UNIVERSITY
OF TAMPERE

This document has been downloaded from
Tampub – The Institutional Repository of University of Tampere

Authors:	Talvensaari Tuomas, Laurikkala Jorma, Järvelin Kalervo, Juhola Martti, Keskustalo Heikki
Name of article:	Creating and exploiting a comparable corpus in cross-language information retrieval
Year of publication:	2007
Name of journal:	ACM Transactions on Information Systems
Volume:	25
Number of issue:	1
ISSN:	1046-8188
Discipline:	Natural sciences / Computer and information sciences
Language:	en
School/Other Unit:	School of Information Sciences

© ACM, 2007. This is the author's version of the work. It is posted here by permission of ACM for your personal use. Not for redistribution. The definitive version was published in ACM Transactions on Information Systems 25 (1), http://doi.acm.org/10.1145/ 1198296.1198300.

URN: http://urn.fi/urn:nbn:uta-3-767
DOI: http://dx.doi.org/10.1145/1198296.1198300

All material supplied via TamPub is protected by copyright and other intellectual property rights, and duplication or sale of all part of any of the repository collections is not permitted, except that material may be duplicated by you for your research use or educational purposes in electronic or print form. You must obtain permission for any other use. Electronic or print copies may not be offered, whether for sale or otherwise to anyone who is not an authorized user.

Preprint from the article:

Talvensaari, T., Laurikkala, J., Järvelin, K. & Juhola M. (2007) "Creating and exploiting a comparable corpus in cross-language information retrieval" to appear in ACM Transactions on Information Systems (ACM TOIS) 25 (2007).

This preprint is made available only for rapid distribution of the results. The final version will be available in the ACM Digital Library after publication in the journal and should be cited as soon as it becomes available.

Creating and exploiting a comparable corpus in cross-language information retrieval

Tuomas Talvensaari[1], Jorma Laurikkala[1], Kalervo Järvelin[2], Martti Juhola[1] and Heikki Keskustalo[2]

University of Tampere

Categories and Subject Descriptors: H.3.1 [**Information Storage and Retrieval**]: Content Analysis and Indexing–*Linguistic processing*; H.3.3 [**Information Storage and Retrieval**]: Information Search and Retrieval–*Query formulation*

General Terms: Algorithms, Languages

Additional Key Words and Phrases: Cross-language information retrieval, comparable corpora, query translation

Abstract *We present a method for creating a comparable text corpus from two document collections in different languages. The collections can be very different in origin: in this study we build a comparable corpus from articles by a Swedish news agency and a U.S. newspaper. The keys with best resolution power were extracted from the documents of one collection, the source collection, by using the relative average term frequency (RATF) value. The keys were translated into the language of the other collection, the*

[1]*Department of Computer Sciences, 33014 University of Tampere, Finland; email: {Tuomas.Talvensaari, Jorma.Laurikkala, Martti.Juhola}@cs.uta.fi*

[2]*Department of Information Studies, 33014 University of Tampere, Finland; email: {Kalervo.Jarvelin, Heikki.Keskustalo}@uta.fi*

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

target collection, with a dictionary-based query translation program. The translated queries were run against the target collection and an alignment pair was made if the retrieved documents matched given date and similarity score criteria. The resulting comparable collection was used as a similarity thesaurus to translate queries along with a dictionary-based translator. The combined approaches outperformed translation schemes where dictionary-based translation or corpus translation was used alone.

1. Introduction

In cross-language information retrieval (CLIR) the aim is to retrieve documents that are written in a language different from the one used for query formulated by the user. Usually the query is translated into the language of the documents. The query language is referred to as the *source language,* and the language of the documents as the *target language.* The two main sources of translation knowledge in CLIR are machine-readable bilingual dictionaries and multilingual corpora [Oard and Diekema 1998]. In dictionary-based cross-language retrieval, the source language query keys are replaced by their target language counterparts in a bilingual dictionary. Using dictionaries alone in CLIR is problematic: some of the translation alternatives of a word may differ from the meaning intended by the user. Their inclusion in the target language query brings ambiguity which in turn damages query performance. Dictionaries are also limited in scope. For example, proper nouns and technical terms are often missing from general purpose dictionaries. For an in-depth look at dictionary-based CLIR methods and problems, see Pirkola et al. [2001].

In corpus-based methods, translation knowledge is derived from multilingual text collections using various statistical methods. Such collections can be aligned or unaligned. In aligned multilingual collections, each source language document is mapped to a target language document. If the paired documents are exact translations of each other, the collection is a *parallel corpus*. *Comparable corpora* consist of document pairs that are not translations of each other, but share similar topics. It can be assumed that terms that are translations of each other – or at least close in meaning – co-occur in these combined or aligned documents. These co-occurrences can be used in a cross-lingual similarity thesaurus (see, for example, Sheridan and Ballerini [1996]), where traditional IR concepts, such as *tf·idf* weighting, are used in a reversed manner: a source language word is thought of as the query, and target language words are retrieved as the answer. The similarity thesaurus can be thought of as a sort of statistical bilingual dictionary. Another approach is to use the aligned collection to do pseudo-relevance feedback cross-lingually, so that instead of the source language documents, the expansion keys are derived from their target language alignment pairs [Braschler and Schäuble 1998]. Document alignments can also be used to disambiguate dictionary-based query translation. Usually this works as follows. A source language query is first translated with a machine-readable dictionary. If multiple translation alternatives occur, the original query is run against the source language documents of the aligned collection. The translation alternatives are then pruned or weighted based on their co-occurrences with the original word in the retrieved alignment pairs. Ballesteros and Croft [1998] and Davis [1998] applied parallel corpora this way.

It is intuitively clear that the more similar the aligned documents are, and the larger the corpus, the more we can rely on the translation knowledge obtained from the corpus. A large parallel corpus would thereby be ideal. However, such collections are hard to come by. United Nations documents [Ballesteros and Croft 1998; Davis 1998] and other official multilingual collections, such as Canadian parliament proceedings [Gale and Church 1991] or EU articles [McNamee and Mayfield 2002] have been used as parallel corpora. Besides their relative scarcity, such collections usually have a quite limited domain, and they cover but a limited number of languages, which makes their use as a source of general translation knowledge problematic. Because of these shortcomings, the creation and use of comparable corpora is often a more feasible option. It is obviously easier to find cross-language text collections with similar topics than it is to find collections that are translations of each other. Comparable corpora have successfully been used as a source of translation knowledge in various studies [Franz et al. 1999; Fung and Yee 1997].

In this paper, we introduce a new way to align two document collections in different languages, and test the effectiveness of several combined CLIR approaches based on comparable corpora, dictionary-based query translation, and pseudo-relevance feedback. The method is outlined in Figure 1. We extract the best keys from the source language documents by means of the relative average term frequency (RATF) developed by Pirkola et al. [2002a]. The keys are translated into the target language using the UTACLIR [Keskustalo et al. 2002] query translation program. The resulting target language queries are run against the target collection with the Lemur retrieval system [Lemur]. An alignment is made if a top-*N*-ranking document – *N* being a relatively small number, for example, 10 or 20 – fits into a given date window and its Lemur similarity score exceeds a given threshold. The method was tested by creating a comparable corpus from a Swedish newswire collection and an American newspaper collection. The collection was used as a similarity thesaurus, and it was applied in translating individual words as well as test topics. In the experiments, combined use of COCOT (our comparable corpus query translation system) and UTACLIR (a dictionary-based query translation system) [Hedlund et al. 2004] clearly outperformed the approaches where the systems were used alone. The COCOT-UTACLIR collaboration also worked better than UTACLIR with pre-translation pseudo-relevance feedback.

