Océ at CLEF 2003

Roel Brand, Marvin Brünner, Samuel Driessen, Pascha Iljin, Jakob Klok

Océ-Technologies B.V.
P.O. Box 101
5900 MA Venlo
The Netherlands
{rkbr, mbru, sjdr, qpilj, klok}@oce.nl

Abstract
This report describes the work done at Océ Research for the Cross-Language Evaluation Forum (CLEF) 2003. This year we participated in seven mono-lingual tasks (all languages except Russian). We used the BM25 model, a probabilistic and (for Dutch only) a statistical model to rank documents. Knowledge Concepts' Content Enabler semantic network was used (for statistical model only) for stemming, and query expansion. Our main goals were to compare the BM25 model and the probabilistic model, and to evaluate the performance of a statistical model that uses 'knowledge' from queries and relevance assessments from previous years. Furthermore, we give some comments on the standard performance measures used by CLEF.

1 Introduction
This is our third participation in the Cross-Language Evaluation Forum (CLEF). In 2001 we have only participated in the Dutch monolingual task. Last year we participated in all mono-lingual tasks, some of the cross-lingual and in the multi-lingual task. The goal for this year was to concentrate on mono-lingual tasks. We aimed to compare the models we constructed during the last two years and get more insight into the possibilities of using 'knowledge' from queries and relevance assessments from previous years in order to construct information retrieval systems for this year. (Due to problems with indexing and a restriction on a maximal number of mono-lingual runs, the runs for Russian language were not carried out.)

2 Methods

2.1 Ranking systems
Three different approaches were used to rank the documents:
1. The BM25 model (for seven languages).
2. A probabilistic model (for seven languages)
3. A statistical model (for Dutch language only)

The BM25 model was used this year again, just like in CLEF 2002 [1]. The probabilistic and statistical models appeared as a result of internal research done last year [2].

2.2 Query
A query is constructed automatically from the *title* and *description* tags of a topic, and all single characters and stop word terms are removed. Further, all remaining terms were brought to lower case. The expansion of query terms was carried out for the statistical approach only. For that model, the morphological collapse (dictionary based stemming) of Knowledge Concepts' Content Enabler semantic network was used to obtain root forms of (not necessarily all) query terms. The root forms were then expanded with the semantic network. The morphological variants of the root form (such as plural form, etc.) were added to the query.

2.3 Indexing
The indexes were built for each of the languages by splitting the documents on non-alphanumerical characters. Single characters were removed from indexes. We have decided to leave the stop words in the indexes because of the following reasons. It is very difficult to construct a universal set of stop words. If it is based on the frequencies within a document collection, it is highly probable that the set of stop words will not be the same for two different document collections. In case it is based on human decisions, a number of important terms from the document collection and/or query will be removed. For example, consider the terms 'new' and 'year' as stop words (they are used quite often). After removing these terms from the document collection and from the

queries, it becomes difficult to find a set of relevant documents for the query '*A New Year tree*'. In order to show that stop word removal is not always good, consider the query '*Who said "To be or not to be?"*'. In this case *all* terms from the query can be defined as stop words. This year we used of the following stop word lists:

- Dutch
- internally developed stop word list
- http://www.unine.ch/Info/clef/

- English
- http://www.searchengineworld.com/spy/stopwords.htm
- http://www.unine.ch/Info/clef/

- Norwegian
- http://snowball.tartarus.org/norwegian/stop.txt

- Spanish
- http://www.unine.ch/Info/clef/
- ftp://ftp.cs.cornell.edu/pub/smart/spanish.stop

- Finnish
- French
- Swedish
- German
- http://www.unine.ch/Info/clef/

3 Activities

3.1 Ranking models

The BM25 model has been described in [1]. Last year we have shown that the performance of the BM25 ranking algorithm depends greatly on the choice of the values of the parameters *k1* and *b*. However, the estimation of those values for the optimal performance is only possible when the document collection, the set of queries and the set of relevance assessments are all available beforehand. As far as the relevance assessments are not known in advance, a choice should be made for the parameters. For this year we have chosen $k1$=1.2 and b=0.6 for all seven languages. This pair of values was optimal for Dutch document collection, set of queries and relevance assessments of 2001.

