Automatic construction of parallel English-Chinese corpus for cross-language information retrieval

Jiang Chen and Jian-Yun Nie
Département d'Informatique et Recherche Opérationnelle
Université de Montréal
C.P. 6128, succursale CENTRE-VILLE
Montreal (Quebec), Canada H3C 3J7
{*chen, nie*}*@iro.umontreal.ca*

Abstract

A major obstacle to the construction of a probabilistic translation model is the lack of large parallel corpora. In this paper we first describe a parallel text mining system that finds parallel texts automatically on the Web. The generated Chinese-English parallel corpus is used to train a probabilistic translation model which translates queries for Chinese-English cross-language information retrieval (CLIR). We will discuss some problems in translation model training and show the preliminary CLIR results.

1 Introduction

Parallel texts have been used in a number of studies in computational linguistics. Brown et al. (1993) defined a series of probabilistic translation models for MT purposes. While people may question the effectiveness of using these models for a full-blown MT system, the models are certainly valuable for developing translation assistance tools. For example, we can use such a translation model to help complete target text being drafted by a human translator (Langlais et al., 2000).

Another utilization is in cross-language information retrieval (CLIR) where queries have to be translated from one language to another language in which the documents are written. In CLIR, the quality requirement for translation is relatively low. For example, the syntactic aspect is irrelevant. Even if the translated word is not a true translation but is strongly related to the original query, it is still helpful. Therefore, CLIR is a suitable application for such a translation model.

However, a major obstacle to this approach is the lack of parallel corpora for model training. Only a few such corpora exist, including the Hansard English-French corpus and the HKUST English-Chinese corpus (Wu, 1994). In this paper, we will describe a method which automatically searches for parallel texts on the Web. We will discuss the text mining algorithm we adopted, some issues in translation model training using the generated parallel corpus, and finally the translation model's performance in CLIR.

2 Parallel Text Mining Algorithm

The PTMiner system is an intelligent Web agent that is designed to search for large amounts of parallel text on the Web. The mining algorithm is largely language independent. It can thus be adapted to other language pairs with only minor modifications.

Taking advantage of Web search engines as much as possible, PTMiner implements the following steps (illustrated in Fig. 1):

1 Search for candidate sites – Using existing Web search engines, search for the candidate sites that may contain parallel pages;

2 File name fetching – For each candidate site, fetch the URLs of Web pages that are indexed by the search engines;

3 Host crawling – Starting from the URLs collected in the previous step, search through each candidate site separately for more URLs;

4 Pair scan – From the obtained URLs of each site, scan for possible parallel pairs;

5 Download and verifying – Download the parallel pages, determine file size, language, and character set of each page, and filter out non-parallel pairs.

2.1 Search for candidate Sites

We take advantage of the huge number of Web sites indexed by existing search engines in determining candidate sites. This is done by submitting some particular requests to the search engines. The requests are determined according to the following observations. In the sites where parallel text exists, there are normally some pages in one language containing links to the parallel version in the other language. These are usually indicated by those links' anchor texts [1]. For example, on some English page there may be a link to its Chinese version with the anchor text "Chinese Version" or "in Chinese".

[1] An anchor text is a piece of text on a Web page which, when clicked on, will take you to another linked page. To be helpful, it usually contains the key information about the linked page.

