Computational Linguistics and Chinese Language Processing
Vol. 13, No. 1, March 2008, pp. 121-140

Analyzing Information Retrieval Results With a Focus on Named Entities

Thomas Mandl[*] and Christa Womser-Hacker[*]

Abstract

Experiments carried out within evaluation initiatives for information retrieval have been building a substantial resource for further detailed research. In this study, we present a comprehensive analysis of the data of the Cross Language Evaluation Forum (CLEF) from the years 2000 to 2004. Features of the topics are related to the detailed results of more than 100 runs. The analysis considers the performance of the systems for each individual topic. Named entities in topics revealed to be a major influencing factor on retrieval performance. They lead to a significant improvement of the retrieval quality in general and also for most systems and tasks. This knowledge, gained by data mining on the evaluation results, can be exploited for the improvement of retrieval systems as well as for the design of topics for future CLEF campaigns.

Keywords: Cross-Lingual Information Retrieval, Evaluation Issues, Named Entities (NEs)

1. Introduction

The Cross Language Evaluation Forum (CLEF) provides a forum for researchers in information retrieval and manages a testbed for mono- and cross-lingual information (CLIR) retrieval systems. CLEF allows the identification of successful approaches, algorithms, and tools in CLIR. Within CLEF, various strategies are employed in order to improve retrieval systems [Braschler and Peters 2004; di Nunzio *et al.* 2007].

We believe that the effort dedicated to large scale evaluation studies 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 past data of evaluation initiatives such as CLEF, TREC [Voorhees and

[*] Information Sci., University of Hildesheim, Marienburger Platz 22, 31141 Hildesheim, Germany.
Tel.: +49-5121-883 ext: 837
The author for correspondence is Thomas Mandl.
E-mail: mandl@uni-hildesheim.de

[Received June 30, 2007; Revised October 8, 2007; Accepted January 12, 2008]

Buckland 2002], INEX [Fuhr 2003], NTCIR [Oyama *et al.* 2003], or IMIRSEL [Downie 2003].

Ultimately, further criteria and metrics for the evaluation of search and retrieval methods may be found. This could lead to improved algorithms, quality criteria, resources, and tools in cross language information retrieval [Harman 2004; Schneider *et al.* 2004]. This general research approach is illustrated in Figure 1.

Topics are considered an essential component of experiments for information retrieval evaluation [Sparck Jones 1995]. In most evaluations, the variation between topics is larger than the variation between systems. The topic creation for a multilingual test environment requires special care in order to avoid cultural or linguistic bias influencing the semantics of topic formulations [Kluck and Womser-Hacker 2002]. It must be assured that each topic provides equal conditions as starting points for the systems. The question remains whether linguistic aspects randomly appearing within the topics have any influence on the retrieval performance. This is especially important, as we observed in some cases, as leaving out one topic from the CLEF campaign changes the ranking of the retrieval systems despite the fact that 50 topics are considered to be sufficiently reliable [Voorhees and Buckley 2002; Zobel 1998].

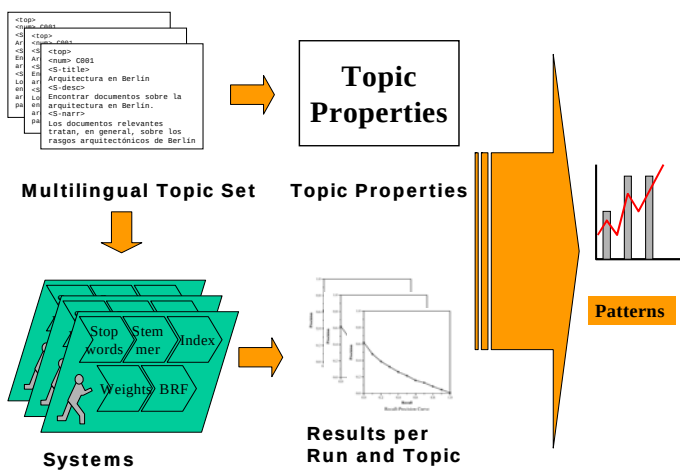

Figure 1. General overview of the research approach.

Most analysis of the data generated in CLEF is based on the average performance of the systems. This study concentrates on the retrieval quality of systems for individual topics. By identifying reasons for the failure of certain systems for some topics, these systems can be optimized. Our analysis identified a feature of the topics which can be exploited for future system improvement. In this study, we focused on the impact of named entities in topics and found a significant correlation with the average precision. Consequently, the goal of this study

is twofold:

(a) to measure the effect of named entities on retrieval performance in CLEF

(b) to optimize retrieval systems based on these results.

Named entities pose a potential challenge to cross language retrieval systems, because these systems often rely on machine translation of the query. The following problems may occur when trying to translate a named entity:

- The named entity may be out of vocabulary for translation
- Copying a named entity into the target language often does not help, as the name may be spelled differently (*e.g.* German: "Gorbatschow" vs. English: "Gorbachev")
- A named entity can actually be translated (*e.g.* "Smith" could be interpreted as a name or a profession and as the latter, translated)

Named entities are a feature which can be easily identified within queries. We consider the systems at CLEF as black boxes and have so far not undertaken any effort to analyze how these systems treat named entities and why that treatment may result in the effects we have observed. The data necessary for such an analysis is not provided by CLEF. The systems use very different approaches, tools and linguistic resources. Each system may treat the same named entity quite differently and successful retrieval may be due to a large number of factors like appropriate treatment as n-gram, proper translation by a translation service, or due to an entry in a linguistic resource. An analysis of the treatment of the named entities would lead merely to case studies. As a consequence, we find a statistical analysis of the overall effect as the appropriate research approach.

