Information Retrieval, 2, 69–80 (2000)

Using Statistical Term Similarity for Sense Disambiguation in Cross-Language Information Retrieval

MIRNA ADRIANI mirna@dcs.gla.ac.uk
Department of Computing Science, University of Glasgow, Glasgow G12 8QQ, Scotland

Received September 10, 1998; Revised May 24, 1999; Accepted September 16, 1999

Abstract. With the increasing availability of machine-readable bilingual dictionaries, dictionary-based automatic query translation has become a viable approach to Cross-Language Information Retrieval (CLIR). In this approach, resolving term ambiguity is a crucial step. We propose a sense disambiguation technique based on a term-similarity measure for selecting the right translation sense of a query term. In addition, we apply a query expansion technique which is also based on the term similarity measure to improve the effectiveness of the translation queries. The results of our Indonesian to English and English to Indonesian CLIR experiments demonstrate the effectiveness of the sense disambiguation technique. As for the query expansion technique, it is shown to be effective as long as the term ambiguity in the queries has been resolved. In the effort to solve the term ambiguity problem, we discovered that differences in the pattern of word-formation between the two languages render query translations from one language to the other difficult.

Keywords: cross-language information retrieval, term disambiguation

1. Introduction

Driven by the increasing availability of electronic documents written in various languages from all over the world, Cross Language Information Retrieval (CLIR) has gained popularity among Information Retrieval (IR) researchers in recent years. The traditional keyword-based monolingual IR systems allow users to search for information resources written in the same language as the query. A search for documents in a number of languages using these systems can only be done by submitting a separate query in each of the languages. CLIR is designed to make such a task much easier by allowing users to retrieve documents in a language or languages different from that of the query. This language barrier problem has brought interesting challenges into the traditional monolingual IR.

One of the most obvious approaches to CLIR is to translate queries into the language of the documents or vice-versa. A number of methods for translating queries or documents have been proposed and studied, namely, through the use of natural language processing (NLP)—based machine translation systems, the use of parallel corpora, and the use of bilingual dictionaries. The first two methods are very labor intensive. Manual hand-coding of linguistic, semantic and pragmatic knowledge is required for an NLP-based machine translation system to produce good translations. This can be quite overwhelming when the domain of coverage is wide. Likewise, a great deal of manual work is also required for building parallel collections, i.e., to translate each of the documents in the collection into its equivalent in another language. The third method, dictionary based translation, has recently become very practical with the increasing availability of machine-readable bilingual dictionaries. In addition, this method offers topic coverage that is less limited than that of parallel corpus as a dictionary typically contains a wider variety of terms than a sample corpus. However, the effectiveness of this method very much depends on the dictionary's comprehensiveness, both in terms of the entry word and the sense (meaning) word coverage. Furthermore, given an ideal dictionary, the effectiveness of this method still depends on its ability to select the right sense of a word from many possible senses provided by the dictionary. This last issue is the main focus of our research presented in this paper. We will not address the quality of the dictionary issue.

In this research, we are studying CLIR for Indonesian-English document retrieval. We are familiar with both languages and have access to a large collection of texts in both languages. The Indonesian language, also known as *Bahasa Indonesia*, is the national language of Indonesia that is spoken by over 210 million of its population. The language has the same root as, and hence shared many aspects with, *Malay* language. Similar to English, it uses the Latin alphabet. It is much simpler than English, in that it does not, among other things, distinguish tenses. However, it uses more word-forms by making use of prefixes, suffixes, multiple-suffixes and combinations of prefixes and suffixes. This renders a stemmer for Indonesian words more complicated than that for English. Since human languages are natural phenomena, we presume that findings in Indonesian-English CLIR would also apply, to a large extent, to CLIR in general.

In this work, we propose a term-sense disambiguation technique for dictionary-based query translation and a query expansion technique which are based on a statistical term-similarity measure. Query translation is chosen as opposed to document translation as it requires much less computational resources. The main objective of this study is to measure the effectiveness of our techniques, in terms of average retrieval precision, and to investigate their effectiveness in alleviating the major problems in CLIR, particularly, the term ambiguity problem.

