

Corpus-based CLIR in retrieval of highly relevant documents

Tuomas Talvensaari, Martti Juhola and Jorma Laurikkala

Department of Computer Sciences, University of Tampere, Kanslerinrinne 1, FIN-33014 University of Tampere, Finland. E-mail: {Tuomas.Talvensaari, Martti.Juhola, Jorma.Laurikkala}@cs.uta.fi

Kalervo Järvelin

Department of Information Studies, University of Tampere, Kanslerinrinne 1, FIN-33014 University of Tampere, Finland. E-mail: Kalervo.Jarvelin@uta.fi

***Abstract.** IR systems' ability to retrieve highly relevant documents has become more and more important in the age of extremely large collections, such as the WWW. Our aim was to find out how corpus-based CLIR manages in retrieving highly relevant documents. We created a Finnish-Swedish comparable corpus and used it as a source of knowledge for query translation. Finnish test queries were translated into Swedish and run against a Swedish test collection. Graded relevance assessments were used in evaluating the results and three relevance criterion levels – liberal, regular, and stringent – were applied. The runs were also evaluated with generalized recall and precision, which weight the retrieved documents according to their relevance level. The performance of our Comparable Corpus Translation system (Cocot) was compared to that of a dictionary-based query translation program; the two translation methods were also combined. The results indicate that corpus-based CLIR performs particularly well with highly relevant documents. In average precision, Cocot even matched the monolingual baseline on the highest rele-*

vance level. The performance of the different query translation methods was further analyzed by finding out reasons for poor rankings of highly relevant documents.

1. Introduction

In cross-language information retrieval (CLIR), the aim is to retrieve documents that are written in different language than the query formulated by the user. The query language is referred to as the *source language*, and the language of the documents as the *target language*. The language barrier can be crossed either by translating the query to the target language or by translating the documents to the source language. Obviously, queries are easier to translate, because they are typically short and can usually be translated as “bag-of-words”, whereas document translations have to obey the complex rules of natural language. We concentrate on query translation. For an overview of different query translation methods, see 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 desired by the user. Their inclusion in the target language query brings ambiguity which in turn damages query performance. Also, dictionaries are limited in scope. Proper nouns and technical terms are often missing from general purpose dictionaries (Pirkola, Hedlund, Keskustalo & Järvelin, 2001).

Corpus-based CLIR methods are based on multilingual text collections, from which translation knowledge is derived 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. *Parallel corpora* consist of document pairs that are exact translations of each other. *Comparable corpora* are made of document pairs that are not translations of each other, but share similar topics. It can be assumed that words that are translations of each other – or at least close in meaning – co-occur in these combined or aligned documents. An aligned collection can thus be used as a source of knowledge for a cross-language *similarity thesaurus*, where word co-occurrence data is used to calculate similarity scores between a source language word and the words in the target language documents. The classic document retrieval approaches can be applied, only this time the roles of documents and terms are reversed. The source language word can be thought of as the query, and instead of documents, target language words are retrieved in response to the query.

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 have been used as parallel corpora (Ballesteros & Croft, 1998; Davis, 1998). Also, official documents from multilingual countries have been used, such as proceedings of the Canadian parliament (Gale & Church, 1991). Because of the scarcity of parallel corpora, there has been a growing interest in building and exploiting comparable corpora. It is obviously easier to find cross-language text collections with

similar topics than to find collections that are translations of each other. Comparable corpora have successfully been used as a source of translation knowledge in various studies (see Franz, McCarley & Roukos, 1999; Fung & Yee, 1997; Braschler, 2004).

Our aim was to find out how corpus-based CLIR – in particular, CLIR based on document-aligned comparable corpora – manages in retrieving highly relevant documents. The fact that an IR system user assesses the relevance of information objects in a multi-leveled manner, according to their worth for a given work task, has become more and more evident in the age of the WWW and other huge document collections. Therefore, it is essential that in evaluating IR systems we reward systems that are able to retrieve documents of high relevance. Use of binary relevance assessments and liberal relevance criteria has been one of the reasons for some researchers to question the validity of the so-called laboratory model of IR (Sormunen, 2002; Kekäläinen & Järvelin, 2002a).

Using graded relevance assessments in laboratory IR research thus seeks to accomplish a more realistic picture of the use of an IR system. We used a four-point relevance scale, presented in Table 1. The scale was introduced by Sormunen (1994) and has subsequently been used by, for example, Kekäläinen and Järvelin (2002b). The scale was used by applying three relevance thresholds, which corresponded to three different relevance criteria: liberal (documents of levels 1-3 were considered relevant), regular (levels 2 and 3 considered relevant) and stringent (only level 1 considered relevant). The evaluation was done separately on each relevance criteria level. Also, we applied the generalized recall and precision measures (Kekäläinen & Järvelin, 2002b) which are directly based on the graded relevance assessments and thus avoid the binarization of the assessments.

