Example-Based Phrase Translation in Chinese-English CLIR

Bin Wang Xueqi Cheng and Shuo Bai
Institute of Computing Technology, Chinese Academy of Sciences
No. 6, Kexueyuan South Road, Zhongguanchun
Beijing 100080 China
+86 010 62587953
{Wangbin,cxq,bai}@ict.ac.cn

ABSTRACT
This paper proposes an example-based phrase translation method in a Chinese to English cross-language information retrieval (CLIR) system. The method can generate much more accurate query translations than dictionary-based and common MT-based methods, and then improves the retrieval performance of our CLIR system.

Categories and Subject Descriptors
H.3.3 [**Information Storage and Retrieval**]: Information Search and Retrieval – *query formulation.*

General Terms
Design, Experimentation.

Keywords
CLIR, Example-Based, Phrase Translation.

1. INTRODUCTION
CLIR is to retrieve documents in one language (target language) providing queries in another language (source language). Most present CLIR methods fall into three categories: dictionary-based, MT-based and corpus-based methods[1]. Although different resources or techniques are used, all these methods try to generate the best target queries.

Example-based method is one of the main approaches used in machine translation (MT) field. In EBMT, old translated pairs are aligned and stored in the example base (EB), and when translating a new object, the most similar example in the EB will be selected to generate the translation. Compared with the other methods used in MT, example-based method can generate much better translation results for the objects that have most similar examples in the EB. However, the similarity between sentences is very difficult to define and compute; this greatly degrades the usability of the method. On the contrary, the similarity computation between phrases seems simpler. As we know, in real world, IR queries are often one to three words long and most of them are phrases. Therefore, when translating these queries, we use example-based method that may generate accurate translations.

SIGIR'02, August 11-15, 2002, Tampere, Finland.
ACM 1-58113-561-0/02/0008. …$5.00.

2. ALIGNMENT
In order to be used in later translation, all translated phrases are aligned at the word level at first[2].

Supposing the Chinese phrase C is $C_1C_2...C_m$, the corresponding English phrase E is $E_1E_2...E_n$, where each C_i ($1 \leqslant i \leqslant m$) or E_j ($1 \leqslant j \leqslant n$) is a word. Four methods are combined[2] to align a Chinese word with an English word. Other alignment modes such as 1-*t, s-t*, s-0, 0-*t* (*s,t*>1)are also considered in our system.

(1) Alignment based on dictionary

For each word in Chinese or English phrase, if one of its translations in the C-E or E-C dictionary appears in the other phrase, the word and its translation can be aligned with each other.

(2) Alignment based on location information

In alignment based on dictionary, one word may have more than one alignment word. This kind of ambiguities can be solved via location(word order) information. The candidate alignment word that has smallest *LocDist* will be selected as the alignment result. *LocDist* between location i in source phrase and location j in target phrase is defined as:

$$LocDist(i,j)=min(|Slope_L-1|,|Slope_R-1|)$$

Where $Slope_L=(j-j_L)/(i-i_L)$, $Slope_R=(j-j_R)/(i-i_R)$, (i_L,j_L) is the closest reliable alignment on the left of i, and (i_R,j_R) is the closest reliable alignment on the right of i.

(3) Alignment based on semantic distance

Because of the limited coverage of bilingual dictionaries, we try to align words based on their semantic similarity. In our system, a Chinese thesaurus named *CILIN*[3] and an English-Chinese dictionary is used to define the semantic distance between Chinese and English words. *CILIN's* categories can be represented as a hierarchical tree as follows:

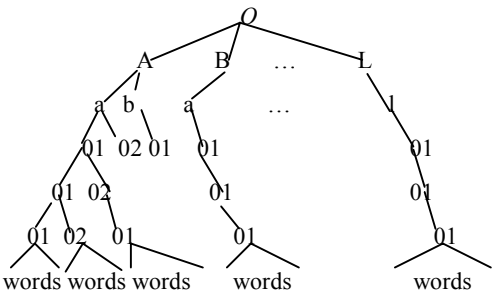

Figure 1. *CILIN's* categories and codes.

The semantic distance between two Chinese words can be defined as the reciprocal of the shortest path length between the two words within the above tree. And the semantic distance between a Chinese word and an English word is defined as the minimal distance between the Chinese word and all the Chinese translation words of the English word in the English-Chinese dictionary.

(4) Alignment based on co-occurrence frequency

The word pair co-occurrence frequency information can be also used to align words throughout the example base collection. A word pair with high co-occurrence frequency can be marked as an alignment.