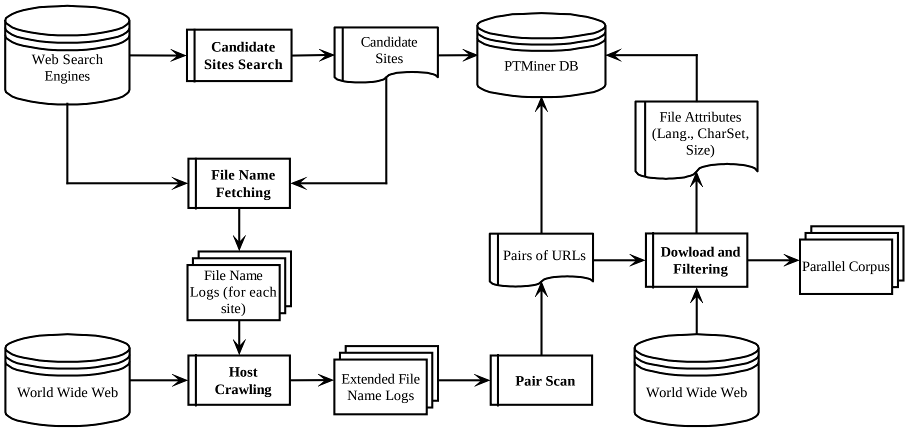

Figure 1: The workflow of the mining process.

The same phenomenon can be observed on Chinese pages. Chances are that a site with parallel texts will contain such links in some of its documents. This fact is used as the criterion in searching for candidate sites.

Therefore, to determine possible sites for English-Chinese parallel texts, we can request an English document containing the following anchor:

$$anchor: "english\ version"\ ["in\ english", ...].$$

Similar requests are sent for Chinese documents.

From the two sets of pages obtained by the above queries we extract two sets of Web sites. The union of these two sets constitutes then the candidate sites. That is to say, a site is a candidate site when it is found to have either an English page linking to its Chinese version or a Chinese page linking to its English version.

2.2 File Name Fetching

We now assume that a pair of parallel texts exists on the same site. To search for parallel pairs on a site, PTMiner first has to obtain all (or at least part of) the HTML file names on the site. From these names pairs are scanned. It is possible to use a Web crawler to explore the candidate sites completely. However, we can take advantage of the search engines again to accelerate the process. As the first step, we submit the following query to the search engines:

$$host: hostname$$

to fetch the Web pages that they indexed from this site. If we only require a small amount of parallel texts, this result may be sufficient. For our purpose, however, we need to explore the sites more thoroughly using a host crawler. Therefore, we continue our search for files with a host crawler which uses the documents found by the search engines as the starting point.

2.3 Host Crawling

A host crawler is slightly different from a Web crawler. Web crawlers go through innumerable pages and hosts on the Web. A host crawler is a Web crawler that crawls through documents on a given host only. A breadth-first crawling algorithm is applied in PTMiner as host crawler. The principle is that when a link to an unexplored document on the same site is found in a document, it is added to a list that will be explored later. In this way, most file names from the candidate sites are obtained.

2.4 Pair Scan

After collecting file names for each candidate site, the next task is to determine the parallel pairs. Again, we try to use some heuristic rules to guess which files may be parallel texts before downloading them. The rules are based on external features of the documents. By external feature, we mean those features which may be known without analyzing the contents of the file, such as its URL, size, and date. This is in contrast with the internal features, such as language, character set, and HTML structure, which cannot be known until we have downloaded the page and analyzed its contents.

The heuristic criterion comes from the following observation: We observe that parallel text pairs usually have similar name patterns. The difference between the names of two parallel pages usually lies in a segment which indicates the language. For example, "file-ch.html" (in Chinese) vs. "file-en.html" (in English). The difference may also appear in the path, such as ".../chinese/.../file.html" vs. ".../english/.../file.html". The name patterns described above are commonly used by webmasters to help organize their sites. Hence, we can suppose that a pair of pages with this kind of pattern are probably parallel texts.

First, we establish four lists for English prefixes, English suffixes, Chinese prefixes and Chinese suffixes. For example: $English\ Prefix = \{e, en, e\_, en\_, e-, en-, ...\}$. For each file in one language, if a segment in its name corresponds to one of the language affixes, several new names are generated by changing the segment to the possible corresponding affixes of the other language. If a generated name corresponds to an existing file, then the file is considered as a candidate parallel document of the original file.

