University of Indonesia Participation at CLIR - CLEF 2005

Mirna Adriani and Ihsan Wahyu
Faculty of Computer Science
University of Indonesia
Depok 16424, Indonesia
(mirna@cs.ui.ac.id, ihsanw101@mhs.cs.ui.ac.id)

Abstract. We present a report on our participation in the Indonesian-English bilingual task of the 2005 Cross-Language Evaluation Forum (CLEF). We chose to translate an Indonesian query set into English using a commercial machine translation tool called *Transtool*, instead of using freely available resources for Bahasa Indonesia on the Internet which are not as complete as those for English. We show that improvement in retrieval effectiveness can be obtained using a query expansion technique.

Keywords: cross-language information retrieval, machine translation, query expansion.

1 Introduction

This year we, the University of Indonesia IR-group, participated in the bilingual 2005 Cross Language Evaluation Forum (CLEF) task, i.e., the English-Indonesian CLIR. We used a commercial machine translation software called *Transtool*[1] to translate an Indonesian query set into English. 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 we could improve our result this time.

2 The Query Translation Process

As a first step, we manually translated the original CLEF query set from English into Indonesian. We then translated the resulting Indonesian queries back into English using *Transtool*.

2.1 Query Expansion Technique

Adding translated queries with relevant terms (query expansion) has been shown to improve CLIR effectiveness [1, 3]. 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 applied this technique in this work. To choose the relevant terms from the top ranked documents, we used the *tf***idf* term weighting formula [4]. We added a certain number of noun terms that have the highest weight scores.

3 Experiment

We participated in the bilingual task with English topics. The English document collection contains 190,604 documents from two English newspapers, the *Glasgow Herald* and the *Los Angeles Times*. We opted to use the query title and the query description provided with the query topics.

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

The query translation process was performed fully automatic using *Transtool*. Using the query titles, the average length of the Indonesian queries was 3.1 words; the average length of the original English queries was 2.6 words; and the average length of the translated English queries was 2.7 words. Using the query descriptions, the average length of the Indonesian queries was 12.1 words; the average length of the original English queries was 9.5 words; and the average length of the translated English queries was 11.3 words. The number of Indonesian words that cannot be translated into English was 10 for the query titles and 26 for the query descriptions.

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 20 documents from the collection to extract the expansion terms. The terms that were used to expand the query were noun only terms. We used the Monty Tagger[2] to identify noun terms in those top 20 documents.

In these experiments, we used Lucene[3] information retrieval system which is based on the *vector space model* [4] to index and retrieve the documents.

4 Results

Our work focused on the bilingual task using Indonesian queries to retrieve documents in the English collections. Table 1 shows the result of our experiments.

Task	**Monolingual**	**CLIR** (translation)	**% Change**
Title	0.2810	0.1582	- 43.70%
Description	0.2364	0.1731	- 26.77%
Title + Description	0.3508	0.1830	- 47.83%

Table 1. Average retrieval precision of the monolingual runs of the title, description and combination of title and description topics and their translation queries using the machine translation.

The retrieval performance of the title-based translation queries dropped 43.70% below that of the equivalent monolingual retrieval (see Table 1). The retrieval performance of the description-based translation queries dropped 26.77% below that of the equivalent monolingual queries. The retrieval performance of using a combination of query title and description dropped 47.83% below that of the equivalent monolingual queries.

Query translation using machine translation (title)	**10 terms added**	**20 terms added**
0.1582 (0%)	0.1135 (-28.25%)	0.1248 (-21.11%)

Table 2. Average retrieval precision of the title-based queries using the query expansion technique with top-20 document method.

The translated title-based queries were then expanded using noun terms from the top 20 documents using the pseudo relevance feedback technique [4]. Adding 10 noun terms reduced the retrieval performance by 28.25%, however, adding 20 noun terms reduced the retrieval performance slightly less, i.e., by 21.11% (see Table 2).

[2] See http://web.media.mit.edu/~hugo/montytagger/.
[3] See http://lucene.apache.org/.

Query translation using machine translation (description)	10 terms added	20 terms added
0.1731 (0%)	0.0936 (-45.92%)	0.0907 (-47.60%)

Table 3. Average retrieval precision of the description-based queries using the query expansion technique with top-20 document method.

Next, the translated description queries were then expanded using noun terms from the top 20 documents using the pseudo relevance feedback technique. Adding 10 noun terms reduced the retrieval performance by 45.92% and adding 20 noun terms reduced the retrieval performance further by 47.60% (see Table 3).

Query translation using machine translation (description + title)	10 terms added	20 terms added
0.1830 (0%)	0.1285 (-29.78%)	0.1190 (-34.97%)

Table 4. Average retrieval precision of the title and the description-based queries using the query expansion technique with top-20 document method.

Finally, the translated title and description-based queries were expanded using noun terms from the top 20 documents using the pseudo relevance feedback technique. Adding 10 noun terms reduced the retrieval performance by 29.78% and adding 20 noun terms reduced the retrieval performance further by 34.97% (see Table 4).

4 Summary

Our results demonstrate that the retrieval performance of queries that were translated using machine translation for Bahasa Indonesia was about 53%-74% of that of the equivalent monolingual queries. The pseudo relevance feedback that is commonly used to improve the retrieval performance did not improve the retrieval performance. In fact, the longer the query is the worse the effect of using the query expansion technique. In our experiments, adding noun terms to the translated queries dropped the retrieval performance to 21%-47% of that of the equivalent monolingual queries. With such a short time available, we were not able to try different approaches to this task. We hope that we will obtain better results in our next participation in CLEF.

5 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. 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).
4. Salton, Gerard, and McGill, Michael J. *Introduction to Modern Information Retrieval*, New York: McGraw-Hill, 1983.
5. Attar, R. and A. S. Fraenkel. *Local Feedback in Full-Text Retrieval Systems*. Journal of the Association for Computing Machinery, 24: 397-417, 1977.