Bilingual tests with Swedish, Finnish and German queries

Turid Hedlund
Heikki Keskustalo
Ari Pirkola
Mikko Sepponen
Kalervo Järvelin

Department of Information Studies
University of Tampere
Finland

e-mail: hedlund@shh.fi, heikki.keskustalo@uta.fi, pirkola@tukki.jyu.fi, mikko.sepponen@uta.fi, kalervo.jarvelin@uta.fi

Abstract

We used a dictionary-based approach, and performed tests in the bilingual track with three language pairs, i.e., Swedish – English (Swe-Eng), Finnish – English (Fin-Eng), and German – English (Ger-Eng). All the source languages are compound languages, i.e., languages rich in compound words. A *compound word* refers to a multi-word expression where the component words are written together. Our main efforts were to develop techniques for the processing of compounds, to study different types of compound languages, and to study the effects query structuring in different languages. We designed and implemented a method for automated query construction in FIN çSWE çGER -> ENG. The goal of this process is to extract automatically topical information from sentences written in one of the source languages (FIN, SWE, GER) and to create a target language (ENG) query. The resulting query may be either structured or unstructured.

Introduction

NLP-techniques have been tested for IR and CLIR for several years. The point of view has been that linguistically motivated indexing would enable the catching of sense in text and in queries differently from the non-linguistic methods used in IR, for example weighting based on word occurrences. Traditional NLP-techniques are extended also to the sub-word level, i.e., morphological decomposition and stemming (Sparck Jones 1999). So far, any great success in increasing the quality of retrieval result due to these techniques have not been reported, compared to statistical methods. The language dependent linguistic features important to IR and CLIR are, for example, the number of homographic word forms, the way to treat compounds and gender features.

The main problems associated with dictionary-based CLIR are 1) phrase identification and translation, 2) source language ambiguity, 3) translation ambiguity, 4) the coverage of dictionaries, 5) the processing of inflected words, and 6) untranslatable keys, in particular proper names spelled differently in different languages (Pirkola et al. 2000)

Our approach to solve the general problems for bilingual CLIR is based on 1) normalisation in indexing, 2) stopword lists, 3) normalisation of topic words, 4) splitting of compounds, 5) recognition of the right components, 6) phrase composition in target language, 7) bilingual dictionaries, and 8) structured queries.

All the source languages we use, Swedish, Finnish and German are languages rich in compounds, therefore it is essential to develop techniques for the processing of compounds. Our other main interest is to compare structured and unstructured queries to solve the ambiguity problem with CLIR. We used a model developed and tested for Finnish - English CLIR by Pirkola (1998; Pirkola et al. 1999).

Research questions

The research question involves designing and implementing of the approach as a method for automated query construction in FIN çSWE çGER -> ENG, using general bilingual dictionaries. The goal of the process is to extract automatically topical information from sentences written in one of the source languages (FIN, SWE, GER) and to create a target language (ENG) query. The resulting query may be either structured or unstructured. We test the relative effectiveness of the approach and method also in comparison to unstructured queries.

The performed tests include three different language pairs, Swedish - English, Finnish - English, German - English. In the bilingual track we have tested morphological analysis programs, dictionary set-ups and translation approaches. All the source languages are rich in compounds and one of our main effort is the morphological decompositioning of compounds into constituents and their proper translation. We hope to be able to show, that in languages rich in compounds the right translation of compounds is a factor that affect the retrieval result. Homographic word forms (Swedish), especially as components in compounds tend to add many translation alternatives to the query. In our model for treating compounds, where we combine every translation alternative for each component as phrase, a great number of translation alternatives multiplies the possible combinations. A rich inflected morphology (Finnish) is also a factor that affects the retrieval result, particularly when trying to identify and handle proper names.

Approach

Our approach for database indexing in the target language is based on word normalisation, and labelling of unrecognised word forms (e.g. proper names) allowing ambiguity and language inconsistency (e.g. seat belt, seat-belt, seatbelt) in the text.

