D-34

Combining Multiple Knowledge Sources for an Efficient Query Expansion in Cross-Language Information Retrieval

Fatiha SADAT†, Masatoshi YOSHIKAWA‡† and Shunsuke UEMURA†

1. Introduction

Cross-Language Information Retrieval (CLIR) consists of providing a query in one language and searching document collections in one or more languages. In the present paper, we focus on query translation and disambiguation to reduce errors associated with polysemy, caused after a simple dictionary translation. Combined query expansion, both before and after translation and disambiguation is used to improve the effectiveness of information retrieval. Therefore, we propose a model using multiple sources for query reformulation through different expansion techniques, such as pseudo-relevance feedback, thesauri and a new feedback strategy named domain-based feedback. Domain-based feedback is based on the extraction of domain keywords from web categories in order to expand original queries. We tested the effectiveness of different combinations of query expansion techniques on French-English information retrieval using TREC data collection. A combined domain-based feedback and thesauri-based expansion showed the greatest improvement for the effectiveness of information retrieval. The rest of this paper is organized as follows: Section 2 gives an overview of the adapted strategy for query translation and disambiguation in CLIR. Query expansion techniques are described in Section 3. Evaluation and results are discussed in Section 4. Section 5 presents the conclusion of this paper.

2. Translation/Disambiguation in CLIR

The translation/disambiguation module of the proposed CLIR system is concerned by the following tasks: Organization of source query terms, dictionary term-by-term translation and disambiguation of target translations. Fig.1 shows an overview of the proposed CLIR system. Missing words in the bilingual dictionary, which are essential for the correct interpretation of the query, can be solved by an automatic compensation through a synonym dictionary or an existing thesaurus related to the source language, in order to find equivalent terms or synonyms. A statistical disambiguation method of target translations is performed, in order to select best translations for each source query term [4]. Co-occurrence tendency using log-likelihood ratio [2] is applied, as follows:

1. Organize source query terms by pairs, i.e. all possible combinations and their co-occurrence tendencies.
2. Select the pair of source terms having the highest co-occurrence tendency.
3. Retrieve all translations related to the selected source terms from the bilingual dictionary.
4. Select and fix best translations for each source term in the combination using a disambiguation method, based on highest co-occurrence tendencies.
5. Go to the next combination of source terms having the next highest co-occurrence tendency and repeat steps 2 through 4 until the translation of every source query term is fixed.

3. Query Expansion

Combined query expansion has proved its effectiveness for information retrieval [1]. In the present research, extracting, selecting and adding terms that emphasize query concepts is performed through different query expansion techniques. We evaluate different combinations using pseudo-relevance feedback with the selection of relevant terms, domain-based feedback with an extraction of domain keywords from web categories and thesaurus-based expansion with a selection of relevant synonyms or multiple word senses from a monolingual thesaurus.

3.1 Pseudo-Relevance Feedback

Pseudo-relevance feedback is applied for the query reformulation and expansion by fixing number of retrieved documents and assuming the top ranked ones as relevant. A fixed number of term concepts are extracted and their co-occurrence tendencies in conjunction with original query terms are estimated. However, any query expansion must be handled very carefully, selecting any expansion term could be dangerous. Our selection is based on the statistical co-occurrence tendency in conjunction with all terms of the original query, rather than with one term.

3.2 Domain-based Feedback

This approach [3] aims to extracting domain keywords from a set of top retrieved documents using pseudo-relevance feedback to expand an original query set. Web categories extracted from *Yahoo!*[1], *AltaVista*[2] and digital libraries such as the *library of congress catalogue* that support some form of subject indexing, are exploited for keywords extraction in domain-based feedback.

3.3 Thesaurus-based Expansion

Thesaurus-based expansion is completed with relevant terms, which are extracted from *WordNet* lexical database [5] for English queries and *EuroWordNet* [6] for French queries. Synonyms and multiple word senses with semantic relations to original query terms are used as expansion candidates.

4. Evaluations and Experiments

We conducted some experiments to evaluate the proposed disambiguation, translation and expansion techniques using French queries to retrieve English documents.

4.1 Linguistic Tools

† Graduate School of Information Science, Nara Institute of Science and Technology (NAIST)
‡ Information Technology Center, Nagoya University