In Section 2, we provide a brief survey on major work in CLIR and other related work. Section 3 describes our sense disambiguation and query expansion techniques. Section 4 discusses the experiments conducted to evaluate the techniques' effectiveness and the results. Finally, in Section 5 we present a summary and conclusion.

2. Related work

Interest in CLIR dates back in the 70's when Salton (1970) conducted his experiments using German and English texts. He demonstrated that queries in English can be easily translated into German using a manually built thesaurus to retrieve documents in German, and vice versa. Work in CLIR has grown wider since then as can be seen by the number of international conference tracks dealing with documents in a wider variety of languages, such as Spanish, German, French, etc. (Harman 1997).

Traditionally, CLIR techniques are grouped into two categories: dictionary based approach and corpus-statistically based approach. The corpus statistics approach makes use of statistical information from the corpus (document collection), particularly term distribution statistics, to map a query—or its representation—from one language to another language. Most corpus-based work uses a *parallel corpus* in two or more languages. A parallel corpus is a collection of the same or equivalent set of documents written in two or more languages. Another type of corpus that is close to parallel corpus, called *comparable corpus*, can be built from sets of documents written in different languages where each set contains documents on the same topic. Sheridan and Ballerini (1996) demonstrated the effectiveness of employing a term similarity thesaurus which was built by comparing terms in each pair of comparable documents written in Italian and German in CLIR. Their German queries were substituted with Italian queries constructed from the terms found in the thesaurus.

Query translation using bilingual dictionaries has been much studied by researchers in the field. Basically, there are two approaches to dictionary-based CLIR, namely, (i) by translating all documents in the collection into the language of the queries (Oard and Hackett 1997), and (ii) by translating the queries into the language of the documents (Adriani and Croft 1997, Ballesteros and Croft 1997, Davis and Dunning 1995, Hull and Grefenstette 1996, Kraaij and Hiemstra 1997). Translating documents is obviously very expensive since there are typically so many documents in a collection and they are quite long. For this reason, many researchers take the second approach.

Since a bilingual dictionary normally contains more than one sense (meaning) for each term, there are chances that wrong senses being selected in the translation process. As a result, the translation queries perform worse than the equivalent manually translated queries. This is called the ambiguity problem in CLIR. Another problem associated with the dictionary-based approach is the problem with translating phrases or multi-word terms in the query. Basically, the problem occurs when words belonging to a phrase are translated independently word-for-word resulting in words with completely incompatible meanings.

A technique that can be used to alleviate the impact of the above problems is to identify phrases in the query and translate them using a phrase dictionary. Such a technique has been shown to improve the performance of CLIR. Hull and Grefenstette (1996) showed that the performance achieved from manually translating phrases in queries is significantly better than that of word-for-word translation using a dictionary. Davis and Ogden (1997) showed that using a phrase dictionary built by extracting phrases from parallel sentences in French and English improved the performance of their dictionary-based CLIR. Work in term disambiguation has been done by Ballesteros and Croft (1998) who demonstrated the effectiveness of translating phrases in Spanish queries into English phrases using terms which co-occur in the English collections.

To further mitigate the effect of mistranslated query terms, many researchers have employed query expansion techniques. Query expansion is a well-known method in IR for improving retrieval performance. Basically, it adds a query with new terms, selected using a certain technique, such that the query becomes more precise, as the added terms clarify the meaning of the original query terms, and its recall is improved, as terms associated with the original query terms are added. In monolingual IR, Sparck Jones (1971) proposed a query expansion technique which adds terms obtained from term clusters built based on co-occurrences of terms within documents. A similar approach using term similarity has also been done by Qiu and Frei (1993). In CLIR, Ballesteros and Croft (1997), and Carbonell et al. (1997) employed pseudo relevance-feedback techniques to obtain terms for query expansions. The pseudo relevance-feedback techniques assume that the top rank documents retrieved using the queries are relevant. Terms appearing in these relevant documents are then added onto the queries. These techniques were demonstrated to be effective in improving retrieval effectiveness.

Ballesteros and Croft (1998) proposed three query expansion methods, namely, pre-translation, post-translation and a combination of post and pre-translation methods using Local Context Analysis (LCA). LCA is a query-expansion technique which extracts expansion terms from the top-ranked text passages retrieved using the original query. They found that post-translation query expansion, i.e., query expansion on the translation queries, and the combination-translation query expansion, i.e., query expansion on both the original and the translation queries, are effective in improving CLIR performance.

