

Highly Relevant Documents Lost in CLIR: Experiments with Dictionary Translation and Pseudo-Relevance Feedback

Raija Lehtokangas, Heikki Keskustalo, and Kalervo Järvelin

Department of Information Studies, FIN-33014 University of Tampere, Finland
Raija.Lehtokangas@uta.fi, Heikki.Keskustalo@uta.fi, Kalervo.Jarvelin@uta.fi

Abstract. Research on cross-language information retrieval (CLIR) has typically been restricted to settings using binary relevance assessments. In this paper, we present evaluation results for dictionary-based CLIR using graded relevance assessments in a best match retrieval environment. A text database containing newspaper articles and a related set of 35 search topics were used in the tests. First, monolingual baseline queries were automatically formed from the topics. Secondly, source language topics (in English, German, and Swedish) were automatically translated into the target language (Finnish), using structured target queries. Effectiveness of the translated queries was compared to that of the monolingual queries. Thirdly, pseudo-relevance feedback was used to expand the original target queries. CLIR performance was evaluated using three relevance thresholds: stringent, regular, and liberal. When regular or liberal threshold was used, a reasonable performance was achieved. Adopting stringent threshold caused a considerable loss of performance when compared to monolingual Finnish performance, with or without pseudo-relevance feedback.

Keywords: cross-language information retrieval, graded relevance assessments, relevance feedback

1. Introduction

A lot of CLIR research has been carried out during the last years, see, e.g., TREC¹, CLEF², and NTCIR³. The research is, however, mainly based on binary relevance assessments. So there is not sufficient knowledge on how CLIR methods treat documents of various relevance levels. In this paper, we concentrate on this aspect of CLIR performance evaluation. At NTCIR, empirical results with graded relevance assessments have been presented (see, e.g., Zhou et al. 2004, Fujii and Ishikawa 2004), but these results have not been interpreted from the point of view we have in this paper. We compare dictionary-based CLIR performance between different levels of relevance and also analyze failures in retrieving highly relevant documents.

Using binary relevance assessments (documents are either relevant or non-relevant) ignores the fact that documents are to different degrees relevant with respect to search requests - considering a marginally relevant document as valuable as a highly relevant one. This is a real problem since a majority of documents relevant in a database may be only marginally relevant (Sormunen 2002).

Normally, searchers prefer documents with a higher degree of relevance. In the present information overload it is more vital than ever to be able to pick the best documents. So, degrees of relevance should be taken into account when evaluating IR systems and methods, and systems and methods able to retrieve the most valuable documents should be credited for this.

Evaluation of IR methods and systems by various relevance levels has recently become possible for two reasons. First, evaluation methods for handling graded relevance data have been developed (Järvelin and Kekäläinen 2000, Kekäläinen and Järvelin 2002). Secondly, test collections exist that provide graded relevance assessments (Sormunen 2000, Sormunen 2002, Kishida et al. 2004, Lee et al. 2002, Voorhees 2001).

This paper presents novel CLIR results based on graded relevance assessments. Our main research question is how well dictionary-based CLIR is able to find documents relevant to different degrees, in particular highly relevant documents. A four-point relevance scale is used in the tests: documents in the test database are highly, fairly or marginally relevant, or non-relevant. CLIR performance is evaluated by precision and recall under three relevance thresholds: 1) *stringent* (only highly relevant documents are accepted) 2) *regular* (both highly and fairly relevant documents accepted), 3) *liberal* (highly, fairly and marginally relevant documents accepted). Performance is also evaluated by *generalized* precision and recall (Kekäläinen and Järvelin 2002) using varying weighting schemes for documents of different levels of relevance.

Moreover, we experiment with expansion of the translated target queries. Query expansion (QE) means query reformulation by changing its search keys (or their weights) to make it better match relevant documents. QE has been studied extensively because the selection of good search keys is difficult but crucial for good results (Efthimiadis 1996, Kekäläinen 1999). QE may be based on external, collection independent knowledge structures (such as thesauri), collection-dependent knowledge structures (e.g., word co-occurrence statistics) or search results. Relevance feedback (RF) is a method based on search results. In interactive RF the searcher examines retrieved documents and gives the IR system feedback at the level of (ir)relevant documents or at the level of candidate search keys extracted from top ranking documents. Harman (1992) argues that several feedback iterations in retrieval are beneficial. In *pseudo RF* (PRF) the IR system assumes the top ranking documents to contain relevant documents and automatically, without user interaction, extracts QE keys by statistical means.

