Some Experiments with the Dutch Collection

Arjen P. de Vries
CWI
Amsterdam
The Netherlands
arjen@acm.org

Anne Diekema
CNLP, Syracuse University
Syracuse NY
USA
diekemar@syr.edu

Abstract

We performed some basic monolingual Dutch and bilingual English–Dutch experiments. The retrieval approach is very basic, without stemming or decompounding, using only a simple language model to rank the documents. In the bilingual task, the English queries are analyzed by a system for question answering. The resulting queries are translated by dictionary lookup and ranked by the same basic retrieval system used in the monolingual task.

1 Introduction

Inspired by shared observations at the first CLEF workshop[1], the authors decided to initiate work on retrieval experiments that help understand the effect of the quality of the translation process on retrieval results.

We are primarily interested in the problem of multi-lingual retrieval from Dutch collections, and as we believe the quality of the resources for translation a significant factor in the retrieval results, we decided to focus on the bilingual English to Dutch task. Limiting ourselves to this task, we could deploy two high quality resources:

- a natural language processing toolkit aimed at (English) Question Answering, developed at CNLP [DLC+01];
- the CD-ROM edition of the (excellent) 'Van Dale Groot woordenboek', a dictionary for English to Dutch translation (and vice versa) [vD97].

2 Experimental setup

For question processing we used the language to logic module from the Center for Natural Language Processing (CNLP) as the front-end of our system. The language to logic (L2L) module converts a natural language query or question into a generic logical representation, which can be interpreted by different search engines. The conversion from language to logic takes place based on an internal query sublanguage grammar, which has been developed by CNLP. Prior to conversion, query processing such as stemming, stopword removal, and phrase and Named Entity recognition take place. Certain terms in the question are expanded with their synonyms or spelling variants, e.g., the term 'Koweit' upon encountering 'Kuwait'.

The terms occuring in the resulting expression are looked up in the dictionary. The Van Dale dictionary has been developed for interactive usage on the desktop only, for example to find a translation while writing a text document with your favourite word processor. Unfortunately, it lacks a command-line interface, which has rendered it unexpectedly difficult to apply as a

[1] 'We want better resources.'

component in an automatic translation system. As a workaround, we developed a screen-scraping tool based on the Win32 modules for the Perl scripting language: we discovered that the Van Dale application supports requests to lookup query words using DDE (a Windows protocol for data exchange). The results, displayed in the results pane, are copied through the Clipboard into our script by emulating the right sequence of keystrokes. The data captured on the Clipboard is then parsed and converted into a query-specific dictionary. Because some terms generate a large number of alternative translations for many different senses, we set some ad-hoc thresholds: a maximum amount of 10 translations per term, from a maximum of 5 different senses, but taking never more than 3 translations per sense.

The results of these two steps (as well as the documents in the collection) are converted to lowercase and stripped from 'strange' characters, and stopwords are removed. The retrieval backend is a database implementation of the simple – but proven effective – language models developed by Hiemstra [Hie01] (more information about the retrieval backend is given in [HdV00]). We intended to perform our experiments with an improved implementation that processes both phrases and disjunctive queries, but we did not finish our implementation work in time, so we have used the simple term-based model.

3 Analysis

Run	Mean average precision
AAmoNLtd	0.399
AAmoNLt	0.348
AAbiENNLtd	0.162
AAbiENNLt	0.133

Table 1: Results of the submitted runs.

The results of our experiments are summarized in Table 1. We submitted four runs, their names encoding the task – monolingual ('mo') or bilingual ('bi') – and the portion of the topics that has been processed: title only ('t') or title and description ('td').

The difference in mean average precision between title-only and title-and-description topics is surprisingly small in both tasks, especially since the title-only topics are quite short (2.5 word on average). A very large difference is found in topic C110, which is however explained easily since the description gives the name 'Kazem Radjavi' and the title does not. Apparently, only a small number of query terms really help retrieving relevant documents, and those query terms do usually occur in both title and description. Further analysis is warranted to explain the small drop in performance.

