Cross-language Retrieval at the University of Twente and TNO.

Dennis Reidsma[1], Djoerd Hiemstra[1], Franciska de Jong[1,2], and Wessel Kraaij[2]

[1] University of Twente, Dept. of Computer Science, Parlevink Group, P.O. Box 217,
7500 AE Enschede, the Netherlands,
{reidsma,hiemstra,fdejong}@cs.utwente.nl
[2] TNO TPD, P.O. Box 155, 2600 AD Delft, the Netherlands,
kraaij@tpd.tno.nl

Abstract. This paper describes the official runs of the Twente/TNO group for CLEF 2002. We participated in the Dutch and Finnish monolingual and the Dutch bilingual tasks. In addition this paper reports on an experiment that was carried out during the assessment of the Dutch results for the CLEF 2002. The goal of the experiment was to examine possible influences on the assessments caused by the use of highlighting in the assessment program.

1 Introduction

This paper has a double focus. The first section describes the CLEF participation of the Twente/TNO group. [3] Section 2 provides the context for the research on multilingual information retrieval carried out at TNO TPD and the University of Twente. Section 3 discusses the applied retrieval model and the results for the Dutch and Finnish runs submitted to CLEF 2002.

The second and main part of the paper starts in section 4. This describes an experiment that has been carried to examine some aspects of the assessment protocol and discusses its results. First the context for the experiment is given and the reasons that led to performing the experiment. After that the experiment itself is described, followed by a summary of results and conclusions.

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), is part of a larger research area that can be described as content-based multimedia retrieval. CLIR is just one of the themes in a series of collaborative projects on multimedia retrieval. The project Twenty-One provided the name

[3] The Twente/TNO group used to participate in earlier CLEF-events under the name Twenty-One. Twenty-One was an information retrieval project funded by the TAP programme of the EU. The project was completed in June 1999.

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 as strong in all projects 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 available 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 projects running currently, for which video and audio retrieval are the major goals, e.g. DRUID and the IST-projects ECHO and MUMIS[4]. 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 to their translation equivalents. Ambiguity resolution and other problems inherent to CLIR-tasks are circumvented in this 'concept search' approach, but there is always the additional user requirement of being 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.

3 Retrieval Experiments for the Dutch and Finnish tasks

The Twenty-One group participated in the Dutch and Finnish monolingual task and the Dutch bilingual task. In this section we present the retrieval model (section 3.1) and discuss the scores for the different tasks (sections 3.2 and 3.3).

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. 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

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

random variable for the query term on position i in the query ($1 \leq i \leq n$, where n is the query length), with the set of all terms in the collection as sample space. 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 $\lambda = 0.15$. Since there is empirical evidence that there is a linear relationship between the probability of relevance of a document and its length, we define the prior $P(D)$ as a constant times the document length [7]. For a description of the embedding of statistical word-by-word translation into our retrieval model, see [4].

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 assessment of the Dutch submissions (cf. section 4). The expected effect of submitting a run for which the queries were manually created from the topics was an increase in 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 [4, 6]. Furthermore we used the VLIS lexical database developed by Van Dale Lexicography and the morphological analyzers developed by Xerox Reserch Centre Europe.

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

Table 1. Mean average precision of the runs on the Dutch and Finnish dataset

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

3.3 The Finnish Run

Since no Finnish morphological analyzer or stemmer was available, it was decided to apply an N-gram approach, which has been advocated as a language independent, knowledge-poor approach by McNamee and Mayfield [9]. 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.