Ballesteros and Croft (1998) and McNamee and Mayfield (2002) recommend pre- and post-translation PRF in CLIR. The latter point out that the benefits of PRF are marginal if the translation resources are good. Xu and Croft (1996) and Mitra, Singhal and Buckley (1998) argue that queries perform poorly in PRF, when no relevant documents are found among the top ranking documents. In the present paper we examine PRF in enhancing the query based on results of an initial dictionary-based CLIR query. We are particularly interested in whether PRF is capable of reducing query ambiguity due to dictionary translation and thereby enhancing the retrieval of highly relevant documents. We employ the RATF formula by Pirkola and Järvelin (2001) in the extraction of candidate QE keys from top ranking initial results.

We evaluate CLIR performance in a laboratory setting, using a best match retrieval system (InQuery) and a test database consisting of Finnish newspaper articles. CLIR queries, having English, German and Swedish as source languages, are translated into the target language by an automated process using morphological analyzers, machine-readable dictionaries and stopword lists (Hedlund et al. 2001). *n*-Gram techniques are applied to words that are untranslatable by the dictionaries, and the target queries are structured by using the synonym operator of InQuery.

We are able to show the graded relevance assessment performance for dictionary-based CLIR. Likewise we are able to show that CLIR performs on a reasonable level when liberal or regular relevance threshold is used. When stringent threshold is used in evaluating the same queries, a loss of performance is observed. PRF is not capable of straightening this.

The paper is organized as follows: test design is presented in Section 2 and findings in Section 3. In Section 4 findings are further discussed. Section 5 concludes the paper.

2. Test Design

2.1. The Training and Test Collections

Our test database TUTK consists of 53,893 Finnish newspaper articles from three newspapers (Sormunen 2000, Kekäläinen and Järvelin 2002). As Finnish is a highly inflectional language and rich in compounds (words written together as one unit), a lemmatizer was used in index building. Words recognized by the lemmatizer were turned into their lemmas in the index, and in addition to this, compounds were split. Finally, all words not recognized by the lemmatizer were put into the index as such (thus typically in inflected forms). The resulting index contains about 241,000 unique recognized words (or compound components) as lemmas and about 118,000 unique unrecognized word forms. There are 35 test topics, each expressing a search request in 1-4 sentences. The themes of the topics are distributed as follows: person (5 topics), organization (12), geographical place (10), general theme (8). The topics are originally expressed in Finnish, but have been translated by professional translators into English, German and Swedish.

For training the PRF process we used a Finnish CLEF collection consisting of 55,344 documents and related topics from the years 2002-2004, one set consisting of 10 topics, for which graded relevance assessments were available and another of 50, for which binary relevance assessments only were available. This training collection was lemmatized in the same way as the test collection.

2.2. Graded Relevance Assessments

A recall base for the 35 TUTK topic requests has been collected by extensive pooling. With respect to the 35 requests, altogether 17,338 documents have been evaluated by human assessors using a four-point relevance scale. Four relevance judges were employed, and the relevance of 20 requests was assessed by two persons, and the remaining 15 requests by one person. (Sormunen 2000, Järvelin and Kekäläinen 2000).

A four-point scale was used in the relevance assessments. Relevance level 0 is used to denote non-relevant documents not about the subject of the request. Relevance level 1 denotes marginally relevant documents - documents referring to the request but not giving more information than the request itself. Relevance level 2 is used to denote fairly relevant documents - documents that contain some new facts with regard to the request. Finally, relevance level 3 is used to denote highly relevant documents - documents that contain valuable information with regard to the request. (Sormunen 2000)

The relevance assessors agreed in 73 % of the parallel assessments. In 21 % of the cases the difference was one point. In the remaining 6 % of the cases the difference was two or three points. Disagreements in judgments were resolved in the following way: if the difference was one point, the assessment was selected from each judge in turn. If the difference was two or three points, the researcher made the final decision about the relevance level. (Järvelin and Kekäläinen 2000)

As a result of the relevance evaluations for the 35 requests, 444 documents are considered highly relevant (relevance level 3), 829 documents fairly relevant (level 2), and 993 documents marginally relevant (level 1). Thus, the recall base contains 2,266 documents evaluated as relevant for the 35 topics. The rest of the database is considered to contain only non-relevant documents with respect to the topics (relevance level 0).