Researches in CLIR have shown that it is possible to solve the language barrier which is the main problem in retrieving documents in language different from the query. Various techniques for sense disambiguation and query expansion are still of interest to CLIR researchers.

3. Algorithms

As mentioned previously, translating queries using a bilingual dictionary gives rise to a number of problems, namely, the ambiguity problem and problems with unidentified acronyms, names or proper nouns. In our work, we concentrate on solving the ambiguity problem by choosing the correct sense for each translated term.

3.1. Term disambiguation

We propose a term-sense disambiguation technique for selecting the best translation sense of a word from the possible senses given by a bilingual dictionary. Basically, given a set of original query terms, we select for each term the best sense such that the resulting set of selected senses contains senses that are closely related—or statistically similar—with one another. Finding such an optimal set is computationally very costly. In this work, we propose an approximate algorithm that requires a reasonable amount of computational time. Given a set of n original query terms $\{t_1, t_2, \ldots, t_n\}$, a set of translation terms, T, is obtained using the following algorithm:

```
1. For each $t_i$ ($i = 1$ to $n$), retrieve a set of senses $S_i$ from the dictionary.
2. For each set $S_i$ ($i = 1$ to $n$), do steps 2.1, 2.2 and 2.3.
2.1 For each sense $t'_j$ ($j = 1$ to $|S_i|$) in $\mathrm{S}_i$, do step 2.1.1
2.1.1 For each set $S_k$ ($k = 1$ to $n$ & $k$ <> $i$), get the maximum similarity, $M_{j,k}$, between
      $t'_j$ and the senses in $S_k$.
2.2 Compute the score of sense $t'_j$ as the sum of $M_{j,k}$ ($k = 1$ to $n$ & $k$ <> $i$).
2.3 Select the sense in $S_i$ with the highest score, and add the selected sense into the
    set $T$.
```

Note that original query terms which are not found in the dictionary are included in the translation set T un-translated. This is typically the case for proper names, technical terms, and acronyms.

We obtain the degree of similarity or association-relation between terms using a term association measure, called *Dice similarity coefficient* (van Rijsbergen 1979), which is commonly used in document and term clustering. The term association value, SIM_{xy}, between term x and y is calculated using the formula described below:

$$SIM_{xy} = 2\sum_{i=1}^{n}(w'_{xi} \cdot w'_{yi}) \Bigg/ \left(\sum_{i=1}^{n} w_{xi}^2 + \sum_{i=1}^{n} w_{yi}^2\right)$$

where

$$\begin{aligned} w_{xi} &= \text{weight of term } x \text{ in document } i \\ w_{yi} &= \text{weight of term } y \text{ in document } i \\ w'_{xi} &= w_{xi} \text{ if document } i \text{ also contains term } y\text{, or 0 otherwise} \\ w'_{yi} &= w_{yi} \text{ if document } i \text{ also contains term } x\text{, or 0 otherwise} \\ n &= \text{the number of documents in the collection.} \end{aligned}$$

The term weight, w_{xi}, of term x in document i is computed using the standard tf*idf weighting formula (Salton and McGill 1983) as follows:

$$w_{xi} = tf_{xi} \cdot idf_x$$

where *idf*, the inverse document frequency, is computed as follows:

$$idf_x = \log(n/df_x)$$

and

$$\begin{aligned} tf_{xi} &= \text{the number of occurrences of term } x \text{ in document } i \\ df_x &= \text{the number of documents containing term } x \text{ in the collection.} \end{aligned}$$

In short, our sense disambiguation algorithm computes the sum of maximum similarity values between each candidate translation of a query term and the translations of other terms in the query. For each query term, the translation term that has the highest sum is chosen as the query term's translation.

Our techniques use a term-similarity matrix built using the statistical term-distribution parameters obtained from our Indonesian corpus—for Indonesian terms, and a subset of our English collections—for English terms. As has been done by other researchers (Sheridan and Ballerini 1996, Ballesteros and Croft 1997, Carbonell et al. 1997), we use the target collections, instead of separate training collections, for this purpose since the statistical term-distribution parameters are readily obtainable from the text retrieval system which is used for indexing the documents.

