A Task-Oriented Extension of the Chinese MeSH Concepts Hierarchy

Xinkai Wang and Sophia Ananiadou
National Centre for Text Mining
School of Computer Science
University of Manchester
United Kindom
E-mail: wangxa@cs.man.ac.uk, Sophia.Ananiadou@manchester.ac.uk

Abstract

The Medical Subject Headings (MeSH) thesaurus is used not only as a vocabulary for indexing and cataloguing biomedical articles, but also constitutes an important linguistic resource for text mining in the biomedical domain. As part of our research, the Chinese translation of English MeSH, Chinese Medical Subject Headings (CMeSH), will be used to provide essential terms and concepts to a Chinese-English cross-lingual information retrieval system. The original MeSH uses a small, accurate and concise set of terms designed for indexing and cataloguing. Such a set of terms does not, however, lend itself well to information retrieval applications, in which the ability to expand queries using term synonyms is important to maximise relevant search results. In this paper, we propose a new approach to extending the MeSH concept hierarchy (the MeSH Tree) based on an online version of the CMeSH term list. We use Google to collect synonyms for Chinese terms and calculate weights for each Chinese term and its synonyms. Our extension has been evaluated on a Chinese-English biomedical information retrieval application. The results indicate that the extension of CMeSH improves the performance of the information retrieval application. Furthermore, the extended resource should also be helpful in other related research.

1. Introduction

The Medical Subject Heading (MeSH) thesaurus, which was developed and released by the National Library of Medicine (http://www.nlm.nih.gov/mesh/), is widely accepted as the standard vocabulary used for indexing, cataloguing, and searching for biomedical and health-related information and documents. For example, Lowe and Banett (1994) report using MeSH to index medical literature. Cooper and Miller (1998) 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, many researchers (Guo et al, 2004; Abdou and Savoy, 2007; Lu et al, 2008) have employed MeSH terms to evaluate or improve biomedical information retrieval applications. In addition, MeSH terms are treated as the standard vocabulary to which terms from other resources are mapped (Elkin et al, 1988; Shultz, 2006). MeSH vocabulary has also been employed in the construction of Chinese medical ontologies. Zhou et al. (2007) attempt to discover novel gene networks and functional knowledge of genes using a significant bibliographic literature database of traditional Chinese medicine. In their research, MeSH disease headings are applied to generate the index data for gene and disease MEDLINE literature.

As part of own research, we have made use of a MeSH-related resource, i.e., Chinese Medical Subject Headings (CMeSH) to evaluate and improve the performance and effectiveness of a Chinese-English biomedical cross-lingual information retrieval application.

CMeSH, which has been translated and is maintained by The Institute of Medical Information of the Chinese Academy of Medical Sciences, retains the terms and concepts of the English MeSH and their relations. 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 (1999) apply CMeSH terms to improve the indexing quality of Chinese abstracts from 1977 concerning family planning and gynecology, whilst Li et al. (2001) develop an information retrieval system with the help of CMeSH terms. The main reason why MeSH terms have not been more widely adopted in Chinese biomedical text processing lies in the philosophy of MeSH design. As MeSH terms are intended to index and catalogue the biomedical literature, they must be represented succinctly, concisely, and accurately. The Chinese translation of MeSH, CMeSH, inherits these features. This means that the original CMeSH terms have no synonyms - each English term has one and only one Chinese translation.

In order to expand the coverage of terms in CMeSH, and also to make it a more useful resource for our research, we have extended the original CMeSH with synonyms and term weights, and have integrated the extended set of terms into the MeSH concept hierarchy, i.e. the MeSH Tree. An online version of the CMeSH (http://www2.chkd.cnki.net/kns50/Dict/dict_list.aspx?firstLetter=A) is the starting point of our work. We have designed an algorithm which exploits the Google search engine to automatically collect synonyms of each original CMeSH term and calculate the frequency of each extracted term. Based on these frequencies, we have defined a formula to compute a weight for each Chinese term. The corresponding English term's weight is not computed, for the reasons discussed in Section 4. Finally, a Chinese-English cross-lingual information retrieval system (CLIR), based on the Lemur toolkit, has been employed to evaluate the extended CMeSH Tree.

2. Background

2.1 MeSH

MeSH consists of a controlled vocabulary, coupled with a hierarchical tree structure. The controlled vocabulary contains several types of concepts, namely Publication types, Geographics, Qualifiers, Descriptors, and Supplementary Concept Records.