We also experimented with the term vector matching strategy of our comparable corpus translation system (Cocot). The cosine coefficient is often used to calculate term vector similarity in cross-language thesauri (Braschler & Schäuble, 1998; Sheridan & Ballerini, 1996). However, it has been shown that the cosine coefficient has some serious limitations when applied to document retrieval (Singhal, Buckley & Mitra, 1996). We found that these limitations also hold true in similarity thesaurus calculation. Accordingly, we applied a term vector matching scheme that is based on the findings of Singhal et al.

We created a comparable corpus by automatically aligning Finnish and Swedish news articles. The corpus was used to translate Finnish test topics into Swedish. The translated queries were run against a different Swedish test collection with InQuery. The runs were evaluated by using graded relevance assessments, and the performance of Cocot was compared to that of a dictionary-based query translation program. Cocot fared considerably better, especially when stringent relevance criteria were applied. Also, we found that the pivoted vector matching scheme brought slight improvement to the performance of Cocot. The results were further analyzed by finding out reasons for poor retrieval rankings of a random sample of highly relevant documents.

2. Creating the comparable corpus

We built a comparable corpus from two independent document collections. The source collection consisted of approximately 50 000 articles by the Finnish newspaper Aamulehti, covering the time period from November 1994 to December 1995. The target collection contained 140 000 newswire articles by the Swedish news agency TT, published

between January 1994 and December 1995. Both collections are part of the Cross-Language Evaluation Forum (CLEF) document collection (Braschler & Peters, 2004). The method for the automatic creation of the comparable corpus described below was first presented (and more thoroughly covered) by Talvensaaari et al. (2004).

In order to extract the best query keys from the source documents, the source collection was first morphologically analyzed with the morphological lemmatization program TWOL, developed by Koskenniemi (1983). TWOL transforms inflected words to their base forms and decomposes compound words into their constituents. Also at this stage, the most frequent words were filtered out by using a stoplist of 773 Finnish 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 every fourth document. The resulting index consisted of about 150 000 words (including unrecognized word forms).

The best query keys – meaning words that best describe the topic of the document – were extracted from each source document by using a combination of document frequency and the Relative Average Term Frequency (RATF, see Pirkola, Leppänen & Järvelin, 2002b), which measures the resolution power of a word. The words of a document were ranked according to their frequency in the document, highest frequency first. Words with RATF values lower than a given threshold were filtered out, after which 22 highest ranking words were chosen to represent the document in the source language query. Each query was translated to the target language by the dictionary-based query translation program

Utaclir (Keskustalo, Hedlund & Airio, 2002). The translated queries were then run against the target collection with InQuery (Callan, Croft & Harding, 1992). The 10 highest ranking documents of the InQuery result were examined to find an alignment pair for the source document. Not all of the source documents were aligned, since most of the source collection documents dealt with local events and topics that were not covered in the target collection. Date-based and score-based filtering was used to find a topically matching counterpart for the source document.

Three different InQuery score thresholds ($\theta_1 < \theta_2 < \theta_3$) were applied to find the alignment pair. First, a document with exactly the same date as the source document was searched for in the top 10 of the InQuery rank. If such a document was found and it had an InQuery score higher than θ_1 it was chosen as the pair. If not, a document published one day later or earlier was searched for. If the pair still was not found, the date difference was increased to two and the threshold was increased to θ_2 . On the fourth round (date difference three) the threshold θ_3 was used. After this, if the alignment pair still remained unfound, the highest ranking document was chosen as the alignment pair if its InQuery score exceeded θ_3 . Otherwise, no alignment was made.

In the alignment scheme, the score threshold increases as it becomes less probable to find a topically similar document. As a last resort, the date-based filtering is abandoned, and only score-based filtering is used. This reflects the fact that two documents need not concern the same event to be topically related.

The alignment method described above produced a comparable collection of 12 045 document pairs, 12 045 Finnish source documents (22 % of all the source documents) paired with 7 422 different target collection documents. The mapping between the collections was thus not bijective. Date-based filtering created 7 381 of the pairs, the rest were made of highest ranking documents.

