Query Expansion Techniques for the CLEF Bilingual Track

Fatiha SADAT [†], Akira MAEDA [† ††], Masatoshi YOSHIKAWA [† ‡] and Shunsuke UEMURA [†]
[†] Graduate School of Information Science, Nara Institute of Science and Technology (NAIST)
8916-5 Takayama, Ikoma, Nara. 630-0101. Japan
[‡] National Institute of Informatics (NII)
[††] CREST, Japan Science and Technology Corporation (JST)
E-mail: {fatia-s, aki-mae, yosikawa, uemura}@is.aist-nara.ac.jp

Abstract

This paper evaluates the effectiveness of a query translation and disambiguation as well as expansion techniques on CLEF Collections, using SMART Information Retrieval System.

We focus on the query translation, disambiguation and methods to improve the effectiveness of an information retrieval. Dictionary-based method with a combination to statistics-based method is used, to avoid the problem of translation ambiguity. In addition, two expansion strategies are tested on their ability to improve the effectiveness of an information retrieval, an expansion via a relevance feedback before and after translation as well as an expansion via a domain feedback after translation.

This method achieved 85.30% of the monolingual counterpart, in terms of average precision.

Keywords:

Cross-Language Information Retrieval, Query Translation, Dictionary-based Method, Disambiguation, Mutual Information, Training Corpora, Domain Keywords, Relevance Feedback.

1 Introduction

This first participation in the Cross-Language Evaluation Forum (CLEF 2001) is considered as an opportunity to better understand issues in Cross-Language Information Retrieval (CLIR) and evaluate the effectiveness of our approach and techniques. We worked on the bilingual track for French queries to English runs, with a comparison to the monolingual French and English tasks.

In this paper, we focus on the query translation, disambiguation and expansion techniques, to improve the effectiveness of an information retrieval by different combinations. Bilingual Machine Readable Dictionary (MRD) is considered as a prevalent method to Cross-Language Information Retrieval. However, simple translations tend to be ambiguous and give poor results. A combination with statistics-based approach for a disambiguation can significantly reduce the error associated with polysemy[1] in dictionary translation. Query expansion is our second interest [4]. As a main hypothesis, combination of query expansion methods before and after the query translation will improve the precision of an information retrieval as well as the recall. Two sorts of query expansion were evaluated: Relevance feedback and Domain feedback, which is an original point in this study. These expansion techniques did not show an improvement comparing to the translation method, as expected.

We have evaluated our system by using SMART Information Retrieval System (version 11.0), which is based on a vector space model, as it is considered to be more efficient than Boolean or probabilistic model.

The rest of this paper is organized as follows. Section 2 gives a brief overview of the dictionary-based method and the disambiguation method. The proposed query expansion and its effectiveness in information retrieval are described in Section 3. An evaluation and results of the conducted experiments are described and discussed in section 4. Section 5 concludes the paper.

---

[1] Polysemy is a word, which has more than one meaning.

2 Query Translation via Dictionary-based Method

Dictionary-based method, where each term or phrase in the query is replaced by a list of all its possible translations, represents a simple and an acceptable first pass for a query translation in Cross-Language Information Retrieval.

In our approach [4], a ***stopping*** phase for French queries, by using a stop list was performed to remove stop words and stop phrases and avoid the undesired effect of some terms, such as pronouns, ... Etc.

A simple ***stemming*** process of query terms was performed before the query translation, to replace each term with its inflectional root, to remove most plural word forms, to replace each verb with its infinitive form and to reduce headwords to their inflectional roots. The next step is a term-by-term ***translation*** using a bilingual machine-readable dictionary. An overview of the ***Query Translation and Disambiguation Module*** is shown in Fig 1.

