The Effects of Query Structure and Dictionary Setups in Dictionary-Based Cross-language Information Retrieval

Ari Pirkola
Department of Information studies
University of Tampere
liarpi@uta.fi

Abstract In this study, the effects of query structure and various setups of translation dictionaries on the performance of cross-language information retrieval (CLIR) were tested. The document collection was a subset of the TREC collection, and as test requests the study used TREC's health related topics. The test system was the INQUERY[1] retrieval system. The performance of translated Finnish queries against English documents was compared to the performance of original English queries against English documents. Four natural language query types and three query translation methods, using a general dictionary and a domain specific (= medical) dictionary, were studied. There was only a slight difference in performance between the original English queries and the best cross-language queries, i.e., structured queries with medical dictionary and general dictionary translation. The structuring of queries was done on the basis of the output of dictionaries.

1 Introduction

In recent years, the development of Internet and related technology has created world-wide multilingual document collections. At the same time information flow across languages has grown due to increased international collaboration. With these factors in the background, IR research has paid increasing attention to cross-language IR (CLIR) systems, where the user presents a query in one language, and in response the system retrieves documents in another language.

The need of CLIR systems in today's world is obvious. Moving from the global perspective to an individual level, CLIR is useful, for example, for the people, who are able to understand a foreign language, but have difficulty in using it actively. Other examples of the use of CLIR are given by Oard and Dorr (1996).

In cross-language IR either documents or queries have to be translated. Research has concentrated on query translation, as it is computationally less expensive than document translation, which requires a lot of memory and processing capacity (Hull & Grefenstette, 1996). Within the query translation framework, basic approaches to CLIR are machine translation (MT), corpus-based methods, and dictionary-based methods.

In current MT systems the quality of translations is often low (Hull & Grefenstette, 1996; Oard & Dorr, 1996; Yamabana et al., 1996). High quality translations can be obtained only when the applicable domain is limited, so that the system can provide sufficient domain knowledge. For IR, the basic problem is, however, that the user requests often are mere sequences of words, without proper internal syntactic structure. Disambiguation in MT systems is, however, based on syntactic analysis. Therefore, MT is not regarded as a promising method for query translation in CLIR.

In corpus-based methods queries are translated and expanded on the basis of multilingual terminology derived from comparable document collections or parallel corpora, these containing similar or identical documents in different languages. A variety of methods based on this approach have been developed, and the performance of queries has varied depending on the method used (Davis, 1997; Davis & Dunning, 1996; Dumais et al., 1996; Sheridan & Ballerini, 1996; Sheridan et al., 1997). Sheridan et al. (1997) found that their corpus-based CLIR queries performed almost as well as the monolingual baseline queries.

Dictionary-based translation is often easier way to implement query translation than the methods based on the comparable documents or the parallel corpora, as these are not readily available. Therefore in the University of Tampere we have adopted the dictionary-based method for our CLIR studies. The dictionaries used in CLIR are often bilingual machine readable dictionaries (MRD), designed for human readers but converted for the purpose of cross-language retrieval, or bilingual or multilingual thesauri, developed specifically for CLIR (see Gilarranz et al., 1997). MRD translation uses a trivial method, in which a source language word is replaced by all of its target language equivalents, all of which are taken to the final query. Studies have shown that the performance of MRD-based system in its simplest form has, roughly, only half of the performance of its monolingual counterpart (Ballesteros & Croft, 1996; Davis & Dunning, 1996; Hull & Grefenstette, 1996).

The basic problems associated with MRD translation are (1) phrase translation, (2) translation polysemy (translation ambiguity), and (3) the coverage of dictionaries. If phrases are not identified, MRD translates phrase constituents instead of full phrases, and the senses of multi-word keywords are lost. This results in decreased precision. Automatic phrase identification methods have been developed for CLIR environment (Ballesteros & Croft, 1997). Translation polysemy is a phenomenon, in which the number of word senses increases when a source language word is translated to a target language by replacing it with all of its target language equivalents.

[1] The INQUERY information retrieval system is provided by the Information Retrieval Laboratory of the Computer Science Department, University of Massachusetts, MA, USA

In the study reported in this paper, the translation polysemy and the dictionary coverage problems were attacked by means of the combination of a general language MRD and a domain specific MRD, i.e., a medical dictionary. The domain was restricted to medicine and health by choosing as test requests TREC's (see Harman, 1996) health related topics. The performance of translated Finnish queries against English documents was compared to the performance of original English queries against English documents. Because the domain was medicine and health, it was assumed that the medical dictionary disambiguates word senses, giving less incorrect senses than the general dictionary, and that it contains such search keys that are not found in the general dictionary.

Besides the special dictionary effect, the effect of query structure on CLIR performance was studied. The main query types were: (1) Natural language sentence based queries ("natural language" as opposed to queries based on concept analysis, i.e., "Boolean" queries). These were regular English sentences. (2) Natural language word and phrase based queries. Both types were divided into two subtypes, structured and unstructured queries. The structuring of queries was carried out mechanically by grouping the target language lexical synonyms of a source language expression into the same facet (Section 3.2., Figure 1)

This paper will demonstrate that the approach adopted in the study solves successfully the translation polysemy and the dictionary coverage problems. The use of the special dictionary and the general dictionary in query translation and structuring of queries are highly effective methods to improve the CLIR performance. The paper will also offer explanations, why these methods have positive effects.

