

A study on automatic creation of a comparable document collection in cross-language information retrieval

Tuomas Talvensaari¹, Jorma Laurikkala¹, Kalervo Järvelin² and Martti Juhola¹

¹*Department of Computer Sciences, 33014 University of Tampere, Finland*

²*Department of Information Studies, 33014 University of Tampere, Finland*

Keywords *Cross-language information retrieval, comparable collection, relevance assessment*

Abstract *We present a new method for creating a comparable document collection from two document collections in different languages. The best query keys were extracted from a Finnish source collection (articles of the newspaper Aamulehti) with the relative average term frequency (RATF) formula. The keys were translated into English with a dictionary-based query translation program. The resulting lists of words were used as queries that were run against the target collection (Los Angeles Times articles) with the nearest neighbor method. The documents were aligned with unrestricted and date-restricted alignment schemes, which were also combined. The combined scheme was found the best, when the relatedness of the document pairs was assessed with a five-degree relevance scale. Of the 400 document pairs, roughly 40% were highly or fairly related and 75% included at least lexical similarity.*

The authors are grateful to the Academy of Finland for financial support of the present research (grants 202185, 206568, 200844, 80771, and 204978).

1. Introduction

In traditional information retrieval tasks, queries and documents are in the same language. Conversely, in cross-language information retrieval (CLIR) (Oard and Diekema, 1998), the language of the queries (source language) and the language of the document collection (target language) are different. The problem is basically similar to that of the single language searches: to find documents in a collection that best match the user's request, but, additionally, we have to somehow cross the language barrier. After the huge growth of the multilingual Internet, cross-language information retrieval has become more and more important (Grefenstette, 1998).

There are various approaches to query formation in CLIR. Oard and Diekema (1998) represent a framework, where query formulation is examined from the viewpoints of different matching strategies and the sources of translation knowledge needed in the matching. Cognate matching does not involve actual translation. Instead, rules to identify similarities in spelling or pronunciation are applied. For instance, proper nouns and technical terms can be similar between languages. Such words often vary a little, which allows the usage of approximate string matching, like *n*-grams (Pirkola *et al.*, 2002a; 2003). Conversely, query translation, document translation, and interlingual matching techniques require deeper translation knowledge, which can be drawn from ontologies, bilingual dictionaries, machine translation lexicons, or corpora. The query translation approach is the most popular in CLIR and is also used in this study. The target document collection could be translated into the source language, but it would be a very complex task and it is far easier to translate concise queries (Oard and Dorr, 1996; Oard and Diekema, 1998). The interlingual techniques convert both the query and the documents into a language-

independent representation. However, some of these techniques, such as the latent semantic indexing, are computationally intensive.

In dictionary-based query translation, the query keys are simply replaced by their counterparts in a bilingual dictionary (Hull and Grefenstette, 1996), while machine translation systems translate the source language request into target language using a lexicon containing information for the automatic analysis, translation, and generation of natural language (Oard and Diekema, 1998). Corpus-based translation utilizes multilingual document collections, where the documents of the two languages are aligned as pairs so that each pair contains a translation of each other (parallel corpora) or at least deal with the same topic (comparable corpora). After having created a cross-language corpus, the queries can be translated, for example, as in (Sheridan and Ballerini, 1996), where the aligned corpora was used to “expand” the source language queries with target language words that co-occur with the query keys in the aligned corpora. Ballesteros and Croft (1998) applied this technique to prune extraneous translation alternatives in dictionary-based translation. Sometimes even unparallel document collections can be useful if the collections share a similar, limited domain (Picchi and Peters, 1998).

The translation techniques are not mutually exclusive, but, on the contrary, can be used jointly. Dictionary-based translation is often the starting point, where all translation alternatives of a word of the source language are given from the dictionary. Using this technique alone is nonetheless problematic according to Ballesteros and Croft (1998), who listed three weaknesses. First, some of the translation alternatives may not correspond to the words of the query in the sense desired by the user. Extraneous terms increase the ambiguity of the query, which in turn damages retrieval performance. Second, dictionar-

ies are limited in scope. Special terms and proper nouns are often absent from general dictionaries. Third, the recognition and translation of phrases constructed from several words can be difficult. However, this is not a major problem in languages where such phrases often form compound words spelled together. The ambiguity introduced by translating the query can be dealt with in many ways (see Pirkola *et al.* 2001). For example, in part-of-speech tagging the translation alternatives that have the same part-of-speech as the source language words are selected.

Extensive parallel corpora are hard to obtain. For example, the minutes of the United Nations (Davis, 1998) and the Bible have been tested as parallel corpora. However, such collections are restricted to particular topics. Because of the problems in constructing such collections, also comparable collections have been used; see for example Sheridan and Ballerini (1996) and Braschler and Schäuble (1998). It is much easier to find document collections sharing similar topics than it is to find collections that are translations of each other. Since translation knowledge can also be obtained from a comparable collection, it would be beneficial to be able to automatically build such collections from two or more collections in the same domain, for instance in the news domain. In this study we propose such a method.

