Evaluating Language Resources for English-Indonesian CLIR

Herika Hayurani, Syandra Sari, and Mirna Adriani

Faculty of Computer Science
University of Indonesia
Depok 16424, Indonesia
{heha51, sysa51}@cs.ui.ac.id, mirna@cs.ui.ac.id

Abstract. We present a report on our participation in the Indonesian-English ad hoc bilingual task of the 2006 Cross-Language Evaluation Forum (CLEF). This year we compare the use of several language resources to translate Indonesian queries into English. We used several readable machine dictionaries to perform the translation. We also used two machine translation techniques to translate the Indonesian queries. In addition to translating an Indonesian query set into English, we also translated English documents into Indonesian using the machine readable dictionaries and a commercial machine translation tool. The results show performing the task by translating the queries is better than translating the documents. Combining several dictionaries produced better result than only using one dictionary. However, the query expansion that we applied to the translated queries using the dictionaries reduced the retrieval effectiveness of the queries.

Keywords: cross-language information retrieval, machine translation, dictionary translation, parallel corpus.

1. Introduction

This year we participate in the bilingual 2006 Cross Language Evaluation Forum (CLEF) ad hoc task, i.e., the English-Indonesian CLIR. As stated in previous work [8], a translation step must be done either to the documents [9] or to the queries [3, 4, 6] in order to overcome the language barrier. The translation can be done using bilingual readable machine dictionaries [1, 2], machine translation techniques [7], or parallel corpora [11]. We used a commercial machine translation software package called *Transtool*[1] and an online machine translation system available on the Internet to translate an Indonesian query set into English and to translate English documents into Indonesian. We learned from our previous work [1, 2] that freely available dictionaries on the Internet could not correctly translate many Indonesian terms, as their vocabulary was very limited. We hoped that using machine translation techniques and parallel documents could improve our result this time.

2 The Translation Process

In our participation, we translate English queries and documents using dictionaries and machine translation techniques. We manually translated the original CLEF query set from English into Indonesian. We then translated the resulting Indonesian queries back into English using dictionaries and machine translation techniques. The dictionaries come from several sources, particularly, *the Indonesian National Research and Technology Office* (*Badan Pengkajian dan Penelitian Teknologi* or BPPT) and online dictionaries on the internet, namely http://www.orisinil.com/kamus.php and http://www.kamus.net/main.php. The machine translation systems that we use are *Toggletext* (www.toggletext.com) and *Transtool*, a commercial software package.

Besides translating the queries, we also translate the English documents from CLEF into Indonesian. The translation process is done using the dictionary from BPPT and *Transtool.* The translation process using the dictionary is done by taking only the first definition for each English word in the document.

2.1 Query Expansion Technique

[1] See "http://www.geocities.com/cdpenerjemah/".

Adding translated queries with relevant terms, known as query expansion, has been shown to improve CLIR effectiveness [1, 3, 12]. One of the query expansion techniques is called the *pseudo relevance feedback* [4, 5]. This technique is based on an assumption that the top few documents initially retrieved are indeed relevant to the query, and so they must contain other terms that are also relevant to the query. The query expansion technique adds such terms into the previous query. We apply this technique to the queries in this work. To choose the relevant terms from the top ranked documents we employ the *tf*idf* term weighting formula [10]. We added a certain number of terms that have the highest weight scores.

3 Experiment

In the experiments, we used *Lemur*[2] information retrieval system which is based on the *language model* to index and retrieve the documents.

We then applied a pseudo relevance-feedback query-expansion technique to the queries that were translated using the machine translation tool. We used the top 10 relevant documents retrieved for a query from the collection to extract the expansion terms. The terms were then added to the original query.

4 Results

The results of our CLIR experiments were obtained by applying the three methods of translation, i.e., using the machine translations, using dictionaries, and using parallel corpus. Table 1 shows the result of the first technique, which shows that translating the English queries into Indonesian using *Toggletext* machine translation tool is better than using *Transtool*.

Table 1. Average retrieval precision in CLIR runs of queries translated using the machine translation tools for title only and combination of title and description.

Task	MT-1 (Toggletext)	MT-2 (Transtool)
Title	0.2538	0.2137
Title + Description	0.2934	0.2397

The result of the second technique, which is translating the English queries into Indonesian using dictionaries, is shown in Table 2. The result shows that using the dictionary from BPPT alone is not as good as using several dictionaries combined.

The last technique that we use is retrieving documents from parallel corpus created by translating the English document collection into Indonesian using the combined dictionaries and *Transtool*. The results, as shown in Table 3, indicate that of retrieving the English version of Indonesian documents that are relevant to an Indonesian query, is much more effective if the parallel corpus is created using the machine translation tool than using the dictionaries.

