Linguistic and Statistical Analysis of the CLEF Topics

Thomas Mandl, Christa Womser-Hacker

University of Hildesheim, Information Science, Marienburger Platz 22
D-31141 Hildesheim, Germany
mandl@rz.uni-hildesheim.de

Abstract. This paper reports on an analysis of the CLEF topics from the year 2001. We investigated especially potential correlations between features of the topics and the performance of retrieval systems. Although there are some weak relations, we can claim as a result, that the properties of the CLEF topics do not introduce any bias for the results of the retrieval systems. We found one correlation for the English topics. The more linguistic challenges the topic texts included, the better the performance of the systems.

1 Introduction

The effort in large scale evaluation studies can only be justified when the results are valid and can serve as a measure for the performance of the systems in real environments. One critical question has been the reliability of the assessors` relevance judgements. Reservations about the individuality of these assessments have been discussed widely. One study showed that the judgements between different jurors are in fact different. However, TREC and equally CLEF are comparison studies which aim at presenting a ranked list of systems. The absolute numbers of the performance values are not meant to exactly represent the quality of the systems. The study which assigned documents to several human relevance assessors showed the absolute numbers of the recall precision values are indeed affected by different judgements. However, the overall ranking of the systems does not change (Voorhees 1998).
In research on information retrieval evaluation, it has been pointed out, that „the quality of requests (and hence queries) appears very important“ (Sparck Jones 1995). In CLEF, there might be a higher uncertainty about assessment done by different people for the same topic as well as about a bias free translation of the topics.
The topic creation for a multilingual environment requires especial care in order to avoid cultural aspects to influence the meaning (Kluck & Womser-Hacker 2002). A thorough translation check of all translated topics was introduced in CLEF to assure that the translators who are only aware of two language versions of the topics do not omit or add any thematic aspects accidentally (Womser-Hacker 2002).

2 Topics in Information Retrieval Evaluation

The topics in information retrieval evaluation should express an information need. The topics are always formulated in natural language. In a real usage scenario, the information need is a state of mind. As a consequence, the transformation of the information need to the query is not tested in evaluation initiatives. The systems are confronted with a natural language formulation approximating the information need.
The creation of appropriate topics or queries for testing systems has been a major concern in the research on the evaluation of information retrieval. In one of the first studies which were based on the Cranfield collection, the queries led to considerable objections against the validity of the results. The queries were derived from the documents and basically each pointed to one document (cf. Womser-Hacker 1997).
TREC and subsequently CLEF have devoted considerable effort toward topic generation. TREC has seen a shift from elaborated topics to more and more shorter topics between TREC-3 and TREC-4 (cf. Sparck Jones 1994). This may have been a result of the web reality where information retrieval systems have been available mostly for free and have been widely used and where the queries are usually very short. In the web-track, only a few words are used as the actual topic while the narrative is only used for the assessment (cf. Hawking 2002).
Retrieval systems can be optimized for short queries by relying more on term expansion techniques (Kwok & Chan 1998). Even proper similarity calculation functions have been developed for short queries (cf. Wilkinson et al. 1995).
An investigation on TREC-1 topics has come to the conclusion that the topics are lexically distant from the relevant documents (Rorvig & Fitzpatri 1998). All judged documents for a topic were plotted in a two dimensional display in a way that the similarities from a similarity matrix were expressed by their distances as

far as possible. Multidimensional scaling was used in this study. The relevant documents lied closer to each other and appeared denser than the non relevant documents. However, when the topics was plotted in the same display, it usually had a relatively large distance from the relevant clusters. It can be concluded that relevance judgements require more than simply counting appearances of words. However, the relevant documents exhibit some formally detectable similarities.

3 Topics of CLEF 2001

As in previous years, the topics in CLEF 2001 have not been constructed e.g. from documents. Rather, they have been created in a natural way in order to closely resemble potential information needs.
This research has been mainly directed toward the following questions:

- What are the important and typical characteristics of the CLEF topics?
- Do these features have any influence on the quality of the retrieval results?
- May this knowledge be exploited for the development of retrieval systems?
- Should the design of the CLEF topics be altered to eliminate a potential bias for the systems?

