A Transitive Model for Extracting Translation Equivalents of Web Queries through Anchor Text Mining

Wen-Hsiang Lu
Institute of Information Science
Academia Sinica;
Dept. of Computer Science and
Information Engineering
National Chiao Tung University
Hsinchu 300, Taiwan, ROC
whlu@iis.sinica.edu.tw

Lee-Feng Chien
Institute of Information
Science, Academia Sinica
Nangang 115, Taiwan, ROC
lfchien@iis.sinica.edu.tw

Hsi-Jian Lee
Dept. of Computer Science and
Information Engineering
National Chiao Tung University
Hsinchu 300, Taiwan, ROC
hjlee@csie.nctu.edu.tw

Abstract

One of the existing difficulties of cross-language information retrieval (CLIR) and Web search is the lack of appropriate translations of new terminology and proper names. Different from conventional approaches, in our previous research we developed an approach for exploiting Web anchor texts as live bilingual corpora and reducing the existing difficulties of query term translation. Although Web anchor texts, undoubtedly, are very valuable multilingual and wide-scoped hypertext resources, not every particular pair of languages contains sufficient anchor texts in the Web to extract corresponding translations in the language pair. For more generalized applications, in this paper we extend our previous approach by adding a phase of transitive (indirect) translation via an intermediate (third) language, and propose a transitive model to further exploit anchor-text mining in term translation extraction applications. Preliminary experimental results show that many query translations which cannot be obtained using the previous approach can be extracted with the improved approach.

1. Introduction

Cross-language information retrieval (CLIR), addressing the special need where users can query in one language and retrieve relevant documents written or indexed in another language, has become an important issue in the research of information retrieval (Dumais et al., 1996; Davis et al., 1997; Ballesteros & Croft, 1998; Nie et al., 1999). However, its application to practical Web search services has not lived up to expectations, since they suffer a major bottleneck that lacks up-to-date bilingual lexicons containing the translation of popular query terms [1] such as proper nouns (Kwok, 2001).

To enable capability of CLIR, existing IR systems mostly rely on bilingual dictionaries for cross-lingual retrieval. In these systems, queries submitted in a source language normally have to be translated into a target language by means of simple dictionary lookup. These dictionary-based techniques are limited in real-world applications, since the queries given by users often contain proper nouns.

Another kind of popular approaches to dealing with query translation based on corpus-based techniques uses a parallel corpus containing aligned sentences whose translation pairs are corresponding to each other (Brown et al., 1993; Dagan et al., 1993; Smadja et al., 1996). Although more reliable translation equivalents can be extracted by these techniques, the unavailability of large enough parallel corpora for various subject domains and multiple languages is still in a thorny situation. On the other hand, the alternative approach using comparable or unrelated text corpora were studied by Rapp (1999) and Fung et al. (1998). This task is more difficult due to lack of parallel correlation between document or sentence pairs.

[1] In our collected query logs, most of user queries contain only one or two words, so we use query term, query or term interchangeably in this paper.

In our previous research we have developed an approach to extracting translations of Web queries through mining of Web anchor texts and link structures (Lu, et al., 2001). This approach exploits Web anchor texts as live bilingual corpora to reduce the existing difficulties of query translation. Anchor text sets, which are composed of a number of anchor texts linking to the same pages, may contain similar description texts in multiple languages, thus it is more likely that user's queries and their corresponding translations frequently appear together in the same anchor text sets. The anchor-text mining approach has been found effective particularly for proper names, such as international company names, names of foreign movie stars, worldwide events, e.g., "Yahoo", "Anthrax", "Harry Potter", etc.

Discovering useful knowledge from the potential resource of Web anchor texts is still not fully explored. According to our previous experiments, the extracted translation equivalents might not be reliable enough when a query term whose corresponding translations either appear infrequently in the same anchor text sets or even do not appear together. Especially, the translation process will be unavailable if there is a lack of sufficient anchor texts for a particular language pair. Although Web anchor texts, undoubtedly, are live multilingual resources, not every particular pair of languages contains sufficient anchor texts.

To deal with the problems, this paper extends the previous anchor-text-based approach by adding a phase of indirect translation via an intermediate language. For a query term which is unable to be translated, our idea is to translate it into a set of translation candidates in an intermediate language, and then seek for the most likely translation from the candidates, which are translated from the intermediate language into the target language (Gollins et al., 2001; Borin, 2000). We therefore propose a *transitive translation model* to further exploit anchor text mining for translating Web queries. A series of experiments has been conducted to realize the performance of the proposed approach. Preliminary experimental results show that many query translations which cannot be obtained using the previous approach can be extracted with the improved approach.

