The Effectiveness of a Dictionary-Based Technique for Indonesian-English Cross-Language Text Retrieval

Mirna Adriani
Fakultas Ilmu Komputer
Universitas Indonesia
Depok 16424
Indonesia
mirna@cs.ui.ac.id

W. Bruce Croft
Center for Intelligent Information Retrieval
Computer Science Department
University of Massachusetts, Amherst
Amherst, MA 01003-4610, USA
croft@cs.umass.edu

May 1997

Abstract

We evaluate the effectiveness of a dictionary-based cross-language text retrieval technique which uses a two-way dictionary for translating queries from their original language into the language of the text documents. As can be expected, the translated queries are not as effective as queries formulated by the users using the same language as the text documents. We then apply a local-feedback technique to expand the translated queries in order to improve their retrieval effectiveness. Our empirical results show that the technique is effective for English-Indonesian and Indonesian-English cross-language retrieval.

1. Introduction

Internet which has connected computer users all over the world has become a major medium for textual information exchange involving users located in various parts of the world. As a result, we see a dramatic increase in the availability and accessibility of text documents in various languages. It opens the possibility for a user to conduct research using published materials or references written in other languages. For instance, a traveller can obtain information about countries that he/she wants to visit.

The problem is that, as with the Internet in general, finding or selecting relevant texts among all that is available online is difficult, let alone if the texts are written in foreign languages. Since, obviously, not many users are fluent in foreign languages to be able find foreign text documents relevant to their information needs, an information retrieval system that takes a query written in the user's native language and retrieves documents written in another language would be very helpful. Such a system is called a cross-language information retrieval (CLIR) system. This situation has motivated the recent increasing interest in CLIR which is focused toward overcoming the above language barriers.

Dictionary-based CLIR techniques are retrieval techniques which involve translations of textual data from one language to the other using a dictionary. There are two main strategies in dictionary-based CLIR, first, by translating the original documents into the language of the queries, and second, by translating the queries into the language of the documents.

In this study, we employ the second strategy, i.e., the query translation strategy. This strategy is more efficient than the first strategy, i.e., the document translation strategy, in that it does not require the expensive overhead cost of translating all documents, especially when new documents are added frequently and not all of the documents are of interest to the users.

We evaluate this strategy on English-Indonesian and Indonesian-English cross-language text retrievals. Our preliminary results, as can be expected, showed that the retrieval effectiveness of the translated queries is lower than that of queries written in the same language as the documents (mono-language retrieval). To improve the queries recall-precision performance, we apply a query expansion techniques which uses local-feedback technique [Attar & Fraenkel, 1977]. A query is expanded by adding potentially relevant keywords to it in order to make it more specific.

2. Other CLIR Research

One of the earlier work in CLIR is by Salton [Salton, 1970]. In his study, the effectiveness of English and German queries is compared with that of queries obtained using a bilingual thesaurus for retrieving documents in both languages.

Much work in CLIR has been done by researchers recently. One of the well-know methods in CLIR is the parallel corpus method. This method requires the availability of documents in both languages, hence the name parallel corpus, which may be obtained using manual or machine translation, or by using comparable documents of the same topic. Landauer and Littman [Landauer & Littman, 1990] apply Cross-Language Latent Semantic Indexing (CL-LSI) to parallel document collections. This technique transforms the correlation among words into a compact vector representation. In this technique, a query (or a document) does not need to be translated into the other language, instead its vector representation is simply compared with the documents' vectors.

A study using parallel corpus has also been conducted by Sheridan and Ballerini [Sheridan & Ballerini, 1996]. They showed that adding terms obtained from thesaurus to the query results in better retrieval performance for multilingual documents.

Another well-known method, besides the parallel corpus method, is the dictionary-based method. There have been many studies which propose dictionary-based CLIR techniques, particularly those that employ the query translation strategy. The effectiveness of translating a query word-by-word based by looking up in a dictionary has been studied by Hull and Grefenstette [Hull & Grefenstette, 1996]. They identified the cause of poor performance of the translated queries such as word ambiguities and the difficulty in recognising phrases.
Ballesteros and Croft [Ballesteros & Croft, 1996] proposes a technique to improve the retrieval effectiveness of the translated query by applying a query expansion technique based on the local feedback technique. They demonstrated that adding terms obtained using the local feedback to the query at pre-translation, at post-translation and at pre-translation and post-translation combined, results in better retrieval performance than the base retrieval performance of the query without expansion.