Previously, the automatic creation of comparable corpora has been studied, for example, by Sheridan and Ballerini (1996) and by Braschler and Schäuble (1998). Sheridan and Ballerini used document meta-descriptors and publishing dates to align German and Italian news stories by the Swiss news agency SDA. Braschler and Schäuble made use of common proper nouns and numbers, dates, and a small dictionary to combine SDA documents in various languages. Although the SDA stories are not translations of each

other, they are quite similar, because they are composed in the same country. Our method was tested with the Aamulehti (in Finnish) and Los Angeles Times newspaper article collections which are very different in origin. Also, we did not use content meta-descriptors in producing the alignments, if the publishing date is not considered as one. Thus, our method can be applied to collections that are less structured and documented than the SDA collection.

Queries were created from the Finnish text documents with a new approach. We selected the best query keys from the Finnish documents by means of the relative average term frequency (RATF) of Pirkola *et al.* (2002b), which is a new approach to create query vocabulary, and translated queries prepared from the source language keys using the UTACLIR (Keskustalo *et al.*, 2002) query translation program. A morphologically complex language, such as Finnish, as a source language in creating a comparable collection, can also be considered as a novel property. Documents were aligned using unrestricted search, search restricted with date differences, and a combination of these alignment schemes, which turned out to be the best method. The relatedness (or similarity) of the document pairs was manually assessed with a five-level scale. The results were promising: Roughly 40% of the pairs were highly or fairly related and 75% included at least lexical similarity.

2. Test collections

We employed two documents collections: one from a large Finnish newspaper Aamulehti (see <http://www.almamedia.fi/home>) and the other from Los Angeles Times. The collections are part of the Cross-Language Evaluation Forum (CLEF) conference test collection (Peters, 2003). The Aamulehti collection consisted of 54 851 news articles with the aver-

age length of 260 words. The articles were published between the 18th of November 1994 and the 31st of December 1995. The target collection consisted of 113 005 articles with the average length of 572 words, all published in 1994. Thus the common period of the collections is only about six weeks, which affected the creation of a comparable collection significantly. To compare the two average lengths or the numbers of the words in documents, note that compound words are much more abundant in Finnish than in English.

The 8 878 Aamulehti articles from 1994 were manually scanned to find documents that in principle could be aligned with a document from the target collection. A total of 682 documents were chosen. The limited number was due to the geographical distance of the two collections and their short common period (about six weeks). Most of the Aamulehti articles were of national and local topics that are not likely to be addressed in a U.S. newspaper. The terms of the source document collection were normalized with the TWOL program (Koskenniemi, 1983) that also split compound words. TWOL lemmatizes words into their morphological base forms and splits compound words to their constituents. The very frequent words were eliminated first by applying a list of 722 stop words, and, then, by removing words that occurred in more than 10 000 documents. Words appearing only once in the collection were removed. Further, all words which had document frequencies equal to their collection frequencies were deleted – these words appear exactly once per every document. After the removal of the rare and very frequent words there were 149 993 keys.

The indexing of the target collection was performed much like the method of Salton and McGill (1983). First, the ‘s’ suffixes of the genitive were removed. Stop words were

removed using a list of 435 words. The words were then returned to their stems with the Porter’s stemming algorithm (Porter 1980). Lastly, all words appearing only once in the collection were removed. The index comprised of 108 654 keys, from which more than a half appeared in less than six documents.

3. Construction and translation of queries

After the initial preprocessing, we continued with a more accurate selection of keys to construct a *query vocabulary* from the source documents. For this task we applied the relative average term frequency (RATF) formula, which has been found useful both in the monolingual and cross-language information retrieval (Pirkola, *et al.*, 2002b). Similar to the straightforward document frequency-based selection, the RATF formula utilizes the key frequencies, but considers them more carefully, and, moreover, the formula may be adapted to the collection. For these reasons, we refined the coarsely selected set of keys with this method. To our knowledge, this is a novel application of the RATF formula.

The relative average term frequency of a key j ($1 \leq j \leq M$) is defined as

$$RATF_j = \frac{cf_j}{df_j} \cdot 10^3 / \ln(df_j + SP)^p = \frac{\sum_{i=1}^N tf_{ij}}{df_j} \cdot 10^3 / \ln(df_j + SP)^p,$$

in which SP is a collection dependent scaling parameter, p is a power parameter, df_j is the document frequency of word j , tf_{ij} is equal to the frequency of key j in document i ($1 \leq i \leq N$), N is the number of documents of the collection, and M equals the number of keys. The formula gives more weight to keys whose average term frequency, that is the ratio of the collection (cf_j) and document (df_j) frequency, is high. The rare words are penalized

with the scaling SP parameter which weights rare words down. We used SP and p values of 3 000 and 3, respectively. These values were determined experimentally. The terms were sorted along with decreasing order given by the preceding formula. Pruning with a threshold value of 2.4, which was also decided experimentally, yielded a query vocabulary of 88 312 keys.

