UTACLIR @ CLEF 2001:

New features for handling compound words and untranslatable proper names

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

Abstract

We participated in CLEF'2001 with four automated bilingual runs. UTACLIR is an automatic query translation and construction system for cross-language information retrieval. The system automatically extracts topical information from request sentences written in one of the source languages and constructs a target language query, based on translations given by a translation dictionary. The new features for the CLIR process from Finnish, Swedish and German to English focus on matching compound words and a new n-gram based technique for matching proper names and other non-translatable words.

The results for all the four runs are good. Average precision for all the queries shows clear improvements.

For German – English we have tested two types of dictionaries (two runs). The first one included all translations from the standard dictionary. The second contained the same data, except that all direct translations of compounds were excluded. The test with two dictionaries for the German runs gives an indication that the new features in the UTACLIR process work well also with a limited dictionary.

1 Introduction

In the query construction phase the right use of windowing techniques and phrase construction is emphasised (Haas & Losee 1994; Jacquemin 1996, Zhou 1999). The study of the formation of compound words and their combinatorial behaviour in general language and the proper handling of them for CLIR translation is an extensive linguistic as well as an IR task (Levi 1978, Spyns & De Wachter 1995). By a compound we mean in this study two or several words (compound components) that are written together. All the source languages we use are rich in this type of compounds, and thus, one of our main efforts is the morphological decomposition of compounds into constituents and their proper translation. In languages rich in compounds right translation of compounds (or their components) is a factor that greatly affects the retrieval results. The new features and the approach for our automated method for query construction are intentionally as far as possible designed to be source language independent (Grefenstette & Segond 1997).

This is true especially for CLIR-queries where compound splitting and translation of components is performed. In our method for handling compounds we have experimented with the window size, and the phrase operator. In last year CLEF-tests we used an operator requiring strict word order and a fairly small window size. This year we allow for a free word order and a broader window size in the phrase construction for compounds.

Proper names often are prime keys in requests, and if not translated by dictionaries, query performance may be ruined. However, the fact that proper names often are form variants of each other allows the use of approximate string matching techniques to find the target language correspondents for the source language names. Approximate matching techniques involve *Soundex* and *Phonix*, which compare words on the basis of their phonetic similarity (Gadd, 1990) and *n-gram based matching* (Pfeifer et al., 1996; Robertson and Willett, 1998;

Zobel and Dart, 1995). N-gram matching is a language independent matching technique. It thus seems to be an ideal approximate matching technique for CLIR systems processing different languages. Moreover, n-gram matching has been reported to be an effective technique among various approximate matching techniques (Pfeifer et al., 1996; Zobel and Dart, 1995).

2 The New Process

We will present the new features of the UTACLIR research process in this year's CLEF. The old process that is used as a base in the development process is described in detail in the Working Notes and Proceedings of the last year's CLEF (Hedlund et al. 2000; 2001). The old process is used for the Finnish – English test run, except that for compounds we use a more flexible proximity operator and a broader window size. The n-gram matching technique is also in use.

Our approach to solve the general problems for bilingual CLIR is based on 1) word normalisation in indexing, 2) stop-word lists, 3) normalisation of topic words, 4) splitting of compounds, 5) recognition of the right components, 6) handling of non-translated words 7) phrase composition of compounds in the target language, 8) bilingual dictionaries and 9) structured queries. For structuring of queries see (Pirkola, 1998) The language pairs used in the bilingual tests are Finnish – English (FIN-ENG), Swedish – English (SWE-ENG) and German – English (GER-ENG).

An overview of the new UTACLIR process is in Figure 1.