2 The test environment

The test environment of the study consisted of:

- TREC's health related topics, documents and relevance assessments for the documents
- TWOL morphological analyser for Finnish
- A medical MRD and a general MRD for Finnish - English translation
- INQUERY retrieval system
- Kstem morphological analyser for English

The test collection consisted of the AP Newswire, Federal Register, and DOE abstracts subsets of the TREC collection. The collection contained 514,825 documents. The size of the basic file was 1,46 GB. The health related test requests, totalling 34, were selected from the TREC topics 1-300.

The TWOL morphological analyser produces automatically basic word forms from inflected word forms and decomposes compound words into their constituents. Inflected word forms of natural language/sentence queries were automatically turned to basic forms, because the dictionary entry words are in basic forms. Compounds were split, because sometimes they are found in dictionaries only as their constituents. Both compounds and their constituents were translated. The TWOL software performs stemming and compound splitting effectively, although Finnish is a morphologically complex language.

The study used two Finnish - English - (Finnish) MRDs, a general dictionary and a medical dictionary. The general dictionary contained 65,000 Finnish and 100,000 English entry words. The medical dictionary contained 67,000 Finnish and English entry words. The commercial versions of the dictionaries were converted automatically to CLIR versions by removing from them all other material except for actual dictionary words. This conversion was done by means of a simple filter program. There were some errors in the output of automatic conversion. As a consequence, some translations failed. However, of all the translations done in the study only a fraction were incorrect.

INQUERY is a probabilistic information retrieval system based on Bayesian inference net model (Broglio et al, 1994). Queries can be formulated as regular English sentences or can be structured; a wide variety of operators is available for structuring. The Kstem morphological analyser, which produces real English words as its stemming output, is part of INQUERY's latest version. The Kstem software was used for stemming the words of the documents of the test collection. Thus the database index included stemmed words.

3 Methods

3.1 The basic test processes

Figure 1 gives an overall picture of the basic test processes. The processes are presented in detail in Section 3.2. To get test queries that are comparable to the original English queries, the English queries were translated into Finnish by a human translator (by the author), and the Finnish queries were retranslated back to English by means of dictionaries. This approach is often used in dictionary-based CLIR studies. Figure 1 also illustrates the query structuring method of the study. This is discussed in Section 3.2.

3.2 Constructing and translating queries

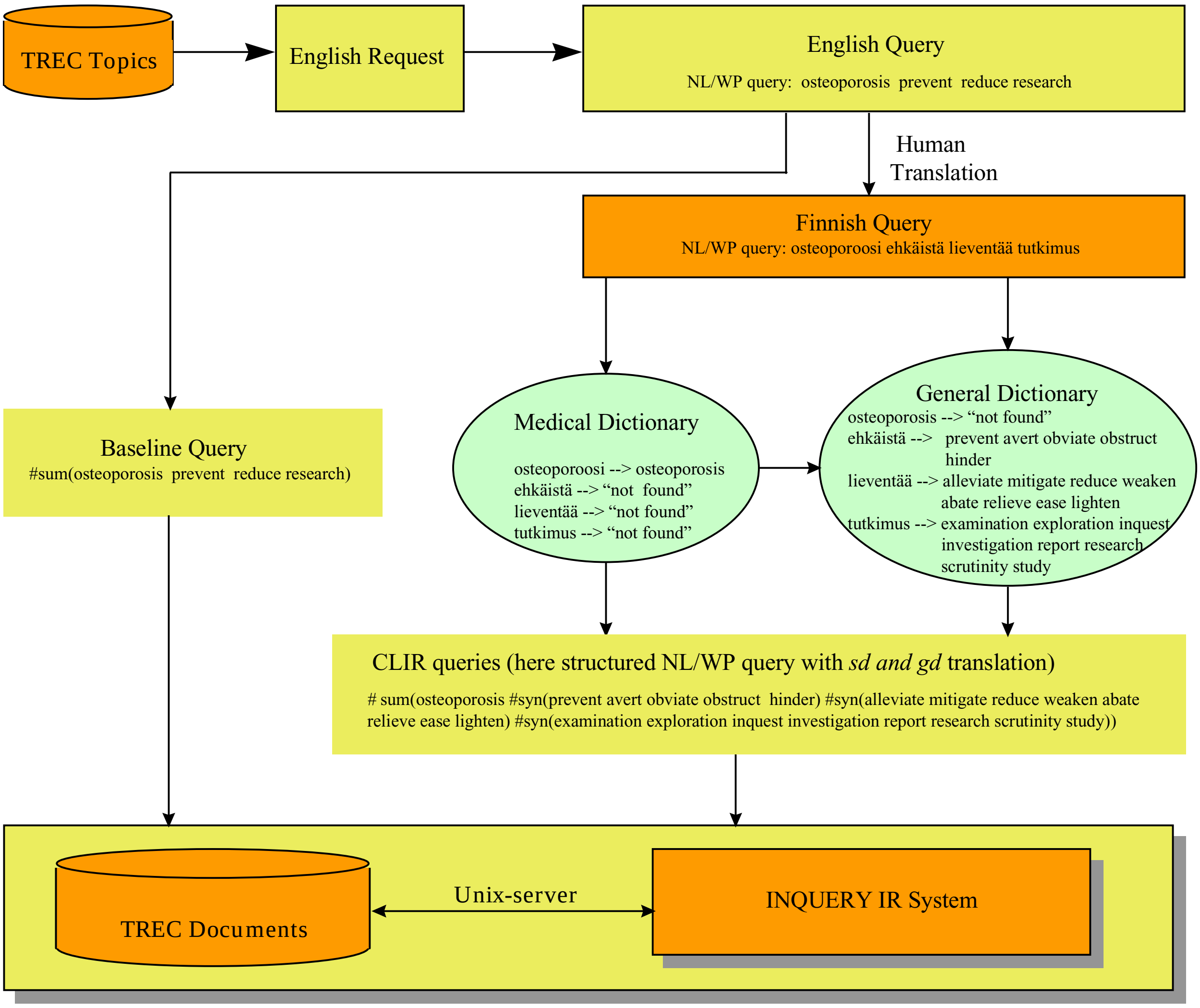
In TREC topics, important words are found in title, description, and narrative fields (some topics do not have narrative fields). On the basis of these fields, the author constructed test requests. The test requests were abridged versions of the TREC topics, consisting of 1-2 natural English sentences (see the Appendix). Hence, the test requests represented the requests that could be used by real users. In many cases the title as such was used as a request.

