

Dictionary-Based Cross-Language Information Retrieval: Problems, Methods, and Research Findings

Ari Pirkola, Turid Hedlund*, Heikki Keskustalo, and Kalervo Järvelin

Department of Information Studies, University of Tampere, Finland

* Swedish School of Economics and Business Administration Library, Helsinki

Email: pirkola@cc.jyu.fi, turid.hedlund@shh.fi, ccheke@uta.fi, likaja@uta.fi

Published in Information Retrieval 4(3/4), 209-230

Abstract

This paper reviews literature on dictionary-based cross-language information retrieval (CLIR) and presents CLIR research done at the University of Tampere (UTA). The main problems associated with dictionary-based CLIR, as well as appropriate methods to deal with the problems are discussed. We will present the structured query model by Pirkola and report findings for four different language pairs concerning the effectiveness of query structuring. The architecture of our automatic query translation and construction system is presented.

1. Introduction

There is an increasing amount of full text material in various languages available through the Internet and other information suppliers. Therefore cross-language information retrieval (CLIR) has become an important new research area. It refers to an information retrieval task where the language of queries is other than that of the retrieved documents. For an overview of the approaches to cross-language retrieval, see (Hull and Grefenstette 1996; Oard and Dorr 1996).

This paper contributes a literature review on dictionary-based CLIR and summarizes our own research at the University of Tampere (UTA). We will discuss the methods, linguistic and technical problems, and empirical findings in dictionary-based CLIR. At UTA, our CLIR research is based on Pirkola's method (Pirkola 1998) which consists of dictionary translation and the structured query model (Kekäläinen and Järvelin 1998, Pirkola 1998). The structuring of queries refers to the grouping of search keys, and the use of proper query operators. We summarize empirical findings

for four different language pairs, namely Finnish to English, English to Finnish, Swedish to English, and German to English query translation are reported. We will discuss briefly other CLIR related research done at UTA, which involves language typology research for CLIR (Pirkola 2001b), and studies of statistical methods for keyword assessment (Pirkola and Järvelin 2001a).

The main problems associated with dictionary-based CLIR are (1) untranslatable search keys due to the limitations of general dictionaries, (2) the processing of inflected words, (3) phrase identification and translation, and (4) lexical ambiguity in source and target languages. The category of untranslatable keys involves new compound words, special terms, and cross-lingual spelling variants, i.e., equivalent words in different languages which differ slightly in spelling, particularly proper names and loanwords. In this paper *translation ambiguity*¹ refers to the increase of irrelevant word senses in translation due to lexical ambiguity in the source and target languages.

The structure of this paper is as follows. Sections 2 to 5 consider the problems of untranslatable search keys, inflected keys, phrase translation, and translation ambiguity, as well as the main techniques to handle the problems in dictionary-based CLIR. Section 6 presents the structured query model. Section 7 reports our findings with language pairs Finnish to English (Pirkola 1998, Pirkola 1999, Pirkola et al. 1999), English to Finnish (Pirkola et al. 2000; Puolamäki et al. 2001), and the CLEF (Peters 2000) experiments Swedish to English, German to English, and Finnish to English (Hedlund et al. 2001b). We report findings concerning the effectiveness of query structuring through synonym sets. Section 8 discusses our CLEF studies. Section 9 summarizes the research done at UTA and presents concluding remarks.

2. Untranslatable search keys

No translation dictionary available for CLIR in any language is as extensive as the lexicon of the language listing all its words. There are theoretical and practical limitations. An apparent one is the fact that new words and word combinations can be generated readily in natural languages. The most important categories of untranslatable search keys generally not found in general dictionaries are new compound words, proper names and other spelling variants, and special terms.

2.1 Compound words

A *compound word* (or a *compound*) is defined as a word formed from two or more words that are written together.² In CLIR, the distinction between compositional, non-compositional, and semi-compositional phrases (Section 4) and compounds is important. Compounds whose meaning can be derived from the meanings of component words are called *compositional compounds* (Akmajian et al. 1990). For instance, the meaning of the Finnish word *kaupunginhallitus* (*city government*) comes from the meanings of the components *kaupungin* (*city*, in genitive) and *hallitus* (*government*). In compositional compounds, a full compound typically is a hyponym (a narrower term) of its headword. A compound whose meaning cannot be deduced on the basis of its components is called *a non-compositional compound*. Here the term *semi-compositional compound* refers to a compound whose meaning is in part interpretable on the basis of the components, for example, a full compound is a hyponym of its headword, but semantically unrelated or only metaphorically related to the other component(s), e.g. the Finnish compound *krokotiilinkyyneleet* (*crocodile tears*).

Due to the productive nature of natural languages, words can be combined into new compound words readily. Some languages, such as German, Swedish, Finnish, and Dutch are characterized by high frequency of compounds (see, respectively, Sheridan and Ballerini 1996, Hedlund et al. 2001a, Pirkola 1999, Bosch and Daelemans 1999). It is obvious that for such languages effective dictionary look-up and the searching of compound words in CLIR cannot solely be based on full compounds but also on their component words. Because translation dictionaries may not include full compounds as such but only their components, the decomposition of compounds and separate translation of component words is often useful. In the case of compositional compounds separate translation will give correct senses.

From the CLIR perspective, a compound word is a more convenient type of expression than a phrase, because compound decomposition is easier than phrase identification (Section 4.1).

Compound splitting can be performed effectively by means of a lexicon-based morphological analyser. For many languages there are morphological programs capable of decomposing compounds, see for instance Xerox's morphological analyser tools.³ To retain the compound sense in the target language, the target language equivalents of the component words can be combined by

¹In CLIR literature the term is used in different senses.

²The term is also used in a broad sense including the definitions of *compound word* and *phrase* of this paper.

³<http://www.xrce.xerox.com/>

a proximity operator (Section 6). This requires that compounds are handled also when indexing the target language collection.

2.2 Proper names and other spelling variants

Translation dictionaries may include some *proper names*, such as the names of capital cities and countries. Most proper names, however, are not covered. Particularly, translation dictionaries do not contain personal names. In many languages proper names are spelling variants of each other. Also *loanwords* may differ slightly in their written form from the original words. For example, Japanese romanized *katakana* words and English words are similar (but not identical) to each other, e.g. *deeta* - *data* (Fujii and Ishikawa 2001). The loanwords may not always be listed in translation dictionaries.

A common method to handle untranslatable words in dictionary-based CLIR is to pass them as such to a CLIR query (the final target language query). However, in the case of spelling variants a source language form does not match the variant form in a database index, causing loss of retrieval effectiveness. To find target language spelling variants for source language words, some alternative method for translation, such n-gram based matching or other approximate string matching technique, or transliteration based on phonetic similarities between languages, has to be applied. In the n-gram method, search keys and the words of documents are decomposed into n-grams, i.e., into the substrings of length n. The degree of similarity between search keys and index terms can then be computed by comparing their n-gram sets (Pfeifer et al. 1996, Robertson and Willett 1998, Salton 1989, Zobel and Dart 1995).

Zobel and Dart (1995) tested different indexing and string matching techniques for personal name variants and spelling errors in English. Matching based on phonetic coding was shown to be a poor method while n-gram matching was effective. The findings also suggested that updating of n-gram indexes is straightforward and index size and retrieval time are acceptable. However, in real IR there is one more problem that must be solved. N-gram matching is non-binary matching that gives a ranked output of strings. From the viewpoint of a topic for which documents are searched, amongst the best matches, say 20 or 30 strings, there are often good and bad keys. Several bad keys even amongst many good keys may deteriorate query performance. Thus, even though from the approximate matching perspective n-gram matching is effective, from the CLIR perspective its effectiveness may be low. This is a similar problem to translation ambiguity in CLIR (Section 5).

