Regular Paper

Exploiting and Combining Multiple Resources for Query Expansion in Cross-Language Information Retrieval

Fatiha Sadat,[†] Akira Maeda,[††] Masatoshi Yoshikawa[†††,†]
and Shunsuke Uemura[†]

As Internet resources become accessible worldwide, the need to develop methods in Cross-Language Information Retrieval for different languages becomes increasingly important. In the present paper, we focus on query expansion techniques to improve the effectiveness of information retrieval. Combination of the dictionary-based translation and statistics-based disambiguation approaches is indispensable in overcoming query translation ambiguity. We therefore propose herein a model, which uses multiple sources for query reformulation, organization, translation and disambiguation, to select target translations and retrieve requested information. Relevance feedback or thesaurus-based expansion, as well as a new feedback strategy, which is based on the extraction of domain keywords to expand an original query, are introduced and evaluated. We tested the effectiveness of the proposed combined method using an application of French-English information retrieval. Experiments using the TREC data collection revealed the proposed combination of disambiguation and query expansion techniques to be highly effective.

1. Introduction

The rapid growth in the number of people who have access to the internet, as well as the increasing availability of distributed information and linguistic resources for research, has made information retrieval such a crucial task to fulfill user's needs, that is to find, retrieve and understand relevant information, in whatever language and form.

Cross-Language Information Retrieval (CLIR) consists of providing a query in one language and searching document collections in one or more languages. Therefore, a translation form is required. In the present paper, we focus on query translation rather than on document translation, which is considered to be an unrealistic translation task because of the enormous number of documents to manage. Our first concern is to find retrieval methods, which perform across languages and do not rely on scarce resources such as parallel corpora. Bilingual Machine Readable Dictionaries (MRDs), which are considered as more prevalent than parallel texts, appear to be a good alternative. However, simple translations tend to be ambiguous and therefore yield poor results. Combining dictionary-based translation and statistics-based disambiguation can significantly reduce errors associated with polysemy[☆] in dictionary translation. Automatic query expansion, one of the most important methods by which to overcome the word mismatch problem in information retrieval, is also considered in the present paper. As an assumption to reduce the effect of ambiguity and errors that arise when using dictionary-based methods, a statistical disambiguation method is performed prior to and after translation. Although we conducted our experiments and evaluations on French-English information retrieval, the proposed techniques are common across different languages.

The main contribution of this research concerns an evaluation and comparison between various combinations of expansion techniques involving pseudo-relevance feedback, thesauri as well as a new feedback named domain-based feedback. Domain-based feedback is based on hierarchical category schemes and pseudo-relevance feedback in order to extract domain keywords and expand original queries. New weighting schemes are proposed for each expansion technique in order to select relevant expansion terms. Moreover, we proposed new disambiguation methods. The first disambiguation method is based on ranking of source query

[†] Graduate School of Information Science, Nara Institute of Science and Technology (NAIST)
[††] Department of Computer Science, College of Science and Engineering, Ritsumeikan University
[†††] Information Technology Center, Nagoya University

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

terms and disambiguation of target translations. The second disambiguation method is based on ranking and disambiguation of target translations.

The remainder of the present paper is organized as follows: Section 2 presents an overview of related works. Dictionary-based translation and the proposed statistical disambiguation techniques are described in Section 3. Query expansion techniques using different combinations are introduced in Section 4. Experiments and evaluations are discussed in Section 5. Section 6 presents the conclusion of the present paper.

2. Related Works

Research in CLIR dates back to the early seventies when Salton[21),24)] established that the performance of English-French CLIR was comparable to the performance of monolingual retrieval when manually developed resources were employed for query translation. Twenty years later, the availability of linguistic resources revealed the possibility of multilingual retrieval with minimal manual intervention.

Current approaches to CLIR may be classified conveniently into four classes: the Machine Readable Dictionary-based (MRD) approach, the Machine Translation-based (MT) approach, the corpus-based approach, which relies on parallel or comparable corpora, and other approaches which are based on other existing linguistic resources, such as the thesaurus.

MT approach is effective; however, a disadvantage of present fully automatic machine translation systems is that these systems are able to produce high-quality translations but only in limited domains[16)]. A certain amount of syntactic error is acceptable if the results of the information retrieval system are not adversely affected; however, MT errors that occur during the translation of concepts can prevent relevant documents from being retrieved, i.e. those using incorrectly translated concepts. An example is the word traitement in French, which would be translated to English as processing rather than as salary, the retrieval process would yield incorrect results.

Corpora-based approaches, which are based on parallel or comparable texts, also have drawbacks. Test corpora are costly to acquire and not readily available. Training corpora must be very large. Moreover, finding previously existing translations of the appropriate type of documents is difficult and translated versions are expensive to create. Nie, et al.[15)] proposes a successful method to gather parallel texts automatically from the Web and construct a training corpus. However, this method would be expensive for any pair of languages or even not applicable for some languages, which are characterized by few amounts of Web pages on the Web.

Therefore, growing interest has been expressed in the potential of knowledge-based technology. Automatic MRD's query translation, on its own, has been found to decrease effectiveness by 40-60% compared to monolingual retrieval[1),9)]. The combination of dictionary-based translation and statistics-based disambiguation has been successfully exploited in several research related to information retrieval[1),2),7),9),11),24),25),29)].

Although, most research on CLIR has concentrated on query translation and disambiguation, and has investigated statistical approaches, no consideration was given to ranking, selection of source query terms or target translations. Moreover, query expansion has been proven effective in improving the performance of information retrieval[1),2),4),13),14)]. Some research studies have attempted to combine various types of thesauri, in essence drawing upon the strengths of each to counter their various weaknesses. Mandala, et al.[13),14)] has proposed the use of heterogeneous thesauri for query expansion, by combining three types of thesauri, a handcrafted thesaurus, a co-occurrence-based thesaurus, and a syntactic-relation-based thesaurus. Experiments using TREC-7 collection have demonstrated the effectiveness of the proposed method. However, the scope of the research completed by Mandala, et al.[13),14)] is limited to monolingual information retrieval and does not include any CLIR environment.

In the present paper, we focus on query translation and disambiguation of multiple target candidates and query expansion using various combinations of query expansion techniques, in order to improve the effectiveness of retrieval across languages. We propose and discuss the application of a statistical disambiguation technique prior to and subsequent to dictionary translation. Extraction, selection and addition of terms that emphasize query concepts are performed using expansion techniques such as, pseudo-relevance feedback and thesaurus-based

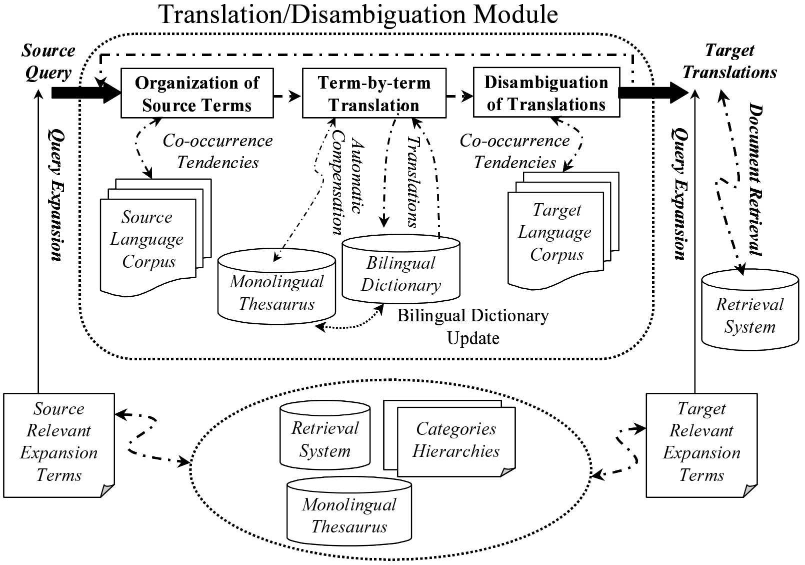


Fig. 1 Overview of the proposed Cross-Language Information Retrieval system.

expansion. Moreover, a new expansion technique, domain-based feedback, which extracts and selects domain keywords in order to expand original queries is introduced and evaluated. Expansion terms will be selected based on weights and ranks using a scheme based on co-occurrence tendencies with all query terms. Weighting schemes depend on the structure of each expansion technique. Linear combinations involving pseudo-relevance feedback, domain-based feedback and thesauri are discussed for the purpose of creating an optimal model for query expansion and enhancing the effectiveness of information retrieval.

