[TMI 99: *Proceedings of 8th International Conference on Theoretical and Methodological Issues in Machine Translation*, August 1999, Chester, UK]

Experiments in Translingual Information Retrieval Using Web-based MT and WordNet

Krzysztof Czuba, Xin Liu
Language Technologies Institute, Carnegie Mellon University
4910 Forbes Avenue, Cyert Hall 254
Pittsburgh, PA 15213, US
kczuba,xliu@cs.cmu.edu

Abstract

Yang *et al.* (1998) suggest a set of corpus-based Translingual Information Retrieval (TLIR) methods and show that they produce good results in comparison to the analogous Monolingual Information Retrieval (MLIR) methods. However, all the TLIR methods require a substantial training document set that must be available in the form of a parallel corpus. In this paper we present results of experiments in which such a parallel corpus was obtained using a publically available MT system. The results show that high retrieval precision can be achieved in English-Spanish cross-lingual retrieval even though a machine translated corpus was used for training.

1 Introduction

Translingual Information Retrieval (TLIR) has been enjoying quite a bit of popularity lately, mostly due to the growing number of documents in various languages available on the Internet. The main idea of TLIR is to provide the user with a means to retrieve documents in a target language (TL) different from the language in which the query is formulated (source language, SL). Retrieval of this kind can be performed e.g. by a librarian doing search in a foreign language or by a user whose passive knowledge of a language does not allow for an effective query formulation in this language, but is sufficient for reading the retrieved documents.

Since the document collection and the query are in different languages, the main task that needs to be solved in order to perform TLIR consists in crossing the language barrier. This can be done in different ways, and it can involve human or machine translation. Translating the query would be one way of crossing the language barrier. However, such a method is not likely to produce good results due to the limited context that is available. This might be particularly problematic for MT-based query translation. Since queries are usually lists of words, it is likely that due to the lack of context the wrong senses of query terms will be picked by the MT system.

In this paper we concentrate on a group of methods considered by Yang *et al.* (1998) that are based on parallel corpora. As Yang *et al.* (1998) show, those methods produce results comparable to Monolingual Retrieval (MLIR) on the same collection, and they achieve results better than simpler methods using dictionary-based query expansion. However, since the size of the required parallel corpus is significant, such corpora-based methods might not offer a cost effective solution when a human translation of a large body of text is required. In this paper we show that it might be sufficient to use a corpus that has been automatically translated. We used the publically available AltaVista translation engine that is powered by the SYSTRAN^{TM} system. SYSTRAN^{TM} has been used in cross-lingual IR before Gachot (1998), but we are not aware of any large scale experiments with the on-line version of the system, which is available to anyone with Web access.

We realize that in a way our approach might appear like a suggestion to solve the IR problem by solving first the more difficult MT problem. However, we do not suggest that a TLIR system be set up by first building a broad-coverage MT system for the respective pair of languages. What we rather intended to show was that given a set of corpus-based TLIR methods, it is possible to set up a TLIR system using an existing MT system in a cost effective way and with good results.

As Collier *et al.* (1998) notice, MT-translated corpora can be viewed as a form of noisy parallel text. This kind of input has also been considered for other NLP applications Fung (1996). Collier *et al.* (1998) use MT in their English-Japanese news article alignment task that is analogous to TLIR in their set-up. They observed a little improvement when MT was used to translate Japanese texts into English in comparison to a dictionary term lookup method. They attribute the improvement to the presence of context in the MT case that allows for at least partial word sense disambiguation.

Another application of MT to TLIR is described by Oh-Woog Kwon *et al.* (1997). They used a lexical transfer system that translates Japanese to Korean to retrieve and translate Japanese patent documents using Korean queries. Although Oh-Woog Kwon *et al.* (1997) did not present any quantitative results, they write that the system "has been successfully in full operation". In their approach, MT is used to translate all documents in the collection and to obtain Korean indexing terms for Japanese documents. The authors attribute the success of their system to the relatively high translation quality that can be obtained for a pair of similar languages like Korean and Japanese.

Our approach is slightly different from what Collier *et al.* (1998) and Oh-Woog Kwon *et al.* (1997) suggest. We do not require the MT system to be part of the retrieval environment. The translation is off-line and only a part of the collection has to be translated. We also present qualitative results based on the UNICEF collection and the set of relevance judgments collected by Yang *et al.* (1998). Since we kept the same experiment set-up as in Yang *et al.* (1998), it is possible to assess the impact of the MT-translated corpus on the retrieval precision.