3.2. *Query expansion*

As described previously in Section 2, query expansion has been known to be effective in improving the retrieval performance of translation queries. We pursue a similar approach in this paper but using a query expansion technique different from those used by previous researchers. First, we perform the query expansion selectively by considering only the top R query terms with the highest *idf*. This is based on the premise that *idf* indicates the degree of importance, or the discriminating power, of a query term. Next, we expand the queries by adding the top M terms that are most similar, in terms of Dice Coefficient, to the selected query term. Through a preliminary experiment, we established the optimal values (with respect to our test collections) of R and M.

4. Experiments

To evaluate the effectiveness of the sense disambiguation and query expansion techniques, we conducted a series of experiments using English and Indonesian corpora. The English corpus contains 242,918 documents of the TREC Associated Press (AP) collections, and the Indonesian corpus contains articles from *Kompas* (an Indonesian daily newspaper), totaling 20,590 documents. For the Indonesian to English CLIR, we manually translated the TREC's 24 queries for cross-language topics into Indonesian. The translation was done by the author who is a native speaker of Bahasa Indonesia. For the English to Indonesian CLIR, we manually created 26 English queries covering a number of major events occurred in Indonesia during the period when the newspaper collection was built. The relevance judgments for the Indonesian documents, with respect to the English queries, were established with the help of three Indonesian student volunteers. First, each of the volunteers identified the relevant documents among the top 100 documents retrieved using an IR system, called INQUERY (Callan et al. 1992), which has been modified to handle Indonesian documents and queries. A document is then marked as relevant, if and only if it is judged as relevant by all of the three volunteers.

We then built an English term-similarity matrix using a subset of the TREC AP collections, namely, the AP89 collection. We did not use the entire collections for this purpose as it would be too costly in terms of computation time and memory space requirement. As for the Indonesian term-similarity matrix, we built the matrix using the entire Indonesian corpus which is much smaller than the English corpus.

In our experiments we used *An English-Indonesian Dictionary* (Echols and Shadily 1992) to translate the Indonesian (English) queries to English (Indonesian) queries. Each query term was replaced manually by all of the senses in the dictionary for that term. Query terms that are not in the dictionary were kept as-is in the query. Phrase translation, to be discussed later, was done manually by matching the sequence of words that are found in the dictionary and replacing them with the translation. The document retrieval and its evaluation were conducted using the INQUERY system and the standard 11 recall-point evaluation method for TREC.

First, we compared the average retrieval precision using the sense disambiguation technique, referred to as the *automatic disambiguation* CLIR method, with that of three other methods, namely, (1) *monolingual*: retrieval using manually translated queries, (2) *simple translation*: retrieval using translation queries obtained by including all possible senses in the dictionary for each of the query terms, and (3) *best-sense translation*: retrieval using translation queries obtained by manually selecting the best sense among the senses in the dictionary for each query term. Since we were also interested in measuring the impact of our technique in mitigating the problem of translating query phrases, we compared the performance of both the *automatic disambiguation* and the *best-sense translation* methods with the same methods but enhanced with phrase translation. Lastly, we measured the effectiveness of our query expansion technique in further improving the retrieval performance of the *automatic disambiguation* method.

4.1. Results

As can be expected, the retrieval performance of the *simple translation* method is worse than that of the equivalent *monolingual* method. Tables 1 and 2 show that the average retrieval precision of these methods for the Indonesian translation queries (queries translated into Indonesian) and the English translation queries (queries translated into English) are 36.3% and 57.4% below that of the *monolingual* method, respectively.

The sense disambiguation technique in the *automatic translation* method helped improve the performance of the *simple translation* method. For the Indonesian translation queries, the average retrieval precision increased by 10.4%, amounting to 25.9% performance drop as compared to the equivalent *monolingual* method (see Table 1). However, it is still worse than the *best-sense translation* method by 7.7%.

A similar result was obtained for the English translation queries. As with the Indonesian translation queries, the average precision of the *automatic translation* method is below that of the *best-sense translation* method, but is better than that of the *simple translation* method by 15.5% (see Table 2).