The significantly decreased mean average precision of the bilingual runs when compared to the monolingual runs demonstrates that the query translation component of our system requires more work. A main cause of our disappointing results is the approach of using the Van Dale dictionary through screen-scraping. First, communication via Clipboard cut-and-paste sometimes malfunctioned: the data would not appear on the Clipboard, probably due to timing problems. This makes it particularly difficult to check whether no translation occurred in the dictionary, or, the answer is not there due to communication problems. For example, a term like 'space probe' is found in the Van Dale dictionary, but the translation was unfortunately 'dropped' by our script (other examples are 'telephone', 'administration' and 'fishing'). Second, the Van Dale application performs a fuzzy match if the query term is not found, but checking whether that has happened would require an additional cut-and-paste of a different text pane. Finally, the copied data is not trivially interpreted by a machine, as it does contain instructions aimed at people, such as 'compare to' or 'see also'.

A deeper problem with our current approach lies in the interaction between different process steps. As a simple example, the front-end adds 'Koweit' as an alternative for 'Kuwait', but this alternative does not exist in the dictionary so is ignored further on in the process. In this particular case, it does not hurt effectiveness during the retrieval step, as 'Koweit' will not be found in the Dutch collection. A more interesting example of this problem is exposed when the additional intelligence in the front-end actually reduces effectiveness. A particularly good example of this case is provided by topic 91, *AI in Latin America*. The L2L module makes two wrong assumptions in this case: 'AI' is not 'Artificial Intelligence', and 'America' does not always imply 'United States'. The final result of the process is the following complex query: ('Artificial Intelligence' $\vee$ ('United States' $\vee$ 'United States of America' $\vee$ 'US' $\vee$ 'USA'). Of course, the original untranslated query terms are added, but given that the retrieval model emphasizes frequent query words, most of the results discuss indeed artificial intelligence in the US, and not Amnesty International in Latin America as desired.

4 Next Steps

Summarizing our current experimental results, we must conclude we have not made much progress since the first CLEF workshop. Still, we find ourselves in need of basic resources such as the dictionary; while the chosen 'Van Dale' seems an excellent tool for interactive usage during a word processing session on a Windows desktop, its application in an automatic retrieval system seems difficult. Although some of the ambiguities in the translation instructions can be interpreted automatically, the current screen-scraping solution is not sufficiently reliable to base further retrieval experiments upon.

A more interesting challenge is to find a balance between the sophisticated (but sometimes mistaken) analysis of the query in the L2L module, and the brute-force term counting of the statistical retrieval model. While we are convinced that it should (some day) be possible to improve retrieval results with an intelligent analysis of the query, it is not yet clear how to detect – without human intervention – that a rule like 'America → US' does not apply for 'Latin America'.

The most urgent work to be done is however to finalize the implementation of an adaptation of our retrieval model that takes into account phrases and disjunctions. Even though most components are in place, more engineering is needed before these experiments can be performed.

References

[DLC$^+$01] A. Diekema, X. Liu, J. Chen, H. Wang, N. McCracken, O. Yilmazel, and E.D. Liddy. Question Answering: CNLP at the TREC-9 Question Answering Track. In E.M. Voorhees and D.K. Harman, editors, *Proceedings of the Nineth Text Retrieval Conference TREC-9*, number 500–249 in NIST Special publications, pages 501–510, 2001.

[HdV00] Djoerd Hiemstra and Arjen de Vries. Relating the new language models of information retrieval to the traditional retrieval models. Technical Report TR–CTIT–00–09, Centre for Telematics and Information Technology, May 2000.

[Hie01] D. Hiemstra. *Using language models for information retrieval.* PhD thesis, Centre for Telematics and Information Technology, University of Twente, 2001.

[vD97] Groot woordenboek Nederlands-Engels/Engels-Nederlands. CD-ROM, 1997. Versie 1.0.