For training the PRF process, a set of 10 topics was selected from the CLEF 2002-2004 topics and the relevance of the documents previously assessed as relevant (using binary relevance) with respect to these topics was reassessed by the researchers themselves using the same four-point relevance scale as discussed above. This set of 10 topics was selected under the condition of having at least 20 relevant documents, to ensure that the different levels of relevance would be represented among them. Each of the researchers assessed each document, the total number of the assessed documents being 299. In 36 % of the 299 assessments there was a complete agreement between the three assessors, and in 63 % the difference was one point (e.g., 3-3-2). So, agreement between the assessors was high. Afterwards, the assessments were compared and the differences resolved, either by majority (difference being one point) or by discussion (difference being two points or more).

2.3. Resources Used

The retrieval system used in the experiments was InQuery (v. 3.1), a probabilistic retrieval system provided by the Center for Intelligent Information Retrieval at the University of Massachusetts (Broglia et al. 1994).

InQuery queries are either natural language queries (e.g. English sentences) or structured queries. Structured queries are constructed by using, e.g., the operator *syn*, which treats all of its arguments as instances of one search key. All operators are preceded by the hash sign #, and the arguments are delimited by parentheses, e.g. *#syn(ship vessel boat)*. If no operator is given, the operator *sum* is used as default. This treats all of its arguments as having an equal influence on the result.

Large machine-readable dictionaries, provided by Kielikone plc., Finland, were used for the word-by-word translations in the language routes English to Finnish, German to Finnish, and Swedish to Finnish. For normalizing source and target language words, morphological analyzers provided by Lingsoft plc., Finland, were used in the respective languages. Novel stop word lists were designed for the present study. Number of words in the stop word lists are as follows: English (402 words), Finnish (737), German (637), Swedish (658).

2.4. Monolingual Queries

The monolingual queries used as the baseline of the study were formed automatically from the topics by lemmatizing their words and forming InQuery synonym sets (*#syn*) for each of (possibly several) lemmas of each word.⁴ If a word was not recognized by the lemmatizer, approximate string matching was applied to find the most similar strings from the target index. We used skip-

grams (see Pirkola et al. 2002) for selecting the two best matching strings. Finally, stop words were removed.

As an example, after processing the Finnish topic *OPEC:n öljyn hintaa ja tuotantomääriä koskevat päätökset* (*The decisions of OPEC concerning oil prices and production levels*) the following baseline query (in InQuery syntax) was formed:

```
#sum( #syn(opec) #syn(n) #syn(öljy) #syn(hinta) #syn(tuotantomääri) #syn(tuotantomäärä)
#syn(päätös) )
```

In the example above, the words *OPEC*, *öljyn* (inflected word form referring to *oil*), *n* (genitive suffix), *hintaa* (inflected word form referring to *price*), *tuotantomääriä* (inflected form referring to *production volume*) and *päätökset* (inflected form referring to *decision*) are normalized successfully. (Note that the word *tuotantomääriä* generates two normalized word forms, *tuotantomääri* and *tuotantomäärä*.) The remaining query words are stopwords (*ja* meaning *and*, *koskevat* - inflected form referring to *related*). Thus they are removed from the query.

2.5. Source Query Word Types for Translation

The following six source query word types are automatically recognized in the UTACLIR query translation framework and processed accordingly in query translation (Hedlund et al. 2001):

- Stop words: source query words belonging to the source stop lists are omitted first. Also, a target stop word list (Finnish) was used to remove remaining stop words from the translated query in each translation route.
- Recognized translatable words: these source words are recognizable (included in the lexicon of the lemmatizer) and translatable (included in the translation dictionary). They are translated, and the translations are treated as synonyms (connected with InQuery's synonym operator).
- Recognized untranslatable and unsplitable words: these source words are untranslatable and cannot be split by the lemmatizer. Typically, this kind of words include proper names and they occur because of the relatively large lexicon of the lemmatizer. As translation is not possible, approximate matching is performed instead to find the most similar strings from the target index.
- Recognized and untranslatable but splittable words: source words belonging to this type are compounds not included in the translation dictionary as whole words. These words are split and translation is attempted for the components.
- Unrecognized but translatable words. These words are rare, because typically the lemmatizers do recognize translatable words. In case such source words exist, they are translated.
- Unrecognized and untranslatable words: typically these words are proper names, acronyms, scientific terms, rare words or new words of the language. As direct translation is not possible, approximate matching is performed as in the third case above.

