Dictionary-based CLIR for the CLEF Multilingual Track

Mirna Adriani
Department of Computing Science
University of Glasgow
Glasgow G12 8QQ, Scotland
mirna@dcs.gla.ac.uk

Abstract

This report describes the work done for our participation in the multilingual track of the Cross-Language Evaluation Forum (CLEF). We use a dictionary-based approach to translate English queries into German, French and Italian queries. We then apply a term disambiguation technique to select the best translation terms from the terms found in the dictionary entries, and a query expansion technique to enhance the queries' retrieval performance.

1 Introduction

This year, we participate in the Cross-Language Evaluation Forum (CLEF) as an opportunity for us to better understand the issues in Cross-Language Information Retrieval (CLIR) by means of experimentation, and, in particular, to measure the effectiveness of our techniques and algorithms. We chose to work on the multilingual track since it brings many new challenges in the IR field. Recently, we have done some work on bilingual CLIR. The new challenges, particularly the merging of results from different language collections, have provided us with insights as to the effectiveness of various approaches.

2 Dictionary-based Approach

We make use of publicly available online bilingual directories to translate English queries into German, France and Italian. The translation method is strait forward, that is by simply replacing each English term with the translation terms found in each of the language dictionaries. Clearly, the quality of the translation depends on the quality of the dictionary. A comprehensive dictionary in machine readable dictionary (MRD) format is very expensive and, so, we opted to use a free dictionary available on the Internet[1]. We consider the limited vocabulary of the dictionary that we use as an additional challenge.

2.1 Term Disambiguation Technique

In order to select the best translation terms from an entry in the dictionary, we apply our term disambiguation technique which is based on the statistical similarity values among terms. The similarity value is measured using the Dice similarity measure, based on the co-occurrences of terms in documents [3]. Basically, given a set of original query terms, we select for each of the terms the best sense such that the resulting set of selected senses contains senses that are mutually related- or statistically similar- with one another. For computational cost consideration, this is done using an approximate algorithm. Given a set of n original query terms $\{t_1, t_2, \ldots, t_n\}$, a set of translation terms, T, is obtained using the following algorithm:

[1] See <http://www.freedict.com>.

```
1. For each $t_i$ ($i$=1 to $n$), retrieve a set of senses $S_i$ from the dictionary.
2. For each set $S_i$ ($i$=1 to $n$), do steps 2.1, 2.2 and 2.3.
  2.1 For each sense $t_j$' ($j$=1 to $|S_i|$) in $S_i$, do step 2.1.1
    2.1.1 For each set $S_k$ ($k$=1 to $n$ and $k <> i$), get the maximum similarity, $M_{j,k}$, between $t_j$' and the senses in $S_{k.}$.
  2.2 Compute the score of sense $t_j$' as the sum of $M_{j,k}$ ($k$=1 to $n$ and $k <> i$).
  2.3 Select the sense in $S_i$ with the highest score, and add the selected sense into the set $T$.
```

Query terms that are not found in the dictionary are included in the translation set *T* as-is. This is typically the case for proper names, technical terms, and acronyms. For a complete explanation of our technique can be found in [2].

2.2 Query Expansion Technique

The resulting translated queries are, of course, worse than the original queries, in terms of their accuracy and retrieval effectiveness. We expand the translated queries by adding related terms to the queries to further improve their retrieval performance. Our query expansion technique uses a Dice similarity matrix which is computed based on the co-occurrences of terms in document passages. We built a database that contains passages of 200 terms from each collection. We then ran each query set to obtain the relevant passages. We used the top 20 passages to create the term similarity matrix. Next, we computed the sum of similarity values between each term in the passages with all terms in the query. Finally, we added the top 10 terms from the relevant passages to the query.

2.3 Rank Merging

The rank merging technique is required because we run the query set for each language collection independent of the other language collections. The results from the four language collections are then merged in a single rank list. We employ a simple method based on an assumption that the highest-rank document in one language is comparable, in terms of relevance to the query, to the highest-rank document in another language. We realized that this assumption is not always true, but, for lacking of time to experiment with other techniques, we thought that it was a reasonable one. With this assumption, we normalize the relevance scores with the highest score in each rank list, and then merge and sort the them into a single rank.

3 Experiment

In the multilingual track, the document collections are in four languages, namely, English, German, French, and Italian. We chose to run the English queries which were then translated using the online dictionaries.

First, we eliminated all stop words from the English queries and stemmed the remaining terms using the Porter stemmer. Each term is then translated into its translation or translations, if more than one is possible, according to the dictionary. We also included translations for terms that are part of phrases in the query. The translation terms were stemmed using the French and German stemmers from the PRISE retrieval system obtained from NIST. Stopwords in the translations were also removed. We then applied our term disambiguation technique to choose the best translation term. The resulting queries were then enhanced by applying the query expansion technique which adds 10 terms from a set of 20 relevant passages that are most closely related to the query terms. The values of 10 and 20 were obtained through a brief preliminary experiment.

Finally, we ran each query set on its respective document collection, including the original English queries on the English collection, and the retrieval results from the sets were then combined into a single document ranking.

In this experiment, we ran two query formats, namely, the title-only and the long (full) query formats. Each query in the long query set contains the title, the description, and the explanation texts of the CLEF query. We chose to do both query sets to see whether the results are consistent across both sets. All the steps in the multilingual task were done fully automatic.

4 Results

We participate in the multilingual task by running the title-only (glatitle) and the long (glalong) query formats. However, only the title-only query run was considered in the CLEF relevance assessment pool.

Run	Task	English	German	French	Italian
glatitle	Monolingual	0.2705	0.2075	0.2260	0.0347
glatitle	Cross Languange	-	0.0810	0.1097	0.0569
glalong	Monolingual	0.3804	0.2790	0.2682	0.1279
glalong	Cross Language	-	0.0932	0.1012	0.1050

Table 1. Average retrieval precision of the monolingual runs and the multilingual runs using English queries that are translated to German, French and Italian queries.

As can be seen in Table 1, we obtained good results for the Italian translation queries, followed by the French translation queries and, lastly, the German translation queries which performed the poorest. Our investigation of the title-only query format revealed that the retrieval performance of each translation query set is proportional with the number of English terms that could not be translated into the target language using the bilingual dictionary. Specifically, for the topic-only query format, our German query set contains 3 untranslated English terms and stand-alone German terms that were supposed to be 19 German compound nouns. The French query set contains 19 untranslated English terms. The Italian query set contains 8 untranslated English terms.

In our previous work [1], we demonstrated that our German queries perform better than the equivalent Spanish queries in retrieving documents from an English collection. The reason being that German compound words have exact meanings in English as compared to Spanish phrases if the phrases are translated word by word using bilingual dictionaries. In other word, the degree of ambiguity of the German queries is less than that of the Spanish queries. However, from this experiment, we learned that translating English queries to German, which involves translating into compound words, is a difficult task.

Lastly, we learned that our rank merging technique also contributed to our poor overall retrieval performance. We hope that we can improve it in the future. We also hope that the next time we will be able to use better machine-readable dictionaries.

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. Using Statistical Term Similarity for Sense Disambiguation in Cross Language Information Retrieval. *Information Retrieval* 2(1), p. 67-78. Kluwer: February, 2000.

[3] van Rijsbergen, C. J. *Information Retrieval*. Second ed. London, UK: Butterworths, 1979.