ISCAS in CLIR at NTCIR-6: Experiments with MT and PRF

Ruihong Huang，Le Sun，Jing Li，Longxi Pan，Junlin Zhang
Institute of Software, Chinese Academy of Sciences
P.O.Box 8718, Beijing, 100080, P.R.China
ruihong05@iscas.cn

Abstract

We participated in the English-Chinese cross-language information retrieval (CLIR) E-C tasks in NTCIR6. Considering the special feature of crossing two different languages in CLIR, our main concerns in our experiment are

1) to evaluate the appropriateness of MT as a means of query translation in CLIR, 2) to evaluate the effect of feedback in retrieval model to the performance of CLIR which has been discussed in some papers. *Besides, we 1) applied Chinese word segmenter with quite high precision to ensure an exact indexing, 2) applied language model with relevance feedback as our retrieval model.*

Keyword: CLIR, Query Translation, Language Model

1 Introduction

In Cross-Language Information Retrieval (CLIR), the candidate documents and the queries are in different languages. Properly translating the queries into the target language used by the candidate documents is necessary and the translation quality may affect the retrieval results greatly. In our E-C CLIR practices, we experimented with two straightforward approaches: the dictionary-based and MT-based translation. Our experiments shows that the MT-based query translation has achieved better results when high performance MT-based translation is applied in CLIR.

The remainder of this paper is organized as follows: Firstly, in section 2, we will describe our considerations in the system design and give an outline of our system, then, in section 3, we will discuss some meaningful results. Finally, we conclude our paper in section 4.

2 Designing and Implementing of our CLIR System

2.1 Document and query indexing

The detailed description of the document set and the query set can be found in [1].

For most of oriental languages, such as Chinese and Japanese, deciding the indexing unit usually isn't so straightforward. As to Chinese, overlapping character n-grams, multi-word phrases and simple words are the main selections. Through the analysis of former experiments, we observed that adopting multi-word phrases as the indexing unit take the risk that the indexing scale is too large to deal with, while overlapping character n-grams may make the system lose some stability in it' s retrieval performance, so we choose the third approach.

In our experiment, we employed the ICTCLAS to segment the documents into the simple words for indexing. ICTCLAS is a free high-performing Chinese word segmenter. It is based on the cascaded Hidden Markov Model

With aim to integrating Chinese word segmenting, POS tagging, disambiguation and OOV recognizing into a complete theoretical framework. The segmenting procedure is mainly comprised of four stages. In stage one, sentence-boundary recognizing and initial word segmenting were done using the category based HMM disregarding the OOV which was dealt with by employing a role based HMM in stage three. For disambiguation, the best N candidates were recalled based on the N-shortest path algorithms in stage two, while the final segmenting sequence wasn't selected until after the OOV recognizing in stage three and POS tagging in stage four.

2.2 Query Translation

In Cross-Language Information Retrieval (CLIR), the final retrieval performance relies on the query translation quality to a large scale. Up to the date, various translation techniques have been proposed. Most of the translation techniques belong to one of the three kinds, machine readable bilingual dictionary based translation [2], MT based translation [3], [4], and parallel corpus statistics based translation [5], [6].

Although some researchers have explored some novel methods in the direction of corpus statistics based translating, and attained good results, the common conditions in practice refrains the use of it for lacking the well-evaluated large corpus. So we think focusing more care on the straightforward approaches is reasonable. Out of the practical consideration, we established our experiment on the other two methods, the dictionary-based translation, and the MT based translation.

We will give some analysis separately here. the dictionary based translating is quite simple, and easy to use, on the other hand, it does little effort in dealing with the two main issues related to the query translation, OOV and disambiguation, which may affect the resulting performance. Machine Translating is another straightforward method applied to the query translation, especially at the time when the MT has attained reasonable translation performance in many language pairs. Furthermore, there are already several successful machine translation systems accessed, the Google provide high performance translation between many language pairs through its website, for example. However, the appropriation of applying MT to query translation in CLIR is questionable, because the goal of MT is to transfer the same meaning in one language different from the original language whose output is continuous and human-understandable while the query translating results in CLIR will be handled in an automatic and discrete way. So it's quite necessary to examine the performance of MT in CLIR.

