Location Identification for the Geographic information Retrieval

Nasikhin and Mirna Adriani

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

Abstract. In this paper we identify location names that appear in queries written in Indonesian using geographic gazeeter. We built the gazeeter by collecting geographic information from a number of geographic resources. We translated an Indonesian query set into English using a machine translation technique. We also made an attempt to improve the retrieval effectiveness using a query expansion technique. The result shows that identifying locations in the queries and applying the query expansion technique can help improve the retrieval effectiveness for certain queries.

Keywords: cross-language information retrieval, geographic information retrieval, query expansion.

1 Introduction

As our participation in the Geographical Information Retrieval of the Cross Language Evaluation Forum (CLEF 2007) task, i.e., for Indonesian-English, we needed to use language resources to translate Indonesian queries 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. Luckily we found a machine translation tool available on the Internet that could help translate the Indonesian queries into English. However, GIR focuses on identifying geographical names [4] that appear in the queries, so we also needed to work on translating the location names from English to Indonesian.

2 The Process of Identifying Location Names

There are many resources that contain geographical information available on the Internet. We made use of the *Gazeeters* to build a location hierarchy map. The location hierarchy was built by extracting the names of countries, province, etc. from the *Gazeeters*. For each location, information about other locations within the area

that it covers was added, such as cities under a province, etc. We obtained the needed geographical information from the Geonames (http://www.geonames.org/) and Wikipedia (http://id.wikipedia.org).

We extracted the names of provinces, their capital cities, the names of mountains, seas etc. from the Geonames and its translation in Bahasa Indonesia from the Wikipedia. Each location has all their alternate names in both English and Bahasa Indonesia. If one location name appears in a query or document, it will be looked up in the gazeeters to find its associated locations that can be used as terms for searching for or indexing the document.

Most documents in the collection contain information about the location of events. For each document, we identified the location where the event mentioned in the document occurred, and added the location information into the document's index entry. For documents that contain more than one location, we choose the location that has the highest frequency in the document. If there is more than one location with the same highest frequency then a location is selected randomly among such locations.

To process the geographical locations further, we identify words that are related to a location name such as *in the (north/south/…) of, in the border of, around* etc. Then we include all location names that fall inside a boxline surrounding the location (city, country etc.) The boxline borders are at certain distance north, south, east, and west of a location.

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* [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 that came with the query topics. The query translation process was performed fully automatic using a machine translation technique. The machine

translation technique translates the Indonesian queries into English using Toggletext[1], a machine translation tool that is available on the Internet.

We then applied a pseudo relevance-feedback query-expansion technique to the queries that were translated using three techniques above. Beside adding terms, we also add location names only that appear on the top documents. In these experiments, we used Lemur[2] information retrieval system which is based on the language model 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.

Table 1. Average retrieval precision of the monolingual runs of the title, their translation queries, and the use of the geographic identification and query expansion on the translated queries.

Task	Monolingual	% Change
Title	0.1767	-
Title (translation)	0.1417	-19.80%
Title (Geoprocessing)	0.1736	-1.75%
Title (Geoprocessing + Geofeedback)	0.1389	-21.39%
Title (Geoprocessing + Pseudofeedback)	0.1936	+9.56%

The retrieval performance of the title-based translation queries dropped 19.80% below that of the equivalent monolingual retrieval (see Table 1). The retrieval performance of location identification process on the queries dropped 1.75% below that of the equivalent monolingual queries. Expanding the queries by adding geographic location from the top documents to the translated queries decreases the retrieval performance by 21.39%. However, adding terms that appear on the top documents on the translated queries improve the retrieval performance by 9.56%.

The retrieval performance of the combination of title and description queries that is translated by machine translation dropped 8.43% below that of the equivalent monolingual retrieval (see Table 2). The identification of the location on the queries improves the average precision 5.91%. Expanding the queries by adding the geographic location that appears from the top documents increase the average precision by 5.65%. However, adding terms from the top documents decrease the average precision by 2.02%.

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

Table 2. Average retrieval precision of the monolingual runs of the combination of title and description topics, their translation queries, and the use of the geographic identification and query expansion on the translated queries.

Task	Monolingual	% Change
Title + Description	0.1979	-
Title + Description (translation)	0.1812	-8.43%
Title + Description (Geoprocessing)	0.2096	+5.91%
Title + Description (Geoprocess + Geofeedback)	0.2091	+5.65%
Title + Description (Geoprocess + Pseudofeedback)	0.1939	-2.02%

5 Summary

Our results demonstrate that identifying location on the queries can have positive and negative effect on the queries. The query expansion technique that was applied to the queries by adding more terms and location names also produced mixed results. For the title queries, the query expansion had a positive impact when the combination of terms and location names were added to the queries. However, the same situation did not work for the combination of title and description queries. It had a positive impact only when the queries were added with terms or location names only. We still need to study further on the effect of location identification because the decreased in retrieval performance was not only caused by the failure in identifying the correct location names but also the failure in translating the words and location names in the queries from one language to another language.

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. Buscaldi, D., Rosso, P., Garcia, P. P. WordNet-based Index Terms Expansion for Geographical Information Retrieval. In the In *Working Notes for the CLEF 2006 Workshop*, 2006.
3. Larson, Ray. Cheshire II at GeoCLEF: Fusion and Query Expansion for GIR. . In *Proceedings of the Geographic Information Retrieval CLEF (GeoCLEF'05).* Vienna, Austria, 2005.
4. Overell, S. E., Ruger, S. Identifying and Grounding Descriptions of Places. In the *Proceedings of the Geographic Information Retrieval workshop (GIR'06).* Seattle, USA, 2006.
5. Salton, Gerard, and McGill, Michael J. *Introduction to Modern Information Retrieval*, New York: McGraw-Hill, 1983.