Proper Name Translation in Cross-Language Information Retrieval

Hsin-Hsi Chen, Sheng-Jie Huang, Yung-Wei Ding, and Shih-Chung Tsai
Department of Computer Science and Information Engineering
National Taiwan University
Taipei, TAIWAN, R.O.C.
hh_chen@csie.ntu.edu.tw

Abstract

Recently, language barrier becomes the major problem for people to search, retrieve, and understand WWW documents in different languages. This paper deals with query translation issue in cross-language information retrieval, proper names in particular. Models for name identification, name translation and name searching are presented. The recall rates and the precision rates for the identification of Chinese organization names, person names and location names under MET data are (76.67%, 79.33%), (87.33%, 82.33%) and (77.00%, 82.00%), respectively. In name translation, only 0.79% and 1.11% of candidates for English person names and location names, respectively, have to be proposed. The name searching facility is implemented on an MT sever for information retrieval on the WWW. Under this system, user can issue queries and read documents with his familiar language.

Introduction

World Wide Web (WWW) is the most useful and powerful information dissemination system on the Internet. For the multilingual feature, the language barrier becomes the major problem for people to search, retrieve, and understand WWW documents in different languages. That decreases the dissemination power of WWW to some extent. The researches of cross-language information retrieval abbreviated as CLIR (Oard and Dorr, 1996; Oard 1997) aim to tackle the language barriers. There are several important issues in CLIR:

(1) Queries and documents are in different languages, so that translation is required.
(2) Words in a query may be ambiguous, thus disambiguation is required.
(3) Queries are usually short, thus expansion is required.
(4) Word boundary in queries of some languages (Chen and Lee, 1996) is not clear, thus segmentation is required.
(5) A document may be in more than one language, thus language identification is required.

This paper focuses on query translation issue, proper name in particular.

The percentage of user queries containing proper names is very high. The paper (Thompson and Dozier, 1997) reported an experiment over periods of several days in 1995. It showed 67.8%, 83.4%, and 38.8% of queries to Wall Street Journal, Los Angeles Times, and Washington Post, respectively, involve name searching. In CLIR, three tasks are needed: name identification, name translation, and name searching. Because proper names are usually unknown words, it is hard to find in monolingual dictionary not to mention bilingual dictionary. Coverage is one of the major problems in dictionary-based approaches (Ballesteros and Croft, 1996; Davis, 1997; Hull and Grefenstette, 1996). Corpus-based approaches (Brown, 1996; Oard 1996; Sheridan and Ballerini, 1996) set up thesaurus from large-scale corpora. They provide narrow but specific coverage of the language, and are complementary to broad and shallow coverage in dictionaries. However, domain shifts and term align accuracy are major limitations of corpus-based approaches. Besides, proper names are infrequent words relative to other content words in corpora. In information

retrieval, most frequent and less frequent words are regarded as unimportant words and may be neglected.

This paper will propose methods to extract and classify proper names from Chinese queries (Section 1). Then, Chinese proper names are translated into English proper names (Section 2). Finally, the translated queries are sent to an MT sever for information retrieval on the WWW (Bian and Chen, 1997). The retrieved English home pages are presented in Chinese and/or English.

1 Name Extraction and Classification

People, affairs, time, places and things are five basic entities in a document. If we can catch the fundamental entities, we can understand the document to some degree. These entities are also the targets that users are interested in. That is, users often issue queries to retrieve such kinds of entities. The basic entities often appear in proper names, which are major unknown words in natural language texts. Thus name extraction is indispensable for both natural language understanding and information retrieval.

In famous message understanding system evaluation and message understanding conferences (MUC) and the related multilingual entity tasks (MET), named entity, which covers named organizations, people, and locations, along with date/time expressions and monetary and percentage expressions, is one of tasks for evaluating technologies. In MUC-6 named entity task, the systems developed by SRA (Krupka, 1995) and BBN (Weischedel, 1995) on the person name recognition portion have very high recall and precision scores (over 94%).

In Chinese language Processing, Chen and Lee (1996) present various strategies to identify and classify three types of proper nouns, i.e., Chinese person names, Chinese transliterated person names and organization names. In large-scale experiments, the average precision rate is 88.04% and the average recall rate is 92.56% for the identification of Chinese person names.

