UTACLIR @ CLEF 2002:
Towards a unified translation process model

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

The UTACLIR query translation system was originally designed for the CLEF 2000 and 2001 campaigns. In the two first years the query translation application consisted of separate programs based on common translation principles for the language pairs Finnish - English, German - English and Swedish - English. The idea of UTACLIR is based on recognizing distinct source key types and processing them accordingly. The linguistic resources utilized by the framework include morphological analysis or stemming in indexing, stop word removal, normalization of topic words, splitting of compounds written together, handling of non-translated words, phrase composition of compounds in the target language, bilingual dictionaries and structured queries.
This year we participated in CLEF with the new UTACLIR system, which is a single program unified for all the languages. The user gives the system the codes of source and target language as well as the query to be translated. The UTACLIR system chooses the language resources upon the codes the user has given. A morphological analyser is used to process the source language words in order to match the words in the translation dictionary. We have utilized an 18-language dictionary (with all possible language pairs) as the main translation resource of the UTACLIR. It is also possible to implement other parallel dictionaries. For the target language it is possible to use either a morphological analyser or a stemmer.

1 Introduction

University of Tampere has participated in the bilingual tasks of CLEF years 2000 and 2001 utilizing the UTACLIR process. UTACLIR has consisted of separate, but similar kind of programs for the language pairs Finnish - English, German - English and Swedish - English. The idea of UTACLIR is based on translating topic words one by one, and then combining the translations into the query.

The source word processing can be described in general level as follows. First the topic words are normalized with a morphological analyser, if possible, and after that source stop words are removed. Then translation is attempted. Translated words are normalized, because it is possible that a dictionary returns words in inflected form (e.g. “United States”). Finally the target stop word removal is done. Normalized translation variants are enveloped with a synonym operator and added to the query. The untranslatable words are mostly proper names and technical terms. Typically words like these are spelling variants of each other in different languages, which allows the use of approximate string matching techniques. These techniques are language-independent. (Pirkola & al. 2002.) The best matching strings are searched from the target index. These are enveloped with a synonym operator and added to the query.

UTACLIR has a special procedure for untranslatable compounds written together. They are first splitted into their constituents and then translated separately. Translated parts are enveloped with a proximity operator. (Hedlund & al. 2001.)

Structuring of queries using the synonym operator, which means grouping of the target words derived from the same source word into the same facet, is applied in the UTACLIR system. This has proved to be an effective strategy in CLIR by earlier studies (Pirkola 1998, 60 - 61).

This year we participated in the Finnish monolingual task, the English – Finnish, English – French and English – Dutch bilingual tasks, and the multilingual task. The monolingual task is a traditional retrieval task, the only novelty being the language, which is not the traditionally used language, English. Finnish is introduced as a target language in CLEF 2002. The bilingual task adds the topic translation to the previous one, as well as some extra problems, for example the problem of non-translatable proper names. The multilingual task involves the result merging phase in addition to the previous one, if the most usual approach, building the separate indexes for all the languages, is followed.

There are at least three possible ways to merge the results. The simplest of them is *the Round Robin approach*, which means that a line of every result set is taken, one by one from each, until there are as many lines as needed. This is based on the fact that the distribution of relevant documents in the lists is not known, because the scores are not comparable, and there is no way to compare them. The second approach is *the raw score approach*, which assumes that document scores are comparable across separate collections. The third is *the rank based approach*. It bases on the fact that the relationship between probability of relevance and the log of the rank of a document can be approximated by a linear function. Merging can subsequently be based on the estimated probability of relevance. Actual score of a document is then applied only to rank documents, but the merging is based on the rank, not on the score. (Hiemstra & al. 2001, 108.)

2 The new UTACLIR process

This year we have a new unified version of UTACLIR in use. The basic process is the same for all the source and target languages. As an input for UTACLIR system the user gives the codes expressing the source and target language, and the source language query. Depending on the codes the system uses external linguistic resources: bilingual dictionaries, morphological analysers, source and target stop lists, and stemmers.

