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/