Learning Inter-Related Statistical Query Translation Models for English-Chinese Bi-Directional CLIR

Yuejie Zhang*, Lei Cen*, Cheng Jin*, Xiangyang Xue*, Jianping Fan+
*School of Computer Science, Shanghai Key Laboratory of Intelligent Information Processing, Fudan University, Shanghai, China
+Department of Computer Science, The University of North Carolina at Charlotte, USA
*{yjzhang, 082024072, jc, xyxue}@fudan.edu.cn, +jfan@uncc.edu

Abstract

To support more precise query translation for English-Chinese Bi-Directional Cross-Language Information Retrieval (CLIR), we have developed a novel framework by integrating a semantic network to characterize the correlations between multiple inter-related text terms of interest and learn their inter-related statistical query translation models. First, a semantic network is automatically generated from large-scale English-Chinese bilingual parallel corpora to characterize the correlations between a large number of text terms of interest. Second, the semantic network is exploited to learn the statistical query translation models for such text terms of interest. Finally, these inter-related query translation models are used to translate the queries more precisely and achieve more effective CLIR. Our experiments on a large number of official public data have obtained very positive results.

1 Introduction

With the explosive growth of multilingual documents, Cross-Language Information Retrieval (CLIR) has received increasing attention in recent years [Wang *et al*., 2006]. Most of current CLIR systems select Query Translation as the main strategy, which has become a pattern with the lowest cost and difficulty [Oard *et al*., 2008]. To facilitate more effective CLIR service in the Internet search engines, it is very important to develop a powerful query translation engine.

The development of large-scale bilingual parallel corpora and Web has provided a good platform for the research on the statistical query translation in CLIR [Gao *et al*., 2006]. It appears that large-scale bilingual parallel corpora from abundant Web information can mitigate the problems of the limited coverage of bilingual dictionary and the restricted translation accuracy. Recently, most of the research on query translation concentrates on mining the useful information from large-scale bilingual parallel corpora [Li *et al*., 2009].

To support more precise query translation and achieve more effective English-Chinese bi-directional CLIR, a novel framework is developed to integrate a semantic network for characterizing the correlations between multiple query terms and learning their inter-related statistical query translation models. The major difference between our approach and other existing methods is that ours exploits the inter-term correlations to learn their statistical translation models simultaneously based on the established semantic network, while the others treat all the query terms equally and independently. Our experiments on a large number of official public data have obtained very positive results.

2 Related Research Work

There are three approaches for query translation, that is, bilingual-dictionary-based, machine-translation-based and bilingual-parallel-corpus-based. The general assumption of these approaches is that the incorrect translations of a few query terms in a query are tolerable and can be remedied via query expansion. For longer queries, it is still possible to retrieve relevant documents in the target language even if there exist a few unknown query terms. Many techniques have also been developed to resolve the term ambiguity associated with query translation [Monz *et al*., 2005].

Bilingual dictionaries have been used in CLIR because of their simplicity and the increasing availability. However, the bilingual-dictionary-based approach may suffer from two problems: (a) the dictionary may have a poor coverage; and (b) it is difficult to select the correct translation of a word among all the translations provided by the dictionary. The machine-translation-based approach translates queries into the document language or all the documents into the query language, and could be a natural way for CLIR. Nevertheless, the translation quality for most existing translation systems is low, and high quality translations can be obtained only when the applicable domain is limited. Disambiguation is accomplished through the syntactic analysis. A number of studies in recent years have explored the possibility of using bilingual parallel texts for query translation in CLIR [Li *et al*., 2009][Wang *et al*., 2006]. One potential advantage of such an approach is that it can provide multiple translations for the same meaning. The translation of a query would contain not only terms for correct translations but also relevant terms. Integrating the thesaurus and the bilingual-parallel-corpus-based statistics technique has been a hot research.

