, υ, ϕ, χ, ψ, ω, A, B, Γ, Δ, E, Z, H, Θ, I, K, Λ, M, N, Ξ, Π, P, Σ, T, Υ, Φ, X, Ψ, Ω, alpha, beta, gamma, delta, epsilon, varepsilon, zeta, eta, theta, iota, kappa, lambda, mu, nu, xi, pi, rho, sigma, tau, upsilon, phi, chi, psi, omega, Alpha, Beta, Gamma, Delta, Epsilon, Varepsilon, Zeta, Eta, Theta, Iota, Kappa, Lambda, Mu, Nu, Xi, Pi, Rho, Sigma, Tau, Upsilon, Phi, Chi, Psi, Omega

Table A.3: Definitions of keyword sets (Part III)

Set name	Keywords
suff	复合征, 症, 病, 综合征, 综合症, 酶, 腈, 烃, 烯, 炔, 醚, 蛋白, 末端, 前段, 肌纤维, 关节, 细胞, 底膜, 基膜, 腺, 囊, 屏障, 鞘, 小管, 袢, 皮质, 髓质, 胶质, 脉, 腔, 窦, 系统, 束, 膜垫, 瓣, 核, 脑室, 神经节, 结节, 通路, 路径, 索, 神经元, 小体, 末梢, 感受器, 感知器, 丛, 连接, 接合处, 接点, 受体, 小球, 脉球, 脉体, 神经, 乳头, 盘, 叉, 前庭, 前部, 后庭, 基质, 房水, 内皮, 结膜, 状体, 上皮, 斑, 窝, 凹, 组织, 脂肪, 淋巴结, 淋巴集结, 亚群, 质体, 芽球, 异常, 血球, 细胞株, 细胞系, 桥粒, 微域, 微区, 联结, 质丝, 纤毛, 鞭毛, 纤毛, 绒毛, 染色体, 真菌, 细菌, 染色质, 包涵体, 周质, 胞质, 间隙, 单体, 核孔, 色素, 纤丝, 分裂器,

