How do Named Entities Contribute to Retrieval Effectiveness?

Thomas Mandl, Christa Womser-Hacker

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

Abstract. Named entities in topics are a major factor contributing to the quality of retrieval results. In this paper, we report on an analysis on the correlation between the number of named entities present in a topic and the retrieval quality achieved for these topics by retrieval systems within CLEF. We found that a medium positive correlation exists for German, English and Spanish topics. Furthermore, we analyze the effect of the document or target language on the retrieval quality.

1 Introduction

Within CLEF, many efforts are made to improve retrieval systems. This body of work allows the identification of successful approaches, algorithms and tools in CLIR (Braschler & Peters 2004).
We believe, the knowledge and work dedicated to this effort can be exploited beyond the optimization of individual systems. The amount of data created by organizers and participants remains a valuable source of knowledge awaiting exploration. Many lessons can still be learned from evaluation initiatives such as CLEF, TREC (Voorhees & Buckland 2002), INEX (Fuhr 2003) or NTCIR (Oyama, Ishida & Kando 2003).
Ultimately, further criteria and metrics for the evaluation of search and retrieval methods may be detected. This could lead to improved algorithms, quality criteria, resources and tools in CLIR (Schneider et al. 2004). This general research approach is illustrated in figure 1. The identification of patterns in the systems' performance for topics with specific items may lead to improvements in system development.

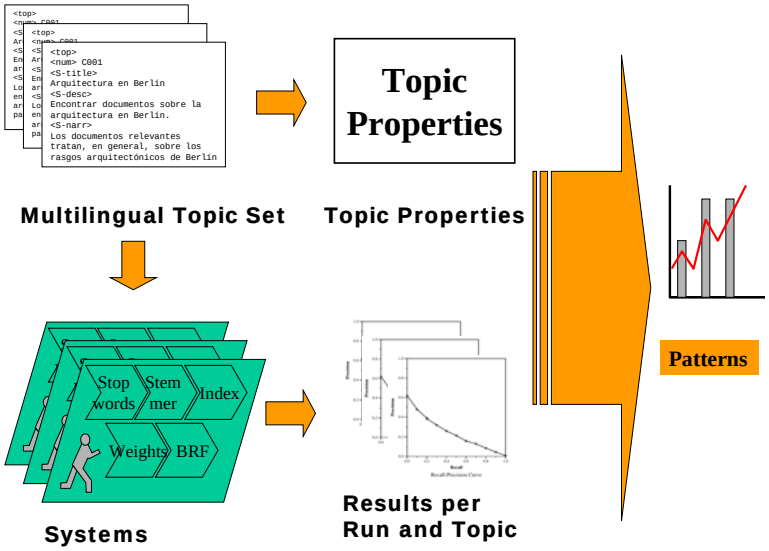

Figure 1. Overview of the approach

Our current analysis concentrates on named entities within the topics of CLEF. Named entities frequently occur in CLEF as part of the topic formulation. Table 1 gives an overview.

Table 1. Name of named entities in the CLEF topics

CLEF year	Number of topics	Total number of named entities	Average number of named entities in topics	Standard deviation of named entities in topics
2001	50	60	1.20	1.06
2002	50	86	1.72	1.54
2003	60	97	1.62	1.18

The large number of named entities in the topic set shows that they are a subject worth studying. The large number may be due to the fact that the document corpus for CLEF consists of newspaper texts. We can also observe an increase of named entities per topic in 2002 compared to 2001. Because of the effect of named entities on retrieval performance (Mandl & Womser-Hacker 2004c), the number of named entities needs to be carefully monitored. Table 2 shows how the named entities are distributed over groups with different numbers of named entities and shows the tasks analyzed in this paper.

Table 2. Overview of named entities in CLEF tasks

CLEF year	Task	Topic language	Nr. runs	Topics without named entities	Topics with one or two named entities	Topics with more than three named entities
2002	Mono	German	21	12	21	17
2002	Mono	Spanish	28	11	18	21
2002	Bi	German	4	12	21	17
2002	Multi	German	4	12	21	17
2002	Bi	English	51	14	21	15
2002	Multi	English	32	14	21	15
2003	Mono	English	11	8	14	6
2003	Mono	Spanish	38	6	33	21
2003	Multi	Spanish	10	6	33	21
2003	Mono	German	30	9	40	10
2003	Bi	German	24	9	40	10
2003	Multi	German	1	9	40	10
2003	Bi	English	8	9	41	10
2003	Multi	English	74	9	41	10

2 Named Entities in Topics and Retrieval Performance

In a study presented at CLEF in 2003, we showed a correlation between the number of named entities present in topics and the systems' performance for these topics (Mandl & Womser-Hacker 2004b). In this paper, we extend the analysis to Spanish and monolingual tasks. In our earlier analysis, the relation was shown for English and German. Including Spanish will show, whether this effect can be revealed for another topic language. By including monolingual tasks, we may be able to compare the strength of the effect between cross- and monolingual retrieval tasks.

