CMU in Cross-Language Information Retrieval at NTCIR-3

Yiming Yang and Nianli Ma
Language Technologies Institute
Carnegie Mellon University
<yiming,manianli>@cs.cmu.edu

Abstract

We participated in the Cross-Language Information Retrieval evaluation at NTCIR-3 for the English-Chinese and English-Japanese tasks. We examined several approaches to query translation, including the use of a commercial machine translation system, a thesaurus that is automatically extracted from a parallel corpus, and a general-purpose online dictionary. The MT-based approach was most effective among these alternatives in our experiments for English-Chinese retrieval on the NTCIR-2 and 3 data. Combined use of machine translation and thesaurus extraction yielded further improvement.

1. Introduction

Finding the most effective way to bridge the language barrier between queries and documents is the central challenge in Cross-Language Information Retrieval (CLIR). Many techniques have been investigated, including the use of machine translation systems, online bilingual dictionaries, automatically extracted bilingual thesauri from parallel corpora, statistical translation models, and vector-space models for the mapping from queries and documents to an "interlingua" [1, 2, 3, 4, 5]. The optimal method of choice for a specific CLIR task, however, does not only depend on the technical strength of a method, but also depends on the availability and quality of the knowledge sources or parallel corpora that a CLIR system can employ for the right domain.

At NTCIR-1 (1999), a large collection of English-Japanese citation records (339,624) of journal articles in the scientific domain is available. Human translation of the title, abstract and keywords for each article can be easily aligned to construct high-quality parallel text that would benefit corpus-based learning approaches, including Example-Based Thesaurus (EBT) extraction. The Berkeley research team, for example, aligned the bilingual keywords in corresponding articles and obtained a high-quality thesaurus for word translation in the scientific domain. This approach was most successful in the evaluation of English-Japanese retrieval at NTCIR-1 [4].

At NTCIR-2 (2001), the data for the English-Japanese retrieval tasks remained the same as those in NTCIR-1, while the data for English-Chinese retrieval tasks were news articles in the two languages, but not translations of each other. In other words, a high-quality parallel corpus was provided for the English-Japanese language pair, but not for the English-Chinese language pair at NTCIR-2.
As a result, an EBT-based method (by JUSTSYSTEM) had the best results in the evaluation of English-Japan retrieval [6], while a MT-based approach (by the Queens College) was more effective than the EBT-based methods for the English-Chinese tasks. It was also shown, by Queens College, that a combined use of MT and EBT obtained a better performance than using MT alone.

We participated in NTCIR-3 as our first experience in CLIR with Asian languages. Our strategy is to systematically examine the effectiveness of alternative approaches to query translation and expansion, including MT-based, EBT-based and Machine-Readable-Dictionary (MRD) based methods. By MT-based we mean to use a general-purposed (often rule-based) machine translation system; by EBT-based we mean to use a thesaurus or dictionary that is algorithmically extracted from a training corpus of parallel text; by MRD-based we mean to use an online-readable dictionary. We used the NTCIR-2 and NTCIR-3 data collections to conduct our experiments for those methods and the combined use of them.

2. Methods

The process consists of four steps:

- Pre-translation query expansion in the source language
- Query translation

- Post-translation query expansion in the target language
- Document retrieval using the translated and expanded query in the target language

2.1 Query expansion

Pseudo-relevance feedback (PRF) is a mechanism for query expansion. Originally developed for monolingual retrieval, it uses the initial query to retrieve a few top ranking documents, assumes those documents to be relevant (i.e., "pseudo-relevant"), and then uses them to expand the original query. Let $\vec{q}$ be the query (a vector of term weights) before the expansion, $\vec{q}'$ be the query after the expansion, $\vec{d}$ be a pseudo-relevant document, and k be the total number of pseudo-relevant documents, the computation is defined to be:

$$\vec{q}' = \vec{q} + \sum_{i=1}^{k} \vec{d}_i$$

In cross-lingual retrieval, PRF has been used to expand the query in both the source language (pre-translation) and the target language (post-translation), and is sometimes referred as *local feedback (LF)* in the literature [7].