2.6. Translated Queries

The translated queries were formed automatically by translating the topics in English, German and Swedish into Finnish.

As an example, after translating the Swedish topic *OPEC:s beslut om priset och produktionsmängderna för olja* (*The decisions of OPEC concerning oil prices and production levels*) the following translated query (in InQuery syntax) is formed:

#sum(#syn(opec roope) #syn(päätöksenteko päätös ratkaisu tuomio) #syn(arpoa arvo hinta kunnia palkinto ylistys) #syn(produktio tuotanto valmistua valmistus) #syn(ainemäärä erä joukko määrä paljous suuruus) #syn(rasvata voidella öljy öljytä))

In the example above, the untranslatable Swedish word *OPEC* is replaced in translation by the first synonym set containing approximate string match results *opec* and *roope*. The source word *beslut* (*decision*) is translatable and is translated by the second synonym set containing the correct dictionary translations (*päätöksenteko*, etc.) for the word. The next word is a stopword (*om* meaning *about*) and is removed. The source word *priset* (inflected form of *pris* meaning *price*) can be normalized and translated, and it is replaced by the third synonym set of the query above. The next source word is a stopword (*och* meaning *and*) and is removed. The next word *produktionsmängderna* is an inflected compound which is untranslatable as a whole. It is automatically split into components (*produktion*, *mängd*) which are individually translated (corresponding fourth and fifth synonym sets in the translated query). Next word is a removable stop word (*för* meaning *for*). Finally, the word *olja* (*oil*) is translated. Compared to the monolingual case, the synonym sets formed by translation typically include several words.

2.7. Pseudo-Relevance Feedback Based Query Expansion

We tested the effect of automatically adding expansion terms into the original translated queries, utilizing pseudo-relevance feedback (PRF). In monolingual PRF original queries are typically short. In our case, on the contrary, original translated queries are *long and noisy*. Therefore, we wanted to investigate whether original queries could be improved by automatically exploring the top retrieved documents and performing a second retrieval round after expansion.

We tested the RATF formula (Pirkola et al. 2002) for term extraction. Originally, the formula was designed to indicate the goodness of query keys. Our results suggest that RATF is also well suited for recognizing the best terms in documents. Its additional advantage is simplicity. The PRF process is described next.

After the initial retrieval, we first collected (inflected) words from the top retrieved documents (thresholds of 10, 20 and 30 documents were used), lemmatized the words and split the compounds. All the words belonging to a short stoplist were removed. Next, the RATF formula was utilized to calculate a goodness value for each remaining word in each document. A fixed number of words was selected as *automatic indexing keys* to represent each document (top 20, 50 or 100 words having the highest value in each document). Finally, as query expansion keys were selected the automatic indexing keys supported by the largest number of documents. As a special case, keywords supported by only one document were never accepted as query expansion keys. The number of expansion keys was 10 or 30.

The expansion keys were added unweighted as the second *#sum* structure, following the first *#sum* part constituting of the original translated query. E.g.,

#sum(
#sum(#syn(opec roope) #syn(päätöksenteko päätös ratkaisu tuomio) #syn(arpoa arvo hinta
kunnia palkinto ylistys) #syn(produktio tuotanto valmistua valmistus) #syn(ainemäärä erä joukko
määrä paljous suuruus) #syn(rasvata voidella öljy öljytä))
#sum(öljy opec tynnyri saudi arabia arabi öljyn saudi-arabia opec-maa kuwait kiintiö irak
öljyntuotanto tuotanto iran hinta dollari öljyministeri järjestö emiiri arabiemiirikunta
öljynviejämaa öljynhintaa viejä tuotantokiintiö raakaöljy öljytä vähentää venezuela tehty))

3. Findings

3.1. Training Runs

The first training runs were performed using the 10 CLEF topics with graded relevance assessments. The effectiveness of the original (unexpanded) baseline queries was compared to that of the expanded queries, using all the 18 combinations of QE alternatives (see 2.7.). On the basis of these results the best combinations were selected for further testing. Further tests were carried out using a larger CLEF topic set (50 topics) with binary relevance assessments. On the basis of these tests, the two best combinations (top 10 documents, top 50 RATF keys, top 30 QE keys; top 20 documents, top 100 RATF keys, top 30 QE keys, i.e., 10_50_30 and 20_100_30) were selected for the final tests in the TUTK collection. These two combinations delivered the best improvements both among the 10 topics for which graded relevance assessments were available and among the larger set with binary assessments.