Therefore it is likely that the same query structuring method - based on defining alternative translations as synonyms - that is applied in the structured query model (Section 6), would also be effective in the case of n-gram based name searching in CLIR. Now the best matches yielded by n-gram matching would be regarded as synonyms and would be combined into the same facet in a query. N-gram matching together with query structuring could be used both in monolingual and cross-lingual name searching. At UTA we will test this method in our current work using a comprehensive test collection with requests containing proper names.

In morphologically complex languages, proper name searching in CLIR is further complicated by inflection (Puolamäki et al. 2001). For example, the name *Gorbachev* is written as *Gorbatshov* in Finnish, and this may take several inflectional forms, like *Gorbatshoville* (allative, *to Gorbachev*), *Gorbatshovin* (genitive, *Gorbachev's*), etc. We indexed the unrecognized words of our Finnish test database (over 119 000 unrecognized forms) as digrams and found that the frequency of such digrams that act as suffixes or are part of suffixes is often high in particular at suffix positions. For example, the *in*-gram is a genitive suffix and common at all positions of strings, but far more common at the last position (65,734 occurrences), i.e., genitive position, than at other positions. At the third last position, for example, its frequency is 10,823. These findings suggest that an *n-gram based stemming* in which the grams with high frequency at suffix positions would be downweighted at suffix positions, would benefit retrieval. For example, without positional downweighting the *in*-genitive forms may be ranked high when such names as *Ingman* and *Lindberg* are searched. In general, the lower the frequency of a specific gram the higher its resolution power.

Fujii and Ishikawa (2001) developed a transliteration method based on phonetic equivalence between Japanese romanized *katakana* words and English words. Transliteration was applied for words not found in the translation dictionary. The researchers reported the method to be useful in Japanese to English CLIR in terms of retrieval performance. Other transliteration methods involve that of Chen et al. (1998) for Chinese and English words and that of Lee and Choi (1997) for Korean and English words.

2.3 Special terms

The use of a special dictionary together with a general dictionary extends the base of translatable words. General dictionaries only rarely translate specific terms. In many queries, however, these are primary search keys. Therefore, a combination of a general and domain specific dictionary would benefit a CLIR system (Pirkola 1998). In fact, in CLIR translation systems it would be possible to

use many dictionaries, each of which have limited content, but which together cover general language and many specific domains. The source language request words together may be used to predict proper special dictionaries, or the user may be asked about the request domain(s).

The use of a special dictionary in CLIR has another advantage: disambiguation effect on CLIR queries. General dictionaries often give many equivalents to a source language word, whereas special dictionaries typically give 1-2 equivalents only. The terms of special dictionaries are often unambiguous. For these reasons, a special dictionary reduces the translation ambiguity problem.

In the case of a CLIR system based on the combination of different dictionaries, there are alternative ways to do translation. Pirkola (1998) studied Finnish to English CLIR and tested two translation methods, sequential and parallel translation. The test environment consisted of a newspaper database, health related topics, and a dictionary combination of a general and medical dictionary (514,000 document TREC subcollection and 34 TREC topics). In *sequential translation*, Finnish search keys were translated by means of 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. In *parallel translation*, translations were taken from both dictionaries. Parallel translation was a clearly better technique both for low performance (unstructured) and high performance (structured) queries. For some keys, sequential translation erroneously gave specific senses. This effect is depressed in parallel translation. The results by Pirkola (1998) suggest that keeping general words and special terms in separate lists does not offer advantages in a database of common language texts (news documents). However, the effectiveness of a translation method surely depends on the text type of a database. News documents contain both everyday and scientific language. This accounts for the effectiveness of the parallel translation technique. It is likely that sequential translation, in which a special dictionary gets more weight, is suited for scientific texts.

3. Word inflection

A commonly used method to deal with *inflected search keys* as well as derivationally related keys is to remove affixes from word forms (Harman 1991, Porter 1980). The method is called *stemming*. The output is a common *root* or *stem* of different forms, which is not necessarily a real word. In lexicon-based *morphological analysis* word forms are *normalized* into base forms which are real words. Morphological analysis also allows the splitting of compounds into their component words.

In CLIR, word form normalization is used as a preprocessing stage for translation to enable the matching of source language keys with dictionary headwords (which are in base forms) also in the case of inflected search keys. Alternatively, source language keys and headwords can be conflated into the same form by a stemmer (Davis and Ogden 1997). One problem related to stemming is that different headwords may be conflated into the same form.

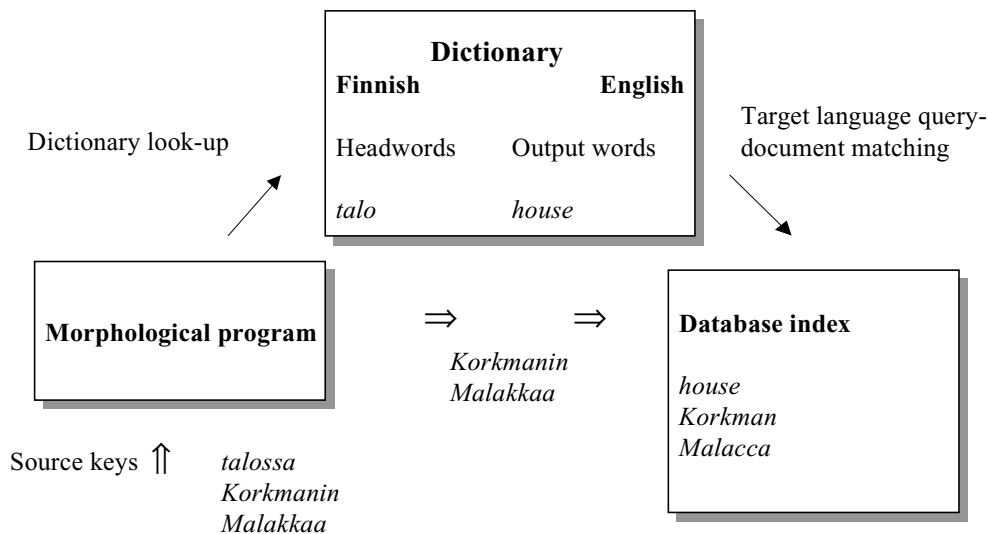


Figure 1. Unrecognized words in morphological analysis

The effectiveness of morphological analysis is limited by the size of a morphological program's lexicon (Hull 1996). As in the case of translation dictionaries, it is impossible to list exhaustively the words of a language in the lexicon. This contributes to the problem of untranslatable search keys. Figure 1 shows how different types of inflected words are handled in dictionary-based Finnish to English CLIR. Most inflected keys are normalized and translated by a dictionary, like the form *talossa* (in the house) which is first normalized into *talo* and then translated into *house*. The unrecognized forms are sent unchanged into a CLIR query. The names *Korkmanin* (personal name in genitive) and *Malakkaa* (a spelling variant of the geographical name *Malacca* in accusative) represent typical cases of words not listed in the lexicon of a morphological program. They do not match with the English index keys. The latter word would not match even if it were normalized. In dictionary-based CLIR, they could be handled similarly as untranslatable spelling variants in the case of a translation dictionary (Section 2.2).

If index terms are stemmed, dictionary output words also have to be stemmed (Davis and Ogden 1997). In the case of the normalized index keys, the normalization of the output words does not seem necessary, but might be useful since some dictionary output words may be in inflected forms, e.g., some phrase component words (Hedlund et al. 2001b).

