Eija Airio

THE EFFECTS OF SEPARATE AND MERGED INDEXES AND WORD NORMALIZATION IN MULTILINGUAL CLIR

UNIVERSITY OF TAMPERE
DEPARTMENT OF INFORMATION STUDIES
Research Notes
RN ● 2005 ● 1

Abstract. Multilingual IR may be performed in two environments: there may exist a separate index for each target language, or all the languages may be indexed in a merged index. In the first case, retrieval must be performed separately in each index, after which the result lists have to be merged. In the case of the merged index, there are two alternatives: either to perform retrieval with a merged query (all the languages in the same query), or to perform distinct retrievals in each language, and merge the result lists. Further, there are several indexing approaches concerning word normalization. The present paper examines the impact of stemming compared with inflected retrieval in multilingual IR when there are separate indexes / a merged index. Four different result list merging approaches are compared with each other. The best result was achieved when retrieval was performed in separate indexes and result lists were merged. Stemming seems to improve the results compared with inflected retrieval.

Keywords: Multilingual information retrieval; result list merging; index merging; stemming.

1 Introduction

Cross-language information retrieval (CLIR) studies retrieval across language barriers. In CLIR, the search request is formulated in a language differing from document languages(s). There are two kinds of CLIR tasks: bilingual and multilingual tasks. In a bilingual task, there is only one document language (called also a target language), while in a multilingual task there are several target languages. Bilingual IR suits a user who is capable to read documents in the target language (or who is able to receive translation help), but for whom it is difficult to express a proper query in the target language. Multilingual IR suits a person capable to read multiple languages: he needs to formulate his query in one language only and perform retrieval only once, instead of formulating queries in multiple languages and performing multiple retrievals. The number and the use of multilingual indexes have grown recently because of the growth of the Internet. Multilingual information retrieval – presenting the request in one language and retrieving documents in multiple languages - is not a mainstream yet. However, multilingual IR would be advantageous for Internet users, because many of them are capable to read multiple languages. Thus research on CLIR, especially multilingual IR, would be valuable, and possibilities to apply CLIR in Internet environments should be tested.

Multilingual IR may be performed in two environments. There may exist separate indexes for separate target languages, or there may be one index only for all the document languages. In the first case, retrieval must be performed in each index separately, and the result lists have to be merged. In the latter case, it is possible to perform a multilingual retrieval or separate retrievals in all the target languages. There are naturally various indexing and translating methods which may be applied in both cases. There is more research on the approach of separate indexes than on the merged index approach. The studies on separate indexes with result list merging show that it is difficult to develop a new result list merging approach which would give significantly better results than simple basic methods (see Callan & al. 1995, Savoy 2002, Chen 2003 and Braschler & al. 2002). Previous research on retrieval in a merged index has given contradictory results. Nie and Jin reported (2002) poor results in the merged index compared with retrieval in separate indexes with result list merging, while Chen's results with the merged index were only slightly worse than those given by separate indexes (see Nie & Jin 2002 and Chen 2002). Thus, because of lack of studies on a merged index, as well as contradictory results, it is reasonable to direct research on this area.

Word inflection is a feature of most natural languages. Depending on the language, nouns, verbs and adjectives, or all of them, inflect. Languages differ in the degree of inflection (see Pirkola 2001). Word inflection has an impact on IR as well, because search requests and documents include natural language words. Various word normalization methods have been utilized in IR to decrease the impact of word inflection. The word normalization methods may be divided roughly in two groups: lemmatizers and stemmers. Lemmatizers return the base form of a word, while stemmers return a stem, for example a truncated word, which is not necessarily a lexical word of the language. Previous research shows that word normalization mostly benefits monolingual non-English retrieval (see Alkula 2001, Kettunen 2004, Popovic & Willet 1992, Braschler Ripplinger 2004, Hollink & al. 2004 and Airio 2005), but the advantages are not remarkable in English retrieval (see Harman 1991, Popovic & Willet 1992, Krovetz 1993 and Airio 2005).

There are few studies about the impact of word normalization on bilingual IR. Decompounding in the indexing phase seems to improve the retrieval result if the target language is a compound-rich language (see Airio 2005). In each target language, the effect of normalization compares to the effect in bilingual IR. It is however interesting to see the co-effects of normalization and merging.

The present research studies the impact of target language stemming on multilingual IR, compared with inflected word form retrieval, as well as two different approaches of multilingual IR: retrieval in separate indexes and retrieval in a merged index. Different result list merging strategies are compared with each other as well.

The structure of the paper is following. Section 2 presents the main approaches in cross-language information retrieval and stemming, as well as results list merging and index merging. Research questions, as well as resources, methods and runs are introduced in Section 3. Results are presented in Section 4 and section 5 includes discussion. Finally, conclusions are presented in Section 6.

2 CLIR, stemming and merging methods

2. CLIR

In CLIR, it is possible to translate search requests into the document language(s), or to translate documents into the source language. The latter alternative would be comfortable for the user, but it is difficult and expensive to implement. Translation of documents is not possible to perform wholly automatically, but manual resources are needed. In addition, the document collection should be re-translated for each possible source language separately. Thus, it is easier and cheaper to translate search requests, and that is the more common CLIR approach, as well.

There are various query translation approaches: corpus-based, machine translation and the dictionary-based approach. In the **corpus-based approach** a probabilistic dictionary is derived from parallel corpora. Probabilistic dictionaries are often narrow because of domain dependence of parallel corpora. They are suitable especially for translating special terminology. (Kraaij 2004, 125.) Probabilistic dictionaries may include inflected word forms, and are thus suitable for translating queries for retrieval in inflected word form indexes.

