RMIT Chinese-English CLIR at NTCIR-4

Ying Zhang, Phil Vines
School of Computer Science and Information Technology, RMIT University
GPO Box 2476V, Melbourne, Australia, 3001.
{yzhang,phil}@cs.rmit.edu.au

February 27, 2004

Abstract

We participated in the Chinese-English CLIR task. We concentrated primarily on the issues of translation disambiguation and automatic translation extraction of Out of Vocabulary (OOV) terms. A new segmentation-free technique has been developed to extract translations of Chinese OOV terms from the Web. Our technique is able to extract the correct translation from the Web in most cases and thus improve translation quality and segmentation accuracy.

Keywords: translation disambiguation, translation extraction.

1 Introduction

The performance of dictionary-based query translation approaches is limited by the accuracy of three factors: phrase translation, translation ambiguity, and dictionary coverage, the last being the most serious problem. The phrase translation problem stems from word by word translation of a phrase which should have been identified as a single entity and translated as such, for example, numbers, personnel names, and calendar times. Second, translation ambiguity arises from the fact that many words have multiple possible translations. Third, the dictionary may lack some terms that are essential for the correct interpretation of the query.

Our current work focuses only on the problems of translation ambiguity and dictionary coverage. Researchers have proposed a number of techniques to resolve the translation ambiguity problem using term co-occurrence [2], mutual information [4] or term-similarity [8]. To deal with out of vocabulary (OOV) terms, researchers have used parallel corpora [5], anchor text [3], or transliteration [6] to extract translations. The major differences in our proposal are: first, we propose an improved disambiguation technique to select the most appropriate English translation for each Chinese query term; second, we develop a method to automatically detect Chinese OOV terms and extract the most appropriate English translations through mining the web text; thus improving the effectiveness of dictionary-based CLIR.

The structure of the paper is as follows. In Section 2, we describe our algorithms for translation disambiguation, English translation extraction and post-translation query expansion. In Section 3, we detail our experiments and the results we obtained; and Section 4 concludes the paper.

2 Methodology

In the section, we describe the three techniques we have used: the translation disambiguation technique, the translation extraction technique, and the query expansion technique.

2.1 Translation Disambiguation

Each set of English translations E is a sequence of words $(e_1, e_2, e_3, ...e_n)$. We use a probability model $P(E) = P(e_1, e_2, e_3, ...e_n)$ to estimate the maximum likelihood (ML) of each sequence of words. We select English translations E with the highest $P(E)$ among all possible translation sets.

Our disambiguation technique is based on *Hidden Markov Model* [7] that has been used widely for probabilistically modelling sequence data.

$$P(e_1, e_2...e_n) = P(e_1)\prod_{a=2}^{n} P(e_a|e_{a-1})$$

In order to compute the probability value , we need to calculate the quantities $P(e)$ and $P(e|e')$.

$$P(e) = \frac{f(e)}{N}, P(e|e') = \frac{P_w(e, e')}{\sum_{e''} P_w(e'', e')}$$

where $f(e)$ is the collection frequency of term e, N is the number of terms occurring in the document collection, and $P_w(e, e')$ is the probability of term e' occurring after term e within a window size w.

The zero-frequency problem arises quite often in the context of probabilistic language models when the model encounters an event in a context where it has never been seen before. Smoothing provides a way to estimate and assign the probability to that unseen event. In this work we use the following absolute discounting and interpolation formula which applies the smoothing method proposed by [1], since the results reported were quite promising.

$$P(e|e') = max\{\frac{f_w(e,e') - \beta}{N}, 0\} + \beta P(e)P(e') \quad (1)$$

where $f_w(e,e')$ is the frequency of term e' occurring after term e within a window size w. Federico and Bertoldi [1] successfully used this formula to compute the frequency of term e' and e within a text window of fixed size through an order-free bigram language model in their work. However, they did not give detailed information about the size of the text window. From our previous experiments [11], we determined that the best results are generally obtained with window size $w = 6$. The absolute discounting term β is equal to the estimate proposed in [9]:

$$\beta = \frac{n_1}{n_1 + 2n_2}$$

where n_k representing the number of terms with the collection frequency k.