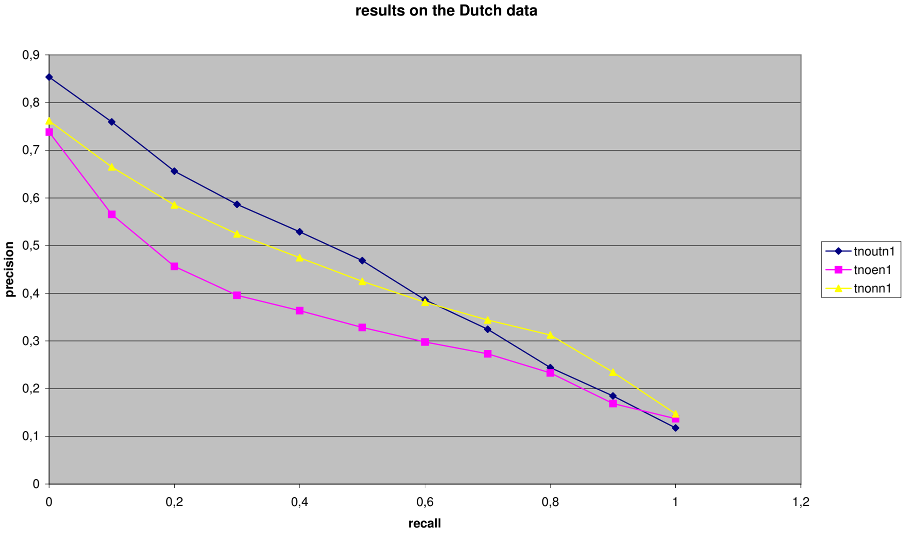

Fig. 1. Precision-recall graphs on the Dutch experiments

4 The Role of Highlighting during Assessment

In a retrieval evaluation campaign such as CLEF, the assessment of the retrieval results plays an important role. For more detail on the methods used to create the test collections on which the evaluations are based and the role of the assessments in this we refer to [1].

The University of Twente was responsible for the assessments of the results for the Dutch CLEF collection, consisting of the 1994-1995 newspaper articles from 'NRC Handelsblad' and 'Algemeen Dagblad'. This meant that for a large number of these articles a judgement had to be made as to whether each article was relevant to the topic for which it was returned as a result by one of the participating retrieval systems.

For the official ranking of the submitted runs the standard assessment protocol was followed. In addition some assessments were repeated under different conditions. The purpose of these additional assessments was to investigate the effect of the use of the highlighting option in the presentation of the retrieved documents on the assessment results. These experiments were kept strictly separate from the official assessments and did not affect the offical rankings presented by the CLEF committee.

This section discusses the motivation for this additional experiment and reports on the findings. First the protocol used to structure the assessment task is discussed. After that some parameters in the protocol are discussed. The last subsections are dedicated to the experiments we carried out on the effect of the highlighting option in the presentation of the documents on the resulting relevance judgments.

4.1 The Assessment Protocol

Within the CLEF campaign the assessment of retrieval results is carried out to produce a high quality, dependable collection of judgments that tell us which documents in the collection are relevant for the given queries. These judgments are used to rank the individual submissions of the participating search engines.

The assessment protocol ensures that the outcomes of the various assessment teams are a solid and reusable basis for the construction of testbeds for the various language specific retrieval tasks.

Variations in several parameters of this protocol can influence the results, such as the number of assessors, the selection of documents that will be manually assessed and the criteria on which a document is judged relevant or non-relevant. The assessment protocol has been standardized based on past research of such parameters, but the fact that the newspaper collections differ for each language and the fact that for each language there is another team doing the assessment task already introduces new variations. On top of that there are several other parameters that are not controlled by the protocol. If experimental results were to indicate that one of these parameters had effects that were previously unknown, the assessment protocol would have to be adapted and the validity of the assessments of previous years would also have to be reconsidered.

The next section discusses some parameters in the assessment protocol and previous research on the influence of variations in these parameters on the resulting system ranking. The rest of the paper is dedicated to our experiments with one of the remaining parameters.

4.2 Protocol Parameters

One of the parameters of the assessments concerns variation in judgement between persons or even for one person on different moments. The question whether this has an adverse effect on the stability of the ranking has been researched for example by Lesk and Salton [8] and Voorhees [13]. For the Dutch testbed similar experiments have been carried out by Hiemstra and Van Leeuwen [5]. Other parameters include pool depth, assessor characteristics, the kind of instruction that the assessors get and the document presentation style. Table 2, taken from [3], gives an overview of variations in the assessment procedures among the various assessment teams for CLEF 2001.

The rest of this paper addresses a parameter that has not yet been researched very extensively: the style of document presentation, more specifically the effect of the use of the highlighting option that the assessment software makes available to the assessors.