The new UTACLIR system has the same basic elements as the old one (see Figure 1). The source word processing has not changed, but there are new features in the target word processing. If translation variants are found, either a morphological analyser or a stemmer is utilized, depending on the index type of the target language. The stemmer produces ready components for the target query, in which case stop word removal is not done. However in case a morphological analyser is used to process the target words, stop word removal is done. Stop words are in a morphologically analysed form, and cannot be utilized in the stop word removal of stemmed target words.

The compound splitting procedure was not yet implemented in UTACLIR during CLEF 2002 runs.

It is possible to use input codes for denoting parallel resources in the new UTACLIR system. In that case the input codes denote not only the source and target language, but also the resource used. If we have for example three different English – Finnish bilingual dictionaries in use, we can easily test their performance with UTACLIR. The source words must be processed by a morphological analyser, not by a stemmer. There is no sense to stem source words, because we do not have dictionaries for stemmed source words at the moment.

The UTACLIR system constructs a three level tree data structure from the source query: 1) Original source keys given by the user; 2) Processed source language strings, for example processed by morphological analysers; 3) Post-processed word-by-word translations. The tree can be traversed and interpreted in different ways, and the final translated query can be constructed by interpreting the tree. (Hedlund & al 2002a, 16 –17.)

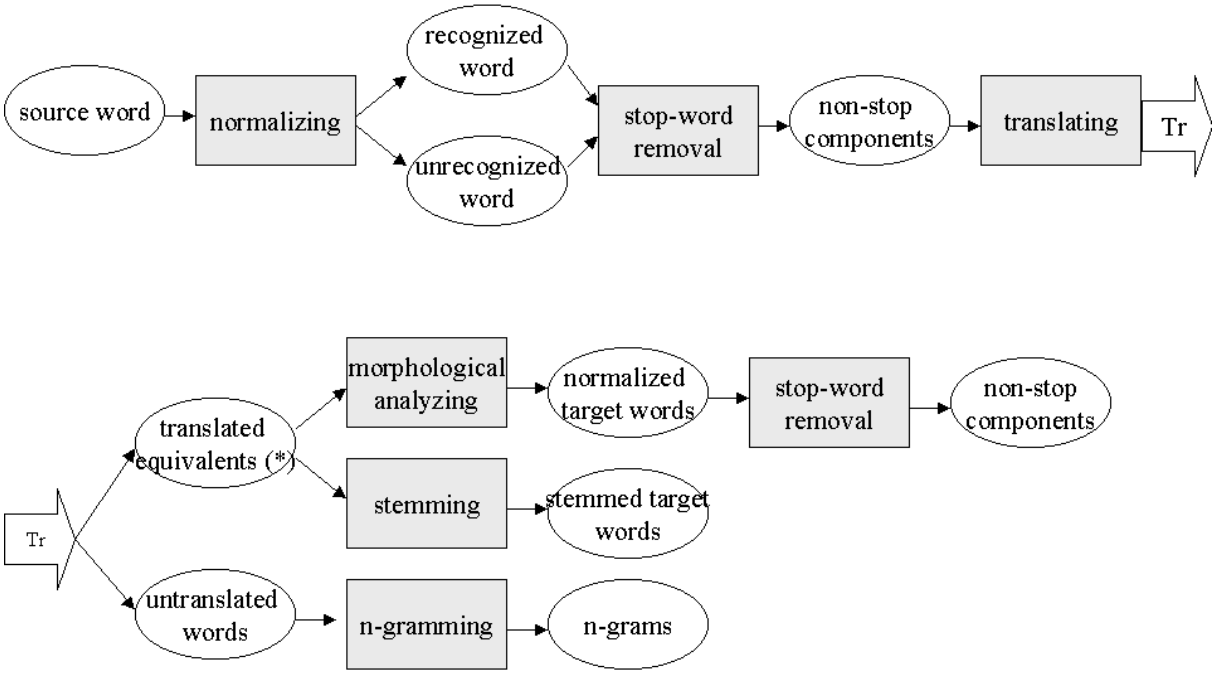

Figure 1. An overview of processing a word in the new UTACLIR process.
(*) Depending on the target language, either morphological analysis or stemming was performed.

