Proper Names in the Multilingual CLEF Topic Set

Thomas Mandl, Christa Womser-Hacker

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

Abstract. The reliability of the topics within the Cross Language Evaluation Forum (CLEF) needs to be evaluated constantly to justify the efforts for experiments within CLEF and to demonstrate the validity of the results as far as possible. The analysis presented in this paper is concerned with several aspects. Continuing and expanding a study from 2002, we investigate the difficulty of topics and the correlation between the retrieval quality for topics and the occurrence of proper names.

1 Introduction

Topics are an essential aspect of experiments for information retrieval evaluation (Sparck Jones 1995). The topic creation for a multilingual environment requires especial care in order to avoid cultural bias to influence the semantics of a topic formulation (Kluck & Womser-Hacker 2002). A thorough translation check of all translated topics in CLEF assures that the translations all include the same semantics (Womser-Hacker 2002).
Such a high level intellectual assessment cannot guarantee that on a lower level some unidentified linguistic properties lead to a bias. Therefore, we continued an analysis of the CLEF 2001 topics and results (Mandl & Womser-Hacker 2002). This time, we concentrated on proper names. In addition, the notion of difficulty of a topic and the validity of the translations is further explored.

2 Analysis of Information Retrieval Evaluation Results

The validity of large-scale information retrieval experiments has been the subject of a considerable amount of research. Zobel 1998 concluded that the TREC (Text REtrieval Conference[1]) experiments are reliable as far as the ranking of the systems are concerned. Voorhees & Buckley 2002 have analyzed the reliability of experiments as a function of the size of the topic set. They concluded that the typical size of the topic set in TREC is sufficient for a satisfactory level of reliability.
Further research is dedicated toward the question whether the expensive human relevance judgements are necessary or whether the constructed document pool of the highest ranked documents from all runs may serve as an reliable approximation of the human judgements. According to a study by Soboroff et al. (2001), the ranking of the systems in TREC correlates positively to a ranking based on the document pool without further human judgement. However, there are considerable differences in the ranking which are especially significant in the highest ranks. As a consequence, the human judgements are necessary to achieve the highest reliability of the system rankings. It has also been shown that different human jurors judge the relevance of documents differently. That means, different jurors find different sets of relevant documents. This disagreement between different jurors does not result in different system rankings (Voorhees 1998). Although the different sets of relevant documents lead to different recall and precision values, the final rankings of the runs remain rather stable.

3 Difficulty of Topics

The notion of difficulty of topics has been of great interest in the IR community. The question of what makes a topic a difficult one remains unsolved.
Voorhees & Harman 1997 measured the difficulty of TREC topics from two perspectives. One was the estimation of experts and the second was the actual outcome of the systems measured as the average precision

[1] http://trec.nist.gov

which systems achieved for that topic. They found no correlation between the two measures. This result was confirmed in a study of the topics of the Asian languages retrieval evaluation NTCIR[2] (Eguchi et al. 2002). Furthermore, Eguchi et al. 2002 tried to find whether the system ranking changes when different difficult levels of topics were considered. They conclude, that changes in the system ranking occur, however, the Kendall correlation coefficient between the overall rankings does not drop below 0.69. For that analysis, the actual difficulty measured by the precision of the runs was used. The overall rankings remain stable but top ranks could be affected. When considering this result, we need to be aware, that the number of topics in the sub-sets with different difficulties is lower than in the overall set. According to the results from Voorhees & Buckley 2002 who analyzed the reliability of experiments as a function of the size of the topic set, such a small set does not lead to completely reliable results.

We conducted a study for the CLEF topics of the year 2003 to investigate the correlation between perceived topic difficulty and actual performance of the systems. We call the average precision of the best run for a topic the *system difficulty*. The best run was chosen to get the best possible retrieval quality that a system can reach within the CLEF campaign. Accordingly, the human estimation of difficulty is called the *intellectual difficulty* of a topic.

The intellectual difficulty was surveyed during the CLEF topic creation meeting. The topic creators were asked to judge whether a topic would be difficult for systems and whether it would be difficult for the human jurors to assess the relevance of the documents. The system difficulty was extracted from the output of the trec_eval program which was mailed to all participants as part of the result.

Surprisingly, only a weak correlation (according to Breavis / Pearson) of 0.14 could be found between the system and the intellectual difficulty. It seems, that the difficulty of a topic cannot even be judged sufficiently well by CLEF experts. However, the judgments about the difficulty of the assessment yielded a stronger positive correlation (0.30) to the topic system difficulty. The CLEF campaign also provides the number of relevant documents for each topic. The judgements about the difficulty of the assessment show a negative correlation (-0.27) to the number of relevant documents found. That means, that for topics which the topic creators consider difficult to assess actual point to fewer relevant documents. The two correlation measures are statistically significant at a level of 95%.

