Biomedical Chinese-English CLIR Using an Extended CMeSH Resource to Expand Queries

Xinkai Wang*, Paul Thompson†, Jun'ichi Tsujii‡, Sophia Ananiadou†

*School of Computer Science
University of Manchester, Manchester, UK
wangxa@cs.man.ac.uk

†School of Computer Science, National Centre for Text Mining
University of Manchester, Manchester, UK
{Paul.Thompson, Sophia.Ananiadou}@manchester.ac.uk

‡Microsoft Research Asia
Beijing, China
jtsujii@microsoft.com

Abstract

Cross-lingual information retrieval (CLIR) involving the Chinese language has been thoroughly studied in the general language domain, but rarely in the biomedical domain, due to the lack of suitable linguistic resources and parsing tools. In this paper, we describe a Chinese-English CLIR system for biomedical literature, which exploits a bilingual ontology, the "eCMeSH Tree". This is an extension of the Chinese Medical Subject Headings (CMeSH) Tree, based on Medical Subject Headings (MeSH). Using the 2006 and 2007 TREC Genomics track data, we have evaluated the performance of the eCMeSH Tree in expanding queries. We have compared our results to those obtained using two other approaches, i.e. pseudo-relevance feedback (PRF) and document translation (DT). Subsequently, we evaluate the performance of different combinations of these three retrieval methods. Our results show that our method of expanding queries using the eCMeSH Tree can outperform the PRF method. Furthermore, combining this method with PRF and DT helps to smooth the differences in query expansion, and consequently results in the best performance amongst all experiments reported. All experiments compare the use of two different retrieval models, i.e. Okapi BM25 and a query likelihood language model. In general, the former performs slightly better.

Keywords: cross-lingual information retrieval, biomedical information retrieval, query expansion, CMeSH

1. Introduction

Most studies on Chinese-English CLIR are focussed on the newswire domain, since linguistic resources and parsing tools designed for this domain are readily available. In contrast, there is a lack of comparable resources and tools for the biomedical domain. In this paper, we describe our approach to biomedical Chinese-English CLIR, using a bilingual MeSH-like ontology to expand queries. To our knowledge, this constitutes the first effort at tackling this problem.

Resources based on the Medical Subject Headings (MeSH) ontology have been widely applied in information retrieval (IR) tasks, for example, Guo et al. (2004), Lu et al. (2009), Abdou and Savoy (2007), Qin and Feng (1999), and Li et al. (2001). However, the Chinese translation of MeSH, i.e. the Chinese Medical Subject Headings (CMeSH) ontology, has rarely been used in CLIR tasks, not only because it is not freely available, but also since CMeSH lacks synonymous terms and term weights, both of which can help to improve retrieval performance. We developed the eCMeSH Tree (Wang and Ananiadou, 2010), which extends the CMeSH Tree by incorporating both synonyms and term weights. In this study, we explore the utility of the eCMeSH Tree in improving the performance of Chinese-English CLIR, through the expansion and translation of queries.

The performance of our approach is compared with two other methods of improving CLIR, i.e. query expansion based on pseudo-relevance feedback (PRF) and document translation (DT). Our results demonstrate that retrieval using the eCMeSH Tree can outperform the PRF method. Additionally, we investigate the improvements in retrieval performance that can be obtained when the three methods are combined in different ways. Our experiments show that the best results are achieved when all three methods are used in combination.

All experiments are conducted using both a probabilistic model (Okapi BM25) (Robertson et al., 1992) and a language model (query likelihood language model (Ponte and Croft, 1998) with Jelinek-Mercer smoothing (Zhai and Lafferty, 2001)). The Lemur toolkit [1] has been used to construct the retrieval system. The document collection is the 2006 and 2007 TREC Genomics Collection. We compare the differences in retrieval performance attained when manual and automatic word segmentation are applied, and discuss the potential drawbacks of using the PRF and DT approaches.

2. Related work

Biomedical CLIR is challenging due to the complex and inconsistent terminology used in biomedical text. Previous approaches aimed at improving biomedical IR tasks (including CLIR) can be summarised as follows:

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

Linguistic approaches Several attempts have been made to improve biomedical CLIR through the incorpation of various resources, such as MeSH terms (Abdou and Savoy, 2007; Hersh et al., 2007), UMLS (Hersh et al., 2007), the Gene Ontology (Hersh et al., 2007), and Entrez gene database (Hersh et al., 2007). In addition, a number of studies have investigated how the linguistic processing steps involved in CLIR can be adapted to the biomedical domain. The steps include tokenization strategies (Jiang and Zhai, 2007; Trieschnigg, 2010), stemming (Zhou and Yu, 2006), and techniques to process numbers, hyphens and parentheses in biomedical texts (Büttcher et al., 2004).

Feedback approaches Relevance feedback methods have been used to develop high-performance biomedical IR (Lin, 2008; Yin et al., 2009; Smucker, 2006; Huang et al., 2007).