3.2. Structured Test Runs

The effectiveness results of the monolingual and bilingual structured runs⁵ are presented in Table 1. Effectiveness is studied separately at *stringent* relevance threshold (Rel = 3), *regular* threshold (Rel = 2,3), and *liberal* threshold (Rel = 1,2,3). At the liberal threshold, difference between the baseline monolingual run and the bilingual runs ranges from -11 % to -19 %. As for the regular threshold, the difference between the monolingual and the bilingual runs is slightly greater, ranging from -14 % to -21 %. The results of the stringent threshold are clearly the worst, -21 % to -35 % below the monolingual baseline.

Above effectiveness was evaluated using binary relevance (yet separately for different relevance levels or their combinations). Performance of the runs was also evaluated using generalized precision and recall (Kekäläinen and Järvelin 2002). By this measure effectiveness can, taking the different degrees of relevance into account, be expressed in one single value. Relevance values originally given to the documents can be reweighted, thus allowing experiments with different user scenarios.

Weighting reflects how documents at different levels of relevance are valued in relation to each other (e.g., if highly relevant documents are valued 10 times as much as marginally relevant, the former get the weight 10, the latter 1). If all the relevance levels are given the same weight, we have the normal binary relevance situation (1,1,1). We experimented by giving different weights to the relevance levels, first having the original weights 3, 2 and 1 (3 for highly relevant, 2 for fairly relevant and 1 for marginally relevant documents), then valuing the highly relevant ones more

(weights $10,4,1$, and $100,10,1$). Results using generalized precision and recall are presented in Table 2. The table presents for each language pair the CLIR query effectiveness and the difference to the monolingual baseline. It can be seen that the more the highly relevant documents are weighted in relation to the less relevant ones, the larger is the difference to the baseline. This is in line with what was observed about the lower performance for the highly relevant documents (Table 1).

3.3. Pseudo-Relevance Feedback Based Expansion Runs

We examined the effect of PRF on the effectiveness of the CLIR queries. Based on the training runs, two of the best combinations were selected for the final test runs, i.e. 10_50_30 and 20_100_30 (see 3.1.). We present results for the former combination because it gave similar but slightly better results. In Table 3, effectiveness of the original and the expanded queries is given for all three relevance thresholds. For all the thresholds and language routes, considerable improvements were achieved by PRF, ranging from 6 % to 20 %. For some language routes, the average effectiveness even exceeded that of the monolingual monolingual baseline. Even though improvements using QE were achieved for all the relevance thresholds, performance of the stringent threshold in relation to the other thresholds could not be raised.

4. Discussion

In our experiments, dictionary-based CLIR was performed on three thresholds of relevance: 1) *stringent* (only highly relevant documents are accepted), 2) *regular* (fairly and highly relevant documents accepted), and 3) *liberal* (marginally, fairly and highly relevant documents accepted). It was found that reasonable CLIR performance can be achieved if liberal or regular relevance threshold is used. However, if the stringent threshold is used, as high performance relative to the monolingual baseline cannot be achieved.

A random sample of 76 highly relevant documents ranked low (representing 30 topics) from the Swedish-Finnish run was selected for a further study. Rankings of these documents ranged from 51 to 983. The vocabulary of the documents was studied to find possible reasons why these documents did not match with the queries and were thus not retrieved earlier.

Quite a common reason for a mismatch between a topic and a newspaper article is that the article takes up specific, concrete things whereas the topic expresses the same on a more general level. For example, talking about environmental investments of the forest industry (the exact wording of a topic), articles may mention by name individual paper mills and real measures taken there - without at all telling that these measures are environmental investments or anything like that.