- Publication types or Publication characteristics are used to indicate the genre of the indexed item, rather than its contents, e.g., '*Historical Article'*.
- Geographics include continents, regions, countries, and other geographic subdivisions; they are not used to characterize subject content but rather physical location.
- Descriptors are the main concepts or headings. They are used to index the catalogue, and to search biomedical documents. Examples include '*Dementia*' and '*Carcinoma in Situ*'.
- Qualifiers, also known as Subheadings, are used for indexing and cataloging in conjunction with Descriptors. There are 95 Qualifiers (MeSH 2009), which provide a convenient means grouping together those documents which are concerned with a particular aspect of a subject. For example, '*Liver/drug effects*' indicates that the article or book is not about the '*liver*' in general, but about the effect of drugs on the liver.
- Supplementary Concept Records (SCRs) are used to index chemicals, drugs, and other concepts. Unlike Descriptors, SCRs have no tree numbers (see below).

The MeSH Tree is a hierarchy of MeSH descriptors, in which each descriptor is allocated a tree number, which represents the position of the node in the tree. The following example illustrates the structure and organization of the MeSH Tree.

```
……
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
……
```

On each line, the text before the semi-colon constitutes a MeSH term. In the remainder of the paper, we refer to these as 'English terms'. After the semi-colon, the string starting with Latin letter and followed by digits and dots represents a tree number, which encodes the term's position within the tree. The version of the MeSH Tree used in this study is the MeSH Tree 2008, which has 24,763 unique terms and 48,442 tree nodes.

2.2 CMeSH

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 official CMeSH contains three parts: a Chinese translation of MeSH, traditional Chinese medical subject headings and Special Classification for Medicine of China Library Classification. The usual usage of CMeSH is to index and catalogue biomedical literature in a library, or to provide standard keywords to describe journal articles and conference papers.

An online version of the CMeSH term list is available at: http://www2.chkd.cnki.net/kns50/Dict/dict_list.aspx?firstLetter=A. The example below illustrates the Chinese counterpart of the English MeSH example presented above.

```
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
```

In contrast to research achievements using the original MeSH, the usage of CMeSH is currently largely limited to acting as a gold standard for indexing and cataloging biomedical documents or for assigning indexing terms in IR systems. There is very little work that reports on evaluating cross-lingual information retrieval with CMeSH or on improving information extraction via CMeSH terms. Analyzing the social or economical factors which limit the usage of CMeSH is the duty of economists; our focus is on the deficiencies of the original CMeSH, based on our task-oriented requirements. From the above example, we can conclude that:

- There are no term weights for CMeSH terms.
- Each English term has one and only one Chinese translation.

Term weights are essential to text mining or NLP algorithms based on probabilistic and statistical models. Without term weights, CMeSH can thus function only as a traditional word list. In the cross-lingual information retrieval task, our experiments have shown the high degree to which term weights contribute towards the improvement of retrieval performance (see section 5.2). Another issue of the original CMeSH is that many Chinese translations are missing. 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', '阿滋海默症', '老年性痴呆', or 'Alzheimer 氏病'. 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.

The research described in this paper attempts to overcome the above-mentioned issues of the original CMeSH. Firstly, Google is used to collect web pages which may contain candidate translations. Following this, linguistic rules and a term extraction tool are applied to identify candidate terms from these web pages, and the frequency of each term is calculated. Finally, each term's weight is computed according to frequency of the term and that of its English equivalent.

3. Related Work on Ontology Evaluation

The extended CMeSH Tree constitutes an ontology. Evaluating the effectiveness of this ontology is a critical step of our research. 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:

- Testing the ontology in an application and evalutating the result (Porzel and Malaka, 2004); also called application-based evaluation;
- Comparing the ontology to a 'gold standard' (Maedche and Staab, 2002);
- Human evaluation of the ontology according to a set of predefined criteria, standards, requirements, etc. (Lozano-Tello and Gómez-Pérez, 2004);
- Comparing the ontology with a set of data (e.g., a collection of documents) from the domain to be covered by the ontology (Brewster et al., 2004); also called data-driven evaluation.