The above alignment also makes a useful statistics-based bilingual dictionary for later use.

3. QUERY TRANSLATION

To find the most similar example to the input Chinese query phase in the EB can be also regarded as an IR process. So some IR techniques are used here.

3.1 Inverted Index

Inverted file data structure is used here to store the example base. Each example can be represented by its words. The word order and word count information within a phrase are also recorded in the posting files. For instance, the phrase "模式识别" should be represent as "模式 1 2" and "识别 2 2" in the corresponding posting files. To speed up the retrieval process, the index file is organized using a hashing function.

3.2 Phrase Matching

When a new phrase is submitted to the example base, the most similar phrase will be selected to generate the translation.

First, the words and word counts in the new phrase are used to picked out those candidate phrases, then the similarity between the new phrase and each candidate phrase is computed as follows:

$$dis(C^{'}, C) = \prod_{i=1}^{m} dis(C_i^{'}, C_i)$$

Where $dis(C_i^{'}, C_i)$ is the semantic distance between $C_i^{'}$ and C that can be computed as in section 2. Our goal is to get the C that minimizes the above distance. At the same time, the minimal distance must be lower than a preset threshold θ.

Sometimes, the new phrase may be longer than all the phrases in the example base, or we cannot find a suitable phrase whose similarity is above θ. Then we can segment the phrase into smaller chunks to use the above formula recursively. The semantics of the last word within a phrase is used to represent the semantics of the whole phrase.

3.3 Phrase Translation

In EBMT, the translation process is very simple. Each translation pair can be viewed as a translation template and the target phrase can be used to generate the translation result through a few words substituting some words. For example, the phrase "文本检索"can be translated into "text retrieval" via translation pair "情报[1] 检索[2]/information retrieval". What we should do is to pick out the best translation of "文本" in the C-E dictionary and put it in the corresponding location of the target phrase. While translating, part-of-speech information should also be considered.

4. EXPERIMENTAL RESULTS

In our experiments, we have about 700,000 phrase pairs that are mostly science and technology terms, most of them are collected from different on-line dictionaries or bilingual corpora. About 69% of them are two-word or three-word phrases. At first, we segment all the Chinese phrases into words, and then we align them with corresponding English phrases. Necessary manual work is also done to correct the misaligned results.

100 multi-word Chinese query phrases are selected from a WEB IR log file. 77 of them are 2-word phrases, 15 are 3-word phrases, and the others are more than 3 words long.

Besides the above phrase translation method, we also use another two methods in our Chinese-English CLIR system: CEMT-based method and dictionary-based method. In CEMT-based method, we use a CEMT system named *TransEasy*[4] to translate the queries into English. In dictionary-based method, we use the statistics-based dictionary to map the queries into English via dictionary lookup.

Out of the 100 queries, our method generates 78 completely correct translation results(θ*=0.01)*, while the num for CEMT-based method and dictionary-based method are respectively 47 and 53. 14 translation results using our method are partially correct and the others cannot be translated because of the unknown words. These results confirm our expectation.

5. CONCLUSIONS

Most IR queries are quite short, i.e., they are most words or phrases. Therefore the main task in CLIR is not translating sentences but translating phrases. Example-based method can provide very good translation results but the similarity computation between sentences is quite complex. Based on the above consideration, we apply example-based query phrase translation in our Chinese-English CLIR system, and the experiments achieve good results.

In our near future, we will apply a chunk recognizer to process more than one phrase long queries and we may cluster all the examples to reduce the phrase search time.

6. REFERENCES

[1] Yiming Yang, Jaime G. Carbonell, Ralf D. Brown, Robert E. Frederking,. Translingual Information Retrieval : Learning from Bilingual Corpora. AI Journal special issue: Best of IJCAI-97.

[2] Bin Wang, Qun Liu, Xiang Zhang, Word Alignment on Chinese-English Bilingual Corpora, in Proceedings of JSCL-99 (Beijing, November 1999), 123-128.

[3] Mei, Jia-Ju, Yi-Ming Zhu, Yun-Qi Gao, Hong-Xiang Yin, Tongyici CiLin (Chinese Synonym Forest), Shanghai Press of Lexicon and Books, 1983 (In Chinese).

[4] Qun Liu, ShiWen Yu, *TransEasy*: A Chinese-English MT system Based on Hybrid Approach, in Proceedings of AMTA'98 (Langhorne, PA. USA, October 1998), 514-517