情 報 学 基 礎 66－11
デジタル・ドキュメント 32－11
（２００２．３．１５）

Applying Multiple Resources for Query Expansion in Cross-Language Information Retrieval

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

In this paper, we focus on query expansion techniques to improve the effectiveness of an information retrieval. A combination to the dictionary-based translation and statistics-based disambiguation approaches is indispensable to overcome query translation ambiguity.

We propose a model using multiple sources for query reformulation, organization, translation and disambiguation to select one target query term and retrieve the information needed by a user. Relevance feedback or 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 tested the effectiveness of the proposed combined method, by an application to a French-English Information Retrieval. Experiments using the TREC data collection proved a great effectiveness of the proposed disambiguation and combined query expansion techniques.

Keywords

Cross-Language Information Retrieval, Query Translation, Query Terms Disambiguation, Query Expansion, Relevance Feedback, Domain Keywords, Thesaurus.

言語横断情報検索における複数の手法による問合せ拡張の適用

ファティアサダト[†]，亮前田[†††]，正俊吉川[†‡]，俊亮植村[†]
[†] 奈良先端科学技術大学院大学 情報科学研究科
〒630-0101 奈良県生駒市高山町 8916-5
[‡] 国立情報学研究所 (NII)
[††] CREST，日本科学および技術株式会社(JST)
{fatia-s, aki-mae, yosikawa, uemura}@is.aist-nara.ac.jp

概要

本稿では，言語横断情報検索の性能向上のための問合せ拡張手法について述べる問合せ翻訳の曖昧性を解消するには，辞書ベースの翻訳と統計ベースの曖昧性解消アプローチの組み合わせが不可欠である．我々は，検索対象言語における適切な検索語を選択するために，問合せの再構成・組織化・翻訳・曖昧性解消を複数資源を用いて行うモデルを提案する．適合フィードバック，シソーラスベースの拡張，さらに利用者の問合せを拡張するドメインキーワードの抽出による新たなフィードバック手法を提案し，評価を行う．フランス語－英語間の検索への応用によって，提案する統合手法の検索性能を確認した．

TREC コレクションを用いた実験において，提案する曖昧性解消と問合せ拡張手法は大幅な性能の向上を確認した．

キーワード

言語横断情報検索，問合せ翻訳，問合せ曖昧性解消，問合せ拡張，適合フィードバック，ドメインキーワード，シソーラス。

1 Introduction

With the growing number of linguistic resources accessible through the World Wide Web (Internet), an information retrieval became such a crucial task to fulfill user's needs. Cross-Language Information Retrieval CLIR, consists of providing a query in one language and searching document collections in one or multiple languages. In this research, we focus on the query translation, rather than on documents translation, which is considered as an unrealistic translation task because of the huge amount of documents to manage. The disambiguation of target translations is a second goal as well as the query expansion to improve the effectiveness of an information retrieval by different combinations. The proposed study is general across languages in information retrieval however; we have conducted experiments and evaluations on a French-English information retrieval.

The rest of this paper is organized as follows. Section 2 gives an overview of the proposed translation/disambiguation approach in Cross-Language Information Retrieval. The proposed query reformulation and expansion 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 Query Terms Translation / Disambiguation

In this research, query translation is performed by a dictionary-based method [5], followed by the disambiguation of translation candidates to select one target translation and retrieve the needed information.

2.1 Dictionary-based Translation

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 at Cross-Language Information Retrieval [4].

In our approach [6], query translation is performed after a simple ***stemming*** process of query terms to replace each term with its inflectional root, to remove most plural word forms, to replace each verb with its infinitive form and to remove stop words and stop phrases. The next step is a term-by-term ***translation*** using a bilingual machine-readable dictionary [5]. 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 the dictionary translation.

2.2 Statistics-based Disambiguation

In the proposed system, a statistical disambiguation method for target query terms is performed, by selecting one best translation, equivalent to each source query term. A co-occurrence frequency: ***Mutual information*** [2][3][6] or ***Log-Likelihood Ratio***, among other estimations could be used. The process of disambiguation is described as following:

1. First an *organization* of source query terms per pairs is necessary,
2. Co-occurrence frequencies will be computed for each combination of source terms,
3. The pair of terms with a highest mutual co-occurrence frequency will be selected for a disambiguation,
4. Next a *translation* of these terms is completed,
5. A second statistical *disambiguation* by co-occurrence tendency is performed to select the best target translation for each source query terms,
6. Go to the next combination with the next highest co-occurrence tendency and repeat steps 3 to 5 until all translations of the source query terms are fixed.