Evaluation of ontologies in general is carried out at three basic levels: vocabulary, taxonomy, and (non-taxonomic) semantic relations. We are not intending to evaluate the *isa* hierarchy (taxonomy) and the non-taxonomic relations (semantic relations) of the extended CMeSH Tree, because our work does not add new tree nodes to the MeSH concept hierarchy. Moreover, based on the fact that MeSH Tree, as a part of The Unified Medical Language System (UMLS), has been assessed by human experts against a set of criteria (Kumar and Smith, 2003; Smith, 2006), our evaluation of the extended CMeSH Tree will serve only to evaluate the enhanced ontology vocabulary. In order to do this, we have tested the extended CMeSH tree within a CLIR application.

4. Extension of CMeSH

The Figure 1 illustrates the workflow of the process of extending CMeSH. In this chart, the dashed lines represent the steps of obtaining the frequencies of English terms, while solid lines correspond to the steps of extending the original CMeSH, including computing the weights of Chinese translations. The details of the algorithm are explained in the subsections that follow.

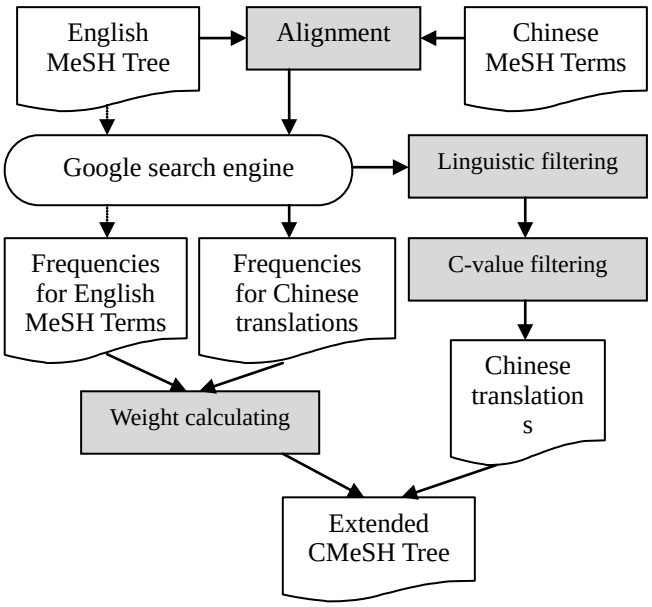

Figure 1: Extension workflow

4.1 Alignment

Alignment is the operation by which the English MeSH Tree terms are matched with the corresponding Chinese MeSH terms. In our experiments, we found that approximately 3% of English terms in the MeSH Tree had no translation in the online CMeSH term list, and that about 8.1% of Chinese terms had no matching MeSH tree terms. In order to resolve this issue, both English terms without Chinese translations, and Chinese CMeSH terms without English counterparts were ignored in the subsequent steps of processing. Following alignment, the MeSH Tree has the following appearance.

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;C10.228.140.380.16 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

4.2 Linguistic Filtering

Each Chinese term in the aligned MeSH tree was queried using Google. This caused a set of relevant documents to be returned, most of which were written in Chinese. In order to extract candidate terms from the returned documents, a set of two-level linguistic rules were applied to filter the document set and to discover candidate terms.

The rules take the form of regular expressions, which are based on the following linguistic characteristics in Chinese biomedical texts.

1). Most Chinese terms have suffixes. In the above example, '-复合征' (meaning 'complex'), '-综合征' (meaning 'syndrome'), and '-病' (the general name for all kinds of disease) are all suffixes.

2). Some Chinese terms contains 'inner' keywords, which can help to identify terms. For example, in '失语, 原发进行性' and '痴呆, 血管性' (Their normal formats are '原发进行性失语' and '血管性痴呆' respectively.), '-性' is an important character that can indicate, when used between two adjacent verbs and nouns, indicates that the first word describes the term after it, thus indicating a high probability of the presence of a term.

3). Compared to the clear suffixes of Chinese terms, it can be more difficult to recognize the start of a term from a stream of characters. Fortunately, terms are often followed by synonyms, which are often indicated using a particular set of phrases. For instance, in the sentence of '阿尔茨海默病(Alzheimer disease, AD), 又称早老性痴呆，……', '又称', which means 'that is' or 'i.e.', can be a good identifier to determine the beginning of the term. Other similar phrases are '简称' (abbreviated as), '也叫' (that is), '也称' (that is), '还叫' (that is), '叫做' (named as or called), etc.