Machine translation (MT) is easy to carry out. However, MT tools are often very expensive, and they are not available for all the language pairs. MT systems give only one translation variant for each source word, which is not advantageous for IR purposes. The **dictionary-based approach** does not have the disadvantages of MT. Machine-readable dictionaries are often free or low-cost, and they give several translations for a single word, which is advantageous for IR. (Kraaij 2004, 124-125.) However, words not present in the dictionary, for example geographical names, cause problems: they are often crucial query words, and losing then may ruin the result.

2.2 Stemming

A **stemmer** is a word normalization tool, which returns a "stem" for each word. The stem is not necessarily any lexical word of a language, but it may be for example a truncated form of a word. Stemming is used for different purposes, and there are various stemming approaches. Some stemming algorithms utilize a suffix list and others a stem dictionary. Many stemming algorithms, whose purpose is to improve IR performance, do not use a stem dictionary, but an explicit list of suffixes, and the criteria for removing suffixes. Stemmers of the perhaps most popular stemmer family today, the Porter stemmers, have adopted this approach. (Porter 1980.)

The very first stemmers were simple: they just stripped off the word endings. For example Lovins created principles for developing stemming algorithms in 1969 (Koskenniemi 1983, 12). The idea of stemming may be illustrated by giving the sample *connect, connected, connecting, connection and connections*. These words have a similar meaning, and it would be reasonable to stem them to a common form. If suffixes *ed*, *ing*, *ion* and *ions* are removed, the stem will be *connect* for all these words. (Porter 1980.)

Stemmers may act in a wrong way from the IR point of view. In **under-stemming** the stemmer removes too small a suffix, for example removing only *s* from the word *babies*. **Over-stemming** is the opposite phenomenon of under-stemming: too a long suffix is removed. **Mis-stemming** happens when the stemmer mis-interprets a part of a word stem to be a suffix. For example suffix *ly* shoud be removed from the word *cheaply*, but not from the word *reply* (Porter 1981).

2.3 Result list merging

Result list merging has been applied in monolingual IR as well as in multilingual IR. The purpose of result list merging is to select from the result lists those documents, which most probably are relevant. There have not been recent breakthroughs in result list merging field: the simplest approaches always seem to be competitive comparing with the new complex ideas. The simplest result list merging approach is the **round robin approach**. A document from every result list is taken by turn. The approach is based on the idea that document scores are not comparable across the collections. If one is ignorant about the distribution of the relevant documents in the retrieved lists, it is reasonable to assume the distribution to be symmetric. (Hiemstra & al. 2001, 108.)

The **raw score approach** bases on the idea of comparable scores across collections: the result lists are merged straightforwardly according to the scores (Hiemstra & al. 2001, 108).

The **normalized score approach** tries to avoid the weakness of the raw score approach by normalizing the scores. The simplest normalizing approach is to divide each score by the maximum score of the topic on the current list. This normalization approach might be advantageous when merging result lists produced by diverse search engines, where the scoring scale varies from list to list. When the result lists are from a single search engine, this approach does not bring any benefit: it simply favours scores which are near the best score of the topic on the list. In the case of a single search engine, normalization could be based for example on collection sizes.

In the **weighted score approach**, weights are based upon document's score and / or the collection ranking information. If the collections are assumed to be different, the collection's score might be used in weight calculation. The collection's score bases on the idea of a collection retrieval inference network (CORI): the query is first used to retrieve a ranked list of collections, and collection scores are given on the basis of this list. (Callan & al. 1995, 22-25). In multilingual retrieval, this should be performed by using translated queries.

2.4 Research on result list merging

Callan and colleagues tested in 1995 several merging methods with monolingual TREC material. Distinct indexes were built for every 17 collections. The normalized score approach was the baseline, because it is equivalent to the single database paradigm. The round robin approach proved to produce the worst results. The raw score approach was much better than the round robin approach, but substantially worse than ranking based on normalized document scores. The weighted score approach produced approximately as good results as the normalized score approach. The results suggest that the quite simple normalized score approach is as accurate as the weighted score approach with its higher computational effort. (Callan & al. 1995, 25-27.)

Voorhees and others (1995) investigated the ability of result merging strategies to learn from the past queries utilizing TREC test collection. A combined index was built, as well as separate indexes. The averaged precision of the run with separate indexes and result merging with round robin approach was about 10 % of the precision of the single collection run. Training queries were used to build a model distribution of relevant documents in separate collections. The effect of the new fusion strategies is dependent on the training set size. The test showed that the fusion techniques can learn to distinguish collections according to the subjects they cover. (Voorhees & al 1995, 172-178.)

Savoy applied the round robin (baseline), the raw score, the CORI and the normalized score strategies in CLEF 2002 runs. When calculating normalized scores, Savoy utilized both the best score and the worst score of each topic. He tested the performance of these strategies both for manually translated and automatically translated queries, with different retrieval systems. The best performance was achieved when utilizing the normalized score approach. The change of mean averages with the raw score and the CORI approaches were between –28.14 % and –45.30 % from the baseline. (Savoy 2002, 39).

Chen tested in 2003 two different result list merging strategies in CLEF 2002. The first strategy was the raw score strategy: combining the ranked lists and sort the list by the raw relevance score. The second strategy was similar to the first one, but the relevance scores were normalized before sorting. The normalizing approach which Chen applied differed from Savoy's approach: normalizing was done simply by dividing the score by the best score of the topic. The average precision of the multilingual CLEF 2002 run was 0.38 using the raw score method, and 0.36 with the normalized score method. (Chen 2003, 45-46).

Braschler and colleagues tried new merging strategies in CLEF 2002. They developed an approach, which they called the collection size based round robin approach. If the collections vary in size, it is reasonable to take this point into consideration when merging results. The round robin approach can be adjusted so, that the portion of result corresponding the collection size is taken from every result set. Their second approach was the feedback merging strategy, which bases on an initial retrieval step. The top ranked documents of the result set were analysed, and an ideal query was built upon them. The ideal query was then compared with the original query: the overlap indicates the degree to which the concepts of the original query are represented in the retrieval result. The result lists were merged in proportion based on the similarity estimates. The new methods seemed to choose different ratios for individual queries than other merging methods, but the result did not differ much from the results of traditional merging methods. (Braschler & al. 2002, 130).