We have observed that shorter distance between two words generally provides stronger correlation and produces more credible results for disambiguation of translation. Gao and Zhou [2] applied a decaying factor to the mutual information calculation, their experiments showed that the decaying factor can be used to discriminate strong and weak term correlation.

$$D(e,e') = \mathbf{e}^{-\alpha \times (Dist(e,e')-1)}$$

where $Dist(e,e')$ is the average distance between e and e' in the document collection and α is determined empirically to 0.8. Therefore, we have added this distance factor $D(e,e')$ into the probability calculation (1) to become:

$$[max\{\frac{f_w(e,e') - \beta}{N}, 0\} + \beta P(e)P(e')] \times D(e,e')$$

The results of using this formula are discussed in Section 3.

2.2 Translation Extraction

Our idea stems from the observation that when new terms, foreign terms, or proper nouns are used in Chinese text, they are sometimes accompanied by the English translation, normally immediately after the Chinese text, e.g. 世紀之毒戴奧辛 (Dioxin), where 世紀之毒戴奧辛 is a sequence of Chinese characters. By mining the web to collect a sufficient number of such instances for any given word and applying statistical techniques, we are then able to infer the appropriate translation with reasonable confidence.

Our procedure consists of three steps: extraction of the Web text, collection of co-occurrence statistics and translation selection.

2.2.1 Extraction of Web Text

First, we extract strings that contain the Chinese query terms and some English text from the Web.

1. When a Chinese query term is missing from the dictionary, we run a script file that uses Google to fetch the top 100 Chinese documents using the Chinese query key terms and save them into a local file using the following command:

```
lynx -source "http://www.google.com.au/
search?q=Chinese-Query&num=100&
lr=lang_zh-TW&cr=countryTW&hl=zh-TW&
ie=UTF-8&oe=UTF-8" > local_file
```

   It should be noted that the side effect of using a search engine is that only higher quality web text is returned. This reduces the likelihood of noisy translations being collected.

2. For each returned document, only the title and the query-biased summary are extracted and saved into a local file.

3. The file is then filtered to remove HTML tags and metadata, leaving only the web text.

For example, suppose a query Q is composed of a sequence of Chinese terms $(c_1, c_2, c_3, c_4, c_5)$, and we have retrieved a series of titles and query-biased summaries of web text that contain both Chinese query substrings $C_{ij} \in Q$ and English terms e, as shown in Figure 1.

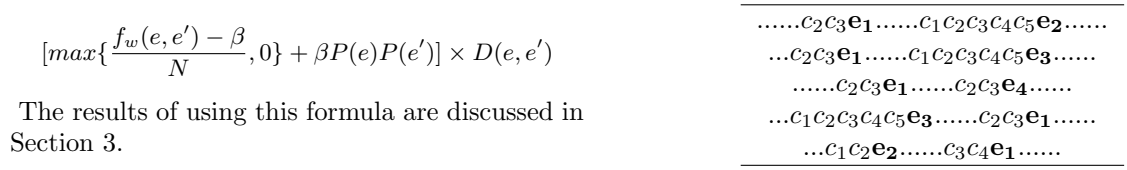

Figure 1: *Web text retrieved*

2.2.2 Collection of Co-occurrence Statistics

We then collect co-occurrence information from the data we obtained, in the following mannner:

1. Scan for the occurrence of English text. Where English text occurs, check the immediately proceeding Chinese text to see if it is a substring of the original Chinese query.

2. We collect the frequency of co-occurrence of the English text and all Chinese query substrings that appear immediately before the English text.

