

# Research Proposal

## Resolving Ambiguity Using Dictionary-based Methods

### In CLIR

#### Background:

In the area of cross-language information retrieval (CLIR), research basically has focused on query translations. Translating all documents is computationally expensive and it's impractical to translate large document collections. So translating the query is more common. There're three main areas in the field: Machine Translation (MT), Dictionary-based methods and Parallel or Comparable Corpora.

MT system is expensive to develop. It tends to need huge amount of time and resources and current translation quality of MT is poor, with the facts that many words may have multiple meanings, some sentences may have multiple grammar structures and what a pronoun refers to is uncertain. So improvements gained via MT may not outweigh the cost of linguistic analysis in MT.

About the corpus-based methods, queries are translated on the basis of multilingual terms extracted from parallel/comparable doc collections. However, obviously the translation performance is dependent on how well the corpora are aligned. Also, the parallel corpora are not always available and those available ones tend to be relatively small or to cover only a small number of subjects.

Compared with parallel corpus, the topic coverage of dictionary-based methods is less limited because a dictionary usually contains a wider variety of terms than a parallel corpus. It may not that expensive and hard to build, compared with MT, and also online dictionaries are widely available now. Also, there're quite a few dictionary-based methods which were mentioned in the papers I read in this area. So I have the feeling that there may be more things can be improved by analysis the dictionary-based methods in this area aiming at resolving ambiguity in query translation.

#### Problem Statements:

From a bunch of papers I looked through, I got to know that ambiguity happened in query translation is one of the biggest issue. This may be caused by three factors which is mentioned in paper "Querying across languages: A dictionary-based approach to multilingual information retrieval": The 1<sup>st</sup> factor is the addition of

extraneous terms to query, because a dictionary entry may list several senses for a term, each of them has one or more possible translations. The 2<sup>nd</sup> one is failure to translate technical terms, because technical terms often are not found in general dictionaries. The 3<sup>rd</sup> one is failure to translate phrases or translate them poorly, because sometimes the meaning of a phrase is quite different from the meaning we got through the translation word by word.

To minimize the negative effect of the addition of extraneous terms to query which is the 1<sup>st</sup> factor to cause query translation ambiguity as I mentioned above, in paper “Dictionary-based methods for cross-lingual information retrieval”, the author introduced an approach—query expansion via local feedback. This method also was used for further disambiguating query translation in the experiments the researchers mentioned in paper “Resolving Ambiguity for Cross-Language Retrieval”. Local feedback is familiar with relevance feedback. A query is modified by the addition of terms found in documents known to be relevant to the query. Applying feedback prior to query translation which is called pre-translation may create a stronger and more specific query for translation by adding terms that emphasize query concepts. Feedback after query translation which is called post-translation may decrease the effect of irrelevant query terms by adding more context specific terms. In their research, they assume that the top retrieved documents are relevant. Here, one problem rises up...

#### Problem 1:

It's very possible that some initial query terms may only have one translation while other terms may have many translations in the target language. When trying to retrieve documents by using this kind of translated query, we may not get the real relevant docs. Then after applying query expansion techniques, the effectiveness may not get improved, or even decreased. The reason being is that usually the query terms which have only one or two translations are the most important words for querying and vice versa, the terms that have many translations are unimportant query words and also tend to cause ambiguity easily. In this case, when trying to retrieve the top relevant documents, those unimportant query keys and those irrelevant translations may dominate the retrieval results and cause the retrieved top N docs are not that relevant at all. Then obviously, the top M terms which were extracted from the top N bad docs might not be good feedback terms which would be used to expand the query. The proposed solution or strategy will be given in next part.

#### Problem 2:

Still in the paper I listed above, the author concluded based on the data got from their designed experiments that “combining pre- and post-translation expansion is most effective and improves precision and recall”. However, I noticed that they only examined a single language pair English to Spanish, and relied on the Collins's

English-Spanish electronic dictionary which is just a general bilingual dictionary. This may cause failure to translate technical query terms, which is just the 2<sup>nd</sup> factor that induces the translation ambiguity issue, because specific technical query terms rarely can be translated in the general dictionary. A proposed solution or strategy will be given in the next part also.

### Proposed Strategies:

#### Strategy 1:

For solving the problem 1 I mentioned in previous part, we may consider giving query terms different weights by increasing the weight of the search keys which have only few translations, since always they are the most important words of a query and ambiguity seldom happened to them, and vice versa. We expect that using this strategy can dramatically avoid that the top retrieved documents are not really relevant to the query during the process of applying local feedback. So the effectiveness may get enhanced and the ambiguity caused by problem 1 will be solved or at least decreased. Let's take a look at a query example I chose from TREC topics.

