Creating a Dutch testbed to evaluate the retrieval from textual databases *

Djoerd Hiemstra
University of Twente, CTIT
P.O. Box 217, 7500 AE Enschede
The Netherlands
hiemstra@ctit.utwente.nl

David A. van Leeuwen
TNO-TM
P.O. Box 23, 3769 ZG Soesterberg
The Netherlands
vanleeuwen@tm.tno.nl

Abstract

This paper describes the first large-scale evaluation of information retrieval systems using Dutch documents and queries. We describe in detail the characteristics of the Dutch test data, which is part of the official CLEF multilingual texttual database, and give an overview of the experimental results of companies and research institutions that participated in the first official Dutch CLEF experiments. Judging from these experiments, the handling of language-specific issues of Dutch, like for instance simple morphology and compound nouns, significantly improves the performance of information retrieval systems in many cases. Careful examination of the test collection shows that it serves as a reliable tool for the evaluation of information retrieval systems in the future.

Keywords: Information Retrieval, XML, Multilinguality

1 Introduction

Information retrieval research has a long-standing tradition of experimental work, starting with the Cranfield experiments [21] and continuing nowadays in the successful Text Retrieval Conference (TREC) series [23]. The main objective of information retrieval is satisfying the user's need for information. Although some aspects of retrieval systems can be evaluated without consulting the user, ultimately some actual or potential users have to be subjects in a controlled information retrieval experiment. Doing an evaluation involving real people is not only a costly job, it is also difficult to control and therefore hard to replicate. For this reason, methods have been developed to design unbiased test collections. These test collections are created by consulting potential users, but once they are created they can be used to evaluate information retrieval systems without the need to consult the users during further evaluations. If a test collection is available, a new retrieval method can be evaluated by comparing it to some well-established methods in a controlled experiment.

For many years, test collections were almost exclusively developed for English, and, since English is a language with a rather simple morphology, many interesting research questions were not addressed by these collections. For instance in languages like Dutch and German, nouns can be glued together almost unrestrictedly to form new words. The English *potable water supply* would be one

*This paper is published as CTIT technical report TR-CTIT-02-04, Feb. 2002, http://www.ctit.utwente.nl

compounded word in Dutch: *drinkwatervoorziening*. Furthermore, language-dependent search techniques which were pioneered for English, like for instance stemming algorithms, have been developed today for many of the world's major languages. To support the evaluation of language specific issues of retrieval, test collections are recently being developed for Asian languages within the NTCIR workshops [7] and for European languages within the CLEF workshops [16]. This paper describes the development of a Dutch test collection within the CLEF workshops.

CLEF (Cross-language Evaluation Forum) is a continuation and expansion of the cross-language system evaluation activity that begun in 1997 at the TREC series. Partners in CLEF are Eurospider (Switzerland), National Institute for Informatics (Japan), National Institute for Standards and Technology (USA), IZ Socialwissenschaften Bonn (Germany), IEI-CNR Pisa (Italy), IEEC-UNED Madrid (Spain), University of Hildesheim (Germany), University of Twente (Netherlands). The objectives of CLEF, which are derived from the TREC objectives, are the following.

- To encourage research in multilingual and cross-language information retrieval;
- to increase communication among industry and academia by creating an open forum for the exchange of research ideas;
- to speed the transfer of technology from research laboratories into commercial products by demonstrating substantial improvements in retrieval methodologies on real-world problems;
- to increase the availability of appropriate evaluation techniques for use by industry and academia, including development of new evaluation techniques more applicable to current systems.

The evaluation investigates the performance of systems that search a static set of documents using textual queries (called *topics*, see the next section) that are previously unseen by the system. Participants return the top 1000 documents retrieved for each topic to CLEF for evaluation. The results of one such an experiment for 50 topics is called a *run*. The right answers (called *relevance judgments*, see the next section) are not known to the participants when they produce their the runs. Participants can choose to participate in two separate tasks. The first task is the monolingual task which concentrates on the use of the Dutch language only: Participants use the Dutch topics to search the document collection. In the second task, the cross-language task, the Dutch document collection should be searched using any topic language except Dutch. Cross-language retrieval constitute a first step towards searching multilingual collections, e.g. searching a mixed collection of Dutch and French documents by either a Dutch or a French query. The system should perform some kind of automatic translation to support this kind of functionality.