4.3 The Effect of Highlighting

The program used to support the assessment task, developed at NIST, offers the option to highlight terms in the retrieved documents. This functionality serves a clear purpose. For example, the assessor can decide on the relevance

Table 2. Aspects of the 2001 judgments [3]

parameter	Dutch	English	French	German	Italian	Spanish
no. of assessors	10	6	5	2	3	4
experienced as user?	yes	some	yes	one	some	no
experienced as assesors?	no	yes	no	one	two	no
written/oral instruction	oral	both	oral	both	both	oral
native topics by assesors	no	yes	yes	one	one	no
translated by assessors	no	no	no	one	one	no
transl. from source lang.?	yes	yes	yes	no	mostly	no
discussion possible?	some	some	yes	yes	yes	yes
single/group opinion?	single	single	?	group	group	group
supervisors involved?	no	no	yes	yes	yes	no
post-assessm. narrative?	no	yes	no	sketchy	sketchy	no

of a document more efficiently if words and phrases related to the topic which occur in the document are highlighted.

Usually the assessor will be told explicitly that the presence or absence of highlighted terms in a document should not be taken as a an *a priori* indication that a document relevant. It is therefore assumed that using highlighting will not influence the assessment results, or more specifically the ranking of the search engines that follows from those results. This assumption however can be questioned. The following subsection explains how highlighting can affect the assessments and why the use of highlighting may influence the ranking of search engines. A simple experiment will be described that we applied in an attempt to detect such effects.

4.4 Possible influences of highlighting on assessment results

We wanted to investigate two different aspects of the assessment results that might be affected by the use of highlighting. The first pertains to the number of documents that are marked as relevant, the second to the ranking of the participating search engines. An effect on the second aspect would clearly have the most impact. We did not expect to find hard statistical evidence for presence or absence of either one of the influences, given the size of the test data. The experiment merely aimed at uncovering possible trends that would warrant further investigation.

The number 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 effect their judgement, it is still possible that this happens unintentionally:

– Assessors might read the documents where terms are highlighted less thoroughly, missing in those documents the relevant parts which do *not* contain highlighted terms.

- Assessors might be biased in favour of documents containing highlighted terms.
- Assessors may uncover documents which they missed because they overlooked the relevant passage without highlighting.

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

4.5 The experiments

The setup of the experiment was straightforward: 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 instructed not to talk to each other about these assessments until all assessments were finished. The extra assessments were used to produce alternative rankings for 29 submissions.

4.6 Results and Conclusions

The number of relevant documents: Table 3 contains some figures on the extra assessments. 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. Given the amount of variation between assessors as found by Voorhees [14], these results suggest that variation in the use of highlighting has no significant influence on the number of documents judged to be relevant.

The rankings: The most important result of the experiments is of course the alternative ranking they produce. As long as the rankings stay more or less the same throughout the variations in the protocol with respect to the highlighting the resulting testbed is still stable and reusable. The official rankings in CLEF are based on the average precision measure shown in figure 2. This figure shows the the influence of *not* using highlighting on the average precision of the submissions as a percentage relative to the average precisions produced using highlighting in the assessment task. The average of the absolute values of these differences is 3.35%.

Table 4 shows two different rankings of some submissions to the Dutch track. The first row shows the official ranking of the 29 submissions that were considered in this experiment. The second row shows the ranking as it would be based on the relevance judgments that were produced *without* using highlighting. Most of the submissions have the same ranking in both cases (though submission 7 and

22 trade places, a minor change). Submission 12 goes 2 places down, but since the initial difference between 11, 12 and 19 was only 0.0144 this is still harmless. Submissions 4 and 16 go 2 places up, but the initial differences in scores between the assessments that are involved are only 0.0127 and 0.0031 respectively. One change is not really trivial anymore: assessment 17 goes 2 places up, 18 goes 3 places up and 21 goes 5 places down. Before the change, the scores of the 6 assessments involved in this change have a difference of 0.0349 between the lowest and the highest (the score of 21 being the highest). After the change this difference is 0.0209 between the score of 21 (the lowest) and the highest. However, the information was not complete enough and the experiments were not extensive enough to permit any fruitful speculation as to the exact causes of this change.

Table 3. Results: no. of relevant found.