The urn model (balls in an urn=terms in a document) has been selected as a basis for the probabilistic model and has been used for all seven languages. The probability model has been selected as the result of theoretical research done in 2002. We aimed to implement a set of clues (we defined) in a '*mathematically correct*' model. That is a model with the assumptions and ranking rules, which are valid from a mathematical point of view. Examples of clues are: presence of terms in the document that are synonyms to the terms from the query, importance of a query field, and length of a document. We found out that a set of clues we defined could not be entirely incorporated in the currently known information retrieval models and be mathematically correct. We found the probabilistic model the most appropriate one to implement several of the defined clues.

Last year we have experimented with the Dutch document collection, the set of queries and the relevance assessments for 2001 and 2002. The statistics for each of these two years were obtained. Two runs have been submitted for the Dutch language using the statistical model, each run used the statistics of one of the two years. Different statistics have been chosen depending on their degree of significance to obtain better performance results. We took into consideration features as the part of speech of the query terms; terms in the document from title or from description; query terms of certain document frequency; terms located in title, lead or description of a document; terms related to the query terms; synonyms to the query terms; proper names. For each feature we calculated a value expressing the expected '*gain*' from its usage. The *gain* value has been defined on the base of known relevance assessments done for (document, topic) pairs. We will not explain the way we calculated the '*gain*' values because it cannot be done in short. The interested reader is referred to [2].

The experiments with the statistical model for CLEF data from 2001 and 2002 resulted in improvements of performance compared to the BM25 and probabilistic models. The proper choice of features to be selected and

their 'gain' values lead to better results. However, this model is strongly dependent on the data collection, queries and relevance assessments. Hence, the results for a set of new documents, new queries and new relevance assessments are unpredictable.

3.2 Different parts of the topic

Last year we experimented with generating queries using different parts of the topics. All possible combinations combined from one to three parts of (title, description, and narrative) have been investigated. Both the runs (for probabilistic and statistical models) and the statistical information indicate that using the narrative makes the performance of the information retrieval engine worse. We suppose that the narrative part contains too many irrelevant terms that add 'noise' to the query. A clever selection of terms from the narrative is needed. However, an automatic selection of 'proper' terms is not an easy task. We did not aim to solve it this year.

3.3 Compound Splitting

For a statistical model, each compound word was split into parts that are independent words themselves. Those words were added to the query as one set. This means only the appearance of all parts of a compound in a document counts. A recursive algorithm takes a word, tries to split a left part that is a meaningful word, and proceeds with the remaining tail of the word.

It has been observed that words are sometimes divided into groups of words that have no logical relation with each other: *Frankrijk (*France*)=Frank+rijk* (Frank+rich) or *neerslag* (precipitation) = *neer+slag* (down+knock). Besides, a number of compounds can be split in several ways. For example, the word 'basketbalkampioenschappen' (basketball championships) has eight (!) ways to be divided into independent nouns. These observations lead us to the development of a statistical rule that reduces senseless splitting. This rule is based on data from the document collection.

3.4 Expansion of the query terms

As was mentioned in the '*Query*' paragraph, we have morphologically expanded query terms for the statistical model only using Knowledge Concepts' Content Enabler semantic network.

3.5 Related terms and synonym expansion

We did the research on using related terms and synonyms. We have found that Knowledge Concepts' Content Enabler is not good enough to create related terms and synonyms for our models, both from the point of view of logic and the performance of the retrieval systems. A measure of 'similarity' between two terms is needed in order to rank the proposed list of related terms and synonyms. The most 'similar' terms only should be used in order to extend the query with related terms and synonyms.