The above approaches can be employed to collect Chinese and English proper name sets from WWW (very large-scale corpora). Identification of proper names in queries is different from that in large-scale texts. The major difference is that query is always short. Thus its context is much shorter than full texts and some technologies involving larger contexts are useless. The following paragraphs depict the methods we adopt in the identification of Chinese proper names.

A Chinese person name is composed of surname and name parts. Most Chinese surnames are single character and some rare ones are two characters. A married woman may place her husband's surname before her surname. Thus there are three possible types of surnames, i.e., single character, two characters and two surnames together. Most names are two characters and some rare ones are one character. Theoretically, every character can be considered as names rather than a fixed set. Thus the length of Chinese person names range from 2 to 6 characters. Three kinds of recognition strategies shown below are adopted:

(1) name-formulation statistics
(2) context cues, e.g., titles, positions, speech-act verbs, and so on
(3) cache

Name-formulation statistics form the baseline model. It proposes possible candidates. The context cues add extra scores to the candidates. Cache records the occurrences of all the possible candidates in a paragraph. If a candidate appears more than once, it has high tendency to be a person name.

Transliterated person names denote foreigners. Compared with Chinese person names, the length of transliterated names is not restricted to 2 to 6 characters. The following strategies are adopted to recognize transliterated names:

(1) character condition
    Two special character sets are setup. The first character of transliterated names and the remaining characters must belong to these two sets, respectively. The character condition is a loose restriction. The string that satisfies the character condition may denote a location, a building, an address, *etc.* It should be employed with other cues (refer to (2)-(4)).
(2) titles
    Titles used in Chinese person names are

also applicable to transliterated person names.

(3) name introducers
Some words can introduce transliterated names when they are used at the first time.

(4) special verbs
The same set of speech-act verbs used in Chinese person names are also used for transliterated person names.

Cache mechanism is also helpful in the identification of transliterated names. A candidate that satisfies the character condition and one of the cues will be placed in the cache. At the second time, the cues may disappear, but we can recover the transliterated person name by checking cache.

The structure of organization names is more complex than that of person names. Basically, a complete organization name can be divided into two parts, i.e., name and keyword. Organization names, country names, person names and location names can be placed into the name part of organization names. Person names can be found by the approaches specified in the last paragraph. Location names will be touched later. Transliterated names may appear in the name part. We use the same character sets mentioned in the last paragraph. If a sequence of characters meet the character condition, the sequence and the keyword form an organization name. Common content words may be inserted in between the name part and the keyword part. In current version, at most two content words are allowed. Besides, we utilize the feature of multiple occurrences of organization names in a document and propose n-gram model to deal with this problem. Although cache mechanism and n-gram use the same feature, i.e., multiple occurrences, their concepts are totally different. For organization names, we are not sure when a pattern should be put into cache because its left boundary is hard to decide.

The structure of location names is similar to that of organization names. A complete location name is composed of a person name (or a location name) and a location keyword. For the treatment of location names without keywords, we introduce some locative verbs. Cache is also useful and N-gram model is employed to recover those names that do not meet the character condition.

We test our system with three sets of MET data (i.e., MET-1 formal run, MET-2 training, and MET-2 dry run). The recall rates and the precision rates for the identification of Chinese organization names, person names and location names are (76.67%, 79.33%), (87.33%, 82.33%) and (77.00%, 82.00%), respectively.

2 Proper Name Translation

Chinese and English are the source language and the target language, respectively, in our query translation. The alphabets of these two languages are totally different. Wade-Giles (WG) and Pinyin are two famous systems to romanize Chinese (Lu, 1995). The proper name translation problem can be formulated as:

(1) Collect English proper name sets from WWW.
(2) Identify Chinese proper names from queries.
(3) Romanize the Chinese proper names.
(4) Select candidates from suitable proper name sets.

