Translation & Transform Algorithm of Query Sentence in Cross-Language Information Retrieval*

Xiao-fei ZHANG[1] Ke-liang ZHANG[1, 2] He-yan HUANG[1]

[1] Research Center of Computer & Language Information Engineering, CAS, Beijing 100097
[2] Nanjing University of Science and Technology, Nanjing 210049
{xiaofei.zhang, keliang.zhang, heyan.huang}@hjtek.com

Abstract. Based on large-scale bilingual corpora and the theories of vector space model and lexical mutual information, this paper explores the application of the traditional monolingual IR technology to converting the translation of query sentence to the computation of the *boost* value of query keyword translations in the bilingual dictionary, so that the target language query sentence is reconstructed. The experiment finds a 92.8% precision rate in the first 10 retrieved documents and an 88.9% precision rate in the first 100 retrieved documents.

Keywords: cross-language information retrieval; query sentence; translation & transform algorithm

1 Introduction

Cross-language information retrieval (CLIR) can be briefly described as follows: (1) The retrieval request is posed in a given language; (2) The computer automatically searches the documents written in another language or other languages; (3) The retrieval results can be delivered, by way of automatic translation, in the language specified by the user. CLIR integrates the traditional monolingual information retrieval technology with machine translation (MT) technology. It is becoming a worldwide subject of crucial importance in the Information Age, as proved in some way by the fact that CLIR is included from time to time as an important subtask in the annual event of Text Retrieval Conference (TREC).

In CLIR systems, the query sentence is usually input as a combination of a series of keywords rather than a sentence in its exact sense. Due to the absence in the series of query keywords of necessary syntactic and semantic information, traditional MT technology [1, 2] can not be readily used for the precise translation of the query sentence. In this paper, a new translation and transform algorithm of query sentence is proposed. This approach is based on large-scale corpora, traditional monolingual IR technology and the theories of vector space model (VSM) and lexical mutual information [3, 4].

2 Query and Query Transform

In cross-language information retrieval, the input is often a combination of a series of keywords rather than a complete sentence. This sequence of query keywords lacks necessary contextual and syntactico-semantic information, so they can not be translated by traditional MT technology in a direct and easy way. Neither can the translation problem be resolved by simply looking up the bilingual dictionary. For example, the English word “bank” corresponds to two Chinese meanings in a typical English-Chinese dictionary: “银行” and “河岸”. Then the problem arises: should the word “bank” be translated into “银行” or “河岸”?

In common sense, when the user inputs {bank; credit} as a query, we may well conclude that he is most likely looking for information about “银行” and “信用”, although we cannot completely exclude

* This research is supported by the National Science Foundation of China under Grant No. 60502048.

the small possibility of interpreting “bank” as the meaning of “河岸” in this context. Therefore, the retrieval algorithm should rank among all the retrieval results the documents containing the keyword “银行” before those containing the keyword “河岸”.

The translation and transform of query sentence in CLIR is formalized as follows:

Suppose there is a SL keyword query sentence:

$$Query_S = W_{S1}\ W_{S2} \cdots W_{Si} \cdots W_{SN} \qquad (1)$$

where W_{Si} represents the i^{th} keyword in the query sentence, for example, the keywords “bank” and “credit” in the query instance “bank credit”.

The SL keyword query sentence is translated and transformed into a TL keyword query sentence, which is represented as follows:

$$\begin{aligned} Query_T = & (W_{T11} \wedge boost_{T11} \quad W_{T12} \wedge boost_{T12} \cdots W_{T1N} \wedge boost_{T1N}) \\ & \cdots (W_{Ti1} \wedge boost_{Ti1}\ W_{Ti2} \wedge boost_{Ti2} \cdots W_{TiN} \wedge boost_{TiN}) \cdots \\ & (W_{TN1} \wedge boost_{TN1}\ W_{TN2} \wedge boost_{TN2} \cdots W_{TNN} \wedge boost_{TNN}) \end{aligned} \qquad (2)$$

where $W_{Ti1}\ \ W_{Ti2} \cdots W_{TiN}$ are the translations in the bilingual dictionary of the keyword W_{Si} of the source language keyword query sentence, and $boost_{Ti1}\ \ boost_{Ti2} \cdots boost_{TiN}$ are the weight of each translation of the keyword W_{Si} in the bilingual dictionary, which is referred to as *boost* value.