There were two main query types. The requests as such were the first type. This type is called natural language/sentence, and is abbreviated to NL/S. The second type was formulated on the basis of the requests by selecting from them the most important words and phrases. It is called natural language/word and phrase, and is abbreviated to NL/WP.

The English NL/S and NL/WP queries were translated into Finnish by the author, who is a native Finnish speaker. The author has worked in the field of medical informatics and has thus understanding of medical terminology. As a translation aid the author used printed dictionaries. The test MRDs were not used in this phase. The English NL/S and NL/WP queries that provided the basis for Finnish queries, were also used as baselines for CLIR queries (see Figure 1).

For each English word a precise equivalent was given. However, it was not always possible to form the Finnish sentences (NL/S queries) that had precisely translated words and were also good Finnish. Therefore, as good Finnish was given priority, some translations were not precise. This probably favoured the baseline queries. According to Hull and Grefenstette (1996) human translation in CLIR experiments is an

Figure 1. The basic test processes


TREC Topics
English Request
English Query
NL/WP query: osteoporosis prevent reduce research
Human Translation
Finnish Query
NL/WP query: osteoporoosi ehkäistä lieventää tutkimus
Baseline Query
#sum(osteoporosis prevent reduce research)
Medical Dictionary
osteoporoosi --> osteoporosis
ehkäistä --> "not found"
lieventää --> "not found"
tutkimus --> "not found"
General Dictionary
osteoporosis --> "not found"
ehkäistä --> prevent avert obviate obstruct hinder
lieventää --> alleviate mitigate reduce weaken abate relieve ease lighten
tutkimus --> examination exploration inquest investigation report research scrutinity study
CLIR queries (here structured NL/WP query with *sd and gd* translation)
# sum(osteoporosis #syn(prevent avert obviate obstruct hinder) #syn(alleviate mitigate reduce weaken abate relieve ease lighten) #syn(examination exploration inquest investigation report research scrutinity study))
TREC Documents
Unix-server
INQUERY IR System

additional source of error. It does not occur in an operational CLIR setting.

Both NL/S and NL/WP queries were divided into two subtypes, structured and unstructured queries. The structured queries had dictionary-based facets, i.e., the words that were derived from the same Finnish word, were grouped into the same facet. Figure 1 illustrates the structuring method applied in the study, showing how the original English query (NL/WP) *osteoporosis prevent reduce research* is transformed into a structured CLIR query (see also the Appendix). The operators of the queries of Figure 1 are described below. It should be noted that unstructured and structured NL/WP queries, as well as unstructured and structured NL/S queries, are comparable, as they are derived from the same Finnish queries. They also have the same baseline. The NL/WP and NL/S queries are not compared, as they do not have identical search key sets.

Compound words are common in Finnish, whereas noun phrases, except for proper name phrases, are relatively rare. A Finnish compound word is very often translated as a noun phrase in English (like *compound word* and *yhdyssana*). This is the main reason why phrases were identified in NL/WP queries (both in the original English queries, or the baseline, and in the Finnish queries). In this way a precise correspondence was obtained between the baseline queries with their many phrases and the Finnish queries with their many compound words. In this respect the correspondence of the original English queries and the Finnish queries was perfect, as there were no such cases where a Finnish (English) expression (a compound, an identified phrase, or a single word) corresponded to an English (Finnish) expression that was a phrase, but which was not identified as a phrase. Phrase identification probably favoured the baseline queries. The effect on CLIR queries was small, as the Finnish queries did not have many phrases.

