RETRIEVING INFORMATION IN ONE LANGUAGE VIA ANOTHER

by Erica Cosijn, Theo Bothma, Heikki Keskustalo, Ari Pirkola & Kalervo Järvelin

Is it possible to retrieve information in one of the official languages of South Africa by using a query in another – and how effective is it? This is the research topic of a joint project by the Department of Information Science of the University of Pretoria and the Department of Information Studies of the University of Tampere, Finland.

Cross-language Information Retrieval Systems (CLIR) involve query translation (from source to target languages) and/or document translation (from target languages to source). The latter requires good machine translation systems (not yet available), while the former requires the user to read the retrieved documents. There are three approaches to CLIR: machine translation, the use of parallel corpora and dictionary translation (bilingual translation dictionaries). The dictionary translation process is represented in → *1*.

The University of Tampere's CLIR framework for isiZulu-English and Afrikaans-English was tested using the InQuery retrieval system. In InQuery, queries can be presented as a "bag of words", or they can be structured using a variety of query operators. All query keys are attached with a belief value. The InQuery query operators are prefix operators marked by the hash sign "#", i.e. #sum(query) or #syn(query). For #sum-operator, the system computes an average weight of query keys (or subqueries). The #syn-operator treats its operand keys as synonymous instances of the same key.

Submitting isiZulu queries to an English database

As isiZulu is a highly inflectional, agglutinative language, and morphological analysers and parsers were unavailable at the time of this research, approximate string matching techniques were used in order to match the word forms in the running text of the query to the normalised word forms (verb roots and noun stems) in the translation dictionary.

The following types of string matching were tested: edit distance, longest common substring matching, digrams, trigrams and skipgrams.

Skipgrams were effective in matching the inflected form to the normalised form of the word. Using skipgrams, the three best matches to the inflected word form were calculated and translated into English in order to form an English language query to be matched to the English language database. The InQuery retrieval system was used for query construction and the original English queries were run as baseline queries. In many CLIR studies, the Pirkola method, i.e. treating translation equivalents as synonyms and combining them by the InQuery syn-operator, has been demonstrated to perform well. The research team therefore formulated two types of syn-structured queries, as well as explicit queries.

In syn1, CLIR queries all the translations of the three best matches of an isiZulu topic word were combined with the syn-operator. For example, the three best matches and their translations for the isiZulu word *esesabekayo* are the following:

- esabeka: capable, feared, fearful, terrible, awe, inspiring, prodigious, wonderful
- sabeka: fearful, terrible, wonderful
- esabela: respond

The syn-statement is as follows: #syn(capable feared fearful terrible awe inspiring prodigious wonderful fearful terrible wonderful respond).

In syn2, CLIR queries the translations of each of the three matches of an isiZulu topic word were combined with the syn-operator. For example, for esesabekayo, the syn-statements are the following: #syn(capable feared fearful terrible awe inspiring prodigious wonderful) #syn(fearful terrible wonderful) #syn(respond).

In both cases (syn1 and syn2) the syn-statements were combined with the sum-operator. The results are listed in the table below.

Table 1: Results: isiZulu-English CLIR

Query type N=50	Precision at 10% recall	Average precision
English - Original English	54.8	34.3
Explicit	5.9	4.0
Syn1	32.1	21.5
Syn2	17.2	11.7

Compared to the English-English baseline, the isiZulu-English results were poor.

Present research in CLIR concentrates on languages with comparable vocabularies in terms of, inter alia, technical and scientific terminology. This research has shown that a set of new problems will be encountered if the language pairs used contain disparate vocabularies.

Submitting Afrikaans queries to an English database

This study reports on the first-ever experiments that apply dictionary-based query translation techniques to Afrikaans queries submitted to an English database. The original English queries were run as baseline queries. Two types of queries were used: a title query (a shorter sentence with few query keys) and a combination of title and description. Combining a bilingual dictionary, a morphological analyser and a stopword list, the Afrikaans queries were translated into English and matched with the English target database. Morphologically, the nature of Afrikaans is quite simple and therefore an Afrikaans morphological normaliser for information retrieval was developed as part of this study.

The Afrikaans to English query translation is based on the University of Tampere's Cross-language Retrieval framework. The main features adopted in this study included normalising the target index, utilising source and target language stopword lists, normalising source keys, splitting source compounds and recognising their components when unable to translate the word as a whole, using a bilingual translation dictionary, performing approximate string matching for untranslatable source keys and structuring the final target query.

The normaliser utilises a word list containing around 82 000 unique Afrikaans single word entries. The ISO 8859-3 character set is used for expressing the characters with diacritics. The normaliser automatically categorises the input string belonging to exactly one of seven distinct key types. This process is represented in → *2*.

The results are promising

The results showed that for the test collection of 35 queries, using both title and descriptor fields, the relative average precision was 60.6%, and 68% precision at 10% recall when compared to the baseline. In comparison to other languages, the results were satisfactory. An analysis of the results shows that CLIR is a useful process, which should be refined further.

