Complementing Dictionary-Based Query Translations with Corpus Statistics for Cross-Language IR

Sung Hyon Myaeng
Dept. of Computer Science,
Chungnam National University
220 Gung-dong, Yusong-gu,
Taejon, Korea 305-764
shmyaeng@cs.chungnam.ac.kr

Myung-Gil Jang
Dept. of Knowledge Information,
Electronics and Telecommunications Research Institute
161 Kajong-dong, Yusong-gu,
Taejon, Korea 305-350
mgjang@etri.re.kr

Abstract

For cross-language information retrieval (CLIR), often queries or documents are translated into the other language to create a mono-lingual information retrieval situation. Having surveyed recent research results on translation-based CLIR, we have convinced ourselves that an effective query translation method is an essential element for a practical CLIR system with a reasonable quality. After summarizing the arguments and methods for query translation and survey results for dictionary-based translation methods, this paper describes a relatively simple yet effective method of using mutual information to handle the ambiguity problem known to be the major factor for low performance compared to mono-lingual situation. Our experimental results based on the TREC-6 collection shows that this method can achieve up to 85% of the monolingual retrieval case and 96% of the manual disambiguation case.

1. Introduction

Cross-language information retrieval (CLIR) enables a user to retrieve documents written in diverse languages using queries expressed in his or her own language. The most common approaches to CLIR have been to translate either queries or documents to overcome the language differences, although other methods without requiring translations, such as cross-language Latent Semantic Indexing (Dumais et al., 1997) have been introduced.

For translation-based CLIR, query translation and document translation are two extreme approaches that have been developed while a hybrid method is possible (McCarley, 1999). Query translation, a more popular method between the two, is much simpler and less expensive compared to document translation. Document translation, on the other hand, has been argued to be more competitive because of rich contextual information in documents, compared to short queries whose translation may involve many errors due to unresolved ambiguities. While it is possible to apply a high-quality machine translation system for documents as in Oard & Hackett (1997), some believe that document translation is simply impractical for large-scale applications using the state-of-art technology (Carbonell et al, 1997). A more recent development of a fast translation algorithm designed for IR makes document translation a viable option (Franz et al., 1999), but experiments have shown that a hybrid method involving both query and document translation gives the best retrieval effectiveness (McCarley, 1999).

Given the assessment of the previous works in CLIR, we focus on query translation in this article because of two compelling reasons. One is that query translation is a more practical solution to CLIR. Even if we need document translation for a higher quality, query translation is still necessary for a hybrid method that has been proven to be the best so far. The other reason has to do with the difficulty of disambiguating queries, which is the main problem of query translation. We believe that by making query translation more accurate through an effective target disambiguation, we can make the simpler approach even more preferable.

To further emphasize on simplicity and practicality, our approach concentrates on dictionary-based translation rather than corpus-based or thesaurus-based ones that suffer from scarcity of resources. Machine-readable bilingual dictionaries for different language pairs are more readily available than parallel or comparable corpora. Bilingual thesauri are much less available.

The major weakness of using a general-purpose bilingual dictionary for query translation is the lack of domain specificity, which results in a high degree of translation ambiguity. It has been known that with a simple use of bilingual dictionaries in some language pairs, retrieval effectiveness can be only 40%-60% of that with monolingual retrieval (Ballesteros & Croft, 1997). It is obvious that other additional resources need to be used for better performance. Our approach is to use a target language corpus as the basis for handling polysemous words in the language. Since the target text database is always available by default, getting hold of a target language corpus is rarely an issue. While the approach taken here is language-independent, our focus has been CLIR between Korean queries and English documents.

Given the aforementioned context, we propose a relatively simple yet effective method for resolving translation ambiguities using mutual information (MI) statistics (Church and Hanks, 1990) obtained only from the target document collection. Mutual information is used not only to select the best candidates but also to assign a weight to one or more translation candidates in the target language.

