Using Mutual Information to Resolve Query Translation Ambiguities and Query Term Weighting

[1] Myung-Gil Jang, [2] Sung Hyon Myaeng and [1] Se Young Park

[1] Dept. of Knowledge Information, Electronics and Telecommunications Research Institute
161 Kajong-Dong, Yusong-Gu,
Taejon, Korea 305-350
{mgjang, sypark}@etri.re.kr

[2] Dept. of Computer Science,
Chungnam National University
220 Gung-Dong, Yusong-Gu,
Taejon, Korea 305-764
shmyaeng@cs.chungnam.ac.kr

Abstract

An easy way of translating queries in one language to the other for cross-language information retrieval (IR) is to use a simple bilingual dictionary. Because of the general-purpose nature of such dictionaries, however, this simple method yields a severe translation ambiguity problem. This paper describes the degree to which this problem arises in Korean-English cross-language IR and suggests a relatively simple yet effective method for disambiguation using mutual information statistics obtained only from the target document collection. In this method, mutual information is used not only to select the best candidate but also to assign a weight to query terms in the target language. 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.

Introduction

Cross-language information retrieval (IR) enables a user to retrieve documents written in diverse languages using queries expressed in his or her own language. For cross-language IR, either queries or documents are translated to overcome the language differences. Although it is possible to apply a high-quality machine translation system for documents as in Oard & Hackett (1997), query translation has emerged as a more popular method because it is much simpler and more economical compared to document translation. Query translation can be done in one or more of the three approaches: a dictionary-based approach, a thesaurus-based approach, or a corpus-based approach.

There are three problems that a cross-language IR system using a query translation method must solve (Grefenstette, 1998). The first problem is to figure out how a term expressed in one language might be written in another. The second problem is to determine which of the possible translations should be retained. The third problem is to determine how to properly weight the importance of translation alternatives when more than one is retained.

For cross-language IR between Korean and English, i.e. between Korean queries and English documents, an easy way to handle query translation is to use a Korean-English machine-readable dictionary (MRD) because such bilingual MRDs are more widely available than other resources such as parallel corpora. However, it has been known that with a simple use of bilingual dictionaries in other 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.

This paper focuses on the last two problems: pruning translations and calculating the weights for translation alternatives. We first describe the overall query translation process and the extent to which the ambiguity problem arises in Korean-English cross-language IR. We then propose a relatively simple yet effective method for resolving translation disambiguation using mutual information (MI) (Church and Hanks, 1990) statistics obtained only from the target document collection. In this method, mutual

information is used not only to select the best candidate but also to assign a weight to query terms in the target language.

1 Overall Query Translation Process

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.

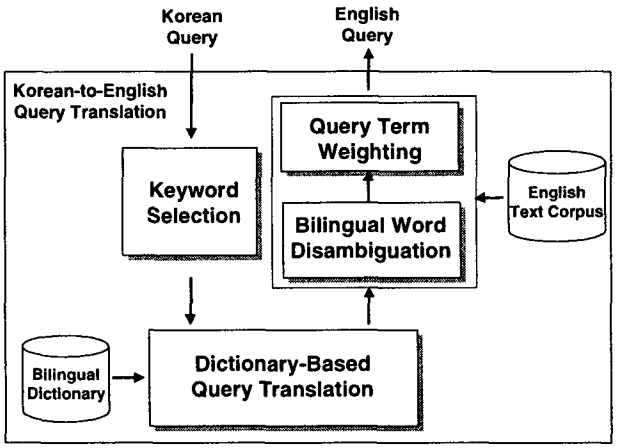

Fig. 1. Four Stages for Korean-to-English Query Translation.

1.1 Keyword Selection

At the first 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 the exact 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.

1.2 Dictionary-Based Query Translation

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. For the correct identification of phrases in a Korean query, it would help to identify the lexical relations and produce statistical information on pairs of words in a text corpus as in Smadja (1993). 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).