	载色体, 囊泡, 溶酶体, 溶菌体, 泡囊, 小泡, 内质网, 线粒体, 糖体, 孢子, 原虫, 滋养体, 脂, 血浆, 血清, 髓液, 内液, 朊, 胚, 胚层, 近交系, 小鼠, 大鼠, 蟾, 蛙, 螈, 显子, 基因, 岛, 斜带, 糖体, 亚基, 次单元, 门, 纲, 目, 科, 属, 亚门, 亚纲, 亚目, 鸭, 鸡, 鹅, 鹅, 鸽, 鹬, 鸠, 隼, 鹫, 鹰, 鹌鹑, 超目, 总目, 禽, 雀, 鹦鹉, 鲤, 鳝, 鳗, 鲡, 鲨, 鳐, 鲽, 鳕, 鲈, 鲷, 鲔, 鲑, 鳟, 鲻, 鲀, 羚, 羊, 驼, 麂, 绵羊, 盘羊, 羚羊, 山羊, 狸, 猞猁, 鬣狗, 貂, 鼬, 豚, 鲸, 鼹, 兔, 貂, 猴, 魈, 狒狒, 猩猩, 猿, 大鼠, 小鼠, 海牛, 松鼠, 猪, 恐龙, 蜥, 蛇, 蛭, 螨, 虱, 蜱, 蜘蛛, 虾, 鲺, 蚤, 蟹, 萤, 甲虫, 象虫, 粉甲, 蟑螂, 蠓, 蝇, 蚊, 蛉, 蚋, 跳蚤, 蚜虫, 蝽, 椿象, 蚁, 蝶, 蛾, 螂, 蜢, 蝗, 蟋蟀, 苔藓, 珊瑚, 海葵, 水母, 水螅, 荨麻, 海参, 瓜参, 人参, 西洋参, 海胆, 刺参, 党参, 海星, 星鱼, 线虫, 头虫, 鞭虫, 蛔虫, 蛲虫, 结虫, 丝虫, 涤虫, 蚴, 吸虫, 涡虫, 轮虫, 蚶, 蛤, 蚌, 贻贝, 蛎, 扇贝, 乌贼, 鱿, 斗鱼, 螺, 蛞蝓, 蜗牛, 海绵, 球虫, 囊虫, 胞虫, 孢子虫, 美虫, 孢虫, 弓形体, 焦虫, 浆虫, 梨形虫, 巴贝西虫, 巴贝虫, 泰勒虫, 袋虫, 毛虫, 草履虫, 膜虫, 纤虫, 仆虫, 黏菌, 绒泡菌, 藻, 眼虫, 衣藻, 滴虫, 贾第虫, 膜虫, 利什曼虫, 锥虫, 锥体虫, 阿米巴, 柄菌, 变形虫, 酵母菌, 纳氏虫, 海带, 小球藻, 地衣, 霉菌, 壶菌, 疫霉, 腐霉, 杆菌, 单胞菌, 卟啉菌, 普雷沃菌, 普氏菌, 绿菌, 蓝细菌, 丝菌, 念珠菌, 螺旋藻, 螺旋藻, 梭菌, 孢菌, 毛菌, 阴性菌, 阳性菌, 边虫, 心虫, 氏体, 螺菌, 螺旋体, 弯曲菌, 披衣菌, 衣原体, 嗜气菌, 需氧菌, 茎菌, 披毛菌, 酸菌, 阿菲菌, 阿菲波菌, 碱菌, 瘤菌, 氏菌, 血清群, 球菌, 厌氧菌, 厌气菌, 氏球菌, 氏杆菌, 胆菌, 弧菌, 色素菌, 色素菌, 包柔体, 弓菌, 沙门菌, 志贺菌, 发光菌, 甾原体, 浆菌, 支原体, 克次体, 放线菌, 噬菌体, 病毒, 麴霉, 曲霉, 孢霉, 壳菌, 角菌, 赤霉, 酵母, 膜霉, 珠霉, 蕈, 伞菌, 菇, 头孢, 顶孢, 僵菌, 虫霉, 莲, 菖蒲, 海棠, 龙舌兰, 丝兰, 腰果, 漆树, 紫玉盘, 茴香, 当归, 芹, 荽, 胡萝卜, 笋, 夹竹桃, 旋花, 茶花, 冬青, 天南星, 芋, 半夏, 五加, 藤, 楸, 田七, 槟榔, 椰, 马兜铃, 桐, 细辛, 靴叶, 杜仲, 菊, 蓟, 阿密, 蒡, 蒿, 苦艾, 菀, 苍术, 针草, 盏花, 红花, 丽花, 曲草, 麴草, 葵, 苣, 葱, 千里光, 蒲公英, 款冬, 锁阳, 凤仙花, 檗, 淫羊藿, 桦, 榛, 薇, 木棉, 芥菜, 拟南芥, 蒜, 芸苔, 芜菁, 荠, 菘蓝, 遏蓝菜, 梨, 铁兰, 乳香, 黄杨,