Topic	no. assessed	rel. w highlighting	rel. w/o highlighting
92	294	38	38
94	412	18	21
100	433	12	14
107	336	5	40
108	217	34,3	48,18
110	296	4	4
115	575	26	25
119	372	31	20
121	274	36	32
125	305	76	124
127	272	20	23
129	390	48	15
130	256	9	20
131	587	40	26
132	481	13	5
135	453	124	128
137	743	5	5
138	564	7,9	7,14
139	416	46	23
140	356	144	109
Total	8032	748	759

Conclusions: Though these variations in the rankings are not large enough to justify the conclusion that variations in the use of highlighting have an adverse effect on the quality of the relevance judgments, they are certainly large enough to warrant further investigation. It would be advisable to do the experiments described in this paper again, involving more topics and doing the assessments for every topic more than once both with and without highlighting. Another aspect that might be included in these new experiments would be the *speed*

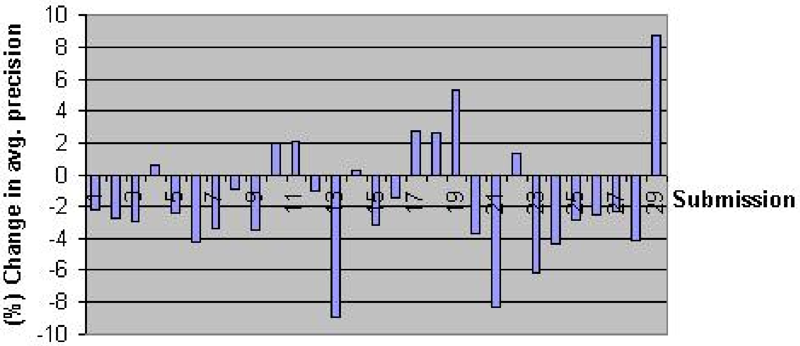

Fig. 2. This figure shows the influence of *not* using highlighting on the average precision of the submissions (percentage relative to the official average precisions).

Table 4. Rankings calculated from assessments using the highlighting functionality vs. rankings calculated from assessments *not* using the highlighting functionality.

with highlighting (1-15)	without highlighting (1-15)	with (16-29)	without (16-29)
29	29	23	18
9	9	21	17
14	14	7	7
13	13	15	15
24	24	25	25
3	3	20	20
26	26	1	1
22	22	4	5
10	10	5	27
19	12	27	4
11	19	16	6
12	21	6	28
18	11	28	16
17	23	2	2
8	8		

with which the assessors work. If the amount of data assessed in the same time using highlighting is for example twice as large as without using highlighting it might be possible that the larger set of relevance judgments compensates for the sligthly lower quality of the individual judgments. We suggest that these experiments be performed to find a definitive answer to the question that has led to this paper: "Does the use of the highlighting option in the NIST system have an unintended effect on the overall quality of the relevance judgments?"

References

[1] M. Braschler and C. Peters. CLEF 2002: Methodology and metrics. *In this volume*, 2002.

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

[3] D. Hiemstra. The CLEF relevance assessments in practice, August 2001. Presentation at the CLEF Workshop 2001, Darmstadt.

[4] D. Hiemstra, W. Kraaij, R. Pohlmann, and T. Westerveld. Translation resources, merging strategies and relevance feedback for cross-language information retrieval. In Peters [10]. LNCS 2069.

[5] D. Hiemstra and D. van Leeuwen. Creating a dutch testbed to evaluate the retrieval from textual databases. Technical report TR-CTIT-02-04, Centre for Telematics and Information Technology, University of Twente, 2002.

[6] W. Kraaij. TNO at CLEF 2001: Comparing translation resources. In Peters [11].

[7] W. Kraaij, T. Westerveld, and D. Hiemstra. The importance of prior probabilities for entry page search. *Proceedings of the 25th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, 2002.

[8] M. Lesk and G. Salton. Relevance assessments and retrieval system evaluation. *Information Storage and Retrieval, 4:242359*, 1969.

[9] P. McNamee and J. Mayfield. A language-independent approach to european text retrieval. In Peters [10]. LNCS 2069.

[10] C. Peters, editor. *Cross-language Information Retrieval and Evaluation.* Springer Verlag, 2000. LNCS 2069.

[11] C. Peters, editor. *Working Notes for the CLEF 2001 Workshop.* 2001.

[12] C. Peters, M. Braschler, J. Gonzalo, and M. Kluck, editors. *Evaluation of Cross-Language Information Retrieval Systems.* Springer Verlag, 2001. LNCS 2406.

[13] E.M. Voorhees. Variations in relevance judgements and the measurement of retrieval effectiveness. *Information Processing & Management, 36:697716*, 2000.

[14] E.M. Voorhees. The philosophy of information retrieval evaluation. In Peters et al. [12]. LNCS 2406.