

Exploiting Thesauri and Hierarchical Categories in Cross-Language Information Retrieval

Fatiha Sadat¹, Masatoshi Yoshikawa^{1,2}, and Shunsuke Uemura¹,

¹ Graduate School of Information Science, Nara Institute of Science and Technology (NAIST), 8916-5 Takayama, Ikoma, Nara 630-0101, Japan

² National Institute of Informatics (NII)
{fatia-s, yosikawa, uemura}@is.aist-nara.ac.jp

Abstract. As Internet resources become accessible to more and more countries, there is a need to develop efficient methods for information retrieval across languages. In the present paper, we focus on query expansion techniques to improve the effectiveness of an information retrieval. A combination to a dictionary-based translation and statistical-based disambiguation is indispensable to overcome translation's ambiguity. We propose a model using multiple sources for query reformulation and expansion to select expansion terms and retrieve information needed by a user. Relevance feedback, thesaurus-based expansion, as well as a new feedback strategy, based on the extraction of domain keywords to expand user's query, are introduced and evaluated. We evaluated the effectiveness of the proposed combined method, by an application to a French-English Information Retrieval.

1 Introduction

Cross-Language Information Retrieval (CLIR), consists of providing a query in one language and searching document collections in one or multiple languages.

In this paper, we focus on query expansion, which has been known to be among the most important methods to overcome the word mismatch problem in information retrieval. The proposed study is general across languages in information retrieval however; we have conducted experiments and evaluations on French and English languages. The rest of this paper is organized as following: Section 2 gives an overview of a translation and disambiguation approach in CLIR. Query expansion techniques with different combinations are introduced in Section 3. Experiments and evaluations are discussed in Section 4. Section 5 concludes the paper.

2 Query Translation / Disambiguation in CLIR

In our approach, a term-by-term *translation* using a bilingual machine-readable dictionary is performed after a simple *stemming* process of query terms to replace

each term with its inflectional root and to remove stop words and stop phrases. Missing words in the dictionary, which are essential for the correct interpretation of the query, can be solved by an automatic *compensation* through a synonym dictionary related to that language or by an existing monolingual thesaurus. This case requires an extra step of looking up the query term in the synonym dictionary or thesaurus, when missing words in the bilingual machine-readable dictionary, to find equivalent terms or synonyms of the concerned query term, before a translation. A disambiguation method [5], [6] using a co-occurrence tendency based on a Log Likelihood Ratio [2] is applied to filter and select best translations, among candidates to create target queries to retrieve documents. An overview of the proposed information retrieval system is shown in Fig.1. Query expansion is completed through a monolingual thesaurus, a relevance feedback (interactive or automatic) or a domain-based feedback.

3 Query Expansion in CLIR

Query expansion has proved its effectiveness in the performance of an information retrieval [1], [3]. We use an approach of combined automatic query expansion before and after translation, with an extraction of expansion terms through the following techniques: Relevance feedback, with the selection of best terms, domain-based feedback with the extraction of domain keywords to add to the original query and thesaurus-based expansion with a retrieval of synonyms from a monolingual thesaurus.