3. Cocot

Our comparable collection query translation system (Cocot in short) was written in C++ and it uses a Berkeley DB index. The index is created by inputting word frequency data of the source language documents and their target language alignment pairs. In the process, very rare words, appearing in only one document, and very common words, appearing in more than a fourth of the documents are filtered out. ,

3.1. Term similarity score

In the classic vector space model of IR, documents and queries are represented by vectors whose elements represent term weights in the documents. The similarity between documents and queries can then be calculated by, for example, applying the cosine coefficient. In a similarity thesaurus, the similarity is calculated between term vectors, whose features represent document weights for the terms. The cosine coefficient seems to be also popular in similarity thesaurus calculation, for example both Braschler and Schäuble (1998) and Sheridan and Ballerini (1996) use it. The cosine coefficient between term vectors t_i and t_j is calculated as

$$\cos(t_i, t_j) = \frac{\sum_{k=1}^n w_{ik} \cdot w_{jk}}{\sqrt{\sum_{k=1}^n w_{ik}^2} \cdot \sqrt{\sum_{k=1}^n w_{jk}^2}}$$

where w_{ik} is the weight of document d_k for the term t_i – usually some variation of the well-known *tf·idf* formula – and n is the number of documents in the collection. However, Singhal et al. (1996) found the cosine coefficient to have serious limitations in document retrieval. Especially, they proved that the cosine coefficient tends to favor short documents, i.e. documents with few unique terms and, accordingly, few non-zero elements in their document vectors. In the same vein, we found that using cosine coefficient in term similarity calculation favors terms that appear in only a small number of documents, and, thus, have feature vectors with few non-zero weights. Therefore, we applied a pivoted vector length normalization scheme proposed by Singhal et al. (1996),

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

where t_i is a term in a source language document, and t_j is a term in a target language alignment pair. The *pivot* value is defined as the mean of the term vector lengths. 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. Also it should 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.

In the pivoted weighting scheme, vectors shorter than the *pivot* value are penalized more than in the cosine weighting scheme, whereas vectors longer than *pivot* are favored. The smaller the *slope* value, the more heavily the long vectors are weighted. In similarity thesaurus calculation, penalizing short term vectors is justified: terms that have few non-zero document weights appear in only a few documents and are usually poor expansion keys. In our experiments, the *slope* value was set to a relatively low value (0.2).

The individual document weights were adapted from Sheridan and Ballerini (1996), and they were calculated as follows: for a document d_j in which a term t_i appears, the weight w_{ij} is

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

where tf_{ij} is the frequency of term t_i in document d_j , $Maxtf_j$ is the largest tf in document d_j , NT is the number of index terms in the collection and dl_j is the length of document d_j , or more precisely, the number of different terms in the document. The weight w_{ij} is zero, if the term t_i does not appear in the document.

3.2. Example translations

When the similarity score is calculated between a source language word and the words in the target language document pairs, we get a similarity rank. The top ranking words should be semantically near the source language word. In Table 2, the results of similarity calculations for various Swedish words are shown. The translation knowledge was extracted from a Swedish-English comparable corpus, built for an earlier study (Talvensaaari, Laurikkala, Järvelin & Juhola, 2004). The score is most successful with nouns, for

example *barn* (meaning *child*), *rysk* (*Russian*) and *bil* (*car*) are translated correctly. A high similarity score indicates a high confidence in the translation, hence the relatively low scores and bad translations for common and rather vague terms such as *draga* (*draw*) and *information*. It should be noted that the system is not meant to be a general purpose translation machine; it is meant to translate queries. Thus, although *Moscow* is not the correct translation of *rysk*, it is a good expansion key and could easily be put in a translated query.

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, since the other top-three ranking words have similarity scores below the threshold.

4. Test runs

4.1. Test collection

The test collection consisted of 161 336 news articles by two Swedish newspapers (*Göteborgs Posten* and *Helsingborgs Dagblad*), originally published in 1994. The index was normalized with SWETWOL, the Swedish version of TWOL, which transformed the document words to their basic forms. Compound words were split but also included in their entirety. Words unrecognized by SWETWOL (mainly rare proper nouns and typing

errors) were used in the index unchanged. The resulting index consisted of 370 000 unique normalized words and 217 000 unrecognized word forms.

The test topic set included 52 topics, 24 of which were part of the 2000 CLEF campaign, and the remaining 28 topics of the 2001 campaign. An example topic is shown in Figure 1. The test queries were formed of the description part of the topics, which were mostly comprised of only one sentence. The Swedish versions of the topics were used for the monolingual baseline runs, while the Finnish versions were used in the bilingual runs.

4.2. Recall base

A recall base for the 52 test topics had been created in an earlier study by Ahlgren (2004). Five different query construction methods were used for each of the topics. The methods varied in respect to the degree of word normalization and the use of expansion keys. A total of 260 runs ($5 \cdot 52$) were executed against the test collection with InQuery. The document pools were created using a document cut-off value (DCV) of 100, which resulted in a total of 9 848 unique documents in the 52 pools. The documents were assessed by four assessors; each of the 52 pools was assessed by one single assessor. The four-point relevance scale, introduced in Table 1, was used in the assessments.