For convenience, we use PTP (standing for PRF, Translation and PRF) as an abbreviation of this cross-lingual retrieval process, in order to avoid the confusion with another CLIR method that uses PRF without translation [1].

2.2 Query translation

We examined three alternative approaches:

- MT based
- MRD based
- EBT based

For MT we used the **TransWiz** software, a general-purpose machine translation system by a Taiwanese company (**http://www.otek.com.tw**), which is the same system used by Queens College at NTCIR-2.

For MRD we obtained the English-Chinese bilingual wordlist (**http://morph.ldc.upenn.edu/Projects/Chinese**) from the Linguistic Data Consortium (LDC), and used that list as a by-directional dictionary.

For ETB we used the parallel corpus of Hong Kong News Parallel Text (18,147 article pairs) provided by LDC (**http://www.ldc.upenn.edu/Catalog/LDC2000T46.html**). A sentence-level aligned corpus (95,740 sentence pairs) was further constructed by USC/ISI using the original corpus. We used the USC/ISI version of the parallel corpus as our primary training set for our EBT method. In addition, we also obtained a small parallel corpus by aligning the corresponding fields (Title, Question, Narrative and Concepts) in the English and Chinese descriptions for the 50 "topics" (queries) at NTCIR-2. We used this corpus as our secondary training set. Applying our EBT extraction algorithm to the two training sets, we obtained two bilingual thesauri and merged them.

The resulting bilingual thesaurus is a collection of "translation" lists, one listing per word in the source language. The derivation of those lists was based on how frequently (measured by counting the document frequency in a parallel corpus) a word co-occurs with the other words, and how far those frequencies away from the expectation by change. We computed the Chi-square statistic [8, 9] for each pair of cross-lingual lexical entries. By thresholding on those Chi-square values, a list of words in the target language was obtained for each word in the source language. We further imposed an additional threshold on any word on the list, i.e., the minimum frequency of documents for that word to co-occur with the corresponding word in the parallel corpus. The two thresholds are parameters of the EBT method and were empirically chosen through validation.

2.3 Combining two approaches

Given that there is more than one way to translate a query, it is natural to ask whether we can use those translations jointly for a better performance. Two obvious answers to this question are: 1) merging the different translations of a query before conducting the search for documents, or 2) using each translation of a query to conduct a search of documents, then merging the retrieved documents in individual searches. Queens College explored both ways in NTCIR-2, and we re-implemented those ideas using the following formula:

$$\vec{q}' = \alpha\vec{q}_1 + (1-\alpha)\vec{q}_2$$

$$s(q,d) = \beta s(q_1,d) + (1-\beta)s(q_2,d)$$

Where q is the original query, q1 and q2 are two translations of the query, q' is the merged, s(q1, d) is the score of document d when using q1 in the search, s(q2, d) is the score of document d when using q2 in the search, and s(q, d) is the merged score for document d with respect to query q. Alpha and beta are the parameters for adjusting the mixture weights for the components to be merged; their values were tuned using validation. Intuitively, if two translations of a query are equally good for cross-lingual retrieval, then their weights will also be equal in the merge; otherwise, the better approach will have a higher weight than the worse approach.

3. Experiments

We conducted cross-lingual retrieval experiments with English-to-Chinese and English-to-Japanese query translation. For comparison purposes, we also evaluated the performance of the same search engine in Chinese and Japanese monolingual retrieval.

3.1 System description and text processing

We used SMART system developed at Cornell [10] as our search engine, and a conventional TF*IDF scheme ("ltc" in the SMART nomenclature) for term weighting:

$$w_t = (\log TF_t) \log \frac{N}{DF_t}$$

Where t is a term and N is the total number of documents in the data collection.