	仙人掌, 腊梅, 薯, 葎草, 忍冬, 接骨木, 蒴藋, 匏, 石竹, 秋罗, 麦瓶草, 繁缕, 蓝菜, 桑, 巧茶, 卫矛, 藜, 苋, 蔷薇, 菠, 榄仁, 跖草, 打碗花, 牵牛, 茱萸, 枣, 葫芦, 栝楼, 菟丝子, 莎草, 荸荠, 薯蓣, 柿, 杜鹃花, 橘, 油桐, 巴豆, 大戟, 蓖麻, 乌桕, 相思子, 合欢, 花生, 黄芪, 紫荆, 木豆, 黄豆, 绿豆, 肉桂, 锦鸡儿, 鹰咀豆, 嘴豆, 蝶豆, 猪屎豆, 金雀花, 扁豆, 刺桐, 皂荚, 槐蓝, 胡枝子, 苜蓿, 木犀, 黎豆, 菜豆, 豌豆, 黄皮, 紫檀, 葛, 槐, 三叶草, 胡芦巴, 蚕豆, 青冈, 榉, 堇, 龙胆, 老鹳草, 香树, 七叶树, 毛纲草, 八角, 核桃, 藿香, 巴洛草, 鞘蕊花, 山香, 薰衣草, 益母草, 薄荷, 紫苏, 黄芩, 水苏, 樟, 月桂, 茉莉, 连翘, 天麻, 肉苁蓉, 芍药, 罂粟, 车前草, 胡椒, 冰草, 麦, 竹, 薏苡, 香茅, 稗, 稻, 甘蔗, 狗尾草, 远志, 鱼腥草, 五味子, 甘草, 地黄, 椿, 茯苓, 辣椒, 曼陀罗, 枸杞, 茄子, 猕猴桃, 可可, 椴, 葡萄, 姜, 蕉, 蔻, 槲, 蒺藜, 蕨, 柏, 松, 杉, 微菌, 脓肿, 血症, 积脓, 感染, 体炎, 膜炎, 结核, 眼炎, 疽, 坏死, 氏病, 芽肿, 氏症, 肺炎, 伤寒, 热, 梅毒, 霍乱, 麻风, 溃疡, 癣, 丹毒, 痈, 疮, 松解症, 关节炎, 髓炎, 椎炎, 柱炎, 盘炎, 传染病, 组织炎, 毒症, 菌病, 体病, 沟炎, 疡, 耳炎, 虫病, 瘟, 脑炎, 管炎, 麻痹, 质炎, 神经炎, 痈, 痘, 巴瘤, 唇炎, 疱疹, 样疹, 痧, 瘤, 癌, 不良, 疣, 红斑, 白斑, 汗斑, 减少症, 肠炎, 畸形, 瘊, 天花, 肝炎, 腺炎, 犬病, 水病, 腔炎, 流感, 感冒, 鼻炎, 血病, 血症, 滋病, 肠病, 症候群, 麻疹, 风疹, 移行症, 虫症, 疟疾, 疖, 曼病, 疥, 囊肿, 硬化, 分化, 转移, 复发, 增生, 肥厚, 皮病, 障碍, 解症, 低下, 功能不足, 亢进, 甲减, 甲亢, 侏儒, 不全, 过距, 过远, 骨病, 发育异常, 化症, 肿大, 减退症, 肥大, 脱钙, 积病, 软化, 疏松, 机能低下, 佝偻, 流失, 骨丧失, 断症, 趾病, 脱位, 骨刺, 赘, 盘移位, 后凸, 后弯, 后突, 前凸, 前弯, 前突, 侧凸, 侧弯, 狭窄, 前移, 崩裂, 弓裂, 平足, 强直, 翻足, 内翻, 外翻, 腭裂, 病变, 病变, 风湿, 挛缩, 包炎, 囊炎, 痉挛, 积血, 脱臼, 积水, 积液, 肌炎, 肌溶解, 鞘炎, 腱压迫, 腱嵌入, 失调, 闭塞, 阻塞, 郁积, 梗阻, 郁滞, 结石, 膨出, 闭锁, 瘘, 反流, 逆流, 憩室, 曲张, 穿孔, 胃炎, 痢疾, 腹泻, 扭结, 套叠, 套迭, 息肉, 失禁, 痔, 痒病, 痒症, 脱垂, 脂肪肝, 昏迷, 压症, 卟啉症, 周炎, 萎缩, 横裂, 口炎, 龈炎, 龈病, 龈症, 腺病,