Of the 9 848 documents, 1 890 were judged at least marginally relevant (Table 3). The number of relevant documents decreases as the relevance level increases. It seems fair to assume that for a given topic, there are more marginally relevant topics than there are highly relevant ones. On all of the relevance levels, there are topics which have no relevant documents. For example, there are eight topics that have do not have a single highly

relevant document in the recall base. However, when the relevance levels are combined, all of the 52 topics have a minimum of two at least marginally relevant documents.

4.3. Conducting the tests

The 52 test topics were run against the test collection with InQuery. Apart from the monolingual baseline, four different query translation methods were tested. In the ‘Cocot-alone’ runs (abbreviated CC) the Cocot system was used by itself. In the Utaclir-Cocot runs (UCCC) Cocot was utilized in tandem with the dictionary-based query translation program Utaclir. Utaclir was also used alone (UC). Also, a Cocot version that uses the old term vector matching strategy without the pivoted vector length normalization was applied (OLDCC) to find out whether the alternative matching strategy brought any improvement in the performance of the system.

In the baseline run, the Swedish versions of the topic descriptions were used. The topic words were normalized with TWOL, which also split compound words. A Swedish stop-list was also applied. The resulting words were tied together with InQuery’s #sum operator. For example, the topic 22 (see Figure 1) was transformed into

```
Q22(MONO) = #sum( landnings @flygplans inträffa plans flyg pågå-  
ende flygplansolycka landningsbana flygning olycka landningsbanan  
bana banan start ).
```

The ‘@’ symbol indicates that TWOL has not recognized the word and has left it unchanged. The unrecognized words are usually rare proper nouns or typing errors. In the example the ‘unknown word’ *flygplans* is included in the query, because it is part of the compound word *flygplansolycka* (*airplane accident*).

The queries for the cross-language runs were formed from the Finnish versions of the test topics. Before translation, the query words were normalized with TWOL. The source language query words that were extracted from topic 22 were

```
kiitorata kiito rata lentotoiminta lento toiminta sattunut sattua
onnettomuustapaus onnettomuus tapaus
```

The Utaclir version used in the UC runs used GlobalDix Finnish-Swedish dictionary and s-gram matching (Pirkola, Keskustalo, Leppänen, Käsälä & Järvelin, 2002) for words not found in the dictionary. S-gram matching is an approximate string matching technique based on a variation of n-grams. The UC query for the example topic was

```
Q22(UC) = #sum(#syn( asfaltsbeläggning start bana) #syn(@kito
@ito) #syn( bana linje) #syn( flykt flygning handling aktivitet
funktionsduglig) #syn( flykt flygning) #syn( handling aktivitet
funktionsduglig) #syn( komma sig råka inträffa) #syn( komma sig
råka inträffa) #syn( olycka fall händelse) #syn( olycka) #syn(
fall händelse) )
```

Utaclir uses the same syntax as InQuery, binding the translation alternatives of a word together with a #syn-operator, which treats its constituent words as synonyms. This kind of query structuring reduces the ambiguity caused by multiple translation alternatives, as shown by Pirkola et al. (2001). For example, the Finnish word *lento* (*flight*) is found in Utaclir's dictionary, and it is replaced by the translation alternatives (*flykt* and *flygning*), bound together with the #syn-operator. The source language word *kiito* (a part of the compound word *kiitorata*, meaning *runway*) is not found in the dictionary and it is translated by s-gram matching. This time the matching is not successful, since it returns two non-words, preceded by the '@' symbol. The failure is understandable, since approximate string matching techniques work best for proper nouns.

In the CC runs, Cocot's WCV value was set to three. The score threshold was set to 3.0, a relatively low level, implicating a high confidence in Cocot's translation abilities. As with Utaclir, the translation alternatives of a word were bound together with InQuery's #syn-operator.

```
Q22(CC) = #sum(#syn( pilot @pilot draken @draken drakenplan
@drakenplan ) #syn( draken @draken pilot @pilot drakenplan
@drakenplan ) #syn( bana @bana lopp @lopp världs @världs )
lentotoiminta #syn( flyg @flyg plan @plan flygplats @flygplats )
#syn( krona @krona kon @kon mil @mil ) #syn( skada @skada olyck
@olycka inträffa @inträffa ) #syn( olycka @olycka skada @skada
inträffa @inträffa ) onnettomuustapaus #syn( olycka @olycka
olycks @olycks omkomma @omkomma ) #syn( vi @vi falla @falla fall
@fall ) )
```

Each word returned by Cocot is also added to the query with the preceding '@' symbol.