Regarding word inflection CLIR effectiveness also depends on the morphological processing of index keys (the monolingual component of CLIR). The effectiveness of stemming depends on the language. In general, *recall* can be expected to improve due to stemming since a larger number of potentially relevant documents are retrieved. Research done in different languages has shown that for many languages stemming also improves *precision*. This holds for English (Hull 1996, Krovetz 1993), French (Savoy 1999), modern Greek (Kalamboukis 1995), and Arabic (Abu-salem et al. 1999). Popovic and Willett (1992) found that stemming resulted in a significant increase in retrieval effectiveness in Slovene language, which is a morphologically complex language. The effectiveness was measured as the number of relevant retrieved documents at document cut-off value of 10. Ekmekcioglu and Willett (2000) used the same evaluation measure and showed that stemming increased retrieval effectiveness in Turkish retrieval. It is likely that morphological normalization has similar effects. In Boolean retrieval, it was discovered that morphological normalization of Finnish may be used to increase recall with a marginal drop in precision (Alkula 2000).

4. Phrases

4.1 Phrase identification

The methods used to identify phrases involve statistical methods, i.e., the use of collocation statistics (Buckley et al. 1996), part-of-speech (POS) tagging (Jing and Croft 1994), and shallow syntactic analysis (Sheridan and Smeaton 1992, Strzalkowski 1995, Zhai et al. 1997). In some retrieval systems syntactic parsing is supplemented with statistical methods to recognize subphrases within phrases (Strzalkowski 1995, Zhai et al. 1997).

Words are often ambiguous in their part-of-speech. For example, the word *run* may be a verb and a noun. In *part-of-speech tagging* words are assigned parts-of-speech in their sentential context. In this way meaningful sequences of words, such as adjacent nouns, can be recognized, and words that are ambiguous in part-of-speech can be disambiguated.

Shallow syntactic analysis, a more demanding task than part-of-speech tagging, aims at detecting phrases and head-modifier relations within phrases (Sheridan and Smeaton 1992, Strzalkowski 1995). Some shallow syntactic analysers identify the functional roles of words in sentences, such as verbs and their arguments. Typical *phrase algorithms* based on syntactic analysis extract noun phrases from syntactically parsed natural texts. The noun phrases or the elements in them, in particular head-modifier pairs, are then used as supplementary search keys and index terms.

4.2 Phrase translation in CLIR

Translation of phrases as full phrases is of prime importance in CLIR. Hull and Grefenstette (1996) studied French to English text retrieval. The test query types and their performance (average precision at 5, 10, 15 and 20 documents) were as follows: queries based on an automatically generated word-based dictionary (0.235), queries based on a manually built word-based dictionary (0.269), and queries based on a manually built multi-word (phrase) dictionary (0.357). The original English queries (baseline) gave the average precision of 0.393. Thus, word-based CLIR queries performed much poorer than the baseline queries, while the gap between phrase-based CLIR queries and baseline queries was small. These findings were corroborated by Ballesteros and Croft (1996) who studied English to Spanish and Spanish to English text retrieval and reported a 55% loss in average precision for queries translated word-by-word compared with the original queries. A 30% loss in performance resulted from translation ambiguity and a 20% loss was due to inaccurate translation of phrases. In another study the researchers showed that in automatic phrase translation the correctness of translations is of crucial importance (Ballesteros and Croft 1997). If phrase translation fails, phrase-based queries may perform poorer than word-based queries.

Phrases are not a major problem for languages in which multi-word expressions are compound words rather than phrases, such as German, Swedish, Finnish, and Dutch (Pirkola 1999). However, if phrases are not identified and translated correctly, the effects on certain queries may be fatal.

Ballesteros and Croft (1998) reported phrase translation by a dictionary to improve retrieval performance. However, not all phrases are listed in dictionaries, which suggests the use of some additional or alternative translation method. Phrase translation based on the word collocation statistics in the target language has been reported to be a useful method (Ballesteros and Croft 1998, Fujii and Ishikawa 2001). Fujii and Ishikawa (2001) explored phrase translation in Japanese to

English retrieval. In Japanese technical terms are often phrases⁴. New technical phrases are generated from existing words, and the new phrases are not generally listed in dictionaries. The researchers constructed a Japanese-English base word dictionary in the domain of technology. Phrases were translated on a word-by-word basis, and appropriate translations were then selected on the basis of word collocation statistics in the target language. The method tended to retain Japanese phrase senses in English. The use of the method improved system performance for all the three retrieval methods tested. For example, for the standard weighting method the average precision of queries based on the phrase translation method was 0.2324 while queries based on the existing dictionary (in which new technical phrases were not listed) gave the precision figure of 0.1785.

5. Translation ambiguity

5.1 Homonymy and polysemy

A lexical item or form with (at least) two entirely distinct meanings is said to be *homonymous*. Thus homonyms are different lexemes with the same form (Lyons 1984).⁵ The senses of homonyms are unconnected. A lexeme which has more than one sense is *polysemous* (Karlsson 1998). The word *board*, for example, has several (sub)senses, e.g., (a) *a thin plank*, (b) *a tablet*, (c) *a table*, and (d) *food served at the table*. The senses of a polysemous word are related to each other, e.g., one subsense may be a metaphorical extension of another subsense.

The main difference between homonymy and polysemy is that homonymy is a relation between two (or more) lexemes, but polysemy is a property of a single lexeme. It is difficult, however, to apply the distinction between homonymy and polysemy consistently (Akmajian et al. 1990, Lyons 1984, Kilgarriff 1993). The term *lexical ambiguity* covers homonymy and polysemy. Based on morphology lexical ambiguity can be divided into base form and inflectional ambiguity. *Base form ambiguity* covers the condition in which two (or more) lexemes (usually two separate headwords of a dictionary) have the same (base) form, as well as the condition in which one lexeme has two or more senses. The ambiguous form may belong to two or more part-of-speech categories. The word *run*, for example, can be used as a verb and a noun. *Inflectional ambiguity* refers to a condition in which two or more lexemes (headwords of a dictionary) share at least one common inflectional

⁴The researchers used the term *compound word*. In this paper the term is used in a strict sense as defined in Section 2.

⁵A *lexeme* is a set of word forms which belong together (Karlsson 1998), or a word considered as a lexical unit, in abstraction from the specific word forms it takes in specific constructions (Matthews 1997).

form. The lexemes may belong to the same POS category or different categories. The form *failing*, for example, is a base form noun, and a second participle form of the verb *fail*.

5.2 The nature of translation ambiguity

Here *translation ambiguity* refers to the increase of irrelevant search key senses due to source and target language lexical ambiguity. Translation ambiguity and difficulty in handling phrases are the main factors for the low effectiveness of plain dictionary-based CLIR queries (Ballesteros and Croft 1996, Grefenstette 1998, Hull and Grefenstette 1996).

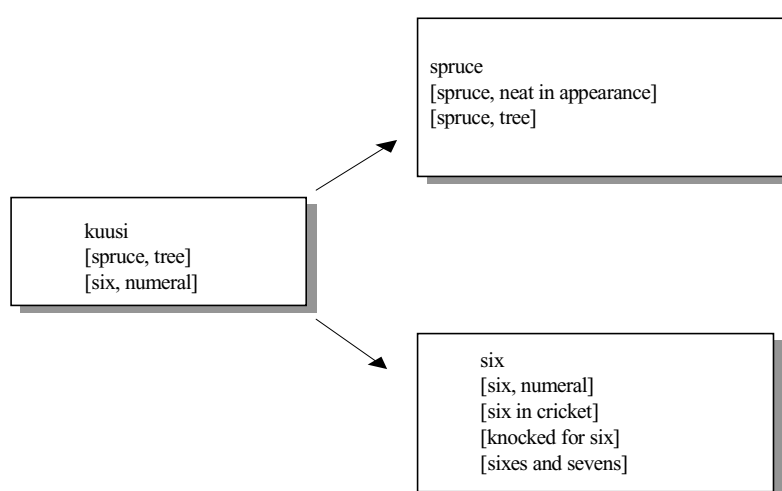


Figure 2. Translation ambiguity

Figure 2 illustrates the increase of ambiguity in a translation process. The Finnish form *kuusi* is homonymous and has two senses: [spruce, tree], [six, numeral]. The English word *spruce* has two and the word *six* four senses (Collins dictionary 1998): [spruce, neat in appearance], [spruce, tree], [six, numeral], [six in cricket], [knocked for six], [sixes and sevens].