Table 1. Average retrieval precision of the Indonesian translation queries.

	Performance & performance drop (%) compared to monolingual
Monolingual	0.2757
Simple translation	0.1757 (−36.3%)
Best-sense translation	0.2256 (−18.2%)
Automatic translation	0.2044 (−25.9%)

Table 2. Average retrieval precision of the English translation queries.

	Performance & performance drop (%) compared to monolingual
Monolingual	0.2804
Simple translation	0.1194 (−57.4%)
Best-sense translation	0.1925 (−31.4%)
Automatic translation	0.1628 (−41.9%)

Table 3. Average retrieval precision of the English translation queries with phrase recognition and translation.

Performance & performance drop (%) compared to monolingual	
Monolingual	0.2804
Simple translation with phrase handling	0.1425 (−49.18%)
Best-sense translation with phrase handling	0.2329 (−16.94%)
Automatic translation with phrase handling	0.1978 (−29.45%)

Table 4. Average retrieval precision of the expanded Indonesian translation queries.

Performance & performance drop (%) compared to monolingual	
Monolingual	0.2757
Automatic translation	0.2044 (−25.86%)
Automatic translation with query expansion	0.2288 (−17.01%)

Table 5. Average retrieval precision of the expanded English translation queries.

Performance & performance drop (%) compared to monolingual	
Monolingual	0.2804
Automatic translation	0.1628 (−41.96%)
Automatic translation with query expansion	0.1627 (−41.97%)

The result of our next experiment indicates that manually translating phrases in the queries, using a phrase dictionary, increased the retrieval performance of the English translation queries for both the *best-sense translation* and *automatic translation* methods (see Table 3). We did not conduct a similar experiment with the Indonesian translation queries as there are no phrases in the English to Indonesian CLIR queries that can be translated using the dictionary.

Next, we investigate the effectiveness of our query expansion technique in further improving the retrieval performance by measuring the performance of the *automatic translation* method enhanced with the query expansion technique. In these experiments the parameters M and R were set to 1 and 5, respectively. The results of this experiment are shown in Tables 4 and 5 for the Indonesian translation queries and the English translation queries, respectively. A significant performance improvement was obtained only for the Indonesian translation queries.

It is interesting to note that expanding the English translation queries after applying the phrase translation did not improve the retrieval performance, as can be seen in Table 6.

4.2. Analyses

The results of our experiments demonstrate that applying our sense disambiguation technique to the dictionary-based query translation method is an effective CLIR approach. The

Table 6. Average retrieval precision of the expanded English translation queries after manual phrase translation.

Performance & performance drop (%) compared to monolingual retrieval	
Monolingual	0.2804
Automatic translation with query expansion	0.1627 (−41.97%)
Automatic translation with phrase handling	0.1978 (−29.45%)
Auto-trans with phrase handling and query expansion	0.1683 (−39.98%)

result of our experiment with manual query phrase translation indicates that, while this technique can reduce the effect of the term ambiguity problem, it does not seem to help mitigate the effect of the phrase translation problem. In other words, the identification and translation of phrases need to be resolved first before the term disambiguation technique can produce good results.

On the other hand, we also observed that the presence of phrases in the input queries can help improving the effectiveness of the sense disambiguation technique. Comparing between the number of phrases in each of the original queries and the relative performance (compared to the *monolingual* method) of the translation query using the *automatic translation* method, we obtained the Spearman's rank correlation coefficient (Mendenhall et al. 1986) of 0.355 for the Indonesian queries and 0.212 for the English queries. The former correlation coefficient (0.355) is significant as it is well above the critical value for 24 observations (queries) and $\alpha = 0.05$, namely 0.343. The data used for the comparisons can be found in Appendix A. This finding suggests that the sense disambiguation technique resolves the ambiguity of words belonging to a phrase more effectively than stand-alone words (words not belonging to a phrase). The logical explanation is that phrase words tend to co-occur and so do their translations.

The result of our experiments with the Indonesian translation queries shows that adding related terms to a query improves its retrieval performance. However, expanding a query with too much ambiguity can have a detrimental effect, i.e., it tends to populate the query with too many irrelevant terms. We conclude, therefore, that the effectiveness of the query expansion technique depends on how much term ambiguity in the queries has been resolved. This is evident in the results of our experiments with the English translation queries (see Table 6).