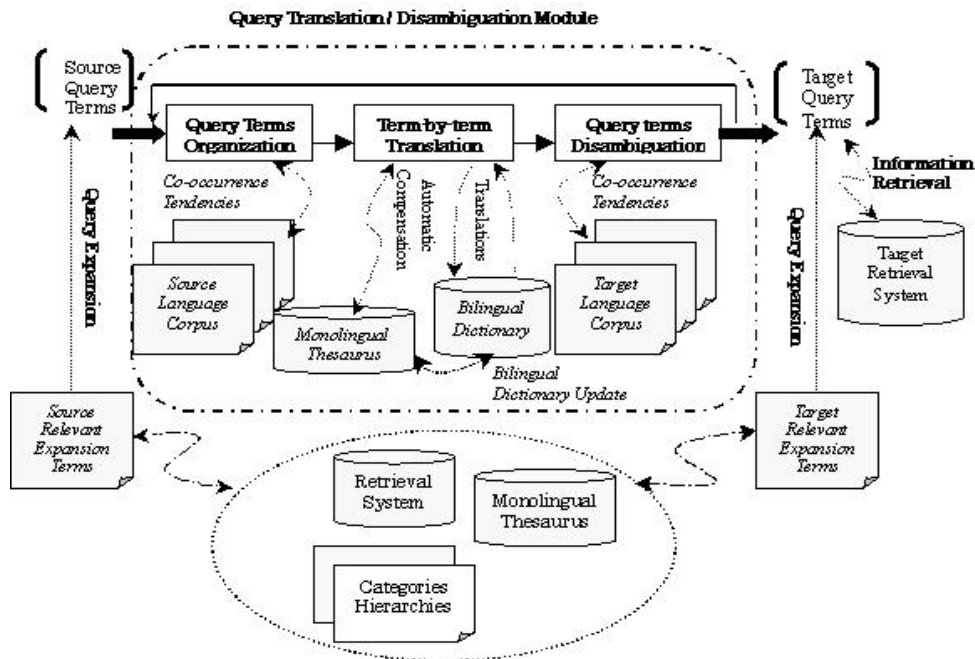


Fig. 1. An overview of an Information Retrieval System using Source and Target Languages

3.1 Relevance Feedback

A fixed number of term concepts will be extracted from the top retrieved documents, and their co-occurrence in conjunction with original query terms will be computed. However, any query expansion must be handled very carefully. Selecting any expansion term could be dangerous. Therefore, our selection is based on statistical co-occurrence frequency in conjunction with all terms of the original query, rather than with just one query term.

Assume that we have a query Q with n terms: $\{term_1 \dots term_n\}$, a ranking factor based on the co-occurrence frequency between each term in the query and the expansion term candidate, is evaluated, such as:

$$Rank(expterm) = \sum_{j=1}^n co-occurrence(term_j, expterm) \quad (1)$$

where, $co-occurrence(term_j, expterm)$ represents the co-occurrence tendency between a query term $term_j$ and an expansion term candidate $expterm$, and can be evaluated by any estimation such as a log-likelihood ratio, etc ... Thus, all co-occurrence values were computed, summed for all query terms and an expansion candidate with the highest rank is selected as an expansion term for the concerned query.

3.2 Domain-based Feedback

We introduce a domain-based feedback [6] as a query reformulation strategy, which consists to extract domain keywords from a set of top retrieved documents, using a classical relevance feedback to expand an original query set. Web directories, such as Yahoo!¹ or AltaVista², are human constructed and designed for human web browsing. They provide a hierarchical category scheme and documents are sorted into the given scheme. Our strategy relies on terms extraction using a classical relevance feedback with a condition that these terms represent a directory or category, which is denoted by a keyword describing its content and thus will be considered as a specific domain to collection of documents. The process is described as following:

- Extract some terms or seed words, by using relevance feedback as well as a ranking strategy to select the expansion term, as explained in the previous section. This set is denoted by set_1 ,
- Collect domain keywords candidates, from categories and directories related to some hierarchical web directories, such as: Yahoo!¹ or AltaVista² or Open Directory³, which is denoted by set_2 ,
- Select a fixed number of domain keywords as seed words from set_1 but also a candidate of set_2 .

¹ <http://www.yahoo.com/docs/pr>

² <http://www.altavista.com>

³ <http://dmoz.org/>

In case of large number of resulting domain keywords, a statistical process will be applied for ranking and selecting of best ones. The resulting set of keywords will be used for a query expansion, which may involve many keywords or just a subset of them.