The participating organisations for this first Dutch evaluation include many of the major Dutch national text retrieval software companies and information retrieval research groups. Participating organisations were Eidetica (Netherlands), Hummingbird (Canada), Johns Hopkins University (USA), Medialab (Netherlands), Océ Technologies (Netherlands), TNO-TPD (Netherlands), Thomson Legal and Regulatory (USA), University of Amsterdam (Netherlands), University of Glasgow (UK), and University of Twente (Netherlands).

This paper serves as an overview of the Dutch retrieval experiments described in detail in the CLEF workshop proceedings [16]. The paper is organised as follows. Section 2 describes the development of the retrieval test data in detail. Section 3 describes the measures that were used for system evaluation. An overview of the results is given in Section 4. The reliability of the results and the quality of the test data for future use is evaluated in Section 5. Finally, Section 6 presents some general lessons learned from this first Dutch evaluation.

2 The test data

An information retrieval test collection has three main ingredients: the *documents*, the textual descriptions of the user's requests for information called *topics*, and the right answers called *relevance judgements*.

2.1 The documents

The documents were kindly provided by PCM Landelijke Dagbladen / Het Parool. They represent a sampling of articles published by the NRC Handelsblad en the Algemeen Dagblad in 1994 and 1995. An example document from the data is shown in Figure 1.

```
<DOC>
<DOCNO>NH19940103-0019</DOCNO>
<HEAD>
<SEC>Binnenland</SEC>
<DAT>19940103</DAT>
<COP>NRC Handelsblad</COP>
</HEAD>
<BODY>
<TI><P>Amsterdam eert Pa Sem met medaille</P></TI>
<LE><P>AMSTERDAM, 3 JAN. J.W. Sijmor, in de Bijlmermeer beter bekend als Pa Sem, heeft
op nieuwjaarsdag uit handen van burgemeester Van Thijn de zilveren eremedaille van de
stad Amsterdam ontvangen.</P></LE>
<TE><P>Pa Sem kreeg het ereteken omdat hij na het neerstorten van een El Al Boeing op
4 oktober 1992 met gevaar voor eigen leven een jongetje redde uit een brandende flat.
Voor de ramp dreef hij een café in de getroffen flat Groeneveen en vervulde hij een
belangrijke rol bij het opvangen van kinderen die in de Bijlmer op straat zwerven. De
eremedaille wordt bij Koninklijk Besluit toegekend aan mensen die zich hebben
onderscheiden door moed, beleid en zelfopoffering.</P></TE>
</BODY>
<INFO>
<PER>Sem, alias Pa</PER>
<PER>Sijmor, J.w.</PER>
<HTR>Onderscheidingen; Algemeen</HTR>
</INFO>
</DOC>
```

Figure 1: An example document

The format of the data uses a labeled bracketing, expressed in the style of SGML (Standard Generalised Markup Language)[1]. The SGML mark-up identifies textual information, like title, leader and text, and bibliographical information like date of publication, newspaper section, and manually assigned keywords. Only the textual information should be used for the official tasks, but bibliographic information is available for later experimentation. An SGML DTD (Document Type Definition) is provided for easy processing of the data.

The Dutch data is part of the multilingual CLEF collection which contains national newspaper and newswire data in English, French, German, Italian and Spanish of the same time period as the Dutch data. Table 1 shows some basic document statistics of the CLEF collection.

[1]SGML is a precursor of XML. The markup does not use any SGML features that are no longer supported in XML.

Language	Source	Size (Mb)	Nr. of records
Dutch	Algemeen Dagblad	241	106,483
	NRC Handelsblad	299	84,121
English	Los Angeles Times	425	110,250
French	Le Monde	157	44,013
	Agence Télégraphique Suisse (SDA)	86	43,178
German	Der Spiegel	63	13,853
	Frankfurter Rundschau	320	139,715
	Schweizerische Depeschenagentur	144	71,803
Italian	La Stampa	193	58,051
	Agenzia Telegrafica Svizzera (SDA)	85	50,527
Spanish	La Agencia EFE	509	215,738
	Total:	2,522	937,732

Table 1: The CLEF multilingual textual database

2.2 The topics

The second ingredient of our test collection consists of natural language statements of the user's need for information. In CLEF, these statements are called *topics*. The topics consist of three parts: the description, the title and the narrative. The description is a natural language statement of the information the user is looking for, typically one sentence. The title consist of the two or three most important words of the description, which comes close to queries as they are actually entered in a real system. The narrative gives a precise definition of the information a document should (or should not) contain in order to be judged as relevant. Figure 2 shows an example topic.

```
<top>
<num> C041 </num>
<NL-title> Pesticide in babyvoeding </NL-title>
<NL-desc> Zoek naar documenten over pesticide in babyvoeding.
</NL-desc>
<NL-narr> Deze documenten geven informatie over ontdekkingen van
pesticide in babyvoeding. Het gaat hierbij om producenten, merken
en supermarkten die verontreinigde voeding hebben aangeboden. De
informatie gaat ook over de maatregelen die tegen de
verontreiniging van babyvoeding met pesticide zijn genomen.
</NL-narr>
</top>
```