2.5 Index merging

In the multilingual merged index, all the documents are indexed into a single index, ignoring the language. Internet indexes are the most familiar merged indexes. There may occur weighting problems when retrieving with a multilingual query in a merged index. The problem might be severe if the number of documents varies between languages. The

terms of a small dataset might be over-weighted, because the weighting schema favours terms with small frequency in the whole index, but large count in individual documents. (Lin & al. 2002, 99).

It is possible that two or more languages share the same word or word form, possibly with separate meaning. To avoid wrong matches, it would be appropriate to add an extra field *language* into the documents, for example <LANGUAGE>fin</LANGUAGE>. The same field notation should be used in queries as well. Another way to use the language code would be to attach it in the indexing phase with each index word. (Nie 2002, 12). For example, the ending *fi* could be inserted with each Finnish index word: *informaatio_fi*. The same ending should be inserted with each query word.

2.6 Research on CLIR in a merged index

Nie and Jin stated in 2002 that the result merging approach is difficult and causes loss of effectiveness. Separating documents in different languages is also an artificial solution, because they in fact co-exist often in the same collection. They suggest the collection merging approach instead. They built a merged collection for different languages preserving the language code in the index. For example they added *_f* for every French index word: French word *chaise* was indexed as *chaise_f*. Similarly, when topics were translated, the language code was inserted to the translated words. The result of CLEF 2002 runs with this merged index was not very promising. Very often one language dominated the result. According to the authors possible reasons could be one of following: 1) the size and coverage of parallel corpora used for translation model training differs between languages; 2) the weights attributed to original query words were not reasonable; or 3) translation is still made independently from retrieval, when they should be considered together. (Nie & Jin 2002, 59-60).

Chen reports on his experiments in result merging and index merging with CLEF 2001. First he made separate indexes for all the languages utilizing stemmers. Retrieval was performed from each index separately, and results were merged according to adjusted scores. Chen made multilingual runs also utilizing a merged index. The translated topics were combined with the English topic. The retrieval was performed using multilingual queries. The average precision of this run was 31.2 %, when title, description and narrative fields were utilized, while it was 34.2 % with separate result lists merged utilizing the weighted score approach. (Chen 2002, 55-57).

3 Research questions, resources, and methods

Research on both multilingual CLIR settings, retrieval in separate indexes and retrieval in a merged index, is needed: usually, indexes are not built for the purpose of CLIR, but CLIR systems are add-ons for existing systems, which may be systems of separate indexes for various languages or systems of a merged index. Even if the earlier results show that retrieval in a merged index gives worse results than retrieval in separate indexes, the first one should be studied. The impact of target word normalization on multilingual IR is not widely examined either.

The current research tries to find out, whether retrieval in a merged index gives comparable results with retrieval in separate indexes. The effect of index and target word stemming was examined as well, and the impact of the language code in the stemmed merged index was tested. Our source language was English, and target languages were English, Dutch, Finnish, French, German, Italian, Spanish and Swedish. We translated the English topics into all the target languages (except English). In order to test the approach of separate indexes, we performed retrieval in eight monolingual indexes with translated queries, and merged the result lists applying various merging approaches. This was performed separately for inflected and stemmed indexes / queries. For testing merged indexes, we created two multilingual indexes: a stemmed index and an inflected word form index. Retrieval in these indexes was performed in two different ways: by formulating merged queries consisting of translations, and by performing retrievals using monolingual translated queries and merging the result lists.

3.1 Research questions

Research on the settings of separate indexes with result list merging has given contradictory results: in some tests the normalized score approach has given the best result, in others the raw score approach has been the best (see Callan & al. 1995, Savoy 2002, Chen 2003 and Braschler & al. 2002). Thus, it is reasonable to test the impact of various merging approaches.

Target word normalization in multilingual CLIR has not been studied widely. It is interesting to know, whether improvements could be attained via introducing target word normalization in either of the CLIR environments.

Thus, following research questions are justified:

1. In the setting of *separate inflected* word form indexes, which of the four merging approaches gives the best result?
2. In the settings of *separate stemmed* indexes, which of the four merging approaches gives the best result?
3. In an *inflected merged* index, which gives better result: retrieval with a merged multilingual query, or separate retrievals in the target languages with result list merging applying the merging approach which gave the best result in the question 1?
4. In a *stemmed merged* index, which gives the best result: retrieval with a merged multilingual query, retrieval with a merged multilingual query with the language code, or separate retrievals in the target languages with result list merging applying the merging approach which gave the result in the question 2?
5. Which gives the best result, when the best runs of the question 1, 2, 3 and 4 are compared with each other?

3.2 Language resources and collections

In this section, we describe the language resources and collections used in this research.

The following language resources were used in the tests:

- Motcom GlobalDix multilingual translation dictionary (18 languages, total number of words 665 000, 25 000 Dutch entries, 44 000 English entries, 26 000 Finnish entries, 30 000 French entries, 39 000 German entries, 32 000 Italian entries, 35 000 Spanish entries, 36 000 Swedish entries) by Kielikone plc. Finland
- Snowball stemmers for English, German, Finnish and Swedish, by Martin Porter
- Stemmers for Spanish and French, by ZPrise
- A stemmer for Italian, by the University of Neuchatel
- A stemmer for Dutch, by the University of Utrecht
- An English lemmatizer ENGTWOL, Lingsoft plc. Finland
- English stop word list (429 stopwords), created on the basis of InQuery's default stop list for English

The Snowball stemmers used in the tests are algorithmic and simple. They do not utilize any dictionaries or exception lists. (Porter 1981). The Spanish stemmer strips off plurals and removes inflectional suffixes of adjectives and verbs. The French stemmer removes plural and feminine off French words, and the Italian stemmer removes plural and accent off Italian words. The Dutch stemmer is an implementation of the Porter stemmer in Dutch: it bases on a suffix stripping algorithm covering Dutch morphology.