4 Runs for 2003

This year we have submitted 16 mono-lingual runs:

- for each of the following languages two runs (using BM25 and probabilistic ranking principles): Finnish, French, German, Italian, Spanish and Swedish two runs;
- for Dutch BM25 ranking, probabilistic ranking, and two statistical rankings (one is based on statistics for 2001 and another one is based on statistics for 2002).

Numerical results of Océ at CLEF 2003

Name of the run	Number of retrieved relevant documents	Average precision	R-precision
Swedish BM25	729 out of 889	0.3584	0.3585
Swedish probabil.	633 out of 889	0.2716	0.2743
Italian BM25	759 out of 809	0.4361	0.4287
Italian probabil.	731 out of 809	0.3805	0.3865
French BM25	894 out of 946	0.4601	0.4273
French probabil.	865 out of 946	0.4188	0.4044
Finnish BM25	417 out of 483	0.3570	0.3230
Finnish probabil.	407 out of 483	0.3031	0.2624
Spanish BM25	2109 out of 2368	0.4156	0.4094
Spanish probabil.	2025 out of 2368	0.3500	0.3696
German BM25	1482 out of 1825	0.3858	0.3838
German probabil.	1337 out of 1825	0.3017	0.3088

Dutch BM25	1438 out of 1577	0.4561	0.4438
Dutch probabil.	1336 out of 1577	0.4049	0.3652
Dutch statist. 2001	1375 out of 1577	0.4253	0.3940
Dutch statist. 2002	1378 out of 1577	0.4336	0.3983

5 Further remarks on the evaluation measures of the information retrieval systems

In *Appendix A* we give some remarks on the evaluation measures used within CLEF.

6 Conclusions

We have compared the BM25 and the probabilistic models on the base of mono-lingual runs. The BM25 model systematically outperforms the probabilistic one. This indicates that striving for mathematical correctness is not the best guideline to obtain better engine performance. At the same time we have observed that a quite simple implementation of the probabilistic model has a satisfactory performance. Furthermore, we conclude that (the construction of) a better retrieval model needs 'knowledge' about the data collection, and the way the user formulates the topics and assesses the documents. All this information is needed in order to tune the statistical retrieval engine.

7 References

[1] Roel Brand, Marvin Brünner, Océ at CLEF 2002, Lecture Notes on Computer Science, to appear in 2003.
[2] Pascha Iljin, Modeling Document Relevancy Clues in Information Retrieval Systems, SAI, to appear in 2004.

Appendix A

The relevance judgements available from the CLEF processing (taking the Dutch data from 2001as an example) are:

> There are 190,604 documents in the document collection. For every query, every participating information retrieval system has provided a ranking of these documents. The top 1000 documents are delivered to the CLEF as the output of each system. Judges select the top N ($N \leq 1000$) documents from this ranked list.
>
> Suppose that there are M participating information retrieval systems. At most N*M documents are read by judges and receive the relevance judgements. In case the same document was in the top N for at least two participating systems, less than N*M obtain relevance judgement values.
>
> There are 16774 relevance judgement values available for 50 queries. This means on average of 335 relevance judgements per query. There are 1224 documents known to be relevant for 50 queries. This means on average of 25 relevant documents per query.

We think that *it is not entirely correct to calculate the statistics for the top 1000 documents in a commonly accepted way*. The very simple reason is that we know the relevance judgement values for much less than 1000 documents. It appears that about 60-70% documents in the top 1000 have unknown values for relevance! Those unread documents get values '*irrelevant*'.

Suppose that a relevant document was not retrieved within top N documents by any of the participating information retrieval systems. Because no value was assigned to this document, it obtains the value 0 (irrelevant) by definition. The paradox is that a *relevant document* gets the value of an *irrelevant document*.

Conclusion

For a *robust comparison* between s information retrieval systems based on the top N documents, all s systems *must* obtain relevance judgements for all N documents, and for every query from the set of queries.

The stated claim can be approximated by reading more documents for every participating system, or having more systems participate in CLEF that rank as many *mutually different* documents as possible within the top N.