For Chinese text processing we used an online Chinese dictionary provided by LDC (**http://www.ldc.upenn.edu/Projects/Chinese/segmenter/Mandarin.fre**), and an algorithm that favors the longest match of characters in word segmentation. We omit the details of this algorithm. After segmentation we removed the "stop words" using a statistic stoplist, generated by the following process: We sorted the lexical entries using their frequencies provided in the dictionary, took the top 100 words, removed the multi-character words and merged the word entrances which are the same word but appear in the forms of character and Pin-yin. This resulted in about 40 most common words. We then repeated this process to add the next 20 most common words to the list.

For Japanese text processing we used a Japanese morphological analyser named ChaSen, which is publicly available (**http://chasen.aist-nara.ac.jp/**). We obtained a statistical stop word list by sorting the words by frequency and threshold at 16,000. That is, any words having an equal or higher frequency are treated as a stop word.

We used the NTCIR-2 data collections for an initial evaluation and parameter tuning, and NTCIR-3 data for the final evaluation. Results represented in this paper are the scores on the NTCIR-3 evaluation collections unless specified otherwise.

For monolingual retrieval, we used pseudo-relevance feedback (PRF) for query expansion in the source language. For cross-lingual retrieval, we used PRF to expand the query before and after its translation.

3.2 Chinese monolingual baselines

Running our search engine over Chinese documents for the Chinese queries generated the monolingual baselines. Table 1 shows the results of our submitted runs using the 'rigid' and 'relax' relevance judgments. We submitted runs for the categories of D (short queries), TDC (medium-length queries) and TDNC (long queries); the performance was measured using the non-interpolated average precision. Figure 1 shows the recall-precision curves (interpolated at 11 grading points of recall) of PRF on the three versions of queries. Clearly, the system had a better performance with longer queries than it did with shorter queries. The relative performance of our system was around the middle among the submitted runs to the C-C evaluation in NTCIR-3.

	Rigid	Relax
Short Queries	0.1794	0.2569
Medium Queries	0.2333	0.3086
Long Queries	0.2667	0.3470

Table 1. Non-interpolated average precision of (baseline) monolingual PRF in Chinese retrieval

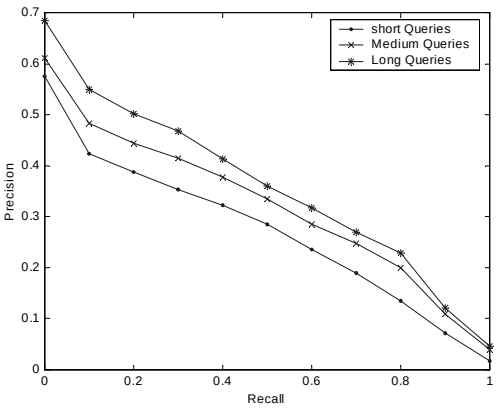

Figure 1. Recall-Precision curve of our monolingual PRF in Chinese retrieval (Relax assessment)

3.3 English-Chinese cross-lingual performance

The task was to retrieve Chinese documents using English queries. We conducted experiments for the PTP approach with the options of MT-based, EBT-based, MRD-based and MT+EBT-based for query translation, and the options of merging the queries or merging the document scores for the combination scheme (see Section 2.2 for details).

Table 2 shows the results of PTP variants in English-Chinese retrieval over NTCIR-3 evaluation set using medium-length queries. Figure 2 shows the recall-precision curves of those methods.

	Rigid	Relax
MT (TransWiz)	0.1540	0.2032
EBT	0.0852	0.1178
MT+EBT merging doc scores	0.1615	0.2223
MT+EBT merging queries	0.1682	0.2322

Table 2. PTP variants in English-Chinese retrieval on NTCIR-3 evaluation set using medium-length queries (The scores are non-interpolated average precisions)