2 Retrieval Methods

In this section we describe the retrieval methods that we used. Since all of the methods are identical to the ones in Yang *et al.* (1998), we based this section on their description. The reader can consult Yang *et al.* (1998) for more details.

The dictionary method makes use of a bilingual dictionary to cross the language barrier (see e.g., Hull *et al.* (1996)). It is a query expansion method that produces a set of TL terms given a query in the SL by looking up all the query terms in a bilingual dictionary. All the resulting translations are then added to the TL query that is used to perform MLIR in the TL. Since MRDs of significant size for different language pairs are becoming more easily available and the method is easy to implement, we decided to adopt it as our baseline.

Performance of the dictionary method (see e.g., Davis (1996)) strongly depends on the quality of the bilingual dictionary. Also, in its simplest form, it will suffer from the general problem faced by most query expansion methods: since at least some of the terms in the SL query have multiple senses, a simple dictionary look-up will return all the translations that correspond to all possible senses. Unless we build a collection-specific thesaurus or do some sense disambiguation, the resulting query will be too general. This might happen when wrong senses are added to the translation and it will result in lower retrieval precision. It is also possible that not all relevant documents will be detected when some translations are missing from the dictionary, resulting in lower recall.

Some extensions of the basic dictionary method have been proposed in the literature and they mostly considered better ways of constructing the dictionary. A general purpose bilingual dictionary might not serve the purpose well, especially when a specialized domain is considered. In such a case, a corpus-specific thesaurus might offer a better solution (Sheridan *et al.* (1996), Brown (1998)). In particular, the EBT method of Yang *et al.* (1998) offers the highest retrieval precision in their experiments.

The methods that we use following Yang *et al.* (1998) attempt to circumvent the above difficulties[1] by using a parallel corpus to cross the language barrier. All the methods are extensions of existing MLIR methods and are based on the Vector Space Model (VSM) Salton (1989). VSM assumes that both queries and documents are represented as vectors of term weights. A similarity measure is defined between two vectors, or between a query and a document. The higher this similarity value, the closer the document is assumed to be to the query and the higher it is ranked in retrieval. For a corpus with m distinct index terms, a query vector $\vec{q}$, a document vector $\vec{d}$ and the similarity measure between them can be expressed using the term weights q_i and d_i as follows Yang *et al.* (1998):

$$\vec{q} = (q_1, q_2, \ldots, q_m)^t, \quad \vec{d} = (d_1, d_2, \ldots, d_m)^t$$

$$sim(\vec{q}, \vec{d}) = \cos(\vec{q}, \vec{d}) = \frac{\sum_{i=1}{}^{m} q_i d_i}{\sqrt{\sum^{m}{}_{i=1} q_i{}^2}\sqrt{\sum_{i=1}{}^{m} d_i{}^2}}.$$

Different weighting and normalization schemes are possible (ntc, atc, ltc and others), and they usually lead to different retrieval results Salton (1989).

The first method that we applied in our experiments is an extension of the Pseudo-Relevance Feedback (PRF) Buckley (1995) to the translingual case. In the translingual version of PRF, the SL query is used to retrieve from the SL part of the training corpus a number of documents in the SL. These documents are then assumed to be relevant to the query and their counterparts in the TL are retrieved from the parallel training corpus. In the next step, a number of most frequent terms is picked from the retrieved

[1]Yang *et al.* (1998) discuss other methods as well. We picked only the ones that were shown to produce high precision results.

TL documents. These terms combined constitute the SL query that is used to perform monolingual retrieval in the SL. Thus, the translingual version of PRF is a form of query expansion that has the following parameters: the number of documents retrieved for feedback (K) and the number of terms that are taken from each of these documents (N). The monolingual retrieval can be done using any method. In our case we used the SMART retrieval system which is based on the VSM. The criterion for document ranking is then as follows Yang *et al.* (1998):

$$\vec{q_t} = \sum_i \{\vec{g_i} | \vec{d_i} \in \text{kNN}(\vec{q_s})\}, \quad sim(\vec{q_s}, \vec{d_t}) = \cos(\vec{q_t}, \vec{d_t})$$

where $\text{kNN}(\vec{q_s})$ is the set of the k highest ranked documents retrieved with the original query, $\vec{d_i}$ is a document in the SL, $\vec{g_i}$ is the corresponding document in the TL.