2. Translation Ambiguities

Although an easy way to find translations of query terms is to use a bilingual dictionary, this method alone suffers from problems caused by translation ambiguities since there are often one-to-many correspondences in a bilingual dictionary. For example, in a Korean query consisting of three words, "자동차 공기 오염" *(ja-dong-cha gong-gi oh-yum)* that means *air pollution caused by automobiles,* each word can be translated into multiple English words when a Korean-English dictionary is used in a straightforward way. The first word "자동차" *(ja-dong-cha)* of the query can be translated into English words with semantically similar but different words like "motorcar", "automobile", and "car". The second word "공기" *(gong-gi),* a homonymous word, can be translated into English words with different meanings: "air", "atmosphere", "empty vessel", and "bowl". And the last word "오염" *(oh-yum)* can be translated into two English words, "pollution" and "contamination".

Retaining multiple candidate words can be useful in promoting recall in monolingual IR system, but previous research indicates that failure to disambiguate the meanings of the words can hurt retrieval effectiveness tremendously. For instance, it is obvious that a phrase like *empty-vessel* would change the meaning of the query entirely. Even a word like *contamination,* a synonym of *pollution,* may end up retrieving unrelated documents due to the slight differences in meaning.

Table 1. The Degree of Ambiguities

	Words			Word Pairs		
	# in Korean	# in English	Average Ambiguity	# in Korean	# in English	Average Ambiguity
Title	48	158	3.29	24	212	8.83
Short	112	447	3.99	91	1459	16.03
Long	462	1835	3.97	423	6196	14.65

As in Jang et al. (1999), Table 1 shows the extent to which ambiguity occurs in our query translation when an English-Korean dictionary is used blindly after the morphological analysis and tagging. The three rows, title, short, and long, indicate three different ways of composing queries from the topic statements in the TREC collection. The left half shows the average number of English words per Korean word for each query, whereas the right half shows the average number of word pairs in English that can be formed from a single word pair in Korean. The latter indicates that the disambiguation process will have to select one out of more than 9 possible pairs on the average, regardless of which part of the topic statements is used for formal query generation.

Slightly different statistical analysis was conducted in a recent work for the Chinese-English pair (Chen et al., 1999). Based on a Chinese thesaurus and Roget's International Thesaurus for English, a Chinese word and an English word have 1.397 and 1.687 senses, respectively. When the top 1000 high frequency words were only considered, different values were obtained: 1.504 for Chinese and 3.527 for English. It should be noted that these numbers are for the monolingual cases (i.e. with a monolingual dictionary or thesaurus) while the Korean-English case reported in Jang et al. (1999) reveals the degree of ambiguity in a bilingual dictionary.

3. Dictionary-based Query Translation

The basic idea in dictionary-based query translation is to replace each term in the query with an appropriate term or set of terms in the target language by a dictionary lookup. Two factors limit the performance of this approach (Ballesteros & Croft, 1997). The first is that many words do not have a unique translation, and sometimes the alternate translations have very different meanings. This translation ambiguity raises a number of problems. The second is that a dictionary may lack some terms that are essential for a correct interpretation of the query. Examples of this case are 1) failure to translate multi-term concepts as phrases or to translate them poorly, 2) failure to translate unknown words in a dictionary entry such as technical terminology, proper names, foreign words and so on. Performance can be improved up to 75% by applying simple language processing techniques such as part-of-speech tagging and phrase indexing (Davis & Ogden 1997, Hull & Grefenstette 1996).

Some researchers have tried to utilize other resources in addition to a dictionary in order to deal with the problems of extraneous translations. The Double MAXimize(DMAX) method proposed by Yamabana et al. (1996) is a statistical word selection method that attempts to solve the disambiguation problem caused by the bilingual lexicon. It selects the word pairs that maximize the co-occurrence frequency among source words and those among target words. Another strategy implemented by Twenty-One system (Kraaij & Hiemstra 1997) tries to use some standard NLP tools as well as bilingual dictionaries from Dutch to other languages such as German, French, English, and Spanish. Translated noun phrases are disambiguated by using noun phrases extracted from the target document corpus. The method proposed by Hull (1997) uses a weighted Boolean model. It calculates term weights and generates a Boolean combination of the terms. Chen et al. (1999) used two monolingual balanced corpora to learn word co-occurrence for word sense ambiguity problems in source and target languages.