Our approach in the query formulation process in the source languages included word form normalisation, the removal of source languge stopwords, and compound splitting with proper component base for recognition in dictionaries (e.g. fogemorphemes in Swedish; inflection in Finnish and German). ” Fogemorphemes are morphemes joining constituents in compounds e.g. “s”. We applied phrase construction in the target language of the compounds in the source languages and identification of unrecognised word forms (e. g. proper names). The unrecognised word forms are used as such, disregarding possible inflection. In all these phases we allow ambiguity, i.e. multiple possible interpretations for the source language word forms. The translation is structured using synonym set to reduce ambiguity effects, and based on bilingual dictionaries.

Research setting

Document collection
The LA Times document database was indexed as document collection. The morphological analysis program ENGTWOL, producing normalised word forms and marking unrecognised word forms, was used as part of the index building process.

Test topics
The provided test topics include title, description and a narrative. For CLIR purposes and automated queries it seems favourable to keep the test requests relatively short 2-3 sentences. Therefore we automatically selected the title and description field only. We used the Finnish, Swedish and German test topics.

Description of the query formulation process

As translation method for all the languages we used the dictionary method, and automated queries. For Swedish, Finnish and German, compound splitting and the translation of constituents were performed. For Swedish, we added a special algorithm for recognising and handling the ”fogemorphemes” and thus allowing us to treat all components as base forms. The morphological analysis program for normalisation and compound splitting used was TWOL (SWETWOL, FINTWOL, GERTWOL, ENGTWOL) by Lingsoft, Inc. The retrieval system used is InQuery, which allows the use of a structuring operator ”SYN” for translation alternatives. We tested the relative effectiveness of structured and unstructured queries over the language pair GER - ENG.

The test topics in the chosen source languages, in our case Swedish, Finnish and German are run through a normalisation process. If a compound is lexicalised and found in the machine-readable-dictionary used, this translation is probably less ambiguous than translating the constituents and is therefore used. For all other compounds, compound splitting is taken place. Compounds in Swedish need special treatment since we know from earlier tests that the morphological analyser for Swedish does need tuning to give proper results for IR purposes (Hedlund et al. 2000). To

solve this problem we have developed an algorithm. All the constituents of a compound should be returned to the lexical base form, which should be a real word and not a stem. In case of German nouns as constituents, they need to get an upper-case initial letter. Proper names and other words not found in the dictionary are added to the query as such.

The structured Swedish-English query processing was implemented using the Tcl programming language. Tcl is convenient language for combining existing software and for processing strings and lists of strings. Swedish -> English translations are obtained using the Motcom dictionary software. Motcom's output contains a lot of information intended for a human reader. The actual translations are obtained from the output of Motcom by a filtering script. For word normalisation we used the SWETWOL program.

The structured German-English query processing was implemented using C programming language for reading and manipulating four input files: CLEF topic file, German stop word file, English stop word file and the Duden German-English translation table for the 40 CLEF topics. German Morphological analysis was accomplished by calling the library function of the analyser software GERTWOL. The making of the German-English translation table was a separate process accomplished by a human analyser following strict syntactic rules for selecting strings from the PC screen. As the selection of the strings was based on the font colour, this process could not be automated.
The unstructured German-English query processing was a simple modification of the corresponding structured German-English processing.

The structured Finnish-English query translation program was also modified from the programming code of the structured German-English query translation. Finnish-English word-by-word translations were accomplished by using a command line type of electronic Finnish-English MOT dictionary as in case of Swedish translation where the corresponding Swedish-English MOT dictionary was used. A filtering script produced, in most cases, a "clean" stream of individual words or phrases as English translation equivalents for a corresponding Finnish word.

The *query structuring* was done by using the *syn* operator provided in the INQUERY retrieval software. Every translation alternative for a word in the MRD is added to the query as a synonym. The Synonym operator's syntax is: #syn(T1 ... Tn), where Ti are terms. The terms in this operator are treated as instances of the same term. I.e., the translation of the word *möte,* #syn(*encounter meeting crossing appointment date*). A compound in the source language that is translated by a dictionary as a phrase need to be marked with a proximity operator. The Ordered Distance operator's syntax is: #N (T1 ... Tn) or #odN (T1 ... Tn), where N is the distance, and Ti terms. The terms within an ordered distance operator must be found within N words of each other in the text in order to contribute to the document's belief score. The #N version is an abbreviation of #odN; therefore #3(health care) is equivalent to #od3(health care).