This is because the target collection has a normalized index, and the word forms not recognized by the normalizing program (TWOL) are in the index preceded by the aforementioned symbol. For example, the word *escobar* is not in the index, but *@escobar* is. We do not know in advance, which words are in the index for unrecognized words and which are not. Words that are not in Cocot's index or whose translation confidence is below the score threshold are left unchanged (such as *onnettomuustapaus*, a quite rare compound version of a word meaning *accident*).

In the UCCC runs, Utaclir was first used in query translation. This time s-gram matching was not utilized. Instead, the words that were not in Utaclir's dictionary were translated by Cocot. It was hypothesized that Cocot could compensate for the limitations of Utaclir's vocabulary. The GlobalDix dictionary that Utaclir used has 26 000 Finnish entries and is indeed quite limited; missing, for example, proper nouns. Cocot's WCV was again

set to three, but the score threshold was increased to 10.0, since it was thought that Utaclir could translate the words whose Cocot translation confidence would be low.

In the example topic, the only word not translated by Utaclir was *kiito* (see above). The translation confidence of this word fell below the increased threshold, so the word was left unchanged in Q22(UCCC). This seems to be a good decision, since in Q22(CC) the words that replace *kiito* (*pilot*, *draken*, *drakenplan* – Draken is a Swedish fighter jet) are generally speaking not good translations or expansion keys.

4.4. Results

The evaluation of the test runs was first done separately at three relevance threshold levels. The levels corresponded to liberal, regular and stringent relevance criteria. Table 4 shows the average precision values for the monolingual baseline and the different translation methods. The 11-point P-R curves corresponding to the relevance threshold levels are shown in Figures 2, 3 and 4.

Evaluation was also done using generalized recall and precision (Kekäläinen & Järvelin, 2002b), which implies giving weights to documents of different relevance levels. We used weights 3, 2 and 1 for the corresponding relevance levels. This means that a highly relevant document is considered three times as valuable as a marginally relevant document. The traditional recall and precision can be thought of as a special case of the generalized scheme where all documents of at least marginal relevance are given the weight 1. Using the generalized scheme is advantageous in a graded relevance research setting, since the number of relevant documents on an individual relevance level can be low, as is

the case with highly relevant documents in the recall base that we used. Collapsing the graded relevance assessments into binary ones can lead to distorted results, as the evaluation is based on the rankings of but a few documents. When generalized recall and precision is used, the evidence behind the results is gained from the whole recall base, while the ability to retrieve highly relevant documents is also rewarded. Table 5 shows the non-interpolated generalized average precisions for the different methods, while Figure 5 presents the corresponding 11-point P-R curve.

The average precision values shown in Tables 4 and 5 were analyzed with the within-subjects repeated-measures ANOVA and the two-tailed paired t test, which are parametric tests assessing differences in means, were applied so that the statistical analysis would be more consistent with the results shown in the average precision versus recall figures. Since the distributions were found positively skewed, a power transformation (Pett, 1997) was performed before the parametric analysis by taking the square root of the precision values. The Kolmogorov-Smirnov one-sample test (Pett, 1997) showed the transformed variables approximately normally distributed. The homogeneity of variances was verified with the Levene test. The Friedman test and the two-tailed Wilcoxon signed ranks test were performed to assure that the results obtained with the parametric tests were sound. These tests are more robust than their parametric counterparts, but they compare medians instead of the means, which are traditionally used to construct the precision versus recall figures.

The multivariate significance testing results (Wilks' Λ) of the within-subjects repeated-measures ANOVA indicated that there was a significant difference ($p < 0.004$) in the means of the transformed average precision values obtained with the five methods using the different relevance thresholds and generalized precision. Similarly, the Friedman test results obtained with the original variables showed the precisions significantly ($p < 0.005$) different. The significant results of both the tests facilitated the pair-wise testing of the transformed and the original variables with the paired t test and the Wilcoxon signed ranks test, respectively.

Tables 4 and 5 show the significant differences on the original significance levels as well as on the significance levels adjusted to guard against the elevated potential of Type I error (null hypothesis is rejected when it is true) in the pair-wise testing. We applied Kounias inequality (Hochberg & Tamane, 1987) to adjust for the multiplicity effect, because the traditional Bonferroni correction (Hochberg and Tamane, 1987; Pett 1997) is due to its independence assumption very conservative when the number of comparisons (k) is large. The mean of the Pearson's correlation coefficients were used to estimate the degree of pair-wise dependency between the variables. Since the mean correlations \bar{r} were close to each other (0.50 - 0.56), Kounias inequality ($k = 10$)

$$\alpha' \geq \frac{\alpha}{k - (k-1)\bar{r}}$$

adjusted the respective original significance levels $\alpha = 0.05$ and $\alpha = 0.01$ to $\alpha' = 0.01$ and $\alpha' = 0.002$ with the generalized as well as all the binarized measures.