In this way, the translation problem is transferred to a phonic string matching problem. If an English proper name denotes a Chinese entity, e.g., Lee Teng-hui denotes "李登輝" (President of R.O.C.), the matching is simple. Otherwise, the matching is not trivial. For example, we issue a query "阿爾卑斯山" in Chinese to retrieve information about Alps. The Pinyin romanization of this name is a.er.bei.si.shan[1]. The string "aerbeisishan" is not similar to the string "alps". We develop several language models incrementally to tackle the translation problem. The first issue we consider is how many common characters there are in a romanized Chinese proper name and an English proper name candidate. Here the order is significant. For example, the Chinese query is '埃斯其勒斯'. Its WG romanization is 'ai.ssu.chi.le.ssu'. The corresponding proper name is Aeschylus. Three characters (shown as follow in underline) are matched in order:

[1] The dot is inserted among romanization of Chinese characters for clear reading. Later, the dot may be dropped when strings are matched.

aeschylus

ais suchilessu

We normalize it by the length of the candidate (i.e., 9), and get a score 0.33. In an experiment, there are 1,534 pairs of Chinese-English person names. We conduct a mate matching: to use each Chinese proper name as a query, and try to find the corresponding English proper name from the 1,534 candidates. The performance is evaluated in such a way that how many candidates should be proposed to cover the correct translation. In other words, the average rank of correct translations is reported. The performances of the baseline model under WG and Pinyin systems are 40.06 and 31.05, respectively. The major problem of the baseline model is: if a character is matched incorrectly, those characters that follow this character will not contribute to the matching. In the above example, chi ('其') will be helpless for translation.

For reducing the error propagation, we consider syllables of the candidate in advance. The matching is done in syllables instead of the whole word. For example, Aeschylus contains three syllables. The matching is shown as follows:

aes	chy	lus
aissu	chi	lessu

The score is increased to 0.67 (6/9). In the similar experiment, the performances of the new language model are improved. The average ranks are 35.65 and 27.32 for WG and Pinyin systems, respectively.

Observing the performance differences between WG and Pinyin systems, we find they use different phones to denote the same sounds. The following shows examples:

(1) vowels
p vs. b, t vs. d, k vs. g, ch vs. j, ch vs. q, hs vs. x, ch vs. zh, j vs. r, ts vs. z, ts vs. c

(2) consonants
-ien vs. -ian, -ieh vs. -ie, -ou vs. -o, -o vs. -uo, -ung vs. -ong, -ueh vs. -ue, -uei vs. -ui, -iung vs. -iong, -i vs. -yi

A new language model integrates the alternatives. The average ranks of the mate match is 25.39. The result is better than those of separate romanization systems.

In the above ranking, each matching character is given an equal weight. We postulate that the first letter of each Romanized Chinese character is more important than others. For example, *c* in *chi* is more important than *h* and *i*. Thus it should have higher score. The following shows a new scoring function:

$$score=\Sigma_i(f_i*(el_i/(2* cl_i)+0.5)+o_i*0.5)/el$$

where

el: length of English proper name,

el_i: length of syllable *i* in English proper name,

cl_i: number of Chinese characters corresponding to syllable *i*,

f_i: number of matched first-letters in syllable *i*,

o_i: number of matched other letters in syllable *i*.

We reduplicate the above example as follows. The first letter is in capital.

aes	chy	lus
AiSsu	Chi	LeSsu

The corresponding parameters are listed below:

el_1=3, cl_1=2, f_1=2, o_1=0, el=9,
el_2=3, cl_2=1, f_2=1, o_2=1,
el_3=3, cl_3=2, f_3=2, o_3=0.

The new score of this candidate is 0.83. Under the new experiment, the average rank is 20.64. If the first letter of a Romanized Chinese character is not matched, we give it a penalty. The average ranks of the enhanced model is 16.78.

Table 1. The Performance of Person Name Translation

1	2-5	6-10	11-15	16-20	21-25	25+
524	497	107	143	44	22	197

We further consider the pronunciation rules in English. For example, *ph* usually has *f* sound. If all the similar rules are added to the language model, the average rank is enhanced to 12.11. Table 1 summarizes the distribution of ranks of the correct candidate. The first row shows the range of ranks. The second row shows the number of candidates within the range. About one-third have rank 1. On the average, only 0.79% of candidates have to be proposed to cover the correct solution. It shows this method is quite effective.

