Nathu IR System at NTCIR-2

Jason S Chang, David Yu, Ching Ting Shen
Afra Cheng, Garfield Shen, Giordano Shen, David Wong
National Tsing Hua University
jschang@cs.nthu.edu.tw

Abstract

This paper describes the NTHU system for monolingual Chinese ad hoc and English-Chinese cross-lingual tasks in NTCIR project. We use NTCIR-2 document collection to research how to deal with quantifier and logical structure in the topics under the vector space model. We also experimented with different methods for translating phrasal query term and proper names.

Keyword: query formulation, query translation, proper name

1 Introduction

The paper describes the participation of the system at National Tsing Hua University in two NTCIR tasks, ad hoc monolingual Chinese task and English-Chinese cross-lingual document retrieval tasks. We used very simple strategy to perform text analysis, index construction and document ranking. Our main focus was to experiment and find effective methods for query formulation and translation.

2 System Overview

The system consists of Chinese segmentation unit, index construction unit, keyword extraction unit, query formulation unit, and query processing and document-ranking unit. The segmentation unit used a general-purpose dictionary without adaptation to the task at hand. Both the documents and topics went through the segmentation unit. The keyword extraction unit was used to extract important phrases in the topics. The index construction unit built an inverted file from the word stream produced by the segmentation unit. For the sake of processing speed, we only indexed the first 500 words of each document [9]. In query formulation unit, we tried to deal with quantifiers and logical statements in the topics. The query-processing unit used the common method for manipulating and ranking list of candidate documents. In the following sections we will describe these units in turns.

3 Segmentation

We used the Academia Sinica's corpus to build a Chinese segmentation unit for unrestricted text. No attempt was made to adapt the dictionary to the NTCIR document collection. In addition to segmentation, we also used a keyword extraction tool to help identify important phrases in the topics. By using word as the index element, we gained efficiency in building the inverted file structure. However, the dictionary is not constructed specifically for the task. Lack of proper names and terminology specific to the NTCIR Chinese Collection might reduce the effectiveness of segmentation. To alleviate part of the problems, we built proper name capability into the segmentation unit [3, 4, 6]. That resulted in a dictionary of words used in the collection containing not only the original word from Academia Sinica's corpus but also unregistered proper names found in NTCIR collection. That is important for query translation of proper names in the English topics, as we will describe in Section 4.

The segmentation unit is biased toward the longest possible match with dictionary entries. Sometimes, that could lead to mismatch between a topic and relevant documents. For instance, longest match strategy favors the long word, "留學生," which will not match a relevant document containing a shorter but related word "留學." Many researchers have long been advocating using both character unigrams and bigrams as basic index unit for that reason.

4 Query formulation and translation

The query formulation for the Chinese ad hoc task focuses mainly on automatic identification of important phrases from the title and description sections. The selection criteria include term frequency and syntactic properties. These phrases were treated differently according to the types of topics they appeared in. We found that the 50 queries can be divided into three types:

Type-A: Relevant documents contain X.
Type-B: Relevant documents contain X but not Y
Type-C: Relevant documents contain X and Y. However, a document containing only X is not relevant.

Basically, we used the vector space model to formulate and process the queries. However, three types of query should be treated differently. For type-A query, we simply assigned weights to phrases according to term's frequency and position in the 50 topics. The query was formulated as a list of weighted term under the vector space model. Type-B queries were treated quite similar to type-A ones except that a term with a negative quantifier was given a negative weight. For type-C query, we produced two formal queries and proceeded in two steps:

(i) Query Y
Retrieve top-ranking document containing Y under the vector space model and call the set of document a range R_Y.

(ii) Query X
Retrieve top-ranking document containing X within the "range" of R_Y. We call R_Y the "range" of search and the query of (X, R_Y) a "range query."

The query translation [1] for the English-Chinese task focuses mainly on selecting among all possible translations for query terms. We concentrated on the translation of phrases and transliteration of proper names.

We noticed that Chinese phrases embodying key concepts in a query or document tend to be lexicalized [5] and lexicalized terms often appear alternatively in abbreviated form. For instances, the term 集會遊行法 (Assembly and Parade Law) is often used in the abbreviated form as 集遊法, much like the use of acronyms in English text. However, one needs to be able to find such abbreviated translations for a query term in order to perform effect English-Chinese CLIR task. The translation of transliteration terms is also a problem, since each transliteration syllable can be translated back in so many ways due to many homophonic characters in Chinese. We dealt with both problems in steps: First, each document goes through a Chinese word segmentation system where words, abbreviation forms and proper names are identified. All the distinct tokens went into a dictionary. Each phrase term in the English topic was then process to find closest translation in the dictionary. That was done by formulating a Boolean query for each query term. The query is subsequently processed as an extended Boolean query to find closest match in the NTCIR dictionary. The query was designed to constrain the phrasal translation so that it covered all constituent words with at least some Chinese morphemes of their alternative translations. For instance, a Boolean query

(組 or 合 or 集 or 會) and
(遊 or 行) and
(法 or 律 or 法 or 則)

was formulated for "Assembly and Parade Law" if parade can be translated alternatively as 組合 or 集會, parade as 遊行, and law as 法律 or 法則. The closest match in dictionary for this extended Boolean turned out to be 集遊法 covering one morpheme each from the translations of "assembly," "parade," and "law."

The transliterated proper names were done similarly. Each syllable in a proper name was lookup in a transliteration table for appropriate codes in Chinese phonetic system and corresponding Chinese characters. These characters form a Boolean query much like the cases for common nouns in order to cover all syllables with one of the alternative Chinese translation. For instance, the Boolean query for finding the translation for "Shih Ming-Teh" was

(失 or 十 or 施 or 時 or ...) and
(名 or 明 or 銘 or ...) and
(的 or 德 or 得 or 的 or 淂)

Although it looks as if there are many combinations of three-character word would satisfy this Boolean expression. However, since the search was restricted within the NTCIR dictionary, correct translation was found in most cases. When more than one satisfied the constraint, we resolved the conflict in favor of the one with higher frequency.