Davis and Ogden [Davis & Ogden, 1997] combined the parallel corpus method and the dictionary-based method. They used a bilingual dictionary to translate the English queries to Spanish for retrieving parallel documents. To reduce ambiguity, they used a statistically-based parts-of-speech tagger to identify the syntactic roles of the terms in the query before translating them. The retrieval is done by comparing the vector representation of the query and the vector representations of the parallel documents.

3. Indonesian-English CLIR study

In this study, we apply a similar technique as the one used by Ballesteros and Croft [Ballesteros & Croft, 1996] for Indonesian-English query. The technique uses the local feedback to expand the queries. The local feedback technique [Attar & Fraenkel, 1977] extracts terms from a number of top documents retrieved by the original query based on a presumption that those documents are relevant. The terms are then added to the query resulting in an expanded query. The query expansion can be performed before, after, or both before and after the query translation. In this study we compare all of the three schemes.

We construct the Indonesian queries based on TREC's Spanish query topics SP26-45 [Harman, 1995] by translating the queries to Indonesian and modifying them to make them relevant with Indonesia's national affairs. The English queries are chosen from TREC topics 151-171. To obtain the terms for the pre-translation query expansion, the local feedback process is performed on the Associated Press collection from TREC (78,321 articles) and on Tempo collections (5,601 articles from Tempo, Indonesian weekly magazine) for, respectively, English-Indonesian and Indonesian-English retrievals. The retrieval performance evaluation and the post-translation query expansion are conducted on the 2GB TREC (vol. 2) English collection and the 68 MB newspaper article collection from Kompas (an Indonesian daily newspaper), respectively, for Indonesian-English and English-Indonesian retrievals.

In the first step of the query translation process, we delete stop phrases, stop words, and translate plural words into their singular forms. Then each word in the queries is replaced with the first definition listed in a two-way Indonesian-English dictionary [Echols & Shaddily, 1992]. Words which do not have any translation are kept in their original form in the query.
Our experiment is conducted using INQUERY information retrieval system. INQUERY is a probabilistic information retrieval system developed at the University of Massachusetts [Callan et.al, 1992]. In order to handle the Indonesian texts, we added an Indonesian word-stemmer module into INQUERY.

3.1 Translation Process

The first step in this study is to do the translation for each word in the queries. Each word is replaced by the first definition found in the dictionary. Note that a dictionary entry may contain a number of terms that do not have the same meaning as the original word such as shown in Table 1.

English word	**Translated Indonesian word from the dictionary**
culture	kesopanan, kebudayaan
control	pengawasan, pengawas, penilikan, pengaturan, penguasaan, pembatasan

Table 1. The Examples of some translated words

The translated queries for both Indonesian to English and English to Indonesian cases result in lower retrieval effectiveness as compared with that of the original language's queries. The terms from the dictionary that are not related to the original query term might have caused this poor performance.

The results (Table 2.) show that the performance of both query translations is below that of the monolingual retrieval. The English to Indonesian query translation shows 33.2% performance drop compared to the original queries. Meanwhile, the Indonesian to English query translation shows greater performance decrease, that is a drop of 62.1%. The cause of such a big drop may be due to the fact that English vocabulary is much larger, consisting of more specific concept words, than Indonesian vocabulary. As a result, mapping words from Indonesian to English is more difficult or less accurate than from English to Indonesia. Moreover, the translation words might not relate to the original query.

	Monolingual Query	Translated Query
English	0.1122	0.0425 (-62.1)
Indonesian	0.1780	0.1163 (-33.2)

Table 2. The average recall-precision from the original queries and the translated queries.

In order to improve the performance of the translated query, we modify the query using the local feedback technique. The modification is done before and after the query translation to add more terms.