This paper is organized as follows. In Section 2, we take a look at previous work done in the automatic creation of aligned corpora. Section 3 introduces the methods and resources used in the study. Section 4 introduces our document alignment method in detail, and Section 5 reports on the experiments and their results. Section 6 provides a brief conclusion and some future directions.

2. Previous work

The automatic creation of comparable corpora has previously been studied by, for example, Sheridan and Ballerini [1996] and Braschler and Schäuble [1998]. Sheridan and Ballerini employed document meta-descriptors and publishing dates to align German and Italian news stories by the Swiss news agency SDA. The SDA documents had fields describing topical content (such as *finance, culture,* or *military*) and the part of the world that the news story handled (for example, *Africa, Germany, United States*). The documents that had matching date and content descriptors were aligned, for example, a German document from 24th August 1994, that dealt with military issues, was aligned with an Italian document from the same day that also had the content descriptor *military*. It is notable that while the SDA stories are not translations of each other, they nevertheless are quite similar, and their alignment is relatively straightforward.

Braschler and Schäuble [1998] also aligned SDA documents, using common proper nouns, numbers and dates, as well as content descriptors. They also aligned English news stories by the news agency AP with the German SDA documents, two collections of quite different origin. They filtered out very common and very rare words from the AP documents, after which they translated the remaining words with a bilingual wordlist that was acquired from "various free sources on the internet". The translated queries were run against the SDA collection. The alignment pair was picked from the top of the rank, after employing date normalization to boost similarity scores of documents that had publication dates near to the source document. Score thresholding was also used to decide whether a pairing should be made.

In a different vein, parallel or comparable documents have also been mined from the web. Resnik [1999] created parallel corpora by detecting structural similarities in multilingual web pages. Typically, when text is presented in many languages in the web, the pages that are translations of each other share a similar structure (headers, paragraphs, hyper links, etc.).

Of the methods mentioned above, Braschler's and Schäuble's method is the most similar to ours, but there are some important differences. Braschler and Schäuble did not use any morphological analysis prior to the source language query formulation, emphasizing that their method did not need expensive linguistic resources. This may work when the source language is a morphologically simple language, such as English. In more complex languages word lemmatization and compound decomposition are needed to gain satisfactory CLIR performance [Pirkola et al. 2001]. In Swedish, for example, compounds are much more abundant than in English. Accordingly, we use the TWOL lemmatizer [Koskenniemi 1983] program to lemmatize inflected source document words, and to decompose compound words.

We use RATF (see Section 3.1) to select source document words to their corresponding queries, while Braschler and Scäuble used raw document frequency, that is, the number of documents in which a word appears in. Comparison of these two techniques is hard; our method may or may not be better than theirs. As mentioned earlier, Braschler and Schäuble used an ad-hoc dictionary to translate the queries. We use UTACLIR, a dictionary-based query translation program (see Section 3.2). UTACLIR uses query structuring to disambiguate translation alternatives and a fuzzy string matching technique to transform words not found in the dictionary, namely proper nouns and technical terms that differ only slightly between languages (for example, Swedish *Gorbatjev* versus English *Gorbachev*) . These techniques clearly improve CLIR results and, likewise, our document alignment method.

Braschler and Schäuble used date normalization in order to find documents that reported the same events as the source document. We examine a small number of top-ranking documents and search for documents that are published near to the source document – the date difference is allowed to be two days in maximum. If no such documents are found, we choose the top-ranking document, provided that its similarity score exceeds a certain threshold. Consequently, the document pairs are not only reports about the same event, but they can be reports about similar incidents that have no apparent relation, for example, a bank robbery in Stockholm reported in the Swedish document collection, paired with a L.A. Times document reporting a similar incident in Los Angeles. These kinds of document pairs, which include similar vocabulary, and thus provide good data for the similarity thesaurus, would be much fewer had we only resorted to date-based alignment.

Our method and Braschler's and Schäuble's method also differ in their applications. Braschler and Schäuble considered the AP-SDA alignments not good enough to be used as a similarity thesaurus, and instead used them for cross-lingual relevance feedback that was more permissive with respect to the quality of the alignments. We show that our alignment method is able to create a comparable collection that can effectively be used as a similarity thesaurus, although the aligned collections are very different in origin.

3. Methods and resources

The TWOL morphological analysis and lemmatization tool, developed by Koskenniemi [1983], was used in normalizing the Swedish source collection and the 5404 English documents of the comparable corpus. Also, the query translation programs UTACLIR and COCOT both use TWOL in pre-processing the input source language words. TWOL transforms inflected words to their base forms and decomposes compound words to their base form constituents.

We used the Lemur language modeling and information retrieval system [Lemur] in aligning the document collections and in the experiment runs. Specifically, we used Lemur's Structured Query Evaluation mode, which applies the InQuery [Allan et al. 1997] query syntax. In our several tests we have used the Lemur system framework and tried it in various modes, including its basic language modelling mode, InQuery mode, and Okapi mode. The InQuery mode has consistently given the best results and this is likely due to the structured query capability (Pirkola method [Pirkola 1998]) enabled by the InQuery mode query language.

3.1. RATF-formula

In order to determine the resolution power of the source document keys (see Section 4.1), we employed the relative average term frequency (RATF), an application of the Kwok [1996] formula developed by Pirkola et al. [2002a].

$$RATF(k) = (cf_k / df_k) \cdot 10^3 / \ln(df_k + SP)^p ,$$

where cf_k is the collection frequency (the number of times the key appears in the collection) of key *k*, and df_k its document frequency (the number of documents in which the key appears in). *SP* and *p* are collection dependent constants, *SP* being a scaling parameter to downweight rare words. To determine values *SP* and *p*, we experimented with the Swedish CLEF 2002 topic descriptions 91-140. The queries were analyzed and normalized with TWOL and the RATF values of the query keys were calculated, based on the index created from the Swedish test collection. A threshold RATF value to filter out keys with low resolution power was chosen and experimented with. *SP* = 1800 and *p* = 3 gave the best results, when the threshold was set to 2.2. These values were used in creating the queries from the source documents.

3.2. UTACLIR

UTACLIR is a dictionary-based query translation and construction method for CLIR [Hedlund et al. 2004]. It is capable of performing query translations between several source and target language pairs, using external resources such as morphological lemmatizers, dictionaries, and stop word lists. It utilizes unified principles for processing basic words, compound words, proper names, phrases, and structuring of the target language queries by the Pirkola method [Pirkola 1998].

The UTACLIR version used in our studies utilizes a GlobalDix Swedish-English dictionary of 36 000 Swedish entries. The dictionary is quite limited, missing, for example, proper nouns. The TWOL lemmatizer is applied to normalize source language words to their dictionary forms, and a Swedish stop word list is used to prune bad query words.