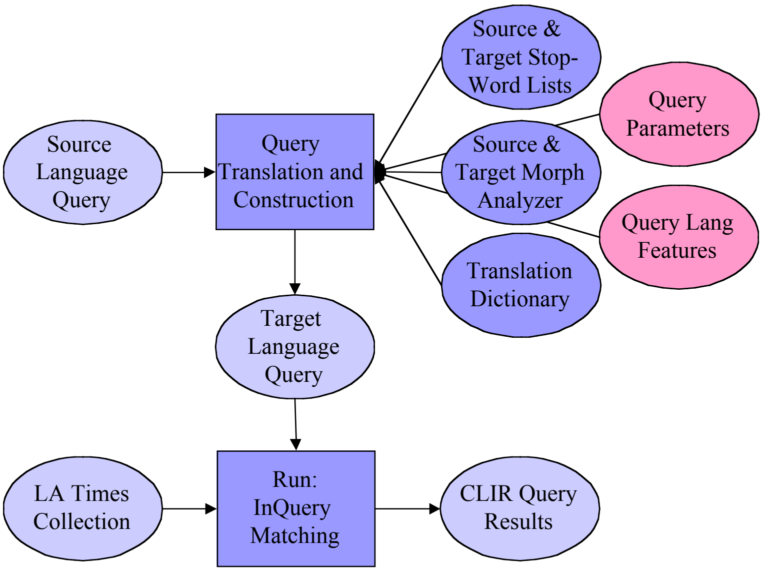

Figure 1. UTACLIR process overview

The new features are as follows:

- A new process for dictionary look-up and translation of compound words
- A new process for matching proper names and other non-translatable words
- New ways of using stop lists
- Normalisation of dictionary output
- Unified process (Swedish and German)

Dictionary look-up and translation of compound words. In the present process, normalised compound words, in case they are not stop words, are first attempted to look up in the dictionary (see Figure 2a). If a translation, or a set of translations, is available, it is likely to be the best alternative for the source word (a compound or non-compound). Such compounds often are non-compositional, i.e., a compound's meaning may be quite different from the meanings of its components (e.g., strawberry). If the translation is a phrase, it will be handled as a phrase in the subsequent phases.

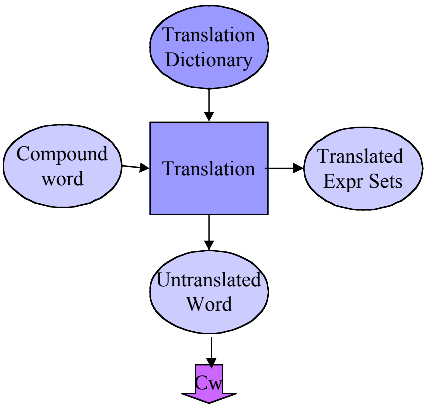

Figure 2 a. Direct translation of compounds

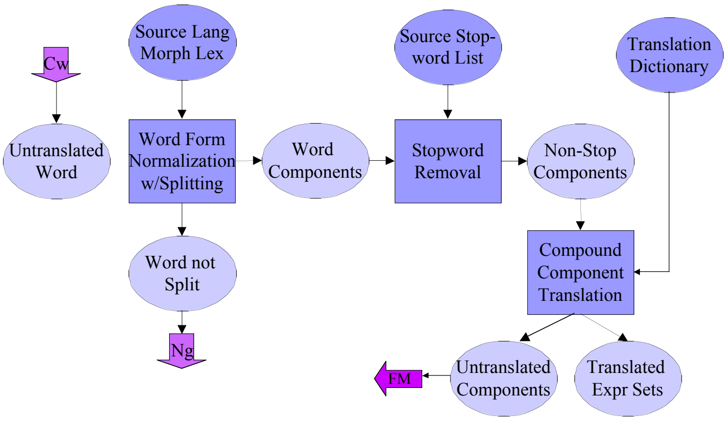

Figure 2 b. Splitting of compounds

Figure 2 a and b. Compound translation (unified process)

Compound words that do not translate are split into their components (see Figure 2b). For Swedish and German, all consecutive component pairs are formed and translated (if possible). For example, for a four-component compound a-b-c-d, the component pairs of a-b, b-c, and c-d are formed. Then these formed component pairs are looked up in the dictionary. In the case of several translations, the equivalents are used as synonyms. If the normalised component did not translate, it was modified by using a fogemorpheme algorithm before a new translation attempt. If it still did not translate, then n-gram method (described below) was used for retrieving the set of six most similar index terms with respect to the component. All combinations of the translation equivalents are formed for the query. The rationale behind this method is that for a multi-component compound word it is hard to know which consecutive components form common established compounds contained in the dictionary.