When the results for the binarized evaluations are examined, it seems that the higher the relevance threshold, the lower the average precision. Only the CC method seems to be an exception to this rule, as its average precision values are quite consistent on all relevance levels. The CC method appears to perform particularly well at the highest relevance threshold level, having a clearly higher average precision than the other translation methods.

Unsurprisingly, the monolingual baseline runs fared better than the bilingual runs, the exception being the CC method at the highest relevance level. However, significant differences were rare between the different translation methods. The findings of the Wilcoxon signed were similar to the t test results. Only the significances altered slightly between some of the methods. At the stringent and regular relevance criteria levels, the CC and UCCC methods were significantly better than UC. At the liberal criteria level, UCCC was better than UC. There was a major difference in average precision between CC and UCCC at the highest relevance level, but the test findings did not confirm this.

Apparently, the CC fared very well on some individual queries, which lifted the average precision to a deceptively high level. This, in turn, can be attributed to the fact that the number of relevant documents on the highest relevance level was relatively low (see Table 2). When there are just one or two relevant documents for a query, the failure to rank them high has a dramatic effect on the average precision of that query.

The OLDCC method (Cocot without the pivoted vector length normalization) fared surprisingly well against the improved Cocot (CC). Only at the highest relevance level there seems to be a difference between the two methods, but the difference was statistically insignificant.

The generalized measures seem to echo the results of the binarized evaluation. Behind the monolingual runs, the CC method appears to be slightly ahead of the rest, at least on low recall levels. Again, though, the significance tests detected few significant differences between the translation methods.

5. Discussion

The retrieval of the highly relevant documents was analyzed more closely for the CC, UCCC and UC translation methods. For each of these methods, 50 highly relevant documents that had a retrieval rank lower than 50 were picked. Then, by modifying the queries, an attempt was made to ‘rescue’ the documents back to the top 50 of the rank. Each of the picked documents had to have been successfully retrieved by some other translation method or by MONO. In this way, the number of documents whose low ranking was based on abnormal vocabulary (or some other reason that made the document hard to find) could be reduced. At least one document was picked for each query, provided that the query had at least one low ranking document. Otherwise, the documents were picked randomly.

The modification of the queries was done iteratively. On each round one or more words were added to or removed from the query. The overall performance of the query (meas-

ured by average precision) was not allowed to drop because of the modifications. Performance-impairing modifications were cancelled. There was no upper limit for the iterations, but the rescuing of a document was aborted when it became clear that it could not be done without hurting the overall query performance.

Table 6 shows the percentages of different query modification actions for the three translation methods. A single occurrence of word addition or removal means one or more words added or removed on a single round. Dictionary expansion is a special case of word addition, where an exact translation of a source language query word is added to the target language query. This modification type also applies to CC, although it does not use a dictionary strictly speaking. Bad s-grams refer to the removal of unsuccessful s-gram matching products in the UC translation method.

The CC method demanded mostly word removals, which could be partly attributed to the low score threshold that was used. Using a low score threshold produces a lot of expansion keys and thus longer queries. Some of the low confidence translations are poor, and their removal helps the highly relevant documents to raise their rank. The other translation methods required more word additions. This seems to suggest that dictionary-based translation needs co-occurrence-based query expansion to be more successful. Also, a larger dictionary is clearly needed for Utaclir. A fourth of the query modifications for the UC method were dictionary expansions. Especially the lack of proper nouns in the dictionary is a serious flaw. The s-gram matching is not an entirely successful technique, but, on the other hand, it does not seem to hurt the performance of the queries either.

5.1. Corpus-based CLIR finds highly relevant documents?

Lehtokangas et al. (2005) pointed out that dictionary-based CLIR is not particularly effective in retrieval of highly relevant documents. In their studies, the performance of a dictionary-based translation system dropped considerably when stringent relevance criteria were applied. Corpus based translation, on the other hand, seems to manage quite well on all relevance criteria levels. One of the reasons for this seems to be that statistical translation brings semantically linked expansion keys to the queries; thus we can speak of a cross-language similarity thesaurus.

Corpus-based translation has its own problems, though. Its performance is closely related to the domain of the underlying corpus. News articles have a fairly general and broad vocabulary, but using this kind of corpus to translate queries concerning, for example, computer programming would not be successful. However, it should be relatively easy to add new document alignments from different domains to a comparable corpus. Expanding dictionaries is arguably a more complex task.