	腺症, 化生, 干燥病, 外流, 疳, 舌炎, 咽炎, 喉炎, 牙斑, 牙垢, 菌斑, 钙化, 错颌, 中毒, 咬合不正, 错颌, 龋, 哮喘, 撕脱, 水肿, 失音, 沉着症, 尘肺, 栓塞, 梗死, 梗塞, 窦炎, 气胸, 积气, 暂停, 气肿, 聋, 尿病, 尿症, 听障, 肌无力, 脑病, 外伤, 损伤, 沉积症, 匹克病, 缺乏, 累积病, 尿症, 尿病, 溢血, 痴呆, 血栓, 血肿, 出血, 癫痫, 头痛, 缺氧症, 内高压, 内低压, 失眠, 鲁病, 鲁症, 增多症, 压迫症, 震颤, 斜视, 失语, 晕厥, 过敏, 肾炎, 流产, 抑郁症, 停搏, 房颤, 早搏, 过速, 衰竭, 绞痛, 包炎, 脉炎, 贫血, 鳞病, 耐受症, 黄疸, 角化, 窒息, 蛋白病, 皮炎, 斑秃, 水疱, 骨折, 休克, 酸, 碱, 化物, 化合物, 酸盐, 酸钙, 化碳, 同位素, 胶体, 胺, 醇, 腺素, 必妥, 喘定, 心安, 酮, 酚, 唑, 苷, 萜, 酸钠, 醛, 酯, 聚糖, 酰, 霉素, 酸脂, 烷, 艻, 肟, 卡因, 酰胺, 脲, 嘧啶, 噻吩, 钾, 钠, 西林, 嘌呤, 吡嗪, 磺胺, 脒, 胍, 铵, 多巴, 他明, 萘, 吲哚, 胆碱, 酐, 呋喃, 杂环, 托品, 喹啉, 肼, 胡萝卜素, 维生素, 激素, 性素, 阻断剂, 抑制剂, 苯, 酞, 蒽, 醌, 芘, 替林, 茚, 苊, 萘, 氮芥, 糖金, 糖铁, 基汞, 汞, 铂, 磷, 腺素, 激素, 氟, 砜, 硫, 沙星, 菲啶, 非林, 沙芬, 吗啡, 诺啡, 吗喃, 衍生物, 咯啶, 茄素, 佐辛, 氮茂, 噁啉, 妥因, 妥英, 噁唑, 红素, 绿素, 黄素, 钴, 太尼, 洛丁, 吡啶, 乳糖, 葡萄糖, 比妥, 萆, 泮, 吡喃, 皮素, 豆素, 喹, 吖啶, 菌素, 氮, 糊精, 肽, 生素, 氨, 甾, 妥辛, 干扰素, 高辛, 米松, 松龙, 复合体, 蛋白, 拮抗剂, 调节剂, 泌素, 张素, 缩素, 瑞林, 长素, 质素, 因子, 酵素, 活化子, 化素, 介素, 配体, 抗原, 抗体, 着剂, 途径, 着素, 通道, 胶原, 菌苗, 疫苗, 标记物, 兴奋剂, 激动剂, 调节剂, 抑制剂, 拮抗剂, 阻断剂, 麻醉药, 镇静剂, 增进剂, 诱导剂, 神经剂, 镇痛剂, 孕药

Table A.4: Definitions of keyword sets (Part IV)

A.2 Rules Defining Keywords

Level	No.	Definition of Rule
1	1	`STOP := stop`
1	2	`SSYM := ssym`
1	3	`LSYM := lsym`
1	4	`RSYM := rsym`
1	5	`NUMB := numb`
1	6	`SUFF := suff`
1	7	`INPT := inpt`
1	8	`ASYM := asym`
1	9	`NSYM := nsym`
1	10	`PREF := pref`
2	1	`SUFF := ([SUFF][SUFF])`

Table A.5: The rules used to define keyword sets

where `:=` the operator, which includes keyword sets
`()` the operator, which indicates the interesting parts of a string
`[]` the operator, which get access to each element in a set