4). Some symbols (e.g. brackets and parentheses) can play the role of delimiters which define the boundaries of term. In the sentence of '…可以减少人们患早老性痴呆(阿尔茨海默氏症)的危险，…', the phrase between parentheses is a term, whose meaning is 'Alzheimer Disease'. Such symbols, may, however, 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. Without special rules, the extracted candidates should be '4-氟苯基' and '1-(4-氟苯基)-1', which are clearly incorrect and not terms. Thus, it is necessary to apply constraints to rules that exploit these symbols.

5). Many Chinese terms start with an English word. For example, '阿尔茨海默病' can also be written as 'Alzheimer 症' or 'AD 症'.

According to these linguistic features, we firstly define four word lists:

- **SUFF** This list contains 347 suffixes, such as '-复合征', '-病', '-腈', '-烃', and etc.
- **SYMB** This is a list of symbols which may function as delimiters, e.g. '(', ')', '[', ']', ',', ' ', '。', and etc.
- **PREF** This list defines the phrases which are considered as prefix indicators, like '又称', '别名', '还叫', etc. Note that these phrases themselves cannot be one part of a term.
- **INPT** This is a list of the special Chinese characters or words whose appearance in a phrase indicates a high probability that the phrase is a term. For example, '-性-', '-化-', '-式-', '-特发-', and etc. are included in the list.

Twenty-three regular expression rules have been constructed based on these four sets of characters. These rules can be grouped into two levels: the first level rules, using maximum length matching strategy, are employed to extract the terms whose penultimate symbols are in SUFF or which are followed by the symbols in the SYMB list. The second level rules, based on the result of the first level rules, determine the start points of candidate terms.

4.3 C-value Filtering

C-value (Frantzi and Ananiadou, 2000) is a simple but effective tool to extract terms, especially cascaded terms, from free texts. We use C-value to discover such cascaded terms and also to filter high-scoring candidate terms. The C-value algorithm requires syntactic features. However, in this study, we do not apply any POS tagging to the results of linguistic filtering. The reasons are: 1) POS taggers trained on Chinese biomedical corpora are not currently available, and taggers trained on newswire are likely to introduce errors and thus affect the performance of the tool. 2) The output of linguistic filtering consists of short phrases, most of which have already been identified as terms or parts of terms. Therefore, in the current work, each candidate term resulting from the linguistic filtering step is assigned the noun phrase POS tag. The maximum number of terms selected from the list is 20. After implementing the C-value processing with the above parameters, the members of the resulting list are considered as the synonyms of the original CMeSH terms.

4.4 Term Weight Calculation

In this study, only Chinese term weights are calculated. This is because the purpose of the extended CMeSH tree is to provide an enhanced set of Chinese terms to improve a Chinese-English CLIR application. Queries are translated or/and expanded using CMeSH terms, and Chinese term weights are directly passed to the translated English query terms. The original weights of English terms, if assigned, were not used in this study.

The weight formula used is a variant of the one proposed by Lynam et al. (2001).

$$w_{ct} = \begin{cases} w'+1.0, & \text{if } f_{ct} > f_{et} > 0 \\ w', & \text{otherwise} \end{cases}$$

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

where f_{ct} is the frequency of Chinese translation and f_{et} is the frequency of English term. Both frequencies correspond to the number of occurrences returned by Google. The weight of the Chinese translation can be computed by the sigmoid function w'. If the frequency of the Chinese translation is greater than that of English term (which means that Chinese translation is more popular than the English equivalents), then we increase its weight.

4.5 The Final CMeSH Resource

After merging the weight values with Chinese and English terms, the final CMeSH Tree has the follwing representation.