Unfortunately, all these existing approaches may still suffer from the following two issues. (a) Inter-Term Correlations for Query Translation–Most existing query translation approaches focus on translating the text terms of the queries independently, and the correlations between the terms are completely ignored. Even many techniques have been developed to solve term ambiguity associated with query translation, it is still very hard for most existing query translation approaches to process the polysemous terms effectively and determine the meaningful phrases from queries precisely, especially for short queries. When multiple terms are used for query formulation, they should have strong correlations and the order of these terms (i.e. syntactic structure of queries) is also meaningful for query translation. (b) Discovering Missing Important Terms–Some important terms may be missed in short queries when unprofessional users cannot find suitable terms to formulate their queries. Such short queries are insufficient to describe users' information needs in a precise and unambiguous way, thus missing the important terms may hinder users to find more relevant documents effectively. Hence, query expansion has become the commonly used approach to solve this problem, which attempts to add strict synonyms and relevant terms with high frequency to the original query. However, the missing important terms do not need to be strict synonyms and the relevance degree for such terms should be measured. For most existing query translation approaches, a good solution hasn't yet been provided to discover missing important terms and translate them into the target language precisely.

In our approach, query translation is modeled as a correlation distribution over bilingual semantic concepts, which in reality can be approximated by the correlation distribution over the senses of query terms and their translations. The most important is to create the semantic similarity between bilingual concepts/senses and measure what degree they are related. First, a semantic network is automatically generated from large-scale English-Chinese bilingual parallel corpora to characterize the correlations between a large number of text terms of interest. Second, the semantic network (i.e., text terms of interest and their inter-term correlations) is exploited to learn the statistical query translation models for a large number of text terms of interest accurately. Finally, these inter-related translation models are used to translate the queries precisely and achieve more effective CLIR.

3 Semantic Network Generation

The semantic network can be considered as an important basis for: (a) characterizing the inter-term similarity contexts precisely; (b) representing the inter-related learning tasks explicitly; and (c) bringing the powerful inferencing scheme to learn the inter-related statistical query translation models with higher discrimination and adaptation power.

3.1 Extraction for Text Terms of Interest

For the semantic network construction oriented to large-scale bilingual parallel corpora, its important basis concerns the extraction for text terms of interest. Intuitively, text terms of interest should be some terms prominent in documents and play important roles in making discrimination between different documents [Sproat *et al*., 2006]. Primarily, text terms of interest should be non-stopwords and focus on some specific terms such as noun, verb and adjective. Moreover, non-stopwords with higher frequency are able to well illustrate and represent the topics of documents and can be taken as text terms of interest. Furthermore, the same text terms are usually avoided to occur with higher repeatability within closer distance in a document, and then the related synonymies can also be considered as text terms of interest. Additionally, if non-stopwords appear in the top positions of a document, they are crucial for the topic description of the document and can be accepted as text terms of interest.

3.2 Inter-Term Semantic Similarity Context

Our semantic network consists of large amounts of text terms of interest and their inter-term similarity contexts. Multiple criteria are utilized to achieve more precise characterization of the inter-term semantic contexts in large-scale bilingual parallel corpora. For two text terms of interest C_i and C_j, their inter-term semantic context $\phi(C_i, C_j)$ consists of the flat inter-term semantic context because of their co-occurrences in bilingual parallel corpora and the hierarchical inter-term semantic context because of their inherent correlation defined by WordNet. $\phi(C_i, C_j)$ is defined as:

$$\phi(C_i, C_j) = -\upsilon \cdot \frac{\theta(C_i, C_j)}{\log\theta(C_i, C_j)} - (1-\upsilon) \cdot \theta(C_i, C_j) \cdot \log\frac{L(C_i, C_j)}{2 \cdot D} \quad (1)$$

where the first part is used to characterize the flat inter-term semantic context; the second part is used to characterize the hierarchical inter-term semantic context; $\theta(C_i, C_j)$ is the co-occurrence probability for C_i and C_j; $L(C_i, C_j)$ is the number of nodes between C_i and C_j on WordNet; υ is the relative importance factors; and D is the maximum number of nodes from root node to leaf node on WordNet.

When a large number of text terms of interest and their inter-term similarity contexts are available, they are used to construct a semantic network. Unlike the one-direction IS-A hierarchy, each text term of interest can be linked with all the other text terms of interest on the semantic network. However, the strengths of the associations between some terms may be very weak, thus it is not necessary for each term to be linked with all the other terms on the semantic network. Eliminating the weak inter-term links can allow machine learning algorithms to concentrate on the most significant inter-term similarity contexts and learn their statistical query translation models accurately. Therefore, each text term of interest is automatically linked with the most relevant text terms of interest with larger values of the inter-term similarity contexts (i.e., their values of $\phi(\cdot, \cdot)$ are above a threshold).