For example, when looking at run EIT01M3N in the CLEF 2001 campaign, we see that it has a fairly good average precision of 0.341. However, for topic 44, which had an average difficulty, this run performs far below (0.07) the average for that topic (0.27). An intellectual analysis of the topics revealed that two of the most difficult topics contained no proper names and were both about sports (Topic 51 and 54).
In more detail, we want to find out whether there are linguistic characteristics of the CLEF topics which may hint toward a better or worse performance of the systems. The rationale for this assumption lies in the retrieval systems. From the literature it is well known that system developers invest great effort in the language processing capabilities. As a result, topics with more linguistic challenges may pose problems for some systems. In this context, it may be interesting to look at systems which perform well and demonstrate weaknesses for topics which are generally solved with good quality (or vice versa). Such an analysis seems to be especially interesting because the deviation between topics is larger than between systems and runs as the following table shows. Some features of the topics may have a higher influence on the result than the retrieval systems parameters. This phenomena is well known from TREC (cf. Womser-Hacker 1997).

Table 1. Overall statistics of average precision

	Average	Std. Deviation	Maximum	Mininum
All Runs	0.273	0.111	0.450	0.013
Topics over all languages	0.273	0.144	0.576	0.018
English Runs	0.263	0.074	0.373	0.104
English Topics	0.263	0.142	0.544	0.018
German Runs	0.263	0.092	0.390	0.095
German Topics	0.263	0.142	0.612	0.005

Table 1 shows that the average of all systems does surprisingly neither differ much form the average of all runs with English as topic language nor from all runs with German. For further analysis we can consider the performance for the topics (all systems for one topic) and the performance of the systems (one system for all topics). In this case the deviation for topics is higher (>0.14) than the deviation for the systems (<0.12) regardless of whether we consider only English topics, German topics or all languages.
However, there do not seem to be overall easy topics nor overall superior runs. The correlation between the average precision and the deviation for a topic over all runs is 0.83. That means, the better a topic is solved in average, the higher is the deviation of systems for that topic. Not all systems have achieved a similar quality of these easier topics. The same is true for runs. The average precision and the deviation for a run over all topics correlate with 0.84. We could draw the conclusion that no run performs well on all topics. In contrary, the better a run performed overall, the higher are the differences between its performance for single topics.
We assume that linguistic phenomena like morphological modifications, abbreviations or proper names pose challenges to the retrieval systems which they might handle with different approaches and consequently different quality. For example, a system performing generally very well may have difficulties when being confronted with

topics containing abbreviations and numbers. In such a case, these difficult queries could be handled by another system because they can be easily recognized automatically.
We carried out an intellectual analysis of the topics during which we assessed the values for the properties in the following table. For some of the features, types as well as tokens were assessed.

Table 2. Topic Properties

German topics:	English topics:
• original topic language • length • compound words • abbreviations • nominal phrases • proper names • negations • subordinate clauses • foreign words • numbers or dates	• original topic language • length • abbreviations • nominal phrases • proper names • negations • subordinate clauses • foreign words • parenthesis • numbers of dates

At first, we looked at the relation between topic features and the average precision of all runs for that topic without considering the systems properties. The overall quality of the run was measured with the following properties:

- Average precision (non- interpolated) for all relevant documents
- Precision at 5 documents
- For some analysis the runs were grouped into five buckets