2.5 Filtering

Next, we further examine the contents of the paired files to determine if they are really parallel according to various external and internal features. This may further improve the pairing precision. The following methods have been implemented in our system.

2.5.1 Text Length

Parallel files often have similar file lengths. One simple way to filter out incorrect pairs is to compare the lengths of the two files. The only problem is to set a reasonable threshold that will not discard too many good pairs, i.e. balance recall and precision. The usual difference ratio depends on the language pairs we are dealing with. For example, Chinese-English parallel texts usually have a larger difference ratio than English-French parallel texts. The filtering threshold had to be determined empirically, from the actual observations. For Chinese-English, a difference up to 50% is tolerated.

2.5.2 Language and Character Set

It is also obvious that the two files of a pair have to be in the two languages of interest. By automatically identifying language and character set, we can filter out the pairs that do not satisfy this basic criterion. Some Web pages explicitly indicate the language and the character set. More often such information is omitted by authors. We need some language identification tool for this task.

SILC is a language and encoding identification system developed by the RALI laboratory at the University of Montreal. It employs a probabilistic model estimated on tri-grams. Using these models, the system is able to determine the most probable language and encoding of a text (Isabelle et al., 1997).

2.5.3 HTML Structure and Alignment

In the STRAND system (Resnik, 1998), the candidate pairs are evaluated by aligning them according to their HTML structures and computing confidence values. Pairs are assumed to be wrong if they have too many mismatching markups or low confidence values.

Comparing HTML structures seems to be a sound way to evaluate candidate pairs since parallel pairs usually have similar HTML structures. However, we also noticed that parallel texts may have quite different HTML structures. One of the reasons is that the two files may be created using two HTML editors. For example, one may be used for English and another for Chinese, depending on the language handling capability of the editors. Therefore, caution is required when measuring structure difference numerically.

Parallel text alignment is still an experimental area. Measuring the confidence values of an alignment is even more complicated. For example, the alignment algorithm we used in the training of the statistical translation model produces acceptable alignment results but it does not provide a confidence value that we can "confidently" use as an evaluation criterion. So, for the moment this criterion is not used in candidate pair evaluation.

3 Generated Corpus and Translation Model Training

In this section, we describe the results of our parallel text mining and translation model training.

3.1 The Corpus

Using the above approach for Chinese-English, 185 candidate sites were searched from the domain *hk*. We limited the mining domain to *hk* because Hong Kong is a bilingual English-Chinese city where high quality parallel Web sites exist. Because of the small number of candidate sites, the host crawler was used to thoroughly explore each site. The resulting corpus contains 14820 pairs of texts including 117.2Mb Chinese texts and 136.5Mb English texts. The entire mining process lasted about a week. Using length comparison and language identification, we refined the precision of the corpus to about 90%. The precision is estimated by examining 367 randomly picked pairs.

3.2 Statistical Translation Model

Many approaches in computational linguistics try to extract translation knowledge from previous translation examples. Most work of this kind establishes probabilistic models from parallel corpora. Based on one of the statistical models proposed by Brown et al. (1993), the basic principle of our translation model is the following: given a corpus of aligned sentences, if two words often co-occur in the source and target sentences, there is a good likelihood that they are translations of each other. In the simplest case (model 1), the model learns the probability, $p(t|s)$, of having a word t in the translation of a sentence containing a word s. For an input sentence, the model then calculates a sequence of words that are most probable to appear in its translation. Using a similar statistical model, Wu (1995) extracted a large-scale English-Chinese lexicon from the HKUST cor-