The semantic network can provide a good environment for: (a) tackling the translation ambiguity; (b) determining more meaningful phrases; and (c) dealing with missing important terms in short queries. Our semantic network for our large-scale bilingual parallel corpora is shown in Figure 1, where each text term of interest is linked with multiple relevant text terms of interest with larger values of $\phi(\cdot, \cdot)$. It is worth noting that different terms of interest can have different numbers of the most relevant terms of interest.

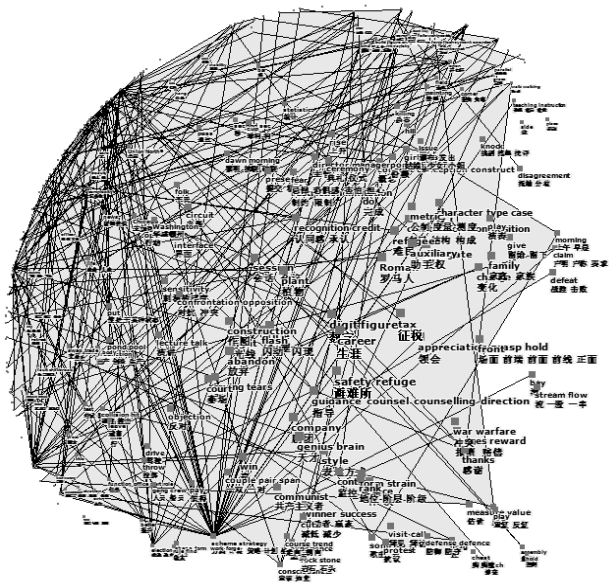

Figure 1. Our semantic network for text terms of interest.

3.3 Combination for Synonymous Terms

Some text terms of interest may be synonymous. The appearances of such terms may result in low recall rates. To address this issue, WordNet is first incorporated to identify the candidates of the synonymous terms in English. These candidate terms are then clustered into multiple groups according to their inter-term similarity contexts. We have incorporated a min-max cut algorithm for term clustering [Shi *et al*., 2000]. Thus the synonymous terms, which have large values of the inter-term similarity contexts, are grouped into the same cluster. These synonymous terms in the same cluster are merged as one *super-term*, and their documents are assigned into the super-term automatically. By combining the synonymous terms and their documents, information seekers can obtain a complete set of the relevant documents.

When multiple synonymous terms $\{c_1, \ldots, c_n\}$ are integrated as one super-term C_s, the inter-term similarity contexts between one given term on the semantic network and C_s largely depend on its inter-term similarity contexts with all these synonymous terms $\{c_1, \ldots, c_n\}$. A novel algorithm is developed for calculating the aggregated similarity contexts between C_s and the other residue terms on the semantic network more effectively. Thus the aggregated semantic context $\hat{\phi}(C_s, C_j)$ between C_s and the residual term C_j on the semantic network is obtained by using Eq.(1) with the cumulative probabilities $\hat{\theta}(C_s, C_j)$ and $\hat{\theta}(C_s)$ and a new inherent correlation $\hat{L}(C_s, C_j)$, which are defined in Eq.(2).

$$\hat{\theta}(C_s, C_j) = \sum_{l=1}^{n} \theta(c_l, C_j), \qquad \hat{\theta}(C_s) = \sum_{l=1}^{n} \theta(c_l) \tag{2}$$
$$\hat{L}(C_s, C_j) = \min\{L(c_l, C_j) \mid l = 1, \cdots, n\}$$

where $\theta(c_l, C_j)$ is the co-occurrence probability for the synonymous term c_l and C_j, $\theta(c_l)$ is the individual occurrence probability of c_l, $L(c_l, C_j)$ is the inherent correlation between c_l and C_j on WordNet.

3.4 Split for Polysemous Terms

Some text terms of interest may be polysemous. The appearances of such terms may result in low precision rates and induce ambiguity for query translation. To address such an issue, WordNet is first incorporated to identify the candidates of the polysemous terms. The polysemous terms are split into multiple sub-terms. For a polysemous term, all its documents are partitioned into multiple clusters automatically, and each cluster may correspond to one *sub-term*. By assigning the documents for the polysemous term into multiple sub-terms, information seekers can obtain more relevant documents according to their real query intentions.