Named entities were intellectually assessed according to the schema of Sekine et al. 2002. The performance of the systems was extracted from the CLEF proceedings. The average precision for a topic is calculated as the average precision of all systems for a individual topic. From the average precision for a topic, we can calculate the average of all topics which contain n named entities. Figure 2 and 3 show the average precision for topics with n named entities for tasks in CLEF 3 and CLEF 4.

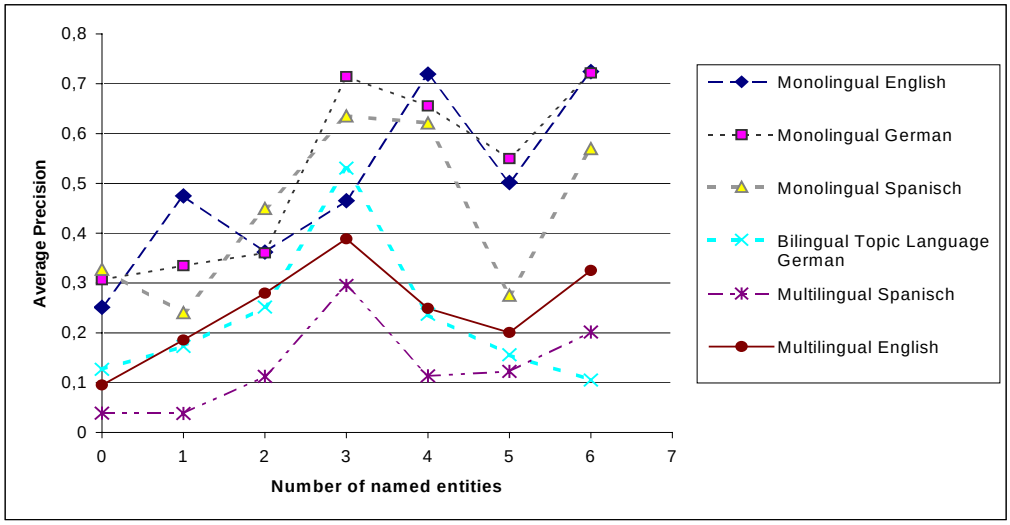

Figure 2. Average precision for topics with n named entities in CLEF 3 (in 2002)

In figure 2 and 3 we can observe that monolingual tasks generally result in higher average precision than cross-lingual tasks. The precision tends to be better when more named entities are present.
The relation previously observed for German and English can also be seen for Spanish.

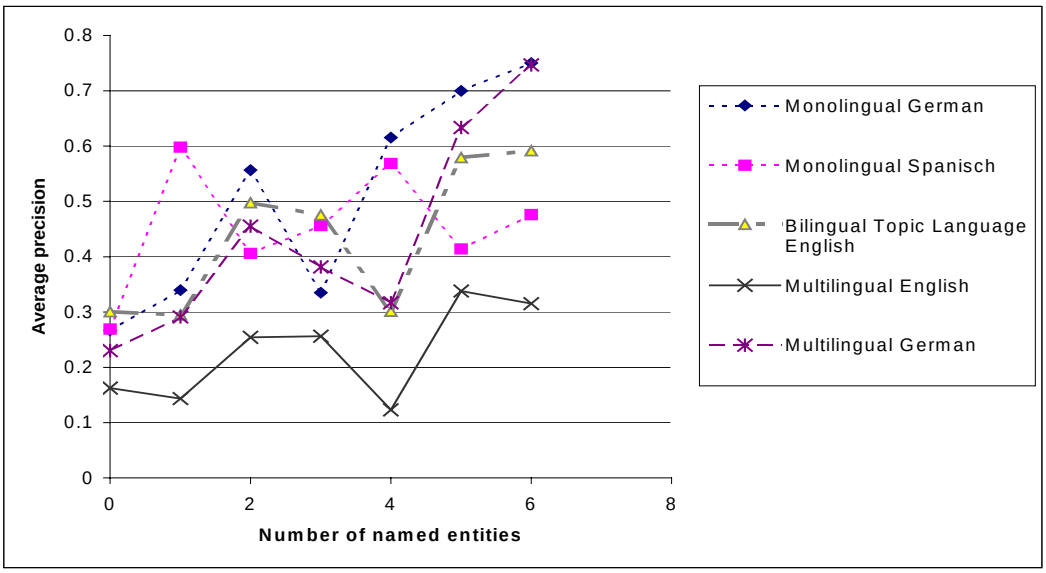

Figure 3. Average precision for topics with n named entities in CLEF 4 (in 2003)

We also calculate the correlation between the number of named entities and the average precision per topic for each of the tasks. The results are presented in table 3 and 4.