Table 1
The morphological and lexical correspondence of English - Finnish - English search keys. The application of the proximity operator uw3

	Original English query	**Finnish query**	**MRD translation (after compound splitting)**	**The elements combined by the proximity operator (uw3)**
1	**Single word** prominent	**Single word** kuuluisa	kuuluisa --> celebrated, famed, famous, noted, prominent, renowned, well-known	prox(well known)
2	**Phrase** heart ailment	**Compound** sydänvaiva	a. sydän --> heart vaiva --> ailment, complaint, discomfort, inconvenience, trouble, vexation b. sydänvaiva --> "not found in the dictionaries"	prox(heart ailment), prox(heart complaint), prox(heart discomfort), prox(heart inconvenience), prox(heart trouble), prox(heart vexation)
3	**Phrase** genetic code	**Phrase** geneettinen koodi	a. geneettinen --> genetical genetic koodi --> code cipher b. geneettinen koodi --> genetic code	a. prox(genetical code), prox(genetic code), prox(genetical cipher), prox(genetic cipher) b. prox(genetic code)

Table 1 shows the morphological and lexical correspondence of English - Finnish - English search keys, and also shows how a proximity operator (prox = uw3, see below) was applied in the test queries. There were other morphological correspondence types in the test data that are not shown in the Table 1, but they were rare (like English compound - Finnish single word). By far the most usual type was the single word - single word case.

The query operators for NL/S and NL/WP queries were the sum-, syn-, and uw3-operators. Search keys contained in the sum-operator have equal influence on search results. The system uses it automatically when no operator is specified. The syn-operator was used in structured CLIR queries; the words of the same facet were combined by the syn-operator. The syn-operator treats its operand search keys as instances of the same key. The uwn-operator (unordered window n) is a proximity operator. It was applied to phrases, with n=3.

The proximity operator uw3 was used in the following cases. First, it was applied to phrases in the NL/WP baseline queries (Table 1: "heart ailment", "genetic code"). Second, it was applied to the English phrases that the dictionaries gave as their output (Table 1: "well known", "genetic code"). Third, Finnish compound words were decomposed, and both full compounds and their constituents were translated, and the operator was applied to the English equivalents of the constituents; those equivalents that were derived from the first part of the compound were joined to those equivalents that were derived from the second part of the compound. All the combinations were generated (Table 1: item 2, the last column). The compound words with three component words were treated similarly. Fourth, also the phrases of Finnish NL/WP queries were translated as full phrases and as phrase constituents. The uw3-operator was then applied to the English equivalents of the phrase constituents as in the case of compound words (Table 1: item 3, item *a* in the last column).

Finnish NL/S and NL/WP queries were run through the TWOL morphological analyser, after which stop-words were removed from the NL/S queries. The stop-word list included Finnish equivalents for the words of the retrieval system's stop-word list. This English list was applied to the baseline queries (and also to the English CLIR queries). Thus, when a certain word was removed from the baseline query, its equivalent was removed from the Finnish query.

Three translation methods were examined in the study. They were the following:

- *gd* translation: Finnish search keys were translated in the general dictionary.
- *sd* → *gd* translation: Finnish search keys were translated in the medical dictionary and the general dictionary, in this order. General dictionary translation was applied after medical dictionary translation only if the latter did not translate a word.
- *sd and gd* translation: Finnish search keys were translated both in the medical dictionary and the general dictionary. Duplicate words were removed.

If a word or a phrase was not found as an entry word in the dictionaries, it was sent unchanged to the final query. These kinds of expressions were English proper names, acronyms, and Finnish words not found in the dictionaries.

The test processes were the following:

- English requests (=NL/S) were constructed from TREC topics
- The English NL/S queries were run as baseline queries through the INQUERY retrieval system
- English NL/WP queries were constructed on the basis of the requests
- The English NL/WP queries were run as baseline queries through the INQUERY retrieval system
- The English NL/S and NL/WP queries were translated into Finnish by the author
- The Finnish NL/S and NL/WP queries were run through the TWOL morphological analyser
- Stop-words were removed from the Finnish NL/S queries
- MRD translation: CLIR queries were formed by translating the search keys of Finnish NL/S and NL/WP queries in (a) the general dictionary (*gd*), (b) the medical dictionary → the general dictionary (*sd* → *gd*), and (c) the medical dictionary and the general dictionary (*sd and gd*)
- MRD translation gave twelve CLIR query types:
    - Unstructured NL/S: *gd*, *sd* → *gd*, *sd and gd* queries
    - Structured NL/S: *gd*, *sd* → *gd*, *sd and gd* queries
    - Unstructured NL/WP: *gd*, *sd* → *gd*, *sd and gd* queries
    - Structured NL/WP: *gd*, *sd* → *gd*, *sd and gd* queries
- CLIR queries were run through the INQUERY retrieval system

- The performance of the unstructured NL/S queries was compared to the performance of the structured NL/S queries, the performance of the unstructured NL/WP queries was compared to the performance of the structured NL/WP queries, and the performance of all the CLIR queries was compared to the performance of the baseline queries.

4 Findings

The performance of test queries was evaluated as precision at 10% recall, average precision at 10%-100% recall, and as precision-recall graphs. The results are presented in Tables 2-3 and Figures 2-5. It should be noted that unstructured and structured NL/S queries, as well as unstructured and structured NL/WP queries, are derived from the same Finnish queries and are thus comparable. The difference is that unstructured queries are unfaceted, but structured queries have dictionary-based facets.