For example, the Swedish phrase ”Jordanien bekräftade att Al-Qaida låg bakom raketattacken mot amerikanska fartyg” (Jordan confirmed that Al-Qaida was behind the rocket attack against American crafts) is translated by UTACLIR as follows:

```
#sum( #syn(  jordanian  jordan)  #syn(  confirm  verify
corroborate)  #syn(  aida  @alidad)  #syn(  sit  lie  law  team
principle)  #syn(  behind)  #syn( missile skyrocket rocket fit
attack)  #syn(  against  towards  v)  #syn(  american)  #syn(
ship  craft  vessel) )
```

First, the inflected source language words are transformed to their dictionary forms and compounds are split by TWOL. Also, stop words, such as the conjunction word *att*, are removed. The normalized source language words are then looked up in the dictionary. A source language word is replaced by all of its translation alternatives in the dictionary.

UTACLIR uses the syntax of the retrieval system InQuery in structuring the target language queries. The translation alternatives of a word are bound together with a SYN operator, which treats its constituent words as synonyms. This type of query structuring reduces the ambiguity caused by multiple translation alternatives, as shown by Pirkola [1998]. All keys within the SYN operator are treated as instances of one key, thus the SYN operator influences the calculation of *tf·idf* values [Rajashekar and Croft 1995]. The probability for operands connected by the SYN operator is calculated by modifying the *tf·idf* function as follows:

$$0.4 + 0.6 \cdot \left(\frac{\sum_{i \in S} tf_{ij}}{\sum_{i \in S} tf_{ij} + 0.5 + 1.5 \cdot \frac{dl_j}{adl}} \right) \cdot \left(\frac{\log\left(\frac{N + 0.5}{df_S} \right)}{\log(N + 1.0)} \right),$$

where tf_{ij} = the frequency of the key i in the document j,

S = a set of search keys within the SYN operator,

dl_j = the length of the document j (as the number of keys),

adl = average document length in the collection,

N = collection size (as the number of documents), and

df_S = the number of documents containing at least one key of the set S.

UTACLIR transmutes the out-of-dictionary words by an effective n-gram matching technique, called the *classified s-gram matching technique* [Pirkola et al. 2002b]. In s-grams (or skip-grams) the n-grams are formed from continuous as well as non-continuous character sequences to better model cross-language word form variation. Skipping is classified into classes by the number of characters skipped (0, 1, 2, ..., *m-n* skipped characters), where *m* is word length and *n* gram length. For digrams, we use a character combination index (CCI) to indicate the number of skipped characters as s-digrams are formed. Table 1 shows the s-digrams with CCI = 0, 1, 2 for the spelling variant pair *pharmacology* and *farmakologian* (the Finnish correspondent for *pharmacology* in a genitive form).

When two words are compared for similarity, their s-gram sets are compared by the DICE formula for each CCI class (Keskustalo et al. [2003]). In the CLIR experiments by

Pirkola et al. [2002b] and Keskustalo et al. [2003], the s-gram techniques outperformed conventional n-gram matching techniques and other conventional string matching techniques.

Since the words *Jordanien* and *Al-qaida* are not found in UTACLIR's dictionary, they are s-gram matched against an English word list. The word list has been created from an English document corpus by using TWOL to lemmatize inflected word forms. The list includes correctly lemmatized, as well as unrecognized word forms that TWOL leaves untouched. The two best matches are returned for each input word. The technique works excellently for the word *Jordanien* (the words *Jordanian* and *Jordan* are returned), but less so for *Al-qaida* (the words *aida* and an unrecognized word *alidad* are returned). This is due to the fact that the word list used in s-gram matching is created from a corpus that predates the arrival of *Al-qaida* to our vocabulary. Obviously, it would have been best for UTACLIR to leave the word unchanged, since *Al-qaida* is spelled identically in English and Swedish.

3.3. COCOT, the comparable corpus query translation program

To obtain translation knowledge from the comparable corpus, we built COCOT, a comparable corpus query translation program. COCOT uses the corpus as a cross-lingual similarity thesaurus, which implies calculating similarity scores between a source language word and the words in the target documents. Like UTACLIR, COCOT can pre-process the input source language words with TWOL.

The similarity thesaurus' similarity score can be calculated by using traditional IR weighting approaches, reversing the roles of documents and words. For a document d_j in which a word t_i appears, the COCOT system calculates the weight w_{ij} as follows:

$$w_{ij} = \begin{cases} 0, & , if\ tf_{ij} = 0 \\ \left(0.5 + 0.5 \cdot \dfrac{tf_{ij}}{Maxtf_j}\right) \cdot \ln\left(\dfrac{NT}{dl_j}\right) & , otherwise \end{cases},$$

where tf_{ij} is the frequency of word t_i in document d_j, $Maxtf_j$ the largest tf in document d_j, dl_j the length of document d_j, or more precisely, the number of unique words in the document. NT can be the number of unique words in the collection, or an approximation of it. The COCOT's similarity score between words t_i and t_k is

$$sim(s_i, t_j) = \frac{\sum_{k=1}^{n} w_{ik} \cdot w_{jk}}{\sqrt{\sum_{k=1}^{n} w_{ik}^2} \cdot \left((1 - slope) + slope \cdot \dfrac{\sqrt{\sum_{k=1}^{n} w_{jk}^2}}{pivot} \right)},$$

where s_i is a word in a source language document, and t_j is a word in a target language alignment pair. The formula employs the pivoted vector length normalization scheme, introduced by Singhal et al. [1996]. The *pivot* value is defined as the mean of the term vector lengths, and *slope* is a constant between 0 and 1 (we chose 0.2). It should be noted that pivoted normalization produces similarity scores that are not in [0, 1]. This makes the use of similarity score thresholding slightly more difficult, since the magnitude of the scores can vary significantly between different collections. It should also be noted that only the target language term vector is normalized with the pivoted normalization scheme. The source language term vector is normalized with the standard cosine normalization. This affects the magnitude of the similarity scores but not, however, the rank of the target language terms. The pivoted scheme was applied because we noticed that the standard cosine normalization penalizes words with long feature vectors (that is, words with high document frequencies) too harshly (see Talvensaari et al. [2006] for elaboration).

Table 2 depicts the results of COCOT similarity calculations for various Swedish words. The score was most successful with nouns, for example *barn* (meaning *child*), *rysk* (*Russian*) and *bil* (*car*) are translated correctly. A high similarity score indicates high confidence in the translation, hence the low scores and bad translations for common and rather vague terms such as *draga* (*draw*) and *information*. It should be noted that not only are the exact translations interesting; many of the top-ranking words are semantically linked to the source language word (for example, *driver* and *vehicle* for *bil* and *Russia, Moscow* and *Yeltsin* for *rysk*), and would thus make good expansion keys.

When COCOT is used to translate queries, a word cut-off value (WCV) and a score threshold is chosen. WCV determines how many of the top ranking target language words are returned per source language word. Score threshold determines the minimum similarity score required for a word to be returned. For example, if WCV is set to three, and score threshold to 10, the words *Russian, Russia* and *Moscow* would be returned for the source language word *rysk* (Table 2). For the word *barn*, only *child* would be returned, because the other top-three ranking words have similarity scores below the threshold. Similar to UTACLIR, COCOT uses the InQuery syntax in structuring its output. All the words returned for a single source language word are tied with InQuery's SYN operator to reduce ambiguity brought by erroneous translations.