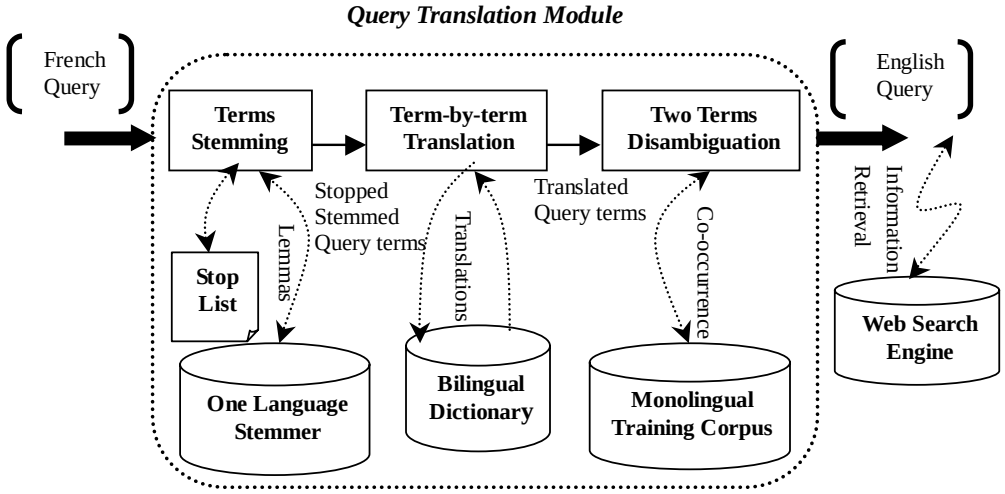

Fig 1. Query Translation and Disambiguation Module Phases

2.1 Query Term Disambiguation using Statistics-based Method

A Word is *Polysemous,* if it has senses that are different but closely related; as a noun, for example, *right* can mean something that is morally approved, or something that is factually correct, or something that is due one. In the proposed system, a disambiguation of the English translation candidates is performed, by selecting the best English term, equivalent to each French query term, by applying a statistical method based on the co-occurrence frequency. For the purpose of this study, we decided to use the *mutual information* [2], which is defined as follows:

$$MI(W_1, W_2) = Log_2 \left(\frac{N \quad f(w_1,w_2)}{f(w_1) \quad f(w_2)} \right)$$

Where N is the size of the corpus, $f(w)$ is the number of times the word w occurs in the corpus and $f(w_1, w_2)$ is the number of times both w_1 and w_2 occur together in a sentence bead.

3 Query Expansion in Cross-Language Information Retrieval

Query expansion, which modifies queries using judgments of the relevance of a few highly ranked documents, has been an important method for increasing the performance of an information retrieval.

In this study, we have proposed two sorts of query expansion: a relevance feedback before and after the translation and disambiguation of query terms, and a domain feedback after the translation and disambiguation of query terms.

3.1 Relevance Feedback before and after Translation

We apply an automatic relevance feedback, by fixing the number of retrieved documents and assuming the top–ranking documents obtained in an initial retrieval. This approach consists to add some term concepts, about 10 terms from a fixed number of the top retrieved documents (about 50 top documents), which occur frequently in conjunction with the query terms, on a presumption that those documents are relevant, to make a new query. One advantage of the use of a query expansion, such as an automatic relevance feedback is to create a stronger base for short queries in the disambiguation process, in the purpose of using co-occurrence frequency approach.

3.2 Domain Feedback

We introduced a domain feedback as a query reformulation strategy, which consists of extracting a domain field from a set of retrieved documents (top 50 documents), through a relevance feedback. Domain key terms will be used to expand the original query set.

4 Information Retrieval Evaluation

The evaluation of the effectiveness of the French-English Information Retrieval System, was performed by using the following linguistic tools :

Monolingual Corpus

The monolingual English part of the *Canadian Hansard corpus* (Parliament Debates) was used in the disambiguation process.

Bilingual Dictionaries

A bilingual French-English COLLINS Electronic Dictionary Data, version 1.0 was used for the translation of French queries to English. Missing words in the dictionary, which are essential for the correct interpretation of the query, such as *Kim, Airbus, Chiapa* was not compensated. We just kept the original source words as target translations, by assuming that missing words could be proper names, such as *Lennon, Kim*, etc…

Stemmer and Stop Words

The stemming part was performed by the English *Porter*[2] *Stemmer*.

Retrieval System

SMART Information Retrieval System[3] was used to retrieve English and French documents. SMART is a vector model, which has been used in many researches for Cross-Language Information Retrieval.

4.1 Submission for the CLEF 2001 Main Tasks

We submitted 4 runs for the bilingual (non-English) task with French as a topic language, and one run for the Monolingual French task :

Bilingual task	Language	Run Type	Priority
RunindexTR	French	Manual	1
RunindexDOM	French	Manual	2
RunindexFEED	French	Manual	3
RunindexORG	French	Manual	4
Monolingual Task			
RunindexFR	French	Manual	1

RunindexORG : The original English query topics are searched against the English Collection (Los Angeles Times 1994 : 113,005 documents , 425 MB).

RunindexTR : The original French query topics are translated to English, disambiguated by the proposed strategy and then searched against the English Collection.