We also make two extra experiments. Given a query, the best model is adopted to find English locations. There are 1,574 candidates in this test. The average rank is 17.40. In other words, 1.11% of candidates have been

proposed. If we merge the person name set and location set, and repeat the experiment, the performance drops to 27.70. It tells us the importance of classification of proper names.

Conclusion

This paper proposes knowledge from character, sentence, and paragraph levels to identify different kinds of proper names. The person name translation problem is formulated as a phonic string matching problem. We consider the length of matching characters, syllables, different romanization systems, pronunciation rules, positive and negative scores in ranking. The name searching mechanism is integrated into a Chinese-English information retrieval system. In this way, languages are transparent to users on the Internet. In current implementation, only 0.79% and 1.11% of candidates for English person names and location names, respectively have to be proposed during name translation.

This model can be employed to set up a bilingual proper name dictionary. We can collect English and Chinese proper names from Internet periodically, and then conduct a mate matching. Human can be involved to select the correct translation. That will reduce the cost to develop a large scale bilingual proper name dictionary for name searching.

References

Ballesteros, L. and Croft, W.B. (1996) "Dictionary-based Methods for Cross-Lingual Information Retrieval." *Proceedings of the 7th International DEXA Conference on Database and Expert Systems Applications*, pp. 791-801.

Bian, G.W. and Chen, H.H. (1997) "An MT-Server for Information Retrieval on WWW." *Working Notes of the AAAI Spring Symposium on Natural Language Processing for the World Wide Web*, 1997, pp. 10-16.

Brown, R.D. (1996) "Example-Based Machine Translation in the Pangloss System." *Proceedings of 16th International Conference on Computational Linguistics*, pp. 169-174.

Chen, H.H. and Lee, J.C. (1996) "Identification and Classification of Proper Nouns in Chinese Texts." *Proceedings of 16th International Conference on Computational Linguistics*, 1996, pp. 222-229.

Hull, D.A. and Grefenstette, G. (1996) "Querying Across Languages: A Dictionary-based Approach to Multilingual Information Retrieval." *Proceedings of the 19th International Conference on Research and Development in Information Retrieval*, pp. 49-57.

Krupka, G.R. (1995) "SRA: Description of the SRA System as Used for MUC-6." *Proceedings of Sixth Message Understanding Conference*, 1995, pp. 221-235.

Mani, I., *et al.* (1993) "Identifying Unknown Proper Names in Newswire Text." *Proceedings of Workshop on Acquisition of Lexical Knowledge from Text*, 1993, pp. 44-54.

McDonald, D. (1993) "Internal and External Evidence in the Identification and Semantic Categorization of Proper Names." *Proceedings of Workshop on Acquisition of Lexical Knowledge from Text*, 1993, pp. 32-43.

Oard, D.W. (1997) "Alternative Approaches for Cross-Language Text Retrieval." *Working Notes of AAAI-97 Spring Symposiums on Cross-Language Text and Speech Retrieval*, pp. 131-139.

Oard, D.W. (1996) *Adaptive Vector Space Text Filtering for Monolingual and Cross-language Applications*, Ph.D. Dissertation, University of Maryland.

Oard, D.W. and Dorr, B.J. (1996) *A Survey of Multilingual Text Retrieval.* Technical Report UMIACS-TR-96-19, University of Maryland, Institute for Advanced Computer Studies. http://www.ee.umd.edu/medlab/filter/papers/mlir.ps.

Paik, W., *et al.* (1993) "Categorizing and Standardizing Proper Nouns for Efficient Information Retrieval." *Proceedings of Workshop on Acquisition of Lexical Knowledge from Text*, 1993, pp. 154-160.

Sheridan, P. and Ballerini, J.P. (1996) "Experiments in Multilingual Information Retrieval Using the SPIDER System." *Proceedings of the 19th ACM SIGIR Conference on Research and Development in Information Retrieval*, pp. 58-65.

Thompson, P. and Dozier, C. (1997) "Name Searching and Information Retrieval." *Proceedings of Second Conference on Empirical Methods in Natural Language Processing*, Providence, Rhode Island, 1997.

Weischedel, R. (1995) "BBN: Description of the PLUM System as Used for MUC-6." *Proceedings of Sixth Message Understanding Conference*, 1995, 55-69.