データベースシステム 121－10
情 報 学 基 礎 58－10
（２０００．５．２５）

Integrating Dictionary-based and Statistical-based Approaches 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)
〒 630-0101. 8916-5 Takayama, Ikoma, Nara, Japan
E-mail: {fatia-s, aki-ma, yosikawa, uemura}@is.aist-nara.ac.jp

Abstract

As Internet resources become accessible to more and more countries, there is a need to develop methods for Cross Language Information Retrieval for European and Asian languages such as English, French, Japanese, Chinese, Arabic, ... Most Cross-Language Information Retrieval researches have focused on dictionary-based method with a combination to statistical approach to avoid the problem of ambiguity. Query expansion is used to improve precision and recall in the process of information retrieval and to dramatically reduce the errors such an approach normally makes. In this paper, we discuss the approach of using a bilingual dictionary, by including an automatic feedback loop in the task of query expansion before translation. A French-English similarity thesaurus is applied after the disambiguation of the translated candidates. We apply this method to a French-English module, in the Multilingual Knowledge Discovery System. A combination of French-English and English-Japanese modules is in prospect.

言語横断情報検索における辞書ベースと統計ベースのアプローチの統合

サダト ファティア，前田 亮，吉川 正俊，植村 俊亮
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

概要

インターネットがより多くの国々で利用可能になるのに伴い，ヨーロッパ言語とアジア言語との間の言語横断情報検索手法が必要になってきている．言語横断情報検索の研究の多くは，辞書ベースの手法に曖昧性を解消するための統計的アプローチを組み合わせるものである．また，適合率および再現率の向上のためと，このような手法によって起きる誤訳を減少させるために，問合せ拡張が用いられる．本稿では，翻訳前の問合せ拡張のために自動適合性フィードバックを用いる対訳辞書ベースのアプローチについて述べる．また，訳語の曖昧性解消の後に，類似性シソーラスによる問合せ拡張が適用される．この手法は，従来から我々が研究を進めている「多言語知識発掘システム」中のフランス語-英語 モジュールとして用いられる．また，今後フランス語-英語，英語-日本語の両モジュールの結合を計画している．

1. Introduction

Interest in Cross-Language Information Retrieval (CLIR) has grown rapidly in recent years. The diversity of information sources and the explosive growth of the Internet Worldwide are compelling evidence of the need for IR systems that can cross language boundaries. Cross-Language Information Retrieval consists of providing a query in one language and searching document collections in one or multiple languages. One can envision many ways to bridge the language barrier between query and collection.

In this paper, we focus on the query translation, disambiguation and methods to improve the effectiveness of information retrieval. We are interested in finding methods for performing cross-lingual retrieval, which do not rely on scarce resources such as parallel corpora. Bilingual Machine Readable Dictionaries (MRDs), more prevalent than parallel texts seem to be a good alternative. However, simple translations tend to be ambiguous and give poor results. A combination with statistical approach for a disambiguation can significantly reduce the error associated with polysemy[1] in dictionary translation. Historically, statistical approach has been viewed as operationally impractical because, while it has helped recall, it generally hurts precision at the top of the result list.

Our main hypothesis is that query expansion via relevance feedback will improve the precision of information retrieval, while a similarity thesaurus will improve the recall. We apply a statistical approach to reduce the errors that a simple dictionary translation will produce. The query expansion is based on a local feedback before the query translation and on similarity thesaurus after the query translation, to improve information retrieval effectiveness. The rest of this paper is organized as follows. Section 2 gives a brief overview of related work. The multilingual knowledge discovery system is described in Section 3. The overview of the proposed information retrieval subsystem and the dictionary-based method with disambiguation are described in Section 4. Query expansion and its effectiveness in information retrieval are described in Section 5. Section 6 concludes the paper.