3 Runs and results

In this chapter, we first describe the language resources used, then the collections, and the indexing strategy adapted. Finally, we report results of the monolingual, bilingual and multilingual runs.

Language resources

- Motcom GlobalDix multilingual translation dictionary (18 languages, total number of words 665 000) by Kielikone plc. Finland
- Motcom English – Finnish bilingual translation dictionary (110 000 entries) by Kielikone plc. Finland
- Morphological analysers FINTWOL, GERTWOL and ENGTWOL by Lingsoft plc. Finland
- Stemmers for Spanish and French, by ZPrise
- A stemmer for Italian, by the Univeristy of Neuchatel
- 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
- German stop word list, created on the basis of the English stop list
- French stop word list, granted by Université de Provence
- Italian stop word list, granted by University of Alberta
- Spanish stop word list, InQuery's default stop list for Spanish

Test collections

The following test collections were used for the tests: English *LA Times*, Finnish *Aamulehti*, French *Le Monde*, French *SDA*, German *Der Spiegel*, German *SDA*, Italian *La Stampa*, Italian *SDA* and Spanish *EFE*. We had to exclude German *Frankfurter Rundschau* because of indexing problems. Next, the indexing of the databases is described.

Lingsoft's morphological analyser FINTWOL was utilized in indexing the Finnish dataset, and GERTWOL in indexing the German datasets. As we did not have morphological analysers for Spanish, Italian and French, we decided to index those databases by utilizing stemmers. We used Zprise's Spanish stemmer, Zprise's French stemmer and the Italian stemmer granted by the Univeristy of Neuchatel.

We built separate index for every dataset instead of indexing by language, for example separate indexes for *Le Monde* and French *SDA*. Thus, we had eight separate indexes instead of five. This choice has an impact on merging phase, and also affects n-gramming. We will discuss these aspects later in this paper.

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

Monolingual runs

We made two monolingual runs, both in Finnish. The approach of these runs was similar to our bilingual runs, only excluding translation (see Figure 2). In the first run topic words are normalized by using Lingsoft's morphological analyser FINTWOL. Compounds written together are splitted into their constituents. If a word is recognized by FINTWOL, it is checked against the stop word list, and the result (the normalized word, or nothing in the case of stop word) is processed further. If the word is not recognized, it is n-grammed. The n-gram function compares the word with the database index contents. It returns the best match form among morphologically recognized index words and the best match form among non-recognized index words, and combines them with InQuery's synonym operator (#syn operator, see Kekäläinen & Järvelin 1998).

The second monolingual Finnish run is similar to the first one, but no n-gramming is done. Unrecognised words are added to the query as such. There was no big difference in performance between the results of our two Finnish monolingual runs.

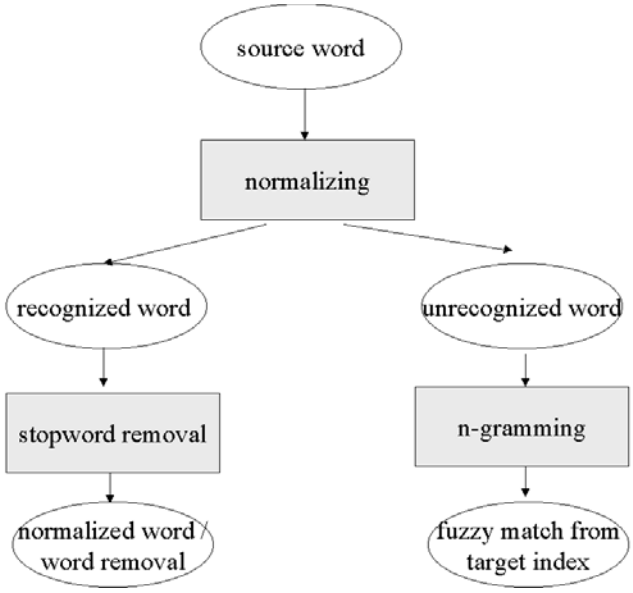

Figure 2. An overview of processing a word in the monolingual run utilizing n-gramming.