When the polysemous term C_p is split into multiple sub-terms $\{\hat{c}_1, \cdots, \hat{c}_m\}$, a novel algorithm is developed to determine their split inter-term similarity contexts with the residue terms on the semantic network more effectively. The split inter-term semantic context between the sub-term $\hat{c}_l$ and any other residue text term of interest on the semantic network is defined by using Eq.(1), where the co-occurrence probability is replaced by the split co-occurrence probability. The split co-occurrence probability $\tilde{\theta}(\hat{c}_l, C_j)$, the split occurrence probability $\tilde{\theta}(\hat{c}_l)$, and the new inherent correlation $\tilde{L}(\hat{c}_l, C_j)$ are defined in Eq.(3).

$$\hat{\theta}(\hat{c}_l, C_j) = \frac{|\hat{S}_l|}{|S_p|}\theta(C_p, C_j), \qquad \hat{\theta}(\hat{c}_l) = \frac{|\hat{S}_l|}{|S_p|}\theta(C_p) \tag{3}$$
$$\hat{L}(\hat{c}_l, C_j) = L(C_p, C_j) + 1$$

where $\theta(C_p, C_j)$ is the co-occurrence probability for C_p and the residue term C_j, $\theta(C_p)$ is the occurrence probability for C_p, $|\hat{S}_l|$ is the size of the document set $\hat{S}_l$ for $\hat{c}_l$, $|S_p|$ is the size of the document set S_p for C_p, $|S_P| = \sum_{l=1}^{m} |\hat{S}_l|$, $L(C_p, C_j)$ is the inherent correlation between C_p and C_j on WordNet.

4 Learning Statistical Query Translation Models

To achieve more accurate query translation, it is very important to develop new algorithms for learning more reliable statistical query translation models for all the text terms of interest on the semantic network. It is necessary to note that all the text terms of interest are correlated and such the inter-term similarity contexts can be represented explicitly by the semantic network and characterized precisely by the strengths of their inter-term similarity contexts $\phi(\cdot, \cdot)$. However, direct modeling of the global inter-term similarity contexts over the whole semantic network becomes computationally intractable. Thus we have developed a structured max-margin learning scheme by incorporating the first-order nearest neighbors (i.e., cliques of the semantic network) and the max-margin Markov networks to exploit the inter-term similarity contexts to learn a large number of inter-related statistical query translation models.

The fully connected text terms of interest on the semantic network are strongly correlated. Thus isolating these terms and learning their statistical translation models independently are not appropriate. In order to exploit the inter-term similarity contexts for model learning, it is very important to

develop new frameworks for integrating multi-task learning with max-margin Markov networks. The idea behind multi-task learning is that if multiple inter-related learning tasks share a common predictive structure, such the structure can be estimated more reliably by considering these inter-related learning tasks together [Taskar, 2004]. One of the most important open problems for multi-task learning is to better characterize what the related tasks are. The idea behind max-margin Markov networks is to exploit both the advantages of the graphical models (i.e., good modeling of inter-term prediction structure) and the Support Vector Machines (SVMs) (i.e., good generalization ability) for achieving more reliable model learning [Lafferty *et al*., 2001].

We have developed a structured max-margin learning scheme by incorporating the semantic network, multi-task learning and max-margin Markov networks to enhance the discrimination power of a large number of inter-related statistical query translation models [Fan *et al*., 2008]. The semantic network is used to identify the inter-related learning tasks precisely, e.g., training the inter-related translation models for the fully connected terms on the semantic network. The inter-task relatedness is characterized explicitly by using the strengths of $\phi(\cdot, \cdot)$, and a common predictive structure is shared among the inter-related translation models. The max-margin Markov networks are integrated to approximate the inter-term similarity contexts.

For a given text term of interest C_i, its first-order nearest neighbors on the semantic network are denoted as Ξ_i (i.e., cliques of the graph). Obviously, the sizes of the first-order nearest neighbors for different text terms of interest could be very different. The joint conditional distribution $P(C_i, X)$ (X refers to one test query term) can be modeled as:

$$P(C_i, X) = \frac{1}{Z}\exp\left(\sum_{C_k \in \Xi_i} f(C_k, X) + \sum_{C_m \in \Xi_i}\sum_{C_n \in \Xi_j} f(C_m, C_n, X)\right) \quad (4)$$

where $f(C_k, X)$ is the basic discriminant function for C_i; $f(C_m, C_n, X)$ is the pairwise discriminant function for the fully connected terms C_i and C_j; Ξ_j is the first-order nearest neighbors of C_j; and Z is a normalizing constant defined as:

$$Z = \sum_{I=1}^{T}\exp\left(\sum_{C_k \in \Xi_i} f(C_k, X) + \sum_{C_m \in \Xi_i}\sum_{C_n \in \Xi_j} f(C_m, C_n, X)\right) \quad (5)$$

where T is the total number of the first-order nearest neighbors on the semantic network. In this paper, T is equal to the total number of text terms of interest on the semantic network because only the first-order nearest neighbors are considered.