---

[2] http://bogart.sip.ucm.es/cgi-bin/webstem/ stem

[3] ftp://ftp.cs.cornell.edu/pub/smart

	Precision				
Recall	**RunindexORG**	**RunindexTR**	**RunindexFEED**	**RunindexDOM**	**RunindexFR**
0.00	0.3233	0.3262	0.3242	0.3304	0.2952
0.10	0.2328	0.2002	0.1843	0.1956	0.2234
0.20	0.1606	0.1361	0.1317	0.1219	0.1695
0.30	0.1292	0.1066	0.1059	0.0974	0.1432
0.40	0.1074	0.0925	0.0864	0.0764	0.1240
0.50	0.0881	0.0731	0.0653	0.0593	0.1059
0.60	0.0714	0.0586	0.0497	0.0481	0.0867
0.70	0.0572	0.0448	0.0373	0.0376	0.0696
0.80	0.0357	0.0295	0.0203	0.0244	0.0590
0.90	0.0250	0.0164	0.0104	0.0117	0.0401
1.00	0.0126	0.0092	0.0063	0.0066	0.0166
Avg. Prec	**0.1014**	**0.0865**	**0.0798**	**0.0780**	**0.1120**
% English Monolingual	**100**	**85.30**	**78.69**	**76.92**	**--**
% French Monolingual	**--**	**77.23**	**71.25**	**69.64**	**100**

Table 1. Results of the Submitted CLEF Bilingual and Monolingual Runs

RunindexFEED : The original French query topics are expanded by a relevance feedback, translated and disambiguated by the proposed method and again expanded, before the search against the English data collection.
RunindexDOM : The translated disambiguated query topics are expanded by a domain feedback, after translation and searched against the English Collection.
RunindexFR : The original French query topics are searched against the French Collection (Le Monde 1994 : 44,013 documents, 157 MB and SDA French 1994 : 43,178 documents , 86 MB).

Query topics were constructed manually by selecting terms from fields <title> and <description> of the original set of queries.

4.2 Results and Performance Analysis

Our participation in CLEF 2001 showed two runs, which contributed to the relevance assessment pool: RunindexTR, the translation and disambiguation method, and RunindexFR, the monolingual French retrieval. The rest of bilingual runs were not judged, because of limited evaluation resources, the result files did not directly contribute to the relevance assessment pool. However, the runs were subject to all other standard processing, and are still scored as official runs. Table 1 shows the average precision for each run.

4.3 Discussion

In our previous research [4], we tested and evaluated two types of feedback loops: a combined relevance feedback before and after translation and a domain feedback after translation. In terms of average precision, we noticed a great improvement of the two methods, comparing to the translation-disambiguation method. As well, the proposed translation-disambiguation method showed an improvement, comparing to a simple dictionary translation method. As a conclusion, a disambiguation method improved the average precision. Moreover, query expansion via the two types of feedback loops, showed greater improvement [4].

However in this study, the submitted bilingual runs did not show any improvement in terms of average precision for a query expansion via a relevance feedback before and after translation or a domain feedback after translation. The proposed translation and disambiguation method RunindexTR, achieved 85.30% accuracy of the monolingual performance RunindexORG (English information retrieval) and 77.23% of the monolingual performance RunindexFR (French information retrieval). This accuracy is higher than that of relevance feedback or domain feedback, as shown in Table 1. The relevance feedback before and after translation RunindexFEED showed a second best result in term of average precision, 78.69% and 71.25% of the monolingual English and

French retrieval, respectively. RunindexDOM, the domain feedback showed a less effective result in terms of average precision, with 76.92% and 69.64% of the monolingual English and French retrieval, respectively. Fig 2 shows the precision-recall curves for the submitted runs to CLEF 2001.

These results were less effective than we expected. This is because; first the bilingual dictionary does not cover technical terms and proper nouns, which are the most useful in improving the total IR accuracy. In this study, the French query set contains 11 untranslated English terms. Using Collins French-English Bilingual dictionary as the only resource for term translations puts the burden of discovering the right translation of proper nouns, which are used in Clef query set. When a word is not found in the dictionary, we just kept the original source word as a target translation one. This method should be successful for some proper names, such as *Lennon, Kim*, etc…but not for others, such as *Chiapa*. Missing words in the bilingual dictionary is one major reason to introduce noise into our results. The second reason is due to the selection of terms to expand the original queries, domain keywords for a domain feedback or terms that occur most often with the original query terms. This made the expansion methods ineffective, comparing to our previous work [4]. We hope to be able to improve the effectiveness of an information retrieval, as explained below:

Bilingual Translation

The bilingual dictionary should be improved to cover all terms described in the original queries. One solution to this problem is to extract terms and their translations through parallel or comparable corpora (non-parallel), and extend the existing dictionary with that terminology, in Cross-Language Information Retrieval.