The next two methods, Latent Semantic Indexing and the Generalized Vector Space Model Wong *et al.* (1985) are direct extensions of the VSM. Since a precise description of these methods is beyond the scope of this paper, the reader is encouraged to consult Yang *et al.* (1998) and the reference therein. Both GVSM and LSI try to take into account some of the semantic correlation that occurs among terms in documents. GVSM tries to achieve that goal by representing terms using the documents as the space basis. LSI goes one step further and represents the indexing terms in a reduced space defined by a linear combination of documents.

Both GVSM and LSI can be described in terms of operations on the term-document matrix. For a training corpus with m terms and n documents, the term-document matrix $A_{m \times n}$ can be seen as a representation of documents using terms (VSM) or a representation of terms using documents (the dual space Sheridan *et al.* (1996), used in GVSM and LSI). Under the assumption that semantically related terms tend to co-occur in a document collection, the representation of terms using row vectors in the term-document matrix can potentially assign similar patterns to semantically related terms.

In GVSM, the term-document matrix is used to represent both the query and the documents in the dual space. For the query this is achieved by computing $\vec{q'} = A^t\vec{q}$, for a document by computing $\vec{d'} = A^t\vec{d}$. Both $\vec{q'}$ and $\vec{d'}$ are now in the dual space and can be compared, e.g., using the cosine measure: $sim(\vec{q}, \vec{d}) = cos(A^t\vec{q}, A^t\vec{d})$.

Yang *et al.* (1998) proposed a novel extension of GVSM to handle the translingual case. They extended the term-document matrix by adding to the SL term-document matrix A the TL term document matrix B in such a way that the corresponding columns in A and B represent parallel document pairs in the training corpus. Thus, vectors $\vec{q'}$ and $\vec{d'}$ share the common base (dual, i.e. document, space) and can be compared even though $\vec{q}$ and $\vec{d}$ originated from documents in different languages. The similarity measure in GVSM can be expressed as: $sim(\vec{q}, \vec{d}) = cos(A^t\vec{q}, B^t\vec{d})$.

In the monolingual version of LSI Deerwester at al (1990), the dimensionality of the space defined by the term-document matrix is first reduced by assuming a new basis that consists of "the most meaningful linear combinations of documents" Yang *et al.* (1998). This new space is claimed to provide a better representation of the semantic content of terms. The reduction is done using singular value decomposition. The new document space is spanned by the orthogonal singular vectors. Both the query and the documents are expressed in this new basis and compared:

$$A = U\Sigma V^t \quad \text{(SVD)}, \quad sim(\vec{q}, \vec{d}) = cos(U^t\vec{q}, U^t\vec{d}).$$

In the translingual version of LSI Dumais (1996), the term-document matrix is extended in the same way as in the translingual GVSM. The similarity is computed as follows:

$$\begin{bmatrix} A \\ B \end{bmatrix} = U_2\Sigma_2{V_2}^t \quad \text{(SVD)}, \quad sim(\vec{q}, \vec{d}) = cos({U_2}^t\vec{q}, {U_2}^t\vec{d})$$

Similarly to PRF, both LSI and GVSM have tunable parameters. In the case of LSI, the number of singular vectors (SV) that are used in the new basis can be adjusted. In the case of GVSM, the number of elements in each document vector (matrix sparsification, SP) can be specified.

3 Experiment design and evaluation

In order to evaluate the performance of the above methods with an automatically translated corpus, we performed a set of experiments that precisely mimicked the experiments described by Yang *et al.* (1998) in which we replaced the human translated corpus with an MT translation. In this way, a direct comparison and an assessment of the influence of the MT translation on the results were possible. Since the set-up was identical to the one used by Yang *et al.* (1998), we based the next paragraphs on their description and the reader is referred to Yang *et al.* (1998) for further details.

