English-Dutch CLIR Using Query Translation Techniques

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

Abstract. We present a report on our participation in the English-Dutch bilingual task of the 2001 Cross-Language Evaluation Forum (CLEF). We attempted to demonstrate that good cross language query translation results can be obtained by combining a dictionary based and parallel corpus based techniques. A parallel corpus based technique was used to choose the best sense from all possible senses found in the dictionary. However, our results demonstrate that a pure dictionary based technique produces the best query translation than a parallel corpus based and the combined techniques. We also show that improvement in retrieval effectiveness can be obtained using a query expansion technique.

1 Introduction

This year we, the University of Glasgow IR-group, participate in the bilingual 2001 Cross Language Evaluation Forum (CLEF) task, i.e., the English-Dutch CLIR. We employ a dictionary based query translation technique using a publicly available dictionary on the Internet. We learned from our previous year's work that this dictionary does not provide good translation terms, as its vocabulary is very limited. We hoped that we could improve on the result using other resources. As a parallel corpus was made available for this forum, we opted to use it for this purpose.

2 The Query Translation Process

Our dictionary-based query translation technique translates each term in a given query to another language by replacing it with the senses of that term in the dictionary. There are well known problems with such translation techniques, mainly, the term ambiguity problem, the phrase translation problem, and the problem with terms not found in the dictionary, such as acronyms or technical terms. These problems result in very poor retrieval performance of the translation queries.

2.1 The Glossary Table

This year we use the RALI parallel corpus (Dutch and English) to create a glossary table. The parallel corpus has been aligned at sentence level. The glossary table contains all possible Dutch translation for each English term found in the corpus.

Since the parallel corpus was created automatically from documents found on the Internet, there are mis-classified documents where Dutch documents are classified as English and vice versa, or English documents contain Dutch words and vice versa. We filtered out mis-classified sentences using stopwords. If a Dutch sentence contains at least one English stopword then it is discarded, and so is an English sentence that contains at least one Ducth stopword. Although this technique may exclude correctly classified source sentences that happen to contain one or two terms of the target language, the detrimental effect of using a mis-classified

sentence is worse than incorrectly excluding a sentence. After this clean up process, the parallel corpus contains 62,536 sentences.

The glossary table entry is created using the following formula:

$$PT = tf_{ED} / (tf_{E} + tf_{D} - tf_{ED})$$

where

PT = the probability that an English term is a translation of a Dutch term
tf_{ED} = the frequency of both the English and the Dutch terms occur together
tf_{E} = the occurrence frequency of the English term
tf_{D} = the occurrence frequency of the Dutch term

For each English term in the sentence, we obtain the Dutch translation term from the parallel sentences with the highest *PT*. Below are some sample entries in the glossary table and their *PT* values:

4.5946 *find-vind*
5.5946 *human-humaan*
6.6868 *indian-indisch*
6.7637 *russia-rusland*

2.2 Choosing the Best Translation Term

The dictionary based query translation technique produces one or more translation terms in the target language for each term in the source language. A sense disambiguation technique is used to choose the best possible translation term. We perform the sense disambiguation process as follows:

1. Obtain the Dutch translation terms of the given English term from the dictionary.
2. Obtain entries in the glossary table for the same English term.
3. Select the Dutch translation term from terms obtained in step 1 that has the highest PT value in the entries obtained in step 2.
4. If the English term is not found in the dictionary but has entries in the glossary table then select the Dutch term from the glossary entries that has the highest *PT* value.
5. If the English term is not found in either the dictionary or the glossary table then it is taken without translation.

2.2 Query Expansion Technique

Expanding translation queries has been shown to improve the CLIR effectiveness. One of the query expansion techniques is called the *pseudo relevance feedback*. This technique is based on an assumption that the top few documents 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 translated queries. We applied this technique in this work. In choosing the good terms from the top ranked documents, we use the *tf*idf* term weighting formula [4]. We add a certain number of terms that have the highest weight values.

3 Experiment

The Dutch document collection contains 190,604 documents from two Dutch newspapers, the *Algemeen Dagblad* and the *NRC Handelsblad*. We participate in the bilingual task using the English topics. We opted to use the query title and the description for all of the available 50 topics. In addition to the combined dictionary and parallel corpus based technique we also conducted an experiment using a pure dictionary based query translation technique, and an experiment using a pure parallel corpus based query translation.