Figure 2: An example topic

Why is CLEF not simply providing the queries? Information retrieval systems might differ considerably in the type of queries they support. Some systems might support complex query languages including the use of e.g. Boolean operators, the explicit marking of mandatory terms, or the explicit marking of phrases. CLEF tries to abstract away as much as possible from the actual functionality that is supported by the systems. Participants are free to use any method to construct the queries from the topics. Two approaches are distinguished: automatic and manual. Automatic queries simply use the words from the title or from the description (or both), but might also automatically reformulate queries based on the documents retrieved initially to perform a second search. Manual queries are

those that require any form of human effort for construction. Because of the human effort involved in manual experiments, care should be taken when comparing them with automatic experiments.

A total of 50 topics were created for the evaluation. For each of the document languages of the CLEF collection between 5 and 8 topics were created. These topics were then translated from the source language to each of the six document languages.

2.3 The relevance judgements

The relevance judgements are of key importance to the test data. Relevance judgements are done by potential users (called *assessors* or *judges*) in a carefully controlled experiment. "Relevance" is not a particularly well-defined term [12, 18]. There are many aspects to relevance that are problematic for the evaluation of retrieval systems. To name a few, relevance of a document may be:

judged on a scale a document might for instance be not useful, somewhat useful, fairly useful, very useful and totally useful to a user;

dependent on time a document that is useful to the user today, might no longer be useful to the user later on;

dependent on other retrieved documents a user that walks down a ranked list might for instance find a document further down the list not useful, because it covers the exact same information as the top ranked document, whereas it would have been useful if the top ranked document was not retrieved;

multifaceted the usefulness of a document might be determined by topicality, credibility, specificity, exhaustiveness, accuracy, recency, clarity, etc.

For test collections and for the calculation of precision and recall (see Section 3) it is standard practice to assume that relevance is a dichotomous decision, either yes (the document is relevant) or no (the document is not relevant). Furthermore, the relevance of one document should not be influenced by the relevance of another document, that is, a document is still relevant even if it is the thirtieth document with the same information. The judges were instructed to do their judgements based on these assumptions. They followed the working definition that a document is relevant if it would be useful for writing a report on the subject of the topic.

For early test collections, relevance judgements were done for each document in the database. This however is, given the size of the Dutch database, impossible. If we assume that each document could be judged in 30 seconds, it would take over 1500 hours to judge the entire Dutch document collection for only one topic. Relevance judgements were therefore done on a sample of the document collection following the so-called pooling method [19]. A document pool is created by taking the top n documents retrieved of each system, removing the duplicates. Only the documents in the pool are judged. Judges do not know which system retrieved a document, nor do they know the rank the system(s) retrieved a document. The pool depth was set to n = 60, based on experimental work by Zobel [25]. The completeness of the pool and the quality of the judgements are evaluated in Section 5.

3 Evaluation

An important element of CLEF is to provide a common approach to evaluate system performance. In this section, we describe a number of standard evaluation measures which are based on precision and

recall. Precision is defined by the fraction of the retrieved documents that is actually relevant. Recall is defined by the fraction of the relevant documents that is actually retrieved.

$$\begin{aligned} \text{precision} &= \frac{r}{n} \qquad r : \text{number of relevant documents retrieved} \\ & \qquad\qquad\quad n : \text{number of documents retrieved} \\ \text{recall} &= \frac{r}{R} \qquad R : \text{total number of relevant documents} \end{aligned}$$

Precision and recall are only applicable to sets of retrieved documents. For the evaluation of ranked retrieval systems, precision and recall have to be averaged somehow over the ranked lists. The idea here is to calculate a number of evaluation measures for different types of users. At one end of the spectrum is the user that is satisfied with any relevant document, for instance a user that searches a web page on last nights football results. At the other end of the spectrum is the user that is only satisfied with most or all of the relevant documents, for instance a lawyer searching for jurisprudence. In CLEF three different evaluation measures are used: precision at fixed levels of recall, precision at fixed points in the ranked list and the average precision over the ranks of relevant documents. A tool for calculating these measures is publicly available from the Smart system [17].

