Evaluating Language Resources for CLEF 2007

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. This is a report on our evaluations of using some language resources for the Indonesian-English bilingual task of the 2007 Cross-Language Evaluation Forum (CLEF). We chose to translate an Indonesian query set into English using machine translation technique, transitive translation technique, and parallel corpus technique. We also made an attempt to improve the retrieval effectiveness using a query expansion technique. The result shows that the best result was achieved by combining the machine translation technique and the query expansion technique.

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

1 Introduction

To participate in the bilingual 2007 Cross Language Evaluation Forum (CLEF) task, i.e., the Indonesian-English CLIR, we needed to use language resources to translate Indonesian queries into English. However, there were not many language resources available freely on the Internet. We sought for some language resources that can be used for the translation process. 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. This lead us to exploring other possible approaches such as using machine translation techniques, and also transitive techniques [3, 4] that perform the translation through some other language, known as pivot language, that has more language resources.

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 machine translation technique, transitive queries technique, and the parallel corpus. For the machine translation technique, we translate the Indonesian queries into English using the available machine translation on the Internet. The transitive

technique uses German and French as the pivot languages. So, Indonesian queries are translated into French and German using bilingual dictionaries, then the German and French queries are translated into English using other dictionaries. The third technique uses a parallel corpus to translate the Indonesian queries. We created a parallel corpus by translating all the English documents in the CLEF collection into Indonesian using a commercial machine translation software called *Transtool*[1].We then created the English queries by taking a certain number of terms from certain number of documents that appear in the top document list.

2.1 Query Expansion Technique

Adding the translated queries with relevant terms (known as query expansion) has been shown to improve CLIR effectiveness [1, 3]. One of the query expansion techniques is called the *pseudo relevance feedback* [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 [5]. We added a certain number of 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. The query translation process was performed fully automatic using a machine translation technique, transitive technique, and the parallel corpus. The machine translation technique translates the Indonesian queries into English using *Toggletext*[2], a machine translation that is available on the Internet.

The transitive technique translates the Indonesian queries into English through German and French as the pivot languages. The translation is done using a dictionary. All of the Indonesian words are translated into German or French if they are found on the bilingual dictionaries, otherwise they stay in the original language.

We then applied a pseudo relevance-feedback query-expansion technique to the queries that were translated using the three techniques above. In these experiments, we used Lemur[3] information retrieval system, which is based on a language model, to index and retrieve the documents.

---

[1] See http://www.geocities.com/cdpenerjemah/.
[2] See http://www.toggletext.com/.
[3] See http://www.lemurproject.org/.

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.

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

Task	Monolingual	Machine Translation (MT)	% Change
Title	0.3835	0.3418	- 10.87%
Title + Description	0.4056	0.3237	- 20.19%

The retrieval performance of the title-based translation queries dropped 10.87% below that of the equivalent monolingual retrieval (see Table 1). The retrieval performance of using a combination of query title and description dropped 20.19% below that of the equivalent monolingual queries.

Table 2. Average retrieval precision of the monolingual runs of the title and combination of title and description topics and their translation queries using the machine translation and query expansion techniques.

Task	Monolingual	MT + Query Expansion	% Change
Title	0.3835	0.3375	- 11.99%
Title + Description	0.4056	0.3878	- 4.38%

The retrieval performance of the title-based translation queries dropped 11.99% below that of the equivalent monolingual retrieval (see Table 2) after applying the query expansion technique to the translated queries. It is reduced the average precision retrieval performance by 1.12% compared to the machine translation only. However, applying query expansion to the combination of the query title and description achieves 4.38% below that of the equivalent monolingual queries. It increases the average retrieval precision of the machine translation technique by 15.81%.

The result of using the transitive translation technique for the combination of the title and description queries is shown in Table 3. Translating the queries into English using German and French as the pivot language decreased the average precision by 30.2% compared to the monolingual queries. Applying the query expansion technique to the resulting English queries resulted in retrieval performance that is 15-18% of the equivalent monolingual queries. If we use only the translated queries resulted from using German as the pivot language and then apply the query expansion technique,

the average retrieval performance is about 14-17% of the equivalent monolingual queries.

Table 3. Average retrieval precision of the monolingual runs of the title and combination of title and description topics and their translation queries using transitive translation.

Task	**Monolingual**	**Transitive Translation**	**% Change**
Title + Description	0.4056	0.2831 (Union)	- 30.20%
Title + Description	0.4056	0.3437 (Intersection+QE 5 docs)	- 15.26%
Title + Description	0.4056	0.3297 (Intersection + QE 10 docs)	-18.71%
Title + Description	0.4056	0.3342 (German only + QE 5 docs)	-17.60%
Title + Description	0.4056	0.3460 (German only + QE 10 docs)	-14.69%

Table 4. Average retrieval precision of the monolingual runs of the title and combination of title and description topics and their translation queries using parallel corpus and query expansion.

Task	**Monolingual**	**PC + QE**	**% Change**
Title + Description	0.4056	0.0374 (top 20 docs)	- 90.77%
Title + Description	0.4056	0.0462 (top 5 docs)	- 88.60%

Next, we obtained the English translation of the queries using the parallel corpus-based technique and applied the pseudo relevance feedback technique using the top 5 and the top 20 documents. The retrieval performance decreased with the increase in the number of top documents considered, i.e., from -88.60% of the equivalent monolingual queries using top 5 documents to -90.77% using top 20 documents.

5 Summary

Our results demonstrate that the retrieval performance of queries that were translated using a machine translation technique for Bahasa Indonesia achieved the best retrieval performance compared to the transitive technique and the parallel corpus technique. The query expansion that is applied to the translated queries improves the retrieval performance of the translated queries. Even though the transitive technique performance was not as good as the machine translation technique, it can be considered as a viable alternative method for the translation process, especially for languages that do not have many available language resources such as Bahasa Indonesia.

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. A. (2000). "Cross Language Retrieval via transitive translation". In: Croft, W. B. (ed.) *Advances in Information Retrieval: Recent Research from the CIIR*, p. 203 – 234. Kluwer Academic Publishers.
4. Gollins, Tim and Sanderson, Mark. Improving Cross Language Retrieval with Triangulated Retrieval. In *Proceedings of SIGIR 2001, p. 90-95*. ACM Publisher.
5. Salton, Gerard, and McGill, Michael J. *Introduction to Modern Information Retrieval*, New York: McGraw-Hill, 1983.