3 Algorithm Analysis

3.1 Basic Theory of Information Retrieval

For an IR system, the kernel problem is ranking, i.e., the so-called relevance computing. Based on vector space model, a relevance computation formula is defined as follows:

$$score(q, d_i) = \frac{\sum_{t \in q} tf(t\ in\ d_i)\, idf(t)\, boost(t)}{\max(score(q, d_i))} \qquad (3)$$

where q represents query sentence, d_i represents the i th document and t represents a query keyword. The denominator $\max(score(q, d_i))$ is a normal factor which does not affect the ranking result but makes comparable the relevance values of different queries. The addition of the normal factor is for the convenience of computing the *boost* value when query keywords are translated and transformed (to be discussed later). In addition, the $idf(t)$ in (3) is defined as

$$idf(t) = -\log p(t) = -\log\left(\frac{N}{M}\right) \qquad (4)$$

where M represents the total number of the documents, and N represents the number of the documents which contain the keyword t.

3.2 *Boost* Value Computation

How to compute the *boost* value of the translations in the bilingual dictionary of a given query keyword is the focus of the algorithm proposed in this paper. Our approach is based on large-scale bilingual corpora, and the computation is implemented by applying the theories of VSM and lexical mutual information to traditional IR.

Supporting Knowledge Base. In addition to the bilingual dictionary, the translation and transform of the query sentence also make use of a large-scale bilingual corpus of aligned sentence pairs. It is meant

to take the bilingual sentence pair as the basic retrieval unit. The corpus used in the experiment contains altogether 162,918 English-Chinese sentence pairs, which amount to approximately two million English words and two million Chinese words. This aligned bilingual corpus is used as the retrieval source for *boost* value computation.

***Boost* Value Computation.** Suppose there is a sequence of SL query keywords:

$$W_{S1}\ W_{S2}\ \cdots W_{Si}\ \cdots W_{SN}$$

and the corresponding sequence of the TL translations of the SL query keywords according to their respective SL entries in the bilingual dictionary:

$$(W_{T11}\ W_{T12}\ \cdots W_{T1N})\cdots(W_{Ti1}\ W_{Ti2}\ \cdots W_{TiN})\cdots(W_{TN1}\ W_{TN2}\ \cdots W_{TNN})$$

Three query sentences are designed for the purpose of *boost* value computation:

$$Query_1 = (W_{Si}\ AND\ W_{Tij})\ AND\ (W_{S1}\ AND\ W_{S2}\cdots AND\ W_{SN}) \quad (5)$$

$$Query_2 = (W_{Si}\ AND\ W_{Tij})\ AND\ (W_{S1}\ OR\ W_{S2}\cdots OR\ W_{SN}) \quad (6)$$

$$Query_3 = (W_{Si}\ AND\ W_{Tij})\ OR\ (W_{S1}\ OR\ W_{S2}\cdots OR\ W_{SN}) \quad (7)$$

Hence the formula for the computation of the *boost* value of W_{Tij}:

$$boost(W_{Tij}) = 2^{\alpha}\times mean(score(q,d_i)) + \beta \quad (8)$$

where $mean(score(q,d_i))$ represents the average relativity of the retrieval results. The relativity computation formula is defined as (3) in 3.1. Also in formula (8), β is equivalent to the datum value of transform and is set to be 0.5; α is the weight coefficient of the query sentence and is determined in the following way:

$$\alpha = \begin{cases} 3, & if\ q == Query_1 \\ 2, & else\ if\ q == Query_2 \\ 1, & else\ if\ q == Query_3 \\ 0, & \text{else} \end{cases} \quad (9)$$

4 Experiment Design and Results

4.1 Test Corpus

We have randomly selected from the Internet some Chinese websites to download the material we need for the experiment. The websites include www.xinhua.com, www.sohu.com, www.sina.com, among others. The downloaded documents amount to 7 gigabyte, consisting of 138,908 webpages.