Out of the analysis above, we examined the performance of Machine Translating as a CLIR query translating approach taking the dictionary based translation as the baseline in our experiments. Specifically, we employed the Google translator in our Machine Translating approach.

2.3 Retrieval Model

There are many researches focusing on the retrieval models, Three of the major retrieval models, Vector Space Model, statistical retrieval model and Language Model [7], have all been studied extensively.

In the former two tracks, the NTCIR4 [9] and the NTCIR5 [8], we had experimented with all kinds of IR models, from the traditional TFIDF model to language model (KL-divergence) with feedback and smoothing. Through the analysis of the former experiments, we have come to the conclusion [8] that language models can attained better retrieval results in Chinese IR and the retrieval performance usually improved when adding feedback and smoothing in the language models. Other researchers [10] had declared that by using pseudo relevance feedback, they can significantly improve cross-language retrieval performance and achieve the level of monolingual retrieval.

So in our experiments, we applied two Language models(KL-divergence), one with pseudo relevance feedback, the other not., both of which is combined with the proposed machine readable dictionary based query translation results and the Machine Translating based query translation results.

3 Experiment Results evaluation

Regretfully, for the lack of communications and coordination among the members of our group, there are some disagreements on the requirements of the first stage and the second stage of the last CLIR track, so in the final submitted experiment results on the second stage, we had used the document collection required on the first test stage, which cause our query results difficult to be evaluated together with other participators', so we had no other choice but to do some limited evaluation manually. Considering the mass workload of comprehensive evaluation, we just did some necessary comparisons referring to the objectives we designed to achieve in the beginning. In the following paragraphs, we'll present our methods in evaluating the results manually and the meaningful conclusions.

As discussed above in 2.2, the focus of our experiments was to evaluate the appropriateness of the Machine Translating as a query translation in CLIR and the effect of pseudo relevance feedback to the Machine Translating query translation. So we concentrated more attentions on the comparisons of the three pair of runs, MT-KL to MT-KL-FB, DT-KL to DT-KL-FB, the better of the former pair to the better of later. Specifically, MT-KL-FB means a combination of Machine Translating and KL-distance retrieval language model with feedback and DT means machine readable dictionary based query translation.

Through our less exact comparison between MT-KL and MT-KL-FB, we noticed that the MT-KL-FB outperformed the former greatly. The same thing was true for DT-KL to DT-KL-FB. Then, when we compared the MT-KL-FB with DT-KL-FB, it was obvious that the first 10 retrieval documents are more relevant to the query on average in MT-KL-FB than in DT-KL-FB.

Besides this, we had also done some monolingual IR experiments on Chinese using the different parts in the query set, the TITLE, the DESC and the DESC+NARR. The run with the DESC attained better results in the Chinese-Chinese monolingual experiments. When we tried to compare the results in MT-KL-FB using the DESC with the C-C monolingual run using the DESC, the first ten retrieval documents in former are somewhat less relevant than those in later on average, while some documents in former are as relevant as the according ones in later sometimes.

Through the above discussion, we can come to the conclusions that the MT-based query translation has achieved reasonable results when high performance MT-based translation is applicable in CLIR. But In our English-Chinese Cross Language IR, even when we included the pseudo relevance feedback in our retrieval model, the retrieval performance of our system was lower than that of the Chinese-Chinese monolingual IR system, which disagree with the declaration in [10], whose experiments are done on English-to-France CLIR systems, that by using pseudo relevance feedback, cross-language retrieval performance improves significantly and can gets to the level of monolingual retrieval.

4 Conclusions and Future Work

In this paper, we mainly described the design and the implementation of our English-Chinese Cross Language IR system. We focus our attentions on the issue in CLIR query translation. The two emphasizes in our experiments are 1) to evaluate the appropriateness of MT as a means of query translation in CLIR, 2) to evaluate the effect of feedback in retrieval model to the performance of MT.