The remainder of this paper is organized as follows. The next chapter provides a brief overview of the research on evaluation results and their validity. Chapter three describes the data for CLEF used in our study. In chapter four, the influence of named entities on the overall retrieval results are analyzed. Chapter five explores the relationship between named entities and the performance of individual systems. In chapter six, we show how the performance variation of systems due to named entities could be exploited for system optimization.

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 concluded that the TREC (Text REtrieval Conference) experiments are reliable as far as the ranking of the systems is concerned [Zobel 1998]. Voorhees and Buckley have analyzed the reliability of experiments as a function of the size of the topic set [Voorhees and Buckley 2002]. They concluded that the typical size of the topic set of some 50 topics in TREC is sufficient for a satisfactory level of reliability.

Human judgments are necessary to evaluate the relevance of the documents. Relevance assessment is a very subjective task. Consequently, assessments by different jurors result in different sets of relevant documents. However, these different sets of relevant documents do not lead to different system rankings according to an empirical analysis [Voorhees 2000]. Thus, the subjectivity of the jurors does not call into question the validity of the evaluation results.

Further research is dedicated toward the question of whether expensive human relevance judgments are necessary or whether the constructed document pool of the most highly ranked documents from all runs may serve as a valid approximation of the human judgments. According to a study by Cahan *et al.*, the ranking of the systems in TREC correlates positively to a ranking based on the document pool without further human judgment [Cahan *et al*. 2001]. However, there are considerable differences in the ranking which are especially significant for the highest ranks.

Another important aspect in evaluation studies is pooling. Not all submitted runs can be judged manually by jurors and relevant documents may remain undiscovered. Therefore, a pool of documents is built to which the systems contribute differently. In order to measure the potential effect of pooling, a study was conducted which calculated the final rankings of the systems by leaving out one run at a time [Braschler 2003]. It shows that the effect is negligible and that the rankings remain stable.

However, our analysis shows that leaving out one topic during the result calculation changes the system ranking in most cases. It has also been noted that the differences between topics are larger than the differences between systems. This effect has been observed in TREC [Harman and Voorhees 1997] and also in CLEF [Gey 2001].

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 one topic (nr. 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 that both topics were from the sports domain (Topic 51 and 54). This effect has been noted in many evaluations and also in CLEF [Hollink *et al*. 2004]. As a consequence, topics are an important part of the design in an evaluation initiative and need to be created very carefully.

Named entities seem to play an important role especially in multilingual information retrieval [Gey 2001]. This assumption is backed by experimental results. The influence of named entities on the retrieval performance is considerable. In an experiment, the removal of named entities from the topic decreased the quality considerably, whereas the use of named entities only in the query led to a much smaller decrease [Demner-Fushman and Oard 2003].

A study for the CLEF campaign 2001 revealed no strong correlation between any single linguistic phenomenon and the system difficulty of a topic. Not even the length of a topic showed any substantial effect, except for named entities. However, the sum of all phenomena was correlated to the performance. The more linguistic phenomena available, the better the systems solved a topic on average [Mandl and Womser-Hacker 2003]. The availability of more variations of a word seems to provide stemming algorithms with more evidence for extraction of the stem, for example.

3. Named Entities in the Multi-lingual Topic Set

The data for this study stems from the Cross Language Evaluation Forum (CLEF) [Peters *et al*. 2003; Peters *et al*. 2004]. CLEF is a large evaluation initiative which is dedicated to cross-language retrieval for European languages. The setup is similar to the Text Retrieval Conference (TREC) [Harman and Voorhees 1997; Voorhees and Buckland 2002]. The main tasks for multilingual, ad-hoc retrieval are:

- The core and most important track is the **multilingual** task. The participants choose one topic language and need to retrieve documents in all main languages. The final result set needs to integrate documents from all languages ordered according to relevance regardless of their language.
- The **bilingual** task requires the retrieval of documents different from the chosen topic language.
- The **Monolingual** task represents the traditional ad-hoc task in information retrieval and is allowed for some languages.

All runs analyzed in this study are test runs based on topics for which no previous relevance judgments were known. For training runs, older topics can be used each year. Techniques and algorithms for cross-lingual and multilingual retrieval are described in the CLEF proceedings and are not the focus of this paper.

The topic language of a run is the language which the system developers use to start the search and to construct their queries. The topic language needs to be stated by the participants and can be found in the appendix of the CLEF proceedings. The retrieval performance of the runs for the topics can also be extracted from the appendix of the CLEF proceedings [Peters *et al.* 2003; Peters *et al*. 2004]. Most important, the average precision of each run for each topic can be retrieved.

3.1 Topic Creation Process

The topic creation for CLEF needs to assure that each topic is translated into all languages without modifying the content while providing equal chances for systems which start with

different topic languages. Therefore, a thorough translation check of all translated topics in CLEF was performed to check if the translations to all languages resulted in the same meaning. Nevertheless, the topic generation process follows a natural method and avoids artificial constructions [Womser-Hacker 2002].