We are not interested in the exact form of the joint conditional probability $P(C_i, X)$, we are rather interested in the statistical query translation model $H_{Ci}(X)$ and its confidence for query translation and term sense selection. Once the translation model for C_i has been fitted on the bilingual training documents, one can do automatic query translation and term sense selection by computing:

$$H_{Ci}(X) = \arg\max\left(\sum_{C_k \in \Xi_i} f(C_k, X) + \sum_{C_m \in \Xi_i}\sum_{C_n \in \Xi_j} f(C_m, C_n, X)\right) \quad (6)$$

This comes from the following fact:

$$P(C_i|X) \propto \exp\left(\sum_{C_k \in \Xi_i} f(C_k, X) + \sum_{C_m \in \Xi_i}\sum_{C_n \in \Xi_j} f(C_m, C_n, X)\right) \quad (7)$$

where $P(C_i | X)$ is the posterior probability for one test query term X to be translated into C_i.

For each text term of interest, we have labeled N bilingual training documents for model learning, and the total number of the training documents for all these M fully connected terms can be $N \times M$. Even the number of the training documents is small for each term, we can always have larger number of training documents for learning the common predictive structure more accurately. Because the common predictive structure can be learned jointly by using the training documents for all these fully connected terms, our structured max-margin learning algorithm can learn a large number of inter-related translation models more effectively. By using a common predictive structure to characterize the relatedness among multiple inter-related discriminant functions, our structured max-margin learning algorithm can have lower computational complexity. By learning from the training documents for other fully connected terms, our structured max-margin learning algorithm can enhance the discrimination and adaptation power of these inter-related translation models significantly. Incorporating the training documents from other inter-related terms for model learning can generalize the translation models from fewer training documents, especially when the available training documents for the given text term of interest may not be representative for large amounts of unseen test query terms.

Our query translation and term sense selection are achieved by a voting from multiple inter-related statistical query translation models for the inter-related text terms of interest to make their errors to be transparent. (a) The test query term is first assigned into the most relevant text term of interest C_p on the semantic network, which has the maximum value of the confidence $P(C_p|X)$. There are T terms on the semantic network and each has a clique, thus the computational cost for this step is $O(T)$. (b) To determine the potential alignment paths for the second term of the test query, its confidences to be assigned into the relevant terms of interest (which are fully connected with the best matched term C_p) are further calculated. The alignment path with the largest value of confidence (i.e., the most possible sense of the term) is selected to translate the second query term. (c) This process is terminated when all these query terms are assigned into the most relevant terms according to their senses.

It is important to note that once the process above is finished, multiple inter-related text terms of interest on the semantic network are selected to interpret the test query in the target language. Especially, our algorithm can provide a good solution to discover the missing important terms more effectively and precisely. For example, if the parent node is matched with the first query term and the grandchild node is matched with the second one, the terms of interest between these two nodes can be selected to fill the missing important terms, so that users can obtain more relevant documents.

5 Experiment and Analysis

5.1 Data Set and Evaluation Metrics

Our English-Chinese bilingual parallel corpora for the semantic network construction are taken from 2009 China Workshop on Machine Translation (CWMT), which has

3,780,000 sentence pairs in total. For the evaluation of English-Chinese query translation and CLIR, 25 English and Chinese bilingual topics (CH55-CH79) and Chinese corpora (126,937 documents from Hong Kong newspapers) from the CLIR task of TREC-9 are utilized. For the evaluation of Chinese-English query translation and CLIR, 50 bilingual topics (351-400) and English corpora (242,918 documents from Associated Press Newswire) from TREC-7 are used[1].

We have incorporated the official criteria for machine translation, which aim at evaluating how well the semantic network assists users on query translation and formulation. The *Precision* (*P*) and *Recall* (*R*) rates are introduced to evaluate how well our translation model supports CLIR.