Finnish is a language rich in compounds written together. Parts of a compound are often content bearing words. (Hedlund & al. 2002b.) In a monolingual run it is reasonable to split a compound into its components, normalize the components separately, and envelope the normalized components with an appropriate operator. In the original run, we used the synonym operator in the monolingual runs for this purpose instead of the proximity operator, which turned out to be not a good approach. For example topic 140 contains the word *matkapuhelin* (mobile phone). The query constructed for this topic contains a synonym clause #syn(matka puhelin), which means, that occurrences of the word *matka* (travel) or *puhelin* (phone) are allowed, instead of a phrase "matka puhelin".

We made an additional run in order to get a more precise view of the effect of the synonym operator in the compounds compared with the proximity operator. There, we replaced the synonym operator with the InQuery's #uw3 operator (proximity with the window size 3) in the cases of compounds. We compared these new results to the corresponding results of our CLEF runs (see table 1). Average precision of this additional run was 30.4 % better in the run using n-grams, and 33.3 % better in the run with no n-grams. We can conclude, that demanding of all the parts of the compound to occur in the document is essential to get better results.

Table 1. Average precision for Finnish monolingual runs using synonym and uw3 operator

	Average precision %	Difference % units	Difference %
Gramming and Synonym operator	27.0		
Gramming and uw3 operator	35.2	+8.2	+30.4
No gramming, Synonym operator	24.0		
No gramming, uw3 operator	32.0	+8.0	+33.3

Common features of bilingual and multilingual runs

Handling of source words by ENGTWOL and the processing of source language stop words were similar in all the bilingual and multilingual runs we made, because we used only English as a source language in all these. GlobalDix dictionary by Kielikone was utilized in all the translations.

We had a beta-version of UTACLIR in use during the CLEF-runs. There were some deficiencies compared to the old version, because all the features of UTACLIR were not yet implemented in the new one. Splitting of compounds was not yet implemented, and non-translated words were handled (using the n-gram method) only in German as a target language. We did not utilize target stop word removal in the case of stemming. Our stop word lists consist of morphologically normalized words at the moment, thus they cannot be used as such to remove the stemmed forms.

The n-gramming functions must be applied separately for each target index. Because we have two distinct indexes in German, French and Italian, we should make eight n-gramming functions. Due to time limitations we made the function only for the German SDA index, and utilized the same with the *Der Spiegel* index. We excluded n-gramming in other cases.

Bilingual runs

We made this year three bilingual runs: English – Finnish, English – Dutch and English – French. The English – Dutch run is not reported because of a severe failure in the indexing of the Dutch database. The result of English – French run was utilized also in the multilingual run.

In the English – Finnish run, FINTWOL was used for normalizing the target words. Target language stop word removal was done after the translation and normalization processes. In the English – French run the stemming approach was used for normalizing the target words in these runs. The French databases were indexed using the stemmer, correspondingly.

The result of the English – Finnish run is in the table 2. The obvious reasons for the quite poor performance of the run would be the defective testing of UTACLIR, and absence of gramming and compound handling. The translations given by the GlobalDix, which were sometimes curious, were doubted to have an impact on the result. We made additional English – Finnish runs to clarify the effect of the dictionary on the result. First we made a run where the untranslatable words were added to the query in two forms: as such and preceded by the character "@" (unrecognised words are preceded by "@" in the index). The average precision was 24.6 %, 21.8 % better than the CLEF run (Table 2). The second comparable run was done utilizing another translation dictionary, MOT with 110 000 Finnish – English entries (compared to 26 000 entries of GlobalDix). The result was 61.4 % better than the original CLEF result. The both dictionaries are from the same producer, Kielikone plc.

