Cross-language Retrieval at Twente and TNO.

Dennis Reidsma[1], Djoerd Hiemstra[1], Franciska de Jong[1,2], Wessel Kraaij[2]
[1]University of Twente, Dept. of Computer Science,
P.O. Box 217, 7500 AE Enschede, The Netherlands
[2]TNO TPD, P.O. Box 155, 2600 AD Delft, The Netherlands
{*reidsma,hiemstra,fdejong*}*@cs.utwente.nl, kraaij@tpd.tno.nl*

Abstract

This paper describes the official runs of the Twenty-One group for CLEF-2002. The Twenty-One group participated in the Dutch and Finnish monolingual and the Dutch bilingual tasks. This paper also reports on an experiment that was carried out during the assessment work. The experiment was designed to examine possible influences on the assessments caused by the use of highlighting in the assessment program.

1 Introduction

This paper describes the CLEF participation of the Twenty-One group. [1]

Section 2 provides the context in which research on multilingual information retrieval is carried out at TNO TPD and the University of Twente. Section 3 discusses the Dutch and Finnish runs that the Twenty-One group submitted to CLEF 2002. First the retrieval model is described (section 3.1), after which our submissions to CLEF 2002 are presented. Section 4 describes and analyses the results of an experiment that has been carried out on some aspects of the assessment protocol and discusses its results.

2 CLIR as an aspect of multimedia retrieval

The work on cross-language information (CLIR) that has been carried out by a joint research group from TNO and the University of Twente since 1997 (TREC-6), has been part of a larger research area that can be described as content-based multimedia retieval. CLIR is just one of the themes in a series of collaborative projects on multimedia retrieval, of which Twenty-One provided the name of the search engine that has been developed and used for the participation in TREC and, later on, CLEF. Though the focus on CLIR-aspects is not in all projects as strong as it used to be in Twenty-One, the possibility to search in digital multimedia archives with different query languages and to identify relevant material in other languages than the query language has always been part of the envisaged functionality. Where the early projects exploited mainly the textual material avaliable in multimedia archives (production scripts, cut lists, etc.), the use of timecoded textual information (subtitles, transcripts generated by automatic speech recognition tools, etc.) has become more dominant in the current running projects, for which video and audio retrieval are the major goals, e.g. DRUID and the IST-projects ECHO and MUMIS[2]. In some projects the CLIR functionality is made available by allowing the users of the demonstator systems to select query terms from a closed list which is tuned to the domain of the media archive to be searched. Translation to other languages is then simply a matter of mapping these query terms

[1]Twenty-One was an information retrieval project funded by the with the TAP programme of the EU. Though the Twenty-One project was completed in June 1999, TNO TPD and the University of Twente still participate in the CLEF events under that name.

[2]For details, cf. http://parlevink.cs.utwente.nl/projects, http://www.tpd.tno.nl/, and [5]

to their translation equivalents. Ambiguity resolution and other problems inherent to CLIR-tasks are circumvented in this concept search like approach. However, there is always the additional user requirement to be able to search for terms that are not in the controlled list. Therefore, even in ontology driven projects such as MUMIS, the type of CLIR functionality that is central to the current CLEF-campaign remains relevant in the multimedia domain.

3 Retrieval experiments on the Dutch and Finnish document set

The Twenty-One group participated in the Dutch and Finnish monolingual task and the Duch bilingual task. In this section we present the retrieval model (section 3.1) and discuss the scores for the different tasks.

3.1 The retrieval model

Runs were carried out with an information retrieval system based on a simple unigram language model. The basic idea is that documents can be represented by simple statistical language models. Now, if a query is more probable given a language model based on document d_1, than given e.g. a language model based on document d_2, then we hypothesise that the document d_1 is more likely to be relevant to the query than document d_2. Thus the probability of generating a certain query given a document-based language model can serve as a score to rank the documents.

$$P(T_1, T_2, \cdots, T_n|D)P(D) \;=\; P(D)\prod_{i=1}^{n}(1-\lambda)P(T_i) + \lambda P(T_i|D) \tag{1}$$

Formula 1 shows the basic idea of this approach to information retrieval, where the document-based language model $P(T_i|D)$ is interpolated with a background language model $P(T_i)$ to compensate for sparseness. In the formula, T_i is a random variable for the query term on position i in the query ($1 \leq i \leq n$, where n is the query length), which sample space is the set of all terms in the collection. The probability measure $P(T_i)$ defines the probability of drawing a term at random from the collection, $P(T_i|D_k)$ defines the probability of drawing a term at random from the document; and λ is the smoothing parameter, which is set to $\lambda = 0.15$. The marginal probability of relevance $P(D)$ is assumed to be uniformly distributed over the documents in which case it may be ignored in the above formula. For a description of the embedding of statistical word-by-word translation into our retrieval model, we refer to [1].

3.2 The Dutch runs

For Dutch three separate runs were submitted. First there was the manual run, in which we had a special interest because of our role in the assesment of all the runs submitted for Dutch (cf. section 4). The expected effect of submitting a run for which the queries were manually created from the topics, was to increase the size and quality of the pool of documents to be assessed. The engine applied was a slightly modified version of the NIST Z/Prise 2.0 system.

The Dutch bilingual run is an automatic run done with the TNO retrieval system (also referred to as the Twenty-One engine) as developed and used for previous CLEF participations [1, 2]. Furthermore we used the VLIS lexical database developed by Van Dale Lexicography and the morphological analyzers developed by Xerox Research Centre Grenoble.