As shown, there is a significant gap between the baseline (BL) and the unstructured NL/S queries (Table 2 and Figure 2). At 10% recall, the precision of the baseline is 37.9%, but only 15.4% for *gd* queries. Special dictionary effect is clear, but baseline queries still perform much better than *sd → gd* and *sd and gd* queries; at 10% recall the precision of *sd → gd* and *sd and gd* queries is, roughly, only half of the precision of the baseline. When average precision is considered, the gap in performance is greater in favour of the baseline.

Structure put in NL/S queries through dictionaries results in a dramatic improvement in performance (Table 2 and Figure 3). At 10% recall the best CLIR queries, *sd and gd*, give the precision figure 35.9%, which is only 2.0% below the precision of the baseline queries (37.9%). The average precision of *sd and gd* queries is 12.9% and that of the baseline queries16.8%. The translation method used is of great importance. At the high precision level (10% recall - 50% recall), *gd* and *sd → gd* queries perform much poorer than *sd and gd* queries.

As shown in Table 3 and Figure 4, the performance of unstructured NL/WP queries is significantly below that of the baseline. Figure 4 is very much like Figure 2, which demonstrates the behaviour of the unstructured NL/S queries. As in NL/S queries, structuring improves the performance of NL/WP queries dramatically (Table 3 and Figure 5). The best structured cross-language queries, *sd and gd,* do almost as well as the baseline. For the former, precision at 10% recall is 31.1%, and for the latter 31.8%. The average precision is practically the same, 12.4% for *sd and gd* queries and 12.5% for the baseline. At three recall levels, 50%, 60%, and 70%, *sd and gd* queries give better precision figures than the baseline. The figures are, respectively, 13.2% and 12.8%, 9.5% and 8.8%, and 6.3% and 6.1% (these figures are not given in the tables of the paper).

Table 2. The performance of NL/S queries

Query type/Translation type	10%-recall P	Average P
Structured, dictionary-based facets		
GD	30,9	10,5
SD --> GD	30,4	11,3
SD and GD	35,9	12,9
Unstructured		
GD	15,4	5,1
SD --> GD	19,2	5,8
SD and GD	20,4	6,3
Structured and unstructured, baseline	37,9	16,8

Table 3. The performance of NL/WP queries

Query type/Translation type	10%-recall P	Average P
Structured, dictionary-based facets		
GD	24,9	9,8
SD --> GD	26,1	10,5
SD and GD	31,1	12,4
Unstructured		
GD	16,5	5,7
SD --> GD	14,6	5,0
SD and GD	19,3	6,5
Structured and unstructured, baseline	31,8	12,5