For each English term e_i with frequency $f(e_i)$, we obtained a group of associated Chinese query substrings C_{ij} with the length $|C_{ij}|$ and the co-occurrence frequency $f(e_i, C_{ij})$. Extending the example from in Figure 1, this information is summarized in Table 1.

e_i	$f(e_i)$	C_{ij}	$\|C_{ij}\|$	$f(e_i, C_{ij})$
e_1	5	c_2c_3	2	4
		c_3c_4	2	1
e_2	2	$c_1c_2c_3c_4c_5$	5	1
		c_1c_2	2	1
e_3	2	$c_1c_2c_3c_4c_5$	5	2
e_4	1	c_2c_3	2	1

Table 1: *The frequency of co-occurrence of English terms and Chinese query substrings*

2.2.3 Translation Selection

We then select the most appropriate translation from Table 1 as follows:

1. Firstly search for *longest* Chinese substring C_t:

   (a) Search for the Chinese query substrings $C_{targets}$, where $|C_{targets}| = max(|C_{ij}|)$.

   (b) Extract the English term e_t and the Chinese query substring C_t, where $f(e_t, C_t) = max(f(e_i, C_{targets}))$.

   (c) Add (C_t, e_t) into the translation dictionary.

2. Then search for the English term $e_{t'}$with the *highest frequency*:

   (a) Search for the English terms $e_{targets}$, where $f(e_{targets}) = max(f(e_i))$.

   (b) Extract the English term $e_{t'}$ and the Chinese query substring $C_{t'}$, where $f(e_{t'}, C_{t'}) = max(f(e_{targets}, C_{ij}))$.

   (c) if $C_{t'} \neq C_t$ and $e_{t'} \neq e_t$, add $(C_{t'}, e_{t'})$ into the translation dictionary.

In the example in Table 1 above, two translation pairs (c_2c_3, e_1) and ($c_1c_2c_3c_4c_5$, e_3) are extracted and added into the translation dictionary. We have extracted at most two translation pairs, which proved to be ample for short queries; and in fact, in most cases, only one translation pair was extracted.

2.3 Post-translation Query Expansion

We applied an automatic feedback query expansion approach that exploits the statistical relationship based on word co-occurrences, namely adding n terms from the top 10 retrieved documents to the original query, on the presumption that those documents are relevant. Our system considered the two elements of automatic feedback as listed below.

1. TF weighting
   We ranked terms from the top 10 retrieved documents based on TF weighting and selected the top 20 terms. TF is the frequency with which term t occurs in top 10 retrieved documents.

2. Mutual Information (MI)
   From these top 20 terms, we selected n terms that have the strongest correlation with original query and added them to make a new query in following manner:

   (a) To measure the MI between a given term t and a original query key term q within a window size $w = 6$, we used:

   $$MI(t, q) = \log_2(\frac{f_w(t, q)}{f_t f_q} + 1)$$

   where
   $f_w(t, q)$ is the frequency that t and q co-occur within a window size of 6 in the document collection;
   f_t is the collection frequency of t;
   and f_q is the collection frequency of q.
   The formula adds 1 to the frequency ratio, so that a zero co-occurrence frequency corresponds to zero MI.

   (b) To measure the MI between a given term t and an original query Q, composed of a sequence of query key terms $\{q_1, q_2, ..., q_n\}$, we uesd:

   $$MI(t, Q) = \sum_{m=1}^{n} \log_2(\frac{f_w(t, q_m)}{f_t f_{q_m}} + 1)$$

(c) In the *title* run the top 10 terms with the highest $MI(t, Q)$ were selected and added to make the new query. In the *desc* run the top 5 terms with the highest $MI(t, Q)$ were selected and added to make the new query.

3 Experiments and results

In this section, we describe the experimental setup, including the document collection, the translation dictionaries we used, pre-processing of the English document collection and Chinese queries, and the experimental design.