It was also noticed that the right sense may be expressed in the document but by a word not in a right form, e.g. a verb may be used in a document when a noun would be needed. Talking, e.g., about incidence of AIDS, all the studied documents (three) used only verb forms ('sairastavat', 'sairastavan' etc., meaning 'to suffer from a disease') referring to 'disease' whereas there was a noun ('sairaus') in the query. A lemmatized index requires the use of precisely the right part-of-speech in the query, as words representing different parts-of-speech normally get separate entries in the index (here: 'sairaus' and 'sairastaa', respectively). Also, the wording of topics is often quite scarce, so

additional words might be needed in the query. Depending on the situation, these could be in hierarchical, associative or synonymous relationship to the words of the original query.

What was said above implies to modifications in queries. Of the two main components in the retrieval process - query and document - attention is here paid to the former because it is the query that is modifiable in the short run. To find out reasons for low rankings in our document sample, we experimented with modifications of the original target queries and tried to raise the rankings of the late retrieved documents. It was decided that the rankings should fall in the range of 1 to 50 after the modifications. There were 76 documents in the sample, and the ranking of all but three documents could be raised. Only modifications that could be carried out without hurting the overall performance of the query (measured in average precision) were accepted (i.e., the performance of the modified query needed to be higher than that of the original query). Sometimes one measure was enough, sometimes two or three different measures together were needed. For each document, all the measures (or combinations of them) that could be found were listed. These lists are, of course, not exhaustive, but could possibly be supplemented. Altogether, there were 196 occurrences of measures (occurring either separately or with others). In 59 % of the occurrences, a word or more had to be added to the query. In 16 % of the occurrences, the wording of the original topic had to be changed, and in 10 %, the dictionary had failed: either an entry or a translation equivalent was missing. In 8 % of the occurrences, there was a special problem connected with a group of compound words, and in 7 %, there were problems with proper names (either proper names were incorrectly interpreted as common nouns of the source language and translated as such, or the inflected forms brought by the *n*-gram process were not exactly those present in the document).

Above, notable is the large proportion of word additions, over half of all occurrences. In 17 % of word additions, the added word and a word in the original query were words of the same root (e.g., one was a derivative of another, or both were derivatives of the same word). This kind of additions could be produced automatically, on the basis of the original query. However, an overwhelming majority of the words added (83 %) did not have a direct relation to the wording of the original target query. Words of this kind should be picked from external sources, such as RF or PRF. Altogether, it should be noted that word additions in these experiments were done intellectually, knowing the vocabulary of the document in question and trying numerous word combinations. Without prior knowledge of the vocabulary in the documents it would have been, in most cases, impossible to know which words to add.

We tried to enhance our original CLIR results by PRF. Even though improvements by using PRF were achieved for all the relevance thresholds, performance of the stringent threshold in relation to the other thresholds could not be raised by this method. It seems that ambiguity brought by the translation process cannot be resolved by this type of relevance feedback only but interaction with the real user is also needed.

5. Conclusion

In this paper, dictionary-based CLIR was tested in a best match retrieval environment, using graded relevance assessments. A four-point relevance scale was used in the test database, which consists of Finnish newspaper articles. Source language queries in English, German and Swedish were translated by an automated process into the target language, using morphological analyzers, machine-readable dictionaries, stopword lists, *n*-gramming of untranslatable words, and structured queries. The effectiveness of the translated queries was compared to that of the monolingual

queries using *stringent*, *regular* and *liberal* relevance thresholds. Reasonable CLIR performance was achieved for *liberal* or *regular* relevance thresholds. Instead, for the *stringent* threshold, i.e. when only highly relevant documents were accepted, equally high performance could not be achieved. When a sample of highly relevant documents ranked low was studied, reasons for the low rankings of these documents were found.

The performance of the translated queries was successfully raised on all the relevance thresholds by using query expansion based on pseudo-relevance feedback. However, performance of the stringent threshold in relation to the other thresholds could not be raised by this method.

Acknowledgements

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

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

FINTWOL (Morphological Description of Finnish): Copyright (c) Kimmo Koskenniemi and Lingsoft plc. 1983-1993.

GERTWOL (Morphological Transducer Lexicon Description of German): Copyright (c) 1997 Kimmo Koskenniemi and Lingsoft plc.

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

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

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

This research was funded by the Academy of Finland, under Project Numbers 177033 and 1209960.