The Weighted Sum operator's syntax is #wsum (Ws W1 T1... Wn Tn), where Ws is the query weight, Wi are term weights for the terms Ti. The terms are considered according to the weight associated with each (Wn). The final belief score is scaled by Ws, the weight associated with the #wsum itself. For example: #wsum(1 1.0 architecture 2.0Berlin) This example weights Berlin twice as heavily as architecture.

The automatic query construction process takes the following 5 resources as inputs:
1) the CLEF topic file in source language
2) a file, or files containing stopwords in the source language (SWE, FIN, GER)
3) a file containing stopwords in the target language (ENG)
4) a bilingual translation dictionary for each language pair
5) a morphological analysis program for each source language

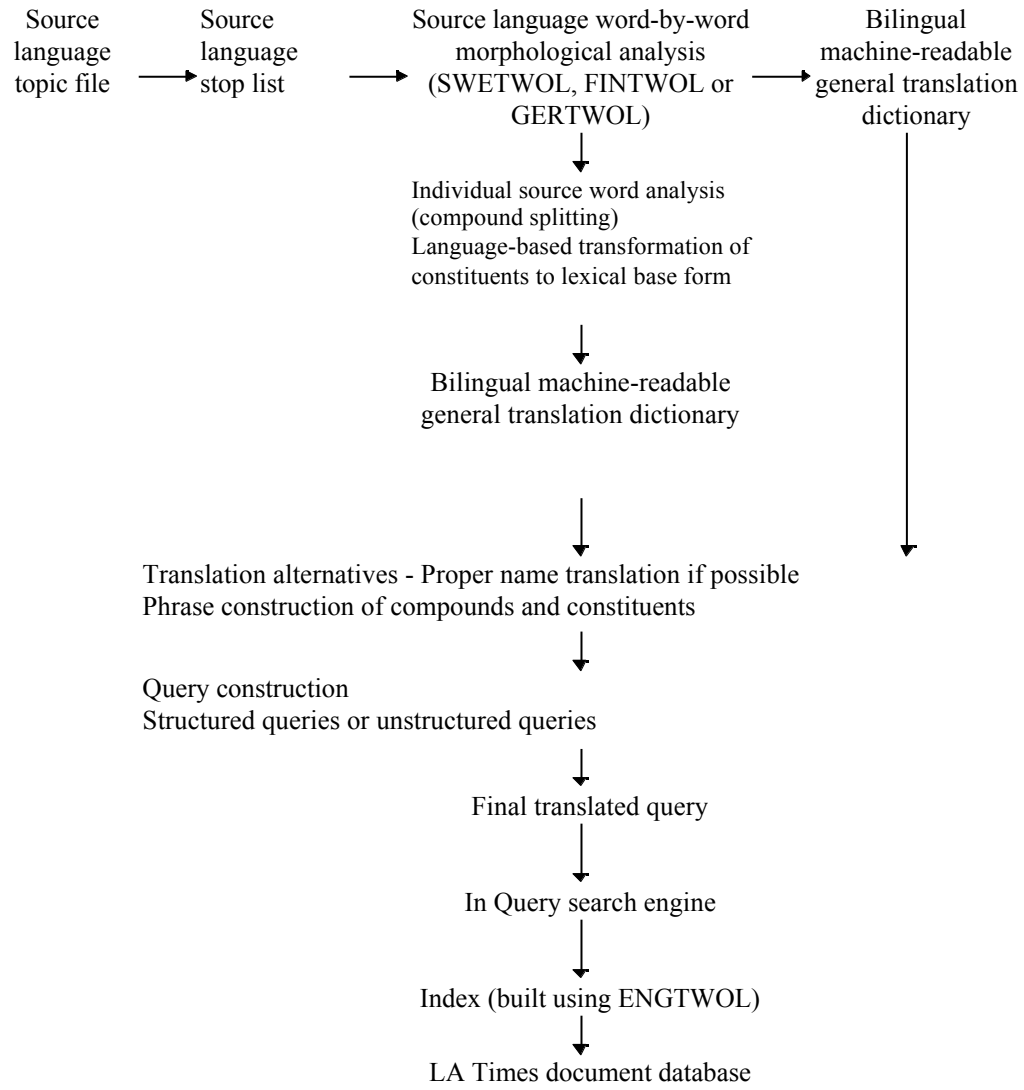

Figure 1. General description of the automatic query forming process

Evaluation

Analysis of the problems in the query formulation process

Technical problems

-proper names do not match the index words in the document database, i.e., the forms "usa" and "united states" are not recognised. The form "unite state" would have matched the LA Times index.