For German and Swedish compounds, we applied the fogemorpheme algorithm as in CLEF'2000 (see Figure 3). For fogemorphemes in Swedish see Hedlund et al. (2001b). Finnish compound processing differed from the earlier process used in CLEF'2000 in that this year we used a more flexible proximity operator and a broader window size. That is, the proximity operator was changed from OD (ordered window) to UW (unordered window) which allows for free word order in the target phrases. The window size was set to 5 + n, where n = the number of spaces between words in the phrase.

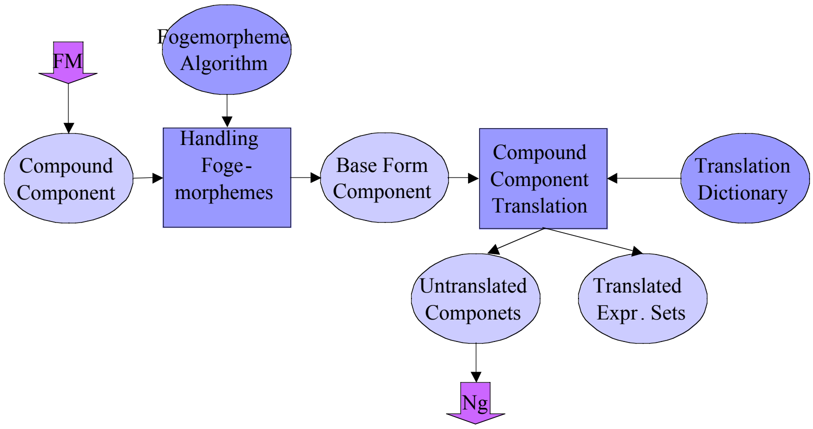

Figure 3. Handling fogemorphemes.

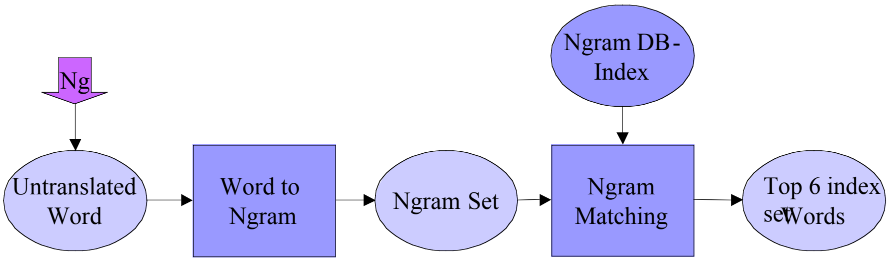

Figure 4. N-gram processing

Untranslatable words. In the present process, proper names and other untranslatable words are handled by an advanced n-gram method (see Figure 4). The method is language independent and is described in detail in (Pirkola & al., 2001). The method is able to find target language spelling variants for source language proper names. Correspondents may be found despite slight variations in characters and/or the number of characters. Proper name translation and matching in CLIR is complicated by the fact that proper names may be inflected similarly as common nouns (particularly in Finnish), and may possess suffixes (representing different case and number features, and other grammatical categories). The n-gram technique was used for all source languages. It was applied for each untranslatable source language word. The six most similar words (the degree of similarity based on similarity calculations associated with the n-gram technique) from the target database index were included into the final query. For Swedish and German, the n-gram technique also was applied for untranslatable components of compounds.