CLEF 2003 datasets (Dutch, English, Finnish, German, Italian, Spanish and Swedish) were used for the tests (see Table 1).

Table 1. CLEF 2003 datasets

Collection language	**Source**	**Number of documents**	**Percentage of documents**	**Size of the corpus (MB)**
Dutch	Algemen Dagblad 1994-1995 NRC Handelsblad 1994-1995	190 604	12.3	540
English	Los Angeles Times 1994 Glasgow Herald 1995	169 477	10.9	579
Finnish	Aamulehti 1994-1995	55 344	3.6	137
French	ATS 1994-1995 Le Monde 1994	85 793	5.5	174
German	Rundschau 1994 Der Spiegel 1994-1995 SDA German 1994-1995	294 809	19.0	668
Italian	La Stampa 1994 AGZ 1994-1995	157 558	10.2	364
Spanish	EFE 1994-1995	454 045	29.3	1 088
Swedish	Tidningarnas Telegrambyrå	142 819	9.2	352

	1994-1995			
Altogether		1 550 449	100.0	3 902

We utilized CLEF 2003 English topics (60 topics in 2003) and relevance assessments in the tests. The *InQuery* system, provided by the Center for Intelligent Information Retrieval at the University of Massachusetts, was utilized in indexing the databases and as the retrieval system.

3.3 Translation and query formulation

The **UTACLIR** query translation system of University of Tampere was used in the test. The system utilizes external language resources (translation dictionaries, stemmers, lemmatizers and stop word lists). Word processing in UTACLIR proceeds as follows. In order to match topic words with the dictionary words, the source language lemmatizer is utilized. Without source word lemmatization, words in their inflected forms are not translated, because translation dictionaries contain only basic form words. The lemmatizer produces one or more basic forms for a token. After normalization, stop words are removed, and non-stop words are translated. If translation equivalents are found, they are normalized utilizing a lemmatizer or a stemmer, depending on the target index. For untranslatable words, two highest ranked words obtained by n-gram matching from the target language index are selected as query words.

Queries are structured utilizing a synonym operator (see Pirkola 1998): the target words derived from the same source word are grouped into the same synonym group. (Airio & al. 2003, 92-93.)

In the current tests, the UTACLIR system was used in the following way. The source language was English, and the **topic words were lemmatized with the English lemmatizer**. Then the lemmatized English words were translated into each target language in turn. When creating queries for the **stemmed indexes**, the translated words were **stemmed with the appropriate stemmer**. Creating queries for the **inflected word form indexes** was not so simple, because we had no word form generators in use. Thus, we simply took **the translated word**, which is most often in the basic form, **as such with the query**. This solution naturally causes loss of recall, because documents including words in their inflected forms only are lost. This was the only possible solution, however, in our case. Tools generating inflected word forms are rare, and this is one of the bottlenecks of multilingual non-normalized retrieval. For **untranslatable** words, **n-gramming** techniques were applied for both query types, stemmed and inflected.

Monolingual English queries were constructed analogously with the translated ones. The English stemmer was applied when creating the stemmed queries. For creating the queries for the inflected word form indexes, we applied the English lemmatizer for word normalization.

3.4 Indexes

We had two separate index settings, where we wanted to test the result list merging approach: the setting of stemmed indexes and the setting of inflected word form indexes. For both of these, eight indexes were built, which makes altogether 16 indexes.

Two merged indexes were built: the stemmed index and the inflected word form index. When creating the stemmed index, eight separate stemmers were utilized (as in the result merging approach), because each language has the stemmer of its own. After stemming, the final merged stemmed index was created. A document specific language code field was included in the stemmed index. For making that possible, a language code tag was inserted into the documents before indexing them. For example the following tag was inserted into each Finnish document: *<LANGUAGE>finnish</LANGUAGE>*.

The word tokenization rules used in indexing were following. First, punctuation marks were converted into spaces. Next, strings broken down by the space character were decoded to be indexable words. Capitals were converted into lower case letters before indexing. No stop word removal was performed in indexing.

3.5 Result list merging

Four result list merging approaches were tested for merging the result lists of retrievals in separate indexes: the round robin approach, the raw score approach, the collection size based approach and the normalized score approach based on the collection sizes.

1. **The round robin approach**. We had eight result lists, each including 1000 documents per a topic, and 1000 documents for each topic were selected for the final merged list. Thus, 125 first documents for each topic were taken from each result list.

2. **The raw score approach**. The documents corresponding the topic number in turn were first merged, and then sorted in descending order (along the score). Finally, 1000 first documents were taken as the result.
3. **The size based approach**. For each language, the number of documents present in the merged list was calculated upon the size of the database. Thus, 290 Spanish, 190 German, 120 Dutch, 110 English, 100 Italian, 90 Swedish, 60 French and 40 Finnish documents were selected for each topic.
4. **The normalized score approach based on collection sizes**. The idea of this normalization approach is that the effect of the known normalization factors in the original scores should be minimized. That was performed using the score formula inversely. The INQUERY score is calculated in the following way (see Allan & al. 1997):

$$0.4 + 0.6 \times (tf_{ij} / (tf_{ij} + 0.5 + 1.5 \times (d_{lj} / adl)) \times (\log ((N + 0.5) / df_i) / \log (N + 1.0)))$$

where tf_{ij} = frequency of search key *i* in document *j*
df_i = the number of documents that contain the search key *i*
dl_j = the length of the document *j*
adl = average document length in the collection
N = number of documents in the collection

The second factor of the product contains numbers, which are hard to calculate (tf_{ij} and dl_j).The third factor of the product is constructed of a division. The numerator of the division again contains a number which we do not know (df_i). If the unknown numbers (the second factor of the product and numerator of the third factor) in the product are marked with x and y, the formula may be introduced in a following way:
0.4 + 0.6 *x * y / log(N+1.0), where N is the collection size, and x and y are unknown factors from our point of view. Thus, if we want to diminish the impact of known factors in this formula, we make the following calculation: $x * y = \frac{score - 0.4}{0.6} * \log(N+1.0)$, and use x*y as a new score.