Figure 2 shows an exemplary topic from CLEF containing a named entity. The topic's structure is built up by a short title, a description with a few words and a so-called narrative with one or more sentences. Participants of CLEF have to declare which parts are used for retrieval.

```
<top lang="ES"> <num>C083</num>
<ES-title> Subasta de objetos de Lennon. </ES-title>
<ES-desc> Encontrar subastas públicas de objetos de John Lennon.</ES-desc>
<ES-narr> Los documentos relevantes hablan de subastas que incluyen objetos que
pertenecieron a John Lennon, o que se atribuyen a John Lennon.</ES-narr>
</top>    <top> <num>C083</num>
<FR-title> Vente aux enchères de souvenirs de John Lennon </FR-title>
<FR-desc> Trouvez les ventes aux enchères publiques des souvenirs de John Lennon.
</FR-desc>
<FR-narr> Des documents pertinents décriront les ventes aux enchères qui incluent les objets
qui ont appartenu à John Lennon ou qui ont été attribués à John Lennon. </FR-narr> </top>
```

Figure 2. Example of a CLEF topic with a named entity

3.2 Data

An intellectual analysis of the results and the properties of the topics had identified named entities as a potential indicator of good retrieval performance. For that reason, named entities in the CLEF topic set were analyzed in more detail.

Named entities were intellectually assessed according a published schema [Sekine *et al.* 2002]. The analysis included all topics from the campaigns in the years 2000 through 2004. The number of named entities in each topic was assessed intellectually. We focused on English, Spanish, and German as topic languages and considered monolingual, bilingual, and multilingual tasks.

Table 1 shows the overall number of named entities found in the topic sets. The extraction was done intellectually by graduate students. We also assessed in which parts of the topic the name occurred, whether found in the title, the description, or the narrative. This detailed analysis was not exploited further because very few runs use a source other than title plus description. In very few cases, the topic narrative includes additional named entities not already present in the title and the description. For our analysis, the sum of named entities in all three parts was used. We analyzed the topic set in three languages, and in some cases, differences between the number of named entities between two versions of a topic occur.

These differences were considered. In 18 cases, a different number of named entities was assessed between German and English versions of topics 1 through 200, and in 49 cases, a difference was encountered between German and Spanish for topics 41 though 200. For example, topic 91 contains one named entity more for German because German has two potential abbreviations for United Nations (UN and UNO) and both are used.

The numbers given in Table 1 are based on the English versions of the topics and consider the number of types rather than tokens of named entities in title, description, and narrative together.

Table 1. Number 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
2000	40	52	1.14	1.12
2001	50	60	1.20	1.06
2002	50	86	1.72	1.54
2003	60	97	1.62	1.18
2004	50	72	1.44	1.30

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
2001	Bi	German	9	16	24	7
2001	Multi	German	5	16	24	7
2001	Bi	English	3	16	24	7
2001	Multi	English	17	17	26	7
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	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	Bi	English	8	9	41	10
2003	Multi	English	74	9	41	10
2004	Multi	English	34	16	23	11

The large number of named entities in the topic set shows their importance. Table 2 shows the number of runs within each task. For the analysis presented in chapter five, we divided the topics into three classes: (a) no named entities, (b) one or two named entities, and (c) three or more named entities. The distribution of topics over these three classes is also shown in Table 2. It can be seen that the three classes are best balanced in CLEF 2002, whereas topics in the second class dominate in CLEF 2003.

Only topics for which no zero results were returned were considered for each sub-task. Since these topics differ between sub-tasks, there are slight differences between the numbers for each class even for one year. For further analysis, only tasks with more than eight runs were considered.

4. Named Entities and General Retrieval Performance

Our first goal was to measure whether named entities had any influence on the overall quality of the retrieval results. In order to measure this effect, we first calculated the correlation between the overall retrieval quality achieved for a topic and the number of named entities encountered in this topic. In the second section, this analysis is refined to single tasks and specific topic languages.

4.1 Correlation Between Average Precision and Number of Named Entities

Table 3. Method a: Best run for each topic in relation to the number of named entities in the topic

Number of named entities	0	1	2	3	4	5
Number of Topics	42	43	40	20	9	4
Average of Best System per Topic	0.62	0.67	0.76	0.83	0.79	0.73
Minimum of Best System per Topic	0.09	0.12	0.04	0.28	0.48	0.40
Standard Deviation of Best System per Topic	0.24	0.24	0.24	0.18	0.19	0.29

Table 4. Method b: Average precision of runs in relation to the number of named entities in the topic

Number of named entities	0	1	2	3	4	5
Number of Topics	42	43	40	20	9	4
Minimum of Average Performance per Topic	0.02	0.04	0.01	0.10	0.17	0.20
Average of Average Performance per Topic	0.20	0.25	0.36	0.40	0.31	0.40
Maximum of Average Performance per Topic	0.54	0.61	0.78	0.76	0.58	0.60
Standard Deviation of Average Performance	0.14	0.15	0.18	0.17	0.14	0.19

First, we determined the overall performance in relation to the number of named entities in a topic. The 200 analyzed topics contain between zero and six named entities. For each number *n* of named entities, we determine the overall performance by two methods: (a) take the best run for each topic and (b) take the average of all runs for a topic. For both methods, we obtain a set of values for *n* named entities. Within each set, we can determine the maximum, the average, and the minimum. For example, we determine for method (a) the following values: best topic for *n* named entities, average of all topics for *n* named entities, and worst topic among all topics with *n* named entities. The last value gives the performance for the most difficult topic within the set of topics containing *n* named entities. The maximum of the best runs is in most cases 1.0 and is, therefore, omitted. The following Tables 3 and 4 show these values for CLEF overall. Figures 3 and 4 show detailed analysis for specific tasks.