3. Translation/Disambiguation in CLIR

We initiated the concept of an information retrieval system involving query translation, disambiguation and expansion, as well as retrieval of documents in the target language[19].

3.1 Overview of the Proposed CLIR System

Figure 1 shows the overall design of the proposed Cross-language Information Retrieval system. Query expansion is applied prior to and/or subsequent to the translation/disambiguation process. Among the proposed expansion strategies are, relevance feedback, domain-based feedback and thesaurus-based expansion. This can be completed simply by translating user's queries into the target language and then selecting the best target translations, which can be a single term or several weighed terms, for each source query term.

In the present study, a term-by-term translation using a bilingual MRD is performed after simple *stemming* 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. Three primary tasks are accomplished using the translation/disambiguation module. First, an *organization of source query terms*, which is considered key to the success of the disambiguation process, will select best pairs of source query terms. Next, a *term-by-term translation* using the dictionary-based method[20], in which each term or phrase in the query is replaced by a list of possible translations, is completed. The problem of missing words in the dictionary, which are essential for the correct interpretation of the query, could be solved. This may occur either because the query deals with a technical topic, which is outside the scope of the dictionary or because the user has entered some form of abbreviations or slang, which is not included in the dictionary[16]. In order to solve this problem, an automatic *compensation* via synonym dictionary or existing thesaurus in the concerned language, is introduced. This case requires an extra step to look up the query term in the thesaurus or synonym dictionary and find equivalent terms or synonyms of the targeted source term, thus performing a query translation. In addition, short queries of one term are handled in this phase. In the third task, *disambiguation of target translations*, best translations related to each source query term are selected. Finally, documents are retrieved in the target language.

3.2 Co-occurrence Tendency

All possible combinations of source query terms are constructed and ranked depending on their mutual co-occurrence in a training corpus. A type of statistical metric called *co-occurrence tendency*[12] can be used to accomplish this task. Methods such as Mutual Information (MI)[3),20)], Log-Likelihood Ratio (LLR)[5)], the Modified Dice Coefficient[12)] or Gale's method[6)] are all candidates for the co-occurrence tendency.

If a word A frequently co-occurs with another word B in a fixed window size of a corpus, we can expect their co-occurrence tendency to be high. Thus, correct pairs of terms tend to co-occur in a corpus and incorrect pairs tend not to co-occur. We use this hypothesis to rank pairs of source query terms according to their co-occurrence tendencies in the source language corpus. As well, correct translations of query terms should co-occur in the target language corpus and incorrect translations should tend not to co-occur. We use again this hypothesis as the foundation for a method to disambiguate target translations according to their co-occurrence tendencies in the target language corpus. Therefore, the selection of pairs of source query terms for translation, as well as the disambiguation of translation candidates in order to select target translations, is performed by applying one of the statistical metrics based on co-occurrence frequency, as follows:

Mutual Information (MI)

This estimation uses *mutual information*[3)] as a metric for significance of word co-occurrence tendency, as follows:

$$MI(w_1, w_2) = \log \frac{Prob(w_1, w_2)}{Prob(w_1)Prob(w_2)}$$

Here, $Prob(w_i)$ is the frequency of occurrence of word w_i divided by the size of the corpus N, and $Prob(w_i, w_j)$ is the frequency of occurrence of both w_i and w_j together in a fixed window size in a training corpus, divided by N, i.e. number of terms in the corpus.

Log-Likelihood Ratio (LLR)

The Log-Likelihood Ratio[5)] has been used in a number of research studies. LLR is expressed as follows:

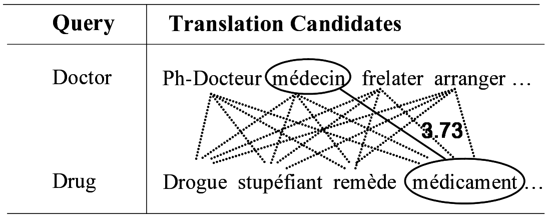


Fig. 2 Two-term disambiguation process (Source/Target languages are English/French).

$$-2\log\lambda = K_{11}\log\frac{K_{11}N}{C_1R_1} + K_{12}\log\frac{K_{12}N}{C_1R_2} + K_{21}\log\frac{K_{21}N}{C_2R_1} + K_{22}\log\frac{K_{22}N}{C_2R_2}$$

where,

$$\begin{aligned} C_1 &= K_{11} + K_{12}, C_2 = K_{21} + K_{22}, \\ R_1 &= K_{11} + K_{21}, R_2 = K_{12} + K_{22}, \\ N &= K_{11} + K_{12} + K_{21} + K_{22}, \\ K_{11} &= \text{frequency of common occurrences} \\ &\quad \text{of word} w_i \text{and word} w_j, \\ K_{12} &= \text{corpus frequency of word} w_i - K_{11}, \\ K_{21} &= \text{corpus frequency of word} w_j - K_{11}, \\ K_{22} &= N - K_{12} - K_{22} \end{aligned}$$

3.3 Disambiguation of Target Translations

A word is *polysemous* if it has senses that are different but closely related. For example, the word *right* can mean something that is morally acceptable, something that is factually correct, or one's entitlement. A two-term disambiguation of translation candidates can be applied[12),20)] following the dictionary-based translation. The standard procedure for *two-term disambiguation* is as follows:

(1) Construct all possible combinations of pairs of terms, from the translation candidates.

(2) Request the disambiguation module to obtain the co-occurrence tendencies. The window size is set to one paragraph of a text document rather than to a fixed number of words.

(3) Choose the translation, which shows the highest co-occurrence tendency, as the most appropriate.

Figure 2 shows an example of the two-term disambiguation process for an English query "*doctor*, *drug*". All possible combinations of translation candidates of both source terms are constructed and ranked based on their co-occurrence tendencies as follows: (*médecin*, *médicament*), (*médecin*, *remède*), (*médecin*, *drogue*), etc. The best combination having the highest co-occurrence tendency is selected for the pair of target French terms "*médecin, médicament*". As illustrated in Fig. 2, the disambiguation procedure is used for two-term queries due to computational cost[12]. In addition, the primary problem concerning long queries involves the selection of pairs of terms as well as the order of disambiguation. We propose and compare two disambiguation methods for long queries (*n-term disambiguation*).

The first method, denoted by *RSDT* (Ranking Source query terms and Disambiguation of Target translations), is based on a ranking of pairs of source query terms before translation and then disambiguation of target translations. The key concept in this step is to maintain a ranking order from the organization phase and perform translation and disambiguation starting from the most informative pair of source terms, i.e. a pair of source query terms having the highest co-occurrence tendency in the source language corpus. These source terms are translated using a bilingual dictionary. Disambiguation of translation candidates is completed following the co-occurrence tendencies in the target language corpus. Thus, co-occurrence tendency is involved within the *RSDT* method in both source and target languages corpora.