<num> Number: 441

<title> Lyme disease

<desc> Description:

How do you prevent and treat Lyme disease?

<narr> Narrative:

Documents that discuss current prevention and treatment techniques for Lyme disease are relevant. Reports of research on new treatments of the disease are also relevant.

For example, we use the description part to do searching. Term 'Lyme' may have only one translation in the bilingual dictionary, but terms 'prevent', 'treat' and so on may have bunch of meanings. And obviously, the term 'Lyme' is a more important query term as a search key. So give it more weight, then rank the retrieved docs based on not only tf\*idf, but also the weights of query terms. About the detailed algorithm, we can analysis the algorithms applied in relevance feedback and choose a most suitable one or just build one based on all the factors which affect the relevance of a document and a query. Finally after computation, we can get a score that a document D is relevant to a query Q, and then based on the scores, rank all retrieved documents and expand the query by using the top M terms appeared in the top N documents.

#### Strategy 2:

For resolving the ambiguity caused by problem 2 I mentioned above, we may try to use some technical dictionaries with the general one. Because special technical query

terms are rarely translated through the general bilingual dictionaries, though they may be crucial search keys. But most likely, the technical dictionaries are able to translate them correctly and uniquely, since the terms in the special dictionaries are often unambiguous. Here, the same query sample is used as above.

In TREC topic #441, the description is “How do you prevent and treat Lyme disease?” The most important search key is clearly the term “Lyme”. This term may only occur in a medical dictionary, but may not occur in a general one. It’s a kind of arthritis.

Based on these analyses, I think a special technical dictionary should alleviate the translation ambiguity. Also a general dictionary is still necessary to be used in query translation. So we expect that using special dictionary and general dictionary can solve problem 2 and then get higher effectiveness. One thing need to be dealt with here is that we need to address automatically the specific query terms to the correct technical dictionaries, because different domains have different terminologies.

#### Further disambiguation:

About the 3<sup>rd</sup> factor which can cause the translation ambiguity as I mentioned above – “failure to translate phrases or translate them poorly”, we may use Co-occurrence Statistic Model which is illustrated in paper “Resolving Ambiguity for Cross-Language Retrieval” to identify and translate phrases with a phrase machine-readable dictionary, though the phrase problem is language specific.

#### Test Environment:

The test environment of this research proposal may consist of:

- ✧ TREC Cross-language topics, documents
- ✧ A general MRD for English-Spanish translation
- ✧ One or more technical MRDs for English-Spanish translation
- ✧ A phrase English-Spanish dictionary
- ✧ A retrieval system, i.e. INQUERY

#### Evaluation:

Compare the average precisions at different recalls for word-by-word translation as a baseline, word-by-word translation augmented by query expansion, and by query expansion with weighting strategy. If the average precisions of using weighting strategy basically are higher than other runs, it means that disambiguation through increasing the weights of relevant query terms is effective. Also we can get the precision of monolingual retrieval which is the best case we wish to achieve in cross-language IR. Then comparing it with the precision got from weighting method,

we can know how effective the weighting method is.

Run and compare the average precision at different recalls for word-by-word translation using only general bilingual MRD as a baseline, general MRD with one or more technical MRD, general MRD with one or more technical MRD and phrase MRD, and general MRD with one or more technical MRD, phrase MRD and Co-occurrence. Then do corresponding analyses as above. Through all the running data, we'll know if the proposed methods take effect in resolving ambiguity of query translation. If so, we also can know how effective each strategy is.

### Reference:

Lisa Ballesteros and W. Bruce Croft. "Dictionary-based methods for cross-lingual information retrieval" Proceedings of the 7th International DEXA Conference on Database and Expert Systems Applications

David A. Hull and Gregory Grefenstette. "Querying across languages: A dictionary-based approach to multilingual information retrieval" Proceedings of the 19<sup>th</sup> International Conference on Research and development in Information Retrieval, page 49-57, 1996

L. Ballesteros, W.B. Croft, "Resolving Ambiguity for Cross-Language Retrieval" Proceedings of ACM SIGIR, 64-71, 1998.

M. Aljlal and O. Frieder, "Effective Arabic-English Cross-Language Information Retrieval via Machine Readable Dictionaries and Machine Translation,"

M. Aljlal, O. Frieder, and D. Grossman, "On Bidirectional English-Arabic Search," Journal of the American Society of Information Science and Technology, 53(13), November 2002.

Lisa Ballesteros and W. Bruce Croft. Phrasal translation and query expansion techniques for cross-language information retrieval.

Leah S. Larkey and Margaret E. Connell. Structured Queries, Language Modeling, and Relevance Modeling in Cross-Language Information Retrieval

David A. Grossman and Ophir Frieder. Information Retrieval: Algorithms and Heuristics