Multilingual experiments of UTA at CLEF 2003
The impact of different merging strategies and word normalizing tools

Eija Airio, Heikki Keskustalo, Turid Hedlund, Ari Pirkola
University of Tampere, Finland
Department of Information Studies
e-mail: eija.airio@uta.fi, heikki.keskustalo@uta.fi, turid.hedlund@shh.fi, pirkola@tukki.jyu.fi

Abstract
There are two main translation approaches in a multilingual information retrieval task: either to translate the topics or to translate the datasets. The first one is an easier and more common approach. There are two indexing approaches: either to index the languages in the same index, or to build separate indexes for different languages. If the latter approach is used, retrieved result sets must be merged. Several merging strategies have been developed, but there have been no breakthroughs. University of Tampere used the raw score approach as the baseline in the CLEF 2003 runs and tested two new merging approaches. No remarkable differences were found between the merging methods. The index words can be stored as such (inflected index), or normalized. The most common normalizing method in IR is stemming. The tests of University of Tampere show that stemming is a good normalizing method for English, which is a language with weak morphology. For Finnish stemming gives remarkably worse results, the index built with the morphological analyzer gave better results.

1 Introduction

A multilingual information retrieval task deals with search requests in one source language and documents in several target languages. Two main approaches are possible. (a) The target documents are translated into the source language, or (b) the search requests are translated into the document languages. The first approach would be comfortable for the user. It is quite an unrealistic approach, however, because translating thousands of documents is an expensive and resource-demanding task. In addition, translation should be done for each possible source and target language separately. The second approach, translating only the search requests, is more realistic.

There are two approaches to index a multilingual document collection: (a) to build separate indexes for each document language, or (b) to build a common multilingual index. If the first approach is followed, then retrieval must be performed from each index separately, and the result lists have to be merged somehow. If a common index is created, there are two approaches: either (a) to perform retrieval by each target language separately, and then merge result lists, or (b) merge the search requests into a single query, and perform retrieval.

In CLEF 2003, University of Tampere (UTA) utilized the UTACLIR (Hedlund & al. 2002) system for topic translation. The approach of separate indexes was followed, and three different merging strategies were tested. The impact of two different word normalizing methods on multilingual information retrieval was investigated.

2 Word normalization methods

The area of linguistics concerned with the internal structure of words is called morphology. Inflectional morphology describes impacts of language syntax on words, for example plural of nouns, or tempus of verbs. Derivational morphology goes beyond the syntax: it may affect word meaning as well. The impact of morphology on information retrieval is language dependent. English, for example, has quite weak morphology, and word inflection does not have a great impact on IR. On the other hand, there are languages with strong

morphology (e.g. Hungarian, Hebrew and Finnish), which may have hundreds or thousands of word form variants. The impact of word inflection on IR is considerable in these cases. (Krovetz 1993, 191.)

There are two main approaches to handle inflection: (a) to normalize index words, or (b) to leave index words inflected and let users handle the problem. The search engines of Internet have mostly pushed the responsibility onto their users, which is understandable because of huge amounts and large diversity of data. Users of Internet search engines are guided either to use truncation (for example Alta Vista, http://www.altavista.com/) or to supply all requested word forms (for example Google, http://www.google.com/).

There are two kinds of tools for normalizing words: stemmers and morphological analyzers. The purpose of stemmers is to reduce morphological variance of words. There are several stemming techniques. The simplest stemming algorithms only remove word plural endings, while more developed ones handle a variety of suffixes in several steps. (Krovetz 1993, 191.) Stemming is a normalizing approach compatible with languages with weak morphology, because their inflection rules are easy to apply in a stemming algorithm. Stemming may not be the best normalizing method for languages with strong morphology, because it is not possible to create any simple rules for them. Morphological analyzers are more sophisticated normalizing tools. They incorporate full description of inflectional morphology of the language and a large lexicon of basic core vocabulary items. (Karlsson 1995, 1.)

3 Merging methods

There are many ways to merge the result lists. One of the simplest is *the Round Robin approach*, which bases on the idea that document scores are not comparable across the collections. Because one is ignorant about the distribution of relevant documents in the retrieved lists, an equal number of documents is taken from the beginning of every result list. (Hiemstra & al. 2001, 108.)