The second method, denoted by *RTDT* (Ranking Target translations and Disambiguation of Target translations), is based on a ranking of target translation candidates, following their co-occurrence tendencies in the target language corpus.

The proposed statistical disambiguation methods could be employed on large-scale, domain-independent test corpora such as collection of newspapers or generalized debates. Specialized domain texts sometimes lack terms, which are necessary for an efficient query disambiguation.

Suppose, Q represents a source query having n terms $s_1, s_2, \ldots, s_n$.

RSDT Method: *Ranking Source query terms and Disambiguation of Target translations*

(1) Construct all possible combinations of terms of one source query: $(s_1, s_2), (s_1, s_3), \ldots, (s_{n-1}, s_n)$.

(2) Rank all combinations, according to their co-occurrence tendencies toward the highest values.

(3) Select the combination (s_i, s_j) having the highest co-occurrence tendency, where at least one translation of the source terms has not yet been fixed.

(4) Retrieve all translations related to this combination from the bilingual dictionary.

(5) Apply a two-term disambiguation process to all possible translation candidates.

(6) Fix the best target translations for this combination and discard the other translation candidates.

(7) Go to the combination having the next highest co-occurrence tendency and repeat steps 3 through 6 until the translation of every source query term is fixed.

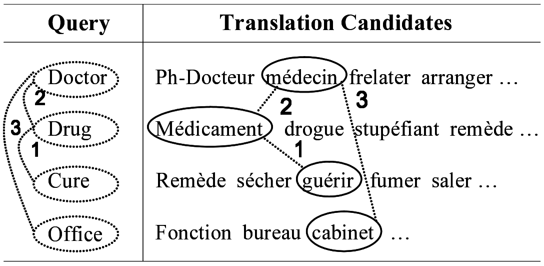

Query
Translation Candidates
Doctor
Drug
Cure
Office
2
3
1
Ph-Docteur médecin frelater arranger ...
Médicament drogue stupéfiant remède ...
Remède sécher guérir fumer saler ...
Fonction bureau cabinet ...
2
3
1

Fig. 3 RSDT disambiguation process (Source/Target languages are English/French).

Table 1 An example for RSDT method.

English Query Terms	Best French Translations		
	Step 1	Step 2	Step 3
Doctor	–	***Médecin***	Médecin
Drug	***Médicament***	Médicament	Médicament
Cure	***Guérir***	Guérir	Guérir
Office	–	–	***Cabinet***

Figure 3 shows an example for the *RSDT* method. Assume a source English query "*doctor drug cure office*". First, all possible combinations of source query terms are constructed and ranked based on their co-occurrence tendencies as follows: (*drug*, *cure*), (*doctor*, *drug*), (*doctor*, *office*), (*doctor*, *cure*). These combinations are translated and their best translations are selected and fixed according to highest co-occurrence tendencies. As a result, the best combination of target French translations is selected as "*médecin médicament guérir cabinet*".

Table 1 shows the different steps to select source terms and fix the best translations.

The second disambiguation method proposed herein is based on ranking and disambiguation of target translation candidates using co-occurrence tendencies.

RTDT Method: *Ranking Target translations and Disambiguation of Target translations*

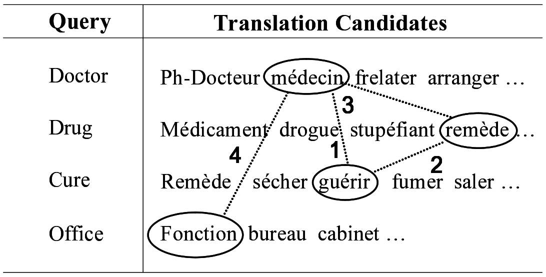

Query
Translation Candidates
Doctor
Drug
Cure
Office
Ph-Docteur médecin frelater arranger ...
Médicament drogue stupéfiant remède ...
Remède sécher guérir fumer saler ...
Fonction bureau cabinet ...
1
2
3
4

Fig. 4 RTDT disambiguation process (Source/Target languages are English/French).

Table 2 An example for RTDT method.

English Query Terms	Best French Translations		
	Step 1	**Step 2**	**Step 3**
Doctor	***Médecin***	Médecin	Médecin
Drug	–	***Remède***	Remède
Cure	***Guérir***	Guérir	Guérir
Office	–	–	***Fonction***

(1) Retrieve all possible translation candidates for each source query term s_i from the bilingual dictionary.

(2) Construct sets of translations $T_1, T_2, \ldots, T_n$ related to each source query term $s_1, s_2, \ldots, s_n$, and containing all possible translations for the concerned source term. For example, $T_i = t_{i1}, \ldots, t_{in}$ is the translation set for the term s_i.

(3) Construct all possible combinations of elements of different sets of translations. For example, $(t_{11}, t_{21}), (t_{11}, t_{22}), \ldots, (t_{ij}, t_{nk})$.

(4) Select the combination having the highest co-occurrence tendency.

(5) Fix these target translations, for the related source terms and discard the other translation candidates.

(6) Go to the next highest co-occurrence tendency and repeat steps 4 through 5 until the translation of every source query term is fixed.

Figure 4 shows an example for the RTDT method with the source English query "*doctor drug cure office*". In this case, all possible combinations of target translation candidates are constructed and ranked based on their co-occurrence tendencies as follows: (*médecin*, *guérir*), (*guérir*, *remède*), (*remède*, *médecin*) (*médecin*, *fonction*), etc. The best French translations are selected and fixed following the highest co-occurrence tendencies. As a result, the best combination of target French translations is selected as "*médecin remède guérir fonction*". **Table 2** shows the different steps to select and fix the best translations.

4. Query Expansion for Information Retrieval

Query expansion has been considered in several studies[1),2),4),13),14)] and the effectiveness of this technique in improving the performance of information retrieval has been proved. Following the research reported by Ballesteros and Croft[1),2)] on the use of local feedback in CLIR, the addition of terms that emphasize query concepts in the pre- and post-translation phases improves both precision and recall. Mandala, et al.[13),14)] proposed the use of heterogeneous thesauri for the purpose of query expansion by combining three types of thesauri. This combined method was effective for monolingual information retrieval.

In the present paper, query expansion is represented by one of the following techniques: pseudo-relevance feedback to select the most highly weighted terms in the relevant documents, domain-based feedback to extract and select domain keywords, and thesaurus-based expansion to retrieve synonyms and multiple word senses from a monolingual thesaurus. Each expansion type is discussed below.

4.1 Pseudo-relevance Feedback

Pseudo-relevance feedback, which is applied to query reformulation and expansion, attempts to fix the number of retrieved documents and assumes that the top-ranked documents are relevant. Test collections including relevance judgments are used to determine relevant documents among the retrieved ones. A fixed number of term concepts are extracted and their co-occurrence frequencies in conjunction with the original query terms are estimated. However, any query expansion must be handled very carefully, because selecting just any expansion term could be dangerous. The proposed expansion method via Rocchio-inspired relevance feedback[18)] is based on statistical co-occurrence tendency in conjunction with all terms in the original query, rather than with only one query term. Assume a query Q having n terms $term_1, \ldots, term_n$. A ranking factor based on the co-occurrence tendency between each query term and the expansion candidate is defined as follows:

$$Rank(Q, expterm) = \sum_{i=1}^{n} w_i \times co\text{-}occurrence(term_i, expterm)$$

where *co-occurrence*$(term_i, expterm)$ represents the co-occurrence tendency between a query term and the targeted expansion candidate. This can be evaluated by any estimation method, such as mutual information or log-likelihood ratio. The value of w_i, which represents the weight of the query term $term_i$ in relation to the expansion term $expterm$, is equal to 1 if a co-occurrence tendency between $term_i$ and $expterm$ exists; otherwise w_i becomes 0.