4 Using Mutual Information for Target Disambiguation

Since the goal of disambiguation is to select the best pair of a source and a target word (or a phrase) among many alternatives, the mutual information statistic is a natural choice in judging the degree to which two translated words or phrases co-occur within a certain text boundary. For a given pair of source terms, it would be reasonable to choose the pair of translations that are most strongly associated with each other, thereby eliminating those translations that are not likely to be correct ones.

The mutual information *MI(x,y)* is defined as the following formula (Church and Hanks, 1990).

$$MI(x,y) = \log_2 \frac{p(x,y)}{p(x)p(y)} = \log_2 \frac{N f_w(x,y)}{f(x)f(y)} \qquad (1)$$

Here x and y are words occurring within a window of w words. It is calculated based on word co-occurrence statistics and used as a measure to calculate correlation between words. The probabilities $p(x)$ and $p(y)$ are estimated by counting the number of observations of x and y in a corpus, $f(x)$ and $f(y)$, and normalizing each by N, the size of the corpus. Joint probabilities, $p(x,y)$, are estimated by counting the number of times, $f_w(x,y)$, that x is followed by y in a window of w words and normalizing it by N. In our application of query translation, the joint co-occurrence frequency $f_w(x,y)$ has 6-word window size which seems to allow semantic relations of query as well as fixed expressions (idioms such as *bread* and *butter).* We ensure that the word x be followed by the word y within the same sentence only.

In our query translation scheme, *MI* values are used to select most likely translations after each Korean query word is translated into one or more English words. Our use of *MI* values is based on the assumption that when two words co-occur in the same query, they are likely to co-occur in the same affinity in documents. Conversely, two words that do not co-occur in the same affinity are not likely to show up in the same query. In a sense, we are conjecturing mutual information can reveal some degree of semantic association between words. These *MI* values were extracted from the English text corpus consisting of 1988 - 1990 AP news, which contains 116,759,540 words.

Our Korean-to-English query translation scheme works in four stages: keyword selection, dictionary-based query translation, bilingual word sense disambiguation, and query term weighting. Although none of the common resources such as dictionaries, thesauri, and corpora alone is complete enough to produce high quality English queries, we decided to use a bilingual dictionary at the second stage and a target-language corpus for the third and the fourth stages. Our strategy was to try not to depend on scarce resources to make the approach practical. Figure 1 shows the four stages of Korean-to-English query translation.

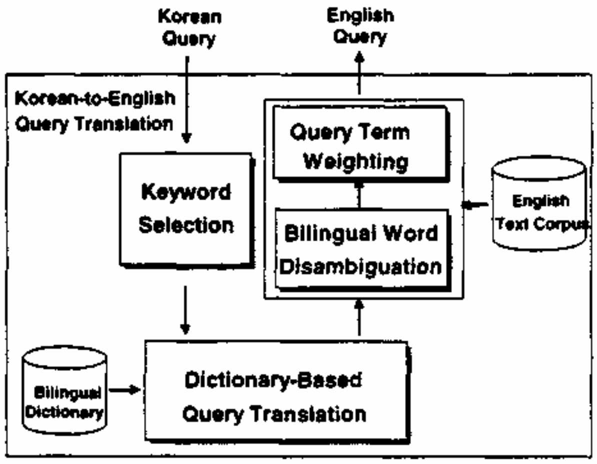

Fig. 1. Four Stages for Korean-to-English Query Translation.