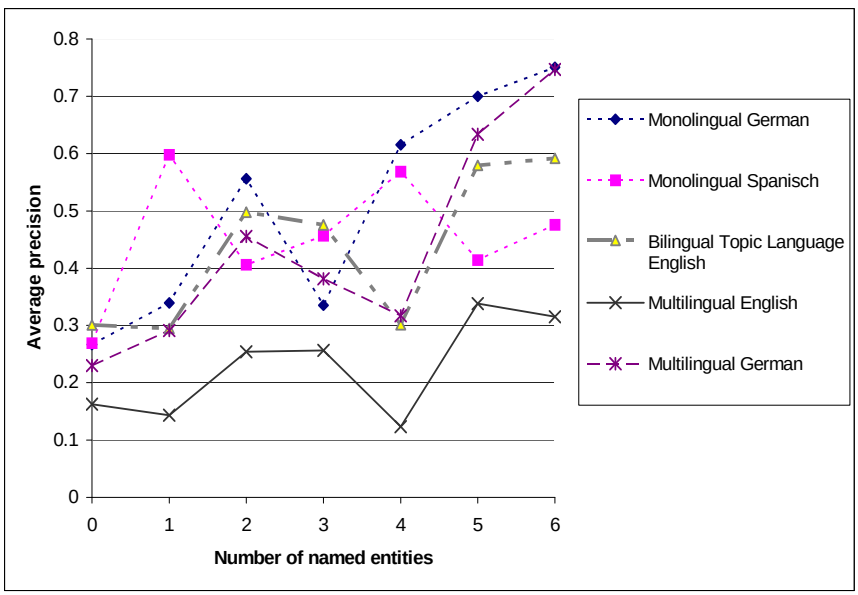

Figure 3. Method a: Average precision for topics with n named entities for CLEF 2002

The CLEF campaign contains relatively few topics with four or more named entities. The results for these values are, consequently, not significant.

It can be seen that topics with more named entities are generally solved better by the systems. This observation can be confirmed by statistical analysis. The average performance correlates to the number of named entities with a value of 0.43 and the best performance with a value of 0.26. Both correlation values are statistically significant at a level of 95%. With one exception, the worst performing category is always the one without any named entities.

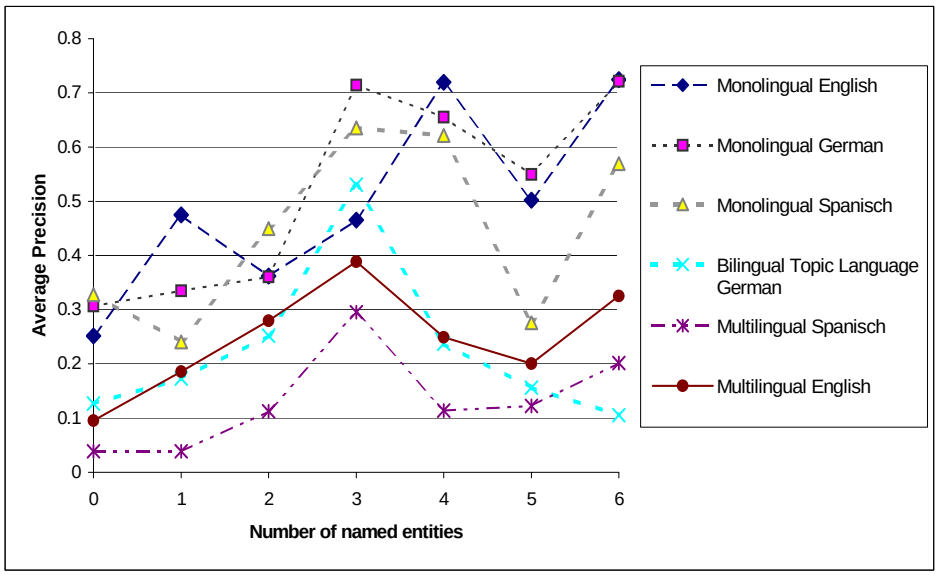

Figure 4. Method b: Relation between system performance and the number of named entities in CLEF 2002

4.2 Correlation for Individual Tasks and Topic Languages

The correlation analysis was also carried out for the individual retrieval tasks or tracks. This can be done by (a) calculating the average precision for each topic achieved within a task, by (b) taking the maximum performance for each topic (taking the maximum average precision that one run achieved for that topic), and by (c) calculating the correlation between named entities and average precision for each run individually and taking the average for all runs within a task. Both measures a and b are presented in Table 5. Except for one task (multilingual with topic language English in 2001), all observed correlations are positive. Thus, the overall effect occurs within most tasks and even within most single runs.

There is no difference in the average strength of the correlation for German (0.27) and English (0.28) as topic language. The average for each language in the last column shows a more significant difference. The correlation is stronger for German (0.19) than for English (0.15) as topic language. Furthermore, there is a considerable difference between the average correlation for the bilingual (0.35) and multilingual run types (0.22). This could be a hint that the observed positive effect of named entities on retrieval quality is smaller for multilingual retrieval.