Improvement of retrieval models Several approaches have concentrated on enhancing retrieval models by adjusting parameters or integrating additional processing. Abdou and Savoy (2006) evaluate both the Okapi BM25 model and the InB2 probabilistic model derived from the *Divergence from Randomness* paradigm and they conclude that the latter model performs better than the Okapi model. Trieschnigg et al. (2010) take a cross-lingual IR perspective to monolingual biomedical information retrieval. They view the mismatch between terms used in a query and terms used in relevant documents in the monolingual IR task as a cross-lingual matching problem.

Some of the major problems faced by CLIR systems operating on the Chinese language concern out-of-vocabulary (OOV) words and translation ambiguity. In terms of attempts to solve the OOV problem, Zhang et al. (2005) propose an approach that exploits the juxtaposition of English text and Chinese text on the web, while Lu et al. (2002) find web pages written in different languages that have hyperlinks pointing to a common page, in order to find potential translations of words. Yang and Li (2002) successfully mine parallel Chinese-English documents from the Web to find the appropriate translations for OOV words, and Chen and Nie (2000) process aligned English-Chinese documents from the Web. To address the problem of translation ambiguity, Gao et al. (2002) apply an improved co-occurrence approach to disambiguate dictionary-based translation. Zhang et al. (2005) use a hidden Markov model (HMM) with distance factor and window size to facilitate disambiguation. Zhang et al. (2000) use a mutual information value matrix to select an English translation, instead of looking up the translation in a Chinese-English dictionary.

3. The eCMeSH Tree

3.1. Overview of CMeSH

CMeSH is published by The Institute of Medical Information of the Chinese Academy of Medical Sciences, and consists of three parts: a Chinese translation of MeSH, traditional Chinese medical subject headings, and Special Classification for Medicine of China Library Classification. CMeSH includes only the translations of each MeSH heading term, its scope note, which consists of several short sentences, and some of the entry terms. To date, there has been little research on improving the performance of CLIR using CMeSH terms. Qin and Feng (1999) used CMeSH terms to improve the indexing quality of Chinese abstracts from 1977 concerning family planning and gynecology. Li et al. (2001) developed a monolingual information retrieval system with the help of CMeSH terms. The reasons that very few studies have explored the use CMeSH to improve IR are likely to be as follows: 1) MeSH terms do not have term weights assigned to them. As the Chinese translation of the original MeSH, CMeSH inherits this limitation. Moreover, 2) in CMeSH, each English MeSH heading term has one and only one Chinese translation. Furthermore, only a subset of the entry terms has been translated, and some of the entry terms belonging to the same tree node are assigned the same Chinese term.

Table 1 illustrates the MeSH Tree terms and their counterparts in the CMeSH Tree. The text before each semicolon is a term, while the part after the semicolon corresponds to the node number in the tree; the relations between terms are represented by the nestedness of the tree node numbers. The translated CMeSH entry terms are not shown in the table, since we use the version of the CMeSH tree that is freely available on the Internet (See Section 3.2.), which only provides heading terms.

The MeSH Tree	The CMeSH Tree
Dementia;C10.228.140.380	痴呆;C10.228.140.380
AIDS Dementia Complex;C10.228.140.380.070	艾滋病痴呆复合征;C10.228.140.380.070
Alzheimer Disease;C10.228.140.380.100	阿尔茨海默病;C10.228.140.380.100
.....	.....

Table 1: Sample MeSH Tree terms and corresponding CMeSH Tree terms

In order to enhance the utility of the CMeSH Tree as a resource to improve biomedical IR system, we previously extended the original CMeSH Tree with synonyms of terms and term weights (Wang and Ananiadou, 2010). We refer to this extended tree as the *eCMeSH Tree*.

3.2. CMeSH Extension Algorithm

Our previous work (Wang and Ananiadou, 2010) provides a detailed discussion of the algorithm used to extend the CMeSH Tree. In the current study, we have enhanced the algorithm, by adding mutual information (MI) filtering after C-value (Frantzi et al., 2000) extraction, as shown in Figure 1, and by connecting MeSH entry terms to eCMeSH Tree terms, as exemplified in Figure 2. The reason for introducing MI filtering is so that irrelevant characters that are affixed or suffixed to some of the terms extracted by C-value method are removed.