In the new process, *stop word lists* are used in a different way than last year. The new stop lists are used after the normalisation of words to base forms. Thus we do not have to include inflected word forms onto the lists.
Some modifications were done to last year's lists. Owing to the change in the process, only base form words were added onto the lists. This is important when dealing with highly inflectional source languages. Stop lists are not used for the target language query in the Swedish-English and the German-English processes.

Normalisation of dictionary output. Dictionary output can include phrases and words in inflected forms. These do not match the normalised index terms. Therefore, because index terms were normalised, dictionary output words also were normalised.

Unified process. Our aim has been to unify the process for all source languages as much as possible. Language dependent features that have to be added to the process are included, but the initial process is unified. We are trying to develop parameters for the query construction phase, including possibilities to change operators, ways to handle components of compounds in the final query. In this year CLEF tests we have tried to unify the Swedish – English and German – English processes, however the process is not yet completely finished. Therefore there are minor differences in the programming solutions for proximity operators. The Finnish process is not yet adapted.

3 Runs and Results

The runs
We participated in CLEF'2001 with four automated bilingual runs (three language pairs), Finnish – English, Swedish – English and German – English. For all runs, queries were constructed on the basis of the title and description field of the topics.

The main resources

- Motcom Swedish – English translation dictionary (60.000 entries) by Kielikone
- Motcom Finnish – English translation dictionary (110.000 entries) by Kielikone
- Oxford Duden German – English translation dictionary (260.000 entries)
- Morphological analysers: SWETWOL, FINTWOL, GERTWOL and ENGTWOL by Lingsoft
- Inquery retrieval system

German – English translation processes

For German – English we have tested two types of dictionaries (two runs).
Using the Duden German-English dictionary (260.000 words) two translation tables for the 50 CLEF topics were created. The first one included all translations from the dictionary. The second translation table contained the same data, except that *all direct translations of compounds were excluded.* The construction of the German-English translation table was a separate process analysed by a human, following strict syntactic rules for selecting strings from the PC screen. German – English process could not be automated because of interface problems and colour fonts used in the Duden dictionary. However, the translation tables were used automatically.

The results

The results of the four UTACLIR runs are presented below and in Figures 5 and 6.

Testrun	Average precision (CLEF 2001)	(CLEF 2000)
TAYfinstr (Finnish structured)	0,3894	0,2275
TAYswestr (Swedish structured)	0,3769	0,2540
TAYgerstr (German structured)	0,3474	0,2665
TAYgershort (German/short structured)	0,3054	--------

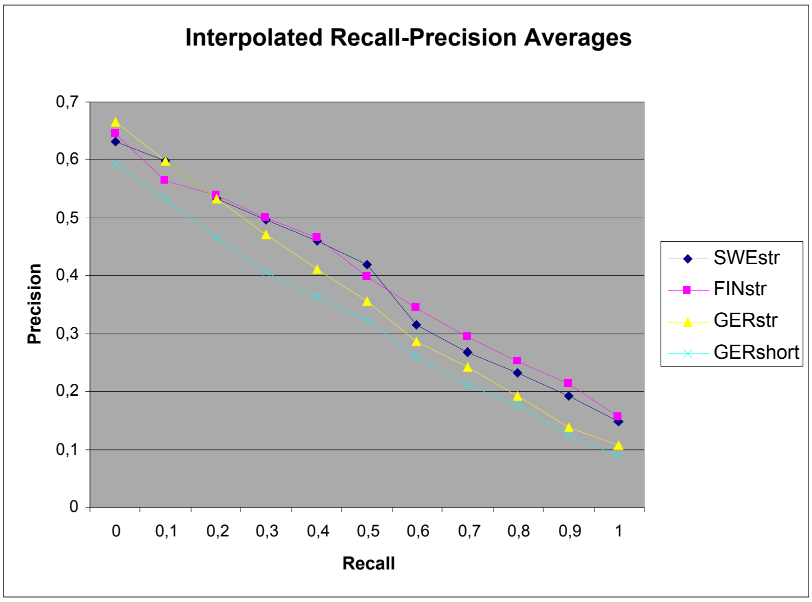

Figure 5. Results: Interpolated Recall-Precision Averages for all four runs.