Dementia;C10.228.140.380 痴呆:0.343881162222 痴呆症:1.425371472485 失智:0.314097771253
AIDS Dementia Complex;C10.228.140.380.070 艾滋病痴呆综合征:0.099850335887 AIDS 痴呆综合征:0.050615806911 AIDS 痴呆症候群:0.018721486519 艾滋病痴呆复合征:0.080638707889 艾滋病痴呆复合症:0.00000853004 爱滋病痴呆复合症:0.00000461272
Alzheimer Disease;C10.228.140.380.100 阿尔茨海默病:0.097398853027 Alzheimer 病:0.04592354867 阿滋海默症:0.18626155397 老年性痴呆:0.244575613782 Alzheimer 氏病:0.073412383701 早老性痴呆:0.074511778 Alzheimer 氏症:0.041794385
Aphasia, Primary Progressive;C10.228.140.380.132 失语, 原发性进行性:0.000000211796 原发性进行性失语:0.317609856764 原发性进行性失语症:0.208916619538
Creutzfeldt-Jakob Syndrome;C10.228.140.380.165 克-亚综合征:0.014768214 Creutzfeldt-Jakob 病:0.324075159688 Creutzfeldt-Jakob 综合征:0.001337330099 早老痴呆症:0.264388092697 克-雅氏综合征:0.005840557652 克-雅氏病:0.119031301389 克雅氏病:0.119031301389 库贾氏病:0.203074988059 牛海绵状脑病:1.398253302721 疯牛病:1.431304403242 皮质-纹体-脊髓变性:0.006199180668 克鲁兹弗得-雅柯病:0.002491919204 库雅氏症:0.029140724567
Dementia, Vascular;C10.228.140.380.230 痴呆, 血管性:0.226335681686 血管性痴呆:0.36594277919 血管梗塞型痴呆症:0.000230068283 血管型失智症:0.081200049502
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:0.00000043303 常染色体显性遗传病合并皮质下梗死和白质脑病:0.00000043303 CADASIL 病:0.02736615094 遗传性多发梗死痴呆病:0.003007920526 伴皮层下梗死和白质脑病的常染色体显性遗传性脑动脉病:0.014594966461 伴皮质下梗死和白质脑病的常染色体显性遗传性脑动脉病:0.01608595828 显性遗传脑动脉病伴皮层下梗死及脑白质病:0.00005469028

The extended CMeSH enriches the original resource with both synonyms and term weights. Moreover, Chinese terms and their synonyms are mapped to English MeSH terms with their tree number.

5. Evaluation

A CLIR system has been employed to evaluate the scope of vocabulary in the extended CMeSH Tree. The difference between cross-lingual information retrieval and mono-lingual information retrieval is that CLIR requires a stage in which either the queries are translated from the source language into the target language (in which the document set is written), or else the document set is translated into the language in which the queries are expressed. In this study, we translate Chinese queries into English and carry out information retrieval on English documents. The CMeSH Tree is used to translate or/and expand the Chinese query terms.

5.1 Experimental description

The toolkit used for constructing the CLIR system is Lemur (http://www.lemurproject.org/). The document collections are the TREC Genomics data from 2006 and 2007 (Hersh et al., 2006, 2007), which contain a total of 162,259 biomedical papers (11.9 GB). The indexing and retrieving algorithm is Okapi BM25; the parameters used for Okapi BM25 are the system default values. All documents are indexed for document level retrieval. Indexing of the TREC Genomics documents does not involve stemming. Stopwords are removed using the PubMed stop list (http://www.ncbi.nlm.nih.gov/bookshelf/br.fcgi?book=helppubmed&part=pubmedhelp&rendertype=table&id=pubmedhelp.T43).

TREC Genomics 2006 and 2007 tasks provide 64 sentences as queries and a gold standard relevance judgement, i.e., each query has been associated with a set of relevant documents in TREC Genomics data set by human experts, except Query 173 and Query 180. There are also scripts and utilities to calculate retrieval performance. In order to obtain the Chinese queries, we make use of the same strategy reported in Levow et al. (2004), in which the original queries are manually translated into other languages. Here, we manually translate TREC Genomics queries from English into Chinese. For instance, Query 161 in the 2006 task is:

'**What is the role of IDE in Alzheimer's disease**'.

The corresponding Chinese translation is as follows:

'在阿尔茨海默病中 IDE 的作用是什么'.

The Chinese query sentences need to be segmented before translating into them into English queries. As our work is intended to evaluate the effectiveness of CMeSH Tree terms rather than the performance of the information retrieval, and also that there is no high-performance POS tagger for the Chinese biomedical domain, we manually segment the Chinese queries into word sequence. This process ensures that errors are not introduced by automatic word segmentation. In the above example, the segmentation result is as follows:

'在 阿尔茨海默病 中 IDE 的 作用 是 什么'.

We then remove words from the query whose grammatical categories do not correspond to one of the

following: nouns or noun phrases, verbs (except link verbs and auxiliary verbs), and adjectives. Words without Chinese characters, like 'IDE', are also retained in the query. By carrying out a preliminary experiment, (i.e. performing mono-lingual information retrieval on TREC genomics data to compare the performances of two word selection policies: all words and selected words), we found that the above categories of words play a more important role on retrieving relevant documents than prepositions, adverbs, postpositions, interrogatives, etc. Following this filtering strategy, the above-mentioned example now contains the following words:

'阿尔茨海默病 IDE 作用'.