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

2. Previous works in CLIR

Work on Cross-Language Information Retrieval dates back to the early seventies when Salton [13] [14], established that performance of English-French cross-language retrieval was comparable to the performance of monolingual retrieval when manually developed resources were employed for query translation. 20 years after these original experiments, the availability of linguistic resources has showed the possibility to attempt multilingual retrieval with minimal manual intervention. Current approaches to Cross-Language Information Retrieval may be classified usefully into three classes: Those which use Machine Readable Dictionary (MRD) translation, those which rely on parallel or comparable corpora and those which employ existing linguistic resources. The problem with using parallel or comparable texts is that test corpora are costly to acquire and not readily available, it is difficult to find already existing translations of the right kind of documents and translated versions are expensive to create. For this reason, there has been more interest recently in the potential of knowledge-based technology. Automatic machine-readable dictionary's query translation, on its own, has been found to lead to a drop in effectiveness of 40-60 % of monolingual retrieval (Hull and Grefenstette [8], Ballesteros and Croft [2]). Yamabana [17],

Ballesteros [2] and Hull [9] have used machine translation approach successfully, for query translation and information retrieval. On the other hand, techniques to improve the precision-recall of information retrieval, such as query expansion via relevance feedback have been used in many researches [2]. A thesaurus such as WordNet is being used extensively in word similarity measurement technique, to enhance the information extraction systems [18]. A thesaurus is a type of an ontology, specialized to the organization of terminology for a language. There are now a number of thesaurus-based systems available commercially. Fujii [5] claims that better result can be obtained by the integration of the two approaches: statistics-based for word sense disambiguation and thesaurus-based approaches. This combined technique is evaluated as an effective method in its application to a word sense disambiguation task [18]. However, most of researches on Cross-Language Information Retrieval were concentrated on European languages while integrating Asian languages is neglected.

3. The Multilingual Knowledge Discovery System (MKDS)

We initiated the idea of a multilingual information system for collecting, indexing and retrieving documents [12]. Some parts such as the user interface and the query translation for Japanese and English are implemented, while other parts are substituted by an existing search engine. Currently, we are working on the integration of a French-English module while other modules combining Asian and European languages such as a Japanese-French module, will be added to the MKD system. The integration of new languages to the system, such as Arabic, is in prospect. An overview of the actual multilingual knowledge discovery system is shown in Fig 1.