After the construction of the vocabulary, queries representing the source documents were formed as follows: for each document the keys that appeared in the query vocabulary were sorted by decreasing frequency within the document and ten top ranking keys were selected into the query representing the document. If less than ten document keys appeared in the vocabulary, all of the keys were chosen. If more than one key shared the tenth place, all such keys were taken into account. The number of keys to include in the queries was decided experimentally. Some of the queries contained rather many keys, which slowed down the later processing. The average length of the 682 queries was 14.6 keys. A total of 93 (14%) queries had more than 20 keys and 13 (2%) queries consisted of more than 30 keys. The long queries could have been shortened simply by selecting, for example, the first ten words from the frequency list, regardless of shared ranks. According to Pirkola and Järvelin (2001), even two or three best query keys may suffice in a monolingual context. We also tried queries containing seven keys, but the results were discouraging.

To translate the queries we applied UTACLIR, a dictionary-based query translation program (Keskustalo *et al.*, 2002). UTACLIR analyses a source language query morphologically with the help of TWOL and removes stop words. Thereafter, it replaces the query keys with their target language translation alternatives. The UTACLIR version

used here included the GlobalDix Finnish-English dictionary, which is rather limited in vocabulary. For example, it does not contain proper nouns, such as country names. This is a problem for our research, since country names and other proper nouns are very common and important in news articles. Thereby, we took advantage of a brief bilingual list of proper nouns that was principally comprised of names of countries and cities in order to complete the translation task. The dictionary should not be too extensive, because additional translation alternatives bring more ambiguity.

Words absent from the dictionary and the word list were handled with approximate string matching. UTACLIR splits words into so-called *s*-grams (Pirkola *et al.*, 2002a) that differ from the ordinary *n*-grams in the sense that also not strictly successive symbols of a string are used, but subsequences of symbols, where one symbol is left out or skipped between a predecessor symbol and its successor symbol. For instance, the Finnish word *Moskova* (*Moscow*) yields digrams $A_0 = \{MO, OS, SK, KO, OV, VA\}$, when zero characters are skipped, $A_1 = \{MS, OK, SO, KV, OA\}$, when one character is skipped etc. Correspondingly, we obtain $B_0 = \{MO, OS, SC, CO, OW\}$ and $B_1 = \{MS, OC, SO, CW\}$, etc., for its English translation. *S*-grams have been found are better than *n*-grams, particularly for short words (Pirkola *et al.* 2001). When computing the similarity with *s*-grams, the sets of digrams obtained by skipping different numbers of characters are compared in a unique way, as explained in Pirkola *et al.* (2002a) and Keskustalo *et al.* (2003). The preceding and trailing spaces may also be taken into account.

4. Assessment of the relatedness of document pairs

The traditional binary relevance assessment is liberal, because it does not quantify the degree of relevance. It only indicates whether the document can be said to be relevant or

irrelevant to the query. Conversely, the graded relevance assessments allow fine-grained analyzes of retrieval performance. Graded relevance has recently been studied, for example, by Voorhees (2001), Sormunen (2002), and Kekäläinen and Järvelin (2002).

Since we felt that a graded scale would be useful also in our study, we evaluated the document pairs with the five-level relatedness scale introduced by Braschler and Schäuble (1998). We use here instead of values “same story”, “related story”, “shared aspect”, “common terminology, and “unrelated” the respective, but more convenient, values “highly related”, “fairly related”, “marginally related”, “weakly related”, and “unrelated”. In the following, the relatedness scale is described with the help of examples:

- **Highly related pair.** The two documents consider the same event, for example:
 - *Angolaan saatiin vihdoin rauhansopimus; Unitan johtaja ei tullut tilaisuuteen, sodan mahdollisuus yhä suuri* (Aamulehti, the 21st November, 1994) [Peace treaty finally signed in Angola; the leader of Unita did not attend, possibility of war still remains significant.]
 - Angola peace treaty signed, but long conflict continues (L.A. Times, the 21st November, 1994).
- **Fairly related pair.** The two documents consider different event more or less, but they are clearly connected to each other. They may also consider the same event, only from somewhat different viewpoints. They may also only partially be same, that is they tell about the exactly same event, but it is a part from the whole story. An example follows:
 - *Kiina 'tutki': Spratlysaaret kuuluvat meille* (Aamulehti, the 6th December, 1994). [China 'investigated': “the Spratly islands belong to us”.]

- China, Vietnam hold formal talks on longtime land and sea disputes (L.A. Times, the 20th August, 1994).
- **Marginally related pair.** The documents deal with, for example, happenings of the same area or the same persons are named. Their relation is weaker than in the preceding relevance stage, likewise:
 - *Jeltsin: politiikka tympii kansalaisia* (Aamulehti, the 19th November, 1994). [Yeltsin: politics bore citizens.]
 - Yeltsin vows to take offensive on slumping economy (L.A. Times, the 27th November, 1994).
- **Weakly related pair.** Correspondence between the documents is slight, but still existent. An example tells about medicine, but not about the same topic:
 - *Japanissa tuhannet saaneet vaarallista verivalmistetta* (Aamulehti, the 23rd November, 1994). [Thousands have received dangerous blood preparation in Japan.]
 - Blood money; Medicine: Tiny hemacare has a potentially promising plasma technique in the fight against AIDS (L.A. Times, the 22nd November, 1994).
- **Unrelated pair.** There is no relation between the documents or it is really negligible.