- the second problem of technical art is that any words translated in English by a dictionary still have to be normalised by ENGTWOL. For example, the query words "taking" and "drugs" never matched any index words of the LA Times database. The reason for this is that the ENGTWOL program used in the index building process produced word forms "take" and "drug", respectively, into the index of the database.

Both these problems can be solved if we run the dictionary translation through the morphological analyser, thus normalising all word forms in the same way as they appear in the document database index.

Semantical problems

-length of the target queries. There appears to be great variation to the length of the queries due to:

1) dictionaries, and the numbers of translation alternatives for a word.

2) compound words in the source language. When splitting compounds into three or four constituents the amount of translation alternatives and their combinations grow rapidly.
3) homographic words with many senses. Frequent words not in the stop list of the source language tend to have many senses, and they also tend to appear as constituents in compound words.

- important concepts are not translated, which tend to ruin the whole query.
due to:
1) dictionaries, if the word is not in the dictionary it is used as such in the query.
2) compound words have constituents that are not translated and due to this the translated phrases come to include words in the source language never appearing together with the translated ones in the document text.

Language specific problems

Swedish:
The morphological analyser needs to be tuned for the normalisation of constituents when splitting compounds. The special algorithm we used for handling fogemorphemes appears to work well in the query formulation process and reduces the number of non-translated words in several topics. However, since we deal with constituents of compounds the actual effect on the search result is dependable also on other factors, such as to what extent the constituent bear important search keys.

German:
The German language has the special feature of capital initial letter in nouns, also the use of the double "s" ß in text. We utilised morphological information of nouns in German in order to match German noun keys more precisely into translation dictionary entries. The capital initial letter was identified in all the input files: CLEF topic file, German stop word file and the Duden German-English translation table for the 40 CLEF topics. When splitting the compounds the noun constituents also had to get the capital initial letter in order to be identified.

Finnish:
The Finnish language is special in having a very rich inflectional morphology, and instead lacking prepositions. The morphological analyser works well and the normalisation process has no greater obstacles. The problems are caused by inflectional forms of proper names. These typically cannot be normalised since the morphological analysis program cannot identify them.

Test runs
The results of the four test runs show comparable performance for three separate source languages.
The best average performance is by the German structured run, and the lowest by the German unstructured.
The average precision over recall levels are as follows:

German structured (Gerstr)	26,7
Swedish structured (Swestr)	25,4
Finnish structured (Finstr)	24,5
German unstructured (Geruns)	21,6