Terms Extraction for a Feedback Loop

According to previous researches [1] [4], query expansion before and after translation improves the effectiveness of an information retrieval. In our case, we used the mutual information [2] to select and add those terms, which occur most often with the original query terms. Previous results showed that results based on the mutual information are significantly worst that those based on the log-likelihood-ratio or chi-square test or modified dice coefficient [3]. For an efficient use of the term co-occurrence frequency in the relevance feedback process, we will select the log-likelihood-ratio for further experiments.

Domain Feedback

A combination of the proposed domain feedback method to a relevance feedback before or after translation could be a solution, to improve the effectiveness of an information retrieval.

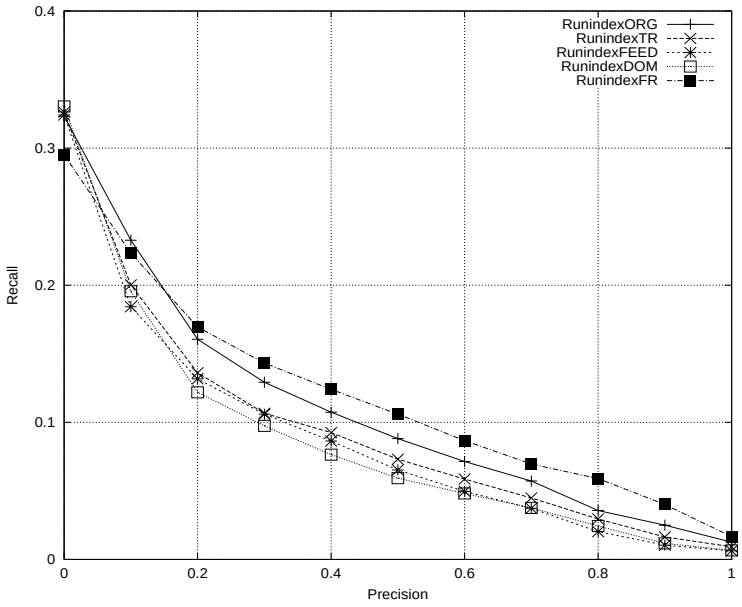

Fig 2. A Precision-Recall curves for the submitted Bilingual and Monolingual runs

5 Conclusion

Our conclusion at this point can only be partial. We still need to perform more experiments to evaluate the proposed query expansion methods. The purpose of our investigation was to determine the efficacy of translating and expanding a query by different methods. The study compares the retrieval effectiveness using the original monolingual queries, translated and disambiguated queries and the alternative expanded user queries on a collection of 50 queries, via a relevance feedback or a domain feedback techniques. An average precision measure is used as the basis of the experiments evaluation.

However, the proposed translation and disambiguation method showed the best result in terms of average precision, comparing to the query expansion methods: via a relevance feedback before and after query translation and disambiguation and via a domain feedback after query translation and disambiguation.

What we presented in this paper is a rather simple study, which highlights some areas in Cross-Language Information retrieval. We hope to be able to improve our researches and find more solutions to fulfill the needs for Information retrieval cross languages.

Acknowledgment

This work is partially supported by the Ministry of Education, Culture, Sports, Science and Technology, Japan, under grants 11480088, 12680417 and 12208032, and by CREST of JST (Japan Science and Technology).

References

[1] Ballesteros, L. and Croft, W. B. : "Phrasal Translation and Query Expansion Techniques for Cross-Language Information Retrieval". Proceedings of the 20th ACM SIGIR Conference, (1997). P 84-91.

[2] Gale, W. A. and Church, K. : "Identifying word correspondences in parallel texts". Proceedings of the 4th DARPA Speech and Natural Language Workshop, (1991). P.152-157.

[3] Maeda, A., Sadat, F., Yoshikawa,M. and Uemura, S. : "Query Term Disambiguation for Web Cross-Language Information Retrieval using a Search Engine". Proceedings of the 5th International Workshop on Information Retrieval with Asian Languages, (Oct 2000). P 25-32.

[4] Sadat, F., Maeda, A., Yoshikawa, M. and Uemura, S. :" Cross-Language Information Retrieval via Dictionary-based and Statistical-based Methods". Proceedings of the 2001 IEEE Pacific Rim Conference on Communications, Computers and Signal Processing (PACRIM'01), (August 2001).