5.2 Experiments on Query Translation

Referring to the translation classification criteria published by National 863 Expert Group of China, the translation for each short query (the terms in the title field) and long query (the terms in both title and description fields) by our model is evaluated, and compared with those by the Bilingual-Dictionary-based [2] and machine-translation-based [3] translation patterns, as shown in Table 1. It can be seen that the translation quality by our approach is more satisfactory.

Translation Pattern	Classification Criteria		
	Satisfactory	With Errors	Illegal
Based on Bilingual Dictionary	56%	32%	12%
Based on Machine Translation	64%	28%	8%
Based on Our Semantic Network	76%	19%	5%

Table 1. Translation evaluation results for queries from TREC-9.

Some examples for short queries are given in Table 2. It can be observed that some special short queries with polysemous terms, phrases and missing important terms can be well translated and formulated with the correct translations and relevant terms through our model. For Query CH56 and 373 with the polysemous terms "*violation*" and "出口" [*exit*/*export*], based on the polysemous correlations of each term on the semantic network, the relevant translation terms can be acquired, such as "违反" [*violation*/*violate*] for "*violation*" and "*export*" for "出口" [*exit*/*export*]. For Query CH55 and 385 with the phrases "*World Trade Organization*" (*WTO*) and "混合燃料" [*hybrid fuel*], based on the hierarchical correlations on the semantic network, the meaningful phrases can be determined and then their corresponding translations can be obtained, such as "世界贸易组织" [*WTO*] for "*World Trade Organization*", and "*hybrid fuel*" for "混合燃料" [*hybrid fuel*]. The final result is much better than those acquired by the bilingual-dictionary-based and machine-translation-based query translation patterns. Especially for some extremely short queries with only two significant terms, such as Query CH73 and 354, both of the direct translation for each single query term and the other meaningful missing important terms are discovered, which are consistent with the illustration in the description and narrative fields of the query topic.

[1] For Chinese topics from TREC-9, their English translations are provided. However, for English topics from TREC-7, no Chinese translations are directly available and each topic is manually translated into Chinese. Each topic includes the title, description and narrative field.
[2] The bilingual dictionary we used is a human compiled bilingual lexicon and involves almost 360,000 lexical entries in total.
[3] The machine translation system we used is a famous free online translator SYSTRAN (*http://www.systranet.com*).

Existent Problem	Topic Number	Query Form	Bilingual-Dictionary-based Query Translation	Machine-Translation-based Query Translation	Our Semantic-Network-based Query Translation
Polysemous Term	CH56	*human rights* ***violations***	人权[*human right*]; **违犯** [***offend against***]	人权[*human right*]; **侵害** [***infringe upon***]	人权[*human right*]; **违反** [***violation/violate***]
	373	加密设备**出口** [*encryption equipment* ***export***]	加密[*encrypt*]; 设备[*equipment*]; **出口**[***exit***]	加密[*encryption*]; 设备[*supposes to*]; **出口**\[***prepare the mouth***]	加密[*encryption*]; 设备[*equipment*]; **出口**\[***export***]
Phrase	CH55	***World Trade Organization*** *membership*	**世界**[***World***]; **交易**[***Trade***]; **组织**[***Organization***]; 成员资格 [*qualification for member*]	**世界贸易组织**[***World Trade Organization***]; 成员国 [*qualification for member*]	**世界贸易组织**[***World Trade Organization***]; 成员国 [*member nation*]
	385	**混合燃料**汽车 [***hybrid fuel*** *cars*]	**混合**[***mix***]; **燃料**[***fuel***]; 汽车 [*automobile*]	**混合**[***blended***]; **燃料**[***fuel***]; 汽车 [*automobile*]	**混合燃料**[***hybrid fuel***]; 汽车[*automobile/ auto/car/vehicle*]
Missing Important Term	CH73	*AIDS in China*	艾滋病[*AIDS*]; 中国[*China*]	艾滋病[*AIDS*]; 在[*in*]; 中国[*China*]	艾滋病[*AIDS*]; **艾滋病毒**[***HIV***]; **传染病**[***pandemic***]; **治疗**[***treatment***]; **医院**[***hospital***]; **办事处**[***authority office***]; **国家**[***country***]; 中国[*China*]
	354	新闻工作者风险[*journalist risks*]	新闻工作者[*journalist*]; 风险[*risk*]	新闻工作者[*journalist*]; 风险[*risk*]	新闻工作者[*journalist*]; **记者**[***correspondent/reporter***]; **不幸/灾难**[***disaster***]; **冲突**[***conflict***]; **暴力**[***violence***]; **恐怖主义**[***terrorism***]; 风险[*risk*]

Table 2. Several translation examples for short query.