Figure 1 shows the workflow used to extend the CMeSH Tree. Firstly, the English MeSH Tree terms are aligned with terms extracted from the version of the CMeSH Tree that is freely available on the Internet [2]. This consists of a list of

[2] http://www2.chkd.cnki.net/kns50/Dict/

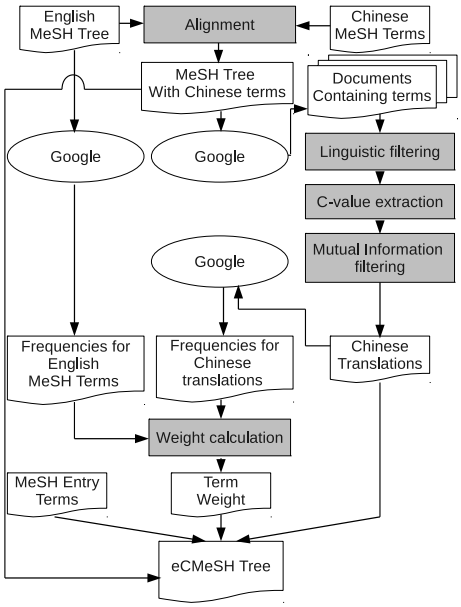

Figure 1: The workflow of extension of CMeSH Tree

Chinese keywords, most of which are translations of original English MeSH terms. However, the online version of CMeSH contains some terms that do not appear in the original MeSH. Terms that have not been aligned are ignored in subsequent processing steps. Secondly, using each pair of aligned terms as search terms, relevant documents in both English and Chinese are retrieved using Google. Thirdly, the retrieved Chinese documents are processed to extract alternative translations (i.e. synonyms) of the original term, through sequential application of the following: a) linguistic rules (discussed below), which identify text segments potentially containing synonyms, b) C-value (Frantzi et al., 2000), which extracts candidate translations from the identified text segments, and c) mutual information filtering, which refines the candidate translations by removing affixes or/and suffixes of terms. Fourthly, the frequencies of each English term and Chinese translation in the documents retrieved by Google are calculated; term weights are computed according to these frequencies, using Equation 1. Finally, using the information gathered from the steps above, the CMeSH Tree is extended to form the new eCMeSH Tree. Figure 2 provides an example of a node in the eCMeSH Tree. Each node includes equivalent heading terms in both languages (shown in boxes). Each heading term has several synonyms. For Chinese, these are the synonyms that were automatically extracted using the steps described above, together with their calculated weights. For English, we have added the MeSH entry terms, which were not included in the original CMeSH Tree.

The linguistic rules used to identify potential synonyms of Chinese terms are extensions of standard regular expressions. Definition rules are firstly used to define a number of sets of keywords that may indicate the suffixes or affixes of Chinese terms. Then, two layers of rules are applied to determine both boundaries of potential terms, using these keyword sets. Finally, the characters between boundaries are extracted as synonymous terms.

$$w_{ct} = \begin{cases} w + 1.0 & \text{if } f_{ct} > f_{et} > 0, \\ w & \text{otherwise.} \end{cases} \quad (1)$$

$$w = e^{-e^{-\frac{\log_{10}\left((f_{ct}+0.5)/(f_{et}+0.5)\right)}{2}}}$$

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

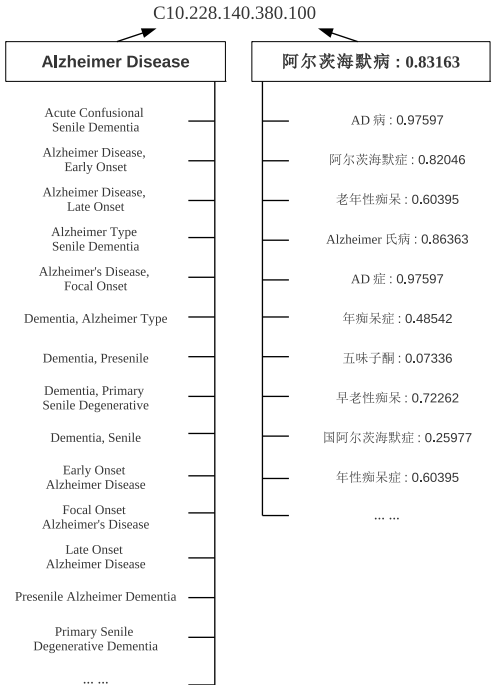

Figure 2: An example of eCMeSH Tree

4. CLIR Using Individual Methods

This section provides details and results of the experiments conducted to evaluate the impact of the three individual methods introduced above (i.e. query expansion using both the eCMeSH Tree and PRF, and the DT method) on the performance of CLIR. Prior to performing the experiments, the queries from the 2006 and 2007 TREC Genomics tracks were manually translated into Chinese and then segmented into words using both manual (Manual WS) and automatic (Automatic WS) methods. Automatic WS was carried out using *BaseSeg* (Zhao et al., 2006). After segmentation, query terms were filtered using the following rules: 1) terms without Chinese characters, such as *P53*, are retained in the query; 2) words including punctuation, like the

dict_list.aspx?firstLetter=A (accessed on 02/03/2011)

names of organic compounds, such as "1-(4-氟苯基)-1,3-二氢-5-异苯并呋喃腈" (citalopram), are retained as query terms; otherwise, 3) punctuation like ";" (semicolon), "。" (full-wide stop mark), "," (half-wide comma), and so on is erased from the query terms; and 4) terms are removed if they are not nouns or noun phrases, verbs (except link and auxiliary verbs), or adjectives.