3.6 Query merging

We had two merged indexes: the stemmed index and the inflected word form index. We created merged queries for both by concatenating the queries in various languages by the sum operator of INQUERY (for INQUERY and its operators, see Callan & al. 1992). When retrieving in the stemmed index with a merged query, it was possible to apply language codes, expressed in the following way: *#FIELD(LANGUAGE=english)*. Following the idea of Nie in 2002, the language code was attached with each query word (Nie 2002, 12). For that we used INQUERY's *Boolean and* (*#band*) operator:

```
#q141 = #sum(#band(#FIELD(LANGUAGE=english) #syn(letter)) #band(#FIELD(LANGUAGE=english) #syn(bomb))
#band(#FIELD(LANGUAGE=english) #syn(kiesbauer bierbauer)) #band(#FIELD(LANGUAGE=english) #syn(find))
#band(#FIELD(LANGUAGE=english) #syn(inform)) #band(#FIELD(LANGUAGE=english) #syn(explos))
#band(#FIELD(LANGUAGE=english) #syn(letter)) #band(#FIELD(LANGUAGE=english) #syn(bomb))
#band(#FIELD(LANGUAGE=english) #syn(studio)) #band(#FIELD(LANGUAGE=english) #syn(tv))
#band(#FIELD(LANGUAGE=english) #syn(channel)) #band(#FIELD(LANGUAGE=english) #syn(pro))
#band(#FIELD(LANGUAGE=english) #syn(7)) #band(#FIELD(LANGUAGE=english) #syn(present))
#band(#FIELD(LANGUAGE=english) #syn(arabella)) #band(#FIELD(LANGUAGE=english) #syn(kiesbauer bierbauer))
#band(#FIELD(LANGUAGE=finnish) #syn(kirj)) #band(#FIELD(LANGUAGE=finnish) #syn(pom pommit))
#band(#FIELD(LANGUAGE=finnish) #syn(bauer esbau))…
```

3.7 Runs

In the setting of separate indexes, retrieval was first performed using appropriate queries, taking into account the index language and the normalization approach. These runs may be divided into **two run groups**: 1) eight monolingual runs in the inflected indexes and 2) eight in the stemmed indexes. Four different result list merging approaches were applied with these two groups, generating altogether eight result lists.

We had two merged indexes: the inflected index and the stemmed index. Two runs were performed in the inflected index. The first was a multilingual retrieval, and the second consisted of eight monolingual retrievals, whose result lists were merged applying the approach which gave the best result in the setting of separate indexes. In the stemmed merged index, we had the same two runs as in the inflected index, and in addition a run with the language code.

The abbreviations of the runs used in this section are explained in the Table 2.

Table 2. The merging runs

Run	Index type	Normalization	Result merging approach	Query type
SepInf-ROB	Separate	Inflected	Round robin	Monolingual
SepInf-SCO	Separate	Inflected	Raw score	Monolingual
SepInf-SIZE	Separate	Inflected	Size based	Monolingual
SepInf-NORM	Separate	Inflected	Normalization based on the collection sizes	Monolingual
SepStem-ROB	Separate	Stemmed	Round robin	Monolingual
SepStem-SCO	Separate	Stemmed	Raw score	Monolingual
SepStem-SIZE	Separate	Stemmed	Size based	Monolingual
SepStem-NORM	Separate	Stemmed	Normalization based on the collection sizes	Monolingual
MerInf	Merged	Inflected	-	Multilingual
MerInfSep	Merged	Inflected	The best result list merging approach	Monolingual
MerStem	Merged	Stemmed	-	Multilingual
MerStemQueryCode	Merged	Stemmed	-	Multilingual with language codes
MerStemSep	Merged	Stemmed	The best result list merging approach	Monolingual

4 Results

4.1 Separate indexes

When merging the result lists of *inflected runs*, the normalized score approach based on collection sizes gave the best result, average precision was 16.2 % (see Table 3 and Figure 1). The other three merging approaches, the round robin, the raw score and the size based approach both gave an equal result, average precision 15.6 %. The differences between the results are not significant with the Friedman's test at the level 0.05.

Table 3. Non-interpolated average precision of various merging approaches for multilingual inflected runs (source language English) for all relevant documents averaged over queries

Run	Average precision %	Diff. % (from the baseline)	Change % (from the baseline)
SepInf-ROB	15.6		
SepInf-SCO	15.6	0.0	0.0
SepInf-SIZE	15.6	0.0	0.0
SepInf-NORM	16.2	+0.6	+3.9

The normalized score approach based on collection sizes gave the best result, average precision 19.7 %, when merging *stemmed* runs (see Table 4), similarly as with inflected runs. The round robin, the raw score and the size based approach robin gave almost equal results in the stemmed runs (average precision 19.1, 19.2 and 19.3 respectively). The differences between the results are not significant with the Friedman's test at the level 0.05.