2 The Previous Approach

For query translation, the anchor-text-based approach is a new technique compared with the bilingual-dictionary- and parallel-corpus-based approaches. In this section we will introduce the basic concept of the anchor-text-based approach. For more details please refer to our initial work (Lu, et al., 2001).

2.1 Anchor-Text Set

An anchor text is the descriptive part of an out-link of a Web page. It represents a brief description of the linked Web page. For a Web page (or URL) u_i, its anchor-text set is defined as all of the anchor texts of the links, i.e., u_i's in-links, pointing to u_i. In general, the anchor-text set records u_i's alternative concepts and textual expressions such as titles and headings, which are cited by other Web pages. With different preferences, conventions and language competence, the anchor-text set could be composed of multilingual phrases, short texts, acronyms, or even u_i's URL. For a query term appearing in the anchor-text set, it is likely that its corresponding translations also appear together. The anchor-text sets can be considered as a comparable corpus of translated texts, from the viewpoint of translation extraction.

2.2 The Probabilistic Inference Model

To determine the most probable target translation t for source query term s, we developed a probabilistic inference model (Wong et al., 1995). This model is adopted for estimating probability value between source query and each translation candidate that co-occur in the same anchor-text sets. The estimation assumes that the anchor texts linking to the same pages may contain similar terms with analogous concepts. Therefore, a candidate translation has a higher chance to be an effective translation if it is written in the target language and frequently co-occurs with the source query term in the same anchor-text sets. In the field of Web research, it has been proven that the use of link structures is effective for estimating the

authority of Web pages (Kleinberg, 1998; Chakrabarti et al., 1998). The model further assumes that the translation candidates in the anchor-text sets of the pages with higher authority may have more reliability in confidence. The similarity estimation function based on the probabilistic inference model is defined below:

$$P(s \leftrightarrow t) = \frac{P(s \cap t)}{P(s \cup t)} = \frac{\sum_{i=1}^{n} P(s \cap t \cap u_i)}{\sum_{i=1}^{n} P((s \cup t) \cap u_i)}$$

$$= \frac{\sum_{i=1}^{n} P(s \cap t \mid u_i) P(u_i)}{\sum_{i=1}^{n} P(s \cup t \mid u_i) P(u_i)}. \qquad (1)$$

The above measure is adopted to estimate the degree of similarity between source term *s* and target translation *t*. The measure is estimated based on their co-occurrence in the anchor text sets of the concerned Web pages U = {u_1, u_2, ... u_n}, in which u_i is a page of concern and $P(u_i)$ is the probability value used to measure the authority of page u_i. By considering the link structures and concept space of Web pages, $P(u_i)$ is estimated with the probability of u_i being linked, and its estimation is defined as follows: $P(u_i) = L(u_i)/\Sigma_{j=1,n} L(u_j)$, where $L(u_j)$= the number of in-links of page u_j. Such estimation is simplified from HITS algorithm (Kleinberg, 1998).

In addition, we assume that *s* and *t* are independent given u_i, then the joint probability $P(s \cap t \mid u_i)$ is equal to the product of $P(s \mid u_i)$ and $P(t \mid u_i)$, and the similarity measure becomes:

$$P(s \leftrightarrow t) =$$

$$\approx \frac{\sum_{i=1}^{n} P(s \mid u_i) P(t \mid u_i) P(u_i)}{\sum_{i=1}^{n} [P(s \mid u_i) + P(t \mid u_i) - P(s \mid u_i) P(t \mid u_i)] P(u_i)}. \qquad (2)$$

The values of $P(s \mid u_i)$ and $P(t \mid u_i)$ are defined to be estimated by calculating the fractions of the numbers of u_i's in-links containing *s* and *t* over $L(u_i)$, respectively.

Therefore, a candidate translation has a higher confidence value to be an effective translation if it frequently co-occurs with the source term in the anchor-text sets of the pages with higher authority.

2.3 The Estimation Process

For each source term, the probabilistic inference model extracts the most probable translation that maximizes the estimation. The estimation process based on the model was developed to extract term translations through mining of real-world anchor-text sets. The process contains three major computational modules: anchor-text extraction, term extraction and term translation extraction. The anchor-text extraction module was constructed to collect pages from the Web and build up a corpus of anchor-text sets. On the other hand, for each given source term *s*, the term extraction module extracts key terms as the translation candidate set from the anchor-text sets of the pages containing *s*. At last, the term translation module extracts the translation that maximizes the similarity estimation. For more details about the estimation process, please refer to our previous work (Lu et al., 2001).

