A Poor Man's Approach to CLEF

Arjen P. de Vries[1,2]

[1] CWI, Amsterdam, The Netherlands
[2] University of Twente, Enschede, The Netherlands
arjen@acm.org

Abstract. The primary goal of our participation in CLEF is to acquire experience with supporting cross-lingual retrieval. We submitted runs for all four target languages, but our main interest has been in the bilingual Dutch to English runs. We investigated whether we can obtain a reasonable performance without expensive (but high quality) resources; we have used only 'off-the-shelf', freely available tools for stopping, stemming, compound-splitting (only for Dutch) and translation. Although our results are encouraging, we must conclude that a poor man's approach should not expect to result in rich men's retrieval results.

1 Goals

The Mirror DBMS [2] aims specifically at supporting both data management and content management in a single system. Its design separates the retrieval model from the specific techniques used for implementation, thus allowing more flexibility to experiment with a variety of retrieval models. Its design based on database techniques intends to support this flexibility without causing a major penalty on the efficiency and scalability of the system. The support for information retrieval in our system is presented in detail in [3], [1], and [4].

The primary goal of our participation in CLEF is to acquire experience with supporting Dutch users. Also, we want to investigate whether we can obtain a reasonable performance without requiring expensive (but high quality) resources. We do not expect to obtain impressive results with our system, but hope to obtain a baseline from which we can develop our system further. We decided to submit runs for all four target languages, but our main interest is in the bilingual Dutch to English runs.

2 Pre-processing

We have used only 'off-the-shelf' tools for stopping, stemming, compound-splitting (only for Dutch) and translation. All our tools are available for free, without usage restrictions for research purposes.

C. Peters (Ed.): CLEF 2000, LNCS 2069, pp. 149–155, 2001.

Table 1. Size of the stoplists used.

Language	#words
Dutch	124
English	95
German	238
French	218
Italian	133

Stopping and Stemming

Moderately sized stoplists, of comparable coverage, were made available by University of Twente (see also Table 1).

We used the stemmers provided by Muscat[1], an open source search engine. The Muscat software includes stemmers for all five languages, as well as Spanish and Portuguese. The stemming algorithms are based on the Porter stemmer.

Dictionaries

The Ergane translation dictionaries[2] were made available by Gerard van Wilgen. To avoid the necessity of a bilingual wordlist for every possible language combination, Ergane uses the artificial language Esperanto as an interlingua. Ergane supports translation from and to no less than 57 languages, although some languages are only covered by a few hundred words. The number of entries in the dictionaries used are summarized in Table 2.

Table 2. Number of entries in the Ergane dictionaries.

Language	#words
Dutch	56,006
English	15,812
French	10,282
German	14,410
Italian	3,793

Because of synonyms, the size of bilinugal dictionaries might actually be bigger than the size of the smallest word-list of a language pair. After removal of multiword expressions, the number of Dutch entries in the bilingual translation lexicons are presented in Table 3.

Note that these dictionary sizes are really small compared to dictionaries used in other cross-language retrieval experiments. For instance, Hiemstra and Kraaij have used professional dictionaries that are about 15 times as large [6].

[1] http://open.muscat.com/

[2] http://www.travlang.com/Ergane/

Table 3. Sizes of the bilingual dictionaries (from Dutch to target language).

Target	#words
English	20,060
French	15,158
German	15,817
Italian	6,922

Compound-Splitting

Compound-splitting was only used for the Dutch queries. We applied a simple compound-splitter developed at the University of Twente. The algorithm tries to split any word that is not in the bilingual dictionary using the full word-list of about 50,000 Dutch words from Ergane. The algorithm tries to split the word in as little parts as possible. It encodes a morphological rule to handle a property known as 'tussen-s', but it does not use part-of-speech information to search for linguistically plausible compounds.

Because the Dutch word-list used for splitting was much larger than the number of entries in the bilingual dictionaries, compound-splitting might result in words that are only partially translated. For example, the Dutch word 'wereldbevolkingsconferentie' (topic 13, English: 'World Population Conference') was correctly split in three parts: 'wereld', 'bevolking' and 'conferentie' of which only the first two words have entries in the Dutch-to-French dictionary.[3]

3 System

For a detailed description of our retrieval system, we refer the interested user to [3]. The underlying retrieval model is best explained in our technical report[4] [5]. It supplements the theoretical basis of the model with a series of experiments, comparing this model with other, more common retrieval models.

4 Results

This section discusses the results obtained with our system. We discuss the retrieval results expressed in average precision, and, the coverage of our translations. After discussing the official runs, we present some tests performed with pre-processing Dutch topics.

4.1 Official Results

All experiments were done using the title and description fields of the topics. The average query length for Dutch was 10.5 after stopping (which is of course

[3] This example also illustrates the 'tussen-s' rule: the 's' between 'bevolking' and 'conferentie' has been correctly removed.

[4] http://wwwhome.cs.utwente.nl/~hiemstra/papers/index.html#ctit

Table 4. Summary of results (after fixes).

	# queries	Average Prec.	R-prec.
English	33	0.4070	0.4163
French	33	0.4090	0.3831
German	36	0.3134	0.3149
Italian	36	0.3980	0.3935
Bi-lingual	32	0.2375	0.2392
Multi-lingual	39	0.1018	0.1448

Table 5. The submitted, flawed results.