Also, using only topically related document alignments (in contrast to using parallel documents that are translations of each other) is sometimes rather noisy. Very rare and, on the other hand, very common source language words are usually not translated correctly, and in such cases the target language query ends up having lots of bad expansion keys. This can to a degree be controlled by applying higher score thresholds, and dictionaries can be used to translate the more common words. The ideal solution would be to have a larger corpus, with more closely related document alignments. This, in turn, could be achieved by mining the WWW for comparable documents.

The relatively poor performance of the dictionary-based query translation could be partly attributed to the small dictionary that was used in the experiments. In future studies, our purpose is to use a broader dictionary that also includes proper nouns. However, a bigger dictionary can mean either more source language entries or more translation alternatives per entry – usually it means both. In CLIR, the former is arguably preferable, since extraneous translation alternatives bring noise to the queries.

6. Conclusions

We created a document aligned comparable corpus of Finnish and Swedish news articles. The corpus was utilized in query translation, using methods of the classic vector space model of IR to calculate similarity scores between source and target language words. The performance of the Comparable Corpus Translation System (Cocot) was evaluated with graded relevance assessments to find out how the system managed in retrieving highly relevant documents. The performance of the system was compared to that of a dictionary-based query translation system. Also, we experimented with combinations of the two systems. The results indicate that, contrary to dictionary-based translation, the performance of Cocot does not drop when stringent relevance criteria are applied. This seems to be caused by the query expansion effect brought by statistical translation.

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 204970.

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

FINTWOL (Morphological Description of Finnish): Copyright © Kimmo Koskenniemi and Lingsoft Oy. 1983-1993. SWETWOL (Morphological Transducer Lexicon Description of Swedish): Copyright © 1998 Fred Karlsson and Lingsoft, Inc. TWOL-R (Runtime Two-Level Program): Copyright © Kimmo Koskenniemi and Lingsoft Oy. 1983-1992.

References

- Ahlgren, P. (2004). The effects of indexing strategy-query term combination on retrieval effectiveness in a Swedish full text database. University College of Borås/Göteborg University. Publications from Valfrid, 28.
- Ballesteros, L., & Croft, W.B. (1998). Resolving ambiguity for cross-language retrieval. In W. B. Croft, A. Moffat, C. J. van Rijsbergen, R. Wilkinson, & J. Zobel (Eds.), *Proceedings of the 21st Annual International ACM SIGIR Conference on Research and Development in Information Retrieval* (pp. 64-71). New York: ACM.
- Braschler, M. (2004). Combination approaches for multilingual text retrieval. *Information Retrieval*, 7(1-2), 183-204.
- Braschler, M., & Peters, C. (2004). Cross-Language Evaluation Forum: objectives, results, achievements. *Information Retrieval*, 7(1-2), 7-31.
- Braschler, M., & Schäuble P. (1998). Multilingual information retrieval based on document alignment techniques. In C. Nikolaou, & C. Stephanidis (Eds.), *Proceedings of the 2nd European Conference on Research and Advanced Technology for Digital Libraries (ECDL '98)* (pp. 183-97). London: Springer.
- Callan, J.P, Croft, W.B., & Harding S.M. (1992). The INQUERY retrieval system. In A. M. Tjoa, & I. Ramos (Eds.), *Proceedings of the 3rd International Conference on Database and Expert Systems Applications (DEXA-92)* (pp. 78-83). Vienna: Springer.
- 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* (pp. 11-22). Kluwer Academic Publishers.

- Franz, M., McCarley, J.S., & Roukos, S. (1998). Ad hoc and multilingual information retrieval at IBM. In Proceedings of the 7th Text Retrieval Conference (TREC-7). Retrieved June 21, 2005, from http://trec.nist.gov/pubs/trec7/t7_proceedings.html.
- Fung, P., & 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 Linguistics (COLING-ACL '98) (pp. 414-420). Morristown: ACL.
- Gale, W. A. & Church, K. W. (1991). A program for aligning sentences in bilingual corpora. In J. Pustejovsky, & S. Bergler (Eds.), Proceedings of the 29th Annual Meeting of the Association for Computational Linguistics (ACL-91) (pp. 177-184). ACL.
- Hochberg, Y., & Tamane, A. J. (1987). Multiple Comparison Procedures. New York: Wiley.
- Kekäläinen, J., & Järvelin, K. (2002a). Evaluating information retrieval systems under the challenges of interaction and multi-dimensional dynamic relevance. In H. Bruce, R. Fidel, P. Ingwersen & P. Vakkari (Eds.), Proceedings of the Fourth International Conference on Conceptions of Library and Information Science (CoLIS 4) (pp. 253-270). Libraries Unlimited.
- Kekäläinen, J., & Järvelin, K. (2002b). Using graded relevance assessments in IR evaluation. *Journal of the American Society for Information Science and Technology*, 53(13), 1120-1129.
- Keskustalo, H., Hedlund, T., & Airio, E. (2002). UTACLIR - general query translation framework for several language pairs. In K. Järvelin, M. Beaulieu, R. Baeza-Yates,

& S. H. Myaeng (Eds.), Proceedings of the 25th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval (pp. 448-448). New York: ACM.