All the following experiments were carried out on queries which were segmented and filtered according to the above description.

The segmented and filtered Chinese queries are translated into English queries using either a domain dictionary or CMeSH Tree, depending on the experiment being undertaken (see next section). English queries are represented as *indri queries* using 'Indri query language' (Strohman, 2005), which is based on 'Inquery query language'. Finally, Lemur's Indri search engine retrieves relevant documents and computes the performance parameters, such as mean average precision (MAP) and average precision (AP).

5.2 Experiments

CMeSH Tree terms were applied to translate Chinese terms into English equivalents. The quality of CMeSH Tree is thus reflected in the performance of CLIR. To fully evaluate the extended CMeSH tree, we designed four experiments. The baseline experiment makes use of a domain dictionary to translate queries. The other three experiments are aimed to evaluating a) the CMeSH Tree terms themselves, b) term weights within the CMeSH Tree, and c) the CMeSH Tree hierarchy.

1. Baseline

For the baseline experiment, we employed a free domain dictionary, '谷歌金山词霸 2.0' (Google and Kingsoft Dictionary 2.0) (http://g.iciba.com/), to translate Chinese query terms into English counterparts. The policy for out-of-vocabulary (OOV) is to ignore all unknown words.

2. CMeSH Tree term translation

For this experiment, the extended CMeSH Tree terms were used to translate Chinese queries. Terms or words which were not present in the CMeSH Tree were ignored during translation. In the following discussion, this experiment is referred to as 'term_t'.

3. CMeSH Tree term translation with weights

This experiment, subsequently referred to as 'term_w', had the aim of evaluating the effectiveness of our term weighting algorithm. Whenever a Chinese term was found in the CMeSH Tree, the weight of that Chinese term was passed to the English translation. The Indri query consists of these translations and their weights. Here, we inherit the OOV processing policy used in the 'term_t' experiment.

4. CMeSH Tree terms translation with hierarchy expansion

In this experiment, we expanded Chinese queries according to the hierarchical structure of the CMeSH Tree. Spasić and Ananiadou (2005) refer to an algorithm to compute the tree similarity (TS).

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

where C_1 and C_2 are the classes related to Term 1 and Term 2 respectively, $common(C_1, C_2)$ denotes the number of common classes in the paths leading from the root to the given classes, and $depth(C)$ is the number of classes in the path connecting the root and the given class. $common(C_1, C_2)$ is subject to the following conditions in this study: Given that C_1 and C_2 denote classes of Term 1 and Term 2 respectively, if the 'common' function value is the depth of C_2, then Term 2 is the parent node of Term 1; the second condition indicates that Term 2 is the sibling of Term 1.

$$common(C_1, C_2) = \begin{cases} depth(C_2) \\ depth(C_1) - 1, \ where \ depth(C_1) = depth(C_2) \end{cases}$$

Using this constraint, the TS algorithm expands a Chinese query term only with its siblings and parent in the CMeSH Tree.

After expanding the original Chinese query terms, the CMeSH term list is used to translate the expanded queries into English. For this experiment, term weights are ignored, because it is intended to evaluate the ontology hierarchy. We name this experiment 'term_h'.

5.2 Results and Discussion

Table 1 illustrates the Mean Average Precision (MAP) of each experiment. This is the mean value of all queries' average precision.

	baseline	term_t	term_w	term_h
2006	0.2622	0.2857	**0.3014**	0.2706
2007	0.1735	0.1813	**0.1899**	0.1712

Table 1: MAPs for the four experiments

According to the results of experiment term_t, we can conclude that the extended CMeSH has a positive effect on IR results. Compared with baseline, the MAPs have been increased by 2.53% (for the 2006 task) and 0.78% (for the 2007 task) with the help of extended CMeSH terms. The experiment term_w proves that the term weighting algorithm can improve the performance of Chinese-English CLIR greatly – from baseline's results of 26.22% (2006 task) and 17.35% (2007 task) to 30.14% (2006 task) and 18.99% (2007 task) respectively. The reasons for these improvements are 1). Our CMeSH

extension provides more Chinese terms than the dictionary; 2). Our term weighting algorithm succeeds in assigning terms with reasonable weight value.

The results of the term_h experiment are not as expected. For the 2006 task, the MAP is only slightly better than that of baseline experiment, whilst for the 2007 task, it is the worst result of all four experiments. The reason for this bad performance is that our simple query expansion technique introduces too many terms into queries, which reduces the precision of the search engine.