4 Proper Names and Retrieval Performance

Much of the work in information retrieval needs to be dedicated to natural language processing. Phenomena like different word forms or compound words face challenges for information retrieval systems. Therefore, the occurence of linguistic phenomena may favor some systems especially suited for these phenomena. In this context, it may be interesting to look at systems which generally perform well but demonstrate weaknesses for topics which are overall solved with good quality (or vice versa).

A study has been carried out for the CLEF campaign 2001 (Mandl & Womser-Hacker 2002). It revealed no correlation between any single linguistic phenomenon and the system difficulty of a topic. Not even the length of a topic showed any effect. However, when calculating the sum of all phenomena assessed, we found a positive correlation. The more phenomena available, the better systems solved a topic in average. This study will be extended to the topics and results quality of all available CLEF runs.

Intellectual analysis had identified proper names as a potential indicator for retrieval performance. Because of that, proper names in the CLEF topic set were analyzed in more detail. The analysis included all topics from the campaigns in the years 2001 and 2002. The number of proper names in the topics were assessed intellectually. In addition, the retrieval performance of the runs for the topics was extracted from the appendix of the CLEF proceedings.

Overall, we found a significant positive relation between the number of proper names present. The more proper names a topic contains, the better the retrieval results are. There is a high deviation, however, the trend is statistically significant at a level of 95%. .

In detail, the following relations were investigated: For the CLEF topics number 41 to 200, the average precision of the best run was available. For the topics 41 throughout 140, the average of all runs for that topic was available. It was calculated as the average of all average precision values. Thus, one best and one average result for each topic was identified. The topics were ordered according to the number of proper names they contained. Each subset of topics and their performance values were assigned to the number of proper names. For these sets, the basic statistics are shown in table 1 and 2.

[2] http://research.nii.ac.jp/ntcir/

Table 1. Best run for each topic in relation to the number of proper names in the topic (topic 41 to 200)

Number of proper names	**0**	**1**	**2**	**3**	**4**	**5**	**6**
Number of Topics	42	43	40	20	9	4	2
Average of Best System per Topic	0.62	0.67	0.76	0.83	0.79	0.73	0.81
Minimum of Best System per Topic	0.090	0.12	0.036	0.28	0.48	0.40	0.63
Standard Deviation of Best System per Topic	0.24	0.24	0.24	0.18	0.19	0.29	0.26

Table 2. Average retrieval of runs for topic in relation to the number of proper names in the topic (topic 41 to 140)

Number of proper names	**0**	**1**	**2**	**3**	**4**	**5**
Number of Topics	33	22	20	13	8	3
Maximum of Average Performance per Topic	0.49	0.52	0.74	0.69	0.58	0.60
Average of Average Performance per Topic	0.21	0.29	0.39	0.39	0.32	0.46
Minimum of Average Performance per Topic	0.02	0.10	0.14	0.12	0.17	0.28
Standard Deviation of Average Performance	0.13	0.13	0.16	0.17	0.15	0.17

It needs to be noted that the number of topics is higher for fewer proper names. There are not enough topics with three to six proper names to draw far reaching conclusions. Figure 1 and 2 visualize the numbers in the tables above. They need to be interpreted the following way: Each sub-set of topics with a certain number of proper names is marked on the X-axis. Within this sub-set, figure 1 and figure 2 show the best and average performance for that topic. Each figure shows the minimum and the average performance for that sub-set. For the average performance for the topic, the performance for the best topic is also displayed in figure 2. It is not shown for the best performance in figure 1 because that equals 1 usually.