4.2 Experiment Scheme

The experiment aims to test the effectiveness of the translation and transform algorithm presented in this paper. For this purpose, it is designed to include the following steps:

Step 1: Design 100 English query sentences with well-chosen keywords. In each of the query sentences, there is at least one query keyword that has clear distinction between its Chinese translations.

Step 2: Subject the 100 query sentences to the CLIR system for cross-language retrieval between English and Chinese.

Step 3: Evaluate artificially the retrieval results. In the experiment, only the first 10 and 100 of the retrieved documents are respectively evaluated. The evaluation is conducted by the following criterion: if a given document contains the correct translation of the query keyword, it is a true retrieval result; or else it is a false retrieval result.

Step 4: Repeat Step 2 and Step 3 for a contrastive experiment. The contrastive experiment is conducted by looking up the English-Chinese dictionary instead of using the translation and transform algorithm. This is referred to as the dictionary lookup approach.

4.3 Results and Analysis

The results of the contrastive experiments are shown in the following table. For the 100 queries, when the first 10 retrieved documents are considered, the number of true retrieval results is 928, and the precision rate is 92.8% which is 15.4% higher than that of the dictionary lookup approach. When the first 100 retrieved documents are considered, the number of true retrieval results is 8891, and the precision rate is 88.9% which is 13.0% higher than that of the dictionary lookup approach. This shows that the translation and transform algorithm is fairly effective.

Tab. Contrastive Experiment Results

Approach	No. of true retrievals for the first 10 documents	p@10	No. of true retrievals for the first 100 documents	p@100
Translation and transform algorithm	928	92.8%	8891	88.9%
Dictionary lookup	774	77.4%	7593	75.9%

5 Conclusion

In this paper, based on large-scale bilingual corpora and the theories of VSM and lexical mutual information, the traditional monolingual IR technology is applied to converting the translation of query sentence to the computation of the "boost" value of query keyword translations in the bilingual dictionary, so that the TL query sentence is reconstructed. The experiment finds a 92.8% precision in the first 10 retrieved documents and an 88.9% precision in the first 100 retrieved documents, which shows the effectiveness of the translation and transform algorithm.

The translation and transform algorithm presented in this paper requires a large-scale bilingual corpus as its supporting knowledge base. However, it is always a tough job to build a large-scale high-quality bilingual corpus. In particular, when it comes to minority languages, the cost becomes almost unbearably expensive. In our further study, some parameter smoothening mechanism will be introduced so that the algorithm can work effectively when a supporting large-scale bilingual corpus is not available.

References

1. Sadat, F., Yoshikawa, M., Uemura, S.: Bilingual Terminology Acquisition from Comparable Corpora and Phrasal Translation to Cross-Language Information Retrieval [A]. In Proceedings of the 41st Annual Meeting of the Association for Computational Linguistics [C], Sapporo, Japan, July 7-12, 2003
2. Sadat, F., Yoshikawa, M., Uemura, S.: Learning Bilingual Translations from Comparable Corpora to Cross-Language Information Retrieval: Hybrid Statistics-based and Linguistics-based Approach [A]. In Proceedings of the Sixth International Workshop on Information Retrieval with Asian Languages [C]. Sapporo, Japan. July 7, 2003
3. Zhang, Min, Ma, Shao-ping, Song, Rui-hua: DF or IDF? Using Dependent Feature Model in Web-based Information Retrieval [J]. Journal of Software. 2005, Vol.16, No.5: 1012-1020
4. Zhang, Xiao-fei, Chen, Zhao-xiong, Huang, He-yan et al.: Retrieval Approach and Candidate Translation Examples in Interactive Hybrid Strategies Machine Translation System IHSMTS [J]. Mini-Micro Systems, 2005, VOL.26, No.3: 330-334