5. Document alignment formation

Using the laboratory model of information retrieval (Hull 1996) to evaluate the similarity of the document pairs obtained was not an option, because we did not know in advance, which documents were relevant from the viewpoint of the documents of the source collection. An exhaustive search to examine all possible relevance relations between the documents would have been far too large a task.

We used nearest neighbor (NN) searching (Mitchell 1997), a well-known search and classification method, to align the documents. Since the NN searching is based on similarities (or distances) between the pairs of objects, it can be effortlessly adapted to the vector space model of information retrieval. Furthermore, the simple NN method was easy to modify to meet our needs. Adjusting the existing search engines, such as InQuery (Callan *et al.*, 1992), would have been more difficult than the simple NN method.

At first, we experimented with matching based on the traditional cosine measure and a document key weighting based on the following variation of the well-known *tf·idf* formula

$$d_{ij} = tf_{ij} \cdot \ln(N / df_j),$$

where tf_{ij} ($1 \leq i \leq N$, $1 \leq j \leq M$) equals the frequency of the j th key in document \mathbf{D}_i and df_j equals the document frequency of the j th key. Our preliminary tests implied the typical weakness of cosine measure; it considerably favors short documents. Consequently, we adopted the method of pivoted length normalization by Singhal *et al.* (1996) to overcome the problem by using the formula:

$$sim(\mathbf{Q}, \mathbf{D}_i) = \frac{\sum_{j=1}^M w_{qj} \cdot w_{ij}}{\left((1 - slope) + slope \cdot \frac{\sqrt{\sum_{j=1}^M (w_{ij})^2}}{pivot} \right) \cdot \sqrt{\sum_{j=1}^M (w_{qj})^2}},$$

in which a similarity value between document \mathbf{D}_i and query \mathbf{Q} vector is assessed. Here w_{qi} and w_{ij} are the query vector and document vector components of the query vector \mathbf{Q} and document vector \mathbf{D}_i . Value *pivot* is a bound: shorter documents than *pivot* are relatively penalized more than in cosine normalization and longer documents are treated

more leniently. The value of *pivot* can be defined as the mean of the lengths of the document vectors in a collection; its values were 15.9 and 14.6 for the Los Angeles Times and Aamulehti collections, respectively. Value *slope* from interval (0,1) also depends on a collection.

In order to find an optimal *slope*, we ran test searches with the CLEF 2002 test topics from 91 to 140. The Finnish topic descriptions were analyzed and translated as the queries representing the source documents (see Section 3). The CLEF queries are naturally available in English, but translations were used with the Los Angeles Times collection, because also the queries created from the source documents are translated. Effectiveness was assessed with the standard 11-point recall-precision values. Table 1 presents the average precisions of the pivoted normalization scheme with eight different *slope* values and improvements (in percents) incurred by pivot normalization compared to the average precision of the standard cosine measure (0.173). The improvements for all the *slope* values are clear, but the differences between them are slight. The *slope* value 0.450 gave the best precision average.

The documents were aligned with the NN searching using four different schemes. First, *unrestricted* searching was performed so that the similarity between each query and all the documents in the target collection was computed. The document most similar to the source document, that is the nearest neighbor, was chosen as the pair of the source document. Secondly, to accommodate the possible different dates due to the geographical remoteness of over 8 000 kilometers between the two newspapers, the NN searching was restricted to target articles whose publication date differed no more than one day from that of the source document (*date-restricted scheme A*). Thirdly, in the *date-restricted*

scheme B, the searches were performed with a wider time window of four days: Documents published at most two days before or after the source document were searched. Fourth, the *combined scheme* performed search in two stages when needed. The search was first performed with the date-restriction A. If the similarity between the source document and its NN exceeded a given threshold, the top ranking target document was posed to be the pair of the source document, else unrestricted search was made and its result was paired with the source document.

6. Results

Before moving to the actual document alignment, we made a small comparison to be sure that the simple NN searching produced good enough results. The Los Angeles Times collection was searched with the NN method and the InQuery search engine (Callan *et al.*, 1992) using the 50 translated CLEF topics (see Section 5). Both the standard cosine normalization and the pivoted length normalization was applied in connection with the NN searches. Besides the bilingual searches, the NN searches were also made with the corresponding English CLEF topics to illustrate the harm caused by the translation to retrieval performance. Figure 1 depicts an 11-point recall-precision curve for the four different retrieval methods with the CLEF material. The results of InQuery and the NN method were fairly equal, whereas the pure cosine measure was inferior. However, we ought to recall that InQuery uses a different index of the Los Angeles Times collection than the others. InQuery's index was normalized using TWOL, whereas the NN searches were made with stemmed index. The comparison illustrates, however, that the NN method delivers query performance that was good enough to be used in this study.