3.4. The RATF-based pseudo-relevance feedback method

Query expansion, especially pre-translation expansion, has been proven to be beneficial in CLIR [McNamee and Mayfield 2002]. In pre-translation expansion the source language query is first expanded and then translated. We employed a pre-translation pseudo-relevance feedback technique, developed by Lehtokangas et al. [2006], that uses the RATF formula to pick out good expansion keys. In relevance feedback, the user examines the top ranking documents and chooses the ones that are relevant with respect to the query. The original query is then expanded with words extracted from the relevant documents. In pseudo-relevance feedback (PRF), the top ranking documents are assumed to be relevant, and the whole process is done automatically, without the involvement of the user. The RATF-based PRF technique is described next.

First, we make the feedback run against a source language document collection and extract words from the top N_d ranking documents. The words are lemmatized and compounds are split by TWOL. Stop words are also removed. The remaining words are ranked according to their RATF values, and the top N_w words are chosen to represent each of the N_d documents. Then the remaining unique words are ranked according to their document frequency in the N_d documents ($1 \leq df_i \leq N_d$ for every word w_i), and the top N_r words are chosen as expansion keys. Words that appear in only one of the documents (for whom $df_i = 1$) are not chosen as expansion keys, even if they are in the top N_r. In our experiments, we used parameters $N_w = 20$, $N_d = 100$ and $N_r = 30$. The feedback runs were made against the Swedish TT collection (the same that we used in the document alignments), and InQuery was used as the retrieval engine.

4. Document alignment

The Swedish document collection consisted of 142819 news articles by the Swedish news agency TT (Tidningarnas Telegrambyrå), published in 1994 and 1995 (Table 3). Of these, the 72260 documents published in 1994 were chosen as the source documents of the comparable collection. The target collection consisted of 113005 articles by the newspaper Los Angeles Times, all published in 1994.

Apart from the geographical distance between the two collections, the difference in their original function makes it harder to find good alignment pairs from them. The TT collection comprises mostly of short newswire reports, which means that a news event may be reported many times during a day. The first report is typically a short “breaking news” notice, while later many separate reports may appear as the events evolve and more details emerge. A separate document may also bring contextual and historical information about an event. A newspaper is usually published at most once a day, which means that a newspaper article is typically longer than a newswire report, and it encompasses all the information that a news agency may publish during a day about a single event. Thus, it can be expected that a bijective mapping between the two collections is not possible; that is, it is not possible to find a unique alignment pair for every source document. This also means that in a reversed situation – if we were to search alignment pairs for newspaper articles in a newswire collection – it might be wise to search multiple pairs for a single source document.

In order to extract the best query keys from the source documents, the Swedish collection was lemmatized with TWOL. Also at this stage, bad index keys were filtered out by using a Swedish stoplist of 499 words. After word form normalization and stop word filtering, words appearing only once in the collection were filtered out, as well as words appearing in more than 30000 documents. The procedure was similar to the index building procedure proposed by Salton and McGill [1983]. The resulting index consisted of 208768 keys, each document containing in average 114 unique unstopped keys.

4.1. Source language query formulation

First, one source language query was formed from each source document. The best query keys were extracted from each source document as follows. The keys of a source document were sorted according to their frequency in the document, highest frequency first. Keys with equal frequencies were sorted by their RATF values (with RATF parameters SP = 1800 and p = 3, see Section 3.1). The keys with RATF values lower than the threshold 2.2 were filtered out. The top 30 keys of the resulting list were included in the query to represent the document. In average, 24 % of each document's unique, unstopped keys were included in the queries and later translated with UTACLIR.

The chosen number of query keys may seem large at first. In monolingual information retrieval, even two or three good keys have been proven to be enough for satisfactory retrieval performance [Pirkola and Järvelin 2001]. However, in our particular setting there are many possible reasons for a good source language key not to make it to the target language query. For example, if the vocabulary of TWOL lacks the source word, the inflected forms of the word are left untouched. In such a case, different forms of the word do not increase its frequency, but instead compete with each other. Furthermore, ambiguity brought by lemmatization or dictionary-based translation may incur errors and compound word decomposition may generate extraneous keys that can override good ones. In the translation phase, the dictionary of the translation program might not include the source language word. Therefore using multiple keys as topical evidence in searching for alignments is effective.

4.2. Finding the alignment pair

Each of the 72260 source language queries were translated by UTACLIR, after which the queries were run against the target collection with Lemur's structured query evaluation mode.

In creating the alignments, three different indicators of similarity between the source document and the target collection documents retrieved by Lemur were used: the publishing date of the documents, the similarity score calculated by Lemur between the query and the target document, and the rank of the target document. In short, if the top-ranking document of the Lemur run had a high similarity score and the same publishing date as the source document, it most likely dealt with a similar topic or the same event as the source document. The top *r* in which we searched for the alignment pair was quickly reduced to a relatively small number – we ended up in 20. It was also quickly observed that a matching date does not necessarily mean that the document would be a good alignment pair. Some source documents simply do not align well, as they may deal with a strictly local topic or otherwise there are no matching topics in the collections. Usually, a low similarity score indicated that the document would not create a good alignment, and, accordingly, we employed a score threshold to eliminate such pairings. We also applied query length normalization, since shorter queries get higher similarity scores. The normalization was done by multiplying the score with the logarithm of the number of keys in the query. Also, different date windows were experimented with. Finally, a combination of the three indicators was chosen as the document alignment scheme.

Three different document score thresholds ($\theta_1 < \theta_2 < \theta_3$) are applied to find the alignment pair. The thresholds are not absolute score values, but percentile ranks. For example, a percentile rank of 60 means that the score is greater than or equal to 60 % of all the scores in the alignment runs (there are $n \cdot r$ scores, n being the number of source documents). The use of percentiles enables the method to be used with different matching algorithms that have different ranges of scores. The alignment scheme works as follows. First, a document with exactly the same date as the source document is searched for in the top r of the Lemur rank. If such a document is found and its score is of higher percentile rank than θ_1 it is chosen as the pair. If not, a document published one day later or earlier is searched for. If the pair still is not found, the date difference is increased to two and the threshold is increased to θ_2. On the fourth round (date difference three) the threshold θ_3 is used. After this, if the alignment pair still remains unfound, the highest ranking document is chosen as the alignment pair in case its score exceeds θ_3. Otherwise, no alignment is made. The score threshold is relatively low in the beginning, but it increases with the date difference, as it becomes less probable to find a topically similar document. The actual thresholds were chosen after experimenting with a sample set of source documents (see below).

4.3. Assessing the alignment techniques

It was our aim to create a mapping between source and target collections that would combine source documents with similar documents in the target collection. Since the two collections were so different in origin and function, it was not reasonable to expect that all source documents would find a satisfactory counterpart in the target collection. Most of the articles in the Swedish source collection handled with local or national topics that would not be addressed in the U.S. target collection. This is often the situation in practice, as well. We tried to find a mapping technique – a document alignment scheme – that would create as many good quality alignments as possible. Hence, there were two criteria in assessing different alignment schemes: the number of alignments created and the quality of the alignments, i.e. how close topically the aligned documents were.

In experimenting with different alignment schemes, we could not rely on traditional IR test collections, which implies using test queries and relevance assessments to calculate recall and precision values. We did not know in advance, which target collection documents were relevant (i.e. shared the same topic or at least some vocabulary) in relation to each source document. Making such relevance assessments for even a fraction of the source documents would have been a huge task. Therefore, we randomly picked 500 source documents and manually assessed the quality of their alignments with a five-step relevance scale.