1.3 Selection of the Correct Translations

At the word disambiguation stage, we filter out the extraneous words generated blindly from the dictionary lookup process. In addition to the POS tagger, we employed a bilingual word disambiguation technique using the co-occurrence information extracted from the collection of target documents. More specifically, The mutual information statistics between pairs of words were used to determine whether English words from different sets generated by the translation process are "compatible". In a sense, we make use of mutual disambiguation effect among query terms. More details are described in Section 3.

1.4 Query Term Weighting

Finally, we apply our query term weighting technique to produce the final target query. The term weighting scheme basically reflects the degree of associations between the translated terms, and we give a high or low term weighting value according to the degree of mutual association between query terms. This is another area where we make use of mutual information obtained from a text corpus. The result from the four stages is a set of query terms to be used in a

vector-space retrieval model.

2 Analysis of Translation Ambiguity

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 ambiguity 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 S. Lan.	# in T. Lang.	Average Ambiguity	# in S. Lan.	# in T. Lang.	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

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.

3 Query Translation and Mutual Information

Our strategy for cross-language IR aims at practicality in that we try not to depend on scarce resources. Along the same line of reasoning, we opted for a disambiguation approach that requires only a collection of documents in the target language, which is always available in any cross-language IR environment. Since the goal of disambiguation is to select the best pair among many alternatives as described above, the mutual information statistic is a natural choice in judging the degree to which two words co-occur within a certain text boundary. It would be reasonable to choose the pair of words that are most strongly associated with each other, thereby eliminating those translations that are not likely to be correct ones.

Mutual information values are calculated based on word co-occurrence statistics and used as a measure to calculate correlation between words. The mutual information $MI(x,y)$ is defined as the following formula (Church and Hanks, 1990).

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

Here x and y are words occurring within a window of w 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.

Table 2 gives some examples of *MI* values for the alternative word pairs for translated queries of TREC-6 Cross-Language IR Track. These *MI* values were extracted from the English text corpus consisting of 1988 ~ 1990 AP news, which contains 116,759,540 words.

Table 2. Example of $MI(x,y)$ Values

Word x	Word y	f(x)	f(y)	f(x,y)	MI(x,y)
respiratory	ailment	716	1134	74	9.272506
teddy	bear	679	7932	262	8.644690
fossil	fuel	676	13176	333	8.381424
air	pollution	52216	4878	890	6.011214
research	development	24278	24213	1317	5.566768
AIDS	spread	18575	10199	212	4.872597
ivory	trade	1885	86608	84	4.095613
environment	protection	7771	13139	36	3.717652
bear	doll	7932	1394	3	3.455646
region	country	21093	103833	358	2.948925
point	interest	30419	51917	107	2.068232
law	terrorism	70182	4762	20	1.944089
treatment	result	13432	38055	22	1.614487
terrorism	government	4762	193977	29	1.299005
opinion	news	9124	82220	21	1.184332
food	life	32222	40625	30	0.984281
copy	price	6803	90594	10	0.638950
labor	information	26571	30245	11	0.468861

When $MI(x,y)$ is large, the word associations are strong and produce credible results for disambiguation of translations. However, if $MI(x,y) < 0$, we can predict that the word x and word y are in complementary distribution.

4 Disambiguation and Weight Calculation

We can alleviate the translation ambiguity by discriminating against those word pairs with low MI values. The word pair with the highest MI value is considered to be the correct one among all the candidates in the two sets. Since a query is likely to be targeted at a single concept, regardless of how broad or narrow it is, we conjecture that words describing the concept are likely to have a high degree of association. Although we use the mutual information statistic to measure the association, others such as those used by Ballesteros & Croft (1998) can be considered.

In the example of Section 2, each Korean word has multiple English words due to translation ambiguity. Figure 2 shows the *MI* values calculated for the word pairs comprising the translations of the original query. The words under *w1*, *w2*, and *w3* 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 among all the possible pairs.