Queries generated from the source documents and translated by UTACLIR were run against the Los Angeles Times collection using the NN searching. From the available 682 queries 400 randomly selected were utilized, because we did not have enough resources to evaluate the relevance of a larger set of document pairs. However, since the sample was large (59%), it is likely that the results are quite similar to those that would have been obtained with the whole source collection.

In the combined alignment scheme, a low threshold (great confidence on date restriction) gave several precise pairs, but also increased the number of bad alignments, while a high threshold (cautious selection that often relies on unrestricted search) decreased the number of bad pairings. To determine a suitable threshold, the first author examined the alignment of a 100 randomly sampled source documents according to the five-degree scale with different thresholds (see Table 2). The selected threshold 1.6 represents a cautious case.

▲ Figure 2 presents the distribution of relatedness assessments for the four alignment schemes. Date restriction gave more results highly related document pairs (“same story”) than the others, but it also generated more unrelated pairs. Increasing the interval of the dates weakens precision as the number of highly and fairly related (“related story”) document pairs decreases. Unrestricted searches decreased the number of unrelated pairs, but increased the number of weakly related pairs (“common terminology”). To compare statistically the alignment scheme results, we applied the χ^2 goodness-of-fit test to the pairs of the relatedness assessment distributions. The tests showed that only the two the date-restricted schemes corresponded to each other ($p = 0.683$). The other alignment

Muotoiltu: englanti
(Iso-Britannia)

schemes produced significantly dissimilar distributions ($p < 0.001$). The combined alignment scheme was judged the best, because it produced the least unrelated pairs.

The 100 document pairs were also presented to an external evaluator, who was informed about the five relatedness grades. Table 3 shows that the evaluation was rather subjective. The difference in the number of the weakly related pairs was not surprising, because their definition is quite open to interpretations. One would expect that the best and worse pairs would be the easiest to identify, but, surprisingly, the second evaluator found twice as many highly related and unrelated pairs as the first evaluator. We investigated the association between the grades of the evaluators with the Spearman rank-order correlation coefficient, which was a better choice than Pearson correlation because of the ordinal variables, and found a positive dependence between the distributions, 0.728 ($p < 0.001$). After all, the distributions agreed better than what it looked like at first glance.

Figure 3 depicts evaluation distributions for the 400 selected document pairs aligned with the combined scheme. Each evaluator handled 200 document pairs so that the 100 pairs mentioned earlier were incorporated into the pairs of the first evaluator. The evaluation results are separately given in Figure 3 as well as joined together. There appears to be similar variation between the evaluators' results as in Table 3; the second evaluator judging more pairs as highly related and unrelated.

7. Discussion

We constructed a comparable document collection by pairing 400 Finnish newspaper articles with the documents of the Los Angeles Times collection. The source collection was normalized with the FINTWOL program, and, thereafter, the query vocabulary was extracted by applying the relative average term frequency (RATF) formula. The 10 most

frequent keys of the vocabulary within each source document were used as the query representing the document. The translation of the Finnish queries into the target language was handed by the UTACLIR program. The documents were aligned with the nearest neighbor (NN) method using unrestricted as well as restricted schemes and a mixture of these. Lastly, the quality of the document pairs was assessed by means of a five-level relevance scale.

The combined alignment scheme produced the best results which were promising: About 40% of the pairs were highly or fairly related and about 75% of the pairs shared at least some vocabulary (see Figure 3). Along with these results, it must be noted that the test collections of this study were much more different in origin, than, for example, the collections used by Sheridan and Ballerini (1996). Expectedly, a highly related alignment pair was most likely found when the source article dealt with a unique or unforeseeable event, such as the discovery of oil from the Windsor Castle grounds in November 1994. The event is clearly distinguished from other news articles; it does not deal with a common news topic of the time, such as the war in Bosnia. The fairly related pairs were not as clearly distinguished as the highly related ones. More common news topics – Bosnian war, the conflict in Chechnya, and the NHL strike – appear. It is difficult to find an exact match from the target collection for such documents.

Most of the marginally relevant document pairs dealt with the common news topics of the day. Distinction between the marginally and weakly related pairs as well as that between weakly related and unrelated pairs was often vague, which complicated searching tasks. As an example, consider a pair where the source document was of the interest policy of the German central bank and the target article gave hints for those planning busi-

ness trips to Germany. There were common factors, Germany and business life, but the articles were fully different. Depending on the final purpose of the resulting comparable collection, however, the distantly related pairs might also be acceptable. After all, even the weakly related pairs have common terminology.