A.3 Rules Identifying the End of Terms

Level	No.	Definition of Rule
3	1	`^([ASYM]+[SUFF]):1:EOT`
3	2	`^([ASYM]+)[PREF]:1:EOT`
3	3	`([ASYM]+)[RSYM]:0:EOT`
3	4	`([[NUMB]+|[INPT]][[ASYM]+[[,|.][ASYM]+]|[INPT]]+) [PREF]|[NSYM]|[RSYM]:1:EOT`
3	5	`([[NUMB]+|[INPT]][[ASYM]+[[,|.][ASYM]+]|[INPT]]+ [SUFF]):1:EOT`
3	6	`([SUFF][NUMB]):1:EOT`

Table A.6: The rules used to determine the end of terms

where ^ the beginning of a sentence

[]+ the operator, which means an element in a set is repeated at least once

: the separator, which is used to separate the matching pattern, the scanning direction, and the tag

1 the right-to-left scanning direction

0 the left-to-right scanning direction

EOT the tag, meaning "end of term"

| the *OR* logic operator

A.4 Rules Detecting the Beginning of Terms

Level	**No.**	**Definition of Rule**
4	1	BOT:1:[PREF]([ASYM]+)
4	2	BOT:0:^|[SSYM]([ASYM]+)
4	3	BOT:0:[LSYM]([ASYM]+)
4	4	BOT:0:^|[PREF]|[NSYM]([[NUMB]+|[INPT]][[ASYM]+[[,|.][ASYM]+]|[INPT]]+)[PREF]|[NSYM]|[RSYM]
4	5	BOT:0:^|[PREF]|[NSYM]([[NUMB]+|[INPT]][[ASYM]+[[,|.][ASYM]+|[INPT]]]+[SUFF])
4	6	BOT:0:([SUFF][NUMB])

Table A.7: The rules used to identify the beginning of terms

where BOT the tag, meaning "beginning of term"

Appendix B

The 2006 and 2007 TREC Genomics Topic Sets

This appendix lists the topic sets used in the experiments reported in Chapter 4 and 5, and their Chinese translations as well.

B.1 The 2006 TREC Genomics Topic Set (English)

Query ID	**Query**
160	What is the role of PrnP in mad cow disease?
161	What is the role of IDE in Alzheimer's disease?
162	What is the role of MMS2 in cancer?
163	What is the role of APC (adenomatous polyposis coli) in colon cancer?
164	What is the role of Nurr-77 in Parkinson's disease?
165	How do Cathepsin D (CTSD) and apolipoprotein E (ApoE) interactions contribute to Alzheimer's disease?
166	What is the role of Transforming growth factor-beta1 (TGF-beta1) in cerebral amyloid angiopathy (CAA)?
167	How does nucleoside diphosphate kinase (NM23) contribute to tumor progression?
168	How does BARD1 regulate BRCA1 activity?
169	How does APC (adenomatous polyposis coli) protein affect actin assembly?

170	How does COP2 contribute to CFTR export from the endoplasmic reticulum?
171	How does Nurr-77 delete T cells before they migrate to the spleen or lymph nodes and how does this impact autoimmunity?
172	How does p53 affect apoptosis?
173	How do alpha7 nicotinic receptor subunits affect ethanol metabolism?
174	How does BRCA1 ubiquitinating activity contribute to cancer?
175	How does L2 interact with L1 to form HPV11 viral capsids?
176	How does Sec61-mediated CFTR degradation contribute to cystic fibrosis?
177	How do Bop-Pes interactions affect cell growth?
178	How do interactions between insulin-like GFs and the insulin receptor affect skin biology?
179	How do interactions between HNF4 and COUP-TF1 suppress liver function?
180	How do Ret-GDNF interactions affect liver development?
181	How do mutations in the Huntingtin gene affect Huntington's disease?
182	How do mutations in Sonic Hedgehog genes affect developmental disorders?
183	How do mutations in the NM23 gene affect tracheal development?
184	How do mutations in the Pes gene affect cell growth?
185	How do mutations in the hypocretin receptor 2 gene affect narcolepsy?
186	How do mutations in the Presenilin-1 gene affect Alzheimer's disease?
187	How do mutations in familial hemiplegic migraine type 1 (FHM1) gene affect calcium ion influx in hippocampal neurons?

