Journal of Chinese Language and Computing15 (1): (25-32)

Research on Lucene-based English-Chinese Cross-Language Information Retrieval

Yuejie Zhang[①] Tao Zhang[②] Shijie Chen[①]
[①]Department of Computer Science and Engineering
Shanghai Key Laboratory of Intelligent Information Processing
Fudan University, Shanghai, China (200433)
yjzhang@fudan.edu.cn
[②]School of Information Management and Engineering
Shanghai University of Finance & Economics, Shanghai, China (200433)
taozhang@mail.shufe.edu.cn

Submitted on March, 2005, Accepted and Revised on December, 2005

Abstract

In this paper, we present our English-Chinese Cross-Language Information Retrieval (CLIR) system. We focus our attention on finding effective translation equivalents between English and Chinese, and improving the performance of Chinese IR. On English-Chinese CLIR, we adopt query translation as the dominant strategy, and utilize English-Chinese bilingual dictionary as the important knowledge resource to acquire correct translations. On Chinese monolingual retrieval, we investigated the use of different entities as indexes and implement our retrieval system based on the Lucene toolkit. On system evaluation, we present an effective method to generate the sets of relevant documents for query topics.

Keywords

Cross-Language Information Retrieval, query translation, bilingual dictionary, monolingual retrieval, Lucene toolkit, relevance feedback.

1. Introduction

Cross-language information retrieval (CLIR) enables users to search in multilingual document collections using their native language, supported by an effective combination of linguistic and information retrieval technologies. English-Chinese CLIR is a major sub-problem within CLIR.

This paper focuses on the techniques and algorithms used in our system. In Section 2, techniques for the query translation approach are discussed. Section 3 introduces the choice of best indexing units for Chinese IR and the implementation of Chinese monolingual retrieval system which is built on top of the Lucene toolkit. Section 4 describes an effective method to construct the set of relevant documents for query topics. Finally, we present our conclusion in section 5.

2. Query Translation

Here, we adopt query translation as the dominant strategy using English query as the translated object, and utilizing English-Chinese bilingual dictionary as the main knowledge resource for translation.

2.1 Knowledge Source Construction

The knowledge source used in English-Chinese CLIR system mainly includes dictionary knowledge and Chinese Synonym Dictionary. In addition, stopword list and word morphological resumption list are also utilized in our system. In fact, dictionary is a carrier of knowledge expression and storage, which involves almost all information about vocabulary, namely static information.

(1) English-Chinese Bilingual Dictionary.

This dictionary is mainly used in translation processing in word level and phrase level. And it consists of three kinds of dictionary component as follows:

- ✧ Basic Dictionary -- A basic knowledge source independent of particular field, which records basic linguistic vocabulary.
- ✧ Technical Terminology Dictionary -- Recording terminology knowledge in a particular technical field, which is mainly referred to Hong Kong commercial terminology knowledge and incorporated in the basic dictionary.
- ✧ Idiom Dictionary -- Recording familiar fixed matching phenomena, such as idiom and phrase.

The whole bilingual dictionary involves almost 50,000 lexical entries. And each entry is established as the following data structure:

English lexical Information	Part-of-Speech Information	Subcategory Information	Concept Number	Matching Information	Semantic Class Code	Chinese lexical Information

An example of particular entry representation form in dictionary is listed as the following:

*happiness || n || ng || 0 || M ;[U]; || bbaaa ||幸福(felicity) ||||

(2) Chinese Synonym Dictionary

Actually, this dictionary is a thesaurus, which involves nearly 70,000 entries. All entries are arranged according to specified semantic relations. It is mainly used in expanding translation that has passed through translation processing, namely query expansion.

(3) Other knowledge bases

While the stopword list is used in tagging the stopwords in English query, and the English morphological resumption list which describes all irregular varieties about vocabulary is used in morphological resumption of words with irregular variety forms.

2.2 Translation algorithm

The basic framework of English-Chinese-oriented translation algorithm is mainly divided into three parts, as shown in Figure 1.