3.1 Document Collection

The English document collection from the NTCIR Workshop 4 CLIR task contains 347,376 news articles from 1998 to 1999. There are 58 Chinese topics, each topic contains four parts: *title*, *description*, *narrative* and *key words* relevant to whole topic.

3.2 Chinese-English Dictionaries

We used two dictionaries in our experiments: ce3 from Linguistic Data Consortium[1], and CEDICT Chinese-English dictionary[2] to translate Chinese queries into English.

3.3 Segmentation Tools

Unlike English and other European languages, Chinese text dose not have a natural delimiter between words. As a consequence, word segmentation is a major issue in Chinese-English translation processing. The inherent errors caused by word segmentation always remains as a problem in Chinese-English information retrieval. In our experiments, high precision segmentations is not the focus of our work. Instead we aim to evaluate the effectiveness of our disambiguation method as long as the errors caused by word segmentation are reasonably low. However, we note that the web mining technique that we have developed could also be used to improve segmentation accuracy.

3.4 Pre-processing

English stop words were removed from the English document collection. We used a stop list that contains 477 entries and the Porter stemmer [10] to reduce words to stems.

[1] http://www.ldc.upenn.edu/

[2] http://www.mandarintools.com/cedict.html

The Chinese queries were processed in the way as follows:

1. Each Chinese query was presented a list of comma separated query key terms. Our assumption is that each query key term is either a phrase or a word. We found 66 out of 163 query key terms cannot be found in the translation dictionaries. We treat all of these as potential Chinese OOV terms, although some of them can be correctly translated word by word.
2. Using these 66 query key terms as queries, we applied our translation extraction technique and added extracted translation pairs into the translation dictionary.
3. We compiled a segmentation dictionary using the two translation dictionaries. We used dictionary-based segmentation with greedy-parsing to segment the Chinese queries.
4. In this stage we used the translation dictionary to replace each query key term by all English translations.
5. Our translation disambiguation technique was used to select the most appropriate translation for each Chinese query.

3.5 Experimental Design

We have submitted four Chinese-English CLIR runs. Two runs used the *titles* of the Chinese topics as queries to retrieve the documents from the English document collection. The other two runs used the texts of *description* fields as queries. A brief description of each run is shown in Table 3.

RunID	*Translation Disambiguation*	*Translation Extraction*	*Query Expansion*
T-01	√	√	–
T-03	√	√	√
D-02	√	√	–
D-04	√	√	√

Table 3: *NTCIR4: Run Description*

Table 2 shows the translations we have extracted from the Web. Among 66 potential Chinese OOV terms, 41 instances can be translated word by word using the translation dictionary. Of 25 Chinese OOV terms, we were able to successfully translate 17. The remaining 8 cases failed for one of two reasons: first, our search technique did not return any English terms associated with some Chinese OOV terms; second, some personnel names that relate to events are no longer topical and could not be found on the Web, such as "小淵惠三" and "花蝴蝶".

Query ID	*Chinese query*	*Chinese OOV Terms*	*Extracted English Translations*	*Given English Translations*
001	秋門	秋門	—	Chiutou
002	約翰走路	約翰走路	Johnnie Walker	Johnnie Walker
003	胚胎乾細胞	胚胎乾細胞	Embryonic Stem Cell	Embryonic Stem Cells
004	葛瑞菲絲 喬納 花蝴蝶	葛瑞菲絲	Griffith — —	Griffith Joyner Flojo
005	戴奧辛	戴奧辛	Dioxin	Dioxin
006	麥可喬丹	麥可喬丹	Michael Jordan	Michael Jordan
007	巴拿馬運河 卡杜條約	巴拿馬運河	Panama Canal —	Panama Canal Torrijos-Carter Treaty
008	威而鋼	威而鋼	Viagra	Viagra
012	黑澤明	黑澤明	Akira Kurosawa	Akira Kurosawa
013	小淵惠三		—	Keizo Obuchi
014	環境荷爾蒙	環境荷爾蒙	environmental hormone	Environmental Hormone
021	電子商務交易	電子商務	Electronic Commerce	Electronic Commercial Transaction
022	起亞汽車	起亞汽車	Kia Motors Corp	Kia Motors
030	動物複製技術	複製	clone	Cloning
034	東京都知事		—	Tokyo provincial governor
038	奈米科技	奈米科技	Nanotechnology	Nanotechnology
046	基因治療	基因治療	Genetic Treatment	Genetic Treatment
048	國際太空站	國際太空站	ISS	International Space Station
051	隱形戰鬥機	隱形戰鬥機 戰鬥機	stealth fighter F117	Stealth Fighter —-
052	皇太子妃 雅子		— —	Crown Princess Masako
058	非接觸式智慧卡	非接觸式智慧卡	Contactless Smart Cards CSC	Contactless SMART Card