For completeness we did a post-evaluation automatic monolingual Dutch run. Mean average precision figures for the three runs are given in Table 1.

3.3 The Finnish run

Since we did not have a Finnish morphological analizer or stemmer, we decided to apply an N-gram approach, which has been advocated as a language independent, knowledge-poor approach

run label	m.a.p.	description
tnoutn1	0.4471	manual monolingual
tnoen1	0.3369	EN-NL dictionary based
tnofifi1	0.4056	automatic monolingual (Finnish)
tnonn1	0.4247	automatic monolingual

Table 1: mean average precision of the runs on the Dutch and Finnish dataset

by McNamee and Mayfield [3]. After applying a stoplist and lowercasing, documents and queries were indexed by character 5-grams. Unlike the JHU approach, the 5-grams did not span word boundaries. This extremely simple approach turned out to be very effective: for almost all topics the score of this run was at least as high as the median score.

4 Assessment of the Dutch results

The University of Twente was responsible for assessing the results for the Dutch newspaper collections (articles from the newspapers 'NRC Handelsblad' and 'Algemeen Dagblad'). Besides assessing all topics in the standard way for the official ranking of the submitted runs, we also repeated some assessments without allowing highlighting of search terms. This section discusses the motivation for this additional experiment and reports on the findings.

4.1 Introduction

The program used to do the assessments is developed at NIST and offers the possibility to highlight terms in the documents. Highlighting words and phrases for which a search engine has detected a relation to the query terms might make it easier for the assessor to decide on the relevance of a document. Usually the assessor will be told explicitly that the presence or absence of highlighted terms in a document is not decisive in marking a document relevant. The assumption is that using or not using highlighting will not influence the assessment results, or more specifically the ranking of the search engines that follows from those results.

We think however that this assumption can be questioned. The following subsection explains that highlighting can affect the assessments and that therefore the use of highlighting may influence the ranking of search engines. A simple experiment will be described that we applied to detect such differences.

If the assessment process would indeed be seriously influenced by the use of highlighting, the implications would be large. Not only the assessment protocol would have to change, but the validity of the assessments of previous years should also have to be reconsidered.

4.2 Possible influences of highlighting on assessment results

We wanted to investigate two different aspects of the assessment results which might be affected by the use of highlighting. The first is the amount of documents that are marked as relevant, the second is the score of the participating search engines. We did not expect to find hard statistical evidence for presence or absence of either one of the influences, given the size of test data, but rather expected some trend to show up, which would warrant further investigation.

The amount of relevant documents Using highlighting might result in more (or less) documents being marked as relevant. Although the assessors are explicitly told not to let the highlighting affect their judgement it is still possible that that happens unintentionally. For example, assessors might read the documents where terms are highlighted less thouroughly, missing in those documents the relevant parts which do *not* contain highlighted terms. Or the assessors might just be biased in favor of documents containing highlighted terms.

The scores of search engines If the assessors are indeed biased towards documents containing highlighted terms this might influence the scores of the search engines. After all, many search engines rely on detecting the presence of query words for marking them as relevant. So in that case, those engines would perform better with the biased assessment than with assessments produced without using highlighting.

4.3 The experiments

The experiment was simple: 18 topics were each assessed at least twice, once with and once without highlighting. These assessments were assigned randomly over 10 people, in such a way that every assessor did some assessments with and without highlighting and no-one assessed one topic twice. The assessors were absolutely not allowed to talk to each other about these assessments until all assessments were finished.

4.4 The results

The results of this experiment were not conclusive. For half of the topics, the assessments with highlighting resulted in more relevant documents than the assessments without highlighting. For the rest of the topics it was the other way around. Viewed from the perspective of the assessors, using highlighting did also not result in significantly more or less relevant documents relative to the other assessors working on that topic.

4.5 Conclusion

There was no trend discernible that confirmed our expectations. However, we could only test the first aspect described above; we did not have the necessary data to test the effect of the highlighting on the scores of the search engines. This second aspect however is where we expected the most interesting results. We recommend therefore testing that as well. If the amount of data is too small to get reliable results, more data should be collected. If the results show a significant change in the scores of the search engines when highlighting is turned off, the assessment protocol should be reconsidered. It is possible then that the benefits of highlighting do not outweigh the adverse effects on the quality of the assessments, in which case highlighting should not be used anymore.

References

[1] D. Hiemstra, W. Kraaij, R. Pohlmann and T. Westerveld. Translation resources, merging strategies and relevance feedback for cross-language information retrieval. In *Cross-language Information Retrieval and Evaluation, Lecture Notes in Computer Science (LNCS-2069)*, Springer-Verlag, pages 102–115, 2000.

[2] W. Kraaij, TNO at CLEF-2001: Comparing Translation Resources. In *Working Notes of CLEF 2001 Workshop, 2001.*

[3] P. McNamee, J. Mayfield. A Language-Independent Approach to European Text Retrieval In *Cross-language Information Retrieval and Evaluation, Lecture Notes in Computer Science (LNCS-2069)*, Springer-Verlag, pages 102–115, 2000.

[4] Djoerd Hiemstra. Using Language Models for Information Retrieval Ph.D. Thesis, Centre for Telematics and Information Technology, University of Twente, January 2001

[5] F. de Jong, J.-L. Gauvain, Dj. Hiemstra, K. Netter. Language-Based Multimedia Information Retrieval. In *Content-Based Multimedia Information Access, RIAO 2000 Conference Proceedings*, 2000, ISBN 2-905450-07-X, C.I.D.-C.A.S.I.S., Paris, 713-722.