3.3 Thesaurus-based Expansion

This approach is based on expanding a query with a fixed number of relevant terms from a structure derived from a lexical database, WordNet [9] for English queries and EuroWordNet [10], [11] for French queries could be seen as powerful tools to study lexical semantic resources and their language-specificity [8], [11]. Our first suggestion is that synonyms of a query term can be used, as expansion candidates. Following the research reported by Voorhees [9] on the use of lexical relations of WordNet for a query expansion, we can proceed by a simple look up to find synsets of a full query, otherwise, we proceed by a term-by-term search in case of non-existence of a full query in the lexical database. A statistical frequency might be used for ranking and selection to avoid words that do not occur frequently with original terms, such as "*reckoner*" which will be removed from the synset list of "computer". An appropriate weighting scheme will allow a smooth integration of these related terms by reducing their influence over the query [4]. Thus, all terms recovered from the thesaurus will be given weights, expressing their similarity to original query terms, based on their position in the conceptual hierarchy (depth = 1) as well as number of terms accompanying them in the same synset. Some strategies were proposed [4], [9] for sense disambiguation and weight assignment to synonyms and other terms in a thesaurus. In this study, weights assigned to any synonym of one synset would be related to an envelope of 0.5 divided by number of terms in the corresponding synset, which is proportional in the same synset. Following these assumptions, the expanded query with synonyms, would contain: {*computer, data processor, electronic computer, information processing system, calculator, figurer, estimator*}.

The proposed weighting factor for retrieved expansion terms from synsets of a conceptual hierarchy related to the WordNet thesaurus, is described as following:

$$Weight(term, exp_j) = \frac{Sim(term, exp_j)}{2 \times M} \quad (2)$$

Where, M is the number of terms that belong to the same synset. $Sim(term, exp_j)$ is the similarity between a $term$ and an expansion candidate exp_j and could be estimated by any similarity measure, such as the Cosine measure [7] as following:

$$Sim(term, exp_j) = \frac{\sum_i v_{si} v_{ti}}{\sqrt{\sum_i v_{si}^2 \sum_i v_{ti}^2}} \quad (3)$$

where, v_{si} and v_{ti} are frequencies of $term$ and exp_j in a corpus, respectively. However, expanding a query with any of those weighted synonyms implies a careful selection and ranking, depending on statistically most weighted terms, in conjunction

with all query terms, not just one term query. For a query Q with k terms $\{term_1, term_2, \dots, term_k\}$, weights factors would be computed for an expansion term candidate, summed for all query terms, if the expansion term appears in the related hierarchy and the highest weighted term is selected for a query expansion, as following:

$$Weigth(query,expterm) = \sum_{j=1}^k Weight(term_j,expterm) \quad (4)$$

3.4 Combining Different Approaches

Following the research reported by Ballesteros and Croft[1] on the use of a local feedback, adding terms that emphasize query concepts in the post and pre-translation phases, improves precision and recall. This combined method is supposed to reduce the ambiguity by de-emphasizing irrelevant terms added by translation and will improve precision and recall of an information retrieval. The new query Q_{new} can be defined as following:

$$Q_{new} = Q_{orig} + \alpha_1 \sum_{bef} T_i + \alpha_2 \sum_{aft} T_j \quad (5)$$

where, Q_{orig} is an original query, $\sum_{bef} T_i$ and $\sum_{aft} T_j$ represent an added set of terms

before and after translation/disambiguation, consecutively. The two parameters α_1 and α_2 , which represent the importance of each expansion strategy, are given by human experimentally at this moment, but could be estimated using an Expectation-Maximization algorithm.

4 Experiments and Evaluation

Conducted experiments were completed on the proposed strategies for query translation disambiguation and expansion, with an application to a French-English information retrieval, i.e. French queries to retrieve English documents. Linguistic tools used in these experiments are described as following:

- *CLEF 2001 test collection*⁴ was used for a cross-language evaluation. Topics composed of fields <title> for title, <desc> for description and <narr> for narrative, are considered. Key terms contained in these fields, which are averaged 20.6 terms by query, are used to generate French and English source queries.
- *Hansard corpora* (Canadian Parliament Debates) are bilingual French-English parallel corpora, containing more than 100 million words of English text and their corresponding French translations. In this study, we have used Hansard as monolingual French / English corpora.