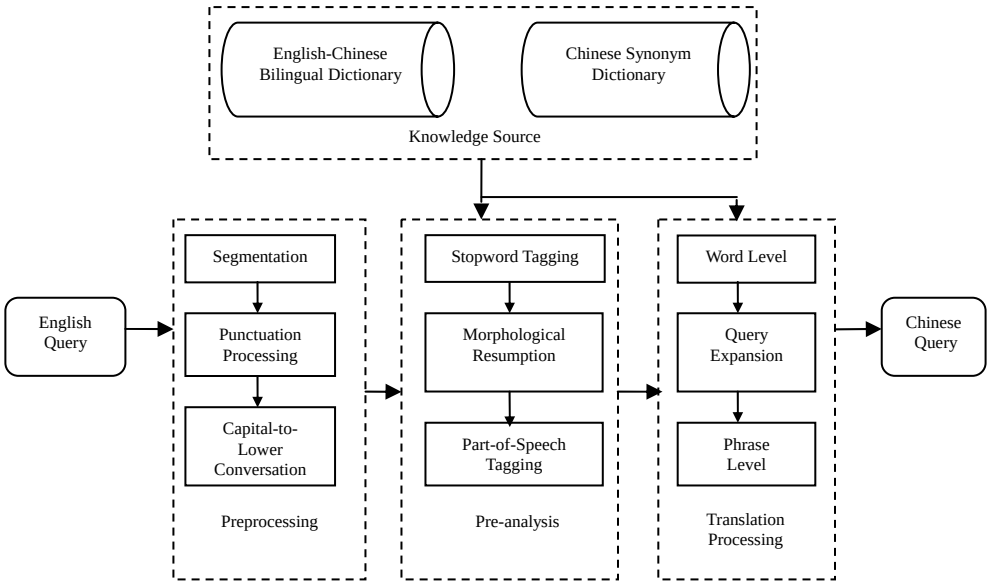

Figure 1. Basic framework of English-Chinese-oriented query translation algorithm

- Preprocessing -- including sentence segmentation, punctuation tagging and capital-to-lower letter conversation for English query.
- Pre-analysis -- including stopwords tagging, word morphological resumption and POS tagging processes.

  Considering that translation processing is related with some stopwords, the stopwords must be tagged by the stopword list. Because there are some words with variety forms in English query, translation knowledge cannot be induced correctly. So by using the English-Chinese bilingual dictionary, the morphological resumption list for irregular variety and heuristics for regular variety, we get words' original form from the process called "morphological resumption". To analyze word part-of-speech, we develop a HMM-based (Hidden Markov Mode) Part-of-Speech Tagger.
- Translation processing -- including translation processes in two levels, that is, word level and phrase level.

  Word level translation -- By using the basic vocabulary part of English-Chinese bilingual dictionary, this process mostly implements translation word by word. For word disambiguation, a word may correspond with several kinds of different sense. Word sense is related with particular word, and cannot be given without particular linguistics environment. The condition of linguistics environment may be syntactic and semantic parameters. When selecting a particular word, the difference mark of word should be chosen. This difference mark represents a certain syntactic and semantic feature, and identifies the sense of word uniquely, namely Concept Code. The concept code together with lexical entry can decide a certain word sense to accomplish word sense disambiguation. For machine translation, word disambiguation should be a very important problem. But in our CLIR system, in some degree, word disambiguation has not taken some obvious affect to retrieval efficiency. At the same time, in order to provide more query information to retrieval system, by using "Chinese Synonym

Dictionary", expansion operation is done for translation knowledge through translation processing. According various synonymous relations described in the dictionary above, all synonyms corresponding with translation knowledge is listed, namely completing query expansion process. Thus, more affluent query information can be provided to retrieval system. So the retrieval efficiency is increased greatly, and the retrieval performance is improved.

Phrase level translation -- This process is implemented based on the idiom dictionary part of English-Chinese bilingual dictionary. The recognition of near distance phrase and far distance phrase is an important problem. Here, by adopting Greedy Algorithm, the recognition and translation processing of near distance phrase is mainly completed, shown as the following:

- Acquiring phrase set which includes some phrases taking current query word as head word from English-Chinese bilingual dictionary.
- Establishing some phrases which take current word as head word and involve the same number of word as the member in phrase set.
- Comparing each one of the established phrases and every member in the correspondent phrase set and finding the matched phrase with the maximum length.

3. Chinese Information Retrieval

3.1 Chinese word segmentation and its effect on Chinese IR

Unlike written in English where spaces are used as word delimiters. Chinese texts do not use spaces to mark word boundaries. Words have been the basic unit of indexing in traditional IR. As Chinese sentence are written as continuous character strings, a pre-processing has to be done to segment sentences into shorter units that may be used as indexes. The basic approaches of Chinese segmentation can be roughly divided into two groups, namely character-based approach and word-based approach (Schubert Foo and Hui Li, 2002).

Dictionary-based approach is a popular word-based approach for text segmentation. In this approach, segmented texts are matched against a dictionary prior to being indexed. Longest match algorithm is often used to solve segmentation ambiguities. Unknown word problem is one of the main problems of dictionary-based approach. Especially, many proper nouns, which play an important role in IR, are not in dictionary, and are not considered as indexed.