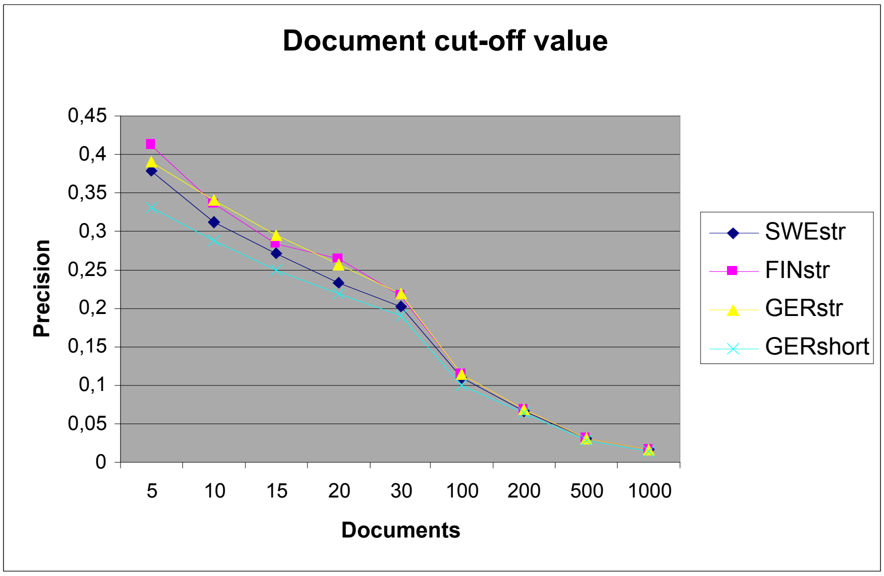

Figure 6. Results: Document cut-off values for the four runs.

Generally the results for all the four runs are good. Improvements are considerable from the last year. Average precision for all the queries shows clear improvements, but there is still great variation in the performance of single queries. Some queries perform exceedingly well getting high scores, but some fail to retrieve relevant documents. This holds for all language pairs.

4 Discussion and Conclusion

After last year CLEF evaluation we identified 1) problems with the translation of proper names (especially true for Finnish) 2) problems with untranslated components in the compound handling process 3) the handling of compounds containing more than two components was not satisfactory 4) normalisation is needed for the dictionary translation in order to match the normalised index of the database.

This year we focused on compound words and proper names. Both are important and common in the source languages and in this year's topics. Therefore the way they are handled considerably affects the results. The new methods for compounds and proper names account for the improvements we achieved this year. However, part of the improvement may be due to the higher number of judged queries.

The choice of using the n-gram technique for all unidentified words was a success. This especially holds for Finnish, in which proper names quite often appear in inflected forms. However, the use of the n-gram technique also increased noise in the form of many nonsense words that were added to the final query. This particularly holds for Swedish and German where all untranslated single words and compound components were handled by the n-gram technique. If the untranslated word was not a proper name, nonsense words were added to the query by the n-gram technique. The Finnish process was different in this aspect since n-gram matching was not used as part of compound processing.

The new compound process was effective for some topics but failed for some topics. The Finnish run did not suffer from nonsense words generated from compounds which increased noise in conjunction with the n-gram technique. On the other hand, because we used last year's process for Finnish (except for the proximity operator), compound components were not translated in all cases. This also had negative effects.

The fogemorpheme algorithm seemed to improve the translations of compound components for both Swedish and German.