Table 5. Correlation of system performance and number of named entities for different tasks

CLEF year	Run type	Topic language	Num-ber of runs	(a) Correlation of average precision per topic to number of NEs	Level of statistical significance (t-distribution) for prev. column	(b) Correlation of max. precision per topic to nr. of NEs
2001	Bilingual	German	9	0.44	-	0.32
2001	Multilingual	German	5	0.19	-	0.24
2001	Bilingual	English	3	0.20	-	0.13
2001	Multilingual	English	17	-0.34	-	-0.36
2002	Bilingual	German	4	0.33	-	0.25
2002	Multilingual	German	4	0.43	-	0.41
2002	Bilingual	English	51	0.40	99%	0.36
2002	Multilingual	English	32	0.29	-	0.37
2002	Monolingual	German	21	0.45	95%	0.34
2002	Monolingual	Spanish	28	0.21	-	0.27
2003	Bilingual	German	24	0.21	-	0.10
2003	Bilingual	English	8	0.41	-	0.47
2003	Multilingual	English	74	0.31	99%	0.27
2003	Monolingual	German	30	0.37	95%	0.28
2003	Monolingual	Spanish	38	0.39	99%	0.33
2003	Monolingual	English	11	0.16	-	0.24
2003	Multilingual	Spanish	10	0.21	-	0.31
2004	Multilingual	English	34	0.33	95%	0.34

It needs to be stressed, though, that the effect does not only occur for systems with overall poor performance. Rather, it can be observed in the top ranked runs as well. Figure 5 shows the strength of the correlation for all runs in one task. The runs are ordered according to their average precision. The correlation between the systems MAP for a topic and the number of named entities present in that topic is also shown in Figure 5.

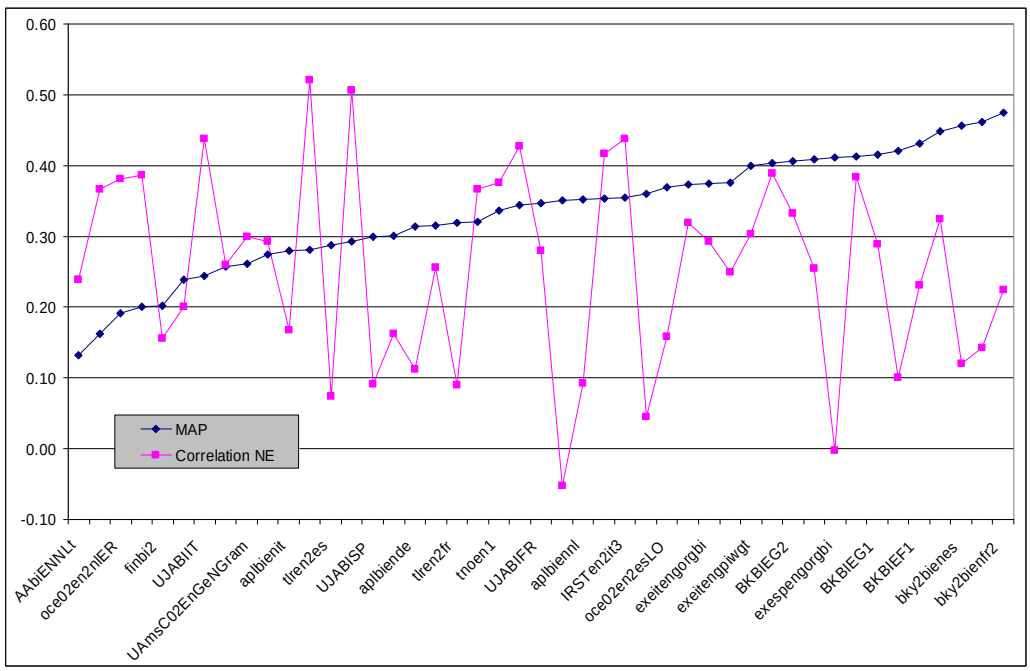

Figure 5. Correlation between named entities and performance for runs in CLEF 2002 (task bilingual, topic language English)

5. Conclusion Performance Variation of Systems for Named Entities

In this chapter, we show that the systems tested at CLEF perform differently 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. As already mentioned, we distinguished three classes of topics: (a) the first class without proper names (called “none”), (b) the second class with one or two named entities (called “few”), and (c) a third class with three or more named entities (called “lots”). This approach is suitable for implementation and allows the categorization before the experiments and the relevance assessment. It requires no intellectual intervention but, solely, a named entity recognition system.

5.1 Variation of System Performance

As we can see in Table 2, the three categories are well balanced for the CLEF campaign in 2002. For 2003, there are only few topics in the first and second categories. Therefore, the average ranking is extremely similar to the ranking for the second class “few”.

Figure 5 shows that the correlation between average precision and the number of named entities is quite different for all runs for one exemplary task. The runs in Figure 6 are ordered

according to the original ranking in the task. We observe a slightly decreasing sensitivity for named entities with higher system performance. However, the correlation is still substantial and sometimes still high for top runs.

A look at the individual runs shows large differences between the three categories. We show the values for three tasks in Figure 6. The curve for many named entities lies mostly above the average curve, whereas the average precision for the class none without named entities in most cases remains below the overall average. Sometimes, even the best runs perform quite differently for the three categories. Other runs perform similarly for all three categories.