An overview of the proposed translation / disambiguation module, which is a part of the Cross-Language Information Access System is described in Figure 1.

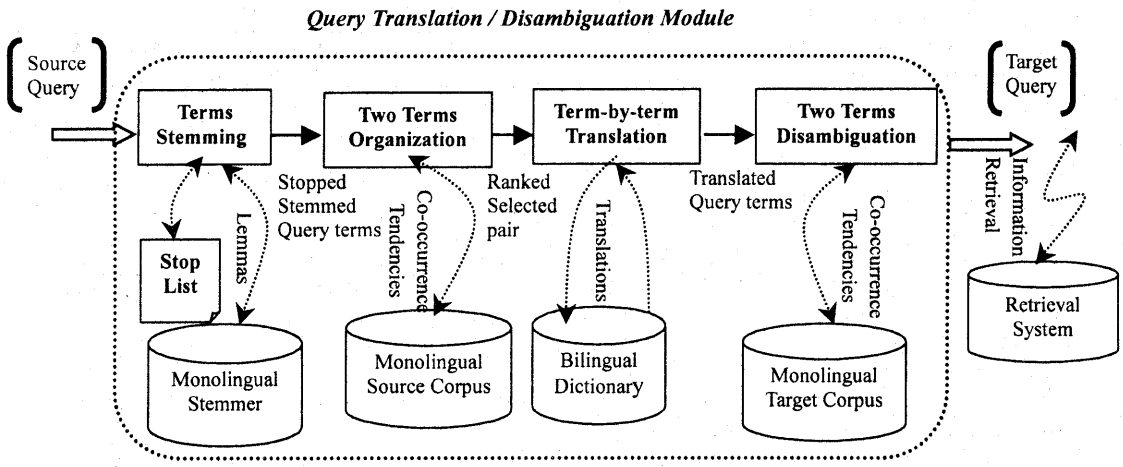

Figure 1. The Translation / Disambiguation Module in Cross-Language Information Retrieval

3 Query Reformulation / Expansion

Query reformulation through an automatic terms expansion is considered as one of the most important method in increasing the performance of an information retrieval. In this research, extracting, selecting and adding terms that emphasize query concepts is performed prior and after the query translation and disambiguation. 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. We apply a combined query expansion prior and after the translation and disambiguation processes. This combined method will reduce the ambiguity by de-emphasizing irrelevant terms added by translation and will improve a precision as well as a recall of an information retrieval.

We propose three sorts of query expansion to modify user's query and retrieve needed documents: a relevance feedback, with its best selected terms, a domain-based feedback which is characterized by an extraction of domain keywords and a thesaurus-based query expansion with a retrieval of synonyms and multiple words senses.

3.1 Relevance Feedback

Relevance feedback is applied as a query expansion, by fixing the number of retrieved documents and assuming the top ranked ones obtained. A fixed number of term concepts, will be extracted from the fixed number of top retrieved documents, with an assumption that these terms occur frequently in conjunction with the original query terms. Finally, these terms are used for the query reformulation. One advantage of the use of a query expansion, such as an automatic relevance feedback is to create a stronger base for short queries, helpful especially in the statistical disambiguation process, when using the co-occurrence frequency estimation.

3.2 Domain-based Feedback

We introduced a domain-based feedback as a query reformulation strategy, which consists of extracting domain keywords from a set of top retrieved documents, through a classical relevance feedback. Domain keywords will be used to expand an original query set [7].

3.3 Thesaurus-based Expansion

We propose the use of a thesaurus for a query reformulation and expansion, where the top fixed and ranked items related to a particular query are retrieved from a monolingual thesaurus and added to the original query. These similar term concepts are used as candidates to emphasize the query, before the retrieval of documents. However, two kinds of expansion