<s id="0000"> <HTML> <HEAD> <META HTTP-EQUIV="Content-type" CONTENT="text/html; charset=iso-8859-1"> <META HTTP-EQUIV="Content-language" CONTENT="Western"> </s>	<s id="0000"> <HTML> <HEAD> <META HTTP-EQUIV="Content-type" CONTENT="text/html; charset=big5"> <META HTTP-EQUIV="Content-language" CONTENT="zh"> </s>
<s id="0001"> <TITLE>Journal of Primary Education 1996, Vol., No. 1&2, pp. 19-27 </TITLE> </HEAD> </s>	<s id="0001"> <TITLE> Journal of Primary Education 1996, Vol., No. 1&2, Page 19-27 </TITLE> </HEAD> </s>
<s id="0002"> <BODY BACKGROUND="../gif/pejbg.jpg" TEXT="#000000" BGCOLOR="#ffffff"> <CENTER> </s> <s id="0003"> <H1>Journal of Primary Education </H1> </s> <s id="0004"> <HR> <B>Volume 6, No 1&2, pp. 19-27 (May, 1996) </B> <HR> </s>	<s id="0002"> <BODY BACKGROUND="../gif/pejbg.jpg" TEXT="#000000" BGCOLOR="#ffffff"> <A HREF="/en/pej/b2g_pej.phtml?URL=%2fen%2fpej%2f0601%2f0601019c.htm"> <IMG SRC="/en/gif/kan.gif" ALT="简体" BORDER=0 ALIGN=R IGHT> </A> <CENTER> </s>
<s id="0005"> <H3>Principles for Redesigning Teacher Education </H3> Alan TOM </CENTER> </s>	<s id="0003"> <H2>初等教育学报</H2> </s> <s id="0004"> <HR>（一九九六年五月）第六卷. </s>
<s id="0006"> <P> <B> <I> Abstract </I> </B> </s>	<s id="0005"> 第一及二期19-27页 <HR> </s>

Figure 2: An alignment example using pure length-based method.

pus which is built manually. In our case, the probabilistic translation model will be used for CLIR. The requirement on our translation model may be less demanding: it is not absolutely necessary that a word t with high $p(t|s)$ always be a true translation of s. It is still useful if t is strongly related to s. For example, although "railway" is not a true translation of "train" (in French), it is highly useful to include "railway" in the translation of a query on "train". This is one of the reasons why we think a less controlled parallel corpus can be used to train a translation model for CLIR.

3.3 Parallel Text Alignment

Before the mined documents can be aligned into parallel sentences, the raw texts have to undergo a series of some preprocessing, which, to some extent, is language dependent. For example, the major operations on the Chinese-English corpus include encoding scheme transformation (for Chinese), sentence level segmentation, parallel text alignment, Chinese word segmentation (Nie et al., 1999) and English expression extraction.

The parallel Web pages we collected from various sites are not all of the same quality. Some are highly parallel and easy to align while others can be very noisy. Aligning English-Chinese parallel texts is already very difficult because of the great differences in the syntactic structures and writing systems of the two languages. A number of alignment techniques have been proposed, varying from statistical methods (Brown et al., 1991; Gale and Church, 1991) to lexical methods (Kay and Röscheisen, 1993; Chen, 1993). The method we adopted is that of Simard et al. (1992). Because it considers both length similarity and cognateness as alignment criteria, the method is more robust and better able to deal with noise than pure length-based methods. Cognates are identical sequences of characters in corresponding words in two languages. They are commonly found in English and French. In the case of English-Chinese alignment, where there are no cognates shared by the two languages, only the HTML markup in both texts are taken as cognates. Because the HTML structures of parallel pages are normally similar, the markup was found to be helpful for alignment.

To illustrate how markup can help with the alignment, we align the same pair with both the pure length-based method of Gale & Church (Fig. 2), and the method of Simard et al. (Fig. 3). First of all, we observe from the figures that the two texts are