3.2 Query Expansion Pre-Translation

The query expansion that is done before the translation process add more terms to the Indonesian and English queries (pre-translation feedback). Furthermore, the new queries are translated word by word to the other language. Table 3. and 4. show some examples of the resulting queries in the pre-translation query expansions for Indonesian and English queries.

English word	Translated Indonesian word from the dictionary
English query	heritage culture indonesia
Translation of English query	[warisan, pusaka] [kesopanan, kebudayaan] indonesia
English query & pre-translation feedback	heritage culture indonesia [indonesian jakarta sumatra jaya archipelago]
Translation of the English query & pre-translation feedback	warisan, pusaka kesopanan, kebudayaan indonesia [orang Indonesia jakarta sumatra jaya nusantara]

Table 3. The Examples of the translated queries using pre-translation feedback for English queries

Indonesian word	**Translated English word from the dictionary**
Indonesian query	tumpahan minyak buruk
Translation of Indonesian query	[something spilled] [oil, grease, fat] [old, worn out, dilapidated]
Indonesian query & pre-translation feed-back	tumpahan minyak buruk translation [menteri barel opec ekonomi kurs]
Translation of the Indonesian query & pre-translation feedback	[something spilled] [oil, grease, fat] [old, worn out, dilapidated] [cabinet minister] barel opec [economics] [a rate of exchange]

Table 4. The Examples of the translated query using pre-translation feedback for Indonesian queries

The results show that the pre-translation feedback for Indonesian to English query translation does not add useful terms to the queries so that the performance is 17-21% lower than the base result (Table 5). The reason is that the Associated Press collection that we use for the training set does not contain enough documents concerning Indonesia. The terms that we get from the local feedback for most of the queries come from the same small set of documents.

No. of terms	0	20	30	40	50
Relevant docs	0	5	5	5	5
Avg. Precision	0.1163	0.0956 (-17.8)	0.0910 (-21.8)	0.0937 (-19.5)	0.0962 (-17.3)

Table 5. The average precision of pre-translation feedback result for English to Indonesian translated queries

On the other hand, the Indonesian to English query translation shows better result (Table 6). The feedback terms help to increase the retrieval performance by 10-15% as compared to the base result.

No. of terms	0	10	20	30
Relevant docs	0	5	5	5
Avg. Precision	0.0425	0.0469 (10.4)	0.0469 (10.4)	0.0489 (15.2)

Table 6. The average precision of pre-translation feedback result for the Indonesian to English translated queries

3.3 Query Expansion Post-Translation Feedback

The query expansion that is done after the translation showed greater improvement than the base query. The effect of adding terms through the local feedback is positive because the documents that are used for the feedback is the same as the documents used for the evaluation (Table 7).

English word	**Translated Indonesian word from the dictionary**
English query	potensial weakness Indonesian navy
Translation of English query	[kesanggupan, tenaga] [kelemahan, kekurangan] [orang Indonesia] [Angkatan Laut]
English query & post-translation feedback	[kesanggupan, tenaga] [kelemahan, kekurangan] [orang Indonesia] [Angkatan Laut] [indonesia laut kompas organisasi negeri luas keamanan]
Translation of the English query & post-translation feedback	[kesanggupan, tenaga] [kelemahan, kekurangan] [orang Indonesia] [Angkatan Laut] [indonesia laut kompas organisasi negeri luas keamanan]

Table 7. The example of the English to Indonesian translated query using post-translation feedback.

The retrieval performance of the English to Indonesian query translation as shown in Table 9. improves by 10-15%. Likewise, the retrieval performance of Indonesian to English query translation improves by 46-68%. The better result in the last experiment is due to that more relevant words from a larger set of relevant documents being added to the queries .

No. of terms	0	10	20	20	30
Relevant docs	0	5	5	10	5
Avg. Precision	0.1163	0.1425 (22.6)	0.1437 (23.6)	0.1432 (23.1)	0.1476 (26.9)

Table 8. The average precision using post-translation feedback for English to Indonesian translated queries.

No. of terms	0	10	10	20	30
Relevant docs	0	30	40	40	40
Avg. Precision	0.0425	0.0717 (68.9)	0.0707 (66.5)	0.0687 (61.8)	0.0622 (46.4)

Table 9. The average precision using post-translation feedback for Indonesian to English translated queries.