The problems of the proposed method can be categorized according to the phases of the study. Firstly, some unsuccessful pairings were caused by the original source documents. Some articles simply did not have a satisfying match in the target collection. Misspellings and repeated sentences in the source documents contributed also to the failures. Secondly, the indexing of the source collection had some shortcomings. Both the FINTWOL program and the selection of the ten top-ranking key produced sometimes excessive search keys which brought ambiguity into the queries. In addition, some words, such as *Tshetshenia (Chechnya)* remained unrecognized, because some rarer proper nouns were missing from the vocabulary of the program. However, the problems related to FINTWOL were minor and the program served quite well the aims of this study. Lastly, some problems were attributable to query translation with the UTACLIR program. As discussed earlier, we had to construct a small bilingual word list, because the dictionary used by UTACLIR had no proper nouns. The list and the approximate string matching feature of UTACLIR may have conflicted, because they both aim to translate words absent from the dictionary. Another problem with the dictionary-based translation was the extraneous translation alternatives, which gave fallacious translations outside the context.

The above mentioned problems are partly such that we can address them in future search. The selection of query keys is one aspect of our research that could be improved. The fine tuning of the RATF value could perhaps more efficiently separate the good and

bad keys. Shortening queries by using always at most ten keys would be a straightforward improvement. The trouble caused by long queries is emphasized in cross-language information retrieval, because a dictionary-based translation usually produces several alternatives for a word of the source language. Thus, the number of words in the target language query can be multiple compared to that of the source language query. The approximate string matching of UTACLIR seemed to be unsuccessful, when used with our own bilingual proper noun list. Perhaps it would be useful to continue our research without it. There are also possibilities to develop the NN searching, for example, by incorporating structured queries, which Pirkola *et al.* (2001) have showed to be useful in cross-language information retrieval. Moreover, it might be reasonable to consider more neighbors than the closest one, because the highest ranking document may not always be the best choice for the alignment pair.

Furthermore, we are going to experiment with larger numbers of document pairs. A problem of our present collections was the shortness of their common time period, about six weeks, and their geographical remoteness, which reduced the number of valid source documents. We shall also widen our research to other languages included in the CLEF project. Further, we will make full-scale CLIR experiments based on the aligned corpora and compare the results to dictionary-based translation.

References

- Baeza-Yates, R. and Ribeiro-Neto, B. (1999), *Modern Information Retrieval*, ACM Press and Addison-Wesley, New York.
- Ballesteros, L. and Croft, W.B. (1998), "Resolving ambiguity for cross-language retrieval", in Croft, W.B., Moffat, A., van Rijsbergen, C. J., Wilkinson, R. and Zobel, J. (Eds.), *Proceedings of the 21st Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, pp. 64-71.
- Braschler, M. and Schäuble P. (1998), "Multilingual information retrieval based on document alignment techniques", in Nikolaou, C. and Stephanidis, C. (Eds.), *Proceedings of the 2nd European Conference on Research and Advanced Technology for Digital Libraries*, Lecture Notes in Computer Science 1513, Berlin, Springer, pp. 183-97.
- Callan, J.P, Croft, W.B. and Harding S.M. (1992), "The INQUERY retrieval system", in Tjoa, A.M. and Ramos, I. (Eds.), *Proceedings of DEXA-92, the 3rd International Conference on Database and Expert Systems Applications*, Vienna, Springer, pp. 78-83.
- Davis, M.W. (1998), "On the effective use of large parallel corpora in cross-language text retrieval", in Grefenstette, G. (Ed.), *Cross-Language Information Retrieval*, Kluwer Academic Publishers, pp. 11-22.
- Grefenstette, G. (1998) "The problem of cross-language information retrieval", in Grefenstette, G. (Ed.), *Cross-Language Information Retrieval*, Kluwer Academic Publishers, pp. 1-9.

- Hedlund, T., Airio, E., Keskustalo, H., Lehtokangas, R., Pirkola, A. and Järvelin, K. (2003), "Dictionary-based cross-language information retrieval: Learning experiences from CLEF 2000-2002", *Information Retrieval*, Vol. 7 No. 1/2, pp. 99-119.
- Hull, D.A. (1996), "Stemming algorithms: a case study for detailed evaluation", *Journal of the American Society for Information Science*, Vol. 47 No. 1, pp. 70-84.
- Hull, D.A. and Grefenstette, G. (1996), "Querying across languages: a dictionary-based approach to multilingual information retrieval", in Frei, H.-P., Harman, D., Schäuble, P. and Wilkinson, R. (Eds.), *Proceedings of the 19th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, pp. 49-57.
- Kekäläinen, J. and Järvelin, K. (2002), "Using graded relevance assessments in IR evaluation", *Journal of the American Society for Information Science and Technology*, Vol. 53 No. 13, pp. 1120-29.
- Keskustalo, H., Hedlund, T. and Airio, E. (2002), "UTACLIR - general query translation framework for several language pairs", in Järvelin, K., Beaulieu, M., Baeza-Yates, R. and Myaeng, S.H. (Eds.), *Proceedings of the 25th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, pp. 448-448.
- Keskustalo, H., Pirkola, A., Visala, K., Leppänen, E. and Järvelin, K. (2003), "Non-adjacent digrams improve matching of cross-lingual spelling variants, in Nascimento, M.A., de Moura, E.S. and Oliveira, A.L (Eds.), *Proceedings of the 10th International Symposium, SPIRE 2003*, Lecture Notes in Computer Science 2857, Berlin, Springer, pp. 252-65.