Thus, all co-occurrence values are computed, summed for all query terms $(i = 1 \ldots n)$ and the expansion candidate of the highest rank is selected as an expansion term for the query Q. Note that the highest rank must be related to the maximum number of terms in the query, if not all query terms:$MAX[\sum_{i=1}^{n} w_i]$

Such expansion may involve a number of expansion candidates or just a subset of the candidates.

4.2 Domain-based Feedback

The domain-based feedback approach[19)] attempts to extract domain keywords from the set of top-retrieved documents using standard relevance feedback to expand the original query set. Web directories, such as *Yahoo!*[☆1] and *AltaVista*[☆2] are human-constructed and are designed for human web browsing. These web directories provide a hierarchical category scheme into which documents are sorted. Digital libraries, such as the *Library of Congress* catalogue support some forms of subject indexing, which is again hierarchical. These kinds of hierarchies can be exploited for keyword extraction and thus query expansion. Our strategy relies on term extraction using a standard relevance feedback under the condition that these terms represent either a directory or category, denoted by a keyword describing its content and are thus considered to be a specific domain for a collection of documents. The process is described as follows:

(1) Extract terms or seed words using relevance feedback as well as the proposed ranking strategy to select the expansion term, as explained in the previous section. This set is denoted as $set1$.

(2) Collect domain keyword candidates, from categories and directories related to hierarchical web directories, such as *Yahoo!*, *AltaVista*, *Open Directory*[☆3] and from hierarchical databases such as the *National Library of Canada*[☆4] and its related subject tree, which is based on the *Dewey Decimal Classification* system. This set of collected keywords is denoted as $set2$.

(3) Select a domain keyword that is a seed word from $set1$ and is also a candidate of $set2$. If the number of terms in the intersection is large, a statistical disambiguation process will be applied to rank the resulting domain keywords and select the best domain keyword.

Table 3 Domain keywords extraction and ranking factors in conjunction with original query terms.

	Domain Keywords	**Source Query Term**	**Ranking (using (LLR)**
1.	Science	design	10.776
2.	Technology	star	8.253
3.	Government	war	8.145
4.	Computers	anti	4.635
5.	Entertainment	missile	0.456
6.	Geography	defense	0.231
7.	Economy	system	0.098
	...		

The resulting set of terms will be used for domain-based feedback, and may involve many expansion terms or just a subset. An example is the TREC query: "*design star wars anti-missile defense system*", which could be expanded using at least one of the following ranked domain keywords: "*science, technology, government, etc.*". **Table 3** shows a set of extracted domain keywords ranked by their co-occurrence tendencies with the domain key term using log-likelihood ratio measure.

A more complex task to extract domain keywords would be based on the classification of retrieved documents. An alternative could involve web document categorization in order to extract relevant documents within a specified category and then select the set of relevant terms to the original query. A set of keywords (path representing hierarchy of categories) related to each document in a category, could be assigned as domain keywords.

4.3 Thesaurus-based Expansion

Using thesauri or other structures such as ontologies, has been the subject of extensive research, and some promising results have been obtained. Recently, some encouraging

[☆1] http://www.yahoo.com/

[☆2] http://www.altavista.com/

[☆3] http://dmoz.org/

[☆4] http://www.nlc-bnc.ca/caninfo/esub.htm findings[13),14),17),26)] have been obtained using WordNet[☆], a large manually-constructed thesaurus. WordNet, an online lexical database reference for English, has been applied very successfully to both monolingual and multilingual information retrieval. The system combines the capabilities of both an on-line thesaurus and an on-line dictionary. WordNet distinguishes different kinds of relationships, such as:

- Hypernymy (*is-a* relation), generalization (ex.: computer ⇒ machine)
- Hyponymy (*has-a* relation), specialization (ex.: computer ⇒ analogue computer),
- Meronymy (*has-part* relation), generalization (ex.: keyboard ⇒ computer).

These concepts serve to organize multiple word senses into a set of hierarchies. Query expansion is possible using a fixed number of descendants/ascendants of the given query that exist in different hierarchies (i.e., hypernyms, hyponyms, meronyms) or in all multiple word senses. In the present study, we investigate the synonymy relation in query expansion.

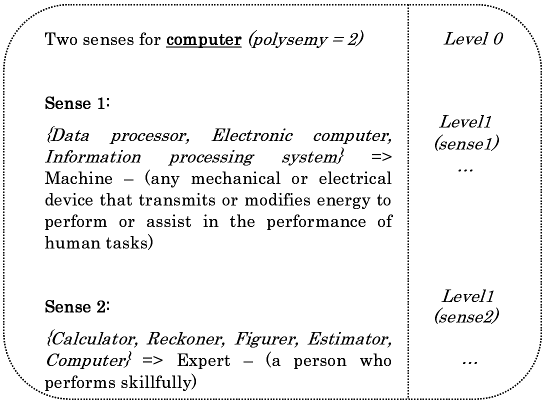
4.3.1 Thesaurus-based Expansion through Synsets

WordNet's basic object is a set of strict synonyms, called a *synset*[26)]. Following the research reported by Voorhees[26)] on the use of lexical relations of WordNet for a query expansion, we can proceed using a simple search to find synsets of the full query. For example, the original query *defense system* could be expanded using the terms *weaponry* or *arm.* Otherwise, in case of non-existence of the full query, we proceed by a term-by-term search in the lexical database. In addition, a simple one-term query can be represented by a compound synonym. In this case, we construct a conjunction between simple terms of the concerned synonym. An example is the simple term *war*, which is first expanded by the compound synonym *military action* and then replaced by the terms *war military action.* Moreover, statistical frequency might be used for ranking and selection in order to avoid words that do not occur frequently together with the original terms. For example, the term *reckoner* would be removed from the synset list of the term *computer.* An appropriate weighting scheme allows smooth integration of these related terms by reducing their influence over the query[17)].

Following these three assumptions, an original query with a term *computer* will be expanded using synonyms and will contain the following terms:

4.3.2 Weighting Expansion Term Candidates

All thesauri such as WordNet provide semantical relations among terms. The related concepts can organize multiple word senses (synonyms or others) into a set of hierarchies, thus query expansion should be possible using a fixed number of descendants/ascendants of a given query, which exists in different hierarchies. Assume that a query term will be represented by level 0 in the conceptual hierarchy. In this case, synonyms would be represented by level 1 in the hierarchy with a fixed number for a synset, such a branch of the conceptual hierarchy, depending on the WordNet definition. Therefore, synonyms could be added to the query. However, some caution is required as these new extracted terms from the thesaurus are not as reliable as the initial terms obtained from users. An appropriate weighting scheme for synonyms will allow smooth integration of these expansion terms by reducing their influence over the query. Accordingly, all terms recovered from the thesaurus will be given weights, expressing their similarity to initial terms, based on their position in the conceptual hierarchy (depth = 1) as well as the number of terms accompanying them in the same synset. These weights range between 0 and 1, and a weight of 1 is assigned to original query terms, which belong to level 0 in the conceptual hierarchy, as shown in **Fig. 5**. Various strategies have been proposed[17),26)] for sense disambiguation and weight assignment to synonyms and other terms of a thesaurus. In the present study, weights assigned to any syn-



Two senses for <u>computer</u> *(polysemy = 2)* *Level 0*

Sense 1:

{Data processor, Electronic computer, Information processing system} => Machine – (any mechanical or electrical device that transmits or modifies energy to perform or assist in the performance of human tasks)

Level1 (sense1)

...

Sense 2:

{Calculator, Reckoner, Figurer, Estimator, Computer} => Expert – (a person who performs skillfully)

Level1 (sense2)

...