<s id="0000"> <HTML> <HEAD> <META HTTP-EQUIV="Content-type" CONTENT="text/html; charset=iso-8859-1"> <META HTTP-EQUIV="Content-language" CONTENT="Western"> </s>	<s id="0000"> <HTML> <HEAD> <META HTTP-EQUIV="Content-type" CONTENT="text/html; charset=big5"> <META HTTP-EQUIV="Content-language" CONTENT="zh"> </s>
<s id="0001"> <TITLE>Journal of Primary Education 1996, Vol., No. 1&2, pp. 19-27 </TITLE> </HEAD> </s>	<s id="0001"> <TITLE> Journal of Primary Education 1996, Vol., No. 1&2, Page 19-27 </TITLE> </HEAD> </s>
<s id="0002"> <BODY BACKGROUND="../gif/pejbg.jpg" TEXT="#000000" BGCOLOR="#ffffff"> <CENTER> </s>	<s id="0002"> <BODY BACKGROUND="../gif/pejbg.jpg" TEXT="#000000" BGCOLOR="#ffffff"> <A HREF="/en/pej/b2g_pej.phtml?URL=%2fen%2fp ej%2f0601%2f0601019c.htm"> <IMG SRC="/en/gif/kan.gif" ALT="简体" BORDER=0 ALIGN=R IGHT> </A> <CENTER> </s>
<s id="0003"> <H1>Journal of Primary Education </H1> </s>	<s id="0003"> <H2>初等教育学报</H2> </s>
<s id="0004"> <HR> <B>Volume 6, No 1&2, pp. 19-27 (May, 1996) </B> <HR> </s>	<s id="0004"> <HR>（一九九六年五月） 第六卷. </s> <s id="0005"> 第一及二期19-27页 <HR> </s>
<s id="0005"> <H3>Principles for Redesigning Teacher Education </H3> Alan TOM </CENTER> </s>	<s id="0006"> <H3>革新教师教育的原则 </H3> Alan TOM </CENTER> </s>
<s id="0006"> <P> <B> <I> Abstract </I> </B> </s>	<s id="0007"> <P> <I> <B> 摘要 </B> </I> <P> </s>

Figure 3: An alignment example considering cognates.

divided into sentences. The sentences are marked by <s id="xxxx"> and </s>. Note that we determine sentences not only by periods, but also by means of HTML markup.

We further notice that it is difficult to align sentences 0002. The sentence in the Chinese page is much longer than its counterpart in the English page because some additional information (font) is added. The length-based method thus tends to take sentence 0002, 0003, and 0004 in the English page as the translation of sentence 0002 in the Chinese page (Fig. 2), which is wrong. This in turn provocated the three following incorrect alignments. As we can see in Fig. 3, the cognate method did not make the same mistake because of the noise in sentence 0002. Despite their large length difference, the two 0002 sentences are still aligned as a 1-1 pair, because the sentences in the following 4 alignments (0003 - 0003; 0004 - 0004, 0005; 0005 - 0006; 0006 - 0007) have rather similar HTML markups and are taken by the program to be the most likely alignments.

Beside HTML markups, other criteria may also be incorporated. For example, it would be helpful to consider strong correspondence between certain English and Chinese words, as in (Wu, 1994). We hope to implement such correspondences in our future research.

3.4 Lexicon Evaluation

To evaluate the precision of the English-Chinese translation model trained on the Web corpus, we examined two sample lexicons of 200 words, one in each direction. The 200 words for each lexicon were randomly selected from the training source. We examined the most probable translation for each word. The Chinese-English lexicon was found to have a precision of 77%. The English-Chinese lexicon has a higher precision of 81.5%. Part of the lexicons are shown in Fig. 4, where t/f indicates whether a translation is true or false.

These precisions seem to be reasonably high. They are quite comparable to that obtained by Wu (1994) using a manual Chinese-English parallel corpus.