At the keyword selection stage, Korean keywords to be fed into the query translation process are extracted from a quasi-natural language query. This keyword selection is done with a morphological analyzer and a stochastic part-of-speech (POS) tagger for the Korean language (Shin *et al.,* 1996). The role of the tagger is to help select a single morpheme sequence from the multiple candidate sequences generated by the morphological analysis. This process of employing a morphological analysis and a tagger is crucial for selecting legitimate query words from the topic statements because Korean is an agglutinative language. Without the tagger, all the extraneous candidate keywords generated from the morphological analyzer will have to be entered into the translation process, which in and of itself will generate extraneous words, due to one-to-many mapping in the bilingual dictionary.

The second stage does the actual query translation based on a dictionary look-up, by applying both word-by-word translation and phrase-level translation. At the moment, phrases in the bilingual dictionary can only be identified and translated; no statistical phrases as in Smadja (1993) are considered. Since the bilingual dictionary lacks some words that are essential for a correct interpretation of the Korean query, it is important to identify unknown words such as foreign words and transliterate them into English strings that need to be matched against an English dictionary (Jeong *et al.,* 1997).

At the word disambiguation stage, we filter out the extraneous translations generated blindly from the dictionary lookup process. Our method makes use of the co-occurrence information extracted from a target document collection that should be always available. More specifically, the mutual information statistics between pairs of words were used to determine whether English words translated from two Korean query terms "compatible". In a sense, we make use of mutual disambiguation effect among query terms. More details are described in Section 5.

The disambiguation stage is finally refined with the query term weighting by which we assign different weights to translated terms. Instead of selecting a single translation, the system allows for multiple translations that participate in the final query. This strategy is based on our observation that the sense disambiguation does not always select the best candidate and that multiple candidates are indeed correct for IR purposes. Besides, allowing multiple translations with some caution has the effect of query expansion. It should be noted that the term weighting strategy has nothing to do with the usual process of assigning weights to query terms based on their relative importance by using IDF values, for example. Instead, it reflects to what extent each translation is to be trusted.

5. The Disambiguation Algorithm

Given sets of translations, each of which corresponds to a query term in the source language, the task is to select one or more of them that best reflect the concept conveyed in the original query. The algorithm first arranges the sets in the sequence that matches the order in which the query terms appear in the source language. It then compares the mutual information values for all pairs of translations between two adjacent sets. It only considers adjacent pairs because the order in which query terms appear in the source language reflects the word order in the sentences in the TREC topic statements. With additional computational cost, all permutations of the terms can be used as in other work (Fung et al, 1999; Chen et al, 1999). Although we use the mutual information statistic to measure the association, others such as those used by Ballesteros & Croft (1998) can be used as well.

Figure 2 shows the *MI* values calculated for the word pairs comprising the translations of the original query. The words under w_1, w_2, and w_3 are the translations from the three query words, respectively. The lines indicate that mutual information values are available for the pairs, and the numbers show some of the significant *MI* values for the corresponding pairs.

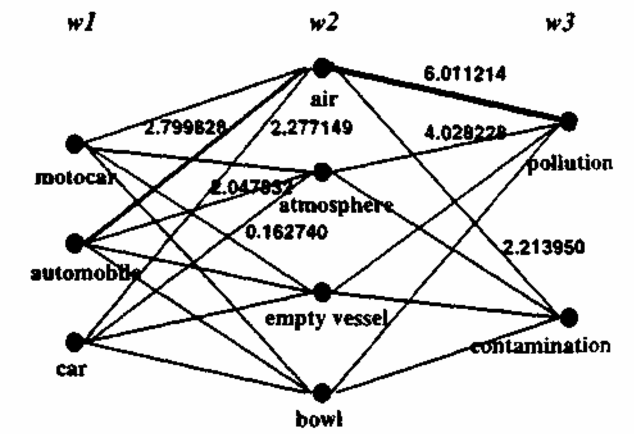

Fig. 2. An Example of Word Pairs with *MI* Values

Our algorithm relies on both relative and absolute magnitudes of the *MI* vales. It first looks for the pair with the highest *MI* value among all the pairs for the query so that we can start with the most reliable translation pair (the <*air, pollution*> pair in Fig.2) The next best pair is chosen among all the pairs as long as it does not conflict with the already selected translations (two at this stage). A conflict occurs when one of the translations of the new pair is a translation that was not chosen for the particular word in the first round. In Fig. 2, the pair <*automobile, atmosphere*> would be such a conflict if it has the second highest *MI* values.