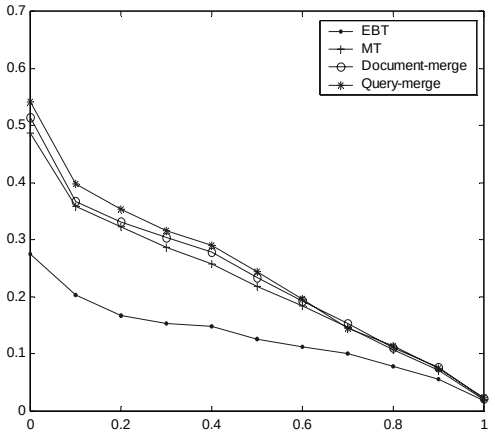

Figure 2. Recall-Precision curves of Cross-lingual PTP variants on medium-length queries (Relax assessment)

Table 3 shows the results of our methods when using long queries instead of the medium-length queries. Table 4 shows the corresponding results of those methods on the NICIR-2 evaluation set when using long queries.

	Rigid	Relax
MT (TransWiz)	0.1716	0.2238
EBT	0.0950	0.1309
MT+EBT merging doc scores	0.1782	0.2348
MT+EBT merging queries	0.1854	0.2488

Table 3. PTP variants in English-Chinese retrieval on NTCIR-3 evaluation set using long queries (the scores are non-interpolated average precisions)

	Rigid	Relax
MRD (LDC word lists)	0.2218	0.2974
MT (TransWiz)	0.3853	0.4573
EBT	0.3290	0.3428
MT+EBT merging doc scores	0.4087	0.4823
MT+EBT merging queries	0.4123	0.5030

Table 4. PTP variants in English-Chinese retrieval on NTCIR-2 evaluation set using long queries (the scores are non-interpolated average precisions)

Several observations can be made from those tables and figures, including:

- The relatively strong performance of the MT-based approach indicates that commercial MT systems are quite good for English-Chinese retrieval in the news domain.
- The EBT-based approach was significantly worse than the MT-based approach, suggesting either the parallel text (Hong Kong news and the topics in NTCIR-2 corpus) are not of sufficient quality from a statistical learning point of view, or our EBT extraction algorithm is sub-optimal.
- With parameters tuned on the validation sets (NTCIR-2 data and NTCIR-3 dry-run data) for alpha (0.67) and beta (0.80), the combination schemes did at least as well as the individual methods or significantly better. The query-merge scheme appears to be a better choice than the document-merge scheme.

Finally, we measured the relative performance of our cross-lingual approach (PTP using MT+EBT and query merge) in the English-Chinese task compared to the performance of monolingual PRF in the Chinese retrieval task. Table 3 shows results: the 11-point average precisions achieved in the cross-lingual task are more than 70% of those obtained in the monolingual task.

	Mono-lingual	Cross-lingual	Relative performance
Relax, MediumQ	0.3086	0.2322	75.2%
Relax, LongQ	0.3470	0.2488	71.7%
Rigid, MediumQ	0.2333	0.1682	72.1%
Rigid, LongQ	0.2667	0.1854	69.5%

Table 3. Cross-lingual retrieval verses monolingual retrieval: performance evaluated using non-interpolated average precision.

3.4 Japanese-related experiments

There was a complication in the evaluations of the English-Japanese and Japanese-Japanese tasks at NTCIR-3. Due to a copyright issue, the Japanese test set contains about 15,000 documents that have the title only but misses the main body of text in each article. The NTCIR organizers decided to discard those documents from evaluation after the submission deadline. As a result, the Japanese-related evaluation results do not accurately reflect the performance of many submitted systems. In particular, for the systems favoring short documents over long documents, their scores would be penalized more than the systems without such a preference. In other words, removing the 15,000 documents (titles only) from the relevance judgments but not removing them from the test set, some systems were "unfairly" penalized more than others in the evaluation. Our

system, unfortunately, was the most-affected system, according to the observation by the NTCIR-3 organizers. Nevertheless, we still present those evaluation results in this paper as reference. Note: Those results should not be interpreted as the representative performance of our system, and should not be used for comparison with any other systems evaluated in NTCIR-3.