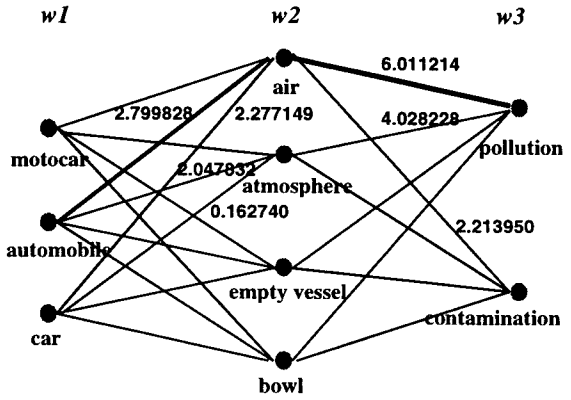

Fig. 2. An Example of Word Pairs with *MI* Values

Our bilingual word disambiguation and weighting schemes rely on both relative and absolute magnitudes of the *MI* vales. The algorithm first looks for the pair with the highest *MI* value and selects the best candidates before and after the pair by comparing the MI values for the pairs that are connected with the initially chosen pairs. This process is applied to the words immediately before or after the chosen pair in order to limit the effect of the choice that may be incorrect.

It should be noted that the words not chosen in this process are not used in the translated query unless the *MI* values are greater than a threshold. As described below, we assume that the candidates not in the first tier may still be useful if they are strongly associated with the adjacent word selected.

For example, the word pair <air, pollution> that has the bold line representing the strongest association in the column is choisen first. Then the three *MI* values for the pairs containing air are compared to select the <automobile, air> pair, resulting in <automobile, air, pollution>. If there were additional columns in the example, the same process would be applied to the rest of the network.

There are three reasons why query term weighting is of some value in addition to the pruning of conceptually unrelated terms. First, our word 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. Second, there may be more than one strong association whose degrees are different from each other by a large magnitude. Third, seemingly extraneous terms may serve as a recall-enhancing device with a query expansion effect.

The basic idea in our term weighting scheme is to give a large weight to the best candidate and divide the remaining quantity to assign equal weights to the rest of the candidates. In other words, the weight for the best candidate, W_b, is either 1 if it is greater than a threshold value or 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)*, gives 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 value in choosing the best candidate and assign 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 word pair candidate between *w1* and *w2* are (motorcar, air), (automobile, air), and (car, air). Here because the weight of the word pairs (automobile, air) is W_b = 0.83, the word "automobile" has a relatively higher term weight than the other two words "motorcar" and "car". Finally the optimal English query set with their term weight, <(motocar,0.085), (automobile, 0.83), (car, 0.085) >, is generated for the translations of *w1*.

5 Experiments

We developed a system for our cross-language IR techniques and conducted 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 so that we can pretend as if the Korean queries had been generated by human users for cross-language IR. In order to compare cross-language IR and mono-language IR, we used the Smart 11.0 system developed by Cornell University.

Our goal was to examine the efficacy of the disambiguation and term weighting schemes in our 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 3 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. The result "Translated Query II" was generated after all the stages of our word disambiguation and query term weighting were done. And the result from the manually disambiguated query set was generated by manually selecting the best candidate terms from the Translated Query I.

Table 3. 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 performances of the translated query set II were about 82%, 85%, and 79% of monolingual retrieval for the three cases, respectively. The performance of the 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.

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

Conclusion

It has been known that query translation using a simple bilingual dictionary leads to a more than 40% drop in retrieval effectiveness due to translation ambiguity. Our query translation method uses mutual information extracted from the 1988 ~ 1990 AP corpus in order to solve the problems of the bilingual word disambiguation and query term weighting. 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.

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

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

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

Lisa Ballesteros and W. Bruce Croft(1997). Phrasal Translation and Query Expansion Techniques for Cross-lingual Information Retrieval, SIGIR'97.

Lisa Ballesteros and W. Bruce Croft(1998). Resolving Ambiguity for Cross-language Retrieval, SIGIR' 98.

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

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.*

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

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