Unless otherwise stated, documents in the document collection are processed by removing HTML tags, and then indexed using a word indexing policy. Okapi BM25 (abbreviated as "BM25") and the query likelihood language model (abbreviated as "LM") are applied to all experiments. Experimental results are measured in terms of *mean average precision* (MAP). Bold numbers indicate the best performing method within each set of experiments.

To statistically determine whether or not a given retrieval approach is better than another, we applied a two-sided *t-test*; the null hypothesis H_0 states that all the retrieval methods being tested are equivalent in terms of performance. The significance level α is equal to 5%. Retrieval methods whose performance is significantly different from the baseline approach are marked with "*".

4.1. Baseline

The baseline system uses an online bilingual dictionary, "the Google and Kingsoft Dictionary 2.0" [3], henceforth referred to as "the Dictionary", to translate Chinese query terms into English. No expansion of queries is used. The translation policy is as follows: 1) If a term has more than one entry in the Dictionary, only the first entry is selected as the translation of that term. 2) Terms without translations are ignored.

The results of baseline experiments are illustrated in Tables 3 and 4 and marked as "Baseline". In the experiments which evaluate the performance of PRF methods, illustrated in Table 2, the baseline experiment is PRF with the Dictionary (marked as "PRF-D"). The percentages in brackets in each of the tables show the difference in performance of the various experiments from the baseline experiments.

4.2. Query Translation Using CMeSH Tree terms

In order to evaluate the improvement attained when using the eCMeSH Tree terms, it is firstly necessary to determine the retrieval performance obtained when using the original CMeSH Tree terms. In this set of experiments, the original CMeSH Tree terms are used to translate (but not expand, since the original CMeSH Tree terms do not include Chinese synonyms) the Chinese query terms into their English equivalents using the following criteria: 1) If a Chinese term is found in the CMeSH Tree, then its corresponding English MeSH heading term is used to replace this Chinese term. 2) If a term cannot be found in the CMeSH Tree, then the Dictionary is used to translate it. 3) If several terms in the query have identical translations, then duplicate translated terms are removed. 4) All untranslatable Chinese terms are ignored, but acronyms or abbreviations written in Latin characters within the original Chinese query are retained.

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

In Tables 3 and 4, the results of the experiments conducted using the original CMeSH tree are indicated as "CMeSH".

4.3. Query Expansion Using the eCMeSH Tree

Experiments exploiting the eCMeSH Tree involve query expansion, according to the following criteria: 1) If a Chinese query term is found in the eCMeSH Tree, any Chinese terms belonging to sibling and child nodes are sorted into a list according to their term weights; the top 20 terms are selected and added to the original query along with their weights. 2) Terms which are not found in the eCMeSH Tree are ignored, except for those without Chinese characters, which are retained in the query, given their likelihood of representing terms. 3) Query terms without a weight (e.g. acronyms written in Latin characters) are assigned a term weight of $1/N$, where N is the total number of query terms (after the word filtering step and before expansion) in a given query.

After expansion, queries are translated using the eCMeSH Tree and the Dictionary: 1) If a Chinese term is found in the eCMeSH Tree (either as "heading term" or one of its synonyms), then its English counterparts in the eCMeSH Tree are used to replace this Chinese term. These counterparts consist of the equivalent English "heading term" and all of its "entry 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 is assigned the term weight of the original term. 3) If several terms in the query have identical translations, then duplicate translations are removed. 4) All untranslatable terms are ignored.

Two sets of experiments were performed, one in which query terms were assigned weights from the eCMeSH tree, as described above (shown as "eCMeSH-W" in Table 3 and 4), and one in which term weights were ignored (shown as "eCMeSH-N" in Tables 3 and 4).

4.4. Query Expansion Using Pseudo-Relevance Feedback

In this set of experiments, the eCMeSH Tree is not used; rather, as a baseline, the Dictionary is used to translate the Chinese query terms into English using a term-by-term translation policy. This baseline is compared to the results obtained when the CMeSH Tree and the eCMeSH Tree are used to carry out the translation. The initial term weight of each term is assigned as $1/N$, where N is the total number of query terms in a certain query (after the word filtering step and before expansion).