Think that in monolingual Finnish and monolingual English searching the correct sense is [spruce, tree]. There is one extraneous sense in Finnish, [six, numeral], as well as English, [spruce, neat in appearance]. But in Finnish to English retrieval there are five extraneous senses, [spruce, neat in appearance], [six, numeral], [six in cricket], [knocked for six], [sixes and sevens]. Thus, lexical ambiguity associated with CLIR queries stems in part from a source language and in part from a target language.

The effectiveness of a CLIR query depends on the number of relevant search key senses in relation

to the number of irrelevant senses in the CLIR query. The proportion is here regarded as an ambiguity measure of *degree of ambiguity* (DA). In the spruce-example above, the degree of ambiguity is increased from 1/1 in both Finnish and English retrieval to 1/5 in Finnish to English retrieval. If there are no incorrect senses, DA for a query is zero. DA illustrates the phenomenon of translation ambiguity, and only refers to the proportion of incorrect senses to correct senses in a query, but does not deal with the frequency of senses in a collection which is another important aspect.

Because of the importance of translation ambiguity in CLIR and differences between languages in lexical ambiguity, CLIR research is confronted with evaluation problems, such as how to compare different systems processing different languages. At UTA we are studying language typologies. The aim is to develop language independent methods to quantify morphological and semantic properties of languages for CLIR system development and evaluation. The question of morphological and lexical-semantic differences between languages is considered in Pirkola (2001b).

5.3 Disambiguation techniques in CLIR

The techniques to handle translation ambiguity in dictionary-based CLIR involve part-of-speech tagging, various corpus-based disambiguation methods, and query structuring (section 6).

Part-of-speech tagging

Davis (1997) studied English to Spanish retrieval and found that dictionary-based translation supplemented with POS disambiguation, or POS and corpus-based disambiguation, improved CLIR effectiveness. In POS disambiguation, only those equivalents which had the same part-of-speech as the source language key were selected for the final query. The average precision figures were as follows: simple dictionary-based method, 0,14; dictionary-based method supplemented with POS, 0,19; dictionary-based method supplemented with POS and corpus-based disambiguation, 0,21; monolingual baseline queries, 0,29. The findings were corroborated by Ballesteros and Croft (1998) who tested POS disambiguation in English to Spanish CLIR. However, Pirkola (2001a) found that the advantages of resolving POS-ambiguous and inflectionally ambiguous source keys in Finnish to English CLIR were negligible. Nevertheless, it is expected that POS disambiguation has different effects in different languages due to variation in lexical ambiguity.

Corpus-based disambiguation

In dictionary-based CLIR it is possible to exploit additional resources to resolve translation ambiguity, in particular source and target language corpora (Ballesteros and Croft 1996, Ballesteros and Croft 1997, Chen et al. 1999, Yamabana et al. 1996), bilingual aligned corpora (Davis 1997, Davis and Dunning 1996), and comparable document collections (Peters and Picchi 1996). Corpus-based disambiguation techniques involve query expansion (QE) to reduce the effects of bad translation equivalents (Ballesteros and Croft 1996, Ballesteros and Croft 1997, Chen et al. 1999), the use of word co-occurrence statistics for selecting the best or correct translations (Chen et al. 1999, Yamabana et al. 1996), and the selection of translation equivalents on the basis of aligned sentences (Davis 1997, Davis and Dunning 1996).

Yamabana et al. (1996) developed an English to Japanese CLIR system, which used a bilingual translation dictionary, but in which disambiguation was based on source and target language corpora. The method selected the best equivalents on the basis of co-occurrence frequencies among source language keys and those among target language keys.

Ballesteros and Croft (1996, 1997) tested the effects of two pseudo relevance feedback techniques, i.e., *local feedback* (LF) and *local context analysis* (LCA), in English to Spanish and Spanish to English CLIR. Local feedback is a query expansion (QE) technique in which all the top ranked documents are assumed to be relevant (Attar and Fraenkel 1977). In LCA the passages of documents are ranked. The system then selects the top ranked words from the top ranked passages as QE terms. Both methods were tested both prior to (in a source language document collection) and after dictionary translation. Besides separate pre- and post-translation experiments, combined pre- and post-translation expansion was studied for LF and LCA. Thus there were 6 tests altogether: pre-translation LF, post-translation LF, pre-translation LCA, post-translation LCA, combined LF, and combined LCA. Both LCA and LF query expansion helped to counteract the negative effects of translation ambiguity. The best method was the combined pre- and post-translation LCA. The queries based on the combined LCA performed markedly better (average precision 0.14) than queries constructed by simple automatic translation (average precision 0.08).

6. The structured query model

Query structuring using the *#syn* and *#uwn* operators (see below) of the InQuery retrieval system (Allan et al. 1997, Turtle 1990) has been shown to be an effective disambiguation method in many CLIR studies (Ballesteros and Croft 1998, Hedlund et al. 2001b, Pirkola 1998, Pirkola 1999, Puolamäki et al. 2001, Sperer and Oard 2000). In this section we present the structured query model by Pirkola (1998). The model includes three search statement types (syn-, proximity (*#uwn*)-, and combined syn/proximity statements). It can be applied for five situations A-E discussed below. Among them, the effects of structuring through synonym sets, i.e., syn-based structuring (Case A) have been tested for four language pairs. The effects of phrase-based structuring (Case C) have been tested for a restricted phrase definition. The evaluation results are discussed in Section 7. The query structure examples presented in this section are taken from these experiments.

InQuery's #syn and #uwn operators

InQuery's *#syn*-operator treats its operand search keys as instances of the same search key. For the keys linked by the *#syn*-operator, an aggregate document frequency is computed instead of individual document frequencies for every key (Sperer and Oard, 2000). The formula for computing the probability of the *#syn*-operator is a modification of the *tf.idf* function as follows (Kekäläinen and Järvelin 1998):

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

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

S = a set of search keys within the *#syn* operator

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

adl = average document length in the collection

N = collection size (as a number of documents)

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

The *#uwn*-operator (unordered window) is a proximity operator. It only retrieves the documents

which contain the arguments of the operator within the defined window. In the structured query model it is used in compound- and phrase-based structuring.

A. Syn-based structuring

In *syn-based structuring* the translation equivalents of each source language key are grouped together by the *#syn*-operator of the InQuery retrieval system. In the case of a source language key giving both single word equivalents and proximity statements (see the cases B-E), all the items are combined by the *#syn*-operator (combined syn/proximity statement).

For the translation equivalents $a_1 \dots a_n$ of a source language key A , a search statement is as *#syn*($a_1 \dots a_n$). For instance, the Swedish word *möte* gives the translations *encounter, meeting, crossing, appointment, date*. These are combined by the *#syn*-operator as *#syn(encounter meeting crossing appointment date)*.

Syn-based structuring alters (relatively) tf.idf weights of keys in a query:

- Important search keys often have 1-2 translations only, and have relatively more weight in structured than in unstructured CLIR queries.
- In unstructured queries semantically insignificant rare keys (i.e., keys with low document frequency) may deteriorate query performance. In structured queries in syn-facets they are downweighted because of the aggregate document frequency (Sperer and Oard 2000)
- The findings by Pirkola et al. (1999) suggest that in syn-based queries correct conjunctive relationships between search keys are enhanced.