Our investigation reveals that there is a dominant pattern of term-formation in Indonesian that differs from that in English, at least with respect to our queries. In the English TREC queries there are 11 single-word terms each of which translates into an Indonesian compound-noun or noun phrase consisting of two or more words (see Table 7), but, there are no English compound-noun or noun phrases in the queries that translate into single-word Indonesian terms. Likewise, in the Indonesian queries used for the monolingual retrieval there are 7 Indonesian compound-noun and noun phrases that translate into single-word English terms, but, again, there are no single-word Indonesian terms that translate into English compound-noun or noun phrases.

All this demonstrates that in Indonesian, a large portion of the language's vocabulary is formed by combining words into compound-noun phrases. Since many words have a wide range of meanings depending on their context of use, both as stand-alone words (i.e., polysemy) and as members of compound-noun phrases, translating an Indonesian term into

Table 7. English terms in the original TREC queries that translate into Indonesian compound-noun and noun phrases.

TREC query no.	English term	Indonesian translation
3.	*drug*	*obat* (substance) *bius* (sedative)
4.	*reusage*	*daur* (cycle) *ulang* (repeated)
5.	*acupuncture*	*pengobatan* (medication) *tusuk* (puncture) *jarum* (needle)
6.	*automobile*	*kendaraan* (vehicle) *bermotor* (engine-powered)
8.	*highway*	*jalan* (way) *bebas* (free) *hambatan* (obstruction)
9.	*logging*	*penebangan* (cutting) *pohon* (tree)
9.	*desertification*	*kegundulan* (boldness) *hutan* (forest)
16.	*resurgence*	*kemunculan* (occurrence) *kembali* (again)
19.	*wine*	*minuman* (liquor) *anggur* (grape)
22.	*chocolate*	*permen* (candy) *coklat* (chocolate)
23.	*fast food*	*makanan* (food) *cepat* (fast) *saji* (served)

English is much more difficult than vice versa. In other words, the degree of ambiguity of an Indonesian word is much higher than that of an English word. More specifically, in the queries used in our experiments, the average number of senses in the dictionary for each Indonesian query word is 11.8 senses, whereas, the average number of senses for each English query word is only 4.3 senses.

5. Summary and conclusion

The availability of machine-readable bilingual dictionaries has made dictionary-based approaches to CLIR practical. One of the central research issues in CLIR, which is also the main problem faced by dictionary-based query translation techniques, is term ambiguity (Hull and Grefenstette 1996). We proposed a sense disambiguation technique based on the concept of statistical term similarity for translating queries in one language to another using a bilingual dictionary. We proposed a sense disambiguation technique based on the concept of statistical term similarity for translating queries in one language to another using a bilingual dictionary.

We have demonstrated that the technique improved the retrieval effectiveness of translation queries in both Indonesian to English and English to Indonesian CLIR experiments, particularly when there are phrases in the queries. Once a reasonably good translation query is obtained, our query expansion technique, which is also based on the term similarity, can improve the retrieval performance further.

Our work with term-similarity based sense disambiguation and query expansion has provided us with a vehicle for investigating the impact of various language-specific term characteristics, including term distribution parameters, on the retrieval performance of CLIR. Along this line, we have presented a finding which indicates that differences in the pattern of word formation between two languages can render automatic query or document translation difficult.

This study is part of an on-going research in an effort to build a model for CLIR. Our short-term objective in this research is to identify the statistical property, or properties, of terms which contributes to the term ambiguity and the phrase translation problems.

Appendix A: Number of phrases vs average retrieval precision

Table A. The number of phrases and the relative performance (the average retrieval precision, compared to the *monolingual* method) of each query's *automatic translation* method.