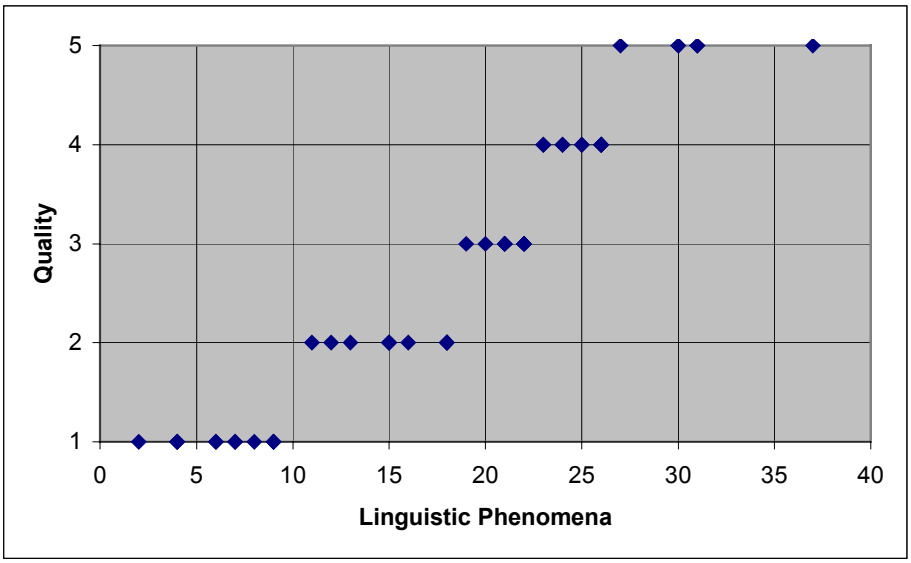

Figure 1. Relation between sum of number of linguistic phenomena and the runs' performance bucket

The correlation calculation between linguistic phenomena and average precision of all runs revealed the following. The only properties that had a correlation higher than 0.2 or lower than -0.2 for the number of phenomena and the precision were proper names. This indicates that the existence of proper names in the topic and consequently in the query makes a topic easier. Maybe proper names have a low frequency in the corpus and the same is true for other low frequency terms.

Table 3. Correlation between number of proper names and precision

	Types	Tokens	Sum of all linguistic phenomena
English	0.446	0.473	0.293
German	0.440	0.464	0.286

The sum of all linguistic phenomena also has a positive but lower correlation with the precision. However, if we consider the bucket analysis where the runs are grouped in five buckets, the correlation reaches 0.96 for the English topics. Each bucket contains 20% of all runs. Bucket one the worst runs and so forth. The relation can be seen in figure 1. The reported relations are not statistically significant since only 50 topics and 70 runs were analyzed.

4 Topics and Runs of CLEF 2001

The global analysis of topics and runs needs to be refined by a more detailed inspection on the relation between topic and run properties. To allow such a investigation, some properties of the runs were also considered in the analysis. They can be found in the CLEF proceedings:

- Multi or bilingual run
- Topic language
- Used topic fields (title, description, narrative)
- Used for the pool or not

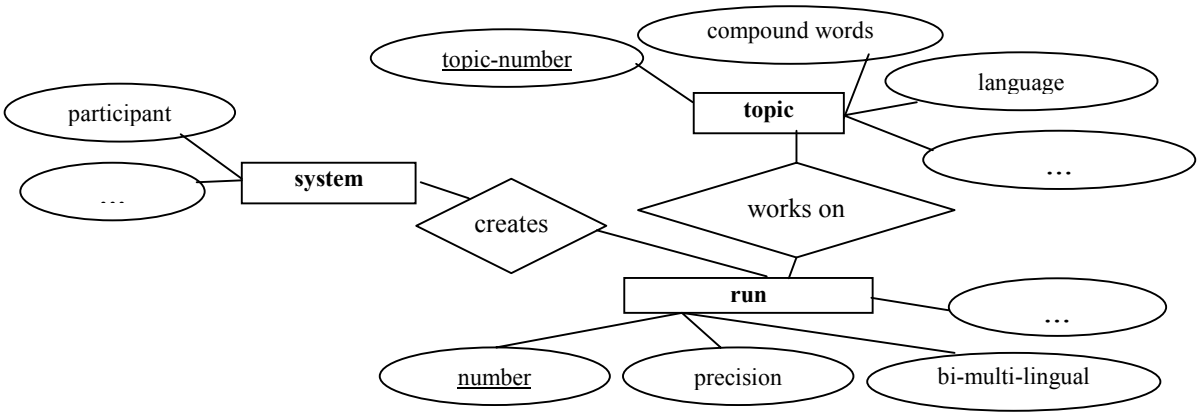

Figure 2. Data model for topics, systems and runs

As a target of a machine learning model, we used the average precision of the run for the topic in question. The basic data model is shown in figure 2.

A statistical analysis was not carried out beforehand, because there is relatively little evidence for a statistical analysis. There were only 600 combinations of runs and German topics and 900 for English topics. As a consequence, we first used machine learning methods to find out whether a formal model can be built which established any relations between topic properties and system performance. If any relations were to be detected, subsequent statistical analysis could have been applied.