terms could be applicable, synonyms that are very close to the query and multiple word senses that describe special lexical or semantical relationships to the query terms. Machine-readable thesauri such as Roget's or WordNet[1] in the case of English, or the French part of the EuroWordNet[2] can be seen as powerful tools for studying lexical semantic resources and their language-specificity. They can be fundamentally applicable to the related languages. WordNet, an online lexical database reference for English, whose design is inspired by current psycholinguistic theories of human lexical memory, has been very successful for Monolingual and Cross Language Information Retrieval. The system has the power of both an on-line thesaurus and an on-line dictionary. WordNet's basic object is a set of strict synonyms called *synsets* for English nouns, verbs, adjectives and adverbs, each representing one underlying lexical concept [8][9]. Our suggestion is that words with a semantic relation, with the original English queries terms, can be added and thus form an expanded English query. Following the research reported by Voorhees [8], on the use of lexical semantics relations of WordNet for a query expansion, we can integrate and combine thesauri before and after translation, as a combined query expansion. The process is performed by a simple look up in the thesaurus or lexical database, to find synsets of the full query, such as the original query: "*Defense System*". Expansion candidates would be "*Weaponry*" or "*Arm*". However, we proceed by a term-by-term search, in case of non-existence of the full query in the lexical database. Also, a simple one-term query can be represented by a compound synonym (containing more than one term). In this case, a conjunction between simple terms of the concerned synonym would be possible. An example is a simple term "*war*" which will be expanded by a compound synonym "*military action*" and replaced by "*war military action*". Figure 2 and 3 show an example of a query expansion with synonyms via WordNet and EuroWordNet thesauri respectively, where a selection of synonyms is based on the highest frequency in conjunction

Orig Query	*WordNet thesaurus synonyms (Ordered by frequency)*
Doctor	Physician Dr. Medico Theologian ...
Drug	Medicine Narcotic Artifact ...
Cure	Remedy Curative Cure ...
Expanded Query	"Doctor Drug Cure Physician Medicine Remedy"

Figure 2. WordNet-based Query Expansion

Orig Query	*EuroWordNet thesaurus synonyms (Ordered by frequency)*
Demander	Poser Questionner Interroger
Questionnaire	Epreuve Composition Examen Interrogation
Expanded Query	"Demander Questionnaire Poser Epreuve "

Figure 3. EuroWordNet-based Query Expansion

with original query terms. One advantage of using WordNet or EuroWordNet thesauri is the existing synonyms order by frequency in the lexical databases.

4 Experiments and Results

The evaluation of proposed methods, based on a query translation / disambiguation and based on query expansion through different combinations, was completed through the following steps:

4.1 Linguistic resources

The following linguistic resources were used to conduct experiments and evaluate results:

Test Data

We used TREC[3] test collection 1. Topics 63-150 queries were considered as English queries, for the conducted experiments and were composed

[1] http://www.cogsci.princeton.edu/~wn/

[2] http://www.hum.uva.nl/~ewn/docs.htm

[3] http://trec.nist.gov/data.html

of several fields. Tags <num>, <dom>, <title>, <desc>, <smry>, <narr> and <con> denote topic number, domain, title, description, summary, narrative and concepts fields. Key terms contained in the field of title <title>, which are averaged 2.8 terms by query, were used to generate English and French queries.

Monolingual Corpora

The Canadian *Hansard* corpus (Parliament Debates) is a bilingual French-English parallel corpus. It contains more than 100 million words of English text and the corresponding French translations. In this study, we have used the Hansard as monolingual corpora for French and English languages.

Bilingual Dictionary

COLLINS French-English dictionary was used for the translation of source queries.

Thesauri

WordNet [8] and *EuroWordNet* [9][10], lexical databases were used for the thesaurus-based query expansion.

Stemmer and Stop Words

Stemming part was performed by the English *Porter*[4] *Stemmer*. A special French stemming was developed and used in these experiments.

Retrieval System

NAMAZU [5] retrieval system (version 2.0.1) was used to retrieve English documents. NAMAZU is a freeware full text retrieval system based on Boolean model and supports basic functions such as a composition of Boolean operators, results ranking and phrasal retrieval [9].

4.2 Conducted Experiments

We have conducted two types of experiments: Based on Query Translation / Disambiguation (QT/D) and based on Query Expansion (QE) with different combinations.

4.2.1 Experiment 1 (based on QT/D)

This part concerns the evaluation of the proposed disambiguation method for short and long queries, combined to the translation.

A monolingual retrieval with original English queries is represented by ***ORIG*** method. The constructed queries were evaluated by the following methods: ***HUM*** for the human French-English translation method, ***AMB-AND*** is a result of using no-disambiguation or using all possible translations for each query term, obtained from the bilingual dictionary. ***DIS-AND*** is the result of the translation and the disambiguation of consecutive pairs of source terms, by applying the classical disambiguation process without neither ranking nor selection. ***N.DIS-AND*** is the result of the proposed disambiguation method: Ranking source query terms, dictionary-translation and statistical disambiguation to select one best target translation. In these evaluations, log-likelihood ratio was used as a co-occurrence tendency for the disambiguation as well as for ranking strategies. Boolean operators (*and* / *or*) were used to combine terms of the query and retrieve documents using the Boolean retrieval system. Results and evaluations are shown in Table 1 and Figure 4.