3.5 Effect of Stopwords

We also found that stop-lists have significant effect on the translation model. Stop-list is a set of the most frequent words that we remove from the train-

English word	t/f	Translation	Probability	Chinese word	t/f	Translation	Probability
a.m.	t	上午	0.201472	办事处	t	office	0.375868
access	f	公开	0.071705	保护	t	protection	0.343071
adaptation	t	适应	0.179633	报告	t	report	0.358592
add	t	补充	0.317435	备	t	prepare	0.189513
adopt	t	采用	0.231637	本地	t	local	0.421837
agent	t	代理人	0.224902	便会	f	follow	0.023685
agree	t	同意	0.36569	标准	t	standard	0.445453
airline	t	航空公司	0.344001	补校	f	adult	0.044959
amendment	t	修订	0.367518	不足	t	inadequate	0.093012
appliance	t	用具	0.136319	部分	t	part	0.313676
apply	t	适用	0.19448	财经	t	financial	0.16608
attendance	t	列席	0.171769	参观	t	visit	0.309642
auditor	f	审核	0.15011	草案	t	bill	0.401997
average	t	平均	0.467646	车辆	t	vehicle	0.467034
base_on	f	计算	0.107304	储蓄	t	saving	0.176695

Figure 4: Part of the evaluation lexicons.

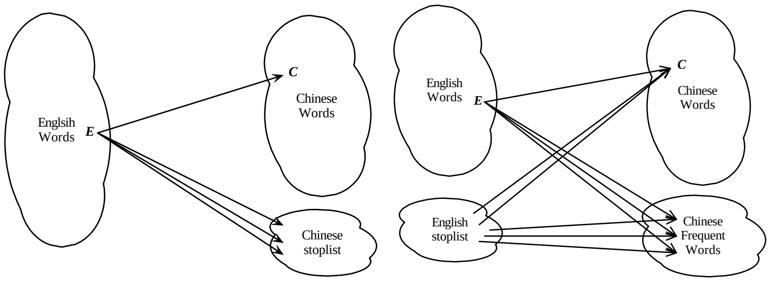

Figure 5: Effect of stop lists in C-E translation.

ing source. Because these words exist in most alignments, the statistical model cannot derive correct translations for them. More importantly, their existence greatly affects the accuracy of other translations. They can be taken as translations for many words.

A priori, it would seem that both the English and Chinese stop-lists should be applied to eliminate the noise caused by them. Interestingly, from our observation and analysis we concluded that for better precision, only the stop-list of the target language should be applied in the model training.

We first explain why the stop-list of the target language has to be applied. On the left side of Fig. 5, if the Chinese word C exists in the same alignments with the English word E more than any other Chinese words, C will be the most probable translation for E. Because of their frequent appearance, some Chinese stopwords may have more chances to be in the same alignments with E. The probability of the translation $E \rightarrow C$ is then reduced (maybe even less than those of the incorrect ones). This is the reason why many English words are translated to "**的**" (of) by the translation model trained without using the Chinese stop-list.

We also found that it is not necessary to remove the stopwords of the source language. In fact, as illustrated on the right side of Fig. 5, the existence of the English stopwords has two effects on the probability of the translation $E \rightarrow C$:

1 They may often be found together with the Chinese word C. Owing to the Expectation Maximization algorithm, the probability of $E \rightarrow C$ may therefore be reduced.

2 On the other hand, there is a greater likelihood that English stopwords will be found together with the most frequent Chinese words. Here, we use the term "Chinese frequent words" instead of "Chinese stopwords" because even if a stop-list is applied, there may still remain some common words that have the same effect as the stopwords. The coexistence of English and Chinese frequent words reduces the probability that the Chinese frequent words are the translations of E, and thus raise the probability of $E \rightarrow C$.

The second effect was found to be more significant than the first, since the model trained without the English stopwords has better precision than the model trained with the English stopwords. For the correct translations given by both models, the model

	C-E CLIR	E-C CLIR
Mono-Lingual IR	0.3861	0.3976
Translation Model	0.1504 (39.0%mono)	0.1841 (46.3%mono)
Dictionary	0.1530 (39.6%mono)	0.1427 (35.9%mono)
TM + DICT	0.2583 (66.9%mono)	0.2232 (56.1%mono)

Table 1: CLIR results.

trained without considering the English stopwords gives higher probabilities.