The methods were evaluated on a set of 30 queries using three sets of documents: the training set, the validation set, and the test set (all obtained from Yiming Yang). All three sets were taken from the same parallel corpus, called the UNICEF collection. The UNICEF collection contains 2255 document pairs that were taken from the UN Multilingual Corpus Graff *et al.* (1994). The documents are available in both English and Spanish. They pertain to UNICEF reports and deliberations and are fairly general in terms of topics and vocabulary. Documents are split into paragraphs with the average of 9 paragraphs in a document. Out of the 2255 document pairs, 1134 pairs were randomly selected and used as the training set. 550 pairs were kept for the validation set, and 571 pairs were added to the test set.

The training set was used as the corpus the methods utilize to cross the language barrier. In the PRF method, the initial set of SL documents and their TL counterparts were retrieved from the training set. In LSI and GVSM, the term-document matrices were created using documents from the training set. The validation set was used to tune the methods' parameters, i.e., all the methods were run with different parameter settings and the setting was chosen for which the 11-pt average precision was highest. Then, all the methods were run with the best setting on the unseen data in the test set. The resulting 11-pt average precision is reported in this paper. The computation of the 11-pt average precision was based on the set of human relevance judgments collected and described by Yang *et al.* (1998). The 11-pt average precision is the interpolated average of precision values with thresholding at recall levels of 0%, 10%, ..., 100% Salton (1989).

As in Yang *et al.* (1998), we experimented with three different training set alignments: sentence alignment, paragraph alignment, and document alignment. The test set was always aligned according to documents.

The bilingual corpus[2] was created by using the AltaVista translation service on the Web (`http://babelfish.altavista.digital.com`) that is driven by the SYSTRAN^{TM} translation engine. Translations from English to Spanish were obtained automatically for the English UNICEF collection for both documents and paragraphs. 14452 paragraphs, 2255 documents were translated and split into test and training sets. The alignment of documents and paragraphs was precisely the same as in the experiments Yang *et al.* (1998) ran, which makes for a good setting for assessing the effect of using a machine translated corpus.

The translations were obtained for single paragraphs that were then combined into documents. The sentence-aligned corpus was obtained from the original English paragraph corpus and its SYSTRAN^{TM} translation. Sentences were extracted only from paragraphs that were easily aligned in both languages using a very simple alignment algorithm. This resulted in a sufficiently clean parallel corpus aligned by the sentence. However, this corpus is not an exact copy of the one used by Yang *et al.* (1998) since a slightly different criterion was used for splitting paragraphs into sentences. In particular, paragraphs that did not align because of a small translation problem or time out were discarded completely, which resulted in a different number of sentences. The sentence-aligned corpus that we generated has 19935 sentence pairs as compared to 20917 sentence pairs in the corpus Yang *et al.* (1998) used (4.7% smaller).

The translation engine sometimes times out. There were on the order of 500 such cases. Time-outs are likely to occur for long sentences or complex phrases and some words might be missing from the translation. In general, there is no way to prevent time outs from happening in a practical setting like this, although it might be possible if the translations were obtained not from the Web but directly from SYSTRAN. The sentences in which a time-out occurred were kept in the corpus. Given the size of the corpus, the effect of timed-out sentences (less than 3% of the total number of sentences, most of them were partially translated) should be negligible.

Although SYSTRAN^{TM} has an impressive lexicon, some of the words remained untranslated, i.e. some English words occur in the Spanish translation. Although we have not checked how many English words were inserted, from looking at the corpus the percentage of such words seems to be negligible.

4 Potential problems with an MT derived corpus

Although using MT is a potentially cheap way of deriving a bilingual corpus, the current MT technology does not allow for high quality translation to be obtained from general-purpose MT. In general-purpose MT, specialized terms might not be present

[2]The discussion in this section is not meant to be an evaluation of the quality of the translations that are produced by the AltaVista service and it is not possible to judge this quality on the basis of the results presented in this paper.

in the system's dictionary for cost reasons or might be omitted from it to prevent often unnecessary ambiguity resolution.

Another potentially serious flaw is the paraphrase problem: most, or even all, current MT systems choose the same translation for a word in a particular sense regardless of how repetitive the translation is. The words *aircraft* and *airplane* are a simple example: in a non-technical document they are likely to be treated as synonyms and only one will be consistently picked by an MT system in a given domain. Since both can be used in a query, a risk that one of them will not be an efficient query term is quite high. This might hurt the performance of IR methods since very often they rely on the different synonyms being present in documents/queries. This particular problem clearly cannot be solved by automatically creating a larger corpus as it might be the case for a human translated corpus. However, source language query expansion can be helpful here.