The closest match in the Chinese dictionary (consisting of all terms from NTCIR Chinese collection) to the extended Boolean query was then used to replace the English query term. Finally, the translated query was run against the Chinese collection.

The translation of proper names was quite effective with over 80% of proper names converted to the correct names in the original language of Chinese. The results were mixed when it came to translation of other phrasal terms. There was time the translated query term was just right. For instance, "water and soil conservation" was formulated as

(土 or 壤 or 泥) and (水 or 開) and
(保 or 守 or 持).

That expression for translating query "water and soil conservation" correctly matched 水土保持 in the dictionary. However, due to the quality of bilingual dictionary for translating singleton word and the limit of the method itself, only half of the terms were translated satisfactorily.

5 Indexing and Query Processing

We used a straightforward inverted file structure to store index terms and their postings. For the sake of processing speed, we only indexed the first 500 words of each document. The inverted file was constructed by using a sort-based method. The query terms were process in increasing order of their document frequency. A fixed upper limit of 5,000 documents was used to guarantee a reasonable time for processing query be done within. When the limit was reached, we simply stop creating space for new document candidates. However, we continued to process remaining query terms and to calculate the relevance score for existing document candidates.

Ranking was done based on term frequency alone without taking document frequency into consideration. Additionally, the position of each term occurrence was also considered important factor of relevance. We favored the term appear near the beginning of the document by assigning a weighting factor in reverse proposition to the square root of distance between the beginning of document and the position of the term.

Table 1 Evaluation results

Runs	Avg. Prec.	R-Precision
Nthu-chir-lo-01	0.5009	0.5050
Nthu-chir-lo-02	0.4847	0.4964
Nthu-chir-lo-03	0.4257	0.4403
Nthu-ecir-lo-01	0.1652	0.1926

Table 2 Interpolated recall vs. average precision for the Nthu-chir-lo-01 run

Recall	Precision	Recall	Precision
0.00	0.9064		
0.10	0.8252	0.20	0.7389
0.30	0.6559	0.40	0.6114
0.50	0.5475	0.60	0.4820
0.70	0.3455	0.80	0.2578
0.90	0.1616	1.00	0.0632
Avg	0.5009		

Table 3 Precision vs. number of documents retrieved

# of doc retrieved	Precision:
5 docs	0.7600
10 docs	0.6880
15 docs	0.6120

# of relevant docs	0.5050

6 Evaluation Results

We carried out three runs to assess the effectiveness of our query formulation strategy (Nthu-chir-lo-02). Table 1 shows that the strategy outperformed straightforward vector space model (Nthu-chir-lo-01) quite significantly.

We also experimented indexing locations and numbers (Nthu-chir-lo-01) in order to be able to accommodate topics that contain reference to indefinite quantities and location. The experimental results show only marginal increase in precision.

7 Conclusions

The NTHU system was built based on very simple techniques. However, it still delivered reasonable performance and allowed us to experiment on the problem of automatic query formulation and translation. We showed that proper treatment of quantifiers indeed improved precision significantly over straightforward bag of word treatment under the vector space model. We also found that coupling a general dictionary with the task-specific dictionary is quite effective for query translation in English-Chinese cross-linguistic task, given that the dictionary is built beforehand with a word segmentation unit that recognizes proper names.

References

[1] Chen, H.H., et al. (1998) "Proper Name Translation in Cross-language Information Retrieval." Proceedings of 17th International Conference on Computational Linguistics and 36th Annual Meeting of ACL.

[2] Croston-Oliver, S.H. and W.B. Dolan (1998) *Less is More.* In Proceedings of Joint International Conference on Computational Linguistics and Annual Meeting of the Association for Computational Linguistics, pp.349-356.

[3] H.H. Chen, Y.W. Ding, and S.C. Tsai (1998), "Named Entity Extraction for Information Retrieval", Computer Processing of Oriental Languages, vol. 12, No. 1, 1998, pp.75-85 .

[4] Jason S Chang, S.D. Chen, S.J. Ker, Y. Chen, and John S. Liu (1994), "A Multiple-Corpus Approach to Recognition of Proper names in Chinese Texts", Computer Processing of Chinese and Oriental Languages, vol. 8, No. 1, June 1994, pp.75-85.

[5] Justeson, John S. and Slava M. Katz. (1995) *Technical Terminology: Some Linguistic Properties and an Algorithm for Identification in Text.* Natural Language Engineering, Vol. 1, Pt. 1, March 1995, pp.9-27.

[6] Liddy E. D., W. Paik, S.E. Yu and M. McKenna (1995) *Document Retrieval Using Linguistic Knowledge*, In Proceedings of Intelligent Multimedia Information Retrieval Systems and Management, pp. 106-114.

[7] Sproat, C. Shih, William Gale, and Nancy Chang (1996), "A Stochastic Finite-State Word-Segmentation Algorithm for Chinese", Computational Linguistics, vol. 22, No. 3, 1996, pp.377-403.

[8] Strzalkowski, Tomek (1996) Natural Language Information Retrieval, Information Processing and Management, Vol. 31, No. 3, pp.397-417, Pergamon Elsevier.

[9] Wasson, Mark. (1998) *Using Leading Text for News Summaries: Evaluation Results and Implication for Commercial Summarization Applications.* In Proceedings of Joint International Conference on Computational Linguistics and Annual Meeting of the ACL, pp.1264-1368.