Figure 1 shows the Average Precision (AP) for each query in all four experiments.

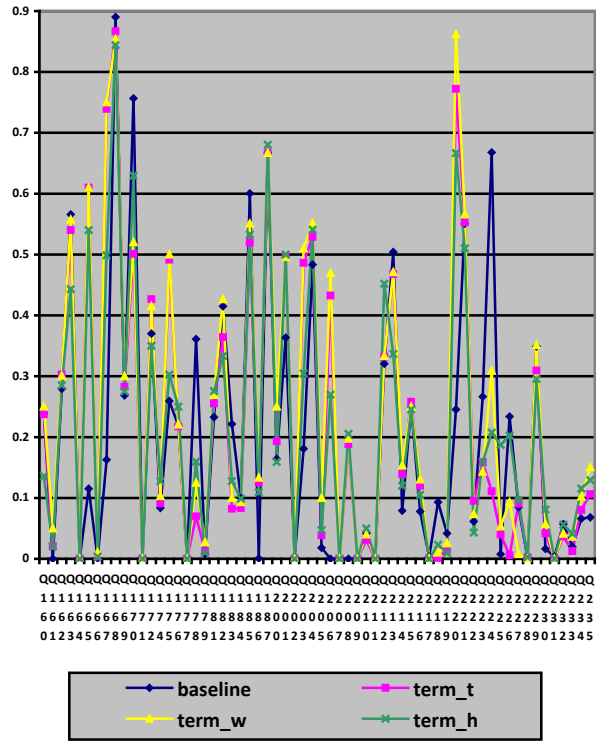

Figure 1: AP of each query in four experiments

Figure 1 provides another perspective on the CMeSH Tree extension. It illustrates how well the CMeSH Tree extension performs on each query. These results are consistent with those presented in Table 1. For example, in most cases, the curve of experiment term_w is higher than other curves, which indicates that CMeSH Tree term translation with term weight is the best result among experiments. This chart also gives the details of performances of each query. For instance, Query 224 obtains the best AP in baseline experiment; CMeSH terms highly decrease its AP by 55.67% (for term_t), 35.87% (for term_w), and 46.02% (for term_h), compared with the result of baseline. Meanwhile, Query 220 is greatly improved by CMeSH term translation with term weight, but the term_h strategy reduces its performance.

6. Conclusions

In this paper, we have proposed a task-oriented extension of the Chinese MeSH Tree. We have employed the Google search engine to collect Chinese synonyms and calculate weights for them. We have evaluated our extension to the tree using a Chinese-English cross-lingual information retrieval system in the biomedical domain. We have analysed the scope and effectiveness of the extended CMeSH Tree terms and their weights. The results of our experiments illustrate that the extended CMeSH Tree significantly improves the performance of the CLIR application. The enhanced performance of the CLIR serves to demonstrate the quality of the extended CMeSH Tree. It is intended that this new linguistic resource will also help others in future research.

Future work will include the following:

1. Ensuring that all terms in the original English MeSH tree have corresponding Chinese translations (a small number are still missing);
2. Computing the weights of English as well as Chinese terms;
3. Finding English synonyms of English MeSH terms;
4. Evaluating the performance of the extended CMeSH Tree in other NLP applications.

7. Acknowledgements

We sincerely thank Mr. Paul Thompson of the National Centre for Text Mining for his help with proof-reading and comments regarding this paper. We also thank the two anonymous reviewers for their helpful comments.

8. References

Abdou, Samir, and Savoy, Jacques, (2007). Searching in MEDLINE: Query expansion and manual indexing evaluation. *Information Processing and Management.* Volume 44, Issue 2, 2008, pp. 781--789.

Brewsster, C., Alani, H., Dasmahapatra, S., and Wilks, Y., (2004). Data driven ontology evaluation. In *Proceedings of International Conference on Langauge Resources and Evaluation, Lisbon*, pp. 24--30.

Cooper, Gregory F., and Miller, Randolph F., (1998). An Experiment comparing Lexical and Statistical Methods for Extracting MeSH Terms from Clinic Free Text. *Journal of the American Medical Informatics Association.* Volume 5, Issue 1, 1998, pp. 62--75.

Elkin, Peter L.; Cimino, James J.; Lowe, Henry J.; Aronow, David B.; Payne, Tom H.; Pincetl, Pierre S.; and Barnett, G. Octo, (1988). 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*, IEEE Comput Soc Press (1988), pp. 185--190.