The relevance scale used in assessing the alignments was adapted from Braschler and Schäuble [1998]. The five levels of relevance are

1. **Same story.** The two documents deal with the same event.
2. **Related story.** The two documents deal with the same event or topic from a slightly different viewpoint. Alternatively, the other document may concern the same event or topic, but the topic is only a part of a broader story or the article is comprised of multiple stories.
3. **Shared aspect.** The documents deal with related events. They may share locations or persons.
4. **Common terminology.** The events or topics are not directly related, but the documents share a considerable amount of terminology.
5. **Unrelated.** The similarities between the documents are slight or non-existent.

Different alignment schemes were tested and assessed using the sample set. Table 4 shows similarity class distributions and the number of alignments created for three alignment schemes. In the first scheme, the score thresholds are at a relatively low level, particularly for date distances 0 and 1, for which there actually is no score threshold (θ_1= 0). This brings in a lot of pairs of similarity classes 4 and 5. In the second scheme, the thresholds are at a higher level, thus a lot of the lower quality pairs are removed, yet nearly all of the pairs of classes 1 and 2 remain. In the third scheme, query length normalization is used, while the thresholds are at the same level as in scheme 2. The query length normalization dramatically reduces the number of bad pairings. This seems to be caused by the fact that very short documents, for example, lottery results or very short sports results, transform into short queries, which in turn get high Lemur scores. However, there usually is no good alignment pair for such documents in the target collection, and hence they make bad alignment pairs if no query length normalization is used. The third scheme of Table 4 was used in creating the comparable corpus.

The chosen thresholds might not be directly applicable in different collections. If the two collections are more similar than in these experiments, lower thresholds could bring in more good alignment pairs. This implies that a similar threshold tuning with a sample set would be necessary every time a comparable corpus is created. We do not consider this necessarily a serious problem, since a comparable collection created with our method could be a long-lasting CLIR resource, and the extra work would thus be small in proportion. However, the “threshold space” could be limited by defining a few threshold levels with which to experiment. Additionally, the quality assessment of the sample set need not be very formal or tightly scrutinized, just a quick “eye-balling” through the sample pairs would do. All in all, it could be estimated that one day’s work by a single assessor would be enough to decide the threshold levels used for creating a comparable corpus. We note that the method presented here (that is, without the threshold levels) was successfully used in another study [Talvensaari et al. 2006] to create a Finnish-Swedish comparable corpus.

The total of 72260 source documents were processed and aligned in the manner described above. The resulting comparable collection consisted of 13142 document alignments, 13142 source documents mapped to 5404 different target collection documents. Thus, about 18 % of the source documents found an alignment pair, compared to the 19 % of the initial tests (97 alignments from 500 source documents).

The relatively low number of unique target documents was partly expected, due to the difference of the two collections (see above). However, the number could be increased with some kind of “collision-handling”. For example, if two source documents were competing for the same target document, the winner could be decided by comparing the dates or the similarity scores, or both. The “loser” would then be tried to re-align with some other target document. This would mean fewer alignments, since not all “losers” can be expected to find a new alignment pair – but more unique target documents, and thus, more lexical coverage in the target language. Alternatively, we could align several target documents with a single source document. In this scheme, collision-handling would not be necessary, since all the documents from the top *r* that fulfilled the alignment criteria could be aligned with the source document in question.

5. Test runs and results

We used CLEF topics of the 2001, 2002, and 2003 campaigns as the topic set, and the Los Angeles Times CLEF collection as the test collection. The documents that were part of the comparable collection were removed from the database and the recall base. Originally, there were 118 topics that had at least one relevant document in the test collection. After removing the COCOT documents from the recall base, we also removed topics that had only one relevant document in the test collection, to gain more reliable results. We were left with 91 topics (see Table 5), of which we used the title and the description part of the topics (see Figure 2). Stop words and redundant phrases (such as *find documents discussing* in the example topic) were removed before further processing.

A total of 7 different CLIR approaches were applied to the test collection. Moreover, monolingual queries were made to establish baseline performance. In the monolingual runs, the English queries were first lemmatized with TWOL, since the target database had a lemmatized index. In the CLIR runs, UTACLIR and COCOT were combined and used separately, affecting the subsets of source query words translated by each system. In the experiments, COCOT's WCV value was set to 5 and score threshold to 15, except in the COCOT-alone runs (CC), where the values of the parameters were 2 and 9, respectively. The threshold was decreased in order to gain more lexical coverage – even at the expense of confidence in the translation – since we had to depend solely on COCOT. The WCV was lowered, on the other hand, because of the lower translation confidence. In the COCOT-UTACLIR (CC-UC) run the query words were first translated with COCOT. The words that were not found in COCOT's index, or whose translation confidence dropped below the threshold, were then translated with UTACLIR. In the UTACLIR-COCOT (UC-CC) run, queries were first translated with UTACLIR, after which words that were not in UTACLIR's dictionary were translated with COCOT. In this run, UTACLIR's s-gram feature was turned off, since COCOT functioned in the same role as s-gram matching, translating out-of-dictionary words. The translation machines were also used in parallel by translating queries both with UTACLIR and COCOT, and including both programs' output in their entirety in the target query (UC+CC). In the PRF+UC run, the Swedish queries were expanded with the RATF-based pseudo-relevance feedback, and then translated with UTACLIR. The PRF+CC-UC run is CC-UC with pre-translation PRF.

Table 6 gives three indications of performance: 1) the non-interpolated average precision over all relevant documents, 2) precision at 10 retrieved documents, 3) and R-precision, the average precision after R retrieved documents, R being the number of relevant documents for a query. Figure 3 shows the standard p-r curves for the monolingual run, the UC, PRF+UC, CC, and CC-UC runs. Only one of the COCOT-UTACLIR combinations (CC-UC) is presented in the figure for the sake of clarity. All of the combined approaches performed quite evenly, so their curves would have piled up and cluttered the figure. The results indicate that combined approaches work best in CLIR. The different combinations of UTACLIR and COCOT outperform the approaches where the systems are used alone. Pre-translation PRF boosts UTACLIR's performance, but, surprisingly, clearly impairs the performance of the CC-UC combination. The differences in the non-interpolated average precision were statistically assessed using the Friedman test [Conover 1999]. COCOT and the PRF+CC-UC combination were excluded from the tests, because they were clearly the worst methods. As expected, the monolingual baseline was significantly ($p < 0.001$) better than the CLIR methods. The significance of the differences between UC and PRF+UC, UC and CC-UC, and UC-CC were 0.06, 0.08, and 0.10, respectively. Although these differences approached significance on level α = 0.05, they suggest that the combining of CLIR methods is beneficial. No significant differences were found between the combinations.

Figures 4 and 5 depict the query-by-query performance of PRF+UC and CC-UC in average precision, compared to that of UC. Each bar represents a single query, and the average precision of the UC run is the zero-level. The figures are quite similar; both translation schemes perform worse than UC in some queries, but on the whole, there are more improved queries. Moreover, the difference in performance seems to be larger in the improved queries than in the queries where CC-UC and PRF+UC performed worse.