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.

Lehtokangas, R., Keskustalo, H., & Järvelin, K. (2005). Dictionary-based CLIR loses highly relevant documents. In D. E. Losada & J. M. Fernández-Luna (Eds.), Proceedings of the 27th European Conference on Information Retrieval (ECIR '05) (pp. 421-432). Heidelberg: Springer.

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

Pett, M.A. (1997). Nonparametric statistics for health care research: Statistics for small samples and unusual distributions. Thousand Oaks: Sage Publications.

Pirkola, A., Hedlund, T., Keskustalo, H., & Järvelin, K. (2001). Dictionary-based cross-language information retrieval: problems, methods, and research findings. Information Retrieval, 4(3), 209-230.

Pirkola, A., Keskustalo, H., Leppänen, E., Käsälä, A.-P., & Järvelin, K. (2002). Targeted s-gram matching: a novel n-gram matching technique for cross- and monolingual word form variants. Information Research, 7(2). Retrieved June 22, 2005, from <http://InformationR.net/ir/7-2/paper126.html>.

Pirkola, A., Leppänen, E., & Järvelin, K. (2002). The RATF formula (Kwok's formula): exploiting average term frequency in cross-language retrieval. Information Re-

search, 7(2). Retrieved June 22, 2005, from <http://InformationR.net/ir/7-2/paper127.html>.

Sheridan, P., & Ballerini, J.P. (1996). Experiments in multilingual information retrieval using the SPIDER system. In H.-P. Frei, D. Harman, P. Schäuble, & R. Wilkinson (Eds.), *Proceedings of the 19th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval* (pp. 58-65). New York: ACM Press.

Singhal, A., Buckley, C., & Mitra, M. (1996). Pivoted document length normalization. In H.-P. Frei, D. Harman, P. Schäuble, & R. Wilkinson (Eds.), *Proceedings of the 19th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval* (pp. 21-29). New York: ACM Press.

Sormunen, E. (1994). *Vapaatekstihaun tehokkuus ja siihen vaikuttavat tekijät sanoma-lehtiaineistoa sisältävässä tekstikannassa*. [The effectiveness of free-text searching in full-text databases containing newspaper articles and abstracts]. Espoo, Finland: Technical Research Centre of Finland, Research Publications 790.

Sormunen, E. (2002). Liberal relevance criteria of TREC – Counting on negligible documents? In K. Järvelin, M. Beaulieu, R. Baeza-Yates, & S. H. Myaeng (Eds.), *Proceedings of the 25th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval* (pp. 324-330). New York: ACM Press.

Talvensaari, T., Laurikkala, J., Järvelin, K., & Juhola, M. (2004). Creating and exploiting a comparable corpus in cross-language information retrieval. Submitted manuscript.

Tables

Table 1. The four point relevance scale.

Table 2. Term similarity calculations for five Swedish words, calculated from a Swedish-English comparable corpus. The correct translation are in bold.

Table 3. Recall base statistics.

Table 4. Average non-interpolated precisions (%) for different query translation methods at three relevance thresholds. Significance of differences (¹ < 5%, ² 5-10%, ³ > 10%) in average both on the original ($\geq = p < 0.05$; $\gg = p < 0.01$) and adjusted ($\geq = p < 0.01$; $\gg = p < 0.002$) significance levels.

Table 5. Generalized average non-interpolated precisions (%) for different query translation methods at three relevance thresholds. Significance of differences (¹ < 5%, ² 5-10%, ³ > 10%) in average both on the original ($\geq = p < 0.05$; $\gg = p < 0.01$) and adjusted ($\geq = p < 0.01$; $\gg = p < 0.002$) significance levels.

Table 6. The number of different query modification measures for three query translation methods.

Figures

Figure 1. CLEF topic 22.

Figure 2. P-R curves for the monolingual baseline and four translation methods for liberal relevance criteria.

Figure 3. P-R curves for the monolingual baseline and four translation methods for regular relevance criteria.

Figure 4. P-R curves for the monolingual baseline and four translation methods for stringent relevance criteria.

Figure 5. Generalized P-R curves for the monolingual baseline and four translation methods.