Figure 2
Precision-recall curves for unstructured NL/S queries

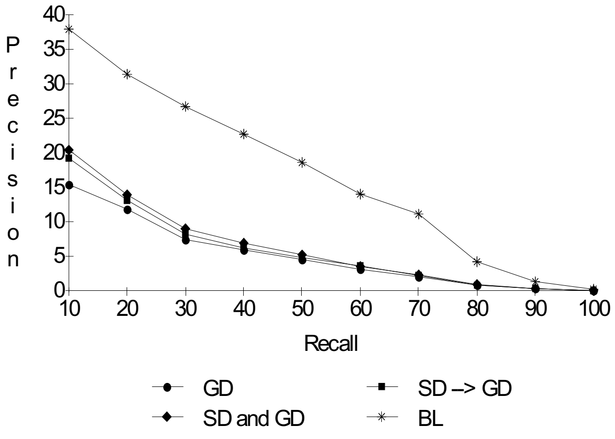


Figure 4
Precision-recall curves for unstructured NL/WP queries

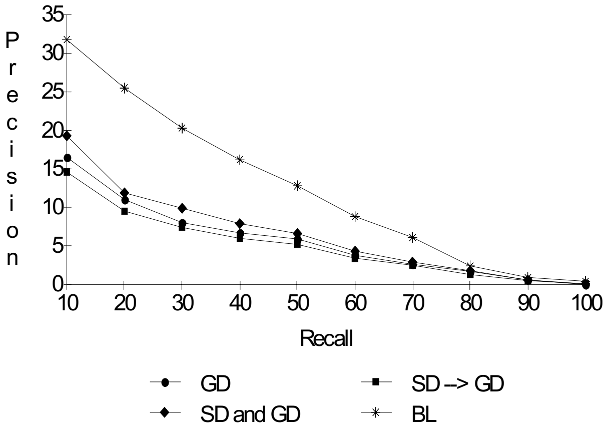


Figure 3
Precision-recall curves for structured NL/S queries

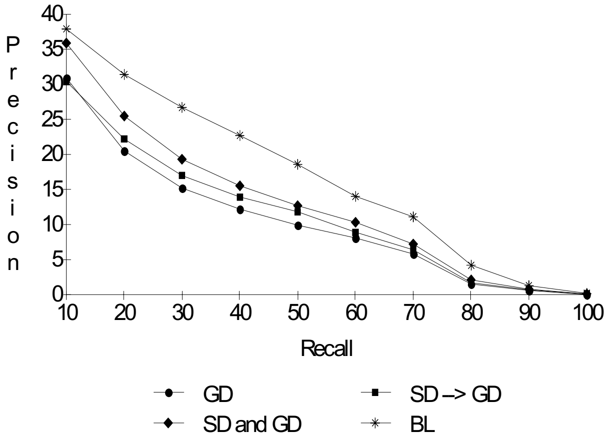
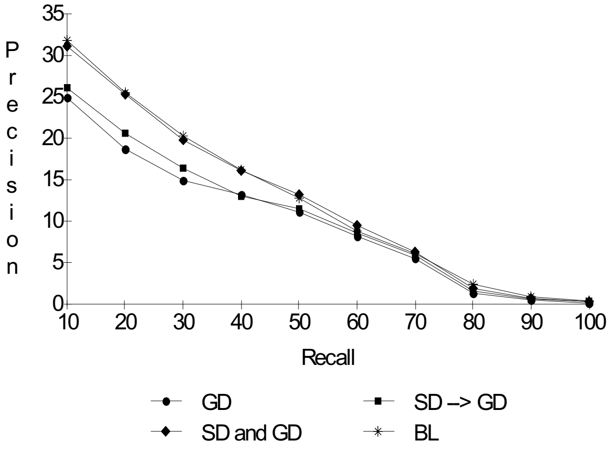


Figure 5
Precision-recall curves for structured NL/WP queries



5 Discussion

The results can be summarised as follows:

- MRD-based cross-language queries can reach the performance level of monolingual queries.
- Structuring of queries is of primary importance in cross-language IR.
- A special dictionary has positive effects in MRD-based cross-language IR.
- The performance of special dictionary/general dictionary based CLIR queries depends on the way in which special and general dictionary translations are applied.

Previous experiments have indicated that there is a significant gap between cross-language and monolingual queries. Hull and Grefenstette (1996) worked with translated French queries and English documents of the TIPSTER collection consisting of newspaper articles. They found that the precision of the queries translated in the automatically generated word-based dictionary, was only 60% of the precision of the original English queries. However, the performance of multi-word cross-language queries based on the manual editing of the dictionary was some 90% of the performance of the original English queries. The findings of Ballesteros and Croft (1996) are consistent with those reported by Hull and Grefenstette. They found a 55% loss in average precision in queries translated word-by-word compared to the original queries. Ballesteros and Croft (1997) studied the effect of corpus-based query expansion on CLIR performance, and found that expansion helped to counteract the negative effects of translation failures. Davis and Dunning (1996) and Davis (1997) also found that the performance of MRD-based CLIR queries was much poorer than that of monolingual queries. However, when MRD translation was supplemented with parts-of-speech (POS) disambiguation, or POS and corpus-based disambiguation, CLIR queries performed much better. The average precision figures were the following: simple MRD-based method, 0.14; MRD-based method supplemented with POS, 0.19; MRD-based method supplemented with POS and corpus-based disambiguation, 0.21; monolingual baseline queries, 0.29 (Davis, 1997). Hull (1997) found that MRD-based Boolean structured queries performed better than MRD-based vector space queries. For the former, the average precision was 0.28, and for the latter 0.20.

Thus, the previous studies show that simple MRD-based CLIR queries perform poorly. However, the performance can be improved by supplemental methods and by structuring of queries.

The results presented in this paper show that MRD-based CLIR queries perform almost as well as monolingual queries, if domain specific MRD is used together with general MRD and queries are structured on the basis of the output of dictionaries. Because the commercial versions of the dictionaries were converted automatically to CLIR versions, with no manual changes done to the dictionaries or the translations, the performance level of the CLIR queries achieved in the study can be achieved in practice in an operational CLIR setting.

The study used a structuring method, in which those words that were derived from the same Finnish word were grouped into the same facet. This method can be implemented automatically unlike structuring based on concept analysis. The latter requires a human interpreter to identify the concepts in the requests.

One possible way by which structuring disambiguates CLIR queries is that it enforces "conjunctive" relationships between search keys. Disambiguation through increasing the weight of relevant search keys is an important way of disambiguation (Hull, 1997). Often those search keys that have only one or two translations are the most important words of a request and, vice versa, those keys that have many translations are unimportant words. Thus, in unstructured CLIR queries unimportant search keys and irrelevant translation equivalents tend to dominate and depress the effect of important keys. As a result, many nonrelevant documents are ranked high. But when search keys are faceted, they no longer are equal, but now facets are treated equally. In this way important search keys get relatively more weight (see the sample queries of the topic "osteoporosis" in the Appendix)

The positive effects of the medical dictionary were mainly due to two factors. First, it contained search keys that were not found in the general dictionary, and second, it disambiguated word senses. Typically, domain specific terms are prevalent in special dictionaries, whereas in general dictionaries they are not common. Therefore general dictionaries only rarely translate specific terms. In many queries, however, these are crucial search keys. Consider the following TREC-topic: "What research is ongoing to reduce the effects of osteoporosis and prevent the disease". The most important search key is obviously the term *osteoporosis*. The term occurs in the medical dictionary used in the study, but does not occur in the general dictionary. Also some other test queries contained terms that are only found in the medical dictionary. The results were not good for these queries in the case of "general dictionary only" -translation. In these cases s*d* → *gd* and *sd and gd* queries performed much better.