To make a difference from the translation process via an intermediate language, the above process is called direct translation, and the adopted model called direct translation model hereafter. Meanwhile, we will use function *Pdirect* in Equation (3) for the estimation of the direct translation.

$$Pdirect(s,t) = \log P(s \leftrightarrow t). \qquad (3)$$

3 The Improved Approach

3.1 The Indirect Translation Model

As mentioned above, for those query terms whose corresponding translations either appear infrequently in the same anchor text sets or do not appear together, the estimation with equation (2) is basically unreliable. To increase the possibility of translation extraction especially for the source terms whose corresponding translations do not co-occur, we add a phase of indirect translation through an intermediate language. For example, as shown in Fig. 1, our idea is to obtain the corresponding target

translation “索尼” in simplified Chinese by translating the source term “新力” in traditional Chinese into an intermediate term “Sony” in English, and then seek for translating “Sony” into a target term “索尼” in simplified Chinese. For both the source query and the target translation, we assume that their translations in the intermediate language are the same and can be found.

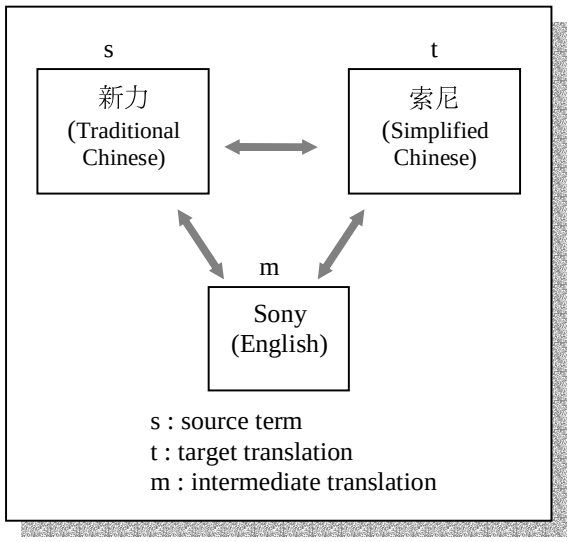

Fig. 1. An abstract diagram showing the concepts of direct translation and indirect translation.

The above assumption is not unrealistic. For example, it is possible to find the Chinese translation of a Japanese movie star through submitting his/her English name to a search engine and browsing the retrieved Chinese pages containing the English name. The Web contains large amounts of multilingual pages, and English is the most likely intermediate language between other languages. Based on this assumption, we extend the probabilistic inference model and propose an indirect translation model as the following formula:

$$\begin{aligned} Pindirect(s,t) &= \log P(s \leftrightarrow m, m \leftrightarrow t) \\ &\approx \log[P(s \leftrightarrow m) \times P(m \leftrightarrow t)] \\ &= \log P(s \leftrightarrow m) + \log P(m \leftrightarrow t). \end{aligned} \quad (4)$$

, where m is the transitive translation of s and t in the intermediate language, $P(s \leftrightarrow m)$ and $P(m \leftrightarrow t)$ are the probability values obtained with the direct translation model which can be calculated by Equation (2).

3.2 The Transitive Translation Model

The transitive model is developed to combine both the direct and indirect translation models and improve the translation accuracy. By combining Equation (3) and (4), the transitive translation model is defined as follows:

$$Ptrans(s,t) = \begin{cases} Pdirect(s,t), & \text{if } Pdirect(s,t) > \theta \\ Pindirect(s,t), & \text{otherwise.} \end{cases} \quad (5)$$

, where θ is a predefined threshold value.

4 Experimental Results

4.1 Analysis of Anchor-Text Sets and Query Logs

In the initial experiments, we took traditional Chinese and simplified Chinese as the source and target language respectively, and used English as the intermediate language. We have collected 1,980,816 traditional Chinese Web pages in Taiwan. Among these pages, 109,416 pages whose anchor-text sets contained both traditional Chinese and English terms were taken as the anchor-text set corpus. We also collected 2,179,171 simplified Chinese Web pages in China and extracted 157,786 pages whose anchor-text sets contained both simplified Chinese and English terms. In addition, through merging the two Web page collections into a larger one, we extracted 4,516 Web pages containing both traditional and simplified Chinese terms. The three comparable corpora provide a potential resource of translation pairs for some Web queries. In order to realize the feasibility in translating query terms via transitive translation, we aim at finding out the corresponding simplified Chinese translations of traditional Chinese query terms via English as the intermediate language.