3.1 Precision at fixed recall levels

For this evaluation a number of fixed recall levels are chosen, e.g. 10 levels: $\{0.1, 0.2, \cdots, 1.0\}$. The levels correspond to users that are satisfied if they find respectively 10 %, 20 %,$\cdots$, 100 % of the relevant documents. For each of these levels the corresponding precision is determined by averaging the precision on that level over the topics.

In practice, the levels of recall do not always correspond with natural recall levels. For instance, if the total number of relevant documents R is 3, then the natural recall levels are 0.33, 0.67 and 1.0. Other recall levels are determined by using interpolation. The interpolation method used determines the precision at recall level l by the maximum precision at all points larger than l. For example, if the three relevant documents were retrieved at rank 4, 9 and 20, then the precision at recall points $0.0, \cdots, 0.3$ is 0.25, at 0.4, 0.5 and 0.6 the precision is 0.22 and at $0.7, \cdots, 1.0$ the precision is 0.15 [4]. Interpolation might also be used to determine the precision at recall 0.0, resulting in a total of 11 recall levels. Average precision at 11 points of recall is visualised in the recall-precision graphs of Figure 3 and Figure 4 in the next section.

3.2 Precision at fixed points in the ranked list

Recall is not necessarily a good measure of user equivalence, for instance if one query has 20 relevant documents while another has 200. A recall of 50 % would be a reasonable goal in the first case, but unmanageable for most users in the second case [6]. A more user oriented method would simply choose a number of fixed points in the ranked list, for instance 9 points at: 5, 10, 15, 20, 30, 100, 200, 500 and 1000 documents retrieved. These points correspond with users that are willing to read 5, 10, 15, etc. documents of a search. For each of these points in the ranked list, the precision is determined by averaging the precision on that level over the topics. Similarly, the average recall might be computed for each the points in the ranked list.

A potential problem with these measures however is that, although precision and recall theoretically range between 0 and 1, they are often restricted to a small fraction of the range for many cut-off points. For instance, if the total number of relevant documents $R = 3$, then the precision at 10 will be 0.3 at maximum. One point of special interest from this perspective is the precision at R documents

retrieved. At this point the average precision and average recall do range between 0 and 1. Furthermore, precision and recall are by definition equal at this point. The R-precision value is the precision at each (different) R averaged over the topics [4].

3.3 Average precision over ranks of relevant documents

The mean average precision measure is a single value that is determined for each topic and then averaged over the topics. The measure corresponds with a user that walks down a ranked list of documents who will only stop after he / she has found a certain number of relevant documents. The measure is the average of the precision calculated at the rank of each relevant document retrieved. Relevant documents that are not retrieved are assigned a precision value of zero. For the example above where the three relevant documents are retrieved at ranks 4, 9 and 20, the average precision would be computed as $(0.25 + 0.22 + 0.15)/3$ = 0.21. This measure has the advantages that it does not need the interpolation method and that it uses the full range between 0 and 1 [4]. The mean average precision measure is used in the analyses of Section 5.

3.4 Other performance issues

Although the CLEF series mainly focus on retrieval effectiveness in terms of precision and recall, participating organisations are encouraged to experiment with other performance issues. Participants of CLEF may try to optimise their systems for instance for memory requirements, query response time, and indexing speed. A special task was designed this year for interactive user experiments [15].

4 Results

This section gives a brief overview of the official evaluation results. Participants could choose to participate in the monolingual task or in the cross-language task.

4.1 Monolingual results

The first task is the monolingual task which concentrates on the use of the Dutch language only: Participants use the Dutch topics to search the document collection. Many of the experiments conducted in this task were designed to test how the system handles Dutch morphological normalisation and compounds. With the exception of Johns Hopkins University, who used an index based on letter 6-grams, all systems used some language-specific software designed to handle Dutch texts. Figure 3 summarises the best recall-precision averages of the participating systems. A brief description of the experiments is given below.