B. Compound-based structuring

In *compound-based structuring* the translation equivalents that correspond to the first part of a source language compound are joined by the proximity operator to those equivalents that correspond to the second part of the compound. All the combinations are generated. Compound words with three or more component words are treated similarly. The proximity statements are

combined by the #syn-operator.

For the Finnish to English and English to Finnish experiments (Section 7) a proximity operator of #uwn was chosen. Unordered window operator allows variable word orders in target language sentences. False co-ordinations seem unlikely if a window size is set small. Particularly, if there is variation in noun phrase structure in a target language, e.g., *information retrieval* and *retrieval of information*, unordered window operator seems to be a proper operator.

The #uwn-operator is a Boolean operator in the sense that it only retrieves the documents which contain the arguments of the operator within the defined window. The main *disambiguation effect* by compound-based structuring is that a source language compound typically gives 1-2 good word combinations in a target language, with the other combinations being *nonsense combinations* that do not retrieve any documents. Therefore, the number of proximity statements in the final query has just minor effects on results.

For the translation equivalents $a_1 \dots a_n$ and $b_1 \dots b_m$ of a source language compound AB , the proximity statements are as #uwn($a_1 b_1$),..., #uwn($a_1 b_m$),..., #uwn($a_n b_1$),..., #uwn($a_n b_m$). For instance, a Finnish compound *sydäntauti* (heart ailment) is decomposed by a morphological analyser into *sydän* (heart) and *tauti* (ailment). The translations and proximity statements are:

sydän → heart

tauti → ailment, complaint, discomfort, inconvenience, trouble, vexation

#uwn(heart ailment), #uwn(heart complaint), #uwn(heart discomfort), #uwn(heart inconvenience), #uwn(heart trouble), #uwn(heart vexation).

C. Phrase-based structuring

The case of source language phrases is treated in the same way as compound words (Case B). For the translation equivalents $a_1 \dots a_n$ and $b_1 \dots b_m$ of a source language phrase $A B$, the proximity statements are as #uwn($a_1 b_1$),..., #uwn($a_1 b_m$),..., #uwn($a_n b_1$),..., #uwn($a_n b_m$).

D. Handling semi-compositional compounds and phrases, and untranslatable component words

Because compound words and phrases may be semi-compositional or may include untranslatable component words, in addition to proximity statements single word translations are included in the final query and are combined by the synonym operator.

For the translation equivalents $a_1 \dots a_n$ and $b_1 \dots b_m$ and an untranslatable key C of a source language compound ABC , the proximity statements are:

$$\#uwn(a_1 b_1), \dots, \#uwn(a_1 b_m), \dots, \#uwn(a_n b_1), \dots, \#uwn(a_n b_m), \dots, \#uwn(a_1 C), \dots, \#uwn(a_n C), \\ \#uwn(b_1 C), \dots, \#uwn(b_m C).$$

The final combined syn/proximity statement is:

$$\#syn(\#uwn(a_1 b_1), \dots, \#uwn(a_1 b_m), \dots, \#uwn(a_n b_1), \dots, \#uwn(a_n b_m), \dots, \#uwn(a_1 C), \dots, \\ \#uwn(a_n C) \#uwn(b_1 C), \dots, \#uwn(b_m C) a_1, \dots, a_n b_1, \dots, b_m C).$$

For instance, the Swedish compound *brandbekämpning* (firefighting) is split into *brand* and *bekämpning* by the morphological program Swetwol. The first component translates into *fire* and *conflagration* but the second component does not translate at all (the Motcom Swedish-English translation dictionary, 60.000 words).⁶ The final statement is:

$$\#syn(\#uwn(fire bekämpning) \#uwn(conflagration bekämpning) fire conflagration \\ bekämpning)$$

Note that the syn-statement would only contain two nonsense proximity statements if this case of the model were not followed.

E. Target language phrases in a dictionary

The proximity operator of $\#uwn$ is applied to the output phrases of a dictionary. For the phrase $a b$ of a source language word A , a search statement is as $\#uwn(a b)$.

⁶ MOT Dictionary Software was used for automatic word-by-word translations. Copyright (c) 1998 Kielikone Oy, Finland.

7. Findings on syn-based structuring

From the cases A-E of the structured query model, we have tested empirically the effects of syn-based structuring (Case A) and phrase-based structuring (Case C) for a restricted phrase definition⁷. The results are presented in Table 1 and for the CLEF experiments also in Figure 3 (where 'str' refers to structured queries and 'uns' to unstructured queries). The UTA CLIR research group participated in CLEF'2000 conference (Section 8). Of the six CLEF test runs in Table 1 and Figure 3, unstructured Swedish to English and unstructured Finnish to English are unofficial runs, the other ones are official.

We used in all studies the *InQuery retrieval system* as a test system. For each study, the size of the test collection and the number of queries were as follows:

- Finnish to English, 514,000 documents, 34 test queries (the case 1 in Table 1). Pirkola (1998).
- English to Finnish, 55,000 documents, 20 test queries (the cases 2-3 in Table 1). Puolamäki et al. (2001).
- CLEF (for all experiments): 113,000 documents, 33 test queries (the cases 4-6 in Table 1). Hedlund et al. (2001b).

Table 1. The effects of syn-based structuring (average precision over 10/11 recall points)

Languages	Unstructured runs	Structured runs
1. Fin → Eng (n=34)	6,5	12,4
2. Eng → Fin (n=20)	18,8	27,4
3. Eng → Fin, syn+phrases (n=20)	18,8	29,0
4. Fin → Eng (n=33)	15,9	22,8
5. Ger → Eng (n=33)	21,6	26,7
6. Swe → Eng (n=33)	21,9	25,4

⁷In the English to Finnish experiment, the adjacent words as well as the words separated by the preposition *of* in English requests that correspond to compound words in Finnish requests were defined as phrases. Thus they were manually marked rather than automatically identified.

As can be seen in Table 1, in all cases the structured queries perform better than the unstructured ones. For Finnish to English (two experiments) and English to Finnish retrieval the performance of structured queries with respect to unstructured queries is better than for German to English and Swedish to English retrieval. In the English to Finnish experiments, phrase-based structured queries (avg. precision 29.0%) performed only slightly better than word-based structured queries (avg. precision 27.4%). It should be noted, however, that we adopted a specific type of phrase definition, i.e., correspondence between phrases and compound words. The results cannot be extended to other types of phrases, e.g., statistical phrases.

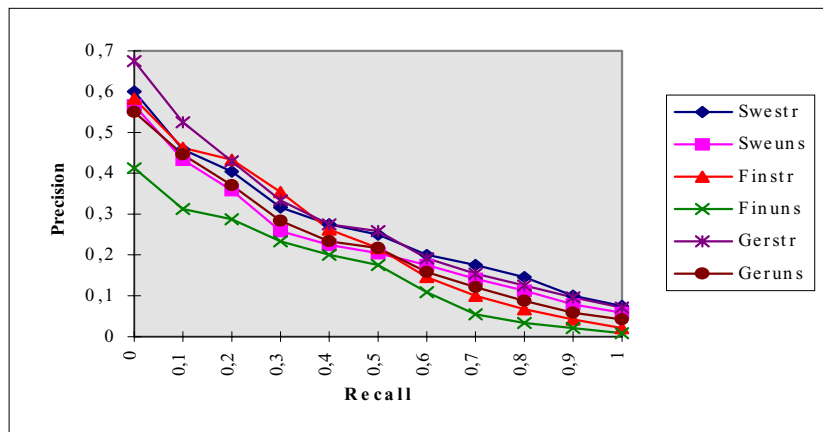


Figure 3. Precision-recall curves for the CLEF runs

We have planned to test the other cases of the structured query model and analyse why syn-based query structuring is more effective in Finnish to English and English to Finnish than in German to English and Swedish to English. We will explore the following issues as likely explanations:

- In the case of a compound language as a source language, the syn- and compound-based structuring methods have overlapping effects. This is because in compound-based structured (i.e., unstructured with respect to syn-facets) queries there often are 1-2 good proximity combinations which precisely correspond to the source language compound word, and several nonsense proximity combinations. In these cases, it does not make much difference whether or not the #syn-operator is used to combine the proximity statements. Thus, the effects of syn-based structuring depend on the proportion of compound words to single words in a source language (query) as well as the proportion of compositional compounds. In these respects,

compound languages differ from each other (Pirkola 2001b).