General dictionaries often give many equivalents to a source language word, whereas special dictionaries usually give only one or two equivalents. The terms of special dictionaries are often unambiguous. For these reasons, a special dictionary alleviates the translation polysemy problem, in which the translation of one source language word to many target language words causes fuzziness in CLIR queries. For example, the general dictionary of this study gives for the Finnish word *leikkaus* 7 equivalents: *cut, cutting, clipping, operation, editing, section,* and *retrenchment*. The number of word senses is more than 7. The medical dictionary gives only two equivalents: *operation* and *surgery*. The number of word senses is more than two, but much less than in the case of the general dictionary.

Because the domain was predetermined as medicine and health, in most cases the medical dictionary gave correct and unambiguous translation equivalents. However, sometimes it had negative effects. First, for two Finnish words it gave incorrect equivalents. Second, sometimes it gave translations that are not used in everyday life. These are the main reasons why s*d* → *gd* queries, for which general dictionary translations were not done if the medical dictionary translated words, did not perform as well as *sd and gd* queries. In the latter, the positive effects of the good words of the general dictionary outweighed the negative effects of the bad words of the medical dictionary.

However, the efficiency of a translation method surely depends on the text type of a database. The document collection of this study included mainly news documents. These kind of texts contain both everyday and scientific language. Therefore special dictionary *and* general dictionary translation turned out to be a good method. It is likely that s*d* → *gd* translation*,* in which the special dictionary gets more weight, is suited for scientific texts.

Duplicate terms were removed in *sd and gd* translation, because the study wanted to test "standard" *sd and gd* translation, or the simplest possible *sd and gd* translation method. If the duplicates would not have been removed, *sd and gd* queries would probably have given better results, because those search keys that are found in both dictionaries seem to be important search keys in health related topics.

In the study, the domain was restricted to medicine and health, but this does not mean that the approach described here would only be applicable to the texts of a specific domain. A wide variety of electronic domain-specific dictionaries is available in many languages, especially the dictionaries where English is in combination with another language. In CLIR translation systems, it is possible to use many dictionaries, each of which have limited content, but which together cover general language issues and many specific domains. It is possible to address automatically the domain specific terms of queries to the correct dictionaries, because different domains have different terminologies. However, polysemous terms, i.e., terms having different meanings in different domains, may be a problem to the approach like this. However, this kind of polysemy is probably not common. Alternatively, the user could inform the system of the domain of her/his query.

Of the three main problems associated with MRD translation, the phrase problem is language specific. MRD translation is easier for such languages as Finnish, German, and Swedish, which are rich in compound words, than for such languages that are rich in phrases. In this respect Finnish as a source language is an easy language for cross-language IR. On the other hand, Finnish is a complex language morphologically. It is not rare that two words (basic forms) have common inflected forms and that a basic form of one word may be an inflected form of another word. Therefore, as a result of stemming and compound splitting, new search keys (basic forms) were generated, in particular for NL/S CLIR queries. Fuzziness in CLIR queries was partly due to morphological analysis. The second problem associated with MRD translation, the translation polysemy, probably concerns most languages equally. The third problem, the coverage of dictionaries is not a linguistic problem and is in principle the same for all languages. Therefore, as the study attacked the translation polysemy and the dictionary coverage problems, the results are applicable to most languages, even though phrases can lower the relative performance of CLIR in some languages.

6 Conclusions

The results reported in this paper show that a cross-language IR system based on MRD translation is able to achieve the performance level of a monolingual system, if queries are structured and if both general terminology and domain specific terminology are available in translation. By these means, it is possible to solve successfully the translation polysemy and the dictionary coverage problems.

Future research should concentrate on finding methods by which the performance of CLIR queries could be improved further. In the University of Tampere we will also study CLIR queries based on concept analysis (i.e., "Boolean" queries). Also weighting methods should be tested: Does weighting affect CLIR queries similarly as monolingual queries? Still another method that would be worth studying is data fusion; different translation methods produce different result lists. Would the results be improved, if the lists are combined?

Acknowledgements I would like to extend thanks to Kalervo Järvelin for his valuable comments and suggestions and other help, Heikki Keskustalo for help with technical issues and interesting discussions on the research problems, Eero Sormunen for useful suggestions and interesting discussions, and Jaana Kekäläinen and Paul Solomon for good comments.

References

Ballesteros, L. & Croft, W.B. 1997. Phrasal Translation and Query Expansion techniques for Cross-Language Information Retrieval. In Working Notes of AAAI Spring Symposium on Cross-Language Text and Speech Retrieval, Stanford, CA, pages1-8.

Ballesteros, L. & Croft, W.B. 1996. Dictionary-based methods for cross-lingual information retrieval. In Proceedings of the 7th International DEXA Conference on Database and Expert Systems Applications, pages 791-801.

Broglio, J., Callan, J. & Croft, W.B. 1994. Inquery system overview. In Proceedings of the TIPSTER Text Program (Phase I), pages 47-67.