AmsNlM The University of Amsterdam (Monz and De Rijke [14]) used their FlexIR system. The focus of their experiments was on the effects of morphological analyses such as stemming and compound splitting on retrieval effectiveness. They found as much as 55 % improvement in average precision. Similar, but less spectacular, results were found if title only queries were used. Ranking is based on the Smart Lnu.ltc weighting algorithm.

aplmon1a Johns Hopkins University Applied Physics Lab (McNamee and Mayfield [11]) combined a regular word-based index with an n-gram-based index. For the index based on n-grams, they used combinations of $n = 6$ characters, so the phrase "prime minister" contains the 6-grams ␣`prime`, `prime`␣, `rime`␣`m`, ..., `nister`, `ister`␣. The use of overlapping character n-grams

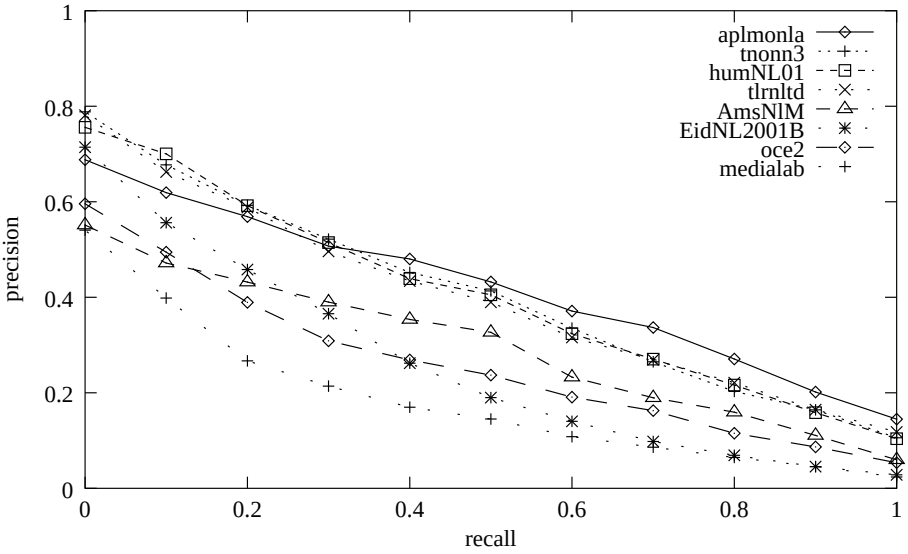

Figure 3: Recall-precision graphs for the best monolingual experiments

provides a surrogate form of morphological normalisation. For instance, the n-gram `minist` could have been generated from "minister", "ministers", "ministerial", "ministry", etc. The ranking function is based on the use of statistical language models. In an interesting new relevance weighting approach, query term-specific relevance weights were computed by a measure called 'residual inverse document frequency'.

EidNL2001B Eidetica (Frizzarin and Groenink [3]) has run six minimalistic tests with its t-repository indexer, doing as little tuning as possible. The indexer performs compound splitting and various types of term extraction based on context-free tagging of words, including proper name recognition. Extracted terms have a length of between 1 and 4 words (or parts of words in the case of compounds).

humNL01 Hummingbird Ltd. (Tomlinson [20]) used their Fulcrum SearchServer. The system uses lexicon-based language processing to stem each distinct word to one or stems. Compound words like e.g. "babyvoeding" produce multiple stems "baby" and "voeding". The ranking function combines term frequency and inverse document frequency in a manner that is similar to the Okapi BM25 approach and the Smart Lnu.ltu approach.

medialab Medialab B.V. (Van der Weerd and Blom [24]) used their standard retrieval engine without any modifications. The system uses a built-in Dutch stemmer, which is based on the Porter stemmer. Compounds are handled at query time by expanding queries with compounds, for instance "roos" is expanded to "klaproos".

oce2 Océ Technologies B.V. (Klok, Driessen, and Brunner [8]) designed a new method for relevance computation based on six heuristics that are quite common in information retrieval, but are seldomly combined into one consistent system. The ranking function uses heuristics based on term frequency, on terms occurring in document fields, on phrases, and on the completeness of the query. The completeness of the query requires that some terms have to present in the document, whereas others are 'interchangeable'. All information is combined in highly structured queries, consisting of several levels of subqueries.

tnonn3 TNO-TPD (Kraaij [9]) compared several approaches to handle lexical variations: fuzzy lookup, Porter stemming and dictionary-based lemmatisation, including some minor experiments with case sensitivity and fuzzy expansion. Dictionary-based lemmatisation, which also splits compounds, performs noticeably better than the stemming approach. The ranking function is based on statistical language models.

tlrnltd Thomson Legal and Regulatory (Molina-Salgado, Moulinier, Knutson, Lund, and Sekhon [13]) used their WIN system which is based on the inference network model. The system uses a morphological analyser and part-of-speech tagger to detect compounds and noun phrases, which are then represented as structured queries at search time. The ranking function, which uses standard *tf.idf* for computing term weights, finds the best dynamic portion in a document and combines the score of the best portion with the score of the whole document.