- The size of a dictionary. The smaller the dictionary the more common the one-to-one relations are, and the closer syn-based structured queries are to unstructured (compound-based structured) queries. If each source key has one equivalent, syn-based structured queries would be identical to unstructured queries.
- The language group (such as a language family). English, Swedish, and German all belong to the same language group, that is, Germanic languages. Finnish belongs to the Finno-Ugric language family. The semantic correspondence between words may be more straightforward (towards one-to-one relation) within a language group than between languages belonging to different groups. In other words, the effects of translation ambiguity may be smaller within a language group than across groups. This assumption is supported by the recent findings by Sperer and Oard (2000) who studied Chinese to English retrieval. The researchers found that the performance of structured CLIR queries based on Pirkola's model was around twice as high as that of unstructured CLIR queries. Chinese and English belong to different language families and differ from each other in many linguistic properties. Chinese is an isolating language having very weak morphology (Whaley 1997). English has much stronger morphology. Chen et al. (1999) reported that in English lexical ambiguity is more common than in Chinese. On the average, an English word had 1.687 senses and a Chinese word 1.397 senses. For the 1000 top high frequency words, the number of senses for English and Chinese words were, respectively, 3.527 and 1.504.
- Case D (semi-compositional compounds and untranslatable component words) of the structured query model was not applied in the CLEF experiments. For some queries, untranslatable component words ruined query performance. This probably dropped CLIR effectiveness and leveled out the differences between the query types.

8. The CLEF experiments at UTA

8.1 Research questions

We participated in the CLEF'2000 conference and studied the following research questions (Hedlund et al. 2001b).

- By what process, using bilingual dictionaries, can we automatically construct effective target language queries from source language request sentences?
- How does retrieval effectiveness vary when source languages vary?
- How does query structure affect CLIR effectiveness when using different source languages? We already presented in Section 7 our CLEF results concerning query structuring.

The first research question involves designing and implementing our method for automated bilingual query construction, using generally available bilingual dictionaries. The method seeks to automatically extract topical information from search topics in one of the source languages and to automatically construct a target language query. The resulting query may either be structured or unstructured.

Regarding the second research question, we made tests in CLEF's bilingual track with three different language pairs, that is, Finnish, Swedish and German as source languages, and English as a target language for each source language. We studied particularly the processing of compound words by morphological programs. All the source languages are rich in compounds and, thus, one of our main efforts was the morphological decomposition of compounds into constituents and their proper translation. In languages rich in compounds the translation of compounds (or their components) is a factor that greatly affects the retrieval results.

The main language and query processing components of the automatic query construction system are (see Figure 4):

- Word form normalization using the Swetwol, Fintwol, and Gertwol morphological analysers (by

Lingsoft, Helsinki)

- The removal of stopwords
- Compound splitting
- The removal of fogemorphemes, i.e., morphemes linking the components of compounds, from Swedish and German compounds and the subsequent normalization of the components
- The labelling of unrecognized words in the database index; proper names and other source query words not found in the dictionaries were added to the final queries unchanged
- The translation of source language words by means of bilingual translation dictionaries
- The construction of proximity statements in the case of compound word translations
- The construction of syn-statements in the case of structured queries
- The construction of the final queries using the #sum-operator

8.2 Compound splitting in the source languages

Compound splitting was performed only if a full compound was not listed in a dictionary. Generally, if a full compound is listed, separate translation of component words probably increases ambiguity. Our earlier tests indicated that the morphological analyser Swetwol for *Swedish* needs tuning in the case of compounds whose components are joined by fogemorphemes (Hedlund et al. 2001a). To solve this problem we developed a fogemorpheme algorithm which seeks to turn all the constituents of a compound into their base forms. *German* has a special feature of the noun initial letters being capital letters. For dictionary look-up, nouns as compound constituents need to get an upper-case initial letter after decomposition. We also removed one common fogemorpheme from German compounds, namely the “s“.