Davis, M. 1997. New experiments in cross-language text retrieval at NMSU's Computing Research Lab. In Harman, D.K. (ed.) Proceedings of the Fifth Text REtrieval Conference (TREC-5), Gaithesburg, MD, pages 447-454.

Davis, M. & Dunning, T. 1996. A TREC-evaluation of query translation methods for multi-lingual text retrieval. In Harman, D.K. (ed.) Proceedings of the Fourth Text REtrieval Conference (TREC-4), Gaithesburg, MD, pages 483-497.

Dumais, S.T., Landauer, T.K. & Littman, M.L. 1996. Automatic cross-linguistic information retrieval using latent semantic indexing. In Grefenstette, G. & Smeaton, A. & Sheridan, P. (eds.), Working Notes of the Workshop on Cross-Linguistic Information Retrieval, ACM SIGIR, Zürich, Switzerland, pages 16-23.

Gilarranz, J., Gonzalo, J. & Verdejo, F. 1997. An Approach to Conceptual Text Retrieval Using the EuroWordNet Multilingual Semantic Database. In Working Notes of AAAI Spring Symposium on Cross-Language Text and Speech Retrieval, Stanford, CA, pages 51-57.

Harman, D. 1996. Overview of the Fourth Text REtrieval Conference (TREC-4). In Harman, D.K. (ed.) The Fourth Text REtrieval Conference (TREC-4), pages 1- 23.

Hull, D. 1997. Using Structured Queries for Disambiguation in Cross-Language Information Retrieval. In Working Notes of AAAI Spring Symposium on Cross-Language Text and Speech Retrieval, Stanford, CA, pages 73-81.

Hull, D. & Grefenstette, G. 1996. Querying across languages. A dictionary-based approach to multilingual information retrieval. In Proceedings of the 19th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval, Zürich, Switzerland, pages 49-57.

Oard, D. & Dorr, B. 1996. A Survey of Multilingual Text Retrieval. Technical Report UMIACS-TR-96-19, University of Maryland, Institute for Advanced Computer Studies.

Sheridan, P., Braschler, M. & Schäuble, P. 1997. Cross-Language Information Retrieval in a Multilingual Legal Domain. In Peters, C. & Thanos, C. (eds.) Research and Advanced Technology for Digital Libraries. First European Conference, ECDL '97, Pisa, Italy, 1-3 September, Proceedings. Lecture Notes in Computer Science, Vol. 1324, pages 253 - 268.

Sheridan, P. & Ballerini, J. 1996. Experiments in Multilingual Information Retrieval using SPIDER system. In Proceedings of the 19th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval, Zürich, Switzerland, pages 58-65.

Yamabana, K., Muraki, K., Doi, S. & Kamei, S. 1996. A Language Conversion Front-end for Cross-Linguistic Information Retrieval. In Grefenstette, G. & Smeaton, A. & Sheridan, P.(eds.) Working Notes of the Workshop on Cross-Linguistic Information Retrieval, ACM SIGIR, Zürich, Switzerland, pages 34-39.

APPENDIX

Sample requests and queries

The test requests were abridged versions of the TREC topics. For many topics, e.g., for the topic 160, the original TREC topic title was used as a request.

Topic 160

The request (=NL/S query): vitamins - the cure for or cause of human ailments

The keywords selected for the NL/WP query: vitamin human ailment cure cause

Topic 216 Osteoporosis

The original TREC topic:

Description: What research is ongoing to reduce the effects of osteoporosis in existing patients as well as prevent the disease occurring in those unafflicted at this time?

The request (=NL/S query): What research is ongoing to reduce the effects of osteoporosis and prevent the disease

The keywords selected for the NL/WP query: osteoporosis prevent reduce research

The following queries (*sd and gd* translation = *sd → gd* translation) of the topic "osteoporosis" represent all CLIR query types of the study and demonstrate the importance of structure in cross-language queries. Note how the term *osteoporosis* has relatively more weight in the structured queries.

Structured NL/WP

#sum(osteoporosis #syn(prevent avert obviate obstruct hinder impede arrest delay retard avoid) #syn(alleviate mitigate reduce weaken abate relieve ease lighten) #syn(examination exploration inquest investigation report research scrutiny study))

Unstructured NL/WP

#sum(osteoporosis prevent avert obviate obstruct hinder impede arrest delay retard avoid alleviate mitigate reduce weaken abate relieve ease lighten examination exploration inquest investigation report research scrutiny study)

Structured NL/S

#sum(osteoporosis #syn(prevention repression restraining restraint obstruction contraception dwarfing stunting) disease #syn(#uw3(bring about) cause create effect #uw3(give rise to) inflict produce) #syn(consequence effect outgrowth result) lieventäminen #syn(join #uw3(join in) ally #uw3(join together) unite #uw3(be connected) #uw3(be linked) belong) #syn(examination exploration inquest investigation report research scrutiny study) meneillä)

Unstructured NL/S

#sum(osteoporosis prevention repression restraining restraint obstruction contraception dwarfing stunting disease #uw3(bring about) cause create effect #uw3(give rise to) inflict produce consequence effect outgrowth result lieventäminen join #uw3(join in) ally #uw3(join together) unite #uw3(be connected) #uw3(be linked) belong examination exploration inquest investigation report research scrutiny study meneillä)