

Dictionary-Based CLIR Loses Highly Relevant Documents

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 both structured and unstructured queries. Effectiveness of the translated queries was compared to that of the monolingual queries. CLIR performance was evaluated using three relevance criteria: stringent, regular, and liberal. When regular or liberal criteria were used, a reasonable performance was achieved. Adopting stringent criteria caused a considerable loss of performance, when compared to monolingual Finnish performance.

1 Introduction

A lot of CLIR research has been carried out during the last years, see, e.g., TREC [15], CLEF [3], and NTCIR [10]. 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., [17], [4]), 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 [14]. 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 [6], [7]. Secondly, test collections exist that provide graded relevance assessments [13], [14], [8], [9], [16].

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 under three conditions: 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). Moreover, performance is evaluated by generalized precision and recall [7] using varying weighting schemes for documents of different levels of relevance.

CLIR performance is evaluated 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. *n*-Gram techniques are applied to words that are untranslatable by the dictionaries. Both structured and unstructured target queries are used.

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 criteria are used. When *stringent* criteria are used to evaluate the same queries, a loss of performance is observed.

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 Test Collection

The target database consists of 53,893 Finnish newspaper articles from three newspapers [13] [7]. As Finnish is a highly inflectional language and rich in compounds (words written together as singular units), a morphological analyzer was used in index building. Words recognized by the analyzer were normalized into their basic forms in the index, and in addition to this, compounds were split. Finally, all words not recognized by the analyzer 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) in basic forms 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), organisation (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.

2.2 Graded Relevance Assessments

A recall base has been collected for the 35 topic requests by extensive pooling. With respect to the 35 requests, altogether 17,338 documents have been evaluated by human assessors using a 4-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. [13], [6]

A 4-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. [13]

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. [6]

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).

2.3 Resources Used

The retrieval system used in the experiments was *InQuery*, a probabilistic retrieval system provided by the Center for Intelligent Information Retrieval at the University of Massachusetts [2].

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 normalizing each word into its basic form by using a morphological analyzer and forming an *InQuery* synonym set (*#syn*) from the normalized forms (each word having possibly multiple lemmas). If a word was not recognized by the analyzer, approximate string matching was applied to find the most similar strings from the target index. We used skip-grams (see [12]) 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* (genetive 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 Translated Queries

The translated queries were formed automatically by translating the topics in English, German and Swedish into Finnish. The query translation framework UTACLIR is based on ideas presented originally in [5]. In the present study, the details of the translation process were fine-tuned using training data. The basic idea of the processing is to utilize morphological analysis for normalizing source words into basic forms, split the untranslatable source compounds into components, and utilize machine-readable dictionaries for bilingual word-by-word translations for translatable words. For untranslatable words, approximate string matching is used for finding, with respect to the source word, the most similar words from the target database index. As in the monolingual case, stop words are removed. Stop word lists are applied for both source and target language words during the translation process. [11]

As an example, after translating the Swedish topic *OPEC:s beslut om priset och produktionsmängderna för olja* 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 valmis-
tua 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 the translations typically include several words (see Section 3).

2.6 Source Query Word Types

The following source query word types are automatically recognized and processed accordingly in query translation [5]:

- 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 morphological analyzer) 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 unsplittable words: these source word are untranslatable and cannot be split by the morphological analyzer. Typically, this kind of words include proper names and they occur because of the relatively large lexicon of the morphological analyzer. 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 morphological analyzers 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.

3 Findings

3.1 Structured runs

General properties with respect to the number of words and synonym sets in the (structured) target queries are presented in Table 1. As we can see, the number

Table 1. General properties of the structured queries (monolingual and translated).

Language route	Topics	Words	Synonym sets	Words/Synonym set
Finnish to Finnish	35	479	459	1.04
English to Finnish	35	5390	517	10.4
German to Finnish	35	2479	616	4.02
Swedish to Finnish	35	1959	647	3.03

of synonym sets varies. Also, in some language routes, the average number of words in a synonym set is larger. On the average, English as a source language produced the largest synonym sets.

The effectiveness results of the monolingual and bilingual structured runs are presented in Table 2, separately for highly relevant (*stringent* relevance criteria accepting relevance level 3 - $Rel = 3$ in Table 2), fairly and highly relevant (*regular* criteria: $Rel = 2,3$), and marginally, fairly and highly relevant (*liberal* criteria: $Rel = 1,2,3$) documents. When all the levels are studied together, difference between the baseline monolingual run and the bilingual runs ranges from -11 % to -19 %. As for the levels 2 and 3, the difference between the monolingual and the bilingual runs is slightly greater, ranging from -14 % to -21 %. The results of the relevance level 3 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 [7]. 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

Table 2. Effectiveness of structured target queries at three relevance thresholds (non-interpolated average precision).

Language route, Rel = 3	Average precision	Difference	Difference (%)
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
Language route, Rel = 2,3	Average precision	Difference	Difference (%)
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
Language route, Rel = 1,2,3	Average precision	Difference	Difference (%)
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

as much as marginally relevant, the former get the weight 10, the latter 1). If all the relevance levels is given the same weight, we have the normal binary relevance situation.

Results using generalized precision and recall are presented in Table 3. 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). The table presents also the binary relevance situation where all the levels are weighted equally (1,1,1), and for each language pair 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 bigger is the difference to the baseline. This is in line with what was observed about the lower performance for the highly relevant documents (Table 2).

