

Using Co-occurrence, Augmented Restrictions, and C-E WordNet for Chinese-English Cross-Language Information Retrieval at CLEF 2001

Wen-Cheng Lin and Hsin-Hsi Chen

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

Abstract. This paper reports the work of NTU in the bilingual-retrieval task at CLEF 2001. In this experiment, we compared the effectiveness of several approaches in Chinese-English cross-language information retrieval. Five models were proposed. Model 1 used co-occurrence information in the target language to disambiguate translation equivalents; Model 2 augmented restriction terms to the original queries to restrict the use of query terms in the target language; Model 3 used a Chinese-English WordNet to translate queries; Model 4 combined Model 3 with Model 2; Model 5 merged the queries constructed by Model 2 and 3.

1 Introduction

The Natural Language Processing Laboratory (NLPL), National Taiwan University (NTU) participated in the bilingual-retrieval task at CLEF 2001. Cross language information retrieval (CLIR) [1][2] deals with the use of queries in one language to access documents in another. Since the languages of queries and documents are different, the performance of using source language queries directly is usually very poor. In order to cross the language barrier, we can translate queries into the language that documents are written in or translate documents into the language that queries are described in or translate both queries and documents into an intermediate language. Query translation is usually employed for efficiency. 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 work, several approaches were proposed. Bian and Chen [3] 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 derived from a target language text collection is used to disambiguate the translation. Mutual information (MI) [4] is used to measure the co-occurrence strength. For a query term, the translation equivalent with the highest MI value is selected. Target polysemy is another problem in CLIR. Chen, Bian and Lin [5] augmented a pseudo context to a query term to restrict its use in the target language. The contextual information is derived from a

source language text collection. Chen, Lin and Lin [6] 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 experiment with the approaches described above. In addition, we propose a combined approach using the C-E WordNet and the 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 the MTIR project [7]. The dictionary gathers 200,037 words, where a word may have more than one translation.
- (2) ASBC [8]
Academic Sinica Balanced Corpus (abbreviated as the 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 [9]
The text collection contains 556,077 English documents, and is about 2.2G bytes.
- (4) Chinese-English WordNet
In our previous work [6], we proposed a method to construct a Chinese-English WordNet automatically. Chinese words in a Chinese thesaurus tong2yi4ci2ci2lin2 ("同義詞詞林") [10] are mapped into WordNet [11]. Following the structures of WordNet, a Chinese WordNet and a Chinese-English WordNet are derived.

When translating queries and selecting augmented restriction terms, the co-occurrence information between words is used to select best translation and appropriate restriction terms. The co-occurrence information for Chinese and English words was derived from the ASBC corpus and TREC6 text collection respectively. We adopted the mutual information formula to measure its strength. For each word, we collected its mutual information value 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 derived from a text collection in the target 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 an automatically constructed C-E WordNet to translate queries. Model 4 combines Models 2 and 3.

3.1 Model 1 – CO Model

When translating queries, a query term may have more than one sense. If all translations of a polysemous word are included in the target query, the incorrect senses are also included and may reduce performance. Therefore, a selection operation should be adopted to select appropriate translations. Bian and Chen [3] proposed a hybrid approach that integrated both lexical and corpus knowledge to translate queries. 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 derived from a text collection in the 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, y) with the highest MI value is extracted, and the translation equivalent x 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 the Target Polysemy Problem

In order to resolve the target polysemy problem, we augmented some words to restrict the use of a translated query term in the target language. In this model, the Chinese queries were translated by the 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 derived from the ASBC corpus, and the mutual information formula was used to measure the strength. We collected the co-occurring terms that have only one translation as candidates. Then we applied the CO model to the translations of these candidates and select one term for each original query term.

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

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

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

where n is the number of words in a query Q ; E_i is the translation of query term C_i ; EW_{ij} is the translation of the 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, the 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. A synset is a set of synonyms that can be used to express a concept. 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:

$$MI(\text{synset}_1, \text{synset}_2) = \sum_{i=1}^m \sum_{j=1}^n MI(t_{1i}, t_{2j}) / (m \times n) \quad (3)$$

where t_{ik} is the k th English word in synset_i . The MI values of any two English words are derived from the TREC6 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 in the Chinese-English WordNet, some query terms were not found. For these query terms, we added their translation equivalents to the English query. The weights of these translation equivalents were 1.

3.4 Model 4 – Combined Approach

In Model 3, we used all translation equivalents for those terms that cannot be found in the Chinese-English WordNet. If a term is polysemous, using all its translation equivalents will introduce noise. In this model, 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 the restriction terms were 1.