We used the Waikato Environment for Knowledge Analysis (WEKA[1]) which allows the testing of many machine learning algorithms within the same framework. WEKA is implemented in JAVA and is provided as open source software.

Neither the German nor the English training data resulted in a satisfying model. This was true for linear algorithms like Naive Bayes as well as for non linear models like neural networks. That means, the existing data cannot be used to build any predictive model to map from system and topic properties to the quality of the run measured by the average precision.

[1] http://www.cs.waikato.ac.nz/~ml/weka/index.html

5 Outlook

The investigations reported in this article have found no bias in the topics of the CLEF campaign. As a result, no reservations toward the validity of the results arise from this research.
At this point, the topics of the 2002 campaign are evaluated. Further languages should also be included in the analysis. Furthermore, the features of the topics of 2002 are assessed once again. The assessment requires human judgement and therefore, we want to eliminate possible individual bias. Additional machine assessable properties like sentence complexity and part of speech frequency analysis are also considered. Domain specific words may also be worth an analysis. Another interesting question could be the comparison of topics with similar difficulty. Are these indeed similar topics?
In other areas of TREC and CLEF, topics properties and tuning system toward them might play an even more crucial role. In the web track, two different types of topics have been developed. In addition to the topical searches adopted from the ad hoc track, homepage finding has been defined as a task. In multimedia retrieval, the topic creation gets more difficult. For example, the video track in TREC comprises several dimensions of topic properties. Topics are characterized formally by the multimedia elements contained. In addition to the mandatory text description, topics may embrace graphic, video or audio files. Furthermore, the topics content is much more diverse than for text retrieval. Their semantics differs greatly. Some demand video with certain graphical features, other require objects, other aim at certain persons during an activity (Smeaton et al. 2002).

References

Hawking, David; Craswell, Nick (2001): Overview of the TREC-2001 Web Track. In: Voorhees & Harman 2001.

Kluck, Michael; Womser-Hacker, Christa (2002): Inside the Evaluation Process of the Cross-Language Evaluation Forum (CLEF): Issues of Multilingual Topic Creation and Multilingual Relevance Assessment. In: 3rd International Conference on Language Resources and Evaluation, Las Palmas, Spain.

Kwok, K. L.; Chan, M. (1998): Improving Two-Stage Ad-Hoc Retrieval for Short Queries. In: Proceedings of the 21st Annual International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR ´98) Melbourne. pp. 250-256.

Rorvig, Mark; Fitzpatri, Steven (1998): Visualization and Scaling of TREC Topic Document Sets. In: Information Processing and Management 34(2-3): pp. 135-149

Smeaton, Alan; Over, Paul; Taban, Ramazan (2002): The TREC-2001 Video Track Report. In: Voorhees & Harman 2001.

Sparck Jones, Karen (1995): Reflections on TREC. In: Information Processing & Management 31(3) pp. 291-314 .

Voorhees, Ellen (1998): Variations in relevance judgments and the measurement of retrieval effectiveness. In: Proceedings of the 21st Annual International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR ´98) Melbourne. pp. 315-223.

Voorhees, Ellen; Harman, Donna (Eds.)(2001): The Tenth Text Retrieval Conference (TREC-10). NIST Special Publication. National Institute of Standards and Technology. Gaithersburg, Maryland. http://trec.nist.gov/pubs/

Wilkinson, Ross; Zobel, Justin; Sacks-Davis, Ron (1995): Similarity Measures for Short Queries. In: Harman, Donna (Ed.) The Fourth Text REtrieval Conference (TREC-4). NIST Special Publication. National Institute of Standards and Technology. Gaithersburg, Maryland, http://trec.nist.gov/pubs/

Womser-Hacker, Christa (1997): Das MIMOR-Modell. Mehrfachindexierung zur dynamischen Methoden-Objekt-Relationierung im Information Retrieval. Habilitationsschrift. Universität Regensburg, Informations-wissenschaft.

Womser-Hacker, Christa (2002): Multilingual Topic Generation within the CLEF 2001 Experiments. In: Peters, Carol; Braschler, Martin; Gonzalo, Julio; Kluck, Michael (Eds.): Evaluation of Cross-Language Information Retrieval Systems. Springer [LNCS 2406] pp. 389-393.