3.2 Unstructured runs

The results of the unstructured translated queries are presented in Table 4. (The only distinction between the structured and unstructured runs is the non-use of synonym sets in the target queries in the latter.) On the whole, the performance level of the unstructured queries is lower, measured both in absolute figures and in relation to the the monolingual baseline. However, differences between the relevance levels are smaller in comparison to the structured queries when measured as a difference to the baseline (and averaged over the three runs at each level). At all relevance levels, runs with English as source language are most

Table 3. 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

affected by not using query structuring in the target queries. There is a clear connection here to the number of words in the different runs: the number of words is by far the largest in the English-Finnish run (see Table 1). Runs with Swedish as source language are remarkably less affected, again the number of words in the Swedish-Finnish target query is clearly smaller than elsewhere.

4 Discussion

In our experiments, dictionary-based CLIR was performed under three conditions: 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 criteria are used. Instead, if stringent criteria are used, i.e. when only highly relevant documents are accepted, as high performance 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.

A quite common reason for a mismatch between a query and a newspaper article is that newspaper articles take up specific, concrete things whereas topics express 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

Table 4. Effectiveness of unstructured target queries at three separate relevance thresholds (non-interpolated average precision).

Language route, Rel = 3	Average precision	Difference	Difference (%)
Finnish-Finnish	28.1	-	-
Swedish-Finnish	15.0	-13.1	-46.6
English-Finnish	14.5	-13.6	-48.4
German-Finnish	12.1	-16.0	-56.9
Language route, Rel = 2,3	Average precision	Difference	Difference (%)
Finnish-Finnish	36.8	-	-
Swedish-Finnish	25.5	-11.3	-30.7
English-Finnish	18.1	-18.7	-50.8
German-Finnish	17.9	-18.9	-51.4
Language route, Rel = 1,2,3	Average precision	Difference	Difference (%)
Finnish-Finnish	37.6	-	-
Swedish-Finnish	26.3	-11.3	-30.1
English-Finnish	18.3	-19.3	-51.3
German-Finnish	19.0	-18.6	-49.5

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 normalized 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 late 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. 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. It is possible that adding words automatically would not be as successful in raising the rankings of late retrieved documents. Therefore it remains an issue whether source or target language query expansion (see [1]) would increase query effectiveness regarding highly relevant documents.

Further research is needed to find out why retrieving highly relevant documents was not as successful as retrieving fairly and marginally relevant documents. When the same queries retrieve documents of other relevance levels quite successfully, it is an interesting question why they fail with respect to the highly relevant ones. Is there something inherent in the highly relevant documents that makes the difference?

5 Conclusion

In this paper, dictionary-based CLIR was tested in a best match retrieval environment, using graded relevance assessments. A 4-point relevance scale was used in the test database, which consists of 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 and unstructured queries. Effectiveness of the translated queries was compared to that of the monolingual queries using *stringent*, *regular* and *liberal* relevance criteria (*stringent*: only highly relevant documents accepted ; *regular*: highly and fairly relevant documents together accepted; *liberal*: highly, fairly and marginally relevant documents together accepted). Reasonable CLIR performance was achieved

when *liberal* or *regular* relevance criteria were used. Instead, when *stringent* criteria were used, 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 were studied, reasons for the low rankings of these documents were found.

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.

The authors thank the anonymous referees for useful suggestions.

References

1. L. Ballesteros, and W.B. Croft. 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, ACM Press, 64–71, 1998.
2. J. Broglio, J. Callan, and W.B. Croft. INQUERY system overview. In *Proceedings of the TIPSTER text program (Phase I)*. San Francisco, CA: Morgan Kaufmann Publishers, 1994.
3. CLEF Homepage. Available: <http://clef.iei.pi.cnr.it>
4. A. Fujii, and T. Ishikawa. 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>
5. T. Hedlund, H. Keskustalo, A. Pirkola, M. Sepponen, and K. Järvelin. Bilingual tests with Swedish, Finnish and German queries: dealing with morphology, compound words and query structure. In C. Peters (ed.) *Cross-language information retrieval and evaluation: Proceedings of the CLEF 2000 workshop*, Lecture Notes in Computer Science: 2069, 210–223, 2001.
6. K. Järvelin, and J. Kekäläinen. 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, ACM Press, 41–48, 2000.

7. J. Kekäläinen, and K. Järvelin. Using Graded Relevance Assessments in IR Evaluation. *Journal of the American Society for Information Science and Technology*, 53(13): 1120–1129, 2002.
8. K. Kishida et al. 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>
9. S. Lee et al. 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>
10. NTCIR Homepage. Available: <http://research.nii.ac.jp/ntcir/index-en.html>
11. A. Pirkola, T. Hedlund, H. Keskustalo, and K. Järvelin. Dictionary-Based Cross-Language Information Retrieval: Problems, Methods, and Research Findings. *Information Retrieval*, 4 (3/4), 209–230, 2001.
12. A. Pirkola, H. Keskustalo, E. Leppänen, A.-P. Käsälä, and K. Järvelin. Targeted s -Gram Matching: a Novel n -Gram Matching Technique for Cross- and Monolingual Word Form Variants. *Information Research*, 7 (2), 2002. Available: <http://InformationR.net/ir/7-2/paper126.html>
13. E. Sormunen. A Method for Measuring Wide Range Performance of Boolean Queries in Full-Text Databases. Dissertation. Tampere, University of Tampere, 2000.
14. E. Sormunen. 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, ACM Press, 320–330, 2002.
15. TREC Homepage. Available: <http://trec.nist.gov/>
16. E. Voorhees. Evaluation by Highly Relevant Documents. In *Proceedings of the 24th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, New York, ACM Press, 74–82, 2001.
17. Y. Zhou, J. Qin, M. Chau, and H. Chen. 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>