- Koskenniemi, K. (1983), *Two-Level Morphology: A General Computational Model for Word-Form Recognition and Production*, Publications of the Department of General Linguistics, University of Helsinki, No. 11.
- Mitchell, T.M. (1997), *Machine Learning*, McGraw-Hill, New York.
- Oard, D.W. and Diekema, A.R. (1998), "Cross-language information retrieval", *Annual Review of Information Science and Technology* (ARIST), Vol. 33, pp. 223-56.
- Oard, D.W. and Dorr, B.J. (1996), "A survey of multilingual text retrieval", Institute for Advanced Computer Studies and Computer Science Department, University of Maryland, Technical Report UMIACS-TR-96-19.
- Peters, C. (2003), "Introduction to the CLEF 2003 working notes", Istituto di Scienza e Tecnologie dell'Informazione (ISTI-CNR), Pisa, Italy, available at: http://clef.iei.pi.cnr.it/2003/WN_web/00.2%20-%20intro.pdf (accessed 30 September 2004).
- Picchi, E. and Peters, C. (1998), "Cross-language information retrieval: a system for comparable corpus querying", in Grefenstette, G. (Ed.), *Cross-Language Information Retrieval*, Kluwer Academic Publishers, pp. 81-92.
- Pirkola, A. and Järvelin, K. (2001), "Employing the resolution power of search keys", *Journal of the American Society for Information Science and Technology*, Vol. 52 No. 7, pp. 575-83.
- Pirkola, A., Hedlund, T., Keskustalo, H. and Järvelin, K. (2001), "Dictionary-based cross-language information retrieval: problems, methods, and research findings", *Information Retrieval*, Vol. 4 No. 3/4, pp. 209-30.

- Pirkola, A., Keskustalo, H., Leppänen, E., Käsälä, A.-P. and Järvelin, K. (2002a), “Targeted s-gram matching: a novel n-gram matching technique for cross- and monolingual word form variants”, *Information Research*, Vol. 7 No. 2, available at: <http://InformationR.net/ir/7-2/paper126.html> (accessed 30 September 2004).
- Pirkola, A., Leppänen, E. and Järvelin, K. (2002b), “The RATF formula (Kwok's formula): exploiting average term frequency in cross-language retrieval”, *Information Research*, Vol. 7 No. 2, available at: <http://InformationR.net/ir/7-2/paper127.html> (accessed 30 September 2004).
- Pirkola, A., Toivonen, J., Keskustalo, H., Visala, K. and Järvelin, K. (2003), “Fuzzy translation of cross-lingual spelling variants”, in Callan, J., Hawking, D., Smeaton, A. and Clarke, C. (Eds.), *Proceedings of the 26th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, ACM Press, New York, pp. 345-52.
- Porter, M.F. (1980), “An algorithm for suffix stripping”, *Program*, Vol. 14, pp. 130-7.
- Resnik, P. (1999), “Mining the web for bilingual text”, in *Proceedings of the 37th Annual Meeting of the Association for Computational Linguistics*, pp. 527-34.
- Salton, G. and McGill, M.J. (1983), *Introduction to Modern Information Retrieval*, McGraw-Hill, 1983.
- Sheridan, P. and Ballerini, J.P. (1996), “Experiments in multilingual information retrieval using the SPIDER system”, in Frei, H.-P., Harman, D., Schaübie, P. and Wilkinson, R. (Eds.), *Proceedings of the 19th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, pp. 58-65.

- Singhal, A., Buckley, C. and Mitra, M. (1996), "Pivoted document length normalization", in Frei, H.-P., Harman, D., Schaübie, P. and Wilkinson, R. (Eds.), *Proceedings of the 19th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, pp. 21-9.
- Sormunen, E. (2002), "Liberal relevance criteria of TREC – Counting on negligible documents?", in Järvelin, K., Beaulieu, M., Baeza-Yates, R. and Myaeng, S.H. (Eds.), *Proceedings of the 25th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, pp. 324-30.
- Turtle, H.R. and Croft, W.B. (1992), "A comparison of text retrieval methods", *The Computer Journal*, Vol. 35 No. 3, pp. 279-90.
- Voorhees, E. (2001), "Evaluation by highly relevant documents", in Croft W.B., Harper, D.J., Kraft, D.H. and Zobel, J. (Eds.) *Proceedings of the 24th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, ACM Press, New York, pp. 74-82.

Table 1. Average precision values for eight *slope* values of the pivot method and improvement (%) achieved compared to the average precision of the standard cosine measure. *Slope* equal to 0.450 produced the highest improvement (in bold).

Slope	0.350	0.400	0.450	0.500	0.550	0.600	0.650	0.700
Precision	0.238	0.239	0.241	0.238	0.240	0.239	0.238	0.235
Improvement	37.80	38.40	39.40	38.10	38.90	38.60	38.10	36.10

Table 2. Document pair relatedness distributions of the combined alignment scheme for 100 documents for different threshold values (with the best results printed in bold).