We also collected popular query terms with the logs from two real-world Chinese search engines in Taiwan, i.e., Dreamer and GAIS[2]. The Dreamer log contained 228,566 unique query terms from a period of over 3 months in 1998, and the GAIS log contained 114,182 unique query terms from a period of two weeks in 1999. There were 9,709 most popular query terms whose frequencies were above 10 in both of the logs and 1,230 of them were English terms. After filtering out the terms which were used locally, we obtained 258 terms. These query terms were taken as the major test set in the term translation extraction analysis. The traditional Chinese translations of the test query terms were determined manually and taken as *the source query set* in the following experiments.

According to our previous work (Lu et al., 2001), there were three methods for term extraction, which is a necessary process step in extracting translations from anchor-text corpus. Since we have not yet collected a query log in simplified Chinese, in the following experiments we adopted the PAT-tree-based keyword extraction method, which is an efficient statistics-based approach that can extract longer terms without using a dictionary (Chien, 1997).

To evaluate the achieved performance of query translation, we used the *average top-n inclusion rate* as a metric. For a set of test query terms, its top-n inclusion rate is defined as the percentage of the query terms whose effective translation(s) can be found in the top *n* extracted translations.

4.2 Performance with the Direct Translation Model

In order to realize the feasibility of the transitive translation model, we carried out some experiments based on the direct translation models and the three different anchor-text set corpora in the first step. Table 1 shows the results of the obtained top-5 inclusion rates, where terms "TC", "SC" and "ENG" represent traditional Chinese, simplified Chinese and English terms respectively. The performance of translating TC into SC is worse than that of the other two since the size of the anchor-text set corpus containing both TC and SC is relatively small in comparison with the others. This is why we are pursuing in this paper to integrate the direct translation with the indirect translation via a third language. However, the performance of the direct translation from TC to SC is used as a reference in comparison with our proposed models in the following experiments.

Table 1. Top-n inclusion rates obtained with the direct translation model and the three specific language pairs corpora.

Type	Top1	Top2	Top3	Top4	Top5
TC=>SC	35.7%	43.0%	46.9%	49.6%	51.2%
TC=>ENG	68.6%	82.2%	85.7%	88.0%	88.8%
ENG=>SC	45.3%	55.8%	59.3%	61.6%	64.0%

4.3 Performance with the Indirect and Transitive Translation Models

To realize the improvement using the transitive translation model, some further experiments were conducted. As shown in Table 2, the indirect and transitive translation models outperform than the direct translation model. As mentioned above, the size of the anchor-text corpus that contains both TC and SC is small. The indirect translation model is, therefore, helpful to find out the corresponding translations for some terms with low-frequency values in the corpora. For example, the traditional Chinese term "西門子" was found can obtain its corresponding translation equivalent "西门子" in simplified Chinese via the intermediate translation "Siemens", which cannot be found only using the direct translation.

By examining the top-1 translations obtained with the three different models, it was found that the inclusion rates can be from 44.2% using the indirect translation to 49.2% using the transitive translation model. Table 3 illustrates some of the translations extracted using the transitive translation model.

[2] These two search engines are second-tier portals in Taiwan, whose logs have certain representatives in the Chinese communities, and whose URL's are as follows: http://www.dreamer.com.tw/ and http://gais.cs.cu.edu.tw/.

Table 2. Top-n inclusion rates obtained with different models.

Model	Top1	Top2	Top3	Top4	Top5
Direct Translation	35.7%	43.0%	46.9%	49.6%	51.2%
Indirect Translation	44.2%	55.1%	58.0%	59.7%	60.5%
Transitive Translation	49.2%	58.1%	60.9%	61.6%	62.0%
Combination of Transitive Translation and Lexicon	55.8%	60.8%	64.0%	65.9%	67.8%

4.4 Performance with an Integration of Lexicon Lookup

An additional experiment was also made to compare with the use of a translation lexicon for query translation. The lexicon contained more than 23,948 word/phrase entries in both traditional Chinese and simplified Chinese. It was found the top-1 inclusion rate that using the lexicon lookup was 12.4% which is obviously lower than the 49.2% that using the proposed transitive translation model. In addition, the top-1 inclusion rate can reach to 55.8% (see the last row of Table 2) if both of the approaches are combined. With the combined approach, the translation(s) of a query term is picked up from the lexicon if such a translation is already in the lexicon, otherwise it is obtained based on the transitive translation model.

5 Concluding Remarks

Anchor-text set corpus is a valuable resource for extracting translations of Web queries. How to exploit such kind of corpora in query translation is a challenging and potential research task. In this paper, we extend our previous approach by proposing a transitive translation model and achieve some improvements on translating those queries whose translations cannot be extracted using the previous approach. The improved approach has been proven particularly useful for the specific language pairs whose anchor texts are insufficient. However, there are still some problems need to be further investigated in the future.