This process of selecting the next best pair continues until at least one translation is chosen for each query word. After two pairs have been selected :':" <$w_{i,}$ w_{i+1}> and <w_j, w_{j+1}>, where w_{i+2} is equal to w_j, the MI values for translation pairs corresponding to w_{i+1} and w_j are ignored because the translations for the two words have been determined already. This process makes sense intuitively because adjacent word pairs in the source query are not always expected to have a strong association. In such a case, our decision can be made with other associations preceding the first translations for the first word and following the second translations for the second word.

It should be noted that no translation for a given source term is chosen in this process when the *MI* values associated with the pairs involving the translations for the term are all below a threshold. That is, when the translations are not strongly associated with preceding or following ones, we assume that we have no ground for selecting a particular translation. In order to handle this situation and to consider multiple translation candidates, we apply the weighting scheme. To be more precise, there are four reasons for employing the weighting scheme. They are:

- Our translation selection method is not guaranteed to give the correct translation. The method would give a reasonable result only when two consecutive query terms are actually used together in many documents, which is a hypothesis yet to be confirmed for its validity.

- There may *be* more than one strong association whose degrees are different from the rest by a large magnitude. It is hard to justify the process of selecting the best one when the second best candidate is almost as good as the best.

- Seemingly extraneous terms may serve as a recall-enhancing device with a query expansion effect. Taking all those above a threshold would be a reasonably safe way of expanding target queries.

- We need to deal with the case where no translation can be selected based on the algorithm.

The basic idea in our weighting scheme is to give a large weight to the best candidate and divide the remaining quantity so that the result is equally assigned to the rest of the candidates. In other words, the weight for the best candidate, W_b, is 1 if the largest *MI* value involving the candidate is greater than a threshold value. Otherwise it is expressed as follows.

$$W_b = \frac{f(x)}{\theta + 1} \times 0.5 + 0.5 \qquad (2)$$

Here x and θ are a *MI* value and a threshold, respectively. The numerator, $f(x)$, is currently the smallest integer greater than the *MI* value so that the resulting weight is the same for all the candidates whose *MI* values are within a certain interval. Once the value for W_b is calculated, the weight for the rest of the candidates are calculated as follows:

$$W_r = \frac{1 - W_b}{n - 1} \qquad (3)$$

where n is the number of candidates. It should be noted that $W_b + \Sigma W_r = 1$.

Based on our observation of the calculated *MI* values, we chose to use 3.0 as the cut-off in choosing the best candidate and assigned a fairly high weight. The cut-off value was determined purely based on the data we obtained; it can vary based on the new range of *MI* values when different corpora are used.

In the example of Fig. 2, the translation pair candidate between w_1 and w_2 are *<motorcar, air>*, *<automobile, air>*, and *<car, air>*. Here because the weight of the translation pairs *<automobile, air>* is W_b = 0.83, the translation "automobile" has a relatively higher term weight than the other two "motorcar" and "car". Finally the optimal English query set with their term weight

<(motocar, 0.0625),(automobile, 0.875),(car, 0.0625) >

is generated for the translations of w_1.

6. Experiments

In order to test the efficacy of the algorithm, we ran some basic experiments using the collection from the Cross-Language Track of TREC 6. The 24 English queries are comprised of three fields: titles, descriptions, and narratives. These English queries were manually translated into Korean queries to mimic a CLIR situation with Korean queries and English document. We used the Smart 11.0 system developed by Cornell University for both mono-lingual IR and CLIR using our query translation scheme.