Table B.1: The 2006 TREC Genomics Topic Set (English)

B.2 The 2006 TREC Genomics Topic Set (Chinese)

Query ID	Query
160	在疯牛病中PrnP的作用是什么?
161	在阿尔茨海默病中IDE的作用是什么?
162	在癌症中MMS2的作用是什么?
163	结肠癌中APC(结肠腺瘤性息肉病蛋白)的作用是什么?
164	在帕金森病中Nurr-77的作用是什么?
165	组织蛋白酶D(CTSD)和载脂蛋白E(ApoE)的反应如何促成阿尔茨海默病?
166	转化生长因子-β1(TGF-beta1)在淀粉样脑血管病(CAA)的作用是什么?
167	核苷二磷酸激酶(NM23)如何促成肿瘤发生?
168	BARD1如何调控BRCA1活性?
169	APC(结肠腺瘤性息肉病)蛋白如何影响肌动蛋白装配?
170	COP2如何促成从内质网导出CFTR?
171	Nurr-77如何在T细胞迁移到脾脏或淋巴结之前去除它们？这对自体免疫有什么影响?
172	p53如何影响细胞凋亡?
173	烟碱受体α7亚单位如何影响乙醇代谢?
174	BRCA1泛素化活性如何促成癌症?
175	L2如何与L1作用而形成HPV11病毒壳体?
176	Sec61介导的CFTR降解如何促成囊性纤维化?
177	Bop-Pes反应如何影响细胞生长?
178	胰岛素样生长因子与胰岛素受体间的反应如何影响皮肤生物学?
179	HNF4与COUP-TF1之间的反应如何抑制肝功能?
180	Ret-GDNF反应如何影响肝脏发育?
181	亨廷顿蛋白基因变异如何影响亨廷顿病?
182	Sonic Hedgehog基因变异如何影响发育障碍?
183	NM23基因变异如何影响气管发育?
184	Pes基因变异如何影响细胞生长?
185	下丘泌素受体2基因变异如何影响发作性睡眠症?
186	早衰蛋白-1基因变异如何影响阿尔茨海默病?

187	I型家族性偏瘫性偏头痛(FHM1)基因变异如何影响钙离子注入海马神经元?

Table B.2: The 2006 TREC Genomics Topic Set (Chinese)

B.3 The 2007 TREC Genomics Topic Set (English)

Query ID	Query
200	What serum [PROTEINS] change expression in association with high disease activity in lupus?
201	What [MUTATIONS] in the Raf gene are associated with cancer?
202	What [DRUGS] are associated with lysosomal abnormalities in the nervous system?
203	What [CELL OR TISSUE TYPES] express receptor binding sites for vasoactive intestinal peptide (VIP) on their cell surface?
204	What nervous system [CELL OR TISSUE TYPES] synthesize neurosteroids in the brain?
205	What [SIGNS OR SYMPTOMS] of anxiety disorder are related to coronary artery disease?
206	What [TOXICITIES] are associated with zoledronic acid?
207	What [TOXICITIES] are associated with etidronate?
208	What [BIOLOGICAL SUBSTANCES] have been used to measure toxicity in response to zoledronic acid?
209	What [BIOLOGICAL SUBSTANCES] have been used to measure toxicity in response to etidronate?
210	What [MOLECULAR FUNCTIONS] are attributed to glycan modification?
211	What [ANTIBODIES] have been used to detect protein PSD-95?
212	What [GENES] are involved in insect segmentation?
213	What [GENES] are involved in Drosophila neuroblast development?
214	What [GENES] are involved axon guidance in C.elegans?
215	What [PROTEINS] are involved in actin polymerization in smooth muscle?
216	What [GENES] regulate puberty in humans?
217	What [PROTEINS] in rats perform functions different from those of their human homologs?
218	What [GENES] are implicated in regulating alcohol preference?
219	In what [DISEASES] of brain development do centrosomal genes play a role?