[1] http://www.yahoo.com
[2] http://www.altavista.com

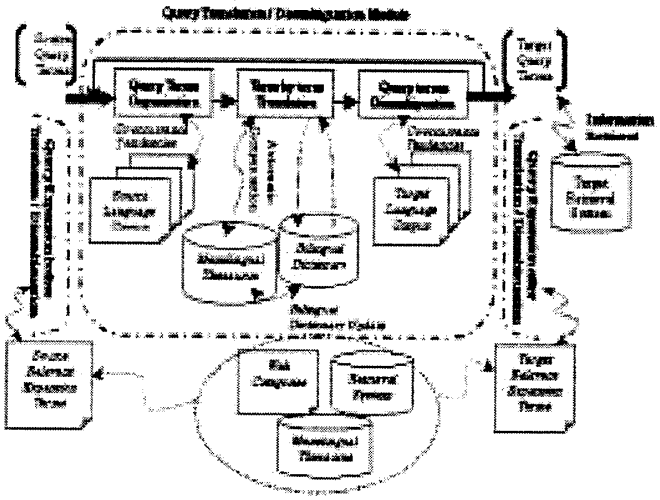

Fig.1. **Overview of the proposed CLIR system**

We used *TREC*[3] volume 1data collection. Topics 51-150 were considered and key terms contained in fields <title> and <description>, which are averaged 5.7 terms by query were used to generate original English and French queries. The Canadian *Hansard* corpus (Parliament Debates) was used for both French and English languages. *COLLINS* French-English dictionary performed the translation of source terms. *WordNet* and *EuroWordNet* were considered for the thesaurus-based expansion. *SMART*[4] information retrieval system, which is based on vector space model, was used to retrieve English documents.

4.2 Experiments and Discussion

Retrieval with original English queries was represented by *Mono_ Eng* method. An average precision measure was used as the basis of evaluation. The proposed statistical disambiguation method *N_DIS* showed a great improvement with 101.94% of average precision of the monolingual retrieval. This suggests that ranking and selecting pairs of source terms is very helpful to the translation and disambiguation methods. Query expansion before or after translation/disambiguation showed a drop in precision/recall. However, combined *Feed.bef_aft* improved the average precision with 102.89% of the monolingual counterpart. Domain-based feedback showed a drop in average precision but combined with a relevance feedback both before and after translation/disambiguation *Feed.bef_dom* improved the average precision. Using a thesaurus for query expansion (WordNet or EuroWordNet) did not show any improvement. Best results were achieved by the combined thesauri and domain-based feedback *Feed.ewn_wn_dom,* with 104.29% in term of average precision of the monolingual counterpart. This suggests that adding domain keywords to generalized thesauri improves the effectiveness of retrieval. Results and performances of different methods are described in Table 1. Thus, main techniques used in this successful proposed method are summarized as follows:

- A combined statistical disambiguation method was applied first prior to translation, in order to eliminate misleading pairs of terms to translate and disambiguate, then after translation in order to select best target translations.
- Each type of query expansion has different characteristics and therefore their combination could provide a valuable resource for query expansion.
- Adding domain keywords to original queries and then selecting thesaurus word senses in order to avoid misleading expansion terms, is considered as a very effective method to improve the effectiveness of information retrieval.

Table1. **Best results for different combinations of query translation, disambiguation and expansion in CLIR**

Method	Avg. Prec	% Mono
Mono_Eng *(baseline)*	0.2628	100
N_DIS	**0.2679**	**101.94**
Feed.aft	0.2663	101.33
Feed.bef_ aft	**0.2704**	**102.89**
Feed.bef_ dom	**0.2725**	**103.69**
Feed.ewn_wn_dom	**0.2741**	**104.29**

5. Conclusion

A combined dictionary-based and statistical-based method has been used efficiently in CLIR. We proposed and evaluated in the present paper an efficient disambiguation method for short and long queries to apply with all source query terms, rather than with one term. Also, combined query expansion techniques prior and after translation and disambiguation provided valuable resources for information retrieval. The proposed domain-based feedback was combined to thesauri-based expansion before and after translation and fulfilled the best improvement in the context of information retrieval. Among ongoing researches, we would like to investigate an approach to learning from category hierarchies to extract and select domain keywords, in order to enhance query expansion techniques in CLIR.

References

[1] Ballesteros, L. and Croft, W.B. 1997. Phrasal translation and query expansion techniques for Cross-Language Information Retrieval. In Proceedings of the 20th ACM SIGIR Conference.

[2] Dunning, T.E. 1993. Accurate methods for the statistics of surprise and coincidence. Computational Linguistics, 19(1): 64-74.

[3] Sadat, F., Maeda, A., Yoshikawa, M. and Uemura, S. 2001. Query Expansion Techniques for the CLEF Bilingual Track. In Proceedings of the CLEF 2001 Cross-Language Evaluation Campaign.

[4] Sadat, F., Maeda, A., Yoshikawa, M. and Uemura, S. 2002. Statistical Query Disambiguation, Translation and Expansion in Cross-Language Information Retrieval. In Proceedings of the LREC 2002 Workshop on Using Semantics for Information Retrieval and Filtering: State of the Art and Future Research.

[5] Voorhees, M.E. 1994. Query expansion using lexical-semantics relations. In Proceedings of the 17th ACM SIGIR Conference.

[6] Vossen, P. 1998. EuroWordNet, a multilingual database with lexical semantic networks. The Kluwer Academic Publishers.

[3] http://trec.nist.gov/data.html
[4] ftp://ftp.cs.cornell.edu/pub/smart