Our goal was to examine how far we can go with the relatively simple but practical disambiguation and weighting schemes for query translation. We ran our system with three sets of queries, differentiated by the query lengths: 'title' queries with title fields only, 'short' queries with description fields only, and 'long' queries with all the three fields. The retrieval effectiveness measured with 11-point average precision was used for comparison against the baseline of monolingual retrieval using the original English query.

Table 2 gives the experimental results from using the four types of query set. The result from "Translated Query I" was generated only with the keyword selection and dictionary-based query translation stages without disambiguation. The result "Translated Query II" was generated after all the stages of our disambiguation and weighting were applied. And the result from the "Manually Disambiguated Query" set was generated by manually selecting the best candidate translations from the Translated Query I.

Table 2. Experimental Results

Query Sets	Title		Short		Long	
	11pt. P	C/M(%)	11pt. P	C/M(%)	11pt. P	C/M(%)
Original Query	0.3251	-	0.3189	-	0.2821	-
Tran. Query I	0.2290	70.44	0.21443	67.20	0.1587	56.26
Tran Query II	0.2675	82.28	0.2698	84.60	0.2232	79.12
M.Disam. Query	0.2779	85.48	0.3002	94 14	0.2433	86.25

The performance of the Translated query set I was about 70%, 67%, and 56% of monolingual retrieval for the three cases, respectively. The performance of the translated query set II was about 82%, 85%, and 79% of monolingual retrieval for the three cases, respectively. The performance of the manually disambiguated queries, 85%, 94%, and 86% of monolingual retrieval for the three cases, respectively, can be treated as the upper limit for the cross-language retrieval. The reason why they are not 100% is attributed to the several factors. They are: 1) the inaccuracy of the manual translation of the original English query into the Korean queries, 2) the inaccuracy of the Korean morphological analyzer and the tagger in generating query words, and 3) the inaccuracy in generating candidate terms using the bilingual dictionary. Compared to the past results where query translation with a simple dictionary-based method can reach slightly more than 50% of the corresponding mono-lingual case, we consider the performance of our method is very promising.

The difference between Translated Query I and Translated Query II indicates that the M1-based disambiguation and the weighting schemes are quite effective in enhancing the retrieval effectiveness. In addition, the results show that the use of the query translation scheme is more effective with longer queries than with shorter queries. This is expected because the longer the queries are, the more contextual information can be used for mutual disambiguation.

7. Conclusion

We have surveyed various CLIR methods based on translations of queries, documents, or both. While there are arguments for document translations or hybrid method where both query and document translations are used, we focus our attention to query translation methods for which we have described some survey results and our own method.

It has been reported that query translation using a simple bilingual dictionary leads to a more than 40% drop in retrieval effectiveness due to ambiguity problems. Our query translation method uses mutual information extracted from the 1988 - 1990 AP corpus in order to alleviate the problems through a translation selection and weighting method. The experiments using test collection of TREC-6 Cross-Language Track show that the method improves retrieval effectiveness in Korean-to-English cross-language IR. The performance can be up to 85% of the monolingual retrieval case. We also found that we obtained the largest percent increase with long queries as expected.

While the experimental results are very promising, there are several issues to be explored. First, we need to test how effectively the method can be applied. Second, we intend to experiment with other co-occurrence metrics, instead of the mutual information statistic, for possible improvement. This investigation is motivated by our observation of some counter-intuitive *MI* values. Third, we also plan on using different algorithms for choosing the terms and calculating the weights.

In addition, we plan to use the pseudo relevance feedback method that has been proven to be effective in monolingual retrieval. Terms in some top-ranked documents are thrown into the original query with an assumption that at least some, if not all, of the documents are relevant to the original query and that the terms appearing in the documents are useful in representing user's information need. Here we need to determine a threshold value for the number of top ranked document for our cross-language retrieval situation, let alone other phenomenon.

References

Lisa Ballesteros and W. Bruce Croft (1997). "Phrasal Translation and Query Expansion Techniques for Cross-lingual Information Retrieval". In *Proceedings of the 20th Annual ACM S1GIR Conference on Information Retrieval,* Philadelphia, PA.