Relatedness	Threshold				
	1.4	1.5	1.6	1.7	1.8
High	17	17	16	15	14
Fair	24	24	24	24	23
Marginal	19	19	21	21	23
Weak	26	26	27	26	26
Unrelated	14	14	12	14	14

Table 3. Document pair relatedness distributions of two evaluators for 100 document pairs.

Relatedness	Evaluator 1	Evaluator 2
High	16	32
Fair	24	21
Marginal	21	18
Weak	27	3
Unrelated	12	26

Figure 1. The recall-precision curves for the bilingual NN searches with the cosine normalization (\blacklozenge), the bilingual NN searches with the pivoted length normalization (\blacksquare), the bilingual InQuery searches (\blacktriangle), and the monolingual NN searches with the pivoted length normalization (\square).

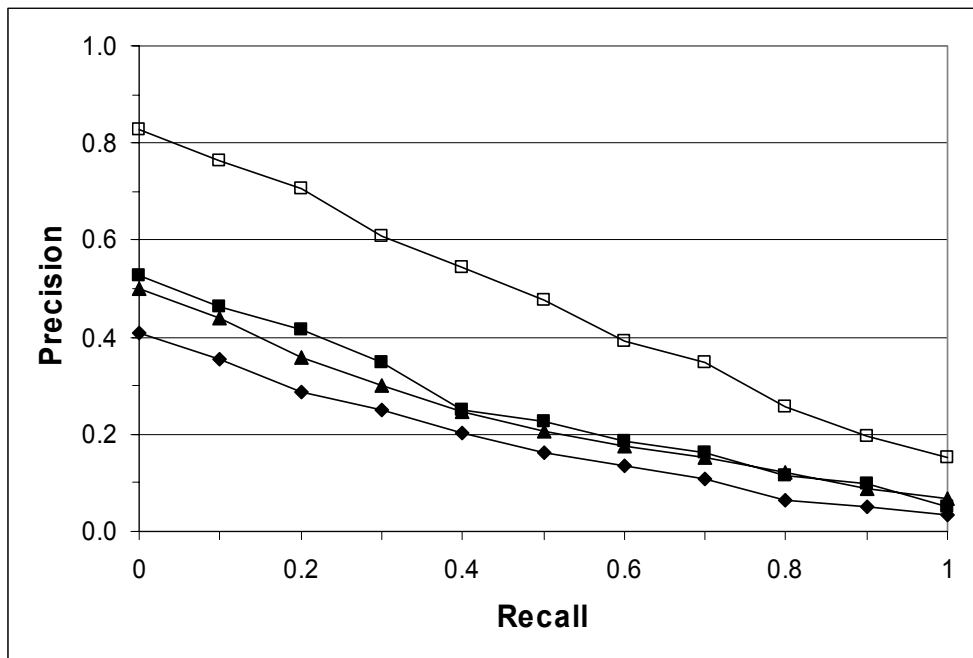


Figure 2. Relevance assessment distributions of 100 document pairs according to four alignment schemes: unrestricted, date restriction with one day (A), date restriction with two days (B), and the combined one.

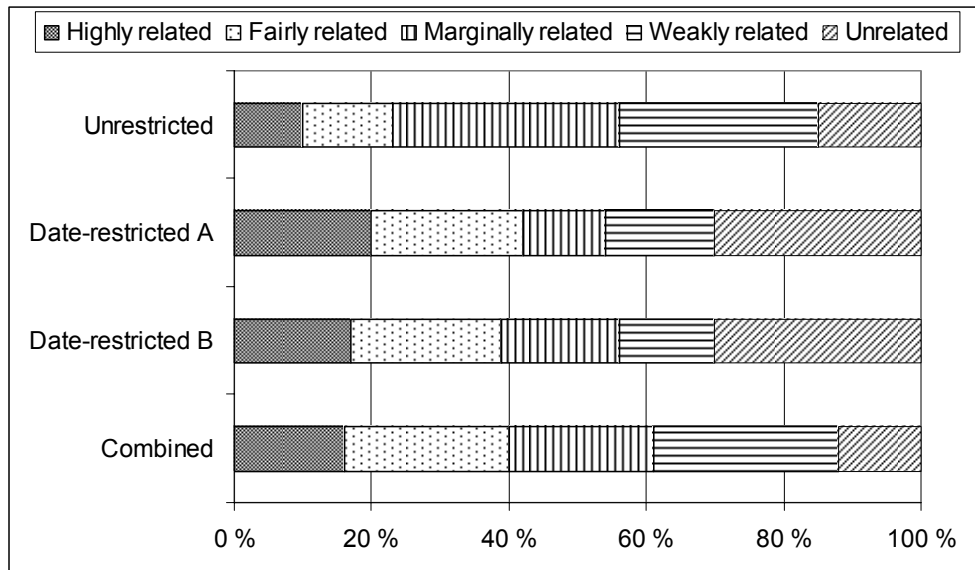


Figure 3. Distribution of the relevance assessments of 200 document pairs evaluated by each evaluator and the distribution of all the 400 relevance assessments.