⁴ <http://www4.eurospider.ch/CLEF/>

- *COLLINS⁵ Series 100 French-English bilingual dictionary* was used to translate source queries. The bilingual dictionary includes 75.000 references and 110.000 translations, which seems to be plenty for research.
- *WordNet* [9] and *EuroWordNet* [11] are used as thesauri to the query expansion and possible compensation, in case of limitation in the bilingual dictionary.
- *Porter⁶ Stemmer* was used for the stemming part.
- *SMART⁷*, an information retrieval system based on vector space model that has been used in many researches for CLIR would retrieve English and French documents.

4.1 Experiments and Results

A retrieval with original French / English queries were represented by *Mono_Fr / Mono_Eng* methods, respectively. *No_DIS* is the result of translation without disambiguation, which means selecting the first translation as target, for each source query term. *N_DIS* refers to the translation and disambiguation method (trans_disambiguation). Expansion methods were represented by: *Feed.bef / Feed.aft*, for a relevance feedback before / after the trans_disambiguation, respectively. *Feed.bef_aft* refers to a combined relevance feedback before and after the trans_disambiguation. Domain-based feedback was evaluated with *Feed.dom*, after the trans_disambiguation. A combined method with a relevance feedback was tested with *Feed.bef_dom*. WordNet-based expansion was evaluated using synsets of target translations with *Feed_wn*, as well EuroWordNet-based expansion on synsets of source queries with *Feed.ewn*. Combined thesauri-based expansion and relevance-feedback, is represented by *Feed.bef_wn* and *Feed.ewn_aft*, with domain-based feedback, by *Feed.dom_wn* and *Feed.ewn_dom*. Combined thesauri-based expansion was tested with *Feed.ewn_wn*. Finally, *Feed.ewn_wn_dom* represents combined thesauri with domain-based expansion. Results and performances of these methods are described in Table 1, with an average precision and a difference comparing to the monolingual counterpart.

4.2 Discussion

The disambiguation method *N.DIS* showed a better improvement in terms of average precision, 101.94% of monolingual retrieval, comparing to *No_DIS* or to monolingual English or French retrieval. *Feed.aft* showed a good help to the average precision, with a 101.33 % of the monolingual counterpart. Combined relevance feedback techniques before and after trans_disambiguation *Feed.bef_aft* showed a better result, with a 102.89% of the monolingual counterpart. This suggests that a combined query expansion before and after translation and disambiguation, improves the effectiveness of an information retrieval.

⁵ Collins Series 100 Bilingual Dictionary

⁶ <http://www.tartarus.org/~martin/PorterStemmer/>

⁷ <ftp://ftp.cs.cornell.edu/pub/smart>

Method	Mono_Fr	Mono_Eng	No_DIS	N_DIS	Feed.aft	Feed.bef_ aft	Feed.bef_ dom
Av. Prec	0.2629	0.2628	0.2214	0.2679	0.2663	0.2704	0.2725
% Mono	100	100	84.24	101.94	101.33	102.89	103.69

Feed_wn	Feed.ewn	Feed.bef_ wn	Feed.ewn_ aft	Feed.dom _wn	Feed.ew n_dom	Feed.ewn_ wn	Feed.ewn _wn_dom
0.2518	0.2579	0.2571	0.2588	0.2540	0.2545	0.2608	0.2741
95.81	98.13	97.83	98.47	96.65	96.84	99.23	104.29

Table 1. Best results and evaluations for different combinations of query translation/ disambiguation and expansion: Relevance feedback, domain-based feedback and thesaurus-based expansion

Domain-based feedback showed a drop in term of average precision comparing to previous methods. However, combined with a relevance feedback before and after trans_disambiguation, a greater result with a 103.69% in terms of average precision was deducted. Thesaurus-based expansion with WordNet *Feed_wn* or EuroWordNet *Feed.ewn* as well as combination to relevance feedback *Feed.bef_wn*, *Feed.ewn_aft* or domain-based feedback *Feed.dom_wn*, *Feed.ewn_dom* showed drops in average precision. In the other side, a combined thesauri-based expansion *Feed.ewn_wn* showed a better result but again a drop in average precision. The best result was achieved by the combined thesauri-based expansion and domain-based feedback *Feed.ewn_wn_dom* with a 104.29% of the monolingual counterpart, in term of average precision. This suggests that adding domain keywords to generalized thesauri improves the effectiveness of retrieval.