One advantage over the basic dictionary-based query expansion that can be expected from an MT-derived corpus is the presence of context. This should have a positive effect in comparison to methods that (automatically) translate queries or are based on simple, non-discriminative dictionary look-up. It is very likely that syntactic disambiguation will be quite successful and this should be beneficial especially in the case of methods like PRF. Also, some semantic disambiguation might be done by the MT system. In an ideal case, the MT system would perform some discourse level analysis leading to even more successful disambiguation. However, in practice current MT systems operate mostly at the sentence level and only very local context is used.

In addition to these issues, some purely technical difficulties can occur. They can include problems with coverage of both lexical items and grammatical structures, inability to handle lengthy sentences, etc.

It should be also pointed out that there should be a match between the topics that occur in the collection to be translated using MT and the type of MT system used. This is similar to the requirement that the training set in all the considered TLIR methods resemble the test set in terms of topics covered, vocabulary and style.[3] The UNICEF collection might be especially suitable for MT translation. It consists of documents containing mostly general terminology with some medical and UNICEF/UN specific terms. In particular, it does not contain technical documents requiring specialized vocabulary and disambiguation methods.

5 Application of WordNet

A form of query expansion might be useful to alleviate the paraphrase problem mentioned above. Adding all synonyms of the terms in the query is one form of expansion. Although this task would usually be accomplished by all the discussed methods when run with a sufficiently large human-translated corpus, it might not be the case when an MT-translated corpus is used. In the current setting, it would probably be best to expand the TL query, but this might be difficult in the case of LSI and GVSM in which the TL query is implicit. Instead, we considered ways of expanding the SL query.

[3] It might be possible to obtain a machine translation of a highly specialized collection by simply fine tuning the MT system (e.g., by adding domain specific terms to the dictionary).

	GVSM			LSI			PRF		
	atc	ltc	ntc	atc	ltc	ntc	atc	ltc	ntc
paragraph-alignment									
Q1	0.3638	0.3860	**0.4229**	0.3316	0.3653	**0.4165**	0.3420	0.3866	**0.4254**
Q2	0.3551	0.3751	0.3832	0.3238	0.3665	0.3988	0.3439	0.3671	0.4046
Q3	0.3498	0.3796	0.3827	0.3090	0.3467	0.3712	0.3280	0.3688	0.4032
document-alignment									
Q1	0.3443	0.3843	**0.4231**	0.3829	0.4222	**0.4428**	0.3120	**0.3797**	0.3779
Q2	0.3327	0.3728	0.3942	0.3682	0.4192	0.4214	0.3010	0.3631	0.3643
Q3	0.3689	0.3733	0.3900	0.3170	0.4251	0.4285	0.2911	0.3605	0.3643
sentence-alignment									
Q1	0.3030	0.3141	0.3231	0.2828	0.2990	0.3001	0.2993	0.3677	**0.4047**
Q2	0.3238	0.3394	0.3555	0.2705	0.2974	0.3162	0.2878	0.3560	0.3747
Q3	0.3263	0.3376	0.3557	0.2690	0.2969	0.3112	0.2876	0.3490	0.2876

Table 1: 11-pt average precision on the validation set

By expanding the SL query we hoped to add to it synonyms that would get different translations in the TL from the original query terms thus increasing the chance that relevant documents are ranked higher.

As with most query expansion methods, it is to be expected that the precision value will drop at the cost of increased recall. Since in our evaluation the recall is always 100% (all the documents are ranked by the SMART system, i.e., all relevant documents will be retrieved; the 11-pt average precision is a measure of how far in the ranking the relevant documents were found), we should expect a drop in the 11-pt average precision due to some spurious documents being ranked higher on the basis of wrong senses picked during expansion. However, we were lead to believe by our initial experiments that the expansion might improve the 11-pt average precision. In these experiments, we did not have the validation and the test set. More precisely, the results were not cross-validated on an unseen test data. In this case, we found that for some methods (LSI, GVSM), query expansion as described below leads to higher 11-pt average precision. However, in the experiments we describe in this paper, this observation did not bear out. In all cases, the 11-pt average precision dropped when query expansion was used. Nevertheless, we present the results with expanded queries for comparison.[4]

The method we used to expand the queries was based on WordNet Fellbaum (ed.) (1998). The original query set (Q1) was expanded using the WordNet synonymy relation. We constructed two new query sets, Q2 and Q3. Q2 contains all the terms

[4]In the initial experiments, the parameters were tuned on a document set which was the sum of the current validation and test sets. It would be interesting to find out why the precision went up with query expansion. We suspect that for some queries in the query set, the expansion is particularly effective and it would be interesting to find those queries. The size of the corpus is most probably the reason why this effect is observed. In the case of the validation and test sets, the number of queries for which the validation set contains relevant documents (26) is different from the number of queries for which relevant documents can be found in the test set (23).