Character-based (or n-grams-based) approach does not require any linguistic knowledge. It segments texts into strings containing one (uni-gram) or two (bi-gram), or more characters. Since 75% of all available and commonly used Chinese words are made up of two characters (Wu, Z.M. and Tseng, G., 1993), bi-grams approach is an effective approach. The most obviously advantage is its simplicity and ease of application. On the other hand, it can skip the unknown word problem. For example, for proper noun that are not in the dictionary, such as 大亚湾 (a place in southern China), word segmentation will segment it into three characters, i.e. 大, 亚, and 湾. When using overlapping bi-grams, it will be segmented into two bi-grams, i.e. 大亚 and 亚湾. If both bi-grams occur in the same document, there is a higher probability that the document concerns 大亚湾, than the documents where the three single characters occur.

According to the experimental result of Microsoft Research China on TREC 5&6 Chinese data (Jianfeng Gao, 2001), they compared the IR performance of different approaches of Chinese segmentation. Combining the bi-grams with uni-grams, the average IR precision is 0.4254. Using longest match with large dictionary (220K entries), the precision is 0.3907. Using longest match, large dictionary and complementing longest words by single characters, the precision is 0.4290. If adding the unknown words recognition to the third approach, the precision is 0.4342.

We get the similar result in our experiment. For query text, we found that first segmenting text into word, then segmenting word into bi-grams will achieve better performance.

3.2 Chinese Monolingual Retrieval System based on Lucene toolkit

Our retrieval approach is based on the vector space mode. The similarity between the query q and the each document d_j is computed as the inner product or cosine of the angle between their associated vectors, as described in the Computation Formulae (1) and (2).

$$sim(d_j, q) = \vec{d}_j \cdot \vec{q} \tag{1}$$

$$sim(d_j, q) = \frac{\vec{d}_j \cdot \vec{q}}{|d_j| \times |q|} \tag{2}$$

The notion of TF-IDF is used for term weight. (3) and (4) are the best known formulae, where N is the number of documents in the collection, n_i is the number of documents containing term t_i and $freq_{i,j}$ is the frequency of term t_i appears in document d_j.

$$idf_i = \log \frac{N}{n_i + 1} + 1 \tag{3}$$

$$tf_{i,j} = \frac{freq_{i,j}}{doc\ length} \tag{4}$$

According to our experiment, we find that using $tf_{i,j} = \sqrt{\frac{freq_{i,j}}{doc\ length}}$ achieve the better performance. The nonlinear TF function is more close to the reality. Document term weight and query term weight are calculated using the formulae (5) and (6).

$$w_{i,j} = tf_{i,j} * idf_i \tag{5}$$

$$w_{i,q} = tf_{i,q} * idf_i, \quad tf_{i,q} = \sqrt{freq_{i,q}} \tag{6}$$

Because the presence of a large portion of the query terms indicates a better match with the

query, we define the *Cooccur*(d_j, q), as described in the Formula (7). This value is multiplied into scores.

$$Cooccur(d_j, q) = \frac{the\ number\ of\ query\ terms\ matched\ in\ the\ document}{the\ total\ number\ of\ terms\ in\ the\ query} \quad (7)$$

Here, we give the final score function, as shown in the Formula (8). We do not normalize d_j, because we have normalized d_j according to the length of document when calculating TF. The normalization of vector q does not affect ranking, but it makes scores from different queries comparable.

$$score(d_j, q) = \vec{d}_j \cdot \frac{\vec{q}}{|q|} * Cooccur(d_j, q) \quad (8)$$

Lucene is an open source toolkit for text indexing and searching (Brian Goetz, 2003; Jakarta Lucene Home Page). Its retrieval approach is based on vector space model. Our rank algorithm can easily be added to it. In our system, two parsers are used to index Chinese document. One is based on bi-grams approach and the other is based on word approach.

4. Evaluation

For an IR system, performance evaluation is important. Retrieval performance evaluation is usually based on a test reference collection (Ricardo Baeza-Yates and Berthier Ribeiro-Neto, 1999). The test reference collection consists of a collection of documents, a set of example information request, and a set of relevant documents for each example information request. In the real environment, i.e. very large collections, for special query, it is expensive or impossible to create sets of relevant judgments. We try an effective method to evaluate the system performance. For each topic, we create a "perfect query" for it. "Perfect query" is a manually modified query which can achieve high retrieval performance. Then we compare the results of Monolingual retrieval and CLIR retrieval with it. Although this method may lose some accuracy, it is inexpensive and can be used in real environment.