Pseudo-relevance feedback (Xu and Croft, 1996) provides an automatic approach to analysing the most relevant documents from those returned by an initial search. The PRF functionality built into *Indri*, the index and retrieval engine of the Lemur toolkit, is applied to expand the translated English query terms. 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 terms extracted from these documents (ranked based on term frequencies) as the terms that will provide the most useful expansion of the original query. The weights used to adjust original query terms and the terms resulting from the

	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	0.2205*	**0.3009**	0.2771	0.2546	**0.2765**	0.2379	0.2193*	**0.3178**	0.2763	0.2539*
		(-12.68%)	(-19.44%)		(-7.91%)	(-15.39%)		(-13.96%)	(-20.69%)		(-13.06%)	(-20.11%)
2007	**0.1654**	0.1275*	0.1085*	**0.2154**	0.1699*	0.1483*	**0.1591**	0.1268*	0.1079*	**0.2123**	0.1691*	0.1477*
		(-22.91%)	(-34.40%)		(-22.52%)	(-31.15%)		(-20.30%)	(-32.18%)		(-20.35%)	(-30.43%)

Table 2: Effects of resource quality on retrieval performance of query expansion using PRF

application of the relevance feedback method are both 0.5. Other parameters of Indri's PRF are configured using their default values.

Our results show that the performance of query expansion using PRF depends on the quality of the linguistic resources that are used to translate queries. We have conducted experiments to compare the retrieval based on PRF using different linguistic resources: the Dictionary (abbreviated as "PRF-D" in Table 2), the eCMeSH Tree terms ("PRF-e"), and the CMeSH Tree terms ("PRF-C"). The results are shown in Table 2, illustrating that the best retrieval performance is achieved when the Dictionary is used to translate the terms. A possible reason for this result is that the Dictionary contains appropriate translations for query terms that are not domain specific. However, it should be noted that using the eCMeSH Tree to perform translation obtains better results than when the original CMeSH Tree terms are used. Since the use of the Dictionary achieves the best results, the PRF experiments shown in Tables 3 and 4 use the Dictionary to perform the translation of the terms.

4.5. Document Translation

In these experiments, the Google translation service [4] is used to translate the document collection into Chinese before retrieval. The translated Chinese documents are indexed using the bigram indexing policy; thus the processed queries are also separated as bigrams. These experiments do not use dictionaries or ontologies to translate or expand the Chinese queries; moreover, no term weights are assigned to the query terms. The results of document translation experiments are shown as "DT" in Tables 3 and 4.

5. CLIR Using Hybrid Approaches

We conducted a further set of experiments, in an attempt to improve CLIR performance using hybrid approaches based on combinations of the methods described in Section 4.. Query expansion using the eCMeSH Tree is combined with PRF and DT in different ways, in order to evaluate their respective contributions to CLIR.

The pre-processing of the document collection and query sets used are the same as those described in Section 4., unless otherwise stated. Table 5 compares the retrieval performance of these hybrid approaches.

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

This two-stage query expansion approach is carried out as follows: 1) The eCMeSH Tree is firstly applied to expand and translate queries, as described in Section 4.3., 2) PRF is subsequently applied, as described in Section 4.4., to further expand the query, using the Dictionary to translate the terms.

In Table 5, this approach is denoted using "e+PRF". This experiment is taken as the baseline of the hybrid methods.

5.2. Query Expansion Using the eCMeSH Tree with Document Translation

In this experiment, all the documents in document collection are first translated into Chinese and indexed using bigrams, as explained in Section 4.5.. Then, the eCMeSH Tree is applied to expand Chinese queries; there is no need to translate queries, because the document collection has already been translated into Chinese. In Table 5, the results of this approach are denoted using"e+DT", which are compared with the results of "e+PRF".

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

Since the translated document collection is represented as bigrams (see Section 4.5.), the results returned from Indri are also bigrams, not terms. However, this hybrid method requires individual terms, in order match them against terms from the eCMeSH Tree. Thus, this set of experiments uses a modified version of the PRF method described in Section 4.4.. Here, the set of candidate terms is extracted from the relevant documents using the same term extraction algorithm used in creating the eCMeSH Tree (described in Wang and Ananiadou (2010)), based on linguistic rules and C-value term extraction. The top ranked 25 terms extracted using this method, which do not appear amongst the terms in the query expanded using the eCMeSH Tree, selected from the top 50 relevant documents, are added to the original queries. The *tf-idf* measure is used to calculate the appropriate weights for final query terms chosen by the PRF method. In Table 5, The results of these experiments are denoted as "e+PRF+DT". The results are compared with those obtained using both the "e+PRF" method (shown using "△" in the table) and the "e+DT" method (shown using "◇" in the table).