4.2.2 Experiment 2 (based on QE)

Evaluations on different query expansion combinations was completed by the following methods:

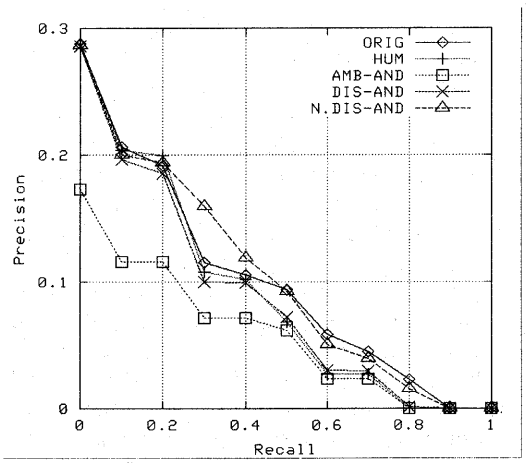

Figure 4. Recall-Precision Curves for QT/D experiments

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

[5] http://www.namazu.com

FEED_B, represents the result of a relevance feedback before the translation and disambiguation, ***FEED_A*** is a result of query translation, disambiguation and then expansion through a relevance feedback. ***FEED_B&A-AND*** is a result of combined query expansion before and after the translation and disambiguation by an *and* Boolean operator. ***FEED_B&A-OR*** is a similar combination with an *or* operator. Using a domain-based feedback after translation was a result of the ***FEED_D*** method. Combined strategies with the relevance feedback were completed by the following methods: ***FEED_B&D-AND*** and ***FEED_B&D-OR*** for a combination using *and* / *or* Boolean operators consecutively. Results and evaluations are shown in Table 2 and Figure 5.

Thesaurus-based expansion was completed with a fixed number of synonyms, ordered by their mutual frequencies in conjunction with query terms. ***FEED_EW-AND/ FEED_W-AND*** are related to expansion methods with EuroWordNet and WordNet synonyms before or after translation respectively with a combination by an *and* Boolean operator. Boolean *or* operator is used as well for the combined query expansion and methods ***FEED_EW-OR / FEED_W-OR***. Relevance and domain-based feedbacks were evaluated as well by a combination to the thesaurus-based expansion. Results of these evaluations are shown in Table 3 and Figure 6.

4.3 Discussion

The purpose of this investigation is to determine the efficacy of query expansion, by different combinations and strategies. An average precision measure is used as the basis of the evaluation. The proposed disambiguation method ***N.DIS*** has shown an improvement in term of average precision by 81.82% and 90.42% for the combination by Boolean operators *or* / *and* respectively, of the monolingual retrieval. Combined query expansion via a relevance feedback before and after translation showed a good improvement with an average precision of 91.95% and 96.19% of the monolingual performance for combinations with *and/or* operators respectively. A performance improvement occurred because better query terms were added to the translated query terms. Those terms tend to refine the original query and de-emphasize inappropriate