Table 4. Non-interpolated average precision of various merging approaches for multilingual stemmed runs (source language English) for all relevant documents averaged over queries

Run	Average precision %	Diff. % (from the baseline)	Change % (from the baseline)
SepStem-ROB	19.1		
SepStem-SCO	19.2	+0.1	+0.5
SepStem-SIZE	19.3	+0.2	+1.0
SepStem-NORM	19.7	+0.6	+3.1

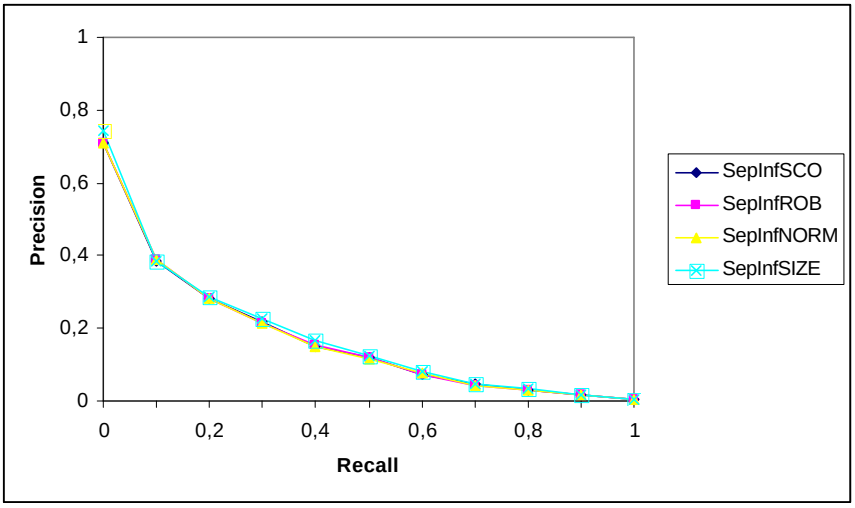

Figure 1. Interpolated average precision values for inflected runs in the setting of separate indexes

The average precision of the bilingual inflected runs varied from 8.2 % to 44.2 %. The average was 23.4 %. Among the stemmed bilingual runs the average precision varied from 20.1 % to 47.2 %, and the average was 29.0 %. Thus, it is obvious that none of the merging methods we applied was complete, because the results of the result list merging runs are much lower than those of the individual runs.

Because the results of the merging approaches are so close to each other it would be interesting to know, whether relevant documents in the merged lists are the same or not. To clarify this, the stemmed merged lists were compared with each other one by one. The comparison was performed only for the stemmed runs. We can conclude that the result would be similar for the inflected runs, because the same merging methods were applied for both list types. The comparison was performed among 100, 500 and 1000 first documents, respectively. With all these comparisons, the number of the same relevant documents varied from 79.3 % to 100 %. Thus, the lists share approximately the same relevant documents.

4.2 Merged indexes

There were only two runs in the inflected merged index: the run with the merged query, and the run where the result lists of individual retrievals were merged with the best merging approach, which was the normalized score approach based on collection sizes (see Table 5 and Figure 2). The latter gave a little better result. Differences between the runs are not statistically significant by the Wilcoxon signed ranks test at the 0.05 level.

Table 5. Non-interpolated average precision of the inflected index merging approaches (source language English) for all relevant documents averaged over queries

Run	Average precision %	Diff. % (from the baseline)	Change % (from the baseline)
MerInf	7.6		
MerInfSep	8.7	+1.1	+14.5

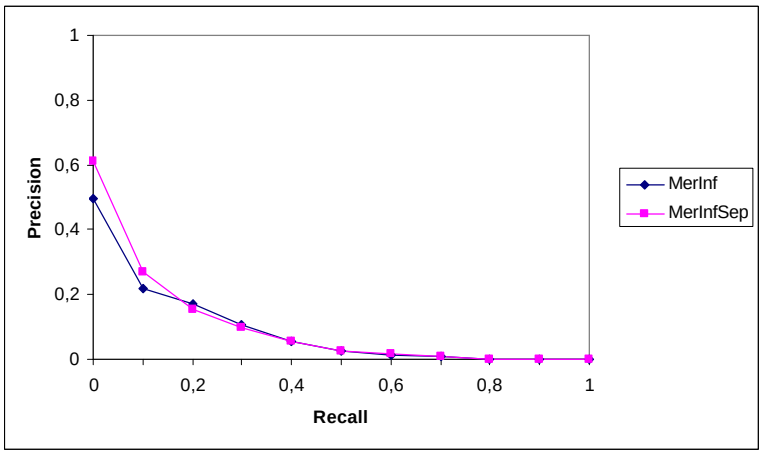

Figure 2. Interpolated average precision values for inflected index merging runs

Among the runs in the merged stemmed index, the best result, average precision 17.8 %, was achieved by the result lists merging approach (see Table 6 and Figure 3). The next best was the merged query without language codes. The run with the languages codes achieved the worst result, 13.1 %. The differences between the results are significant with the Friedman's test at the level 0.05.

Table 6. Non-interpolated average precision of the stemmed index merging approaches (source language English) for all relevant documents averaged over queries

Run	Average precision %	Diff. % (from the baseline)	Change % (from the baseline)
MerStem	14.2		
MerStemQueryCode	13.1	-1.1	-7.7
MerStemSep	17.8*	+3.6	+25.4

* Differences from the other runs are statistically significant by the Friedman's test at the 0.05 level.