6. Discussion

6.1. Retrieval Improvements Obtained Using the eCMeSH Tree

In Tables 3 and 4, the best retrieval performance achieved using the eCMeSH method is 0.3058 for the 2006 Track and 0.1901 for the 2007 Track. The performance of this

[4] http://translate.google.com/

	Automatic WS						Manual WS					
	Baseline	CMeSH	eCMeSH-N	eCMeSH-W	PRF	DT	Baseline	CMeSH	eCMeSH-N	eCMeSH-W	PRF	DT
2006	0.2309	0.1976 (-3.33%)	0.2503 (8.40%)	0.2647 (14.64%)	0.2737 (15.03%)	**0.2985*** (29.27%) (9.06%)^{D1}	0.2622	0.2503 (-4.54%)	0.2857 (8.96%)	0.3058 (16.63%)	0.3009 (13.04%)	**0.3368*** (28.45%) (11.93%)^{D1}
2007	0.1353	0.0911* (-32.67%)	0.1435 (6.06%)	0.1415 (4.58%)	0.1654 (22.25%)	**0.1800*** (33.04%) (8.83%)^{D1}	0.1735	0.1344 (-22.54%)	0.1813 (4.50%)	0.1901 (9.57%)	0.2154* (24.15%)	**0.2305*** (32.85%) (7.01%)^{D1}

Table 3: Experimental results using Okapi BM25 for retrieval

	Automatic WS						Manual WS					
	Baseline	CMeSH	eCMeSH-N	eCMeSH-W	PRF	DT	Baseline	CMeSH	eCMeSH-N	eCMeSH-W	PRF	DT
2006	0.2278	0.1935 (-15.06%)	0.2497 (9.61%)	0.2390 (4.92%)	0.2765 (21.38%)	**0.2791** (22.52%) (0.94%)^{D1}	0.2619	0.2418 (-7.67%)	0.2842 (8.51%)	0.2925 (11.68%)	0.3178 (21.34%)	**0.3216** (22.79%) (1.20%)^{D1}
2007	0.1330	0.0789* (-40.68%)	0.1341 (0.83%)	0.1375 (3.38%)	0.1591 (19.62%)	**0.1683*** (26.54%) (5.78%)^{D1}	0.1695	0.1154* (-31.92%)	0.1799 (6.14%)	0.1899 (12.04%)	0.2123 (12.04%)	**0.2257*** (33.16%) (6.31%)^{D1}

Table 4: Experimental results using the language model for retrieval

method on the 2006 Track exceeded the performance of the PRF method, when Okapi BM25 was used. Although document translation and, in most cases, PRF, produce better performance than the use of the eCMeSH Tree, they suffer from a number of drawbacks, such as the following: 1) Document translation is computationally expensive and thus it is not suitable for cross-lingual information retrieval where documents are added or removed frequently, or the content of documents is subject to change. According to our experiments, for example, it takes about four months to translate entire the 2006 TREC document collection (162,259 articles, about 11.9GB) into Chinese, when a computer equipped with a 1.44GHz Intel Dual Core CPU and 3.0 GB memory is used. 2) The quality of the linguistic resources that are used to translate queries plays an important role in the CLIR performance of query expansion using PRF. Table 2 compares the retrieval performance attained when different resources are applied to assist query expansion when using PRF approach. According to the table, the best performance is achieved when translation is carried out using the Dictionary on the 2006 Track data (0.3009). When the eCMeSH Tree terms are applied to translate Chinese queries into English, the retrieval performance decreases by 7.91%, to 0.2771. However, since there is a further considerable decrease in the retrieval performance when the CMeSH Tree is used instead of the eCMeSH Tree (15.39% less than when the Dictionary is used), our results clearly show that the eCMeSH tree can have a positive effect on retrieval performance. Whilst the Dictionary may have a wider coverage of query terms that are not domain specific, and hence achieves slightly superior performance to the eCMeSH tree when used on its own, the eCMeSH tree can help to provide a greater number of translations for domain specific terms. This is illustrated by the higher retrieval performance (0.3304) obtained on the same dataset when PRF using the Dictionary is combined with the eCMeSH tree (e+PRF in Table 5). In Table 5, it can be observed that the best performing hybrid method is the combination of the eCMeSH Tree with PRF and document translation. This achieves the best retrieval results of all experiments on both the TREC 2006 Track (0.3782) and the 2007 Track (0.2524). Moreover, it can be observed from the table that the eCMeSH Tree smooths the differences in performance between the PRF and the DT approaches. Consider the results obtained for the 2006 Track, using Okapi BM25 and manual word segmentation. Table 3 shows that the difference in performance between the DT and PRF methods (shown as "D1" in the table) is 11.93%. However, after combining these approaches with the use of eCMeSH Tree terms, the difference between "e+DT" and "e+PRF" (marked as "D2" in Table 5) is 6.72%. In this case, D2 is 43.67% smaller than D1, which further demonstrates the valuable contribution made by the eCMeSH Tree terms, in that they are able to reduce the differences in performance between various approaches to CLIR. In all but a few cases, the combination of the eCMeSH tree with document translation (e+DT) results in improvements over the combination of the eCMeSH tree with PRF (e+PRF). We do not discuss the differences between the results obtained using the "e+PRF+DT" and "e+PRF" configurations, because the PRF method described in Section 5.3., is different from that used in the "e+PRF" experiment, described in Section 5.1., meaning that a direct comparison is not possible.