Table 2: *NTCIR4: Extracted English translations of Chinese OOV terms*

4 Conclusions

In this work, we have looked at in detail at two factors that degrade Chinese-English CLIR: the translation ambiguity and the dictionary coverage. We have applied an improved disambiguation technique to improve dictionary-based query translation, and developed a new technique to extract English translations of Chinese OOV terms through mining the Web. We have also showed that it can be used to improve Chinese segmentation accuracy.

References

[1] M. Federico and N. Bertoldi. Statistical cross-language information retrieval using n-best query translations. In *Proceedings of the 25th annual international ACM SIGIR conference on Research and development in information retrieval*, pages 167–174, Tampere, Finland, 2002.

[2] J. Gao, M. Zhou, J. Nie, H. He and W. Chen. Resolving query translation ambiguity using a decaying co-occurrence model and syntactic dependence relations. In *Proceedings of the 25th annual international ACM SIGIR conference on Research and development in information retrieval*, pages 183–190, Tampere, Finland, 2002.

[3] W. Lu, C. Tung, L. Chien and H. Lee. Translation of web queries using anchor text mining. *ACM Transactions on Asian Language Information Processing*, Volume 2, Number 1, pages 159 – 172, 2002.

[4] A. Maeda, F. Sadat, M. Yoshikawa and S. Uemura. Query term disambiguation for web cross-language information retrieval using a search engine. In *Proceedings of the 5th International Workshop on Information Retrieval with Asian Languages*, pages 25–32, 2000.

[5] C. J. A. McEwan, I. Ounis and I. Ruthven. Building bilingual dictionaries from parallel web documents. In *Proceedings of the Twenty-Fourth European Colloquium on Information Retrieval Research*, pages 303–323, Glasgow, Scotland, UK, 2002.

[6] H. Meng, B. Chen, S. Khudanpur, G. Levow, W. Lo, D. Oard, P. Schone, K. Tang, H. Wang and J. Wang. Mandarin-English Information (MEI): investigating translingual speech retrieval. In *Computer Speech and Language*, 2003.

[7] D. R. Miller, T. Leek and R. M. Schwartz. A hidden Markov model information retrieval system. In *Proceedings of SIGIR-99, 22nd ACM International Conference on Research and Development in Information Retrieval*, pages 214–221, Berkeley, US, 1999.

[8] A. Mirna. Using Statistical Term Similarity for Sense Disambiguation in Cross-language Information Retrieval. *Information Retrieval*, Volume 2, Number 1, pages 67–68, 2000.

[9] H. Ney, U. Essen and R. Kneser. On structuring probabilistic dependences in stochastic language

modelling. *Computer Speech and Language*, Volume 8, Number 3, pages 1–38, 1994.

[10] M. F. Porter. An algorithm for su#x stripping. *Automated Library and Information Systems*, Volume 14, Number 3, pages 130–137, 1980.

[11] Y. Zhang and P. Vines. Improved use of contextual information in cross-language information retrieval. In *Proceedings of the 7th Australasian Document Computing Symposium, ADCS2002*, pages 129–132, Sydney, Australia, 2002.