NTU at CLEF 2001: Chinese-English Cross-Lingual Information Retrieval

Hsin-Hsi Chen and Wen-Cheng Lin

Department of Computer Science and Information Engineering
National Taiwan University
Taipei, TAIWAN, R.O.C.
E-mail: hh_chen@csie.ntu.edu.tw, denislin@nlg2.csie.ntu.edu.tw

Abstract

This paper reports the work of NTU on bilingual-retrieval task at CLEF 2001. We proposed five models. Model 1 used co-occurrence information to disambiguate translation equivalents; Model 2 augmented restriction terms to the original queries; Model 3 used C-E WordNet to translate queries; Model 4 combined Model 3 with Model 2; Model 5 merged the queries constructed by Model 2 and 3. The best one is Model 5. The average precision of Model 5 is 0.1135, which is 53.06% of monolingual information retrieval.

1. Introduction

National Taiwan University (NTU) Natural Language Processing Laboratory (NLPL) participated bilingual-retrieval task at CLEF 2001. In this experiment, we used Chinese queries to retrieve English documents and query translation was adopted to unify the language of queries and documents.

In our previous works, several approaches were proposed. Bian and Chen (1998) proposed a hybrid approach that integrated both lexical and corpus knowledge to translate queries. A bilingual dictionary provides the translation equivalents of each query term, and the word co-occurrence information trained from a target language text collection is used to disambiguate the translation. Target polysemy is another problem in CLIR. Chen , Bian and Lin (1999) augmented a pseudo context to a query term to restrict its use in the target language. The contextual information is trained from a source language text collection. Chen, Lin and Lin (2000) proposed a method to construct a Chinese-English WordNet automatically. We used this C-E WordNet and a bilingual dictionary to translate queries. In this paper, we propose a combined approach that use C-E WordNet and augmented restrictions to construct target queries.

2. Resources

In this work, we used four linguistic resources:

(1) Chinese-English dictionary

The bilingual dictionary is integrated from four sources, including LDC Chinese-English dictionary, Denisowski's CEDICT, BDC Chinese-English dictionary v2.2 and a dictionary used in query translation in MTIR project (Bian and Chen, 2000). The dictionary gathers 200,037 words, where a word may have more than one translation.

(2) ASBC (Huang, *et al.*, 1995)

Academic Sinica Balanced Corpus (abbreviated as ASBC corpus) is a POS-tagged Chinese balanced corpus. The major topics include philosophy (10%), science (10%), society (35%), art (5%), life (20%), and literary (20%). This corpus is composed of five million words.

(3) TREC6 text collection (Harman, 1997)

The text collection contains 556,077 English documents, and is about 2.2G bytes.

(4) Chinese-English WordNet

In our previous work (Chen, Lin and Lin, 2000), we proposed a method to construct a Chinese-English WordNet automatically. Chinese words in tong2yi4ci2ci2lin2 ("同義詞詞林") (Mei, *et al.*, 1982) are mapped into WordNet (Fellbaum, 1998). Following the structures of WordNet, a Chinese WordNet and a Chinese-English WordNet are derived.

The co-occurrence information of Chinese and English words was trained from ASBC corpus and TREC6 text collection respectively. We adopted mutual information (MI) (Church, *et al*., 1989) to measure its strength. For each word, we collected its mutual information with other words within a window of size 3.

3. Query translation

We adopted query translation to unify the language of queries and documents. The Chinese queries were translated into English. The translated English queries were used to retrieve English documents using a monolingual information retrieval system. We proposed four models to translate queries. Model 1 uses co-occurrence information trained from a text collection in source language to select the best translation equivalents of source language query terms. Model 2 tries to resolve the target polysemy problem by augmenting some restriction words. Model 3 uses automatic constructed C-E WordNet to translate queries. Model 4 combines Model 2 and 3.

3.1. Model 1 – CO Model

At first, the Chinese queries were segmented. For each Chinese word, we collected the translation equivalents by looking up a Chinese-English bilingual dictionary. Then the best translation equivalents were selected by using the co-occurrence information. The mutual information was trained from a text collection in target language, i.e. TREC6 text collection. For a query term, we compare the MI values of all the translation equivalent pairs (x, y), where x is the translation equivalent of this term, and y is the translation equivalent of another query term within a sentence. The word pair (x_i, y_j) with the highest MI value is extracted, and the translation equivalent x_i is regarded as the best translation equivalent of this query term. Selection is carried out based on the order of the query terms.