Fig. 5 WordNet simple search for the term "computer" (only synonyms are represented).

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

onym of one synset are related to an envelope of 0.5 divided by the number of terms in the corresponding synset. The weight of a term is proportional in the same synset. For example, if a synset contains M terms, the weight of each term will be $1/M$. Thus, we propose a weighting factor to be assigned to the retrieved expansion term from a synset in the conceptual hierarchy related to WordNet thesaurus, as follows:

$$Weight(term, expterm_j) = \frac{Sim(term, expterm_j)}{2M}$$

where M is the number of terms which belong to the same synset. $Sim(term, expterm_j)$ is the similarity between the $term$ and an expansion candidate $expterm_j$ and can be estimated by any similarity measure, such as the Cosine measure[23] as follows:

$$Sim(term, expterm_j) = \frac{\sum_k v_{ik} v_{jk}}{\sqrt{\sum_k v_{ik}^2 \sum_k v_{jk}^2}}$$

v_{ik} and v_{jk} are the frequencies of $term$ and $expterm_j$ in a corpus, respectively.

In addition, a Term Frequency (TF) and Inverse Document Frequency (IDF) measures[22] can be used as a suitable similarity measure, as follows:

$$Sim(term, expterm_j) = TF(expterm_j) \times IDF(expterm_j)$$

where $TF(expterm_j)$ is the term frequency of term $expterm_j$ in the context of the query term, and $IDF(expterm_j)$ is the inverse document frequency of the expansion term $expterm_j$, which is represented by $\log(N/n_j)$. N is the maximum frequency of any word in a corpus and n_j is the total number of occurrences of term $expterm_j$, in a corpus.

However, expanding a query using any of those weighted synonyms implies a careful selection and ranking, depending on the statistically most-weighted terms in conjunction with all query terms, rather than just a one-term query. For a query Q using k terms $term_1, term_2, \ldots, term_k$, weight factors are computed for an expansion term candidate and summed for all query terms ($1 \leq i \leq k$), if the query term appears in the related hierarchies. The highest weighted term is then selected for query expansion, as follows:

$$Weight(query, expterm) = \sum_{i=1}^{n} weight(term_i, expterm)$$

4.4 Combining Different Approaches

Following the research on the use of local feedback reported by Ballesteros and Croft[1),2)], adding terms that emphasize query concepts in the post- and pre-translation phases improves precision and recall. This combined method has been reported to reduce the ambiguity by de-emphasizing irrelevant terms added by translation and therefore should improve precision and recall in information retrieval. The new query Q_{new} can be defined as follows:

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

where Q_{orig} is the original query translated and disambiguated, and $\sum_{bef} T_i$ and $\sum_{aft} T_j$ represent sets of terms added before and after translation/disambiguation, respectively. Although the two parameters α_1 and α_2, which are considered as query dependant and represent the importance of each expansion strategy have been set arbitrarily in the present study, their values may be estimated using an Expectation-Maximization algorithm. The query expansion technique may be one of the methods described in Section 4.

5. Experiments and Evaluation

Experiments and evaluations of the effectiveness of the proposed strategies for query translation, disambiguation and expansion were conducted using a large-scaled test collection related to French-English information retrieval, i.e. French queries to retrieve English documents.

5.1 Linguistics Resources

Test Data

Experiments were conducted using the *Tipster volume 1* data collection. This data collection was used in earlier research studies related to TREC-1[☆]. This data collection contains approximately 164,600 documents from the Wall Street Journal[8)], the size of which is approximately 65 megabytes. Topics 51-150 queries, which are composed of several fields, are considered as English queries for the present experiments. The tags <num>, <dom>, <title>, <desc>, <smry>, <narr> and <con> denote

[☆] http://trec.nist.gov/data.html

the topic number, the domain, title, the description, the summary, the narrative and concept fields. Key terms contained in the title <title> and description <desc> fields, which averaged 5.7 terms per query, are used to generate English source queries. French version of the queries was constructed by manually translating the original English query set by a native speaker.

Monolingual Corpus

Possibly the most well-known parallel corpus is the Canadian *Hansard*☆1 Corpus consisting of debates from the Canadian Parliament, which have been published in the country's official languages, English and French. This corpus has been used in research for many years, among others by Gale and Church6) for testing their alignment algorithm as well as other research studies. The present paper is based on *Hansard* corpora, which contains more than 100 million words of English text and the corresponding French translations. In the present study, we used *Hansard* as a monolingual corpus for the French and English languages.

Bilingual Dictionary

COLLINS☆2 Series 100 French-English dictionary was used to translate source queries. This bilingual dictionary includes 75,000 references and 110,000 translations, which are considered to be sufficient for the present research.

Thesauri

*WordNet*26) and *EuroWordNet*27),28) are used for thesaurus-based expansion and possible compensation for limitations in the bilingual dictionary.

Stemmer and Stop Words

Stemming was performed using the *Porter Stemmer*☆3.

Retrieval System

SMART☆4, an information retrieval system based on a vector space model that has been used in enormous studies concerning CLIR was used to retrieve English documents.

5.2 Experiments and Results

A retrieval using original English queries was represented by *Mono_Eng* method. We conducted two types of experiments: Those related to the query translation/disambiguation and those related to the query expansion before and after translation. Document retrieval was performed using original and constructed queries, by the following methods: *All_Tr*, which is the result of using all possible translations for each term in the source query as obtained from the bilingual dictionary, and *No_DIS*, which uses no disambiguation, meaning that the first translation is selected as the target for each source query term. We used a log-likelihood ratio as an estimation for co-occurrence tendency for all disambiguation methods, as follows: *Bi_DIS*20) was used for disambiguation of consecutive pairs of source terms without any ranking or selection. *RSDT* is a result of the first proposed disambiguation method, i.e. ranking source query terms and thus translation and disambiguation. *RTDT* is the result of the second proposed disambiguation method, i.e. ranking and disambiguation of translation candidates.

Query expansion using different combinations is evaluated before and after translation and disambiguation by the *RSDT* method, which we denote by trans_disambiguation. Pseudo-relevance feedback was evaluated using the test data, Tipster volume 1 data collection. Fixed experimental parameters are determined following the previous research by Ballesteros and Croft1). Approximately 50 top-ranked documents obtained from the initial retrieval are assumed relevant. Term concepts are extracted from the set of retrieved documents, and a fixed number (up to 10) of top-ranked terms are used as expansion candidates. Therefore, *Feed.bef* is obtained by adding a number of terms to the original queries and then performing a trans_disambiguation. *Feed.aft* is obtained by query trans_disambiguation and then expansion via pseudo-relevance feedback. *Feed.bef_aft* is obtained by combined pseudo-relevance feedback both before and after trans_disambiguation. A domain-based feedback was tested using *Feed.dom* after trans_disambiguation. A combined method including pseudo-relevance feedback was tested using *Feed.bef_dom*. Thesaurus-based expansion was evaluated using synsets related to each query term. A weighting factor was calculated to rank and select best-candidate expansion terms to be added to the original query set. Thus, WordNet-based expansion was evaluated on target translations using *Feed_wn*, and the EuroWordNet-based expansion was evaluated on source queries using *Feed.ewn*. Com-

☆1 http://www.isi.edu/natural-language/download/hansard/

☆2 Collins Series 100 Bilingual Dictionary

☆3 http://www.tartarus.org/~martin/PorterStemmer/

☆4 ftp://ftp.cs.cornell.edu/pub/smart

Table 4 Explanations on the translation, disambiguation and expansion methods.