5.3 Experiments on English-Chinese Bi-Directional CLIR

To investigate the effect of our model for CLIR, 14 runs are carried out. (1) *E-C_ShortCLIR*1–English Short Query (ESQ) and the Bilingual-Dictionary-based Translation (BDT); (2) *E-C_ShortCLIR*2–ESQ and the Machine-Translation-based Translation (MTT); (3) *E-C_ShortCLIR*3–ESQ and our model; (4) *E-C_LongCLIR*1–English Long Query (ELQ) and BDT; (5) *E-C_LongCLIR*2–ELQ and MTT; (6) *E-C_LongCLIR*3–ELQ and our model; (7) *C-C_MonoIR*–Chinese Long Query (CLQ), monolingual IR; (8) *C-E_ShortCLIR*1–Chinese Short Query (CSQ) and BDT; (9) *C-E_ShortCLIR*2–CSQ and MTT; (10) *C-E_ShortCLIR*3–CSQ and our model; (11) *C-E_LongCLIR*1–CLQ and BDT; (12) *C-E_LongCLIR*2–CLQ and MTT; (13) *C-E_LongCLIR*3–CLQ and our model; (14) *E-E_MonoIR*–ELQ, monolingual IR. The *P-R* curves and Median Average Precision (MAP) values are shown in Figure 2 and 3.

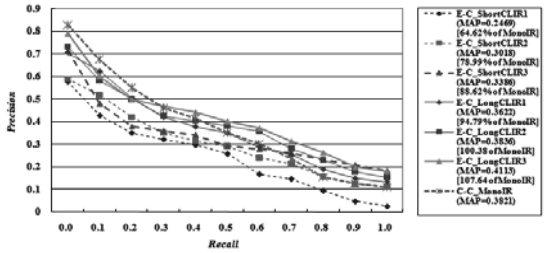

Figure 2. Results for English-Chinese CLIR runs.

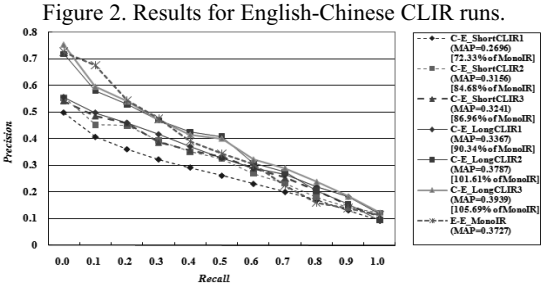

Figure 3. Results for Chinese-English CLIR runs.

It can be found that the best runs for English-Chinese and Chinese-English CLIR are *E-C_LongCLIR*3 and *C-E_LongCLIR*3 respectively. By adopting our translation model, the English-Chinese bi-directional CLIR for long query has gained the significant improvement on the whole retrieval performance. We can even achieve 107.64% of Chinese monolingual effectiveness for English long query in English-Chinese CLIR and 105.69% of English monolingual effectiveness for Chinese long query in Chinese-English CLIR. This performance is comparable to those reported by other English-Chinese or Chinese-English CLIR systems [Gao *et al*., 2006][Wang *et al*., 2006]. Compared with the traditional query translation patterns based on bilingual dictionary and machine translation, the semantic network that covers abundant bilingual inter-term correlations is exactly a better way for query translation. Through comparing the results of *E-C_ShortCLIR*1~3 or *C-E_ShortCLIR*1~3, it can also be further confirmed that our model can support English-Chinese bi-directional CLIR for short query.