6.2. Other Factors Effecting Retrieval

In the majority of our experiments show, Okapi BM25 retrieval model is slightly superior to the language model. For instance, the retrieval performance achieved on the 2006 Track with the "eCMeSH-W" configuration, using manual word segmentation, is 0.3058 with Okapi BM25 and 0.2925 with the language model. Thus, compared with the

	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 (-5.22%)^{D2}	**0.3018** (2.20%)$^{\triangle}$ (7.82%)$^{\diamond}$	0.3304	0.3526 (6.72%)^{D2}	**0.3782** (14.47%)$^{\triangle}$ (7.26%)$^{\diamond}$	0.2941	0.2890 (-1.73%)^{D2}	**0.2973** (1.09%)$^{\triangle}$ (2.87%)$^{\diamond}$	0.3278	0.3239 (-1.19%)^{D2}	**0.3779** (15.28%)$^{\triangle}$ (16.67%)$^{\diamond}$
2007	0.2095	0.2100 (0.24%)^{D2}	**0.2172** (3.68%)$^{\triangle}$ (3.43%)$^{\diamond}$	0.2375	0.2401 (1.10%)^{D2}	**0.2514** (5.85%)$^{\triangle}$ (4.71%)$^{\diamond}$	0.2064	0.2083 (0.92%)^{D2}	**0.2103** (1.89%)$^{\triangle}$ (0.96%)$^{\diamond}$	0.2349	0.2366 (0.72%)^{D2}	**0.2497** (6.30%)$^{\triangle}$ (5.54%)$^{\diamond}$

Table 5: Comparisons of hybrid approaches

language model, the use of Okapi BM25 improves the retrieval performance by 4.55%. However, in a small number of cases, the language model achieves superior performance. This is the case for the 2006 Track data, when the "PRF" method is used, in conjunction with manual word segmentation, where the retrieval performance is 0.3178.

According to our experiments, the automatic segmentation tool, which is trained using a newswire corpus, has a significantly negative impact on the retrieval results, compared to the use of manual segmentation. As an example, the retrieval performance of the "eCMeSH-W" configuration on the 2006 Track data, using the language model, decreases from 0.2925 (manual word segmentation) to 0.2390 when automatic word segmentation is used, i.e. a drop in performance of 22.38%.

Our experiments also show that the use of the weights assigned to the eCMeSH Tree terms helps to improve the performance of CLIR. In Tables 3 and 4, a comparison of the experiments using eCMeSH terms without weights (eCMeSH-N) with those in which the weights are used (eCMeSH-W) reveals that weights improve performance in all cases.

All experiments illustrate that there is a significant difference between the retrieval performances on the 2006 Track data and the 2007 Track data. For the the 2007 Track, the queries consist of a set of short questions for a question and answering task. In contrast, the queries in the 2006 Track are declarative sentences describing the information request. After the application of filtering of the queries to remove unnecessary words, the number of terms remaining in the queries for the 2007 track is much smaller than for queries in the 2006 Track, due to the removal of interrogatives from the 2007 queries. Since the 2007 queries are more diverse in terms of query types, and they are also more general than the declarative sentences in the 2006 Track, this makes retrieval more difficult, and leads to the drop in retrieval performance on the 2007 Track.

7. Conclusions

In this paper, we have described the application of a Chinese-English CLIR system to biomedical articles. Query expansion using the eCMeSH Tree was compared with query expansion based on pseudo-relevance feedback, and document translation. Different combinations of these individual approaches to CLIR were also investigated. In terms of individual methods, the overall retrieval performance achieved using the eCMeSH query expansion method is comparable to that of pseudo-relevance feedback query expansion approach.

For the most part, the results achieved by the hybrid approaches are better than those achieved by individual approaches. Combining the two methods of query expansion, i.e. the use of the eCMeSH Tree and pseudo-relevance feedback, resulted in superior retrieval performance, compared to the individual use of these methods. Furthermore, when these two methods are further combined with document translation, the best results amongst all experiments are achieved.

Our experiments show that the strategy for segmenting terms has a significant impact on the retrieval performance, i.e. manual word segmentation significantly outperforms the automatic approach. Since the automatic approach was based on a segmenting tool trained on the newswire domain, further research in needed into adapting or developing segmentation tools for the biomedical domain. Finally, our results show that the Okapi BM25 model performs slightly better than the language model.

8. References

S. Abdou and J. Savoy. 2006. Report on the TREC 2006 Genomics Experiment. In *Proceedings of the 15th Text REtrieval Conference (2006)*, Gaithersburg, Maryland, USA. National Institute of Standards and Technology (NIST).

S. Abdou and J. Savoy. 2007. Searching in MEDLINE: Query Expansion and Manual Indexing Evaluation. *Information Processing and Management*, 44(2):781–789.

S. Büttcher, C. L. A. Clarke, and G. V. Cormack. 2004. Domain-Specific Synonym Expansion and Validation for Biomedical Information Retrieval. In *Proceedings of the 13th Text REtrieval Conference, TREC-2004*, Gaithersburg, Maryland, USA. National Institute of Standards and Technology (NIST).