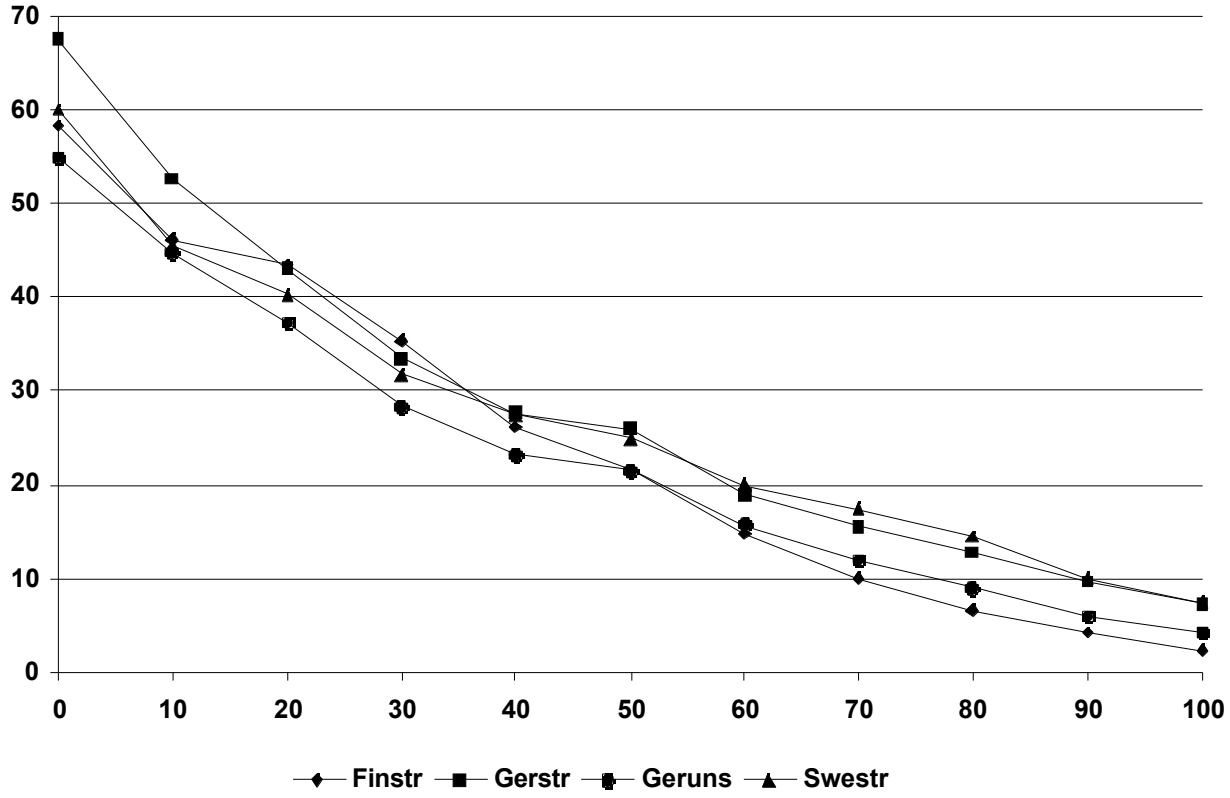

Figure 2. Interpolated recall - precision averages

Examining the query performance of our runs for each 33 topics we find that our results in general tend to be above the median value for all the participating runs. On the other hand we can report very good results for some topics and then complete failures for some, the variation is quite large. This is true for all the language pairs. We have discussed some of the reasons for this above, when discussing problems in the query formulating process.

Comparing the results for the structured and unstructured German - English queries, we get a better performance for the structured queries. Our earlier findings (Pirkola 1998) with Finnish - English CLIR suggest that the difference in performance for this language pair is larger. We have been testing structured / unstructured queries also for the other language pairs as extra runs after we got the relevance assessments from CLEF, and it seems that the Finnish - English queries also now tend to differ more. For Swedish - English structured / unstructured queries the difference is about the same as for German - English. We are not in this early stage able to say if this is a language dependent phenomenon or if there is some other reason for this.

Acknowledgements

InQuery (TM) SOFTWARE Modifications Copyright (c) 1998-2000 by the Center for Intelligent Information Retrieval (CIIR) at the University of Massachusetts at Amherst. All rights reserved. InQuery (TM) Copyright (c) 1996-2000 by Dataware Technologies, Inc., Hadley, Massachusetts, U.S.A. (413-587-2222; http://www.dataware.com). All rights reserved. The InQuery (TM) software was developed in part at the Center for Intelligent Information Retrieval (CIIR) at the University of Massachusetts at Amherst (For more information, contact 413-545-0463 or http://ciir.cs.umass.edu). InQuery (TM) is registered trademark of Dataware Technologies, Inc.

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 Oy. 1983-1993.

References:

Hedlund, T., Pirkola, A. and Järvelin, K. (2000). Aspects of Swedish Morphology and Semantics from the Perspective of Mono- and Cross-language Information Retrieval. Forthcoming in *Information Processing & Management* vol. 37/1 pp.147-161 dec. 2000.

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., Hedlund, T., Keskustalo, H., Järvelin, K. (2000). Cross-Lingual Information Retrieval Problems: Methods and findings for three language pairs. Accepted for publication in *ProLISSa Progress in Library and Information Science in Southern Africa. First biannual DISSAnet Conference.* Pretoria, 26-27 October 2000.

Sparck Jones, K. (1999). What is the role of NLP in text retrieval. In T. Strzalkowski (Ed.) *Natural language information retrieval.* Dordrecht: Kluwer Academic Publishers.