The Raw Score approach is based on the assumption that document scores are comparable across collections (Hiemstra & al. 2001, 108). The lists are sorted directly according to the document scores. The raw score approach has turned out to be one of the best basic methods (Chen 2002a, Moulinier & al. 2002, Savoy 2001).

Different methods for normalizing the scores have been developed. A typical *Normalized Score approach* is to divide the score by the maximum score of the collection. Some other balancing factor can be utilized as well.

Several more sophisticated approaches have been developed, but there have not been any breakthroughs.

4 The UTACLIR process

The UTACLIR translation process we used in CLEF 2003 was the same that we utilized in our CLEF 2002 runs. The user gives the source and target language codes and the search request as input to UTACLIR. The system calls external resources (bilingual dictionaries, morphological analyzers, stemmers, gramming functions and stop lists) according to the language codes.

UTACLIR processes source words as follows:

1) the word is normalized utilizing a morphological analyzer
2) the source stop words are removed
3) the normalized word is translated
4) if the word is translatable, the translated words are normalized (by a morphological analyzer or a stemmer, depending on the target language code)
5) the target stop words are removed (in the case that a morphological analyzer was applied in phase 4)
6) if the word is untranslatable in phase 4, two highest ranked words obtained in n-gram-matching are selected as query words.

The system assumes that there is a morphological analyzer for the source language, because source words in the translation dictionaries are in their base forms. Target language words can be normalized either by a morphological analyzer or a stemmer, depending on the target index. The target language code expresses the target normalizing type, as well as the target language. Target stop word removal is done only when a

morphological analyzer is applied in the target language normalizing phase, because stop list words are in their basic forms.

5 Runs and results

In this section, we first describe the language resources used, then the collections, and the merging strategies adapted. Finally, we report results of our CLEF 2003 multilingual runs and the additional runs we performed.

Language resources

We used the following language resources in the tests:

- Motcom GlobalDix multilingual translation dictionary (18 languages, total number of words 665 000) by Kielikone plc. Finland
- Morphological analyzers FINTWOL, SWETWOL and ENGTWOL by Lingsoft plc. Finland
- 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
- Stemmers for English, German, Finnish and Swedish, SNOWBALL stemmers by Dr Martin Porter
- English stop word list, created on the basis of InQuery's default stop list for English
- Finnish stop word list, created on the basis of the English stop list
- Swedish stop word list, created at University of Tampere

Test collections and indexes

The test collections of "Large-multilingual" (Multilingual-8) track of CLEF 2003 were used for the tests.

Eleven indexes were built for the test. For English, Finnish and Swedish we built two indexes: one utilizing a stemmer, and one utilizing a morphological analyzer. For Dutch, French, German, Italian and Spanish we built a stemmed index for each.

The *InQuery* system, provided by the Center for Intelligent Information Retrieval at the University of Massachusetts, was utilized in indexing the databases and as a test retrieval system.

Merging methods applied

We made tests with our CLEF 2002 result lists in order to select good, but simple merging strategies for CLEF 2003. We used raw score method as the baseline. On the basis of our tests, two more approaches were selected: the *dataset size based method* and the *score difference based method.*

The *dataset size based method* is based on the fact that it is likely that more relevant documents are found in a large dataset than in a small dataset. The number of document items taken from single result sets was calculated as follows: T * n / N, where T is the number of document items per topic in the single result list (in CLEF 2003 it was 1000), n is the dataset size and N is the total number of documents (the sum of documents in all the collections). 185 German, 81 French, 99 Italian, 106 English, 285 Spanish, 120 Dutch, 35 Finnish and 89 Swedish documents were selected for every topic in CLEF 2003 runs.

In *score difference based method* every score is compared with the best score of the topic. Only documents with the difference of scores under the predefined value are taken to the final list. This bases on the assumption that documents whose scores are much lower than the score of the top document, may not be relevant.