	Indonesian queries		English queries	
Query no.	Phrases	Relative performance (%)	Phrases	Relative performance (%)
1.	1	115.62	0	17.11
2.	0	0.00	0	55.15
3.	1	37.32	0	44.23
4.	1	216.46	1	100.00
5.	1	73.30	1	200.00
6.	1	82.43	0	97.20
7.	1	100.44	1	200.00
8.	3	163.86	0	164.20
9.	2	15.77	1	14.50
10.	1	18.75	1	99.27
11.	2	2.73	1	94.95
12.	2	91.21	1	47.25
13.	3	106.67	1	103.89
14.	1	62.48	1	1.18
15.	2	16.81	1	26.70
16.	2	162.50	1	95.84
17.	4	92.88	2	100.00
18.	2	35.70	0	2.09
19.	2	147.98	1	217.97
20.	2	101.88	1	46.59
21.	1	0.00	1	100.00
22.	1	4.99	2	0.00
23.	1	86.72	2	104.03
24.	1	17.06	2	127.51
25.	–	–	1	100.00
26.	–	–	1	200.00

Acknowledgments

We would like to thank Prof. W. Bruce Croft at University of Massachusetts in Amherst who allowed us to use the INQUERY system for this research. Our appreciation also goes to Hendarto and Andrey Andoko at Harian Kompas (http://www.kompas.com) who made the newspaper collections available to us.

References

Adriani M and Croft WB (1997) The effectiveness of a dictionary-based technique for Indonesian-English cross-language text retrieval. CIIR Technical Report IR-170. University of Massachusetts, Amherst.

Ballesteros L and Croft WB (1997) Phrasal translation and query expansion techniques for cross-language information retrieval. In: Belkin NJ, Narasimhalu AD and Willet P, Eds., Research and Development in Information Retrieval, pp. 84–91.

Ballesteros L and Croft WB (1998) Resolving ambiguity for cross-language retrieval. Research and Development in Information Retrieval, pp. 64–71.

Callan JP, Croft WB and Harding SM (1992) The inquery retrieval system. International Conference on Database and Expert Systems Applications.

Carbonell J, Yiming Y, Frederking R, Brown RD, Geng Y and Lee D (1997) Translingual information retrieval: A comparative evaluation. International Joint Conference on Artificial Intelligence (IJCAI).

Davis M and Dunning TE (1995) A TREC evaluation of query-translation methods for multi-lingual text retrieval. In: Harman D, Ed., The Fourth Text Retrieval Conference (TREC-4). NIST, Gaithersburg, MD.

Davis MW and Ogden WC (1997) Free resources and advanced alignment for cross-language text retrieval. In: Harman DK, Ed., The Sixth Text Retrieval Conference (TREC-6). NIST, Gaithersburg, MD.

Echols JM and Shadily H (1992) An Indonesian-English dictionary. Gramedia Pustaka Utama, Jakarta.

Kraaij W and Hiemstra D (1997) Cross language retrieval with the twenty-one system. In: Donna H, Ed., The Sixth Text Retrieval Conference (TREC-6). NIST, Gaithersburg, MD.

Hull DA and Grefenstette G (1996) Querying across languages: A dictionary-based approach to multilingual information retrieval. Research and Development in Information Retrieval, pp 49–57.

Harman D (1997) Overview of the sixth text retrieval conference. In: Harman D, Ed., The Sixth Text Retrieval Conference (TREC-6). NIST, Gaithersburg, MD.

Mendenhall W, Scheaffer RL and Wackerly DD (1986) Mathematical Statistics with Applications, 3rd ed. Duxbury Press, Boston.

Oard DW and Hackett P (1997) Document translation for cross-language text retrieval at the University of Maryland. In: Harman D, Ed., The Sixth Text Retrieval Conference (TREC-6). NIST, Gaithersburg, MD.

Qiu Y and Frei HP (1993) Concept based query expansion. Research and Development in Information Retrieval, pp. 160–169.

van Rijsbergen CJ (1979) Information Retrieval, 2nd ed. Butterworths, London.

Salton G (1970) Automatic processing of foreign language documents. Journal of the American Society for Information Science, 21:187–194.

Salton G and McGill MJ (1983) Introduction to Modern Information Retrieval. McGraw-Hill, New York.

Sheridan P and Ballerini JP (1996) Experiments in multilingual information retrieval using the SPIDER system. Research and Development in Information Retrieval, pp. 194–208.

Sparck Jones K (1971). Automatic Keyword Classifications for Information Retrieval. Butterworths, London.