Thus, key techniques used in this successful method can be summarized as following:

- A statistical disambiguation method based on co-occurrence tendency is crucial to avoid wrong sense disambiguation and select best target translations,
- Adding domain keywords to the original query and then selecting thesaurus word senses, to avoid wrong sense disambiguation, is considered as an effective approach for the retrieval of any information,
- Each type of query expansion has different characteristics and therefore their combinations could provide a valuable resource for query expansion and showed the greatest improvement in term of average precision,

5. Conclusions and Future Works

Linguistics resources are readily available to achieve an efficient and effective query translation method in Cross-Language Information Retrieval. What we proposed and evaluated in this paper could be summarized as following: first, a query disambiguation, which is considered as a valuable resource for query translation.

Second, combined query expansion techniques before and after the translation and disambiguation methods, through a relevance feedback or domain-based feedback, has showed its effectiveness compared to the monolingual retrieval, the simple word-by-word dictionary translation or the translation and disambiguation. Third, thesauri-based with synonyms and domain-based feedbacks showed the greatest improvement to an information retrieval.

Our ongoing work involves a deeper investigation on different relations of WordNet and EuroWordNet thesauri, beside synonymy, and use multiple word senses for query expansion. An approach of learning from documents categorization or classification, not necessarily web documents, to extract relevant keywords for a query expansion, is among our future researches. Finally, our main interest is to find more effective solutions to fulfill needs of an information retrieval across languages.

References

1. Ballesteros, L. and Croft, W. B.: Phrasal Translation and Query Expansion Techniques for Cross-Language Information Retrieval. In Proceedings of the 20th ACM SIGIR Conference (1997) 84-91.
2. Dunning, T.: Accurate Methods for the Statistics of Surprise and Coincidence. Computational Linguistics, vol.19. No.1 (1993) 61-74.
3. Loupy, C., Bellot, P., El-Beze, M. and Marteau, P.-F.: Query Expansion and Classification of Retrieved Documents. In Proceedings of TREC-7. NIST Publication (1998).
4. Richardson, R., Smeaton, A.F.: Using WordNet in Knowledge-based Approach to Information Retrieval. In Proceedings BCS-IRSG Colloquium, CREWE (1995).
5. Sadat, F., Maeda, A., Yoshikawa, M. and Uemura, S.: Integrating Dictionary-based and Statistical-based Approaches in Cross-Language Information Retrieval. IPSJ SIG Notes, 2000-DBS-121/2000-FI-58 (2000) 61-68.
6. Sadat, F., Maeda, A., Yoshikawa, M. and Uemura, S.: Query Expansion Techniques for the CLEF Bilingual Track. In Proceedings of the CLEF 2001 Cross-Language System Evaluation Campaign (2001) 99-104.
7. Salton, G and McGill, M.: Introduction to Modern Information Retrieval. New York: McGraw-Hill (1983).
8. Yamabana, K., Muraki, K., Doi, S. and Kamei, S.: A Language Conversion Front-End for Cross-Linguistic Information Retrieval. In Proceedings of SIGIR Workshop on Cross-Linguistic Information Retrieval, Zurich, Switzerland (1996).
9. Voorhees, M. E.: Query Expansion using Lexical-Semantic Relations. In Proceedings of the 17th ACM SIGIR Conference (1994) 61-69.
10. Vossen, P.: EuroWordNet, A Multilingual Database for Information Retrieval. In Proceedings of the DELOS Workshop on CLIR, Zurich (1997).
11. Vossen, P.: EuroWordNet, A Multilingual Database with Lexical Semantic Networks. The Kluwer Academic Publishers (1998).