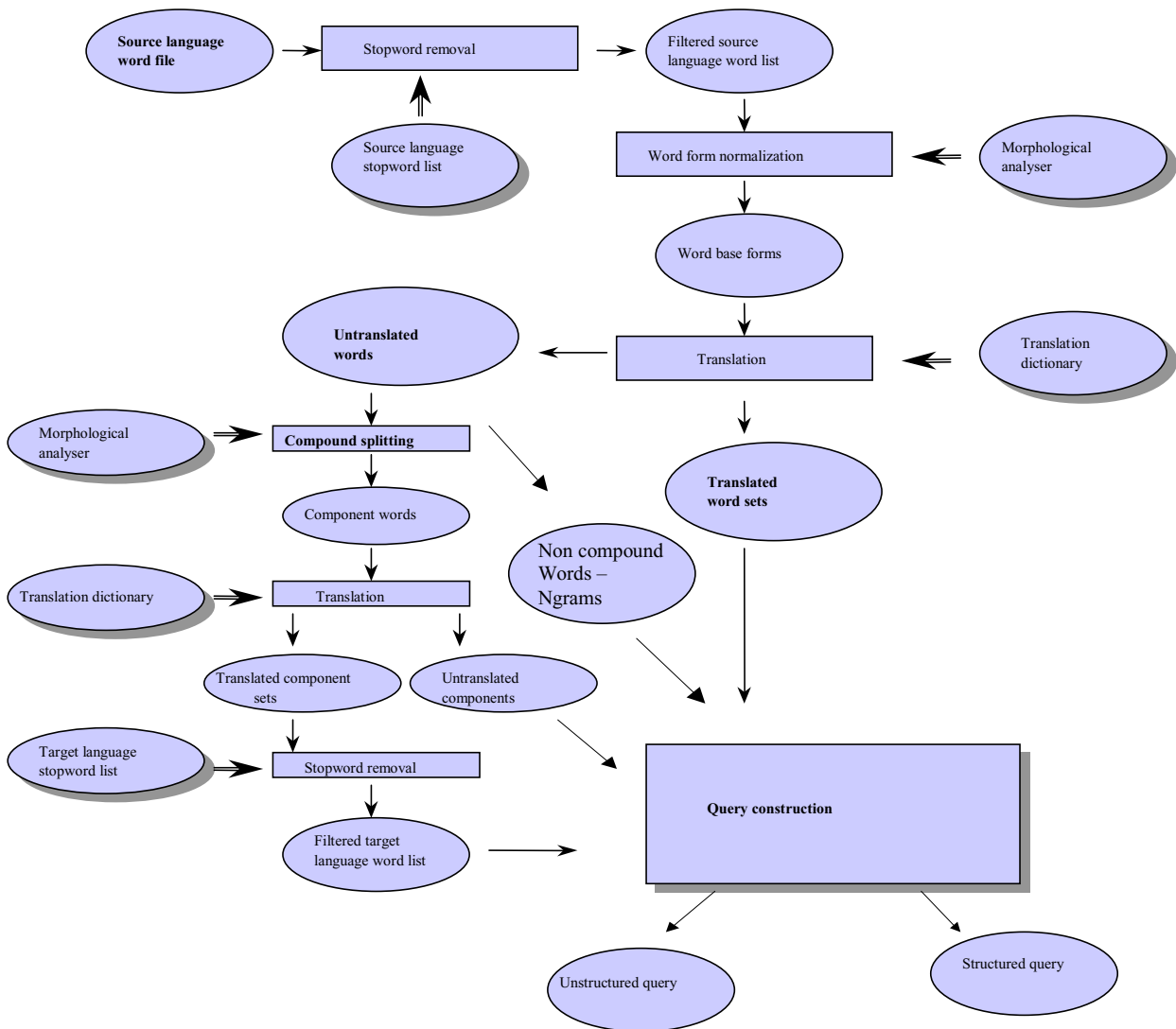


Figure 4. The main elements of the automatic query construction system

9. Discussion and conclusions

In the literature review of this paper we considered the main problems associated with dictionary-based CLIR, which are (1) untranslatable search keys due to the limitations of general dictionaries, (2) the processing of inflected words, (3) phrase identification and translation, and (4) lexical ambiguity in source and target languages. Many of the effective methods developed by the research community to handle such problems have been discussed in this paper. The techniques involve the use of special dictionaries to extend the base of translatable words (Pirkola 1998, Pirkola 1999), stemming and morphological analysis to handle inflected words (Hull 1996, Krovetz 1993, Porter

1980), part-of-speech tagging for phrase translation (Ballesteros and Croft 1997) and for removing bad translation equivalents (Ballesteros and Croft 1998, Davis 1997, Davis and Ogden 1997), corpus-based disambiguation methods (Ballesteros and Croft 1997, Ballesteros and Croft 1998, Chen et al. 1999, Davis 1997, Fujii and Ishikawa 2001), and query structuring for the ambiguity problem (Ballesteros and Croft 1998, Hedlund et al. 2001b, Hull 1997, Pirkola 1998, Pirkola 1999, Sperer and Oard 2000).

We reported recent findings of CLIR research done at the University of Tampere. In summary, our main findings suggest that:

- Query structuring through synonym sets is a simple and essential tool for dictionary-based CLIR effectiveness; query structuring performs disambiguation indirectly
- The parallel use of general and special dictionaries improves effectiveness; in different types of collections, e.g., domain specific collections, the sequential application of dictionaries may perform better
- Word by word translation of natural language request sentences yields performance comparable to (or better than) that achieved by selecting source keys and phrases; languages rich in compounds have an additional advantage of source language compounds trivially suggesting target language phrases
- Proper names are generally not translatable and may pose matching problems due to differing spelling and inflection; proper names thus require other processing techniques such as those based on n-grams
- When compound word components are in inflected form or are joined together by special morphemes (fogemorphemes), compound splitting must recognize the correct base form of the components; otherwise component translation is endangered.

Other CLIR related research done at UTA involves morphological typology research (Pirkola 2001b), the studies of automatic assessment of keywords (Pirkola and Järvelin 2001a), and the development of Interactive Query Performance Analyser (QPA, see the article by Sormunen et al. 2001 in this issue). QPA serves IR research and instruction. The system automatically analyses and compares the performance of individual queries. QPA includes several bilingual translation

dictionaries as well as morphological programs, and allows thus to perform CLIR queries. It is possible to choose, for example, automatic query translation from English to Finnish.

The *morphological typology* research provides a set of computable variables for characterizing the morphological features of natural languages from the CLIR perspective, and to be used in CLIR research in system development and evaluation. We are experimenting with different languages in our CLIR research project and have the opportunity to study the utilization of the variables. The morphological IR typology will be complemented with semantic and syntactic IR typologies. In this way a more complete picture of linguistic differences between languages will be achieved.

Our statistical studies of automatic assessment of keywords indicate that the use of *average term frequency* is useful in IR. For monolingual IR, we have devised a method, based on average term frequency, which automatically and with good reliability identifies the most important keys of requests. We have shown that structural weighting of the most important keys in automatically structured queries will improve query performance. Average term frequency could be exploited in two ways in CLIR: to identify the most important search keys of a request as in monolingual retrieval (Pirkola and Järvelin 2001a) and to prune out the keys of low average term frequency, which probably are bad keys (Pirkola and Järvelin 2001b). As discussed in this paper, a standard method in dictionary-based CLIR is to replace each source language key by all of its target language equivalents given by a dictionary. The number of irrelevant keys in a target language is usually high. Therefore, the utilization of average term frequency seems to apply particularly for CLIR.

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 Arto Voutilainen and Juha Heikkilä.

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

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

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

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

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

This research is part of the research project *Query structures and dictionaries as tools in concept-based and cross-lingual information retrieval* funded by the Academy of Finland (Research Projects 44703; 49157). This research was completed while the last author was on a leave at the Department of Information Studies, University of Sheffield, UK, fall 2000.

References

Abu-Salem H, Al-Omari M and Evens MW (1999) Stemming methodologies over individual query words for an Arabic information retrieval system. *Journal of the American Society for Information Science*, 50(6): 524-529.

Akmajian A, Demers R, Farmer A and Harnish R (1990) *Linguistics: an introduction to language and communication*. The MIT Press, Cambridge, MA.

Alkula R (2000) *Merkkijonoista suomen kielen sanoiksi*. PhD Dissertation. University of Tampere, Department of Information Studies. *Acta Universitatis Tamperensis* 763. [in Finnish]

Allan J, Callan J, Croft B, Ballesteros L, Broglio J, Xu J and Shu H (1997) INQUERY at TREC 5. In: The Fifth Text REtrieval Conference (TREC-5), Gaithersburg, MD.

Available at: http://trec.nist.gov/pubs/trec5/t5_proceedings.html

Attar R and Fraenkel AS (1977) Local feedback in full-text retrieval systems. *Journal of the Association for Computing Machinery*, 24(3): 397-417.

Ballesteros L and Croft WB (1996) Dictionary-based methods for cross-lingual information retrieval. In: *Proceedings of the 7th International DEXA Conference on Database and Expert Systems Applications*, pp. 791-801.

Ballesteros L and Croft WB (1997) Phrasal translation and query expansion techniques for cross-language information retrieval. In: *Working Notes of the AAAI Spring Symposium on Cross-Language Text and Speech Retrieval*, Stanford, CA.

Available at: <http://www.ee.umd.edu/medlab/filter/sss/papers/>

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*, Melbourne, Australia, pp. 64-71.

Bosch A van den and Daelemans W (1999) Memory-based morphological analysis. In: *Proceedings of the 37th Annual Meeting of the Association for Computational Linguistics*, University of Maryland, MA, pp. 285-292.

Buckley C, Singhal A, Mitra M and Salton G (1996) New retrieval approaches using SMART: TREC-4. In: *The Fourth Text REtrieval Conference (TREC-4)*, Gaithersburg, MD.

Available at: http://trec.nist.gov/pubs/trec4/t4_proceedings.html

Chen H-H, Huang S-J, Ding Y-W and Tsai S-C (1998) Proper name translation in cross-language information retrieval, In: *Proceedings of the 36th Annual Meeting of the Association for Computational Linguistics*, pp. 232-236.

Chen H-H, Bian G-W and Lin W-C (1999) Resolving translation ambiguity and target polysemy in cross-language information retrieval, In: Proceedings of the 37th Annual Meeting of the Association for Computational Linguistics, University of Maryland, MA, pp. 215-222.

Collins Cobuild English Dictionary (1998). HarperCollins Publishers.

Davis M and Dunning T (1996) A TREC evaluation of query translation methods for multi-lingual text retrieval. In: The Fourth Text REtrieval Conference (TREC-4), Gaithersburg, MD.

Available at: http://trec.nist.gov/pubs/trec4/t4_proceedings.html

Davis M and Ogden W (1997) QUILT: implementing a large-scale cross-language text retrieval system. In: Proceedings of the 20th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval, Philadelphia, PA, pp. 92-98.