Table 3. Correlation between the number of named entities in topic and the average system performance per topic for tasks in CLEF 3

Monolingual German	Monolingual Spanish	Bilingual Topic Language English	Multilingual German	Multilingual English
0.449	0.207	0.399	0.428	0.294

Table 4. Correlation between the number of named entities in topic and the average system performance per topic for tasks in CLEF 4

Monolingual German	Monolingual Spanisch	Monolingual English	Bilingual Topic Language German	Multilingual Spanisch	Multilingual English
0.372	0.385	0.158	0.213	0.213	0.305

We can observe that the correlation is in most cases higher for the monolingual task. That would mean, that named entities help systems more in monolingual retrieval than in cross-lingual retrieval. However, English seems to be an exception in CLEF 4, because the correlation is almost twice as strong in the multilingual task.

3 Potential for Optimization based on Named Entities

The systems tested at CLEF perform differently well for topics with different numbers of named entities. Although proper names make topics easier in general and for almost all runs, the performance of systems varies within the three classes of topics based on the number of named entities. We distinguished three classes of topics, (a) the first class with no proper names called *none*, (b) the second class with one and two named entities called *few* and (c) one class with three or more named entities called *lots*. The patterns of the systems are strikingly different for the three classes. As a consequence, there seems to be potential to improve system by fusion based on the number of named entities in a topic. Many systems already apply fusion techniques.
We propose a simple fusion rule. First, the number of named entities is determined for each topic. Subsequently, this topic is channeled to the system with the best performance for this named entity class. The best system is a combination of at most three runs. Each category of topics is answered by the optimal system within a group of systems for that number of named entities. The groups were selected from the original CLEF ranking of the runs in one task. We used a window of five runs. That means, five neighboring runs by systems which perform similarly well overall are grouped and fused by our approach. Table 5 shows the improvement by the fusion based on the optimal selection of a system for each category of topics.
The highest levels of improvement are achieved for the topic language English. For 2002, we observe the highest improvement of 10% for the bilingual runs.

Table 5. Improvement through named entity based fusion

CLEF year	Run type	Topic language	Average. precision best run	Optimal average precision name fusion	Improvement over best run
2001	Bilingual	German	0.509	0.518	2%
2001	Multilingual	English	0.405	0.406	0%
2002	Bilingual	English	0.494	0.543	10%
2002	Multilingual	English	0.378	0.403	6.5%
2003	Bilingual	German	0.460	0.460	0%
2003	Bilingual	English	0.348	0.369	6.1%
2003	Multilingual	English	0.438	0.443	1.2%

This approach regards the systems as black boxes and requires no knowledge about the treatment of named entities within the systems. Considering the linguistic processing within the systems might be even more rewarding. Potentially, further analysis might reveal which approaches, which components and which parameters are especially suited for topics with and without named entities.
This analysis shows that the performance of retrieval systems can be optimized by channeling topics to the systems best appropriated for topics without, with one or two and with three and more names. Certainly, the application of this fusion on the past results approach is artificial and the number of topics in each subgroup is not sufficient for a statistically reliable result (Voorhees & Buckley 2002). Furthermore, in our study, the number of named entities was determined intellectually. However, this mechanism can be easily implemented by using an automatic named entity recognizer. We intend to apply this fusion technique in an upcoming CLEF task as one element of the fusion framework MIMOR (Mandl & Womser-Hacker 2004a, Hackl et al 2004).

4 Named Entities in Topics and Retrieval Performance for Target Languages

So far, our studies have been focused to the language of the initial topic which participants used for their retrieval efforts. Additionally, we have analyzed the effect of the target or document language. In this case, we cannot consider the multilingual tasks where there are several target languages. The monolingual tasks have already been analyzed in section 2 and are also considered here. Therefore, this analysis is targeted at bilingual retrieval tasks. We grouped all bilingual runs with English, German and Spanish as document language. The correlation between the number of named entities in the topics and the average precision of all systems for that topic was calculated. The average precision may be interpreted as the difficulty of the topic. The following table shows the results of this analysis.