	Methodś notation	**Techniques used in the method**
1.	Mono_Eng (*baseline*)	Monolingual English IR
2.	No_DIS	IR using simple translation: Selecting the first translation candidate in the dictionary
3.	All_Tr	IR using all translation candidates
4.	Bi_DIS	IR using a disambiguation method based on co-occurrence tendency between consecutive pairs of terms
5.	RTDT	IR using the proposed disambiguation method: Ranking and disambiguation of target translations
6.	RSDT	IR using the proposed disambiguation method: Ranking source query terms and disambiguation of target translations
7.	Feed.bef	IR using query expansion: Pseudo-relevance feedback before the RSDT method
8.	Feed.aft	IR using query expansion: Pseudo-relevance feedback after the RSDT method
9.	Feed.bef_aft	Combination: 7+8
10.	Feed.dom	IR using query expansion: Domain-based feedback after the RSDT method
11.	Feed.bef_dom	Combination: 7 + 10
12.	Feed_wn	IR using query expansion: Thesaurus-based expansion using *WordNet* after the RSDT method
13.	Feed.ewn	IR using query expansion: Thesaurus-based expansion using *EuroWordNet* before the RSDT method
14.	Feed.bef_wn	Combination: 7 + 12
15.	Feed.ewn_aft	Combination: 13 + 8
16.	Feed.dom_wn	Combination: 10 + 12
17.	Feed.ewn_dom	Combination: 13 + 10
18.	Feed.ewn_wn	Combination: 13 + 12
19.	Feed.ewn_wn_dom	Combination: 13 + 12 + 10

Table 5 Results and evaluations of different combinations using translation, disambiguation and expansion techniques.

	Method	**Avg. Prec.**	**% Monolingual**	**% Improvement**
1.	Mono_Eng (*baseline*)	0.2628	**100**	–
2.	No_DIS	0.2214	84.24	−15.76
3.	All_Tr	0.2160	82.19	−17.81
4.	Bi_DIS	0.2259	85.95	−14.05
5.	RTDT	0.2387	90.82	−9.18
6.	RSDT	**0.2679**	**101.94**	**+1.94**
7.	Feed.bef	0.2309	87.86	−12.14
8.	Feed.aft	0.2663	101.33	+1.33
9.	Feed.bef_ aft	**0.2704**	**102.89**	**+2.89**
10.	Feed.dom	0.2328	88.58	−11.42
11.	Feed.bef_ dom	**0.2725**	**103.69**	**+3.69**
12.	Feed_wn	0.2518	95.81	−4.19
13.	Feed.ewn	0.2579	98.13	−1.87
14.	Feed.bef_ wn	0.2571	97.83	−2.17
15.	Feed.ewn_ aft	0.2588	98.47	−1.53
16.	Feed.dom_wn	0.2540	96.65	−3.35
17.	Feed.ewn_ dom	0.2545	96.84	−3.16
18.	Feed.ewn_wn	0.2608	99.23	−0.77
19.	Feed.ewn_wn_dom	**0.2741**	**104.29**	**+4.29**

bined methods involving feedback techniques are represented as *Feed.bef_wn*, for relevance feedback and thesaurus-based expansion prior to and subsequent to trans_disambiguation, and as *Feed.ewn_aft*, for EuroWordNet-based expansion and relevance feedback prior to and subsequent to trans_disambiguation. Similar evaluations use a domain-based feedback: *Feed.dom_wn* and *Feed.ewn_dom*. A combined thesauri-based expansion method was tested using *Feed.ewn_wn*. Finally, *Feed.ewn_wn_dom* represents combined thesauri (WordNet and EuroWordNet) and domain-based expansion. A description of these techniques is shown in **Table 4**. The performances of these methods are presented along with the results **Table 5**. **Figure 6** shows the query translation/disambiguation using log-likelihood ratio for co-occurrence tendency. **Figures 7** and **8** show the query expansion techniques using different combinations.

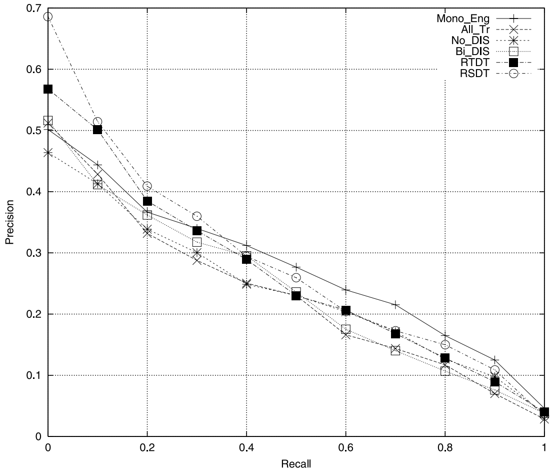


Fig. 6 Recall/Precision curves for query translation and disambiguation (*using RSDT and RTDT disambiguation methods*).

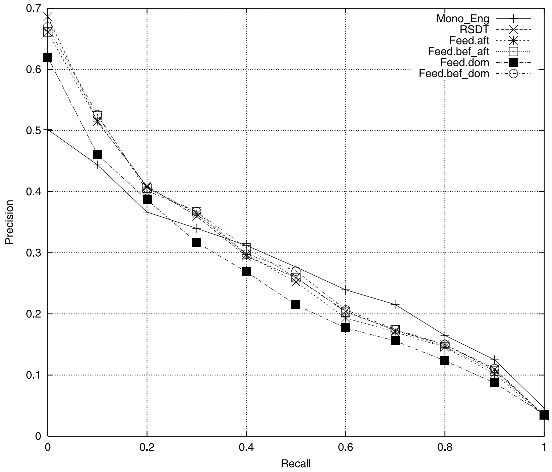


Fig. 7 Recall/Precision curves for query translation, disambiguation and expansion (*using pseudo-relevance feedback and domain-based feedback*).

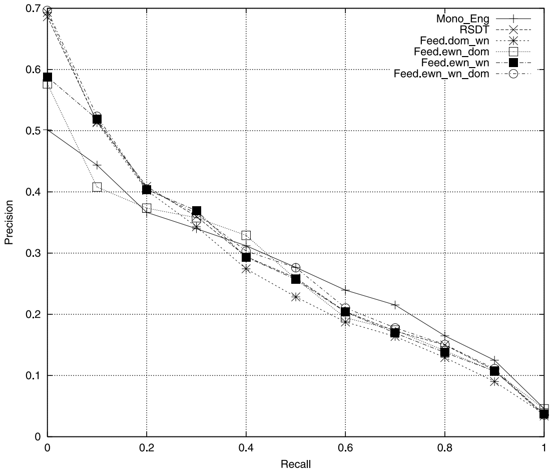


Fig. 8 Recall/Precision curves for query translation, disambiguation and expansion (*using domain-based feedback, WordNet-based and EuroWordNet-based expansion*).