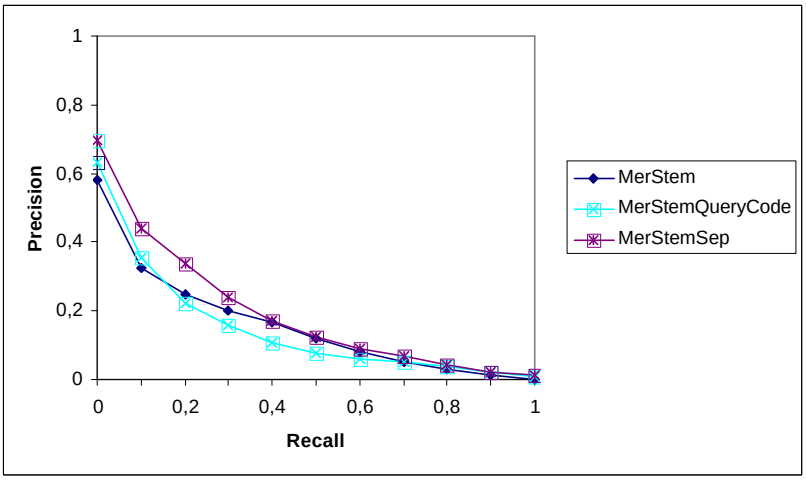

Figure 3. Interpolated average precision values for stemmed index merging runs

4.3 Comparison of the best runs in each run group

Next, the best runs of each run group (inflected / stemmed in separate indexes, inflected / stemmed in merged indexes) were compared with each other (see Table 7). The run in the separate stemmed indexes with result list merging gave the best result, 19.7 % average precision. The next was the run with separate queries and result list merging in the stemmed merged index with average precision 17.8 %. The differences between the results are significant with the Friedman's test at the level 0.05.

Table 7. Non-interpolated average precision of best runs of each run group

Run	Average precision %	Diff. % from SepInfNORM	Change % from SepInfNORM	Diff. % from SepStemNORM	Change % from SepStemNorm	Diff. % from MerInfSep	Change % from MerInfSep
SepInfNORM	16.2						
SepStemNORM	19.7	+3.5	+21.6				
MerInfSep	8.7*	-7.5	-46.3	-11.0	-55.8		
MerStemSep	17.8	+1.6	+.9.9	-1.9	-9.6	+9.1	+104.6

* Differences from the other runs are statistically significant by the Friedman's test at the 0.05 level.

Thus, the settings of separate indexes gave better results than settings of merged indexes. Inside both settings, stemming is advantageous. With separate indexes, the change from the inflected result to the stemmed result is +21.6 %, and with merged indexes 104.6 %.

5 Discussion

Previous research on CLIR in separate indexes has shown that simple merging methods give comparable results with complex new approaches (see Callan & al. 1995, Savoy 2002, Chen 2003 and Braschler & al. 2002). This study shows congruent results. In the both separate index settings, inflected word form and stemmed, the merging approach based on normalization according to collection size gave the best result. The differences to simpler approaches, like the raw score and the round robin approach were minor, however. There is an apparent explanation for the fact that it is difficult to remarkably outperform the results of those simple approaches. When merging result lists there usually is no outside information available, and we must rely on the information in the lists. It is obvious, however, that score normalization is not possible without any exterior information. The score calculation formula includes statistical information about the data sets, and the result lists include only the score and the rank of the document in the list. Based on this information, normalization is out of range. The only variable we can change is the number of documents we select from each list. There are two selection approaches: either to select an equal number of documents per each topic, or to vary the number from topic to topic. In the latter case, we should ground selection on ranks and scores, which do not offer much information for the task.

Score normalization could be possible if collection statistics are available. Any perfect normalization is seldom possible, however, because all needed information is not within reach. In probabilistic retrieval systems, the belief scores are calculated as a combination of term frequency (tf) and inverse document frequency (idf) weights. The collection size and the length of the document are used in the INQUERY formula as well. (Broglio & al. 1994, 554-55.)

Retrieval with a merged, multilingual query in a merged index did not perform very well: the results were much worse than those in the settings of separate indexes. Our results were congruent with those of Nie and Jin in 2002 (Nie & Jin 2002): they also reported poor performance of a merged index. Chen achieved quite good results with a merged index in 2002, but those could not be attained in our tests (Chen 2002, 55-57). Retrieval with separate queries with result list merging performed quite well in the stemmed merged index: the result was 3.6 % better than the result with the merged query, and only 1.9 % worse than the result of the run with separate indexes and result list merging. The approach of separate queries in the inflected merged index did not give much better result than the approach of the merged query.

Nie presented in 2002 the idea of the language code when retrieving in a merged index with a merged query (Nie 2002, 12). The purpose of the code was to decrease the overweighting problem caused by similar translations. According to our tests, the language code does not bring any benefit, because reasons for the poor performance of the merged query do not lie in language confusion.

The performance of both the merged index approach and the separate index approach could be enhanced by stemming. The impact of stemming seemed to be more important in the merged index than in the separate indexes. If there were word form generators available, the inflected results would be better than they are now. Word form generators are rare, however, and their use requires language recognition as well. Presumably stemming would be an easier and cheaper way to enhance the results than word form generation.

6 Conclusions

Previous research has addressed mostly CLIR in separate indexes. Research on CLIR in a merged index has mainly given quite disappointing results. This research gives similar results with an inflected word form index. With a stemmed index, results are better. When monolingual queries with result list merging are used in retrieval, the result almost achieves the result of separate indexes.

In the environment of separate indexes, score normalization is usually impossible to perform, unless collection statistics are available. Simple merging approaches, like the round robin and the raw score approach, give comparable results with more complicated approaches. Thus, there is no need to direct research on developing new result list merging approaches.

Stemming is advantageous for both of the CLIR approaches, the merged index approach and the separated index approach. Stemming is difficult to perform in broad, evolving environments, like Internet. In smaller, more controlled environments stemming is a highly recommendable option.

Increasing numbers of people use Internet daily, and would benefit of CLIR possibilities. It is possible to construct a CLIR system on a multilingual Internet index. User tests should be performed in order to clarify the usability of CLIR in the Internet.

Acknowledgements

The author wishes to thank Prof. Jaana Kekäläinen and Prof. Kalervo Järvelin for their comments in preparing this article.

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

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

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

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

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

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

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

The SNOWBALL stemmers by Martin Porter.