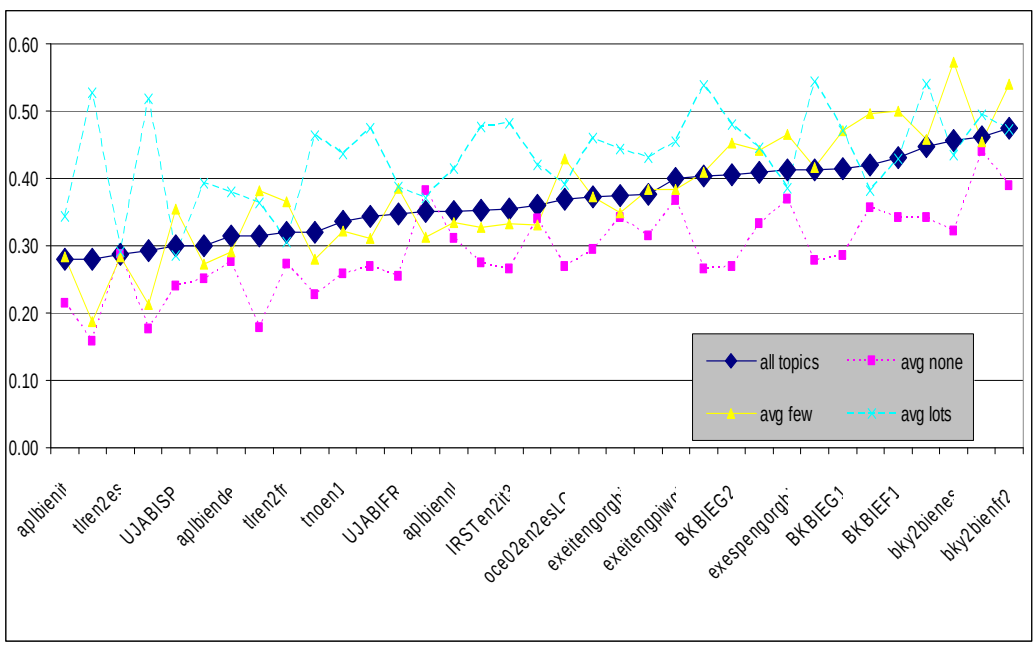

Figure 6. Performance variation of runs in CLEF 2002 (task bilingual, topic language English) depending on number of named entities in topic

5.2 Correlation of System Rankings

The performance variation within the classes leads to different system rankings for the classes. An evaluation campaign including, for example, only topics without named entities may lead to different rankings. To analyze this effect, we determined the rankings for all runs within each named entity class, *none*, *few*, and *lots*. Table 6 shows that the system rankings can be quite different for the three classes. The difference is measured with the Pearson rank correlation coefficient.

For most tracks, the original average system ranking is most similar to the ranking based only on the topics with one or two named entities. For the first and second categories, the rankings are more dissimilar. The ranking for the top ten systems in the classes usually differs more from the original ranking. This is due to minor performance differences between top runs.

Table 6. Correlation of full system ranking to ranking based on topic sub-set

Sub-Task				Topic sub-set		
CLEF year	Run type	Topic language	Number of runs	No NEs	few NEs	lots NEs
2001	Bilingual	German	9	0.92	0.93	0.92
2001	Multilingual	English	17	0.98	0.93	0.75
2002	Bilingual	English	51	0.88	0.93	0.74
2002	Multilingual	English	32	0.94	0.99	0.98
2003	Bilingual	German	24	0.81	0.99	0.91
2002	Multilingual	English	74	0.86	1.00	0.93

These findings are not always statistically significant because each category contains only few topics. As stated by Buckley and Voorhees, some 50 topics are necessary to create a reliable ranking [Buckley and Voorhees 2002].

6. Optimization by Fusion Based on Named Entities

The patterns of the systems are strikingly different for the three classes. As a consequence, there seems to be potential for the combination or fusion of systems.

We propose the following simple fusion rule. For each topic, the number of named entities is determined. Subsequently, this topic is channeled into 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 for that number of named entities. By simply choosing the best performing system for each topic, we can also determine a practical upper level for the performance of the retrieval systems. This upper level can give a hint about how much of the potential for improvement is exploited by an approach. Table 6 shows the optimal performance and 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 the year 2002, we observe the highest improvement of 10% for the bilingual runs. For this task, there is also the highest figure for potential, 53%. Figure 7 shows the results of the optimization.