The test with two dictionaries for the German runs gives an indication that the new features in the UTACLIR process work well also with a limited dictionary. On the other hand the advantage of a direct translation of compounds is inevitable. Our method for handling compounds works as a good and necessary complement, since no dictionary even a comprehensive one holds entries for all compounds. Compound splitting is needed in several queries in the TAYgerstr-run where a comprehensive dictionary is used. The queries in the TAYgershort-run become very long since all compounds were split into their components. When all the alternative translations for the components are combined to a phrase in the target language query, the number of combinations may be high. Nonsense combinations also occurred quite frequently. On the other hand, generally the process can be said to work as expected because of the relevant combinations.

Acknowledgments

ENGTWOL (Morphological Transducer Lexicon Description of English): Copyright (c) 1989-1992 Arto Voutilainen and Juha Heikkilä.
FINTWOL (Morphological Description of Finnish): Copyright (c) Kimmo Koskenniemi and Lingsoft Oy. 1983-1993.
GERTWOL (Morphological Transducer Lexicon Description of German): Copyright (c) 1997 Kimmo Koskenniemi and Lingsoft, Inc.
SWETWOL (Morphological Transducer Lexicon Description of Swedish): Copyright (c) 1998 Fred Karlsson and Lingsoft, Inc.
TWOL-R (Run-time Two-Level Program): Copyright (c) Kimmo Koskenniemi and Lingsoft Oy. 1983-1992.
MOT Dictionary Software was used for automatic word-by-word translations. Copyright (c) 1998 Kielikone Oy, Finland.

The *Inquery* search engine was provided by the Center for Intelligent Information Retrieval at the University of Massachusetts. This research is part of the research project *Query structures and dictionaries as tools in concept-based and cross-lingual information retrieval* funded by the Academy of Finland (Research Projects 44703; 49157).

References

Gadd, T. 1990. Phonix: The algorithm. *Program*, 24(4), 363-369.

Grefenstette, G., Segond, F. (1997) Multilingual natural language procesing. *International Journal of Corpus Linguistics* 2(1), 153-162.

Haas, S. W., Losee, R. M. Jr (1994) Looking in text windows: Their size and composition. *Information Processing and Management* 30(5), 619-629.

Hedlund, T., Keskustalo, H., Pirkola, A., Seppänen, M., Järvelin, K. (2000) Bilingual tests with Swedish, Finnish and German queries. Working Notes for CLEF Workshop http://www.iei.pi.cnr.it/DELOS/CLEF/Notes.html

Hedlund, T., Keskustalo, H., Pirkola, A., Seppänen, M., Järvelin, K. (2001a) Bilingual tests with Swedish, Finnish and German queries: Dealing with morphology, compound words and query structuring.
In Carol Peters (ed). *Cross-Language Information Retrieval and Evaluation*: Proceedings of the CLEF 2000 Workshop, Lecture Notes in Computer Science 2069, Springer 2001, pp 211-225, forthcoming.

Hedlund, T., Pirkola, A. and Järvelin, K. (2001b). Aspects of Swedish Morphology and Semantics from the Perspective of Mono- and Cross-language Information Retrieval. *Information Processing & Management* vol. 37/1 pp.147-161.

Jacquemin, C. (1996) What is the three that we see through the window: A linguistic approach to windowing and term variation. *Information Processing & Management* 32(4), 445-458.

Levi, J. N. (1978) *The syntax and semantics of complex nominals* London: Academic Press.

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

Pirkola, A., Keskustalo, H., Leppänen, E., and Järvelin, K. 2001. Targeted s-gram matching: a novel n-gram matching technique for cross- and monolingual word form variants. Manuscript.

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*

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

Spyns, P., De Wachter, L. (1995) Morphological analysis of Dutch medical compounds and derivations*. ITL review of applied linguistics Institute of applied linguistics* 109-110, 19-35.

Zhou, J. (1999) Phrasal terms in real-word applications. In Thomek Strzalkowski (ed). *Natural language informations retrieval*. Dordrecht: Kluwer 1999.

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