3.4 Combining Pre- and Post-Translation Feedback

We also studied the possibility of adding terms before and after the translation were conducted. At first, we applied expanded the queries before the translation (pre-translation feedback). Then we applied the feedback technique again after the expanding queries were translated (post-translation feedback).

No. of terms	0	5	5	5	5
Relevant docs	0	20	30	40	50
Avg. Precision	0.1163	0.0673 (-42.1)	0.0627 (-46.1)	0.0752 (-35.3)	0.0688 (-40.8)

Table 10. The average precision of the combined feedback result for English to Indonesian translated queries.

The result of the combined method for the English-Indonesian queries is worse that that of the base queries as shown in Table 10. The added terms from the pre-translation feedback which are not relevant to the original query got even bad terms after being expand on post-translation feedback.

No. of terms	0	5	10	30	50	10
Relevant docs	0	10	10	20	30	40
Avg. Precision	0.0425	0.0479 (12.8)	0.0497 (17.0)	0.0628 (47.9)	0.0664 (56.3)	0.0668 (57.3)

Table 11. The average precision of the combined feedback result for Indonesian to English translated queries.

On the other hand, we got a positive improvement on the Indonesian-English queries. The words from the combined feedback produced more than 50% increased in average precision.

4. Conclusion

Our study shows that translating the query using dictionary can perform well in CLIR. The ambiguity on the translated query can be improved by adding words using local feedback query expansion strategy. From the result of the query expansion that is done before the translation, we learn that the characteristic of the training set such as the period of the articles and its domain is important in getting good terms.

The query expansion that is done after the translation shows a greater performance from the base queries. The number of relevant files on the English document collections that is bigger than the number of relevant judgement files in Indonesian document collection has increased the retrieval performance. The combined feedback also improves the performance of the translated queries even though as not good as post-translation feedback.

Our future research includes a study on the retrieval performance of translating terms in phrases found in the query instead of a term-by-term translation as in the current study.

Acknowledgements

We would like to thank Hendarto, Andrey Andoko, Hardanto from Kompas (http://www.kompas.com) and Saiful Basri from Tempo who provide the Indonesian text collections. We would also like to thank Prof. John U. Wolff who provides the Indonesian-English electronic dictionary on this study.

This material is based on work supported in part by the National Science Foundation, Library of Congress and Department of Commerce under cooperative agreement number EEC-9209623. Any opinions, findings and conclusions or recommendations expressed in this material are the author(s) and do not necessarily reflect those of the sponsor.

References

Attar, R. and A. S. Fraenkel. *Local Feedback in Full-Text Retrieval Systems*. Journal of the Association for Computing Machinery, 24: 397-417, 1977.

Ballesteros, Lisa and W. Bruce Croft. *Dictionary Methods for Cross-Lingual Information Retrieval*. Proceedings of the 7th International DEXA Conference on Database and Expert Systems Applications, 791-801, 1996.

Callan, J. P., W. Bruce Croft, S. M. Harding. *The Inquery Retrieval System*. Third International Conference on Database and Expert Systems Applications, 1992.

Hull, David A. and Gregory Grefenstette. *Experiments in Multilingual Information Retrieval*. In Proceedings of the 19th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval, 1996.

Davis, Mark W. and William C. Ogden. *Implementing Cross-Language Text Retrieval Systems for Large-scale Text Collections and the World Wide Web*. AAAI Symposium on Cross-Language Text and Speech Retrieval, 1997.

Echols, John M. and Hasan Shadily. *An Indonesian-English Dictionary*. Third ed. Cornell University, 1992.

Harman, Donna. *Overview of the Fourth Text Retrieval Conference (TREC-4)*. Proceedings of the Fourth Text Retrieval Conference (TREC-4), 1995.

Salton, Gerard. *Automatic Processing of Foreign Language Documents*. Journal of the American Society for Information Science, 21 : 187-194, 1970.

Sheridan, P. and J. P. Ballerini. *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, Zurich, Switzerland, August 1996.