In the document and query Indexing phrase, we selected simple words as the indexing unit and employed a high-performance Chinese word segmenter to the documents before indexing.

In the retrieval phrase, based on our former experiments, we chose the Language Model (KL-distance) directly as our retrieval model. In order to evaluate the effect of feedback in retrieval model to the performance of MT, we examined the two categories of Language Model (KL-distance), one with feedback, the other without.

Through our experiments, we came to two conclusions. One is that the MT-based query translation has achieved reasonable results when high performance MT-based translation is applicable in CLIR. The other is that in English-to-Chinese Cross Language IR, the use of pseudo relevance feedback hasn't promoted the performance of our CLIR system to the same level as our Chinese-Chinese monolingual IR system.

As we know, OOV and disambiguation are the major challenges in CLIR, however, the dictionary based query translation is too simple to conquer any of the challenges, while the Machine Translating based query translation may be powerful enough to deal with both of the two issues, it is a black box to us and uncontrolled, and as discussed in 2.2, its goal isn't the same as ours. In the future, we plan to contribute more efforts into developing high-performance and easy-to-use techniques to discard the bad influences caused by OOV and disambiguation.

5 Acknowledgments

Our work was supported by the Beijing New Star Plan of Technology&Science Fund of China under Grant No. H020820790130 and the National Natural Science Foundation of China under Grant No.. 60203007.

References

[1] K. Kishida, K. -H. Chen, S. Lee, K. Kuriyama, N. Kando, H. -H. Chen, and S. -H. Myaeng. Overview of CLIR task at the sixth NTCIR workshop. In Proceedings of the Sixth NTCIR Workshop.

[2] Hull, D.A., & Grefenstette, G. (1996). Querying across languages: a dictionary-based approach to multilingual information retrieval. In *Proceedings of the Nineteenth Annual International ACM SIGIR Conference on Research and Development in Information Retrieval* (pp. 49-57).

[3] Gachot, D.A., Lange, E., & Yang, J. (1998). The SYSTRAN NLP browser: An application of machine translation technology in cross-language information retrieval. In G. Grefenstette (Ed.), *Cross-Language Information Retrieval* (Chapter 9). Boston, MA: Kluwer Academic Publishers.

[4]Gey, F.C., Jiang, H., Chen, A., & Larson, R.R. (1999). Manual queries and machine translation in cross-language retrieval and interactive retrieval with Cheshire II at TREC-7. In E.M. Voorhees and D.K. Harman (Eds.), *Information Technology: The Seventh Text REtrieval Conference (TREC-7)*, (NIST Special Publication 500-242, pp. 527-540). Washington, DC: U.S. Government Printing Office.

[5] Sheridan, P. & Ballerini, J.P. (1996). Experiments in multilingual information retrieval using the spider system. In *Proceedings of the Nineteenth Annual International ACM SIGIR Conference on Research and Development in Information Retrieval* (pp. 58-65).

[6] Carbonell, J., Yang, Y., Frederking, R., Brown, R.D., Geng. Y., & Lee, D. (1997). Translingual information retrieval: A comparative evaluation. In *Proceedings of the Fifteenth International Joint Conference on Artificial Intelligence* (pp. 708—714).

[7] A.Berger and J.Lafferty. Information retrieval as statistical translation. In Proceedings of the 1999 ACM SIGIR Conference on Research and Development in Information Retrieval. Pages 222-229, 1999.

[8] Jinming Min, Le Sun, Junlin Zhang. ISCAS in English-Chinese CLIR at NTCIR-5. In *Proceedings of the Fifth NTCIR Workshop*.
[9] Z. Junlin, S. Le, Z. Yongchen and S. Yutang. Applying language model into IR task. In *Working Notes of the Fourth NTCIR Workshop Meeting, Cross-Lingual Information Retrieval Task*, 2004.
[10] Qu, Y., A.N. Eilerman, H. Jin, and D.A. Evans. The Effect of Pseudorelevance Feedback on MT-Based CLIR. In Proceedings of the Recherche d'Informations Assistée par Ordinateur (RIAO 2000) (2000)