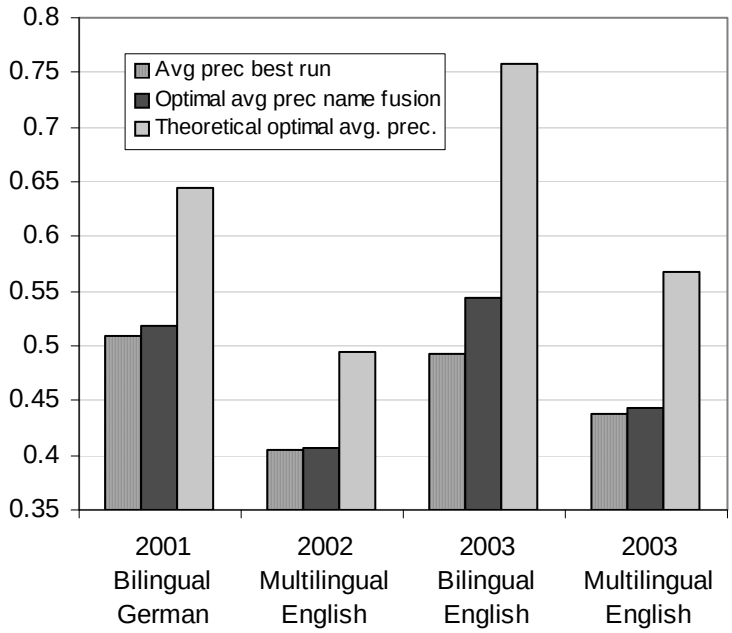

Figure 7. Optimization potential of named entity based fusion

Table 7. Improvement by fusion based on named entities for several tasks

CLEF year	Run type	Topic language	Average precision best run	Optimal average precision name fusion	Improve-ment over best run	Practical optimal average precision.	Improve-ment over best run
2001	Bilingual	German	0.509	**0.518**	**2%**	0.645	27%
2001	Multilingual	English	0.405	**0.406**	**0%**	0.495	22%
2002	Bilingual	English	0.4935	**0.543**	**10%**	0.758	53%
2002	Multilingual	English	0.378	**0.403**	**6.5%**	0.456	21%
2003	Bilingual	German	0.460	**0.460**	**0%**	0.622	35%
2003	Bilingual	English	0.348	**0.369**	**6.1%**	0.447	28%
2003	Multilingual	English	0.438	**0.443**	**1.2%**	0.568	30%

The previous analysis showed that our fusion approach has the potential to boost even top runs. Consequently, this technique may also be beneficial for lower-ranked runs. We applied the optimization through fusion for all runs. In the ordering of all runs according to the average precision (original CLEF ranking), we chose a window of three and five neighboring runs. From these three to five runs, we chose the best results for each of the three classes of number of proper names (none, few, or lots). Again, the best run for each class is chosen and

contributes to the fusion result. Table 6 shows the average improvement for this fusion technique. This analysis shows that the performance of retrieval systems can be optimized by channeling topics to the systems best appropriated for topics with none, one or two and three and more proper names. Certainly, the application of this fusion on the past results approach is artificial and, 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.

7. Named Entities in Topics and Retrieval Performance for Target Languages

So far, our studies have been focused on 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. However, the monolingual tasks have already been analyzed and are also considered here. The additional analysis is targeted at bilingual retrieval tasks. We grouped all bilingual runs with English, German, and Spanish as document languages. 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. Table 8 shows the results of this analysis.

Table 8. 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. These 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.

8. Resume

Research on failure and success stories for individual topics is a promising strategy for the analysis of information retrieval results. Several current research initiatives are focusing on this strategy and are looking at retrieval results beyond average precision [Harman 2004; SIGIR 2005 query difficulty workshop]. We identified named entities in topics as one transparent predictor in multi- and mono-lingual retrieval. Further analysis on named entities should also take the frequency and distribution of the named entities in the corpora into account.

References

Allan, J., and H. Raghavan, "Using part-of-speech Patterns to Reduce Query Ambiguity," In *Proceedings of the Annual International ACM Conference on Research and Development in Information Retrieval (SIGIR ´02)*, Tampere, Finland, Aug. 11-15, 2002, pp. 307-314.

Braschler, M., "CLEF 2002 - Overview of Results," *Evaluation of Cross-Language Information Retrieval Systems. Third Workshop of the Cross Language Evaluation Forum CLEF 2003)*, Trondheim. Springer (Lecture Notes in Computer Science).

Braschler, M., and C. Peters, "Cross-Language Evaluation Forum: Objectives, Results, Achievements," *Information Retrieval*, 2004, 7, pp. 7-31.

Cahan, P., C. Nicholas, and I. Soboroff, "Ranking Retrieval Systems without Relevance Judgments," In *Proceedings of the Annual International ACM Conference on Research and Development in Information Retrieval (SIGIR '01)*, New Orleans, USA, Sep. 9-13, 2001, pp. 66-73.

Cronen-Townsend, S., Y. Zhou, and B. Croft, "Predicting Query Ambiguity," In *Proceedings of the Annual Intl. ACM Conference on Research and Development in Information Retrieval (SIGIR '02)*, Tampere, Finland, 2002, pp. 299-306.

Demner-Fushman, D., and D. Oard, "The effect of bilingual term list size on dictionary-based crosslanguage information retrieval," In *Thirty-Sixth Hawaii International Conference on System Sciences*, (Hawaii, Jan 6-9, 2003).

Di Nunzio, G., N. Ferro, T. Mandl, and C. Peters, "CLEF 2006: Ad Hoc Track Overview," In *Evaluation of Multilingual and Multi-modal Information Retrieval. 7th Workshop of the Cross-Language Evaluation Forum, (CLEF 2006)*, Alicante, Spain, Revised Selected Papers. Berlin et al.: Springer (Lecture Notes in Computer Science 4730) 2007, pp. 21-34.

Diaz, F., and R. Jones, "Using Temporal Profiles of Queries for Precision Prediction," In *Proceedings of the Annual International ACM Conference on Research and Development in Information Retrieval (SIGIR '04)*, Sheffield, UK, 2004, pp. 18-24.

Downie, S., "Toward the Scientific Evaluation of Music Information Retrieval Systems," In *International Symposium on Music Information Retrieval (ISMIR)*, Washington, D.C., and Baltimore, USA 2003, http://ismir2003.ismir.net/papers/Downie.PDF.

Evans, D., J. Shanahan, and V. Sheftel, "Topic Structure Modeling," In *Proceedings of the Annual Intl. ACM Conference on Research and Development in Information Retrieval (SIGIR '02)*, Tampere, Finland, 2002, pp. 417-418.

Fuhr, N., *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.

Gey, F., "Research to improve Cross-Language Retrieval. Position Paper for CLEF," In *Cross-Language Information Retrieval and Evaluation. Workshop of Cross-Language Evaluation Forum (CLEF 2000)*, Lisbon, Portugal, September 21-22, 2000). Berlin et al.: Springer [LNCS 2069] 2001, pp. 83-88.