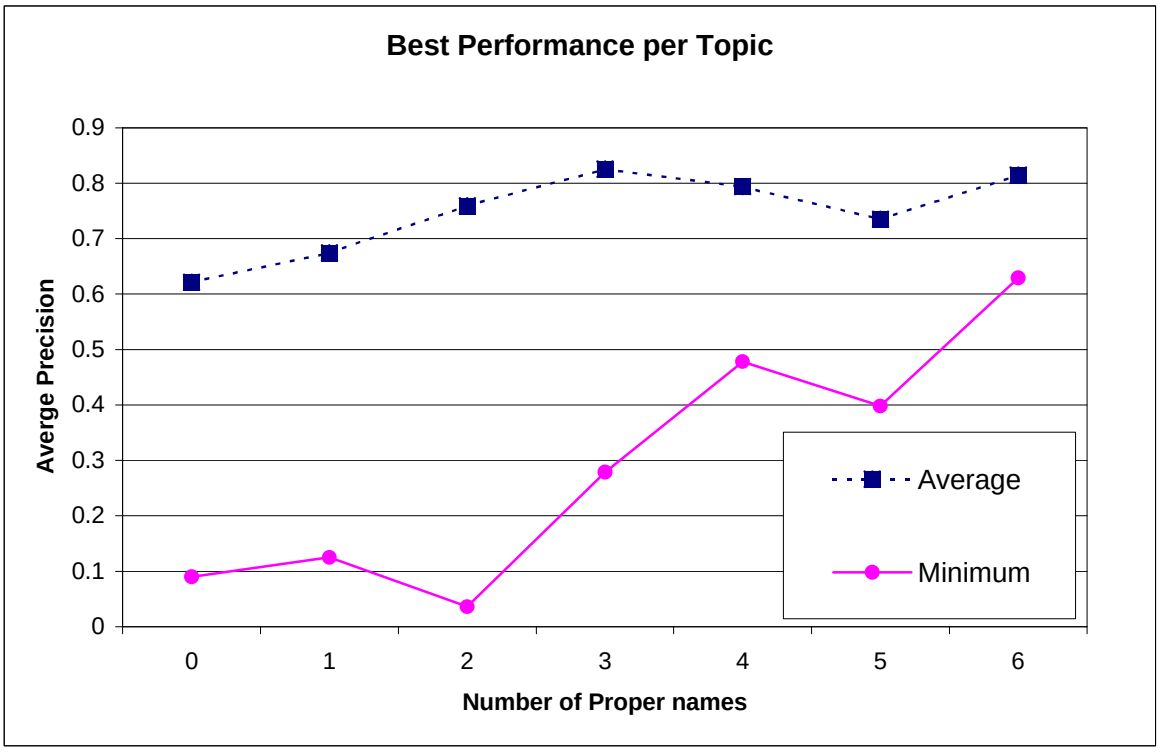

Figure 1. Relation between number of proper names and retrieval performance for the best runs

The graphic representation of the values of table 1 and 2 in figure 1 and 2 respectively suggests a positive correlation between the number of proper names present in a topic and the difficulty of that topic expressed as

the performance of the systems for that topic. This impression is confirmed by a mathematical analysis. The average performance correlates with a value of 0.43 to the number of proper names and the best performance with a value of 0.26. Disregarding the topics from the campaign in 2003 leads to a correlation coefficient of 0.35. All relations are statistically significant. The assumption of independence of the two parameters can be rejected with a probability of over 95%.

This study suggests that the presence of proper names in queries enhances the chances of retrieval systems to find relevant documents. As a consequence, the presence of proper names in the CLEF topic set should be carefully monitored. The system rankings for topics with and without proper names should be evaluated.

This analysis of proper names could be extended to consider the number of tokens, to include the number of individual word present in the proper names as well as abbreviations of proper names.