Davis M (1997) New experiments in cross-language text retrieval at NMSU's Computing Research Lab. In: The Fifth Text REtrieval Conference (TREC-5), Gaithersburg, MD.

Available at: http://trec.nist.gov/pubs/trec5/t5_proceedings.html

Ekmekcioglu FC and Willett P (2000) Effectiveness of stemming for Turkish text retrieval. Program, 34(2): 195-200.

Fujii A and Ishikawa T (2001) Japanese/English cross-language information retrieval: exploration of query translation and transliteration. Computers and the Humanities, to appear.

Grefenstette G (1998) The problem of cross-language information retrieval. In: G Grefenstette, ed. Cross-Language Information Retrieval. Kluwer Academic Press, pp. 1-9.

Harman D (1991) How effective is suffixing? Journal of the American Society for Information Science, 42(1): 7-15.

Hedlund T, Pirkola A and Järvelin K (2001a) Aspects of Swedish morphology and semantics from the perspective of mono- and cross-language information retrieval. Information Processing & Management, 37(1): 147-161.

Hedlund T, Keskustalo H, Pirkola A, Sepponen M and Järvelin K (2001b) Bilingual tests with Swedish, Finnish and German queries: dealing with morphology, compound words and query structure. Lecture Notes in Computer Science by Springer, to appear.

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

Hull D and 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, pp. 49-57.

Hull D (1997) Using structured queries for disambiguation in cross-language information retrieval. In: *Working Notes of the AAAI Spring Symposium on Cross-Language Text and Speech Retrieval*, Stanford, CA.

Available at: <http://www.ee.umd.edu/medlab/filter/sss/papers/>

Jing Y and Croft WB (1994) An association thesaurus for information retrieval, Technical Report UMASS-CS-94-17, University of Massachusetts.

Kalamboukis TZ (1995) Suffix stripping with modern Greek. *Program*, 29(3): 313-321.

Karlsson F (1998) Yleinen kielitiede [General linguistics]. Yliopistopaino, Helsinki [In Finnish].

Kekäläinen J and Järvelin K (1998) The impact of query structure and query expansion on retrieval performance. In: *Proceedings of the 21st Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, Melbourne, Australia, pp. 130-137.

Kilgarriff A (1993) Dictionary word sense distinctions: an enquiry into their nature. *Computers and the Humanities*, 26: 365-387.

Krovetz R (1993) Viewing morphology as an inference process. In: *Proceedings of the 16th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, Pittsburg, PA, pp. 191-203.

Lee JS and Choi K-S (1997) A statistical method to generate various foreign word transliterations in multilingual information retrieval system. In: Proceedings of the 2nd International Workshop on Information Retrieval with Asian languages, pp. 123-128.

Lyons J (1984) Language and linguistics: an introduction. Cambridge University Press.

Matthews PH (1997) The concise Oxford dictionary of linguistics. Oxford University Press, Oxford, New York.

Oard D and Dorr B (1996) A survey of multilingual text retrieval. Technical Report UMIACS-TR-96-19, University of Maryland, Institute for Advanced Computer Studies.

Peters C and Picchi E (1997) Using linguistic tools and resources in cross-language retrieval. In: Working Notes of the AAAI Spring Symposium on Cross-Language Text and Speech Retrieval, Stanford, CA.

Available at: <http://www.ee.umd.edu/medlab/filter/sss/papers/>

Peters C (2000) CLEF - Cross-Language Evaluation Forum.

<http://galileo.iei.pi.cnr.it/DELOS/CLEF/clef.html>

Pfeifer U, Poersch T and Fuhr N (1996) Retrieval effectiveness of proper name search methods. Information Processing & Management, 32(6): 667-679.

Pirkola A (1998) The effects of query structure and dictionary setups in dictionary-based cross-language information retrieval. In: Proceedings of the 21st Annual International ACM Sigir Conference on Research and Development in Information Retrieval, Melbourne, Australia, pp. 55-63.

Pirkola A (1999) Studies on linguistic problems and methods in text retrieval: the effects of anaphor and ellipsis resolution in proximity searching and translation and query structuring methods in cross-language retrieval. PhD Dissertation, University of Tampere, Department of Information Studies, Acta Universitatis Tamperensis 672.

Pirkola A (2001a) The effects of part-of-speech and morphological disambiguation in Finnish-English cross-language retrieval. Manuscript.

Pirkola A (2001b) Morphological typology of languages for IR. *Journal of Documentation* 57 (3).

Pirkola A, Keskustalo H and Järvelin K (1999) The effects of conjunction, facet structure, and dictionary combinations in concept-based cross-language retrieval. *Information Retrieval*, 1(3): 217-250.

Pirkola A, Hedlund T, Keskustalo H and Järvelin K (2000) Cross-lingual information retrieval problems: methods and findings for three language pairs. ProLISSa Progress in Library and Information Science in Southern Africa. First biannual DISSAnet Conference. Pretoria, 26-27 October 2000.

Pirkola A and Järvelin K (2001a) Employing the resolution power of search keys. *Journal of the American Society for Information Science* 52 (8).

Pirkola A and Järvelin K (2001b) Exploiting average term frequency and word distribution statistics in text retrieval. Submitted.

Popovic M and Willett P (1992) The effectiveness of stemming for natural-language access to Slovene textual data. *Journal of the American Society for Information Science*, 43(5): 384-390.

Puolamäki D, Pirkola A and Järvelin K (2001) Applying query structuring in cross-language retrieval. Submitted.

Porter MF (1980) An algorithm for suffix stripping. *Program*, 14: 130-137.

Popovic M and Willett P (1992) The effectiveness of stemming for natural-language access to Slovene textual data. *Journal of the American Society for Information Science*, 43(5): 384-390.

Robertson AM and Willett P (1998) Applications of n-grams in textual information systems. *Journal of Documentation*, 54(1): 48-69.

Salton G (1989) Automatic text processing: the transformation analysis and retrieval of information by computer. Addison-Wesley.

Savoy J (1999) A stemming procedure and stopword list for general French corpora. *Journal of the American Society for Information Science*, 50(10): 944-952.

Sheridan P and Smeaton AF (1992) The application of morpho-syntactic language processing to effective phrase matching. *Information Processing & Management*, 28(3): 349-369.

Sheridan P and 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, pp. 58-65.

Sormunen E, Halttunen K and Keskustalo H (2001) A query performance analyser - a tool for information retrieval research and instruction. Submitted to *Information Retrieval*.

Sperer R and Oard DW (2000) Structured translation for cross-language IR. In: *Proceedings of the 23rd Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, Athens, Greece, pp. 120-127.

Strzalkowski T (1995) Natural language information retrieval. *Information Processing & Management*, 31(3): 397-417.

Turtle HR (1990) Inference networks for document retrieval. Ph.D. dissertation. Computer and information Science Department, University of Massachusetts. COINS Technical Report 90-92.

Whaley LJ (1997) *Introduction to typology: the unity and diversity of language*. Thousand Oaks - London - New Delhi, Sage Publications.

Yamabana K, Muraki K, Doi S and Kamei S (1996) A language conversion front-end for cross-linguistic information retrieval. In: *Working Notes of the Workshop on Cross-Linguistic Information Retrieval*, ACM SIGIR, Zürich, Switzerland, pp. 34-39.

Zhai C, Tong X, Milic-Frayling N and Evans DA (1997) Evaluation of syntactic phrase indexing - CLARIT NLP track report. In: The Fifth Text REtrieval Conference (TREC-5), Gaithersburg, MD. Available at: http://trec.nist.gov/pubs/trec5/t5_proceedings.html

Zobel J and Dart P (1995) Finding approximate matches in large lexicons. *Software - practice and experience*, 25(3): 331-345.