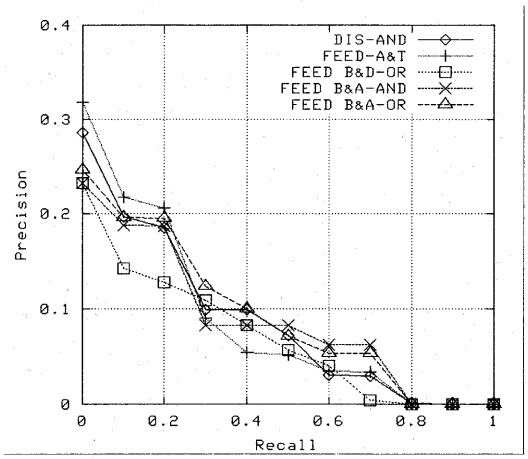

Figure 5. Recall-Precision Curves for QE experiments via Feedback loops

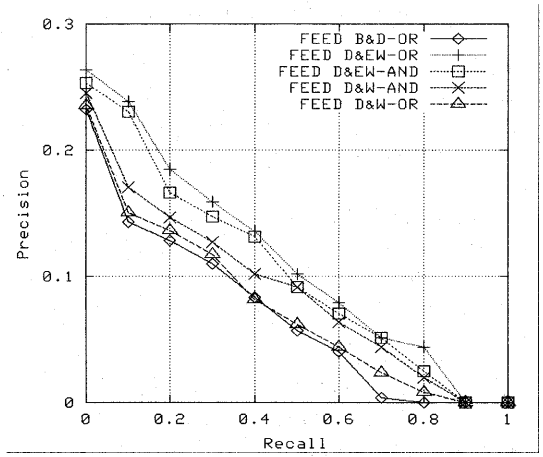

Figure 6. Recall-Precision Curves for QE experiments via Thesaurus Synsets

definitions. Feedback with domain keywords combined to the relevance feedback was helpful to the average precision with 99.13% of the monolingual performance. As for a thesaurus-based expansion, the best result is related to the combined domain-based feedback and thesaurus-based expansion, with a 99.56 % of the monolingual counterpart, in term of average precision. This suggests that adding domain keywords to the thesaurus-based expansion reduces the negative effects of ambiguity caused by inappropriate term definitions, extracted from the thesaurus and thus makes special selection of expansion terms

Method	ORIG	HUM	AMB-AND	AMB-OR	DIS-AND	DIS-OR	N.DIS-AND	N.DIS-OR
Avg. Prec	**0.0919**	0.0839	0.0623	0.0745	**0.0798**	0.0642	**0.0831**	**0.0752**
% Mono	**100**	91.29	67.79	81.06	**86.83**	69.85	**90.42**	**81.82**

Table 1. Evaluation of Translation / Disambiguation Combinations

Method	Fee_A	Feed_D	Feed_B	Feed_B& A_AND	Feed_B& A_OR	Feed_B& D_AND	Feed_B& D_OR	Feed_D& A_AND	Feed_D& A_OR
Avg.Prec	0.0829	0.0697	0.0793	**0.0845**	**0.0884**	0.0650	**0.0874**	0.0688	**0.0911**
% Mono	90.20	75.84	86.29	**91.95**	**96.19**	70.72	**95.10**	74.86	**99.13**

Table 2. Evaluation of Feedback Strategies and Combinations

Method	Feed_W_AND	Feed_W_OR	Feed_B& W_AND	Feed_B& W_OR	Feed_D& W_AND	Feed_D &W_OR	Feed_D& EW_AND	Feed_D& EW_OR
Avg. Prec	0.0497	0.0592	0.0281	0.0390	**0.0887**	**0.0875**	**0.0915**	**0.0856**
% Mono	54.08	64.41	30.57	42.43	**96.51**	**95.21**	**99.56**	**93.14**

Table 3. Evaluations of Feedback and Thesaurus-based Combinations

5 Conclusion

In this paper, we have presented a new model in Cross-Language Information Retrieval. We demonstrated by combining existing approaches to CLIR and Machine Translation, it is possible to improve the effectiveness of an information retrieval and thus the overall system. Combined techniques before and after translation through a domain-based feedback and thesaurus-based expansion, have proved its efficiency on the simple word-by-word dictionary translation.

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).

Reference

[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). P 84-91.

[2] Church, K. W. and Hanks, P. : "Word association Norms, Mutual Information and Lexicography". Computational Linguistics, Vol 16 No1, (1990). P 22-29.

[3] 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.

[4] Grefenstette, G. : " Cross-Language Information Retrieval". The Kluwer International Series on Information Retrieval, Vol. 2, Kluwer Academic Publishers, (1998).

[5] Hull, D. and Grefenstette, G. : "Querying across languages. A Dictionary-based Approach to Multilingual Information Retrieval". In proceedings of the 19th ACM SIGIR Conference, (1996). P49-57.

[6] 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. P 61-68.

[7] Sadat, F., Maeda, A., Yoshikawa, M. and Uemura, S. :"Query Expansion Techniques for the CLEF Bilingual Track". Proceedings of the CLEF 2001 Cross-Language System Evaluation Campaign, (2001). P.99-104.

[8] Voorhees, M. E. : "Query Expansion using Lexical-Semantic Relations". Proceedings of the 17th ACM SIGIR Conference, (1994). P61-69.

[9] Vossen, P. : "EuroWordNet, A Multilingual

Database for Information Retrieval". Proceedings of the DELOS Workshop on Cross-language Information Retrieval, March 5-7, (1997), Zurich.

[10] Vossen, P. : "EuroWordNet, A Multilingual Database with Lexical Semantic Networks". The Kluwer Academic Publishers (1998).