Table 2. Average precision for English - Finnish bilingual runs using alternative resources

	Average precision %	Difference % units	Difference %
GlobalDix + no mark-up of unrecognised words	20.2		
GlobalDix + mark-up of unrecognised words	24.6	+4.4	+21.8
MOT + mark-up of unrecognised words	32.6	+12.4	+61.4

As we did not have an alternative English – French –dictionary to translate from English to French, we could not compare the effect of the dictionary on the results. However, some considerations can be done examining the topic translations. The GlobalDix dictionary seems to return some odd translations. As an example, topic 101 deals with Cyprus. The proper name "Cyprus" is translated to the French word "cyprè", which means a cypress in English. The right translation would be "Chypre". Also untranslatable proper names cause problems in retrieval. For example, topic number 94 achieved a poor result, because it includes the proper name "Solzenitsyn", which does not exist as such in the French dataset: the French layout is "Soljenitsyne". Better results will presumably be achieved with the French n-gramming function.

Multilingual runs

University of Tampere participated for the first time in the multilingual task this year. The main goal was to gain experience for developing a general query translation framework.

The topics were in English, so the beginning of the process was similar in every language: topic words were normalized using ENGTWOL and after that the source stop words were removed. TheGlobalDix dictionary was used to translate normalized source words to the target languages. As we have a morphological analyser for German, GERTWOL by Lingsoft, it was used for normalizing the target words. For Spanish, French and Italian we had no morphological analysers, thus we chose to utilize stemmers instead. We used ZPrise's Spanish and French stemmers, and the Italian stemmer of the Univeristy of Neuchatel. Target stop word removal was done only for morphologically analysed target queries (so it was done only in the German run).

There are several different strategies to merge the results obtained from distinct databases. In the first run we applied merging method described by Voorhees and others: treating the similarity values across the collections as they were comparable, and selecting 1000 greatest similarities across all collections (Voorhees & al 1995, 96). It's obvious that the similarity values are not comparable in all the cases, but we chose this approach because of its simplicity.

Our second multilingual run was similar to the first one, except that a different merging strategy was applied. This was the Round Robin approach: from every result set one line was taken by turn, beginning from the top.

As described in the chapter dealing with databases earlier, we made distinct indexes for all the data sets. So we have eight indexes: one English, one Spanish, two French, two Italian and two German, which means, that we have eight result sets to merge, too. When we have distinct result sets for every data set, we in a way favour the languages which have more than one dataset: French, Italian and German. Whether this is good or not depends on the topic.

We calculated the average precision for the bilingual subtasks present in the multilingual task. The average precision for the English run was 47.6 %, English – French 23.9 %, English – German 13.5 %, English – Italian 20.1 %, and English – Spanish 21.8 %. The absence of one German dataset affects the poor result of the English – German run. Implementing the Italian and Spanish dictionaries was not ready when making the runs. We can expect better result with those languages after some development of UTACLIR.

The average precision of our multilingual run with raw score merging method was 16.4 %, and with the Round Robin method 11.7 %. We have not tested any other merging methods, but probably it would be possible to achieve better results with a more developed method.

4. Discussion and conclusion

Cross-lingual information retrieval has become a significant part of information retrieval research last years, driven mostly by the growth of Internet documents and users. The ultimate goal of cross-lingual information retrieval research is to achieve a situation, where the user can retrieve documents in any language typing a single search topic in one language.