Notes

1. TREC Homepage. Available: <http://trec.nist.gov/>
2. CLEF Homepage. Available: <http://clef.iei.pi.cnr.it>
3. NTCIR Homepage. Available: <http://research.nii.ac.jp/ntcir/index-en.html>
4. #syn clauses are, of course, not needed for unary arguments. This is however due to using the same UTACLIR process both for monolingual and CLIR queries.
5. In this paper, only results for the structured test runs are presented. The unstructured runs performed clearly worse, and the results for these runs are presented in Lehtokangas et al. 2005.

References

- Ballesteros L and Croft WB (1998) Resolving Ambiguity for Cross-language Retrieval. In: Proceedings of the 21st Annual International ACM SIGIR Conference on Research and Development in Information Retrieval. New York, NY: ACM Press, 64-71.
- Broglio J, Callan J and Croft WB (1994) INQUERY system overview. In: Proceedings of the TIPSTER text program (Phase I). San Francisco, CA: Morgan Kaufmann Publishers.
- Efthimiadis, EN (1996) Query expansion. In: Williams, M.E. (1996), ed. Annual Review of Information Science and Technology, vol. 31 (ARIST 31). Medford, NJ: Learned Information for the American Society for Information Science, 121-187.
- Fujii A and Ishikawa T (2004) Cross-Language IR at University of Tsukuba: Automatic Transliteration for Japanese, English, and Korean. In: Working Notes of NTCIR-4, Tokyo, 2-4 June, 2004. Available: <http://research.nii.ac.jp/ntcir-ws4/NTCIR4-WN/index.html>
- Harman, DK (1992) Relevance feedback revisited. In: Proceedings of the 15th ACM SIGIR Conference on Research and Development in Information Retrieval. New York, ACM Press, 1-10.
- Hedlund T, Keskustalo H, Pirkola A, Sepponen M and Järvelin K (2001) Bilingual tests with Swedish, Finnish and German queries: dealing with morphology, compound words and query structure. In: Peters C, ed. Cross-language information retrieval and evaluation. Proceedings of the CLEF 2000 workshop, Lecture Notes in Computer Science, 2069. Berlin: Springer, 210-223.
- Järvelin K and Kekäläinen J (2000) IR evaluation methods for retrieving highly relevant documents. In: Proceedings of the 23rd Annual International ACM SIGIR Conference on Research and Development in Information Retrieval. New York, NY: ACM Press, 41-48.
- Kekäläinen J (1999) The effects of query complexity, expansion and structure on retrieval performance in probabilistic text retrieval. Tampere, Finland: University of Tampere, Department of Information Studies. Ph.D. Thesis. Acta Universitatis Tamperensis 678. Available: <http://www.info.uta.fi/tutkimus/fire/archive/QCES.pdf>
- Kekäläinen J and Järvelin K (2002) Using Graded Relevance Assessments in IR Evaluation. Journal of the American Society for Information Science and Technology, 53(13): 1120-1129.
- Kishida K, Chen K, Lee S, Kuriyama K, Kando N, Chen HH, Hyon Myaeng S and Eguchi K (2004) Overview of CLIR Task at the Fourth NTCIR Workshop. In: Working Notes of NTCIR-4, Tokyo, 2-4 June, 2004. Available: <http://research.nii.ac.jp/ntcir-ws4/NTCIR4-WN/index.html>
- Lee S, Hyon Myaeng S, Kim H, Seo JH, Lee B and Cho S (2002) Characteristics of the Korean Test Collection for CLIR in NTCIR-3. In: Working Notes of NTCIR-3, Tokyo, October 8-10, 2002. Available: <http://research.nii.ac.jp/ntcir/workshop/OnlineProceedings3/index.html>
- Lehtokangas R, Keskustalo H and Järvelin K (2005) Dictionary-Based CLIR Loses Highly Relevant Documents. In: Losada D and Fernandez-Luna J, eds. Advances in Information Retrieval,