The query translation process is performed fully automatic. Stopwords are removed from the English queries and the remaining terms are stemmed using the Porter stemmer. For the pure dictionary based translation technique,

we simply include all possible translation terms found in the dictionary, i.e., without any sense disambiguation. If phrases in the queries are not found in the dictionary, they are translated by translating the individual constituent terms. For the pure parallel corpus based translation technique, we simply select the translation term in the glossary entry that has the highest *PT* value. In these two techniques, English terms that are not found in the dictionary or in the glossary table are taken without translation.

We use a machine-readable dictionary downloaded from the Internet at http://www.freedict.com. This dictionary contains short translations of English words in a number of languages. We realized that the dictionary is not ideal for our purpose, as most of its entries contain only one or two senses. However, its free availability outweighs its limitation. We reformatted the dictionary files so that our query translator program can read them. The dictionary contains 9,972 entries.

Then we apply the pseudo relevance feedback query expansion technique to the combined dictionary and parallel corpus based technique. We used the top 20 and 30 documents to extract the expansion terms.

In these experiments, we used the INQUERY information retrieval system to index and retrieve the documents. Terms in the Dutch queries and documents are stemmed using the Dutch stemmer from the Muscat system.

4 Results

Our work concentrates on the bilingual task using English queries to retrieve documents from the Dutch collections. The result that we submitted (code-named glaenl) is the one from the combined dictionary and parallel corpus based technique. Table 1 shows the result of our experiments. The retrieval performance of the translation queries obtained using the dictionary based technique falls 34.55% below that of the monolingual query (see Table 1). The performance of the query translation using parallel corpus only is the worst, i.e., 60.46% below that of the monolingual query. The retrieval performance of the combined method is 40.80% below the monolingual performance. This indicates that the parallel corpus based sense disambiguation technique drops the performance of the dictionary based translation queries by 5.25%.

Task	P/R	% Change
Monolingual	0.3238	-
Dictionary	0.2119	-34.55
Parallel corpus	0.1280	-60.46
Dictionary & P corpus	0.1917	-40.80

Table 1. Average retrieval precision of the monolingual runs and the bilingual runs using English queries that are translated to Dutch using dictionary only, parallel corpus only, and combined dictionary and parallel corpus.

Query translation using Dict & PC	5 terms	10 terms	20 terms	30 terms
0.1917	0.2002 (+4.43%)	0.2003 (+4.49%)	0.2048 (+6.86%)	0.2074 (+8.20%)

Table 2. Average retrieval precision of the query expansion using the top 30 document method.

Query translation using Dict & PC	5 terms	10 terms	20 terms	30 terms
0.1917	0.2123 (+10.72%)	0.2111 (+10.13%)	0.2116 (+10.39%)	0.2205 (+15.01%)

Table 3. Average retrieval precision of the query expansion using the top 20 document method.

The performance of the translated queries using only the dictionary is better than that of using only the glossary table. This correlates with the number of Dutch query terms that are not found the dictionary and the glossary table. Out of 569 English terms, there are 135 terms that are not found in the dictionary and 260 terms that are not found in the glossary table.

Dutch is a language that contains compound words as German. It has been shown that applying a compound-word splitter results in better retrieval performance [3]. Unfortunately, we do not have any Dutch compound-word splitter which could have improved the entries of our glossary table. In our previous work [1, 2], we showed that German queries can be better translated to English than Spanish queries because German compound words have exact meanings in English as compared to Spanish phrases which have to be translated word by word using a dictionary. In other words, the degree of ambiguity of the German queries is less than that of the Spanish queries. On the other hand, translating queries from English to German is a difficult task as it involves translating multi-word terms into single-word compound terms is a very difficult task. Such is also the difficulty in translating English queries to Dutch.

4 Summary

Our results demonstrate that using freely available bilingual dictionaries can produce superior cross-language retrieval performance as compared to using more expensive parallel corpora. However, this also depends on the breadth of topic coverage of the dictionary and the parallel corpus used. Given more sophisticated linguistic tools such as compound-word splitter and part of speech taggers, a parallel corpus based technique is likely to perform as effective, if not more, than the dictionary based approach.

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. Hiemstra, Djoerd, Wessel Kraaij, Renee Pohlmann, and Thijs Westerveld. Twenty-One at CLEF-2000: Translation resources, merging strategies and relevance feedback. In *CLEF 2000 Working Note Workshop*. Portugal, September 2000.
4. Salton, Gerard, and McGill, Michael J. *Introduction to Modern Information Retrieval*, New York: McGraw-Hill, 1983.