4 English-Chinese CLIR Results

Our final goal was to test the performance of the translation models trained on the Web parallel corpora in CLIR. We conducted CLIR experiments using the *Smart* IR system.

4.1 Results

The English test corpus (for C-E CLIR) was the AP corpus used in TREC6 and TREC7. The short English queries were translated manually into Chinese and then translated back to English by the translation model. The Chinese test corpus was the one used in the TREC5 and TREC6 Chinese track. It contains both Chinese queries and their English translations.

Our experiments on these two corpora produced the results shown in Tab. 1. The precision of monolingual IR is given as benchmark. In both E-C and C-E CLIR, the translation model achieved around 40% of monolingual precision. To compare with the dictionary-based approach, we employed a Chinese-English dictionary, CEDICT (Denisowski, 1999), and an English-Chinese online dictionary (Anonymous, 1999a) to translate queries. For each word of the source query, all the possible translations given by the dictionary are included in the translated query. The Chinese-English dictionary has about the same performace as the translation model, while the English-Chinese dictionary has lower precision than that of the translation model.

We also tried to combine the translations given by the translation model and the dictionary. In both C-E and E-C CLIR, significant improvements were achieved (as shown in Tab. 1). The improvements show that the translations given by the translation model and the dictionary complement each other well for IR purposes. The translation model may give either exact translations or incorrect but related words. Even though these words are not correct in the sense of translation, they are very possibly related to the subject of the query and thus helpful for IR purposes. The dictionary-based approach expands a query along another dimension. It gives all the possible translations for each word including those that are missed by the translation model.

4.2 Comparison With MT Systems

One advantage of a parallel text-based translation model is that it is easier to build than an MT system. Now that we have examined the CLIR performance of the translation model, we will compare it with two existing MT systems. Both systems were tested in E-C CLIR.

4.2.1 Sunshine WebTran Server

Using the Sunshine WebTran server (Anonymous, 1999b), an online English-Chinese MT system, to translate the 54 English queries, we obtained an average precision of 0.2001, which is 50.3% of the mono-lingual precision. The precision is higher than that obtained using the translation model (0.1804) or the dictionary (0.1427) alone, but lower than the precison obtained using them together (0.2232).

4.2.2 Transperfect

Kwok (1999) investigated the CLIR performance of an English-Chinese MT software called Transperfect, using the same TREC Chinese collection as we used in this study. Using the MT software alone, Kwok achieved 56% of monolingual precision. The precision is improved to 62% by refining the translation with a dictionary. Kwok also adopted pre-translation query expansion, which further improved the precison to 70% of the monolingual results.

In our case, the best E-C CLIR precison using the translation model (and dictionary) is 56.1%. It is lower than what Kwok achieved using Transperfect, however, the difference is not large.

4.3 Further Problems

The Chinese-English translation model has a far lower CLIR performance than that of the English-French model established using the same method (Nie et al., 1999). The principal reason for this is the fact that English and Chinese are much more different than English and French. This problem surfaced in many phases of this work, from text alignment to query translation. Below, we list some further factors affecting CLIR precision.

- The Web-collected corpus is noisy and it is difficult to align English-Chinese texts. The alignment method we employed has performed more poorly than on English-French alignment. This in turn leads to poorer performance of the translation model. In general, we observe a higher variability in Chinese-English translations than in English-French translations.

- For E-C CLIR, although queries in both languages were provided, the English queries were not strictly translated from the original Chinese ones. For example, 人权状况 (*human right situation*) was translated into *human right issue*. We cannot expect the translation model to translate *issue* back to 状况 (*situation*).
- The training source and the CLIR collections were from different domains. The Web corpus are retrieved from the parallel sites in Hong Kong while the Chinese collection is from *People's Daily* and *Xinhua News Agency*, which are published in mainland China. As the result, some important terms such as 最惠国 (*most-favored-nation*) and 一国两制 (*one-nation-two-systems*) in the collection are not known by the model.