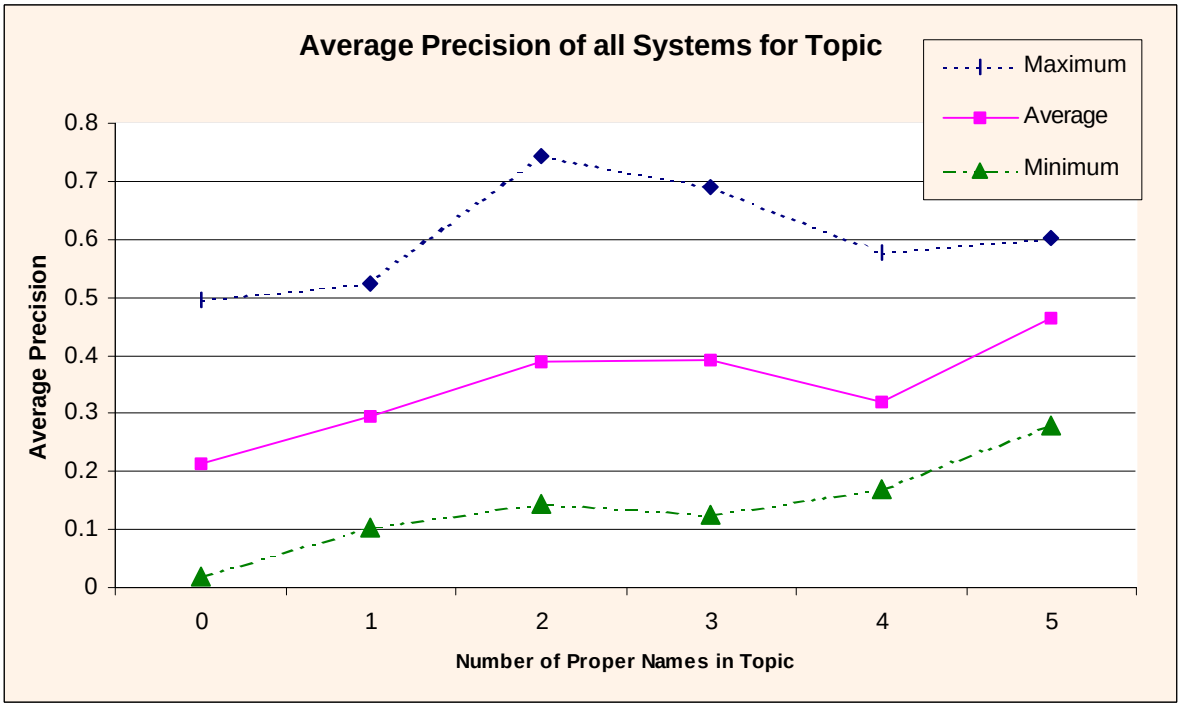

Figure 2. Relation between number of proper names and retrieval performance for the average of all runs for a topic

5 Proper Names in the Multilingual Topic Set

An objection against the reliability of the CLEF experiments could arise from the multilingual aspects of the topic set. Are the topics in one language easier for the systems than in another? Does this lead to a bias in the multilingual track where different languages are used as starting point? For example, due to the stylistic modifications in a language, more occurences of proper names or more heterogeneous word forms may be introduced in the topic formulation. Because proper names were identified as an important aspect, we analyzed the stability of the distribution of proper names over the topic set. The results showed large differences between the number of proper names identified in the topic sets among different languages. This fact seems surprising and several potential explanations seem plausible.

One reason could be the different proficiency level of the graduate students analyzing the topics. Better knowledge of one language could lead to the detection of more proper names.

Another explanation could be that the stylistic and grammatical necessities lead to different numbers of names. Sometimes, in one language more proper names referring to the same entity are commonly used than in other languages (for example: *United Kingdom, England, Great Britain* in English vs. *Inglaterra* in Portuguese or *United States of America, United States, America, US , USA* in English vs. *Estados Unidos, EAU* in Portuguese). For stylistic reasons, several of these different lexical entries may be used. Definitely, further analysis is necessary.

Because the number of proper names seems to have substantial influence on the retrieval quality as shown in chapter 4, this fact needs to be considered. The performance of systems utilizing from different topic languages needs to be analyzed.

6 Outlook

The investigations reported in this article have found no bias in the topics of the CLEF campaign. No reservations about the validity of the results arise from this research. This work needs to continue throughout the campaign and has to be extended to further aspects in the future.

Acknowledgements

We would like to thank Martin Braschler for providing the crucial data for our study. We also acknowledge the participation of the CLEF organization team during the topic creation meeting for their judgements on the difficulty of the topics. Furthermore, we acknowledge the work of several students from the University of Hildesheim who contributed to this analysis as part of their course work, especially Kathrin Wünnemann, Nikolaus Küster and Carsten Spichal.

References

Eguchi, Koji; Kuriyama, Kazuko; Kando, Noriko (2002): Sensitivity of IR systems Evaluation to Topic Difficulty. In: Proceedings of the 3rd International Conference on Language Resources and Evaluation (LREC 2002), Vol.2, Las Palmas de Gran Canaria, Spain, pp.585-589

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.

Mandl, Thomas; Womser-Hacker, Christa (2002): Linguistic and Statistical Analysis of the CLEF Topics. In: Peters, Carol; Braschler, Martin; Gonzalo, Julio; Kluck, Michael (eds.): Evaluation of Cross-Language Information Retrieval Systems. Proceedings of the CLEF 2002 Workshop. Berlin et al.: Springer [LNCS]

Soboroff, Ian; Nicholas, Charles; Cahan, Patrick (2001): Ranking Retrieval Systems without Relevance Judgements. In: Proc Annual Intl ACM Conference on Research and Development in Information Retrieval (SIGIR ‘01) New Orleans. pp. 66-73.

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; Buckley, Chris (2002): The Effect of Topic Set Size on Retrieval Experiment Error. In: Proc Annual Intl ACM Conference on Research and Development in Information Retrieval (SIGIR ‘02) Tampere, Finland. pp. 316-323.

Voorhees, Ellen; Harman, Donna (1997): Overview of the Sixth Text Retrieval Conference. In: Voorhees, Ellen; Harman, Donna (Eds.): The Sixth Text Retrieval Conference (TREC-6). NIST Special Publication. National Institute of Standards and Technology. Gaithersburg, Maryland. http://trec.nist.gov/pubs/

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.

Zobel, Justin (1998): How Reliable are the Results of Large-Scale Information Retrieval Experiments? In: Proc Annual Intl ACM Conference on Research and Development in Information Retrieval (SIGIR ‘98) Melbourne. pp. 307-314.