Frantzi, Katerina, Ananiadou, Sophia, and Mima, Hideki, (2000). Automatic recognition of multi-word terms: the C-value/NC-value method. *International Journal of Digital Library*, 3(2), pp. 117--132.

Guo, Y., Harkema, H., and Gaizauskas, R., (2004). Sheffield University and the TREC 2004 genomics track: Query expansion using synonymous terms. In *Proceedings of the Thirteenth Text REtrieval*

Conference. Gaithersburg, MD: Department of Commerce, National Institute of Standards and Technology, pp. 16--19.

Hersh, W., Cohen, A. M., Roberts, P., and Rekapalli, H. K. (2006), TREC 2006 genomics track, In *Proceedings of the Fifteenth Text REtrieval Conference*, Gaithersburg, MD: Department of Commerce, National Institute of Standards and Technology.

Hersh, W., Cohen, A. M., Roberts, P., and Rekapalli, H. K. (2007), TREC 2007 genomics track overview, In *Proceedings of the Sixteenth Text REtrieval Conference*, Gaithersburg, MD: Department of Commerce, National Institute of Standards and Technology.

Kumar, A., and Smith, B., (2003). The Unified Medical Language System and the Gene Ontology: Some Critical Reflections. In *KI2003: Advances in AI (2003)*, pp. 135--148.

Levow, Gina-Anne, Oard, Douglas W., and Resnik, Philip, (2004). Dictionary-Base Techniques for Cross-Language Information Retrieval, *Information Processing & Management* 41(3), pp. 523--547.

Li, Danya, Hu, Tiejun, Zhu, Wenyan, Qian, Qing, Ren, Huiling, Li, Junlian, and Yang, Bin, (2001). Retrieval System for the Chinese Medical Subject Headings (in Chinese). *Chinese Journal of Medical Library*, Issue 4, 2001.

Lozano-Tello, A., and Gómez-Pérez, A., (2004). Ontometric: A method to choose the appropriate ontology. *Journal of Database Management*, 15(2), pp. 1--18.

Lowe, H. J., and Barnett, G., O., (1994). Understanding and using the medical subject headings (MeSH) vocabulary to perform literature searches. *Journal of the American Medical Association*, 271(14): pp. 1103--1108.

Lu, Zhiyong, Kim, Won, and Wilbur, W. John, (2008). Evaluation of query expansion using MeSH in PubMed, *Information Retrieval*, Volume 12, Issue 1, 2009, pp. 69--80.

Lynam, T. R., Clarke, C. L. A., and Cormack, G. V., (2001). Information Extraction with Term Frequencies. In *Proceedings of the First International Conference on Human Language Technology Research*, pp. 1--4.

Maeche, A., and Staab, S., (2002). Measuring similarity between ontologies. In *Proceedings of the 13th International Conference on Knowledge Engineering and Knowledge Management. Ontologies and the Semantinc Web*, pp. 251--263.

Porzel, R., and Malaka, R., (2004). A task-based approach for ontology evaluation. In *Proceedings of ECAI 2004 Workshop on Ontology Learning and Population*.

Qin, Yayun, and Feng, Qichang, (1999). 中文医学主题词表(机读版)在文献标引中的应用(in Chinese), *Journal of Medical Intelligence*, Issue 5, 1999.

Smith, B., (2006). From concepts to clinical reality: an essay on the benchmarking of biomedical terminologies, *Journal of Biomedical Informatics*, 39(3), pp. 288--298.

Shultz, M., (2006). Mapping of medical acronyms and initialisms to medical subject headings (mesh) across selected systems. *Journal of the Medical Library Association*, Volume 94, Issue 4, pp. 410--414.

Spasić, I. and Ananiadou, S., (2005). A Flexible measure of contextual similarity for biomedical terms. *Pacific Symposium on Biocomputing 10*, pp. 197--208.

Strohman, T., Metzler, D., Turtle, H., and Croft, W. B., (2005). Indri: A language-model based search engine for complex queries. In *proceedings of International Conference on Intelligence Analysis*, May 2-6, extended paper.

Zhou, Xuezhong, Liu, Baoyan, Wu, Zhaohui, and Feng, Yi, (2007). Integrative mining of traditional Chinese medicine literature and MEDLINE for functional gene networks. *Artificial Intelligence in Medicine*, 41(2), pp. 87--104.