Table 2: Results: Afrikaans-English CLIR

Query type N=35	Precision at 10% recall	Average precision
Title only English – English	44.9	28.5
Title only Afrikaans – English	25.0	13.6
Title and description English – English	48.2	32.0
Title and description Afrikaans – English	32.8	19.4

An electronic bilingual translation dictionary was engineered to be used for CLIR specifically. The process has not yet been optimised, and the dictionary can be filtered even further to exclude Afrikaans words in the translated #syn sets, and to reduce the length of dictionary entries. The normaliser performed quite well in terms of the identification and normalisation of plural forms and past tense verbs, and longest common substring matching to normalise words with a variety of suffixes and compound splitting (except in cases where individual parts of the compounds were not found in the dictionary or where the beginning of a proper noun is also an entry in the dictionary).

Hyphenated compounds were not handled correctly. This will be corrected in a future version of the normaliser. The stopword list was found to be adequate, with a few exceptions. A frequency analysis should be done on homographs that are also stopwords, and the list modified accordingly.

Future research

CLIR in the South African context is an important area of research. The process can be useful in indigenous knowledge systems, where it is often necessary to capture indigenous knowledge in the original language and make it accessible to users not fluent in that language. Future projects include building a parallel corpus of English-Zulu texts, Sepedi-English CLIR, fine-tuning the Afrikaans dictionary and morphological analyser and filtering the Afrikaans dictionary to establish "common sense" word usage.

Further reading

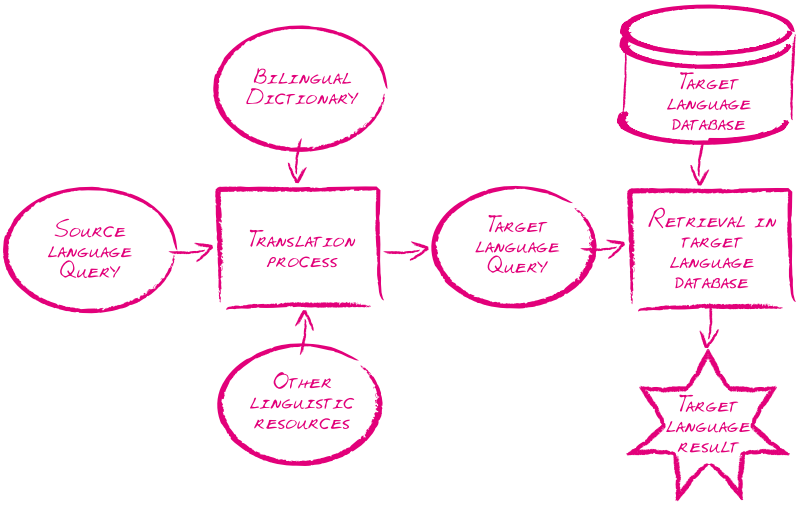
Cosijn, E., Keskustalo, H., Pirkola, A., De Wet, K. & Järvelin, K. 2004. Afrikaans-English cross-language information retrieval. In: Progress in library and information science in Southern Africa: proceedings of the third biennial DISSAnet Conference (ProLISSA3). Edited by Bothma, T. & Kaniki, A. Pretoria: Infuse: pp. 97-110.

Cosijn, E., Pirkola, A., Bothma, T.J.D. & Järvelin, K. 2002. Information access in indigenous languages: A Zulu case study. In: Emerging frameworks and methods: proceedings of the Fourth International Conference on Conceptions of Library and Information Science (CoLIS4). Edited by Bruce, H. et al. Seattle, WA, USA: Libraries Unlimited: pp. 221-238.

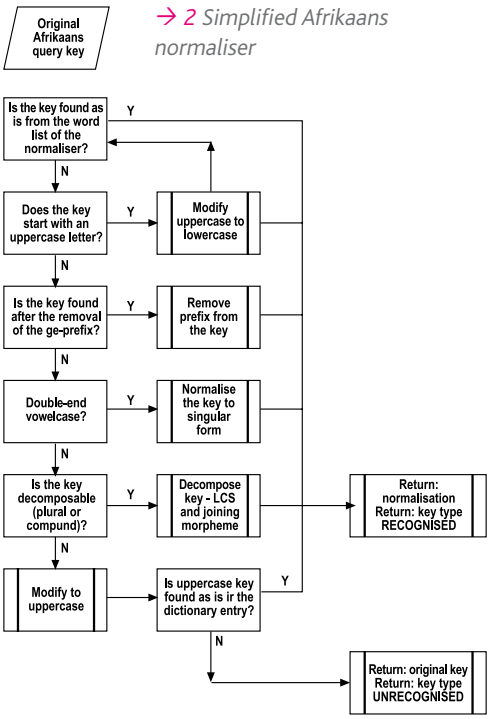
Cosijn, E., Pirkola, A., Bothma, T.J.D. & Järvelin, K. 2002. Information access in indigenous languages: a Zulu case study. South African Journal of Libraries and Information Science, 68(2): 103-114.

Dr Erica Cosijn *is a senior lecturer and Prof. Theo Bothma is head of the University of Pretoria's Department of Information Science. Heikki Keskustalo, Ari Pirkola and Kalervo Järvelin are associated with the Department of Information Studies of the University of Tampere, Finland.*

erica.cosijn@up.ac.za



→ *1 Cross-language information retrieval using a bilingual translation dictionary to translate queries*



→ *2 Simplified Afrikaans normaliser*