5.1. Detailed analysis

What are the reasons for the better performance of the combined approaches over pure dictionary-based translation (or more precisely, dictionary-based-translation with approximate cognate matching)? One would assume that ability to translate out-of-vocabulary (OOV) words would be one of the reasons. As mentioned earlier, the translation dictionary of the UTACLIR version used in the experiments lacks proper nouns, and it can not be expected that s-gram matching could translate all of the OOV words correctly. In many of the test topics, however, proper nouns are essential topical words, and failure to translate them would seriously hurt query performance. This would suggest that UTACLIR equipped with a larger dictionary could bring a major improvement in results. However, a larger dictionary usually means both more source language entries and more translation alternatives per entry. In CLIR, the former is arguably preferable, since extraneous translation alternatives bring noise to the queries.

Another reason for the success of the combined approaches might be good expansion keys – linked either semantically or by real-world association – brought in by COCOT (see Section 3.3). In the UC-CC run, there were 19 queries that performed significantly better (> 5 % absolute difference in average precision) than in the UC run. We define this set of queries as UC-CC's "improvement set" $I_{UC\text{-}CC}$. As mentioned, COCOT was only used to translate words that were out-of-vocabulary for UTACLIR in the UC-CC run. Thus, it can be assumed that the queries of $I_{UC\text{-}CC}$ performed better, because COCOT managed to translate OOV words. In the CC-UC run, on the other hand, there were 29 queries that performed significantly better than UC. Of these, 14 queries were also part of UC-CC improvement set. The set $I_{CC\text{-}UC}$ – $I_{UC\text{-}CC}$ has thus 15 queries, which can be assumed to have some other reason for improvement than OOV word translation. Table 7 shows the sizes of the sets I_M and I_M - $I_{UC\text{-}CC}$ for five translation methods *M*. Each improvement set I_M consists of queries where method *M* performed better than UC. The figures indicate that COCOT boosts dictionary-based translation not only because it translates some OOV words. Presumably, the improvement also stems from its ability to bring semantically linked expansion keys into the query. The relatively large size of the set I_{PRF+UC} - $I_{UC\text{-}CC}$ affirms that analysis based on improvement sets is realistic: it seems obvious that the boost brought by PRF is not due to translation of OOV words (after all, the dictionary is the same as in UC), but to good source language expansion keys.

The poor performance of PRF+CC-UC surprised us, and we can only speculate about the reasons for it. Perhaps the reason lies behind the fact that similarity thesaurus translation translates queries word-by-word; that is, it does not try to capture the semantics of the whole query. Thus, when the query has lots of keys – we added 30 keys to the original query in our PRF experiments – each of the keys pulls the translation to different directions. As the number of source language keys increases, the number of bad translations increases also. Perhaps there comes a saturation point when the bad keys outweigh the good ones, and the semantics of the original query are lost. Probably a different approach to comparable corpus translation, for example, cross-lingual PRF (see Section 1), would work better in this kind of approach.

6. Conclusions and future work

We propose a method for creating a multilingual comparable collection from two monolingual document collections. The source language collection is morphologically analyzed and the best content descriptors are extracted from each source document to be used as query keys by using the relative average term frequency (RATF) formula. The resulting queries are translated with a dictionary-based query translation program, and the translated queries are run against the target collection. The source documents are aligned with target documents by using date restriction and similarity score thresholds. All source documents are not aligned, because for some of them satisfactory counterparts do not exist in the target collection. It is notable that we knew in advance only that the source collection consisted of news stories from the same time period as the target collection. The topics were unknown, and no content meta-descriptors were used in the alignment process, unless the publication date is considered as such. The method can thus be used with collections that are less marked up, that is, no separate descriptors of content are needed. These features support comparable corpus alignment in practical environments where the corpora only partially match each other and may be of different types.

We created automatically a Swedish-English comparable collection with this method, aligning 13142 Swedish documents with 5404 different English documents. The collection was used as a cross-lingual similarity thesaurus. The translation was quite successful, especially with terms that have good resolution power, such as nouns. However, the current system is inadequate, when used alone in query translation. Clearly, there should be more documents to bring more statistical evidence to the translation. This could be achieved, for example, by mining comparable documents from the WWW. However, even the current comparable collection can be used effectively in combination with other query translation approaches, as shown in our study. The translation approach based on the similarity thesaurus requires a relatively noise-free corpus – that is, the aligned documents should be highly similar. Cross-lingual PRF, as used by Braschler and Schäuble [1998], among others, is probably more permissive in this respect. In the future, it might be interesting to compare the performance of different approaches to comparable corpus translation. Furthermore, experimenting with different languages, especially non-Indo-European ones would be a challenging task.

The alignment method could also be improved in various ways. More unique target documents could be obtained by utilizing collision-handling in the alignment process. The score threshold tuning could also be made easier by defining a few threshold levels to experiment with.

7. Acknowledgements

This study was funded in part by Tampere Graduate School in Information Science and Engineering (TISE) and by Academy of Finland under the grant number 204978.

ENGTWOL (Morphological Transducer Lexicon Description of English): Copyright (c) 1989-1992 Atro Voutilainen and Juha Heikkilä.

SWETWOL (Morphological Transducer Lexicon Description of Swedish): Copyright (c) 1998 Fred Karlsson and Lingsoft, Inc.

TWOL-R (Run-time Two-Level Program): Copyright (c) 1983-1992 Kimmo Koskenniemi and Lingsoft Oy.

MOT Dictionary Software was used for automatic word-by-word translations. Copyright (c) 1998 Kielikone Oy, Finland.

The InQuery search engine was provided by the Center for Intelligent Information Retrieval at the University of Massachusetts.

References

ALLAN, J., CALLAN, J., CROFT, B., BALLESTEROS, L., BROGLIO, J., XU, J., and SHU, H. 1997. INQUERY at TREC 5. In *The Fifth Text Retrieval Conference (TREC-5)* (Gaithesburg, MD). E. M. Voorhees, D. K. Harman, Eds. NIST Spec. pub. 500-238. National Institute of Standards and Technology, Gaithesburg, MD, 119-132.

BALLESTEROS, L. and CROFT, W.B. 1998. Resolving ambiguity for cross-language retrieval. In *Proceedings of the 21st Annual International ACM SIGIR Conference on Research and Development in Information Retrieval* (Melbourne, Australia), W. B. Croft, A. Moffat, C. J. van Rijsbergen, R. Wilkinson, J. Zobelpp, Eds. ACM Press, New York, NY, 64-71.

BRASCHLER, M. and SCHÄUBLE, P. 1998. Multilingual information retrieval based on document alignment techniques. In *Proceedings of the 2nd European Conference on Research and Advanced Technology for Digital Libraries* (Heraklion, Greece), C. Nikolaou, C. Stephanidis, Eds. Springer-Verlag, Berlin-Heidelberg, 183-197.

DAVIS, M.W. 1998. On the effective use of large parallel corpora in cross-language text retrieval, in G. Grefenstette, Ed. *Cross-Language Information Retrieval*, Kluwer Academic Publishers, 11-22.

FRANZ, M., MCCARLEY, J.S. and ROUKOS, S. 1999. Ad hoc and multilingual information retrieval at IBM. In *The 7th Text Retrieval Conference (TREC-7).* (Gaithesburg, MD). E. M. Voorhees, D. K. Harman, Eds. NIST Spec. pub. 500-242. National Institute of Standards and Technology, Gaithesburg, MD, 157-168.

FUNG, P. and YEE, L. Y. 1998. An IR approach for translating new words from nonparallel, comparable texts. In *Proceedings of the 17th International Conference on Computational Linguistics and 36th Annual Meeting of the Association for Computational Lingustics* (Montreal, Canada), ACL / Morgan Kaufmann Publishers, San Francisco, CA, 414-420.