4 The IR System

Our Information Retrieval system is based on the vector space model. The index terms are English words, and the term weighting function is $\text{tf} \cdot \text{idf}$. 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 the cosine vector similarity formula to measure the similarity of a query and a document. A 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 using Models 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 (51.80%)	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
ASWTW2	0.1011(47.27%)	0.1051(51.54%)	522
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”, which 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 of the lack of the translation of “歐斯基爾肯”, the relevant document of query 75 cannot 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 noise. Some augmented restriction terms are related to the query terms that the restriction terms are augmented to, but are not relevant to the queries. Thus, these terms become noise.

When we used C-E WordNet, the performance was 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 cannot find any document that is relevant to “Chinese Currency Devaluation”. In run ASWTW, the weights of translation equivalents obtained from the dictionary are lower than those of translations in selected synsets. The reason is that we try to reduce the interference of inappropriate translations. But this also reduces the importance of correct translations. We adjusted the weights of these translations equivalents in a new run ASWTW2. The translations obtained from dictionary and synsets are assigned the same weight, i.e. 3. The average precision is 0.1011.

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 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 tried another combination method. We simply merged the target queries that are constructed by Model 2 and 3. The last row of Table 2 shows the result. The average precision of run TPWN2 is 0.1135.

In the paper [5], we tried to resolve the target polysemy problem by augmenting a pseudo context to a query term to restrict its use in the target language. We experimented on the TREC6 text collection and the result showed that the performance of this method is slightly better than the CO model (0.0918 vs. 0.0831). Chen, Lin and Lin [6] proposed a method to construct a Chinese-English WordNet automatically. We used this C-E WordNet and a bilingual dictionary to translate queries. The average precision was increased to 0.1010. We participated in the TREC9 Cross-Language track last year [12]. We used English queries to retrieve Chinese documents. The performances of the CO model and the A1W model were very close. In the A1W model, an original query term was augmented by all unambiguous terms that frequently co-occur with it. In CLEF 2001, the performance of the CO model is slightly better than that of A1WCO model and ASWTW model. Reviewing the results of these experiments, the performance of augmenting restriction terms is close to the CO model. But it takes more time to select restriction terms. The Chinese-English WordNet is a useful resource, but it still contains errors. If the Chinese-English WordNet is revised by humans, the performance obtained by using this C-E WordNet to translate query should improve.

6 Conclusions

At 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 the 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 cannot find its 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.

References

1. Oard, D.W.: Alternative Approaches for Cross-Language Text Retrieval. In: Working Notes of AAAI-97 Spring Symposiums on Cross-Language Text and Speech Retrieval, (1997) 131-139.
2. Oard, D.W. and Dorr, B.J.: 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>. (1996).
3. Bian, G.W. and Chen, H.H.: 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. Springer-Verlag (1998) 250-265.
4. Church, K.W., *et al.*: Parsing, Word Associations and Typical Predicate-Argument Relations. In: Proceedings of International Workshop on Parsing Technologies (1989) 389-398.
5. Chen, H.H., Bian, G.W. and Lin, W.C.: Resolving Translation Ambiguity and Target Polysemy in Cross-Language Information Retrieval. In: Proceedings of 37th Annual Meeting of the Association for Computational Linguistics (1999) 215-222.
6. Chen, H.H., Lin, C.C., and Lin, W.C.: Construction of a Chinese-English WordNet and Its Application to CLIR. In: Proceedings of the Fifth International Workshop on Information Retrieval with Asian Languages. Hong Kong: ACM (2000) 189-196.
7. Bian, G.W. and Chen, H.H.: Cross language information access to multilingual collections on the Internet. Journal of American Society for Information Science, 51(3) (2000) 281-296.
8. Huang, C.R., *et al.*: Introduction to Academia Sinica Balanced Corpus. In: Proceedings of ROCLING VIII. Taiwan, (1995) 81-99.
9. Harman, D.K.: TREC-6 Proceedings. Gaithersburg, Maryland, (1997).
10. Mei, J., *et al.*: tong2yi4ci2ci2lin2. Shanghai Dictionary Press (1982).
11. Fellbaum, C. (ed.): WordNet: An Electronic Lexical Database. Cambridge, MA: MIT Press (1998).
12. Lin, C.J., Lin, W.C. and Chen, H.H.: Description of NTU QA and CLIR Systems in TREC-9. In: Proceedings of The Ninth Text REtrieval Conference (TREC 9). NIST Special Publication 500-249, Gaithersburg, Maryland, (2000) 389-398.