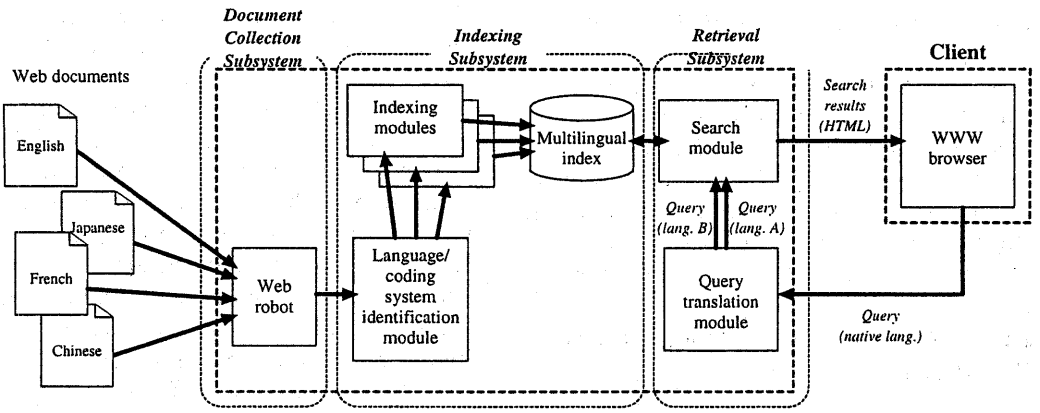

Fig 1. Overview of the Multilingual Knowledge Discovery System

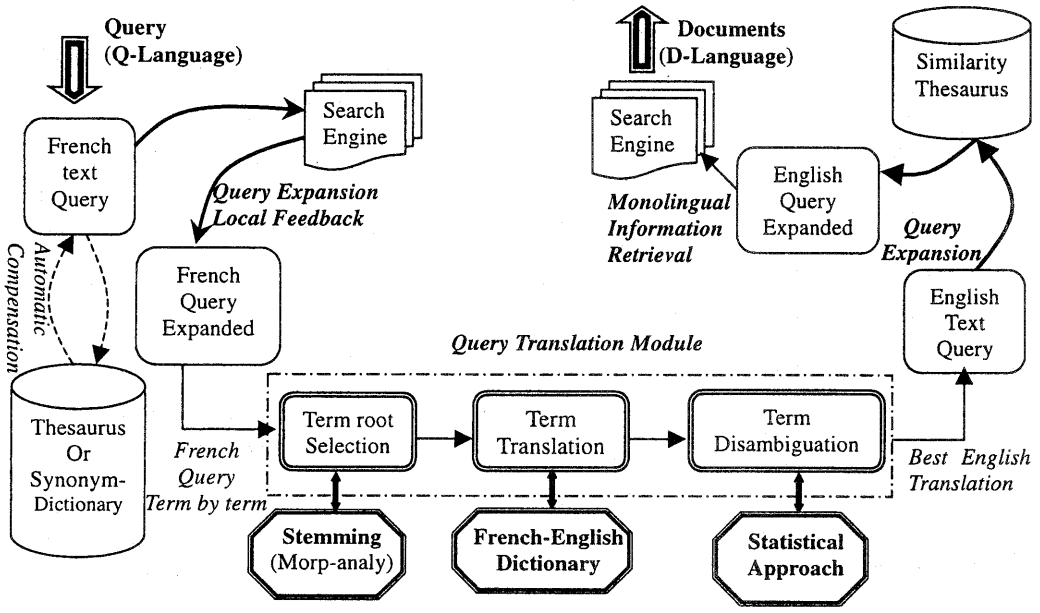

Fig 2. Overview of the proposed French-English Information Retrieval Subsystem.

4. Overview of the proposed Information Retrieval Subsystem

The Information Retrieval is a part of the multilingual knowledge discovery system.
First, user's requests are translated to the language the user specified, then documents are retrieved using a search engine and returned to the user. An overview of the proposed retrieval system is shown in Fig 2.

4.1 The Dictionary-based Method for Query Translation

Cross-Language Information Retrieval requires some form of query translation. It has been shown that dictionary-based method, where each term or phrase in the query is replaced by a list of all its possible translations, represents an acceptable first pass at cross-language information retrieval although such simple methods clearly show performance below that of monolingual retrieval [7].

In our study, query translation is performed after 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 term list built from a bilingual machine-readable dictionary. However, there are two factors that limit the performance of such approach. First is the limitation of general-purpose dictionaries, especially for specialized vocabulary. The second factor is related to the presence of spurious translations or polysemy, which is further discussed in Section 4.2. 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 *thesaurus. This case* requires an extra step of looking up the query term in the synonym dictionary, when missing words in the bilingual machine-readable dictionary, to find the equivalent terms or the synonyms of the concerned term in the query, before query translation. When using a thesaurus, the task of finding equivalent terms in the target language is identical to finding synonyms, in fact it is looking for synonyms but in the same language. Consider for instance, lexical items or lemmas in one language that do not have equivalents in another. In order to provide an exact translation, we must find the synonyms of this lemma or term and then translate it to the target language. An example is *canapé*, a term in French which doesn't have an equivalent English translation in Collins Bilingual French-English Dictionary. The closest lexical term is *meuble*, which means *furniture* in English. A retrieval system looking for *canapé* in this case will retrieve documents related to the term *meuble.*

4.2 Using Statistical approach for a disambiguation

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. This word is considered as ambiguous.
The standard problem caused by polysemy when retrieving by exact term matching is that, besides pertinent documents, we will also retrieve others unsuitable to the query. The expansion of ambiguation makes cross-language retrieval much harder than its monolingual counterpart.

In our system, disambiguation of the English translation candidates was performed, by selecting the best English term, equivalent to each French query term, by applying a statistical process called "Co-occurrence Frequency" [16]. There are some possible estimation functions based on co-occurrence frequency: Mutual Information method [4], Gale's method [6] and Kay's method [10].