First step, we need to create the "perfect query". We use the translated Chinese queries of each topic as original query. After several times of query expansion, new terms can be added to the query and the weight of term can be adjusted. For the topic "China's Protection of Pandas" ("中国对熊猫的保护"), the "perfect query" may be "大熊猫 熊猫 国宝 野生 繁殖 四川 环境 濒危 保护".

Second step, we need to determine the number of relevant documents. We use a bi-search strategy. For example, we submit the "perfect query" to the IR system and there are N documents returned by system. Near the position of N/2, we select 10 documents. If more than 4 of these 10 documents are related to the topic, we continue to examine the documents at the position of 3/4N. Otherwise, we go to the position of 1/4N. Finally, we converge to a place n. The first n documents are regarded as relevant documents.

Our CLIR Evaluation task is based on prepared English Topics and Chinese collections. The first contains 25 English topics. Each topic includes title, description, narration and corresponding Chinese translation. The second contains three Chinese Hong Kong news sets -- HKCD, HKDN and TKP, totally 127,938 documents. We generate the "perfect query" for

each topic. Then it can be used as judgment. Many classical evaluation algorithms can still be used.

We use two kinds of index units for Chinese document, bi-grams and words. To compare the influence of different kind of index units, we test the Chinese long queries (include title and description) and get two monolingual results. Then we test English long queries and get two CLIR results. Figure 2 shows the recall and precision curve of the monolingual IR and CLIR performance of our system under different index unit. Considered users preferring to use short query, we do the same test on short query (only use title). Table 1 give the Comparison results of long query and short query.

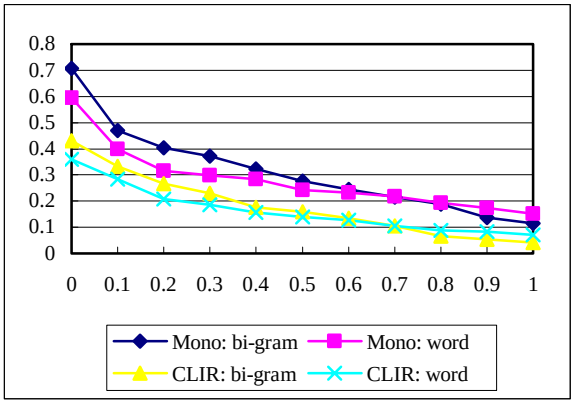

Figure 2. Recall vs. precision curve on long query

	Long query (<TITLE> + <DESC>)	Short query (<TITLE> only)
Mono: bi-gram	0.3140	0.2546
Mono: word	0.2822	0.2364
CLIR: bi-gram	0.1813	0.1527
CLIR: word	0.1643	0.1428

Table 1. Comparison results of different query length

The results show that using bi-grams as index unit is better than words. Long query is better than short query. The performance of the cross-lingual retrieval is about 58% of the monolingual performance. The main reason is that some key concept terms in some topics were either not translated at all or improperly translated due to the limited coverage of the bilingual wordlist we used or improperly translated. The second reason is that the Chinese corpus we used is from Hong Kong news while the dictionary we used is in Chinese mainland style. There are some mismatches between some concept terms. It increases the difficult of query translation.

5. Conclusion

In this paper, we explored English-Chinese CLIR. On Chinese monolingual retrieval, we found that using bi-gram indexing for documents will achieve better result. The main

performance-limiting factor is the limited coverage of the dictionary used in query translation. Some of the key concepts were either improperly translated or not translated. If there are no sets of relevant judgments, manually modified queries can be used to evaluate the performance of system.

Acknowledgement

This paper is supported by National Natural Science Foundation of China (no. 60203010, no.70501018, no. 60533100) and "211 Project" of Shanghai University of Finance & Economics (2004).

References

Brian Goetz, 2003, The Lucene search engine: Powerful, flexible, and free, http://www.javaworld.com.

Jakarta Lucene Home Page, http://jakarta.apache.org/lucene/.

Jianfeng Gao, 2001, An empirical study of CLIR at MSRCN, *International workshop ILT&CIP-2001 on Innovative language technology and Chinese information processing*, Shanghai, April 6-7.

Ricardo Baeza-Yates and Berthier Ribeiro-Neto, 1999, Modern Information Retrieval, Addison-Wesley.

Schubert Foo and Hui Li, 2002, Chinese word segmentation and its effect on information retrieval, *Information Processing & Management*.

Wu, Z.M. and Tseng, G., 1993, Chinese text segmentation for text retrieval: Achievements and problems, *Journal of the American Society for Information Science*, vol. 44, no. 1, pp. 532-542.