	# queries	Average Prec.	R-prec.
German	37	0.1794	0.2032
Multi-lingual	39	0.0864	0.1330

rather long compared to the average query size people enter in e.g. web search engines).

Table 6 summarizes our results. The second column shows the number of queries with hits in the monolingual runs; the third and fourth columns show the mean average precision[5]. The monolingual results for English have been based on the bilingual qrels. The last column summarizes the drop in average precision that can be attributed to the translation process.

Table 6. Official results (after fixes).

	# queries	Monolingual	Dutch → X	relative
English	33	0.4070	0.2303	57%
French	34	0.4090	0.1486	36%
German	37	0.3134	0.1050	34%
Italian	34	0.3980	0.0989	24%

We hypothesize from the relatively low average precision (0.3134) on the monolingual German task that we really have to perform compound-splitting of this corpus. Another possible cause of the lower score for German is that we had to merge the runs from the two subcollections, which were handled separately. But, our experiments on TREC-8 showed that this cannot really explain such a performance drop.

We attribute the large drop in performance for e.g. the bilingual Italian task (only 24% of the average precision of the monolingual task) to the small coverage of our translation dictionaries. The coverage of the topic translations produced has been summarized in table 7.

[5] The mean average precision for the bilingual runs as given by `trec_eval`, normalized for the number of queries with hits in the monolingual case.

Together, the inferior results on German and Italian explain the disappointing average precision obtained on the multilingual retrieval task (0.0864).

Table 7. Coverage of the translations (40 queries).

experiment	total terms	not translated	relative
Dutch → English	420	92	22%
Dutch → French	420	138	33%
Dutch → German	420	115	27%
Dutch → Italian	420	199	47%

4.2 Morphological Normalisation and Compound-Splitting

Our primary goal with CLEF participation is to test whether we could provide a Dutch interface to our retrieval systems. To confirm our intuition about stemming and compound-splitting, we performed some test runs to analyze the effects of morphological normalisation and compound-splitting for Dutch. We either performed stemming or not, and performed compound-splitting or not, resulting in four variants of the system:

nlen1: base-line translation using full-form dictionary
nlen2: translation using Dutch stemmer and a dictionary with stemmed entries
nlen3: translation using compound-splitter for Dutch and full-form dictionary
nlen4: translation using compound-splitter and dictionary with stemmed entries

The results of these runs are summarized in Table 8. We conclude that compound-splitting is very important, and stemming seems a useful pre-processing step.

Table 8. Results on Dutch runs (33 queries).

run	average precision	improvement
nlen1	0.1726	
nlen2	0.2228	29%
nlen3	0.1912	11%
nlen4	0.2303	33%

To support these conclusions, Table 9 summarizes the coverage of the various translations used in the Dutch runs. Compound-splitting and morphological stemming of Dutch words nearly triples the relative coverage of the translation dictionaries. The total of 92 untranslated Dutch terms in the English queries

Table 9. Coverage of the translations (40 queries).

experiment	total terms	not translated	relative
nlen1	366	201	57%
nlen2	366	130	36%
nlen3	420	160	38%
nlen4	420	92	22%

include about 13 proper names like 'Weinberg', 'Salam' and 'Glashow' (topic 2) and a few terms that were left untranslated in the Dutch topics like 'Académie Française' (topic 15) and 'Deutsche Bundesbahn' (topic 40).

5 Conclusions and Future Work

Summarizing our experiments, we may conclude that our retrieval models works well for all monolingual runs, except for German. Future experiments will have to confirm whether a process like compound-splitting will indeed bring our monolingual results to a level comparable to the other languages. The influence of compound-splitting of Dutch topics on the bilingual results raises our expectations on this end.

We were not at all unhappy with our bilingual results. But, from the coverage of the translations, we still have to conclude that a poor man's approach should not expect to result in rich men's retrieval results. However, we cannot blame it all on the dictionaries. The current version of our retrieval system does not use query expansion techniques to improve mediocre translations; it remains to be seen if better statistical techniques can bring us closer to the results obtained with 'proper' linguistic tools.

Acknowledgements

Without Djoerd Hiemstra's help, I would never have obtained CLEF results: he helped me with most of the pre-processing. More importantly, our discussions about his and competing retrieval models has improved significantly my understanding of Information Retrieval. I should also like to thank Gerard van Wilgen for making available the Ergane dictionaries.

References

1. A.P. de Vries. Mirror: Multimedia query processing in extensible databases. In *Proceedings of the fourteenth Twente workshop on language technology (TWLT14): Language Technology in Multimedia Information Retrieval*, pages 37–48, Enschede, The Netherlands, December 1998.
2. A.P. de Vries. *Content and multimedia database management systems.* PhD thesis, University of Twente, Enschede, The Netherlands, December 1999.

3. A.P. de Vries and D. Hiemstra. The Mirror DBMS at TREC. In *Proceedings of the Seventh Text Retrieval Conference TREC-8*, Gaithersburg, Maryland, November 1999.
4. A.P. de Vries and A.N. Wilschut. On the integration of IR and databases. In *Database issues in multimedia; short paper proceedings, international conference on database semantics (DS-8)*, pages 16–31, Rotorua, New Zealand, January 1999.
5. 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.
6. D. Hiemstra and W. Kraaij. Twenty-One at TREC-7: Ad-hoc and cross-language track. In E.M. Voorhees and D.K. Harman, editors, *Proceedings of the Seventh Text Retrieval Conference TREC-7*, number 500-242 in NIST Special publications, 1999.