Estimation Function 1

If two elements often co-occur in the corpus, then they have a high chance to be the best translations among the candidates of the queries term. This estimation uses *mutual information* as the Metric for significance of word co-occurrence frequency [4]. It is defined as follows:

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

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.

Estimation Function 2

Gale's method [6] has been proposed as an alternative to MI and will tend to favor high frequencies more. This is calculated by the phi-squared (ϕ^2) as follows:

$$\phi^2 = \frac{(ad-bc)^2}{(a+b)(a+c)(b+d)(c+d)} \quad \text{where } 0 \le \phi^2 \le 1$$

Where the feature a is the number of times the two words occur in a sentence bead. The feature b is the number of times the first word occurs but the second does not. Similarly, c is

the number of times the second word occurs but the first does not. Finally, *d* is the number of times both words do not occur in the sentence bead.

Estimation Function 3

The validity of word correspondence w_1 and w_2 is estimated, depending on Kay's method [10] as follows :

$$H(w_1, w_2) = \frac{2\, f(w_1, w_2)}{f(w_1) + f(w_2)}$$

Two factors limit the statistical approach. The first one concerns the choice of a corpus and the second factor concerns the case of short queries of one word for information retrieval. Incorporating the Hansard[2] corpus [7], which is one of the only existing corpora for both French and English, can be considered as a solution to the statistical approach problem. However, huge amount of text is necessary to acquire an effective knowledge, since the co-occurence tendency is based on examples and statistics, the larger the corpus size, the more accurate the translation quality. Is there a corpus larger than the Internet Natural Language? The proposed method treats the WWW homepages as though they are a constructed-corpus stored in a local machine. For this purpose, collecting documents from the WWW, by using a web robot is the best alternative to make a corpus.

Queries tend to be very short (case of one headword) which is inconvenient for the co-occurence tendency measure. Local feedback, before translation is a method to expand the query. Specifically, the query will be modified using information derived from documents whose relevance to the query is known. Typically terms found in the relevant documents are added to the query.

[2] Hansard is a bilingual French-English parallel corpus, from the Proceedings of the Canadian Parliament.

5. Query Expansion

Relevance feedback, which modifies queries using judgements of the relevance of a few highly-ranked documents, has historically been an important method for increasing performance of information retrieval systems. We apply a local feedback before translation for query expansion. Following the research reported by Ballesteros and Croft [2], on the use of relevance feedback, adding terms that emphasize query concepts in the pre-translation phase, improves precision. After translation, a similarity thesaurus is used to reduce the ambiguity by de-emphasizing irrelevant terms added by translation. This method will improve recall.

5.1 Query expansion before translation

This query expansion is performed before translation. ***Local feedback*** is different from relevance feedback only in the way that we fix the number of retrieved documents, for example by assuming the top–ranking documents obtained in an initial retrieval without human judgements. The standard approach is 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.

As a new process in our query translation module, following the research reported by Ballesteros and Croft [2], on the use of relevance feedback, is the disambiguation of translated queries after query expansion. An important distinction here is that without disambiguation, the expanded collection of the

translated candidates in the query, increase exponentially the problem of relevant documents and thus besides the pertinent documents, an useless ones will be retrieved. As a second advantage, creating a stronger base for the short queries in the disambiguation process, in the purpose of using co-occurrence frequency approach.

5.2 Query expansion after translation

So far, one of the best known and tested approaches to CLIR are dictionary-based using a thesaurus, although these are generally used in information retrieval, where each document is indexed with keywords from thesaurus. This is another distinction point to the research reported by Ballesteros and Croft [2], where the post-translation phase concerns a local feedback loop. The motivation for integrating a thesaurus is based on the fundamental idea of doing query expansion and measuring retrieval effectiveness. A similarity thesaurus is a term-vs-term similarity matrix, based on how the terms of the collection are indexed. In query expansion, the top 10 ranked associated thesaurus items are retrieved for a particular query and are added into the query. These similar term concepts are used as candidates to emphasize the query, before the retrieval of documents. This approach improves recall and retrieval performance significantly.