Table 2. Average retrieval precision in CLIR runs of queries translated using dictionaries for title only and combination of title and description.

Task	Dic-1	Dic-2
Title	0.1423	0.2063
Title + Description	0.1101	0.2650

[2] See “http://www.lemurproject.org/”.

Table 3. Average retrieval precision in the monolingual runs for title only and combination of title and description on a parallel corpus created by translating English documents into Indonesian using dictionaries and a machine translation tool.

Task	Parallel-DIC	Parallel-MT
Title	0.0201	0.2060
Title + Description	0.0271	0.2515

Lastly, we also attempted to improve our CLIR results by expanding the queries translated using the dictionaries. We were unable to do the expansion process to all the translated queries because of time limitation. The results is as shown in Table 4, which indicates that adding the queries with 5 terms from the top-10 documents obtained from a pseudo relevance feedback technique hurt the retrieval performance of the translated queries.

Table 4. Average retrieval precision for the title only and combination of title and description using the query expansion technique with top-10 document method.

Task	Dic-2	Query Expansion (5 terms added)
Title	0.2063	0.1205
Title + Description	0.2650	0.1829

5 Summary

Our experiments demonstrate that translating queries using machine translation tools is better than translating documents. The retrieval performance of queries that were translated using machine translation tools for Bahasa Indonesia was about 14.28%-18.83% of that of retrieving the documents translated using machine translation. There was no significant difference in retrieval performance between the two machine translation tools that we used.

Taking the first definition in the dictionary when translating an English query into Indonesian appeared to be effective. The result of combining several dictionaries is much better than only using one dictionary.

In order to improve the retrieval performance of the translated queries, we expanded the queries with the terms extracted from the top-10 documents. However, the pseudo relevance feedback technique that is known to improve the retrieval performance did not improve the retrieval performance of our queries.

References

1. Adriani, M. and C.J. van Rijsbergen. Term Similarity Based Query Expansion for Cross Language Information Retrieval. In *Proceedings of Research and Advanced Technology for Digital Libraries*, Third European Conference (ECDL'99), p. 311-322. Springer Verlag: Paris, September 1999.
2. Adriani, M. Ambiguity Problem in Multilingual Information Retrieval. In *CLEF 2000 Working Note Workshop*. Portugal, September 2000.
3. Adriani, M. English-Ducth CLIR Using Query Translation Techniques. In *Proceedings of the Workshop of the Cross-Language Evaluation Forum (CLEF 2001)*, Lecture Notes in Computer Science. Darmstadt, Germany, September 2001.

4. Ballesteros, L, and Croft, W. Bruce. (1998). Resolving Ambiguity for Cross-language Retrieval. In *Proceedings of the 21st International ACM SIGIR Conference on Research and Development in Information Retrieval* (pp.64-71).
5. Attar, R. and Fraenkel, A. S. *Local Feedback in Full-Text Retrieval Systems*. Journal of the Association for Computing Machinery, 24: 397-417, 1977.
6. Davis, Mark and Dunning, Ted. *A TREC Evaluation of Query Translation Methods for Multi-Lingual Text Retrieval*. Ed. D. K. Harman. The Fourth Text Retrieval Conference (TREC-4), November. NIST, 1995.
7. Jones, Gareth and Lam-Adesina, Adenike M. Exeter at CLEF 2001: Experiments with Machine Translation for Bilingual Retrieval. In *Proceedings of the Workshop of the Cross-Language Evaluation Forum (CLEF 2001)*, p. 105-114. Darmstadt, Germany, September 2001.
8. McCarley, J. Scott. Should We Translate the Documents or the Queries in Cross-Language Information Retrieval?. In Proceedings of the *Association for Computational Linguistics (ACL'99)*, p. 208-214, 1999.
9. Oard, D. W. and Hackett, P. G. Document Translation for Cross-Language Text Retrieval at the University of Maryland. In Proceedings of *the Sixth Text Retrieval Conference (TREC-6)*, Gaithersburg MD, November 1997.
10. Salton, Gerard, and McGill, Michael J. *Introduction to Modern Information Retrieval*, New York: McGraw-Hill, 1983.
11. Sheridan, P. and Ballerini, J. P. Experiments in Multilingual Information Retrieval using the SPIDER System. *Proceedings of the 19th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, Zürich, Switzerland, August 1996.
12. Xu, Jinxi and Croft, W. Bruce. Query expansion using local and global document analysis. In *Proceedings in the 19th Annual International ACM SIGIR Conference on Research and development in Information Retrieval*. 1996.