Runs

UTA did five multilingual runs in CLEF 2003. In the first three runs retrieval was performed from eight separate indexes: the morphologically analyzed English, Swedish and Finnish indexes, and the stemmed French, German, Dutch, Italian and Spanish indexes. The last two runs were merged from runs with eight stemmed indexes (English, Swedish, Finnish, French, German, Dutch, Italian and Spanish). The results were merged utilizing raw score, dataset size based and score difference per topic based methods. The difference value in utamul3 was 0.07, and in utamul5 0.08. These values were selected on the basis of test with CLEF 2002 lists.

There are no remarkable differences between the results of our different multilingual runs (Table 1). The best results, 18.6 % average precision, were achieved by two runs, utamul1 and utamul4, applying approaches totally different from each other. In the first one we used morphologically analyzed / stemmed indexes, and applied raw score merging strategy. The latter was performed with stemmed indexes, and dataset size based merging strategy was applied.

Table 1. Average precision of official multilingual runs of UTA in CLEF 2003

	Index type	**Merging strategy**	**Average precision %**	**Difference %**
utamul1	morphologically analyzed / stemmed	raw score	18.6	
utamul2	morphologically analyzed/ stemmed	dataset size based	18.3	-1.6
utamul3	morphologically analyzed / stemmed	score difference per topic	18.2	-2.1
utamul4	stemmed	dataset size based	18.6	0.0
utamul5	stemmed	Score difference per topic	18.5	-0.5

We made three additional multilingual runs to clarify the impact of different indexing and merging strategies on the result (Table 2). In the first one we applied the raw score merging strategy for stemmed indexes. In the run utamul1 we applied the raw score strategy for morphologically analyzed / stemmed indexes, and achieved the best result of our runs. The average precision of our first additional run was 18.3 %, which is worse than the result of utamul1, but does not differ much from the results of our official runs.

In the second and third of our additional runs we applied the round robin merging approach. In the second run we used morphologically analyzed / stemmed indexes, and in the third one stemmed indexes. The result of both these runs were 18.4 % - again a result, which is very near our official results.

Table 2. Average precision of additional multilingual runs

Index type	**Merging strategy**	**Average precision %**
stemmed	raw score	18.3
morphologically analyzed / stemmed	round robin	18.4
stemmed	round robin	18.4

We can order the merging strategies applied in our official and additional runs according to the result they give with the morphologically analyzed / stemmed and stemmed indexes. With morphologically analyzed / stemmed indexes the raw score method gives the best result, 18.6 % average precision. The second is the round robin strategy, which gives 18.4 % average precision. The third is the dataset size based method (18.2 %), and the last the score difference per topic method (18.2 %). The order is different with stemmed indexes. The dataset size based method is the best with 18.6 % average precision, and the score difference per topic is the second (18.5 %). The third is round robin strategy (18.4), and the last is raw score (18.3 %). It is possible that the differences between the results of morphologically analyzed / stemmed and stemmed indexes are caused by change, because

they are very small. More testing should be done to clarify whether the effectiveness of a merging strategy depends on the index type.

The results of our official and additional multilingual runs do not show the performance difference of morphologically analyzed / stemmed vs. stemmed indexes. To test the impact of the indexing method on the retrieval result we made six additional runs: two monolingual English and four bilingual runs (two English – Swedish and two English – Finnish) (Table 3). The average precision of monolingual English run was 1.5 % better with the stemmed index than with the index built by means of a morphological analyzer. The results of bilingual English – Finnish runs are opposite: the stemmed index performed badly. The average precision with the stemmed index was 44.1 % lower than with the index built by means of the morphological analyzer. The results of English – Swedish runs show similar trends, although the difference (-29.9 %) is smaller than in English – Finnish runs but still significant. These results confirm the assumption that stemming is a competitive normalizing approach for languages with weak morphology, but not for languages with strong morphology.