220	What [PROTEINS] are involved in the activation or recognition mechanism for PmrD?
221	Which [PATHWAYS] are mediated by CD44?
222	What [MOLECULAR FUNCTIONS] is LITAF involved in?
223	Which anaerobic bacterial [STRAINS] are resistant to Vancomycin?
224	What [GENES] are involved in the melanogenesis of human lung cancers?
225	What [BIOLOGICAL SUBSTANCES] induce clpQ expression?
226	What [PROTEINS] make up the murine signal recognition particle?
227	What [GENES] are induced by LPS in diabetic mice?
228	What [GENES] when altered in the host genome improve solubility of heterologously expressed proteins?
229	What [SIGNS OR SYMPTOMS] are caused by human parvovirus infection?
230	What [PATHWAYS] are involved in Ewing's sarcoma?
231	What [TUMOR TYPES] are found in zebrafish?
232	What [DRUGS] inhibit HIV type 1 infection?
233	What viral [GENES] affect membrane fusion during HIV infection?
234	What [GENES] make up the NFkappaB signaling pathway?
235	Which [GENES] involved in NFkappaB signaling regulate iNOS?

Table B.3: The 2007 TREC Genomics Topic Set (English)

B.4 The 2007 TREC Genomics Topic Set (Chinese)

Query ID	Query
200	哪些血清蛋白改变与狼疮高疾病活性相关联的表达?
201	哪些Raf基因变异与癌症相关联?
202	哪些药物与神经系统溶酶体异常相关联?
203	哪些细胞或组织类型表达了在其细胞表面的血管活性肠肽(VIP)受体结合位点?
204	些神经系统的细胞或组织类型在脑内合成神经类固醇?
205	哪些焦虑障碍的体征和症状与冠心病有关?
206	哪些毒性与唑来膦酸关联?
207	哪些毒性与依替膦酸关联?
208	哪些生物物质被用来测量唑来膦酸响应毒性?
209	哪些生物物质被用来测量依替膦酸响应毒性?
210	哪些分子功能归因于聚糖修饰?
211	哪些抗体被用来检测蛋白质PSD-95?
212	哪些基因参与昆虫分节?
213	哪些基因参与果蝇成神经细胞发育?
214	哪些基因参与秀丽隐杆线虫的轴突导向?
215	哪些蛋白质参与平滑肌肌动蛋白的聚合?
216	哪些基因调控人类青春期?
217	哪些鼠类蛋白质执行和人类同系物不同的功能?
218	哪些基因涉及酒精偏爱的调节?
219	中心体基因在哪些脑部发育疾病中起作用?
220	哪些蛋白质参与PmrD的激活或识别机制?
221	哪些途径是由CD44介导的?
222	哪些分子功能是LITAF参与的?
223	哪些厌氧细菌菌株抗万古霉素?
224	哪些基因参与人类肺癌黑素生成?
225	哪些生物物质包含clpQ表达?
226	哪些蛋白质构成鼠科动物信号识别颗粒?
227	哪些基因是由糖尿病小鼠的LPS引入的?

228	当在宿主基因组中改变时, 哪些基因改善异源表达蛋白质的可溶性?
229	哪些体征和症状是由人类细小病毒感染引起的?
230	哪些途径参与尤文氏肉瘤?
231	在斑马鱼中发现了哪些肿瘤类型?
232	哪些药物抑制HIV1型感染?
233	在HIV感染过程中, 哪些病毒基因影响膜融合?
234	哪些基因构成NFkappaB信号途径?
235	参与NFkappaB信号传导的哪些基因调控iNOS?

Table B.4: The 2007 TREC Genomics Topic Set (Chinese)