5.3 Discussion

The second column of Table 5 indicates the average precision, which is used as the basis for the evaluation. The third column indicates the average precision as compared to the monolingual counterpart. The monolingual performance *Mono_Eng* was lower than the best results of TREC-2, but higher than other results[8]. We first carried out a set of experiments to investigate the impact of the proposed disambiguation methods on query translation. Our results confirmed our intuition. Results showed that an efficient disambiguation improves CLIR performance significantly and consistently when using large linguistic resources such as the *COLLINS* bilingual dictionary or the *HANSARD* French and English corpora. Therefore, *All_Tr* and *No_DIS* showed no improvement in terms of precision or recall compared to the monolingual English retrieval, whereas simple disambiguation of consecutive pairs of source terms *Bi_DIS* increased recall and precision by 1.71% over the average precision as compared to the simple dictionary translation without any disambiguation. On the other hand, the proposed disambiguation method *RTDT* appears to be helpful in enhancing the precision, 90.82% of the average precision but no gain in recall was obtained. The second proposed disambiguation method *RSDT* showed a better improvement in terms of average precision, 101.94% of the monolingual retrieval, and is thus an effective method for information retrieval. This suggests that ranking and selecting pairs of source query is very helpful for statistical disambiguation, especially for long queries containing at least four terms. Query expansion before translation *Feed.bef* did not improve the average precision; however, after trans_disambiguation *Feed.aft* increased the average precision, 101.33% of the monolingual counterpart. Combined relevance feedback techniques before and after trans_disambiguation *Feed.bef_aft* showed the best result, 102.89% of the monolingual counterpart. This suggests that combined query expansion before and after the proposed translation and disambiguation method improves the effectiveness of information retrieval, using log-likelihood ratio as an estimation of co-occurrence tendency. Domain-based feedback showed a drop in terms of average precision compared to previous methods. However, when combined with relevance feedback before and after trans_disambiguation, average precision increased to 103.69%. Thesaurus-based expansion using WordNet (*Feed_wn*) or EuroWordNet (*Feed.ewn*) improved the recall but resulted in a reduction in average precision. Combined relevance feedback, domain-based feedback and/or thesaurus-based expansion (*Feed.bef_wn, Feed.ewn_aft, Feed.dom_wn* and *Feed.ewn_dom*) reduced average precision but improved recall slightly. On the other side, a combined thesauri-based expansion showed a better result; however, average precision was again reduced. The best result was achieved using the combined thesauri-based expansion and domain-based feedback *Feed.ewn_wn_dom*, which resulted in an average precision of 104.29% compared to the monolingual counterpart. This suggests that adding domain keywords to generalized thesauri improves the effectiveness of retrieval. A statistical t-test[10] was used to evaluate whether the improvement by method X over method Y is significant, by computing a *p-value*. The smaller the *p-value*, the more significant is the improvement. In general, if the p-value is small enough (p-value < 0.05, i.e., less than 5%), we can conclude that the improvement is statistically significant in studies of this type. The following algorithm depicts the ***Paired t-test:***

Let X_i and Y_i be the scores of retrieval methods X and Y for query i where $i = 1 \ldots n$.

We define $D_i = Y_i - X_i$.

$$t = \frac{\bar{D}}{s(D_i)/\sqrt{n}}$$

where

$$\bar{D} = \frac{1}{n}\sum_{i=1}^{n} D_i,$$

$$s(D_i) = \sqrt{\frac{1}{n-1}\sum_{i=1}^{n}(D_i - \bar{D})^2}$$

Table 6 Paired sample t-test significance values.

Method	P-value	Degree of Significance
RSDT	0.005	(0.5%)
Feed.bef_aft	0.0075	(0.75%)
Feed.bef_dom	0.0096	(0.96%)
Feed.ewn_wn_dom	0.0112	(1.12%)

Table 6 shows degrees of significance for the proposed disambiguation and expansion methods over the baseline or monolingual retrieval. The improvement by using the proposed disambiguation method RSDT is statistically significant (p-value = 0.005) compared to the monolingual retrieval. Combination of pseudo-relevance feedback before and after translation Feed.bef_aft is statistically significant (p-value = 0.007) over the monolingual retrieval. Combinations of pseudo-relevance and domain-based feedback before and after translation Feed.bef_dom is statistically significant (p-value = 0.009) compared to the monolingual retrieval. Finally, the improvement obtained with the combination of thesauri-based and domain-based expansion techniques Feed.ewn_wn_dom is statistically significant (p-value = 0.011) compared to the monolingual retrieval. These results confirm our previous conclusion on the effectiveness of the proposed disambiguation method, the combined pseudo-relevance feedback with domain-based feedback and the combined thesauri with domain-based.

Thus, the key techniques used in the proposed methods can be summarized as follows:

- Combined statistical disambiguation based on co-occurrence tendency was applied first prior to the translation in order to eliminate misleading pairs of terms to translate and disambiguate, and then subsequent to translation in order to avoid incorrect sense disambiguation and select best target translations.
- Ranking and careful selection are indispensable for the success of a query translation.
- Log-likelihood ratio was found to be an efficient estimation for query disambiguation, when evaluated using all terms of the original query, rather than using just one query term.
- Adding domain keywords to the original query and then selecting thesaurus word senses, in order to avoid wrong sense disambiguation, is considered to be a kind of Word Senses Disambiguation (WSD).
- Each type of query expansion has different characteristics, and therefore different combinations of these expansion techniques provides a valuable resource for query expansion and allows an improvement in terms of average precision.
- The present results showed that CLIR could outperform monolingual retrieval. The

combination of different methods for query disambiguation and expansion before and after translation has confirmed that monolingual performance is not necessarily the upper bound for CLIR performance[7]. These methods have completed each other and the proposed query disambiguation had a positive effect during the translation and thus retrieval. In addition, combination of query expansion largely affected the translation because related words could be added.

The two proposed disambiguation methods before and after translation were based on the selection of one target translation to retrieve documents. Setting a threshold to select more than one target translation is possible using a weighting scheme for the selected target translations, in order to eliminate misleading terms and optimize the query to retrieve documents.

6. Conclusions and Future Works

Dictionary-based translation has been widely used in CLIR because of its simplicity and the increasing availability of MRDs. However, ambiguity arising due to failure to translate queries is largely responsible for large drops in effectiveness below monolingual performance[1]. Our approach in CLIR combines various types of query expansion, before and after translation and disambiguation processes. In the present paper, we proposed and evaluated efficient disambiguation methods for short and long queries, applied using all terms of an original query rather than with just one query term. These methods provided valuable resources for query translation and thus information retrieval. In addition, a new technique based on relevance feedback and hierarchical category schemes in order to extract domain keywords from the set of top-retrieved documents was proposed. This expansion technique was found to be effective when combined with pseudo-relevance feedback before and after translation and disambiguation. Moreover, simple and efficient weighting schemes based on co-occurrence tendency were proposed. These weighting schemes depend on the structure of the expansion technique. Finally, combining thesauri-based expansion with domain-based feedback showed the greatest improvement for information retrieval. Detailed evaluation for the statistical test of significance will be performed in the near future using comparisons between different CLIR methods with one another.

Our ongoing work involves a deeper investigation of the different relations of WordNet and EuroWordNet thesauri, in addition to synonymy. We would like to investigate the use of multiple word senses for query expansion. A weighting scheme to select relevant expansion terms and eliminate misleading terms so as to construct an optimal query will be proposed in a future study. For the domain-based feedback, an approach involving learning from documents categorization or classification, not necessarily web documents, in order to extract relevant keywords for a query expansion, is among the subject of future research. Setting a threshold to select more than one target translation should be considered. Finally, our primary goal is to find more effective solutions for information retrieval across languages.

Acknowledgments The present research is supported in part by the Ministry of Education, Culture, Sports, Science and Technology of Japan and by the CREST program of the JST Corporation (Japan Science and Technology). The authors would like to thank Dr. Koji EGUCHI and all anonymous reviewers for their valuable comments regarding this paper. The authors are also grateful to Dr. Eric GAUSSIER and Dr. Herve DEJEAN, XRCE (XEROX Research Centre, Europe) and Dr. Masaki MURATA, Communications Research Laboratory (CRL, Japan) among other researchers for their fruitful comments on this research.

References

1) Ballesteros, L. and Croft, W.B.: Phrasal Translation and Query Expansion Techniques for Cross-Language Information Retrieval, *Proc. ACM SIGIR'97*, pp.84–91 (1997).

2) Ballesteros, L. and Croft, W.B.: Resolving Ambiguity for Cross-Language Retrieval, *Proc. ACM SIGIR'98*, pp.64–71 (1998).