Table 3. Average precision of monolingual English, bilingual English –Swedish and bilingual English - Finnish runs using alternative indexes

Language	Index type	Average precision %	Difference % units	Difference %
English	morphologically analyzed	45.6		
English	stemmed	46.3	+0.7	+1.5
Finnish	morphologically analyzed	34.0		
Finnish	stemmed	19.0	-15.0	-44.1
Swedish	morphologically analyzed	27.1		
Swedish	stemmed	19.0	-8.1	-29.9

4 Discussion and conclusion

The impact of different normalizing methods, stemming and morphological analyzing, on the IR performance has not been investigated widely. The reason for that is presumably the fact that English is the traditional indexing language in IR tests. English is a language with weak morphology, which implies that stemming is an adequate word form normalization method. Our monolingual English tests with CLEF 2003 data support this, the result with the stemmed English index is even a little better than the result with the normalized index. The bilingual test we made with Finnish and Swedish indexes show opposite results. The results with stemmed indexes in these languages are much worse than the results with the index built utilizing a morphological analyzer. When high precision is demanded, stemming is not an adequate normalizing method with languages with strong morphology.

On the other hand, the most used IR systems in the real life are the search engines of Internet. They use inflected indexes, which means that the users have to handle inflection. Truncation is possible with some search engines, while others guide their users to supply all the possible forms of their search words. Loss of recall is liable, but recall may not be important in www searching. In many cases more important is precision, which may be good even if the user has not perfect language skills.

In most multilingual experiments separate indexes are created for different languages, and various result merging strategies are tested. The results of test made with a merged index are not very promising (Chen 2002b, Nie & Jin 2002). In the real life, the situation of separate indexes and result merging, occurs quite rarely, however. This would be a reason to direct the research towards the strategies of the merged index approach.

Acknowledgements

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

ENGTWOL (Morphological Transducer Lexicon Description of English): Copyright (c) 1989-1992 Atro Voutilainen and Juha Heikkilä.
FINTWOL (Morphological Description of Finnish): Copyright (c) Kimmo Koskenniemi and Lingsoft plc. 1983-1993.
GERTWOL (Morphological Transducer Lexicon Description of German): Copyright (c) 1997 Kimmo Koskenniemi and Lingsoft plc.
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.

This work was partly financed by CLARITY (Information Society Technologies Programme, IST-2000-25310).

References

Airio, Eija & Keskustalo, Heikki, & Hedlund, Turid, & Pirkola, Ari.. 2002. Cross-language Retrieval Experiments at CLEF 2002. In *Working Notes for the CLEF 2002 Workshop*, Italy 2002, pp. 5-20.

Chen, Aitao. 2002a. Cross-language Retrieval Experiments at CLEF 2002. In *Working Notes for the CLEF 2002 Workshop*, Italy 2002, pp. 5-20.

Chen, Aitao. 2002b. Multilingual Information Retrieval Using English and Chinese Queries. In *Evaluation of Cross-Language Information Retrieval Systems. Lecture notes in computer science; Vol. 2406*. Springer-Verlag, Germany 2002, pp. 44-58.

Hedlund, Turid & Keskustalo, Heikki & Pirkola, Ari .& Airio, Eija & Järvelin, Kalervo. 2002. Utaclir @ CLEF 2001 – Effects of compound splitting and n-gram techniques. In *Evaluation of Cross-language Information Retrieval Systems. Lecture Notes in Computer Science; Vol. 2406*. Springer-Verlag, Germany 2002, pp. 118-136.

Hiemstra, Djoerd & Kraaij, Wessel & Pohlmann, Renée & Westerveld, Thijs. 2001 Translation resources, merging strategies, and relevance feedback for cross-language information retrieval. In *Cross-language information retrieval and evaluation. Lectures in computer science 2069*. Springer-Verlag, Germany 2001, pp. 102-115.

Karlsson, Fred. 1995 SWETWOL: A Comprehensive Morphological Analyser for Swedish. In *Nordic Journal of Linguistics 15*, pp. 1-45.

Krovetz, Robert. 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.

Moulinier, Isabelle & Molina-Salgado, Hugo. 2002. Thomson Legal and Regulatory Experiments for CLEF 2002. In *Working Notes for the CLEF 2002 Workshop*, Italy 2002, pp. 91-96.

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

Savoy, Jacques & Rasolofo, Yves. 2001. Report on the TREC-9 Experiment: Link-Based Retrieval and Distributed Collections. In *Proceedings of the Ninth Text Retrieval Conference, NIST Special Publication 500-249*, Department of Commerce, National Institute of Standards and Technology, pp. 579 – 588.