References

Ballesteros, L. and Croft, W. B. (1997) *Phrasal Translation and Query Expansion Techniques for Cross-Language Information Retrieval*, Proceedings of ACM-SIGIR '97, pp. 84-91.

Borin, L. (2000) You'll Take the High Road and I'll Take the Low Road: Using a Third Language to Improve Bilingual Word Alignment, Proceedings of the 18th COLING, pp. 97-103.

Brown, P., Pietra, S. A. D., Pietra, V. D. J., Mercer, R. L. (1993) The Mathematics of Machine Translation, Computational Linguistics, 19(2), pp. 263-312.

Chakrabarti, S., Dom, B., Gibson, D., Kleinberg, J., Raghavan, P., Rajagopalan, S. (1998) *Automatic Resource List Compilation by Analysing Hyperlink Structure and Associated Text*, Proceedings of the seventh World Wide Web Conference.

Chien, L. F. (1997) *PAT-Tree-Based Keyword Extraction for Chinese Information Retrieval*, Proceedings of ACM-SIGIR '97, pp. 50-59.

Dagan, I., Church, K. W., Gale, W. A (1993) *Robust Bilingual Word Alignment for Machine Aided Translation.* Proceedings of the Workshop on Very Large Corpora, pp. 1-8.

Davis, M. and Ogden, W. C. (1997) *Quilt: Implementing a large-scale cross-language text retrieval system*, Proceedings of ACM-SIGIR'97 Conference, pp. 92-98.

Dumais, S. T., Landauer, T. K., Littman, M. L. (1996) *Automatic Cross-linguistic Information Retrieval Using Latent Semantic Indexing*, SIGIR'96 Workshop on Cross-Linguistic Information Retrieval, pp. 16-24.

Fung, P. and Yee, L. Y. (1998) *An IR Approach for Translating New Words from Nonparallel, Comparable Texts*, Proceedings of The 36th Annual Conference of the Association for Computational Linguistics, pp. 414-420.

Gollins, T., Sanderson, M. (2001) *Improving Cross language Information with Triangulated Translation*, Proceedings of ACM-SIGIR2001 Conference, pp. 90-95.

Kleinberg, J. (1998) *Authoritative Sources in a Hyperlinked Environment*, Proceedings of 9th ACM-SIAM Symposium on Discrete Algorithms.

Kwok, K. L. (2001) *NTCIR-2 Chinese, Cross Language Retrieval Experiments Using PIRCS,* Proceedings of NTCIR workshop meeting.

Lu, W. H., Chien, L. F., Lee, H. J. (2001) *Anchor Text Mining for Translation of Web Queries,*

Proceedings of The 2001 IEEE International Conference on Data Mining.

Nie, J. Y., Isabelle, P., Simard, M., and Durand, R. (1999) *Cross-language Information Retrieval Based on Parallel Texts and Automatic Mining of Parallel Texts from the Web*, Proceedings of ACM-SIGIR'99 Conference.

Rapp, R. (1999) *Automatic Identification of Word Translations from Unrelated English and German Corpora*, Proceedings of The 37th Annual Conference of the Association for Computational Linguistics.

Smadja, F., McKeown, K., Hatzivassiloglou, V. (1996) *Translating Collocations for Bilingual Lexicons: A Statistical Approach*, Computational Linguistics, 22(1), pp. 1-38.

Wong, S. K. M., Yao Y. Y. (1995) *On Modeling Information Retrieval with Probabilistic Inference*, ACM transactions on Information Systems, Vol.13, pp. 38-68.

Table 3. Some examples of extracted target translations with the three different models. (the asterisk indicates the correct translation)

Source terms in traditional Chinese	Top-5 extracted target translations in simplified Chinese		
	Direct Translation Model	Indirect Translation Model	Transitive Translation Model
西門子(Siemens)	Not available	西门子* (Siemens) 公司(Company) 中国(China) 网站(website) 合作(cooperation)	西门子* (Siemens) 公司(Company) 中国(China) 网站(website) 合作(cooperation)
康柏(Compaq)	Not available	康柏* (Compaq) 电脑公司(computer company) 公司(company) 美国(America) 电脑(computer)	康柏* (Compaq) 电脑公司(computer company) 公司(company) 美国(America) 电脑(computer)
新力(Sony)	索尼* (Sony) 本公司(our company) 新力* (Sony) 唱片(record) 中文版(Chinese version)	索尼* (Sony) 新力* (Sony) 电影网(movie site) 娱乐(entertainment) 唱片公司(record company)	索尼* (Sony) 新力* (Sony) 电影网(movie site) 娱乐(entertainment) 唱片公司(record company)