3) Church, K.W. and Hanks, P.: Word Association Norms, Mutual Information, and Lexicography, *Computational Linguistics*, Vol.16, No.1, pp.22–29 (1990).

4) de Loupy, C., Bellot, P., El-Bèze, M. and Marteau, P.-F.: Query Expansion and Classification of Retrieved Documents, *Proc. 7th Text REtrieval Conference* (*TREC-7*), pp.443–450 (1998).

5) Dunning, T.: Accurate Methods for the Statistics of Surprise and Coincidence, *Computational Linguistics*, Vol.19, No.1, pp.61–74 (1993).

6) Gale, W.A. and Church, K.: Identifying Word

Correspondences in Parallel Texts, *Proc. 4th DARPA Speech and Natural Language Workshop*, pp.152–157 (1991).

7) Gao, J., Xun, E., Zhou, M., Huang, C., Nie, J.-Y. and Zhang, J.: Improving Query Translation for Cross-Language Information Retrieval Using Statistical Models, *Proc. ACM SIGIR2001*, pp.96–104 (2001).

8) Harman, D.: Overview of the first TREC conference, *Proc.ACM SIGIR'93*, pp.36–47 (1993).

9) Hull, D. and Grefenstette, G.: Querying Across Languages — A Dictionary-based Approach to Multilingual Information Retrieval, *Proc. ACM SIGIR'96*, pp.49–57 (1996).

10) Hull, D.A.: Using Statistical Testing in the Evaluation of Retrieval Experiments, *Proc. ACM SIGIR'93*, pp.329–338 (1993).

11) Hull, D.A.: A Weighted Boolean Model for Cross-Language Text Retrieval, *Cross-Language Information Retrieval*, Grefenstette, G. (Ed.), chapter 10, pp.119–135, Kluwer Academic Publishers (1998).

12) Maeda, A., Sadat, F., Yoshikawa, M. and Uemura, S.: Query Term Disambiguation for Web Cross-Language Information Retrieval using a Search Engine, *Proc. 5th International Workshop on Information Retrieval with Asian Languages* (*IRAL2000*), pp.25–32 (2000).

13) Mandala, R., Tokunaga, T. and Tanaka, H.: Combining Multiple Evidence from Different Types of Thesaurus for Query Expansion, *Proc. ACM SIGIR'99*, pp.191–197 (1999).

14) Mandala, R., Tokunaga, T. and Tanaka, H.: The Exploration and Analysis of Using Multiple Thesaurus Types for Query Expansion in Information Retrieval, *Natural Language Processing*, Vol.7, No.2, pp.117–140 (2000).

15) Nie, J., Simard, M., Isabelle, P. and Durand, R.: Cross-Language Information Retrieval based on Parallel Texts and Automatic Mining of Parallel Texts from the Web, *Proc. ACM SIGIR'99*, pp.74–81 (1999).

16) Oard, D.W.: Alternative Approaches for Cross-Language Text Retrieval, *Working Notes of the AAAI Symposium on Cross-Language Text and Speech Retrieval* (1997).

17) Richardson, R. and Smeaton, A.F.: Using WordNet in a Knowledge-Based Approach to Information Retrieval, *Proc. 17th BCS-IRSG Colloquium on Information Retrieval* (1995).

18) Rocchio, J.: Relevance Feedback in Information Retrieval, *The SMART Retrieval System — Experiments in Automatic Document Processing*, Salton, G. (ed.), pp.313–323, Prentice Hall (1971).

19) 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, pp.61–68 (2000).

20) Sadat, F., Maeda, A., Yoshikawa, M. and Uemura, S.: Query Expansion Technique for the CLEF Bilingual Track, *Working Notes for the CLEF 2001 Workshop*, pp.99–104 (2001).

21) Salton, G.: Experiments in Multi-Lingual Information Retrieval, *Inf. Process. Lett.*, Vol.2, No.1, pp.6–11 (1973). TR 72-154 at http://cs-tr.cs.cornell.edu.

22) Salton, G. and Buckley, C.: Term-weighting approaches in automatic text retrieval, *Inf.Process.Manage.*, Vol.24, No.5, pp.513–523 (1988).

23) Salton, G. and McGill, M.J.: *Introduction to Modern Information Retrieval*, McGraw Hill (1983).

24) Salton, G.: Automatic Processing of Foreign Language Documents, *J. Am. Soc. Inf. Sci.*, Vol.21, No.3, pp.187–194 (1970).

25) Utsuro, T., Ikeda, H., Yamane, M., Matsumoto, Y. and Nagano, M.: Bilingual Text Matching using Bilingual Dictionary and Statistics, *Proc. International Conference on Computational Linguistics*, pp.1076–1082 (1994).

26) Voorhees, E., M.: Query expansion using Lexical-Semantic Relations, *Proc. ACM SIGIR'94*, pp.61–69 (1998).

27) Vossen, P.: EuroWordNet: a Multilingual Database for Information Retrieval, *Proc. DELOS workshop on Cross-language Information Retrieval*, Zurich, Switzerland (1997).

28) Vossen, P. (ed.): *EuroWordNet: A Multilingual Database with Lexical Semantic Networks*, Kluwer Academic Publishers (1998).

29) Yamabana, K., Muraki, K., Doi, S. and Kamei, S.: A Language Conversion Front-End for Cross-Linguistic Information Retrieval, *Proc. SIGIR Workshop on Cross-Linguistic Information Retrieval*, pp.34–39 (1996).

(Received March 20, 2002)
(Accepted July 4, 2002)

（Editor in Charge *Koji Eguchi*）

Fatiha Sadat is a doctoral student at Nara Institute of Science and Technology. She received a degree of Engineering in computer science from University of Science and Technology Houari Boumédiène, Algiers, Algeria in 1994 and M.E. degree in Information Science from Nara Institute of Science and Technology in 2000. She was a research visitor at XEROX Research Centre, Europe for three months in 2001. Her current research interests include Cross-Language Information Retrieval, information retrieval, multilingual information processing, computational linguistics, and natural language processing.

Akira Maeda is an associate professor at the Department of Computer Science, Ritsumeikan University. He received B.A. and M.A. degrees in Library and Information Science from University of Library and Information Science in 1995 and 1997, respectively, and received Ph.D. degree in Engineering from Nara Institute of Science and Technology in 2000. He was a visiting scholar at the Virginia Polytechnic Institute and State University from 2000 to 2001. He has won the IPSJ Best Paper Award in 1999. His research interests include multilingual information processing, information retrieval, and digital libraries. He is a member of ACM, IEICE and IPSJ.

Masatoshi Yoshikawa received the B.E., M.E. and Ph.D. degrees in Information Science from Kyoto University in 1980, 1982 and 1985, respectively. From 1985 to 1993, he was with Kyoto Sangyo University. In 1993, he joined Nara Institute of Science and Technology as an Associate Professor of Graduate School of Information Science. In June 2002, he became a full professor of Information Technology Center, Nagoya University. His current research interests include XML databases, databases on the Web, and multimedia databases. Dr. Yoshikawa is an Area Editor of Information Systems (Elsevier/Pergamon). He is a member of ACM, IEEE Computer Society and IPSJ.

Shunsuke Uemura is a professor at the Graduate School of Information Science, Nara Institute of Science and Technology, Japan. He has been a member of the research staff at MITI's Electrotechnical Laboratory, a visiting researcher at MIT's Electronic Systems Laboratory and a professor at the Tokyo University of Agriculture and Technology. He received a BE, an ME, and a Dr. Eng. from Kyoto University in 1944, 1946 and 1975 respectively. His research interests include database systems and bioinformatics. He is a senior member of the IEEE Computer Society, and a member of ACM, IEICE and the fellow of the Information Processing Society of Japan.