Lisa Ballesteros and W. Bruce Croft (1998). "Resolving Ambiguity for Cross-language Retrieval". In *Proceedings of the 21st Annual ACM SIGIR Conference on Information Retrieval.*

J.G. Carbonell, Y. Yang, R.E. Frederking, R.E Brown, Yibing Geng and Danny Lee (1998). "Translingual Information Retrieval: A Comparative Evaluation". In *Proceedings of the Fifteenth International Joint Conference on Artificial Intelligence.*

Hsin-His Chen, Guo-Wei Bian and Wen-Cheng Lin (1999). "Resolving Translation Ambiguity and Target Polysemy in Cross-Language Information Retrieval In *Proceedings of the 37th Annual Meeting of the Association for Computational Linguistics,* College Park, MD.

Kenneth W. Church and Patrick Hanks (1990). "Word Association Norms, Mutual Information, and Lexicography". *Computational Linguistics,* Vol. 16, No. 1, pp. 22-29.

Kenneth W. Church and Patrick Hanks (1990). "Word Association Norms, Mutual Information, and Lexicography". *Computational Linguistics,* Vol. 16, No. 1. pp. 22-29.

Mark Davis and W.C. Ogden (1997). "Implementing Cross-Language Text Retrieval Systems for Large-Scale Text Collections and the World Wide Web" In *Proceedings of AAAI Symposium on Cross-Language Text and Speech Retrieval.*

Pascale Fung, Xiaohu Liu and Chi Shun Cheung (1999). "Mixed Language Query Disambiguation" In *Proceedings of the 37th Annual Meeting of the Association for Computational Linguistics,* College Park, MD.

Gregory Grefenstette (1998), *Cross-Language Information Retrieval,* Kluwer Academic Publishers.

David Hull (1996). "A Weighted Boolean Model for Cross-Language Text Retrieval". In *Proceedings of the 19th Annual ACM SIGIR Conference on Informtion Retrieval,* Zurich, Switzerland.

Myung-Gil Jang, Sung Hyon Myaeng and Se Young Park (1999). "Using Mutual Information to Resolve

Query Translation Ambiguities and Query Term Weighting". In *Proceedings of the 37th Annual Meeting of the Association for Computational Linguistics,* College Park, MD.

K. S. Jeong, Y. H. Kown, and S. H. Myaeng, (1997). "Construction of Equivalence Classes through Automatic Extraction and Identification of Foreign Words". In *Proceedings of NLPRS'97,* Phuket, Thailand.

Wessel Kraaij and Djoerd Hiemstra. (1997). "Cross Language Retrieval with the Twenty-One System". In *Proceedings of the Sixth Text Retrieval Conference (TREC-6),* NIST.

J Scott McCarley (1999). "Should we Translate the Documents or the Queries in Cross-Language Information Retrieval?" In *Proceedings of the 37th Annual Meeting of the Association for Computational Linguistics,* College Park, MD.

Douglas W. Oard and Paul Hackett (1997). "Document Translation for the Cross-Language Text Retrieval at the University of Maryland". In *Proceedings of the Sixth Text Retrieval Conference (TREC-6),* NIST.

Frank Smadja (1993). "Retrieval Collection from Text: Xtract". *Computational Linguistics,* Vol. 19, No. 1. pp.143-177.

Joong-Ho Shin, Young-Soek Han, Key-Sun Choi (1996). "A HMM Part of Speech Tagger for Korean with Word Phrasal Relations". In *Proceedings of Recent Advances in Natural Language Processing.*

Kiyoshi Yamabana, Kazunori Muraki, Shinichi Doi and Shin-ichiro Kamei (1996). "A Language Conversion Front-End for Cross-Language Information Retrieval". In *Proceedings of the 19th Annual ACM S1GIR Conference on Information Retrieval,* Zurich, Switzerland.