Hackl, R., R. Kölle, T. Mandl, A. Ploedt, J.-H. Scheufen, and C. Womser-Hacker, "Multilingual Retrieval Experiments with MIMOR at the University of Hildesheim," In *Evaluation of Cross-Language Information Retrieval Systems. Proceedings CLEF 2003 Workshop*, Trondheim, Norway, Revised Selected Papers. Berlin et al.: Springer [LNCS 3237] 2004, pp. 166-173.

Harman, D., "SIGIR 2004 Workshop. RIA and Where can we go from here?," *ACM SIGIR Forum*, 38(2), pp. 45-49.

Harman, D., and E. Voorhees, "Overview of the Sixth Text REtrieval Conference," In *The Sixth Text REtrieval Conference (TREC-6)*, National Institute of Standards and Technology, Gaithersburg, Maryland, 1997, http://trec.nist.gov/pubs/.

Hollink, V., J. Kamps, C. Monz, and M. de Rijke, "Monolingual Document Retrieval for European Languages," *Information Retrieval*, 7(1-2), pp. 33-52.

Kluck, M., and C. Womser-Hacker, "Inside the Evaluation Process of the Cross-Language Evaluation Forum (CLEF): Issues of Multilingual Topic Creation and Multilingual Relevance Assessment," In *Proceedings of the Third International Conference on Language Resources and Evaluation (LREC 2002)*, Las Palmas de Gran Canaria, Spain, May 29-31, 2002, ELRA, Paris, 2002, pp. 573-576.

Lempel, R., and S. Moran, "Predictive Caching and Prefetching of Query Results in Search Engines," in *Proceedings of the Twelfth International World Wide Web Conference (WWW 2003)*, Budapest, Hungary, May 20-24, 2003. pp. 19-28.

Mandl, T., and C. Womser-Hacker, "Linguistic and Statistical Analysis of the CLEF Topics," In *Advances in Cross- Language Information Retrieval: Third Workshop of the Cross-Language Evaluation Forum (CLEF 2002)*, Rome, Italy, September 19-20, 2002 Springer, LNCS 2785. 2003 pp. 505-511.

Mandl, T., and C. Womser-Hacker, "A Framework for long-term Learning of Topical User Preferences in Information Retrieval," *New Library World*, 105(5/6), pp. 184-195.

Oyama, K., E. Ishida, and N. Kando, *NTCIR Workshop 3: Proceedings of the Third NTCIR Workshop on research in Information Retrieval, Automatic Text Summarization and Question Answering*, Tokio, 2003.
http://research.nii.ac.jp/ntcir/workshop/OnlineProceedings3/index.html

Peters, C., M. Braschler, J. Gonzalo, and M. Kluck, "Evaluation of Cross-Language Information Retrieval Systems," In *Third Workshop of the Cross Language Evaluation Forum (CLEF 2002)*, Rome. Berlin et al.: Springer (Lecture Notes in Computer Science 2785) 2003.

Peters, C., J. Gonzalo, M. Braschler, and M. Kluck, "Comparative Evaluation of Multilingual Information Access Systems," In *4th Workshop of the Cross-Language Evaluation Forum (CLEF 2003)*, Trondheim, Norway, August 21-22, 2003, Revised Selected Papers. Springer Lecture Notes in Computer Science 3237, 2004.

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

Sekine, S., K. Sudo, and C. Nobata, "Extended Named Entity Hierarchy," In: *Proceedings of Third International Conference on Language Resources and Evaluation (LREC 2002)*, Las Palmas, Canary Islands, Spain.

Soboroff, I., C. Nicholas, and P. Cahan, "Ranking retrieval systems without relevance judgements," In *Proceedings of the 24th Annual International ACM SIGIR Conference of Research and Development in Information Retrieval*, New Orleans, pp. 66-73.

Sparck, J.K., "Reflections on TREC," *Information Processing and Management*, 31(3), pp. 291-314.

Voorhees, E., and C. Buckley, "The Effect of Topic Set Size on Retrieval Experiment Error," In *Proceedings of the Annual International ACM Conference on Research and Development in Information Retrieval (SIGIR '02)*, Tampere, Finland, Aug. 11-15, 2002, ACM Press, pp. 316-323.

Voorhees, E., "Variations in Relevance Judgements and the Measurement of Retrieval Effectiveness," *Information Processing and Management*, 36(5), 2000, pp. 679-716.

Voorhees, E., and L. Buckland, *The Eleventh Text Retrieval Conference (TREC 2002)*, National Institute of Standards and Technology, Gaithersburg, Maryland. Nov. 2002.
http://trec.nist.gov/pubs/trec11/t11_proceedings.html.

Womser-Hacker, C., "Multilingual Topic Generation within the CLEF 2001 Experiments," In *Evaluation of Cross-Language Information Retrieval Systems: Second Workshop of the Cross-Language Evaluation Forum (CLEF 2001)*, Peters C.; Braschler M.; Gonzalo, J.

and Kluck, Michael (Eds.). 2002, Darmstadt, Germany, September 3-4, 2001. Springer, LNCS 2406, pp. 389-393.

Zobel, J., "How Reliable are the Results of Large-Scale Information Retrieval Experiments?" In *Proceedings of the 21st Annual International ACM Conference on Research and Development in Information Retrieval (SIGIR '98)*, Melbourne, Australia, 1998, pp. 307-314.