GALE, W. A. and CHURCH, K. W. 1991. A program for aligning sentence in bilingual corpora. In *Proceedings of the 29th Annual Meeting of the Association for Computational Linguistics (ACL-91)* (Berkeley, CA), ACL, Morristown, NJ, 177-184.

HEDLUND, T., AIRIO, E., KESKUSTALO, H., LEHTOKANGAS, R., PIRKOLA, A. and JÄRVELIN, K. 2004. Dictionary-based cross-language information retrieval: learning experiences from CLEF 2000-2002. *Information Retrieval*, 7, 1-2, 99-119.

HULL, D.A. 1996. Stemming algorithms: a case study for detailed evaluation. *Journal of the American Society for Information Science*, 47, 1, 70-84.

KESKUSTALO, H., HEDLUND, T. and AIRIO, E. 2002. UTACLIR - general query translation framework for several language pairs. In *Proceedings of the 25th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval* (Tampere, Finland), K. Järvelin, M. Beaulieu, R. Baeza-Yates, S. H. Myaeng, Eds. ACM Press, New York, NY, 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 *Proceedings of the 10th International Symposium* (SPIRE 2003) (Manaus, Brazil), M.A. Nascimento, E.S. de Moura, A.L. Oliveira, Eds. Springer, Lecture Notes in Computer Science 2857, Berlin, 252-265.

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.

KWOK, K.L. 1996. A new method of weighting query terms for ad-hoc retrieval. In *Proceedings of the 19th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, (Zurich, Switzerland), H.P. Frei, D. Harman, P. Schaüble, R. Wilkinson, Eds. ACM Press, New York, NY, 187-195.

LEHTOKANGAS, R., KESKUSTALO, H. and JÄRVELIN, K. 2006. Experiments with dictionary-based CLIR using graded relevance assessments: improving effectiveness by pseudo-relevance feedback. *Information Retrieval,* 10, to appear.

LEMUR. The homepage of the Lemur toolkit for language modeling and information retrieval. http://www.lemurproject.org/.

MCNAMEE, P. and MAYFIELD, J. 2002. Comparing cross-language query expansion techniques by degrading translation resources. In *Proceedings of the 25th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval* (Tampere, Finland), K. Järvelin, M. Beaulieu, R. Baeza-Yates, S. H. Myaeng, Eds. ACM Press, New York, NY, 159-166.

OARD, D.W. and DIEKEMA, A.R. 1998. Cross-language information retrieval, *Annual review of Information Science and Technology* (ARIST), 33, 223-256.

PETERS, C. 2004. What happened in CLEF 2004? Introduction to the working notes. Istituto di Scienza e Tecnologie dell’Informazione (ISTI-CNR), Pisa, Italy.

Available on-line at: http://www.clef-campaign.org/2004/working_notes/WorkingNotes2004/CLEF2004WN%20-%20intro.pdf.

PETT, M.A. 1997. Nonparametric Statistics for Health Care Research: Statistics for Small Samples and Unusual Distributions. SAGE Publications, Thousand Oaks.

PIRKOLA, A. 1998. The effects of query structure and dictionary setups in dictionary-based cross-language information retrieval. In *Proceedings of the 21st Annual International ACM SIGIR Conference on Research and Development in Information Retrieval* (Melbourne, Australia), W. B. Croft, A. Moffat, C. J. van Rijsbergen, R. Wilkinson, J. Zobelpp, Eds. ACM Press, New York, NY, 55-63.

PIRKOLA, A., HEDLUND, T., KESKUSTALO, H. and JÄRVELIN, K. 2001. Dictionary-based cross-language information retrieval: problems, methods, and research findings. *Information Retrieval*, 4, 3/4, 209-230.

PIRKOLA, A. and JÄRVELIN, K. 2001. Employing the resolution power of search keys. *Journal of the American Society for Information Science and Technology*, 52, 7, 575-583.

PIRKOLA, A., KESKUSTALO, H., LEPPÄNEN, E., KÄNSÄLÄ, A.-P. and JÄRVELIN, K. 2002b. Targeted s-gram matching: a novel n-gram matching technique for cross- and monolingual word form variants. *Information Research*, 7, 2. Available on-line at: http://InformationR.net/ir/7-2/paper126.html.

PIRKOLA, A., LEPPÄNEN, E. and JÄRVELIN, K. 2002a. The RATF formula (Kwok's formula): exploiting average term frequency in cross-language retrieval. *Information Research*, 7, 2. Available on-line at: http://InformationR.net/ir/7-2/paper127.html.

RAJASHEKAR, T. B. and CROFT, W. B. 1995. Combining automatic and manual index representations in probabilistic retrieval. *Journal of the American Society for Information Science*, 46, 4, 272–283.

RESNIK, P. 1999. Mining the web for bilingual text. In *Proceedings of the 37th Annual Meeting of the Association for Computational Linguistics* (*ACL '99*) (College Park, MD), ACL / Morgan Kaufmann Publishers, San Francisco, CA, 527-34.

SALTON, G. and MCGILL, M.J. 1983. Introduction to Modern Information Retrieval. McGraw-Hill.

SHERIDAN, P. and BALLERINI, J.P. 1996. Experiments in multilingual information retrieval using the SPIDER system. In *Proceedings of the 19th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, (Zurich, Switzerland), H.P. Frei, D. Harman, P. Schaüble, R. Wilkinson, Eds. ACM Press, New York, NY, 58-65.

SINGHAL, A., BUCKLEY, C., and MITRA, M. 1996. Pivoted document length normalization. In *Proceedings of the 19th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, (Zurich, Switzerland), H.P. Frei, D. Harman, P. Schaüble, R. Wilkinson, Eds. ACM Press, New York, NY, 21-29.

TALVENSAARI, T., LAURIKKALA, J., JÄRVELIN, K. AND JUHOLA, M. 2006. Corpus-based CLIR in retrieval of highly relevant documents. *Journal of the American Society for Information Science* (to appear).

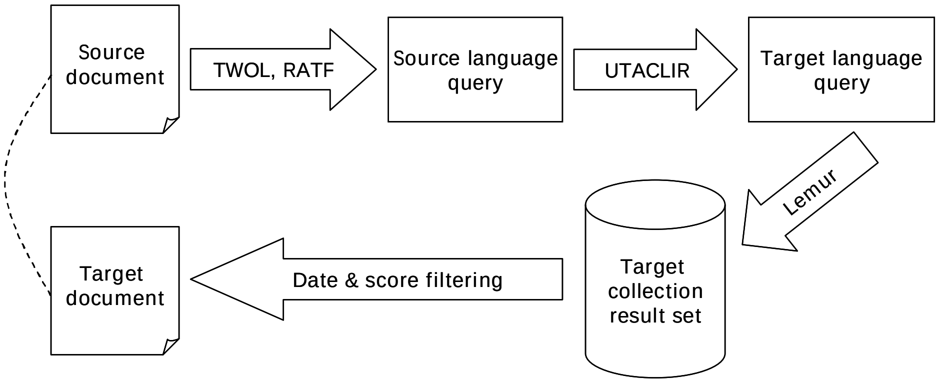

Figure 1. The alignment process.

Table 1. The s-digrams with CCI = 0, 1, 2 for the spelling variant pair *pharmacology* and *farmakologian* (the Finnish correspondent for pharmacology in a genitive form).