5 Summary

The goal of this work was to investigate the feasibility of using a statistical translation model trained on a Web-collected corpus to do English-Chinese CLIR. In this paper, we have described the algorithm and implementation we used for parallel text mining, translation model training, and some results we obtained in CLIR experiments. Although further work remains to be done, we can conclude that it is possible to automatically construct a Chinese-English parallel corpus from the Web. The current system can be easily adapted to other language pairs. Despite the noisy nature of the corpus and the great difference in the languages, the evaluation lexicons generated by the translation model produced acceptable precision. While the current CLIR results are not as encouraging as those of English-French CLIR, they could be improved in various ways, such as improving the alignment method by adapting cognate definitions to HTML markup, incorporating a lexicon and/or removing some common function words in translated queries.

We hope to be able to demonstrate in the near future that a fine-tuned English-Chinese translation model can provide query translations for CLIR with the same quality produced by MT systems.

References

Anonymous. 1999a. Sunrain.net - English-Chinese dictionary. http://sunrain.net/r_ecdict_e.htm.

Anonymous. 1999b. Sunshine WebTran server. http://www.readworld.com/translate.htm.

P. F. Brown, J. C. Lai, and R. L. Mercer. 1991. Aligning sentences in parallel corpora. In *29th Annual Meeting of the Association for Computational Linguistics*, pages 89–94, Berkeley, Calif.

P. F. Brown, S. A. Della Pietra, V. J. Della Pietra, and R. L. Mercer. 1993. The mathematics of machine translation: Parameter estimation. *Computational Linguistics*, 19:263–311.

S. F. Chen. 1993. Aligning sentences in bilingual corpora using lexical information. In *Proceedings of the 31th Annual Meeting of the Association for Computational Linguistics*, pages 9–16, Columbus, Ohio.

Paul Denisowski. 1999. Cedict (chinese-english dictionary) project. http://www.mindspring.com/paul_denisowski/cedict.html.

William A. Gale and Kenneth W. Church. 1991. A program for aligning sentences in bilingual corpora. In *Proceedings of the 29th Annual Meeting of the Association for Computational Linguistics*, pages 177–184, Berkeley, Calif.

P. Isabelle, G. Foster, and P. Plamondon. 1997. SILC: un système d'identification de la langue et du codage. http://www-rali.iro.umontreal.ca/ProjetSILC.en.html.

M. Kay and M. Röscheisen. 1993. Text-translation alignment. *Computational Linguistics*, 19:121–142.

K. L. Kwok. 1999. English-chinese cross-language retrieval based on a translation package. In *Workshop of Machine Translation for Cross Language Information Retrieval, Machine Translation Summit VII*, Singapore.

P. Langlais, G. Foster, and G. Lapalme. 2000. Unit completion for a computer-aided translation typing system. In *Applied Natural Language Processing Conference (ANLP)*, Seattle, Washington, May.

Jianyun Nie, Michel Simard, Pierre Isabelle, and Richard Durand. 1999. Cross-language information retrieval based on parallel texts and automatic mining parallel texts from the Web. In *ACM SIGIR'99*, pages 74–81, August.

Philip Resnik. 1998. Parallel stands: A preliminary investigation into mining the Web for bilingual text. In *AMTA'98*, October.

Michel Simard, George F. Foster, and Pierre Isabelle. 1992. Using cognates to align sentences in bilingual corpora. In *Proceedings of TMI-92*, Montreal, Quebec.

Dekai Wu. 1994. Aligning a parallel English-Chinese corpus statistically with lexical criteria. In *ACL-94: 32nd Annual Meeting of the Assoc. for Computational Linguistics*, pages 80–87, Las Cruces, NM, June.

Dekai Wu. 1995. Large-scale automatic extraction of an English-Chinese lexicon. *Machine Translation*, 9(3-4):285–313.