Table 5 shows the results of our submitted runs for the Japanese monolingual task at NTCIR-3. We used the monolingual PRF method, the same one that we used for the Chinese monolingual retrieval task.

	Rigid	Relax
Short Queries	0.2218	0.2772
Medium Queries	0.2803	0.3523
Long Queries	0.2992	0.3773

Table 5. Japanese retrieval results (in 11-pt average precision) of our monolingual PRF method

For the English-Japanese cross-lingual task we used the MRD-based PTP approach. We currently do not have access to a Japanese MT system, and we do not have a sufficiently large English-Japanese parallel corpus in the news-stories domain.
Thus we cannot test the MT-based and EBT-based options as we did for the English-Chinese task. For the Machine-Readable Dictionary (MRD) we used EDICT (**http://avenue.tutics.tut.ac.jp/pubdict/edict.html**).
Table 6 shows the cross-lingual scores of our submitted runs in the NTCIR-3 evaluation.

	Rigid	Relax
Short Queries	0.0510	0.0552
Medium Queries	0.1438	0.1693

Table 6. English-Japanese cross-lingual results

4. Concluding Remarks

We examined MT-based, EBT-based and MRD-based query translation for CLIR tasks at NTCIR-3. Our experiments in the English-Chinese retrieval task confirmed the observations by Queens College at NTCIR-2, i.e., general-purpose MT systems work well for retrieving Chinese news stories using English queries, while a joint use of MT and EBT can improve the performance even further. Those observations, however, do not necessarily mean that MT is the optimal choice for other language pairs or in other domains. As is evident in the evaluations of English-Japanese retrieval at NTCIR-1 and NTCIR-2, the best performing systems were both EBT-based. An important question for future research is: Is there a principled way for automatic identification of the optimal method, and the knowledge source (including parallel corpus) to use within a CLIR task in a new language pair and new domain? Good answers for this question would have both theoretical and practical values, given that more and more MT-systems and parallel text are becoming available in the Internet for different languages and in many domains.

References

[1] Yang, Y., Carbonell, J. G., Brown, R. and Frederking, R. E.: Translingual Information Retrieval: Learning from Bilingual Corpora. Artificial Intelligence Journal special issue: Best of IJCAI-97, 1998, pp323--345.
[2] M. Franz, J.S. McCarley, S. Roukos (IBM T.J. Watson Research Center): Ad hoc and Multilingual Information Retrieval at IBM, page 157-168
[3] J. M. Ponte, W. Bruce Croft: A Language Modeling Approach to Information Retrieval. SIGIR 1998: pp275-281
[4] A. Chen, F. C. Gey, K. Kishida, H. Jiang, and Q. Liang: Comparing multiple methods for Japanese and Japanese-English text retrieval. Proceeding of the First NTCIR Workshop on Research in Japanese Text Retrieval and Term Recognition pp49-57
[5] K. L. Kwok: NTCIR-2 Chinese, Cross Language Retrieval Experiments Using PIRCS. Proceeding of the Second NTCIR Workshop on Research in Chinese & Japanese Text Retrieval and Text Summarization pp97-104
[6] Sumio FUJITA: Notes on the Limits of CLIR Effectiveness NTCIR-2 Evaluation Experiments at Justsystem. Proceeding of the Second NTCIR Workshop on Research in Chinese & Japanese Text Retrieval and Text Summarization pp181-188
[7] L. Ballesteros, W. Bruce Croft: Phrasal Translation and Query Expansion Techniques for Cross-language Information Retrieval. SIGIR 1997: pp84 – 91.
[8] Yang, Y., Pedersen, and J.P.: A Comparative Study on Feature Selection in Text Categorization. Proceeding of the Fourteenth International Conference on Machine Learning (ICML'97), 1997, pp412-420.
[9] M. Rogati, Y. Yang: CMU PRF using a Comparable Corpus. Working Notes for the CLEF 2001 Workshop pp81-86
[10] C. Buckley: SMART version 11.0, 1992. **ftp://ftp.cs.cornell.edu/pub/smart**