4.2 Cross-language results

In the second task, the cross-language task, the Dutch document collection is searched using any topic language except Dutch. The participants of this task mainly experimented with different resources for the automatic translation English queries to Dutch. Resources used were bilingual word lists, lexical databases and parallel (bilingual) corpora from which translation are extracted automatically. Figure 4 summarises the best recall-precision averages of the participating systems. A brief description of the experiments is given in the remainder of this section.

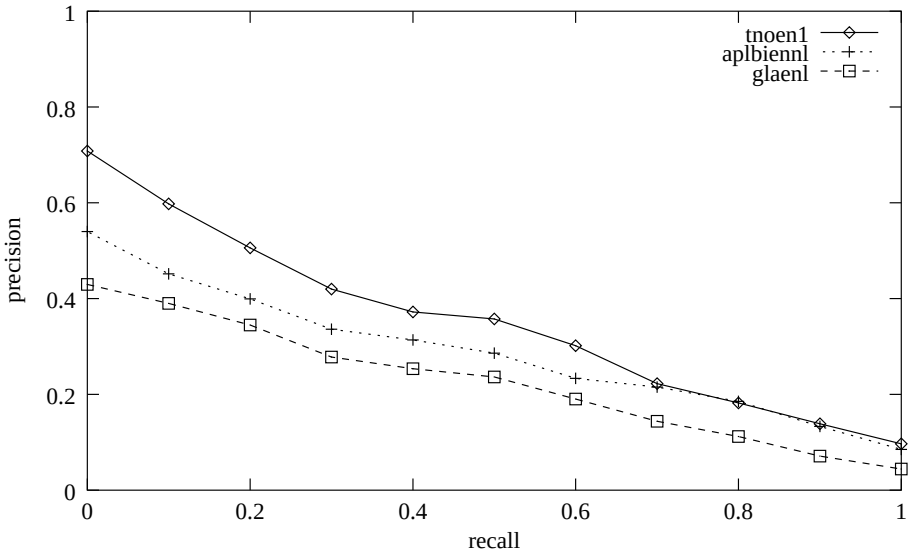

Figure 4: Recall-precision graphs for the best cross-language experiments

aplbiennl Johns Hopkins University Applied Physics Lab (McNamee and Mayfield [11]) compared several approach to query translation: publicly available bilingual wordlists, bilingual wordlists automatically extracted a Dutch-English parallel corpus, and untranslated queries. The last is likely to succeed when languages share root words, because of the 6-gram approach (see previous section). The ranking function is based on the use of statistical language models.

glaen1 University of Glasgow (Adriani [1]) combined a publicly available bilingual word list with translations extracted from a Dutch-English parallel corpus. The corpus translations perform

much worse than the translations from the word list. Translated queries are further expanded by assuming that top few documents retrieved are relevant to the query. Ranking is based on a standard *tf.idf* formula.

tnoen1 TNO-TPD (Kraaij [9]) compared three translation resources for query translation: a commercial machine translation system, a commercial lexical database, and parallel corpora. Surprisingly, the corpus translations perform at the level of the commercial products. Ranking is based on a statistical language model for cross-language retrieval that combines statistical translation with retrieval. The system is able to process structured queries in which possible translations of a source language are grouped and weighted according to their translation probability.

5 Quality of the testbed

The results reported here, as well as the usefulness of the Dutch test collection for future experiments, depend heavily on the quality of the relevance judgements. We investigated two main concerns: the consistency of the judgements and the completeness of the judgements. Consistency addresses the following question: Do the results of the experiments critically depend on the particular choice of human judges, that is, would we have gotten different results if we had a different group of judges? Completeness addresses the question: Do the results of the experiments critically depend on the particular choice of documents that were selected in the pool (see Section 2.3), that is, would we have gotten very different results if we had a different group of participating systems? This section shows that both the consistency and the completeness of the relevance judgements are as good as those of the TREC collections.

5.1 Consistency of the judgements

For the Dutch collection, 10 different judges were used, so the judgements are not particularly dependent on the choices of one particular judge. However, relevance judgements are inherently subjective: They are known to differ across judges, and even to differ for the same judge at different times. Some researchers therefore question if one can draw valid conclusions from experimental work using test collections [5]. A pragmatic reply to this critisism is that system improvements developed on test collections do prove to be beneficial in operational settings. Experimentation using test collections is, and has been for many years, *the* way to conduct information retrieval research.[2] A second reply to critisism on test collections is that a number of studies show that comparative performance of two retrieval strategies is stable when evaluated using two or more independent sets of judgements by different judges. Lesk and Salton [10] and Voorhees [22] used assessments by two or more independent judges. They showed that the relative performance of two systems did not change significantly if a different set of judgements was used.