- Proceedings of the 27th European Conference on IR Research, ECIR 2005, Santiago de Compostela, Spain. Lecture Notes in Computer Science, 3408. Berlin: Springer, 421-432.
- 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. New York, NY: ACM Press, 159-166.
- Mitra M, Singhal, A and Buckley C (1998) Improving automatic query expansion. In: Proceedings of the 21st Annual International ACM SIGIR Conference on Research and Development in Information Retrieval. New York, NY: ACM Press, 206-214.
- 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, 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, Keskustalo H, Leppänen E, Käsälä AP and Järvelin K (2002) Targeted *s*-Gram Matching: a Novel *n*-Gram Matching Technique for Cross- and Monolingual Word Form Variants. *Information Research*, 7(2). Available: <http://InformationR.net/ir/7-2/paper126.html>
- Sormunen E (2000) A Method for Measuring Wide Range Performance of Boolean Queries in Full-Text Databases. Tampere, Finland: University of Tampere, Department of Information Studies. Ph.D. Thesis. Available: <http://acta.uta.fi/pdf/951-44-4732-8.pdf>
- Sormunen E (2002) Liberal Relevance Criteria of TREC - Counting on Negligible Documents? In: Proceedings of the 25th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval. New York, NY: ACM Press, 320-330.
- Voorhees E (2001) Evaluation by Highly Relevant Documents. In: Proceedings of the 24th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval. New York, NY: ACM Press, 74-82.
- Xu J and Croft WB (1996) Query expansion using local and global document analysis. In: Proceedings of the 19th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval. New York, NY: ACM Press, 4-11.
- Zhou Y, Qin J, Chau M and Chen H (2004) Experiments on Chinese-English Cross-language Retrieval at NTCIR-4. In: Working Notes of NTCIR-4, Tokyo, 2-4 June, 2004. Available: <http://research.nii.ac.jp/ntcir-ws4/NTCIR4-WN/index.html>

Table 1. Effectiveness of structured target queries at three relevance thresholds (non-interpolated average precision)

STRINGENT	Average precision	Difference	Difference (%)
Language route			
Finnish-Finnish	28.4	-	-
Swedish-Finnish	20.7	-7.7	-27.1
English-Finnish	22.5	-5.9	-20.8
German-Finnish	18.5	-9.9	-34.9
REGULAR	Average precision	Difference	Difference (%)
Language route			
Finnish-Finnish	36.9	-	-
Swedish-Finnish	31.9	-5.0	-13.6
English-Finnish	31.3	-5.6	-15.2
German-Finnish	29.2	-7.7	-20.9
LIBERAL	Average precision	Difference	Difference (%)
Language route			
Finnish-Finnish	37.6	-	-
Swedish-Finnish	33.4	-4.2	-11.2
English-Finnish	32.8	-4.8	-12.8
German-Finnish	30.3	-7.3	-19.4

Table 2. Effectiveness of structured target queries using different weighting schemes for relevance levels (generalized interpolated average precision (GP) over 11 recall points)

Language route	GP (w=1,1,1)	Difference (%)	GP (w=3,2,1)	Difference (%)
Finnish-Finnish	39.5	-	31.5	-
Swedish-Finnish	34.9	-11.6	26.2	-16.8
English-Finnish	34.5	-12.7	26.8	-14.9
German-Finnish	32.5	-17.7	24.5	-22.2
Language route	GP (w=10,4,1)	Difference (%)	GP (w=100,10,1)	Difference (%)
Finnish-Finnish	27.8	-	26.2	-
Swedish-Finnish	21.6	-22.3	18.7	-28.6
English-Finnish	23.2	-16.6	20.7	-21.0
German-Finnish	20.5	-26.3	17.3	-34.0

Table 3. Effectiveness of the original and the PRF expanded target queries

STRINGENT	Original	Expanded	Diff	Diff (%)
Language route		10 50 30		
Swedish-Finnish	20.7	22.9	+ 2.2	+ 10.6
English-Finnish	22.5	26.6	+ 4.1	+ 18.2
German-Finnish	18.5	19.6	+ 1.1	+ 5.9
REGULAR	Original	Expanded	Diff	Diff (%)
Language route		10 50 30		
Swedish-Finnish	31.9	36.6	+ 4.7	+ 14.7
English-Finnish	31.3	37.2	+ 5.9	+ 18.8
German-Finnish	29.2	32.3	+ 3.1	+ 10.6
LIBERAL	Original	Expanded	Diff	Diff (%)
Language route		10 50 30		
Swedish-Finnish	33.4	38.3	+ 4.9	+ 14.7
English-Finnish	32.8	39.2	+ 6.4	+ 19.5
German-Finnish	30.3	34.5	+ 4.2	+ 13.9