Word	CCI	S-digram set
Pharmacology	(0) (1) (2)	{ph,ha,ar,rm,ma,ac,co,ol,lo,og,gy} {pa,hr,am,ra,mc,ao,cl,oo,lg,oy} {pr,hm,aa,rc,mo,al,co,og,ly}
Farmakologian	(0) (1) (2)	{fa,ar,rm,ma,ak,ko,ol,lo,og,gi,ia,an} {fr,am,ra,mk,ao,kl,oo,lg,oi,ga,in} {fm,aa,rk,mo,al,ko,og,li,oa,gn}

Table 2. Term similarity calculations for five Swedish words. The correct translations are shown in bold.

	barn		bil		rysk		information		draga	
1	child	12.51	car	17.10	russian	22.14	find	3.75	support	4.65
2	find	7.42	driver	10.17	russia	19.09	kill	3.71	peace	4.35
3	family	7.40	vehicle	10.02	moscow	17.47	service	3.64	clear	4.20
4	life	6.57	kill	9.45	yeltsin	15.18	send	3.52	talk	4.09
5	woman	6.42	drive	9.21	soviet	13.96	spokesman	3.48	clinton	3.90
6	live	6.33	police	9.17	boris	13.01	large	3.47	control	3.88
7	year-old	6.32	motor	8.82	russ	11.54	center	3.38	war	3.87
8	found	6.29	auto	8.75	military	9.98	life	3.34	area	3.86
9	mother	6.25	truck	7.96	kremlin	9.72	military	3.31	secretary	3.84
10	kill	6.16	hour	7.93	republic	9.22	associate	3.29	organization	3.77

Table 3. The document collections used in the study (Peters, 2004)

Collection	Number of documents	Time span	Median length of documents (tokens)
L.A. Times	113005	Jan 1994 – Dec 1994	421
TT	142819	Jan 1994 – Dec 1995	183

Table 4. Alignment quality distributions for three alignment schemes. A sample of 500 Swedish documents was aligned with the English collection.

	θ_1 = 0, θ_2 = 42, θ_3 = 99		θ_1 = 75, θ_2 = 94, θ_3 = 95		θ_1 = 75, θ_2 = 94, θ_3 = 95, query length normalization	
	N	%	*N*	%	*N*	%
Class 1	22	8	22	20	21	22
Class 2	23	9	19	17	20	21
Class 3	52	20	35	31	33	34
Class 4	75	28	21	19	19	20
Class 5	92	35	15	13	4	4
	264	100	120	100	97	100

Table 5. Recall base statistics. The documents of the comparable collection were removed from the recall base.

Number of topics	91
Number of relevant documents for all topics	1392
Min number of relevant documents per topic	2
Max number of relevant documents per topic	106
Median number of relevant documents per topic	7
Average number of relevant documents per topic	15.3

```
<top>
<num> C145 </num>
<EN-title> Japanese Rice Imports </EN-title>
<EN-desc> Find documents discussing reasons for and consequences of the first
imported rice in Japan. </EN-desc>
<EN-narr> In 1994, Japan decided to open the national rice market for the first time to
other countries. Relevant documents will comment on this question. The discussion can
include the names of the countries from which the rice is imported, the types of rice,
and the controversy that this decision prompted in Japan. </EN-narr>
</top>
```

Figure 2. A sample topic.

Table 6. Mean average precision, precision at recall-point 10 and R-precision for the monolingual baseline and 7 CLIR approaches.

	Monolingual	UC	PRF+UC	CC	CC-UC	PRF+CC-UC	UC-CC	UC+CC
Average precision	0.394	0.219	0.252	0.208	0.272	0.198	0.257	0.265
Precision at 10 docs	0.348	0.221	0.237	0.151	0.251	0.199	0.225	0.250
R-Precision	0.366	0.215	0.239	0.191	0.250	0.197	0.251	0.255

Figure 3. The standard p-r curves for the monolingual baseline and three CLIR approaches.

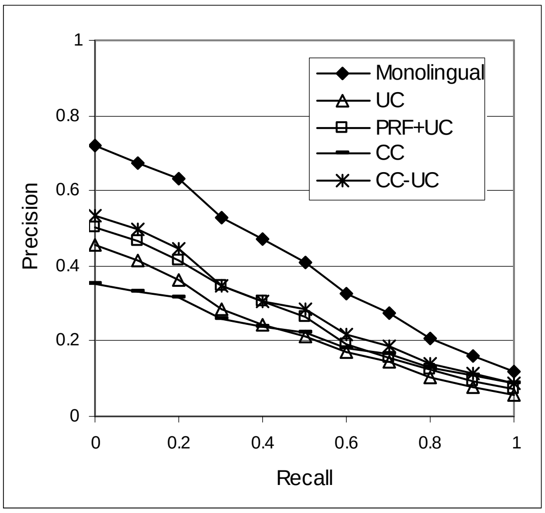

Figure 4. PRF+UC compared to UC query-by-query.

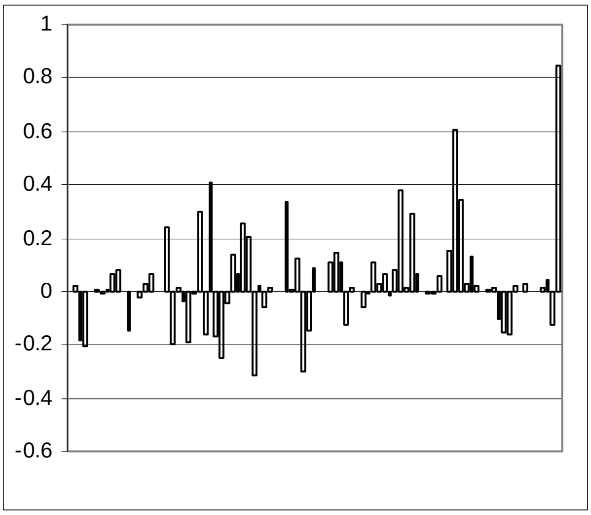

Figure 5. CC-UC compared to UC query-by-query.

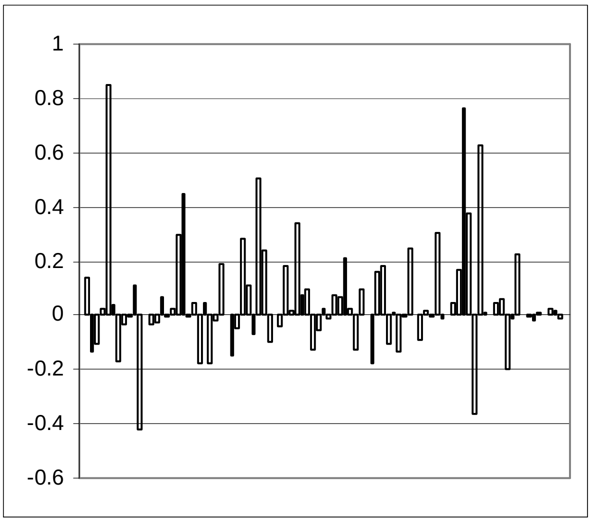

Table 7. Sizes of the improvement sets.

M	$\|I_M\|$	$\|I_M - I_{UC\text{-}CC}\|$
UC-CC	19	0
PRF+UC	28	21
CC	21	7
CC-UC	29	15
CC+UC	24	11