To have some idea of the consistency of our judges, we created an independent set of judgements for a small sample of 10 topics of the total set of 50 topics. Lesk and Salton [10] introduced the *overlap* of the relevant document sets to quantify the amount of agreement among differents sets of relevance assessments. Overlap is defined as the size of the intersection of the relevant documents sets divided by the size of the union of the relevant documents sets. Recent computations of the overlap of different sets of relevance assessments for one of the TREC collections resulted in overlaps of between

[2] About 80 % of the SIGIR 2001 conference publications [2] use test collections for experimentation.

0.421 and 0.494 [22]. On our small sample, the overlap is 0.465.[3] This indicates that the consistency of our judges is as good as the consistency of the judges of the reported TREC collection.

5.2 Completeness of the judgements

The completeness of the judgements addresses the reliability of the official CLEF experiments, but also the reliability of the test collection for future experiments. As said, the test collection should be a reliable tool for future use as well. It is very likely that future experiments will retrieve some documents in the top for which no relevance judgements are available because none of the official CLEF systems retrieved them. So, what if my run is not judged?

To have some idea of the completeness of the relevance judgements, we conducted the following experiment. Each official run that contributed to the pool was evaluated both with (standard evaluation) and without the relevant documents that the run uniquely contributed to the pool. The difference between the two evaluations will give an idea of how reliable the collections are for future work. The comparison shows that on average, we expect an unjudged run to have only 0.0031 higher average precision after judging.

run name	average precision		difference		unique
	unjudged	judged			rel.
ut1	0.4222	0.4230	0.0008	0.2 %	55
aplmonla	0.3943	0.4002	0.0059	1.5 %	29
tnonn3	0.3914	0.3917	0.0003	0.1 %	2
humNL01x	0.3825	0.3831	0.0006	0.2 %	5
tlrnltd	0.3760	0.3775	0.0015	0.4 %	10
tnoen1	0.3246	0.3336	0.0090	2.8 %	32
AmsNlM	0.2770	0.2833	0.0063	2.3 %	32
aplbiennl	0.2692	0.2707	0.0015	0.6 %	7
oce2	0.2363	0.2405	0.0042	1.8 %	21
glaenl	0.2113	0.2123	0.0010	0.5 %	8
oce1	0.2024	0.2066	0.0042	2.1 %	23
medialab	0.1600	0.1640	0.0040	2.5 %	23
EidNL2001A	0.1339	0.1352	0.0013	1.0 %	8
		max:	0.0090	2.8 %	55
		mean:	0.0031	1.2 %	20
		standard deviation:	0.0027	1.0 %	15

Table 2: Average precision of runs with and without unique judgements

Table 2 gives the complete overview of the differences. The runs labeled as `ut1` is a run using manually formulated queries, done at the University of Twente using the NIST Z/Prise system. It was submitted to have at least one high-quality run that contributes many relevant documents to the pool of documents to be judged. This run retrieved most relevant documents (55) that were not retrieved by any of the other systems, but each system uniquely retrieved some relevant documents. If any of the systems would not have participated, we would have missed its uniquely retrieved relevant documents

[3] We followed Voorhees' [22] procedure by taking a random sample of 200 documents that the primary assessors judged not relevant. The reported overlap is the mean of 10 random samples, for which the overlap ranges between 0.443 and 0.489.

during the judgement process. However, this would not have affected the evaluation results, that is, the relative ordering based on the system's retrieval effectivenesses does not change.

6 Conclusion and the future

This paper reports on the first large-scale evaluation of information retrieval systems on Dutch data. The evaluation showed that language-specific modules like stemmers and compound-splitters result in significant improvements in many cases.

We showed that the Dutch testbed is a valuable tool for future evaluations of search technology. Our judges were as consistent as the judges that developed the TREC collections. The evaluation of the completeness of the judgements show that none of the runs uniquely found a significant number of relevant documents. Removing the uniquely found relevant documents for each run does not change the relative ordering of the systems, showing that the Dutch collection serves as a valuable tool for the development of search and language technology for the Dutch market.

There will be two more successive rounds of retrieval experiments within the CLEF workshop series in 2002 and 2003. The test collection described in this paper is available for official participants of the CLEF workshops.

Acknowledgements

The research presented in this paper was funded in part by the DELOS Network of Excellence on Digital Libraries and by the Dutch Telematics Institute project DRUID. We are most grateful to PCM Landelijke Dagbladen / Het Parool for providing the Dutch data. We like to thank Arjan van Hessen and Roeland Ordelman of the University of Twente, and Wessel Kraaij and Renée Pohlmann of TNO-TPD for their help and advice. Many thanks go to the organisations that participated in the official Dutch CLEF experiments.