The building of a similarity thesaurus, which identifies the term relationship in the whole training collection such as corpora, is possible but not easy to implement. In this case the query expansion module is simply a matter of thesaurus look up. In this first step of our study, we use an existing thesaurus. A thesaurus-construct is one of our further plans in Cross-Language Information Retrieval field.

5.3 Combined local feedback and similarity thesaurus expansion

We believe that a combination of a query expansion based on a local feedback before translation, with a query expansion based on a similarity thesaurus after translation, will improve the precision-recall and then the effectiveness of the information retrieval. However, poor translations can counteract any improved gained by the local feedback loop. Statistical approach for the disambiguation of the translated candidates for each query term will improve the quality of translations.

6. Conclusion and Future work

Dictionary-based method is attractive because it is cost effective and easy to perform, resources are readily available and performance is similar to that of other Cross-Language Information Retrieval methods. Ambiguity from failure to translate queries is largely responsible for the large drops in effectiveness below monolingual performance [2].

What we presented in this paper, is a rather straightforward integration of existing system modules and a combination to new concepts to achieve an original basis for Cross-Language Information Retrieval. However, further implementation for the evaluation of the performance and the effectiveness of this study are considered, as one of our next steps in Cross-Language Information Retrieval field.

Our ongoing work involves the combination of existing query translators, such as French-English and English-Japanese query translators, by considering an intermediate language (Ex. English). However, The integration of Arabic in our system [12] is a real challenge in Cross-Language Information Retrieval field and offers an excellent opportunity for further works.

It must be stressed that what we reported here, represents our first step in the direction of Cross-Language Information Retrieval.

Our aim is to establish a significant step towards fulfilling the need for Cross-Language Information Retrieval.

Reference

[1] Baeza-Yates, R. and Ribeiro-Neto, B. : "Modern Information Retrieval". ACM Press Books, Addison Wesley Publishers, (1999).

[2] 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.

[3] Buckley, C., Salton, G., Allan, J. and Singhal, A. : "Automatic Query Expansion Using Smart". Proceedings of Third Text Retrieval Conference (Trec3), NIST Special Publication 500-225, (1995).

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

[5] Fujii, A., Hasegawa, T., Tokunaga, T. and Tanaka, H. : "Integration of Hand-crafted and statistical resources in Measuring word similarity". Proceedings of ACL/EACL'97 Workshop on Automatic Information Extraction and Building of Lexical Semantic resources. (1997)

[6] 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.

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

[8] 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.

[9] Hull, D. : "A weighted boolean model for Cross-Language text Retrieval". In G. Grefenstette editor: Cross-Language Information Retrieval, chapter 10. Kluwer Academic Publishers, (1998).

[10] Kay, M. and Röscheisen, M. : "Text Translation Alignment". Computational Linguistics, Vol 19, No1, (1993). P 121-142.

[11] Kurohashi, S. and Nagao, M. : "A method of case structure analysis for Japanese sentences based on Examples in case frame dictionary". IEEE Transactions on information and Systems, E77-D(2), (1994). P 227-239.

[12] Maeda, A., Guan, Q. and Uemura, S. : "Towards a Multilingual Knowledge Discovery System". IPSJ Sig Notes (1999).

[13] Salton, G. : "Automatic Processing of Foreign language documents". Journal of the American Society of Information Science, (1970).P 187-194.

[14] Salton, G. : "Experiments in Multilingual Information Retrieval". Technical report, (1972). P 72-154. Cornell University, Ithaca, New York.

[15] Sheridan, P., Wechsler, M. and Shauble, P. : "Cross-language Speech Retrieval: Establishing a Baseline Performance". In proceedings of the 20th ACM SIGIR Conference, (1997). P 99-108.

[16] Utsuro, T., Ikeda, H., Yamane, M., Matsumoto, Y. and Nagao, M. : "Bilingual Text Matching using Bilingual Dictionary and Statistics". Proceedings of the 15th International Conference on Computational Linguistics, Aug. (1994). P.1076-1082.

[17] 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).

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