Internet indexes are enormous fusions of documents around the world, written in multiple languages. Internet is too large and too unstable to be used as a test environment. The CLEF test data offer suitable possibilities for interpreting bilingual and multilingual retrieval in an environment simulating real retrieval. The bilingual CLEF task is simple: translating the topics to the target language, or translating the documents to the topic language, and performing the retrieval. The multilingual task includes an extra problem compared to the bilingual task: what to do with the distinct datasets? Most of CLEF participants build distinct indexes for the different languages and then merge the results. Actually this approach differs from that of Internet. If we want to simulate Internet, merging the indexes would be reasonable, not merging the results. The idea of merging the indexes was introduced by Chen in CLEF 2001, as well as an idea of translating the documents and building a monolingual index (Chen 2001). In addition that result merging differs from the Internet approach it is an obvious source of errors (Nie 2002, 11).

It is possible to merge the indexes of different languages, and preserve the language information as well. It can be done for example so that English index words get language code "_e": "chair_e". (Nie 2002, 12). This method helps in recognizing the languages, but still differs from the real situation in Internet.

We are participating the multilingual task first time this year, and our approach is the most usual: merging the results, not indexes. Our main goal in CLEF is to test the new unified UTACLIR system this year. The questions of index building alternatives was not current for us, but in future we may address this topic.

We learnt many important points in the CLEF process this year. Our Finnish monolingual runs repeated the fact, that using a proximity operator instead of the synonym operator with phrases improves the result remarkably. The English – Finnish runs with different translation dictionaries revealed the significance of the dictionary for the result. In general, our multilingual runs prove that a unified process for different languages is possible. The CLEF runs raised many interesting questions concerning the development of UTACLIR. Should we develop a dictionary for stemmed words? If so, we could utilize UTACLIR process with stemmed source languages, without demanding the morphological analyser. Would it be reasonable to construct stemmed stop list? Then we could have the target stop word removal with stemmed target languages as well. A further issue is, what is the implication of result merging on the multilingual run result? Would it be possible to do without merging?

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.

References

Chen, A. 2001. Multilingual information retrieval using English and Chinese queries. Working notes for the CLEF 2001 workshop. *http://www.ercim.org/publication/ws-proceedings/CLEF2/a-chen.pdf*

Hedlund, T., Keskustalo, H., Pirkola, A., Airio, E., Järvelin, K. 2001. UTACLIR @ CLEF 2001: New features for handling compound words and untranslatable proper names. Working notes for the CLEF 2001 workshop. *http://www.ercim.org/publication/ws-proceedings/CLEF2/hedlund.pdf*

Hedlund, T., Keskustalo, H., Airio, E., Pirkola, A. 2002a. UTACLIR – An extendable query translation system. 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. 15 - 18.

Hedlund, T., Pirkola, A., Keskustalo, H., Airio, E. 2002b. Cross-language information retrieval using multiple language pairs. Accepted for presentation at the ProLISSA conference 24 - 25 October 2002, Pretoria.

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: Proceedings of the CLEF 2000 Workshop, Lectures in computer science 2069.* Springer-Verlag, Germany 2001, pp. 102-115.

Kekäläinen, J, Järvelin, K. 1998. The impact of query structure and query expansion on retrieval performance. In *Proceedings of 21st ACM/SIGIR Conference*, pp. 130 – 137.

Nie, J. 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.

Pirkola, A. 1998. The effects of query structure and dictionary setups in dictionary-based cross-language information retrieval. In *Proceedings of the 21st ACM/SIGIR Conference,* pp. 55-63.

Pirkola, A. Keskustalo, H., Leppänen, E., Känsälä, A. P. and Järvelin, K. 2002. Targeted s-gram matching: a novel n-gram matching technique for cross- and monolingual word form variants. In *Information Research*, **7**(2). (http://InformationR.net/ir/7-2/paper126.html)

Voorhees, E.M., Gupta, N. K, Johnson-Laird, B. 1995. The collection fusion problem. In *Proceedings of TREC'3*, pp. 95-104. Gaithersburg: NIST Publication #500-225. (Also http://trec.nist.gov/pubs/trec3/t3_proceedings.html)