References

[1] M. Adriani. English-Dutch CLIR using query translation techniques. In Peters [16], pages 143–146.

[2] W.B. Croft, D.J. Harper, D.H. Kraft, and J. Zobel, editors. *Proceedings of the 24th ACM Conference on Research and Development in Information Retrieval (SIGIR)*. ACM Press, September 2001.

[3] T. Frizzarin and A. Groenink. Minimalistic test runs of the Eidetica indexer. In Peters [16], pages 181–182.

[4] D. K. Harman. Evaluation techniques and measures. In *Proceedings of the third Text Retrieval Conference TREC-3*, pages A5–A13, 1995.

[5] S. P. Harter. Variations in relevance assessments and the measurement of retrieval effectiveness. *Journal of the American Society for Information Science*, 47(1):37–49, 1996.

[6] D. Hull. Using statistical testing in the evaluation of retrieval experiments. In *Proceedings of the 16th ACM Conference on Research and Development in Information Retrieval (SIGIR'93)*, pages 329–338, 1993.

[7] N. Kando, editor. *Proceedings of the 2nd NTCIR Workshop Meeting on Evaluation of Chinese and Japanese Text Retrieval and Text Summarization*. National Institute of Informatics, March 2001. http://research.nii.ac.jp/ntcir/workshop/OnlineProceedings2/

[8] J. Klok, S. Driessen, and M. Brunner. Some terms are more interchangeable than others. In Peters [16], pages 189–194.

[9] W. Kraaij. TNO at CLEF-2001: Comparing translation resources. In Peters [16], pages 29–40.

[10] M. Lesk and G. Salton. Relevance assessments and retrieval system evaluation. *Information Storage and Retrieval*, 4:242–359, 1969.

[11] P. McNamee and J. Mayfield. JHU/APL experiments at CLEF: Translation resources and score normalization. In Peters [16], pages 121–132.

[12] S. Mizzaro. Relevance: The whole story. *Journal of the American Society for Information Science 48*(9), 810–832, 1997.

[13] H. Molina-Salgado, I. Moulinier, M. Knutson, E. Lund, and K. Sekhon. Thomson Legal and Regulatory at CLEF-2001: monolingual and bilingual experiments. In Peters [16], pages 147–152.

[14] C. Monz and M. de Rijke. The University of Amsterdam at CLEF-2001. In Peters [16], pages 165–170.

[15] D.W. Oard and J. Gonzalo. The CLEF-2001 interactive track. In Peters [16], pages 203–214.

[16] C. Peters, editor. *Results of the CLEF 2001 Cross-language System Evaluation Campaign: Working Notes for the CLEF 2001 Workshop*. European Research Consortium for Informatics and Mathematics (ERCIM), September 2001. (to appear as Springer Lecture Notes in Computer Science; draft version available at: http://www.ercim.org/publication/ws-proceedings/CLEF2/)

[17] Smart. *ftp-site*. ftp://ftp.cs.cornell.edu/pub/smart/, 1994.

[18] T. Saracevic. Relevance: A review of and a framework for the thinking on the notion in information science. *Journal of the American Society for Information Science 26*, 321–343, 1975.

[19] K. Sparck-Jones and C.J. van Rijsbergen. Report on the need for and provision of an "ideal" information retrieval test collection. British Library Research and Development Report 5266, Computer Laboratory, University of Cambridge, 1975.

[20] S. Tomlinson. Hummingbird's Fulcrum searchserver at CLEF-2001. In Peters [16], pages 171–180.

[21] B.C. Vickery. *Techniques of Information Retrieval*. Butterworths, 1970.

[22] E.M. Voorhees. Variations in relevance judgements and the measurement of retrieval effectiveness. *Information Processing & Management*, 36:697–716, 2000.

[23] E.M. Voorhees and D. K. Harman. Overview of the 9th text retrieval conference. In *Proceedings of the 9th Text Retrieval Conference TREC-9*, pages 1–14. NIST Special Publication 500-249, 2001.

[24] P. van der Weerd and W. Blom. First experiments with CLEF. In Peters [16].

[25] J. Zobel. How reliable are the results of large-scale information retrieval experiments? In *Proceedings of the 21st ACM Conference on Research and Development in Information Retrieval (SIGIR)*, pages 307–314, ACM Press, 1998.