Table 6. Correlation for target languages for CLEF 3 and 4

CLEF year	Task type	Target language	Number of runs	Correlation between number of named entities and average precision
2003	Mono	English	11	0.158
2002	Bi	English	16	0.577
2003	Bi	English	15	0.187
2002	Mono	German	21	0.372
2003	Mono	German	30	0.449
2002	Bi	German	13	0.443
2003	Bi	German	3	0.379
2002	Mono	Spanish	28	0.385
2003	Mono	Spanish	38	0.207
2002	Bi	Spanish	16	0.166
2003	Bi	Spanish	25	0.427

First, we can see a positive correlation for all tasks considered. Named entities support the retrieval also from the perspective of the document language. This results for the year 2002 may be a hint, that retrieval in English or German document collections profits more from named entities in the topic than Spanish. However, in 2003, the opposite is the case and English and Spanish switch. For German, there are only 3 runs in 2003. As a consequence, we cannot yet detect any language dependency for the effect of named entities on retrieval performance.

5 Outlook

In this paper a strong relation between named entities in topics and the performance of retrieval systems for these topics was confirmed. This finding allows us to formulate a hint for searchers and users of retrieval systems: Whenever you can think of a name related to your retrieval problem, consider including it in the query.
In addition, our results encourage further analysis of other topic features. We are especially considering a part of speech (POS) analysis of the CLEF topics.

Acknowledgements

We would like to thank Martin Braschler for providing the crucial data for our study. Furthermore, we acknowledge the work of several students from the University of Hildesheim who contributed to this analysis as part of their course work.

References

Braschler, Martin; Peters, Carol (2004): Cross-Language Evaluation Forum: Objectives, Results, Achievements. In: Information Retrieval. no. 7. pp. 7-31.

Fuhr, Norbert (2003) (ed.): Initiative for the Evaluation of XML Retrieval (INEX) : INEX 2003 Workshop Proceedings, Dagstuhl, Germany, December 15-17, 2003. http://purl.oclc.org/NET/duett-07012004-093151

Hackl, René; Kölle, Ralph; Mandl, Thomas; Ploedt, Alexandra; Scheufen, Jan-Hendrik; Womser-Hacker, Christa (2004): Multilingual Retrieval Experiments with MIMOR at the University of Hildesheim. To appear in: Peters, Carol; Braschler, Martin; Gonzalo, Julio; Kluck, Michael (eds.): Evaluation of Cross-Language Information Retrieval Systems. Proceedings of the CLEF 2003 Workshop. Berlin et al.: Springer [Lecture Notes in Computer Science]

Mandl, Thomas; Womser-Hacker, Christa (2004a): A Framework for long-term Learning of Topical User Preferences in Information Retrieval. In: New Library World. vol. 105 (5/6). pp. 184-195.

Mandl, Thomas; Womser-Hacker, Christa (2004b): Analysis of Topic Features in Cross-Language Information Retrieval Evaluation. In: 4th International Conference on Language Resources and Evaluation (LREC) Lisbon, Portugal, May 24-30. Workshop Lessons Learned from Evaluation: Towards Transparency and Integration in Cross-Lingual Information Retrieval (LECLIQ). pp. 17-19.

Mandl, Thomas; Womser-Hacker, Christa (2004c): Proper Names in the Multilingual CLEF Topic Set. To appear in: Peters, Carol; Braschler, Martin; Gonzalo, Julio; Kluck, Michael (eds.): Evaluation of Cross-Language Information Retrieval Systems. Proceedings of the CLEF 2003 Workshop. Berlin et al.: Springer [Lecture Notes in Computer Science] Preprint: http://clef.iei.pi.cnr.it:2002/2003/WN_web/53.pdf

Oyama, Keizo; Ishida, Emi; Kando, Noriko (2003) (eds.): NTCIR Workshop3 Proceedings of the Third NTCIR Workshop on research in information Retrieval, Automatic Text Summarization and Question Answering (September2001-October2002) http://research.nii.ac.jp/ntcir/workshop/OnlineProceedings3/index.html

Schneider, René; Mandl, Thomas; Womser-Hacker, Christa (2004): Workshop LECLIQ: Lessons Learned from Evaluation: Towards Integration and Transparency in Cross-Lingual Information Retrieval with a special Focus on Quality Gates. In: 4th International Conference on Language Resources and Evaluation (LREC) Lisbon, Portugal, May 24-30. Workshop Lessons Learned from Evaluation: Towards Transparency and Integration in Cross-Lingual Information Retrieval (LECLIQ). pp. 1-4.

Sekine, Satoshi; Sudo, Kiyoshi; Nobata, Chikashi (2002): Extended Named Entity Hierarchy. In: Proceedings of Third International Conference on Language Resources and Evaluation (LREC 2002); Las Palmas, Canary Islands, Spain.

Voorhees, Ellen; Buckland, Lori (2002) (eds.): The Eleventh Text Retrieval Conference (TREC 2002). NIST Special Publication 500-251. National Institute of Standards and Technology. Gaithersburg, Maryland. November 2002. http://trec.nist.gov/pubs/trec11/t11_proceedings.html

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.

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.