J. Chen and J. Y. Nie. 2000. Parallel Web Text Mining for Cross-Language IR. In *Proceedings of RIAO-2000: Content-Based Multimedia Information Access*, pages 188–192. CollCge de France, Paris, France.

K. Frantzi, S. Ananiadou, and H. Mima. 2000. Automatic Recognition of Multi-word Terms: the C-value/NC-value Method. *International Journal of Digital Library*, 3(2):117–132.

J. Gao, M. Zhou, J. Y. Nie, H. He, and W. Chen. 2002. Resolving Query Translation Ambiguity Using a Decaying Co-Occurrence Model and Syntactic Dependence Rela-

tions. In *Proceedings of the 25th Annual International ACMSIGIR Conference on Research and Development in Information Retrieval*, SIGIR '02, pages 183–190, Tampere, Finland. ACM Press, New York.

Y. Guo, H. Harkema, and R. Gaizauskas. 2004. 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. National Institute of Standards and Technology Special Publication 500-261.

W. Hersh, A. Cohen, L. Ruslen, and P. Roberts. 2007. TREC 2007 Genomics Track. In *Proceedings of the 16th Text REtrieval Conference (TREC-16)*, Gaithersburg, Maryland, USA. National Institute of Standards and Technology (NIST).

X. Huang, D. Sotoudeh-Hosseini, H. Rohian, and X. An. 2007. York University at TREC 2007: Genomics Track. In *Proceedings of the 16th Text REtrieval Conference*, TREC-16, Gaithersburg, Maryland, USA. National Institute of Standards and Technology (NIST).

J. Jiang and C. X. Zhai. 2007. An Empirical Study of Tokenization Strategies for Biomedical Information Retrieval. *Information Retrieval*, 10(4-5):341–363.

D. Li, T. Hu, W. Zhu, Q. Qian, H. Ren, J. Li, and B. Yang. 2001. Retrieval System for the Chinese Medical Subject Headings (in Chinese). *Chinese Journal of Medical Library*, 4.

J. Lin. 2008. Pagerank without Hyperlinks: Reranking with PubMed Related Article Networks for Biomedical Text Retrieval. *BMC Bioinformatics*, 9(270).

W.-H. Lu, L.-F. Chien, and H.-J. Lee. 2002. Translation of Web Queries Using Anchor Text Mining. *ACM Transactions on Asian Language Information Processing (TALIP)*, 2(1):159–172.

Z. Lu, W. Kim, and W. J. Wilbur. 2009. Evaluation of Query Expansion Using MeSH in PubMed. *Information Retrieval*, 12(1):69–80.

J. M. Ponte and W. B. Croft. 1998. 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.

Y. Qin and C. Feng. 1999. 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.

S. E. Robertson, S. Walker, M. Hancock-Beaulieu, A. Gull, and M. Lau. 1992. Okapi at TREC. In D. K. Harman, editor, *Proceedings of Text REtrieval Conference (TREC)*, pages 21–30, Gaithersburg, MD. National Institute of Standards and Technology Special Publication 500-207.

M. D. Smucker. 2006. UMass Genomics 2006: Query-Biased Pseudo Relevance Feedback. In *Proceedings of the 15th Text REtrieval Conference (TREC 2006)*, Gaithersburg, Maryland, USA. National Institute of Standards and Technology (NIST).

D. Trieschnigg, D. Hiemstra, F. de Jong, and W. Kraaij. 2010. 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. ACM.

D. Trieschnigg. 2010. *Proof of Concept: Concept-Based Biomedical Information Retrieval*. Ph.D. thesis, University of Twente.

X. Wang and S. Ananiadou. 2010. 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.

J. Xu and W. B. Croft. 1996. Query Expansion Using Local and Global Document Analysis. In *Proceedings of the 19th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR '96, pages 4–11, New York, NY, USA. ACM.

C. C. Yang and K. W. Li. 2002. 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. ACM Press, New York.

X. Yin, X. Huang, and Z. Li. 2009. 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. IEEE Computer Society.

C. X. Zhai and J. D. Lafferty. 2001. A Study of Smoothing Methods for Language Models Applied to Ad-hoc Information Retrieval. In *Proceedings of the 24th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 334–342, New Orleans, Louisiana, United States.

Y. Zhang, L. Sun, L. Du, and Y. Sun. 2000. 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.

Y. Zhang, P. Vines, and J. Zobel. 2005. 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.

H. Zhao, C.-N. Huang, and M. Li. 2006. 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.

W. Zhou and C. T. Yu. 2006. TREC Genomics Track at UIC. In *Proceedings of the 15th Text REtrieval Conference TREC 2006*, Gaithersburg, Maryland, USA. National Institute of Standards and Technology (NIST).