The Spanish and French stemmers by ZPrise

The Italian stemmer by the University of Neuchatel

The Dutch stemmer by the University of Utrecht

References

Airio, E., Keskustalo, H., Hedlund, T. & Pirkola, A. (2003). UTACLIR @ CLEF2002 – Bilingual and multilingual runs with a unified process. In Peters, C., Braschler, M., Gonzalo, J. & Kluck, M. (Eds.), Advances in cross-language information retrieval. Results of the cross-language evaluation forum - CLEF 2002. Lecture Notes in Computer Science 2785 (pp. 91-100). Springer-Verlag, Germany.

Airio, E. (2005). Word normalization and decompounding in mono- and bilingual IR. Information Retrieval, to appear.

Alkula, R. (2001). From plain character strings to meaningful words: Producing better full text databases for inflectional and compounding languages with morphological analysis software. Information Retrieval 4(3-4), 195-208.

Allan, J., Callan, J., Croft B,, Ballesteros L., Broglio J., Xu J. & Shu H. (1997). INQUERY at TREC 5. In Voorhees, E. & Harman, D. (Eds.), Proceedings of the fifth text retrieval conference (TREC-5), NIST Special Publication 500-238 (pp. 119-132). Retrieved March 2005. Available from <http://trec.nist.gov/pubs/trec5/t5_proceedings.html>.

Braschler, M., Göhring, A. & Schäuble, P. (2002). Eurospider at CLEF 2002. In Peters, C. (Ed.), Working Notes for the CLEF 2002 Workshop (pp. 127-132). Rome, Italy.

Braschler, M. & Ripplinger, B. (2004). How effective is stemming and decompounding for German text retrieval? Information Retrieval 7, 291-316.

Broglio, J. & Callan, P. & Croft, W.B. (1994). INQUERY system overview. In Proceedings of the TIPSTER Text Program (Phase I) (pp. 47-67). CA: Morgan Kaufman Publishers Inc, San Francisco.

Callan, J.P., Croft, W.B. & Harding, S.M. (1992). The INQUERY retrieval system. In Proceedings of the Third International Conference on Database and Expert Systems Applications (pp. 78-83). Springer-Verlag , Valencia, Spain.

Callan, J.P., Lu, Z. & Croft, W.B. (1995). Searching distributed collections with inference networks. In Proceedings of the 18th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval (pp. 21-28). ACM, New York.

Chen, A. (2002). Multilingual information retrieval using English and Chinese queries. In Peters, C., Braschler, M, Gonzalo, J. & Kluck, M. (Eds.), Evaluation of Cross-Language Information Retrieval Systems. Lecture notes in computer science 2406 (pp. 44-58). Springer-Verlag, Germany.

Chen, A. (2003). Cross-language retrieval experiments at CLEF 2002. In Peters C., Braschler, M., Gonzalo, J. & Kluck, M. (Eds.), Advances in Cross-Language Information Retrieval. Lectures in computer science 2785 (pp. 28-48). Springer-Verlag, Germany.

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

Hiemstra, D., Kraaij, W., Pohlmann R. & Westerveld, T. (2001). Translation resources, merging strategies, and relevance feedback for cross-language information retrieval. In Peters, C. (Ed.), Cross-language information retrieval and evaluation. Lectures in computer science 2069 (pp. 102-115). Springer-Verlag, Germany.

Hollink, V., Kamps, J., Monz, C. & De Rijke, M. (2004). Monolingual document retrieval for European languages. Information Retrieval 7, 33-52.

Kettunen, K., Kunttu, T. & Järvelin, K. (2004). To stem or lemmatize a highly inflectional language in a probabilistic IR environment? Journal of Documentation, 61(xxx), xxx-xxx, accepted with minor revision.

Koskenniemi, K. (1983). Two-level morphology: A general computational model for word-form recognition and production. University of Helsinki, Finland. Publications No. 11.

Kraaij, W. (2004). Variations on language modeling for information retrieval. CTIT PhD. thesis No. 04-62, University of Twente.

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 (pp. 191-202). ACM, New York.

Lin, W. & Chen, H. (2002). Merging in multilingual information retrieval. In Peters, C. (Ed.), Working Notes for the CLEF 2002 Workshop (pp. 97-102). Rome, Italy.

Nie, J. (2002). Towards a unified approach to CLIR and multilingual IR. In SIGIR 2002 Workshop I, Cross-language information retrieval: a research map (pp. 8–14.). University of Tampere.

Nie, J. & Jin, F. (2002). Merging different languages in a single document collection. In Peters, C. (Ed.), Working Notes for the CLEF 2002 Workshop (59-62). Rome, Italy.

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, (pp 55-63). ACM, New York..

Pirkola, A. (2001). Morphological typology of languages for IR. Journal of Documentation, 57 (3), 330-348.

Popovic, M. & Willet, 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.

Porter, M. (1980). An algorithm for suffix stripping. Program, 14(3), 130-137. Retrieved January 2005. Available from <http://telemat.det.unifi.it/book/2001/wchange/download/stem_porter.html>.

Porter, M. (1981). Snowball: A language for stemming algorithms. Retrieved January 2004. Available from <http://snowball.tartarus.org/texts/introduction.html>.

Powell, A. L., French, J. C., Callan, J., Connell, M. & Viles, C. L. (2000). The impact of database selection on distributed searching. In Proceedings of the 23rd Annual International ACM SIGIR Conference on Research and Development in Information Retrieval, (pp. 232-239). ACM, New York.

Savoy, J. (2002). Report on CLEF-2001 experiments: Effective combined query-translation approach. In Peters, C., Braschler, M, Gonzalo, J. & Kluck, M. (Eds.), Evaluation of Cross-Language Information Retrieval Systems. Lecture notes in computer science 2406 (pp. 27-43). Springer-Verlag, Germany.

Voorhees, E.M., Gupta, N.K. & Johnson-Laird, B. (1995). Learning collection fusion strategies. In Proceedings of the 18th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval, (pp. 172-179). ACM, New York.