GVSM			LSI			PRF		
Q1	Q2	Q3	Q1	Q2	Q3	Q1	Q2	Q3
0.4191	0.3741	0.4105	**0.4450**	0.4202	0.4159	**0.3966**	0.3714	0.4138

Table 2: Results on the test set, best settings, all alignments

	best MT	best human	% best human	% dictionary
GVSM	0.4191	0.4585	90.75	107.43
LSI	0.4450	0.4234	105.1	114.07
PRF	0.3966	0.4203	94.36	101.66

Table 3: Results on the test set with the best settings overall including alignment

from the original queries. In addition, for each term we added the first (most frequent) synonymous term returned by WordNet. Q3 was created similarly by adding all the synonymous senses returned by WordNet.

Since we did not get better retrieval results, we did not refine the above rather crude method. In particular we did not use phrases that WordNet sometimes returns. Also, in Q3 we kept all the terms returned by WordNet, regardless of their part of speech. In particular, some common adverbs (*fairly*, *thoroughly*) were added to the expanded queries. Such items usually lead to lower precision values.

6 Results

Table 1 presents comprehensive results of our experiments. The values were obtained by running the methods with different parameter settings. Different alignments and weighting schemes are shown. As discussed in section 2, the methods have additional parameters. The values shown correspond to the settings for which the 11-pt average precision was highest. The values in bold face correspond to the overall best parameter settings including the alignment and term weighting. Table 2 contains the 11-pt average precision for all the methods on the test set on different query sets with the settings underlined in Table 1.

We found that GVSM and LSI performed best on the document alignment with the ntc term weighting. PRF performed best on the paragraph alignment, with the same weighting. The sparsification SP for GVSM was 80 and the number of singular values SV in LSI was 400[5]. For PRF, K was 10 and N was 5 was.

Table 2 shows the 11-pt average precision for the best settings (the ones corresponding to the bold face figures in Table 1; includes the alignment). Table 3 show the final comparison with the results reported in Yang *et al.* (1998) (given in the 'best human' column). As can be seen the effect of using an MT translated corpus is rather surprising. All methods performed well. Although Yang *et al.* (1998) report GVSM to perform best, it suffers a 10% hit in precision in our experiments. The precision of PRF

[5]Arguably, this is a very high number of singular values leading to expensive computation. The next best result was for SV = 180, 0.4365 on training set, 0.4241 on test set

drops by only about 5%. Surprisingly, LSI performed better on the machine translated corpus than on the human translated corpus. We do not have a good explanation of this result. The last column contains a comparison with the dictionary method as run by Yang *et al.* (1998). The additional effort required to implement the technique described here results in an improvement over the dictionary method that is particularly strong for GVSM and LSI.

7 Conclusions

As the above results show, general purpose MT can be used to set up in a cost effective way a new TLIR system using corpus-based TLIR methods. However, as in many other IR experiments, the results cannot be easily generalized to other query sets, document collections, and other language pairs. It would be therefore desirable to carry out a similar set of experiments on a different query and document sets. The only other collection that we have access to is the MEDLARS collection. Since this collection was also used by Yang *et al.* (1998), a direct assessment of the impact the MT-translated corpus has on the results would be possible. However, we do not think that the collection is suitable for the approach described here, mostly due to the large number of technical terms from the medical domain that most likely would not be translated reliably by a general purpose MT engine. To what extent this would hurt the precision value remains to be determined.

Another issue that arises is the actual benefit from doing the translation. Although we believe that the syntactic disambiguation performed by the MT system leads to a better correspondence between terms than in the case of a simple dictionary look-up, we cannot rule out the possibility that the relatively good results are due to the high quality of the SYSTRAN™ dictionary. It is possible that, if we could extract the full SYSTRAN™ dictionary, we could run the much simpler dictionary method with comparable results. Clearly, this experiment is out of our reach.