3.2. Model 2 – Resolving Target Polysemy Problem

In order to resolve target polysemy problem, we augmented some words to restrict the use of a translated query term in target language. In this model, the Chinese queries were translated by CO model, and the translation equivalents of augmented words were added to target language queries. The augmented restriction words of a source language query term are those words that frequently co-occur with it within a window. The

co-occurrence information was trained from ASBC corpus, and the mutual information was used to measure the strength. We collected the co-occurred terms that have only one translation as the candidates. Then we apply CO model to the translations of these candidates and select one term for each original query term.

The translations of original query terms and augmented restriction terms were assigned different weights. They were determined by the following formula:

$$\text{weight}(E_i) = \sum_{k=1}^{n} m_k \quad (1)$$

$$\text{weight}(EW_{ij}) = 1 \quad (2)$$

Where n is number of words in a query Q; E_i is the translation of query term C_i; EW_{ij} is the translation of augmented restriction term CW_{ij} and m_k is the number of words in a restriction for C_k.

3.3. Model 3 – Using Chinese-English WordNet

In this model, Chinese-English WordNet was used to construct English queries. First, a Chinese query was tagged by a POS tagger. After removing stop words, we looked up the Chinese-English WordNet for the remaining Chinese words. A set of synsets was retrieved for each Chinese query term. We computed the mutual information for the sets of synsets, and selected a synset for each Chinese query term. The mutual information of two synsets is defined as follows. Let $synset_1$ and $synset_2$ be synsets for two query terms. Assume $synset_1$ and $synset_2$ are composed of m and n English words, respectively.

$$MI(synset_1, synset_2) = \sum_{i=1}^{m}\sum_{j=1}^{n} MI(t_{1i}, t_{2j})/(m \times n) \quad (3)$$

Where t_i is the English word in $synset_i$. The MI values of any two English words are trained from TREC-6 corpus. All English words in the selected synsets were used to construct the target query. The translation equivalents in the selected synsets were assigned larger weights. The weights of translation equivalents in the selected synsets were 3 and that of other words were 1.

When looked up Chinese-English WordNet , some query terms can't be found. For these query terms, we add their translation equivalents to the English query. The weights of these translation equivalents were 1.

3.4. Model 4 – Combined Approach

Consider the terms that can't be found in Chinese-English WordNet in Model 3, we used translations and restriction terms obtained in Model 2 instead of all translation equivalents retrieved from our bilingual dictionary. The weights of these translations were 3 and that of restriction terms were 1.

4. IR system

Our Information Retrieval system is based on vector space model. The index terms are English words, and the term weighting function is tf*idf. When a query is submitted to this IR system, it computes the similarities of this query and all documents, then returns top rank documents. We adopt cosine vector similarity formula to measure the similarity of a query and a document. Higher score means that the query and the document are

more similar.

5. Results

We submitted four runs: NTUco, NTUa1wco, NTUaswtw and NTUtpwn. The English queries of these four runs were constructed by Model 1, 2, 3 and 4, respectively. In our experiments, only the Title and Description fields were used to generate queries. The results are shown in Table 1. There were some bugs in our IR system. Only the documents in January, February and March were indexed. We re-indexed all documents and did four new runs: CO, A1WCO, ASWTW and TPWN. We also did an unofficial run: MONO, a monolingual run. The results are shown in Table 2.

Table 1. Results of official runs

Run	Average precision	R-Precision	Rel_ret
NTUco	0.0254	0.0292	134
NTUa1wco	0.0255	0.0297	135
NTUaswtw	0.0224	0.0328	149
NTUtpwn	0.0195	0.0301	141

Table 2. Results of new runs

Run	Average precision	R-Precision	Rel_ret
MONO	0.2139	0.2039	611
CO	0.1108 (51.8%)	0.1214 (59.54%)	482
A1WCO	0.1107 (51.75%)	0.1198 (58.75%)	485
ASWTW	0.0816 (38.15%)	0.0814 (39.92%)	472
TPWN	0.1080 (50.49%)	0.1172 (57.48%)	491
TPWN2	0.1135 (53.06%)	0.1201 (58.90%)	512

The average precision of run CO is 0.1108, which is 51.8% of monolingual information retrieval. The performances of some queries were very bad. Word segmentation errors may be one of the reasons. Take “史特加” as an example. The word “史特加” (Schneider) was segmented into “史”, “特” and “加”, and were translated into “history”, “unusual” and “recruit” respectively. Dictionary coverage is another problem. Some proper nouns are not included in our bilingual dictionary. For example, “歐斯基爾肯” (Euskirchen) is not included in the dictionary. Because the lack of the translation of “歐斯基爾肯”, the relevant document of query 75 can’t be retrieved.