5.4 Analysis and Discussion

Through analyzing the results for query translation and retrieval, it can be found that the translation quality is highly related to the following aspects. (1) The translation acquisition is associated with the segmentation results for original query, especially for short query in Chinese. For example, in Short Query 396 in Chinese, "病态建筑综合症" [*sick building syndrome*], its segmentation results are "病态" [*morbidity*], "建筑" [*building*/*construction*] and "综合症" [*syndrome*], in which this terminology is segmented into multiple words and some noisy terms may be introduced. (2) Syntactic relations for query terms should be considered for original query. For example, in Short Query CH57 in English, "*organic or natural foods*", the important word "*or*" that indicates the parallel relation between "*organic*" and "*natural*" is removed as a stopword in the preprocessing, and then the subsequent query translation and retrieval results are influenced. (3) The special translation mechanism for OOV query term should be combined with our model. Although our semantic network can cover a large number of text terms of interest and their correlations, some OOV terms are not yet involved. For example, "绝经后" [*postmenopausal*], "雌激素" [*estrogen*] and "体外授精" [*in vitro fertilization*] in Short Query 356 and 368 in Chinese, and "*Daya Wan*" in Short Query CH62 in English. Fusing our existing technique for OOV term translation can solve this problem with better ability. (4) Some extremely short queries contain only one single query term, which may cause that the missing important terms cannot be discovered correctly. For example, in Short Query 379 in Chinese, there is only one query term "主流" [*mainstreaming*]. This is the most stubborn problem. We consider making reasonable pre-expansion for such extremely short query and then utilizing our model efficiently. (5) With the growth of more available bilingual parallel corpora from Web, our semantic network should be updated and adjusted dynamically. Therefore, it can provide more accurate description for semantic relations and translation information, and facilitate precise translation model learning.

6 Conclusions

In this paper, a new framework is introduced to support more precise query translation for English-Chinese bi-directional CLIR. To address the issues of polysemous terms, phrases and missing important terms in query translation and formulation, and enhance the CLIR performance, a novel algorithm is developed by integrating the semantic network and the statistical query translation models to efficiently exploit correlations and semantic similarity between text terms of interest. Our experiments on a large number of data from CWMT2009 and TREC have provided promising results. Our future work will focus on making our new system available online, so that more users can join our study.

Acknowledgments

This work is supported by Shanghai Natural Science Fund (No. 09ZR1403000), 973 Program of China (No. 2010CB327906), National Natural Science Fund of China (No. 60873178), Shanghai Leading Academic Discipline Project (No. B114) and Shanghai Municipal R&D Foundation (No. 08dz1500109).

References

[Fan *et al*., 2008] J.P. Fan, Y.L. Gao, and H.Z. Luo. Inte-Concept Ontology and Multi-Task Learning to Achieve More Effective Classifier Training for Multi-Level Image Annotation. *IEEE Trans. on Image Processing*, 17(3):407-426, 2008.

[Gao *et al*., 2006] J.F. Gao and J.Y. Nie. A Study of Statistical Models for Query Translation: Finding a Good Unit of Translation. In *Proc. of SIGIR 2006*, pages 194-201, 2006.

[Lafferty *et al*., 2001] J. Lafferty, A. McCallum, and F. Pereira. Conditional Random Field: Probabilistic Models for Segmenting and Labeling Sequence Data. In *Proc. of ICML 2001*, pages 282-289, 2001.

[Li *et al*., 2009] Q. Li, Y.Z. P. Chen, S.H. Myaeng, Y. Jin, and B.Y. Kang. Concept Unification of Terms in Different Languages via Web Mining for Information Retrieval. *Information Processing & Management*, 45(2):246-262, 2009.

[Monz *et al*., 2005] C. Monz and B.J. Dorr. Iterative Translation Disambiguation for Cross-Language Information Retrieval. In *Proc. of SIGIR 2005*, pages 520-527, 2005.

[Oard *et al*., 2008] D.W. Oard, D.Q. He, and J.Q. Wang. User-Assisted Query Translation for Interactive Cross-Language Information Retrieval. *Information Processing & Management*, 44(1):181-211, 2008.

[Shi *et al*., 2000] J.B. Shi and J. Malik. Normalized Cuts and Image Segmentation. *IEEE Transactions on PAMI*, 22(8):888-905, 2000.

[Sproat *et al*., 2006] R. Sproat, T. Tao, and C.X. Zhai. Named Entity Transliteration with Comparable Corpora. In *Proc. of COLING-ACL 2006*, pages 73-80, 2006.

[Taskar, 2004] B. Taskar. *Learning Structured Prediction Models*: *A Large Margin Approach*. PhD Thesis, 2004.

[Wang *et al*., 2006] J.Q. Wang and D.W. Oard. Combining Bidirectional Translation and Synonymy for Cross-Language Information Retrieval. In *Proc. of SIGIR 2006*, pages 202-209, 2006.