8 Acknowledgements

We would like to thank Yiming Yang, Jaime Carbonell and Ralf Brown for many useful ideas and comments. We would also like to thank them for allowing us to use their computing facilities, software and document collections.

References

Brown, R.: 1998, 'Automatically-Extracted Thesauri for Cross-Language IR: When Better is Worse', in *Proceedings of the First Workshop on Computational Terminology*, pp. 15–21.

Ballestros L., B. Croft: 1997, 'Phrasal Translation and Query Expansion Techniques for Cross-Language IR', in *SIGIR'97 Proceedings*, pp. 85–91.

Buckley C., Salton G., Allan J., Singhal A.: 1995, 'Automatic query expansion using SMART: TREC 3', in *Overview of the Third Text REtrieval Conference (TREC-3)*, pp. 69–80.

Collier N., H. Hirakawa and A. Kumano: 1998, 'Machine Translation vs. Dictionary Term Translation - a Comparison for English-Japanese News Article Alignement', COLING-ACL'98.

Davis M., T. Dunning: 1996, 'A TREC evaluation of query translation methods for multilingual text retrieval', in *The 4th Text Retrieval Conference (TREC-4).*

Deerwester S, S. T. Dumais, G. W. Furnas, T. K. Landauer, R. Harshman: 1990, 'Indexing by latent semantic analysis', in *Journal of the Am. Society of Inf. Science*, pp. 391–407.

Dumais S., Landauer T., Littman M.: 1996, 'Automatic Cross-Linguistic Information Retrieval using Latent Semantic Indexing', in *SIGIR'96 Workshop on Cross-Linguistic IR.*

Fung P. and K. McKeown: 1996, 'A technical word and term translation aid using noisy parallel copora across language groups', *Machine Translation - Special Issue on new Tools for Human Translators*, pp. 53–87.

Gachot *et al.*: 1998, 'The SYSTRAN NLP Browser: An Application of Machine Translation Technology in Cross-Language Information Retrieval', in Grefenstette (1998).

Graff D., R. Finch: 1994, *Multilingual text resources at the Linguistic Data Consortium*, in *Proceedings of the 1994 ARPA Human Lanuage Technology Workshop.*

Grefenstette, G.: 1998, *Cross-Language Information Retrieval*, Kluwer.

Hull D., Grefenstette G, 'Querying across languages: a dictionary-based approach to multilingual information retrieval', in *SIGIR'96 Proceedings.* pp. 49–57.

Landauer T., Littman M.: 1995 'A statistical method for language-independent representation of the topical content of text segements', Bellcore.

Littman M., Keim G.: 1997, 'Cross-Language Text Retrieval with Three Languages', CS-1997-16, Duke University.

Littman M., Jiang F.: 1998, 'A comparison of Two Corpus-Based Methods for Translingual Information Retrieval'.

Fellbaum Christiane (ed.): 1998, *WordNet: An Electronic Lexical Database*, The MIT Press Books.

Oh-Woog Kwon, Insu Kang, Jong-Hyeok Lee and Geunbae Lee, 'Cross-Language Text Retrieval Based on Document Translation Using Japanese-to-Korean MT System' in *NLPRS'97 Proceedings.*

Salton, G: 1989, *Automatic Text Processing: the Transformation, Analysis, and Retrieval of Information by Computer*, Addison-Wesley, Reading, Pennsylvania.

Salton G, C. Buckley: 1990, 'Improving retrieval performance by relevance feedback', *Journal of American Society for Inf. Science*, 41, pp. 288–297.

Sheridan P., J. P. Ballerini: 1996, 'Experiments in multilingual information retrieval usind the Spider system', in in *SIGIR'96 Proceedings*, pp. 58–65.

Wong S. K. M., Ziarko W., Wong P. C. N.: 1985, 'Generalized Vector Space Model in Information Retrieval', in *SIGIR'95 Proceedings*, pp. 18–25.

Yang Y., J. Carbonell, R. Brown, R. Frederking: 1998, 'Translingual Information Retrieval: Learning from Bilingual Corpora', in *AI Journal, Special issue: Best of IJCAI-97*