The performance of run A1WCO is almost the same as run CO. In Model 2, we add some restriction terms to the original queries. The augmented restriction terms help us to retrieve more relevant documents, but the average precision decrease. When we add words to the original queries, we may also introduce noises. Some augmented restriction terms are related to the query terms that the restriction terms are augmented to, but it

is not relevant to the queries. Thus, these terms become noises.

When we used C-E WordNet, the performance is not good. While constructing C-E WordNet, some Chinese words may have been mapped to wrong synsets. For example, "中國" (China) was mapped to the synset that only contain "Kyushu". Thus we can't find any document that relevant to "Chinese Currency Devaluation". In Model 3, we used all translation equivalents of the words that are not included in C-E WordNet. In this way, some inappropriate translations were also added to the target queries. In Model 4, we used the translations and restriction terms that obtained from Model 2. The result shows that the performance is improved. The average precision of run TPWN is 0.1080, which is 50.49% of monolingual information retrieval. It is better than run ASWTW, but still worse than other runs. We try another combination method. We simply merge the target queries that constructed by Model 2 and 3. The last row of Table 2 shows the result. The average precision of run TPWN2 is 0.1135.

6. Conclusions

In CLEF 2001, we proposed five models. Model 1 used a hybrid approach that integrated both lexical and corpus knowledge to translate queries. The word co-occurrence information is used to disambiguate translation equivalents; Model 2 augmented some restriction terms to the original queries to deal with target polysemy problem; Model 3 used C-E WordNet to translate queries; Model 4 combined Model 3 with Model 2; Model 5 merged the queries constructed by Model 2 and 3. The best one is Model 5. The average precision of Model 5 is 0.1135, which is 53.06% of monolingual information retrieval.

Dictionary coverage is a problem while translating queries. Since the important words of some queries are not included in our bilingual dictionary, the performances of these queries were bad. Word segmentation error is another problem. If a word is not segmented correctly, we can't find it's correct translation. In Model 3, we found that the C-E WordNet has errors. Some Chinese words may have been mapped to wrong synsets. In the future, we will refine the bilingual dictionary and C-E WordNet and conduct more precise error analysis.

Reference

Bian, G.W. and Chen, H.H. (1998) "Integrating Query Translation and Document Translation in a Cross-Language Information Retrieval System." *Machine Translation and Information Soup*, Lecture Notes in Computer Science, No. 1529, Spring-Verlag, 250-265.

Bian, G.W. and Chen, H.H. (2000). "Cross language information access to multilingual collections on the Internet." *Journal of American Society for Information Science*, 51(3), 281-296.

Chen, H.H., Bian, G.W. and Lin, W.C. (1999) "Resolving Translation Ambiguity and Target Polysemy in Cross-Language Information Retrieval." In *Proceedings of 37th Annual Meeting of the Association for Computational Linguistics*, 215-222.

Chen, H.H., Lin, C.C., and Lin, W.C. (2000). "Construction of a Chinese-English WordNet and Its Application to CLIR." In *Proceedings of the Fifth International Workshop on Information Retrieval with Asian Languages*, 189-196. Hong Kong: ACM.

Church, K.W., *et al*. (1989). "Parsing, Word Associations and Typical Predicate-Argument Relations." In *Proceedings of International Workshop on Parsing Technologies*, 389-398.

Fellbaum, C. (Ed.). (1998). *WordNet: An Electronic Lexical Database*. Cambridge, MA: MIT Press.

Harman, D.K. (1997) *TREC-6 Proceedings*, Gaithersburg, Maryland.

Huang, C.R., *et al*. (1995) "Introduction to Academia Sinica Balanced Corpus." In *Proceedings of ROCLING VIII*, Taiwan, 81-99.

Mei, J., *et al*. (1982). *tong2yi4ci2ci2lin2*. Shanghai Dictionary Press.