Domain-Specific Track CLEF 2006: Overview of Results and Approaches, Remarks on the Assessment Analysis

Maximilian Stempfhuber and Stefan Baerisch

Informationszentrum Sozialwissenschaften (IZ), Lennéstrasse 30, 53113 Bonn, Germany
{ stempfhuber, baerisch}@iz-soz.de

Abstract

The CLEF domain-specific track uses databases in different languages from the social science domain as the basis for retrieval of documents relevant to a user's query. Predefined topics simulate the user's information needs and are used by the research groups participating in this track to generate actual queries. The documents are structured into individual elements so that they can be used for optimizing the search strategies. One type of the retrieval task was on cross-language retrieval, the finding of information using queries in a different language than the actual language of the documents. English, German and Russian were used as languages for queries and documents. Queries therefore had to be translated from one of the languages to one or more of the other languages. Besides the cross-language tasks also monolingual retrieval tasks were carried out were query and document are in the same language. The focus here was to map the query onto the language and internal structure of the documents.

This paper gives an overview of the domain-specific track and reports on noteworthy trends in approaches and results as well as on the topic creation and assessment process.

Categories and Subject Descriptors

H.3 [Information Storage and Retrieval]: H.3.1 Content Analysis and Indexing; H.3.3 Information Search and Retrieval

General Terms

Measurement, Performance, Experimentation

Keywords

Information Retrieval, Evaluation

1 Introduction

The CLEF domain-specific task is focused on cross-language retrieval in English, German and Russian. Queries as well as documents are taken from the social science domain and are provided in all three languages. This section presents the domain-specific task with the focus on the corpora used. It gives an overview on the 2006 track. The following sections are on the methods used and results achieved in the retrieval and on the processes of topic-creation and assessment. The paper concludes with an outlook on the future development of the domain-specific track.

2 The CLEF Domain-specific track

The domain-specific track can divided into 3 subtasks depending on the language of the query and the documents.

- In the monolingual subtask queries and documents are in the same language, the Russian subtask uses the INION Collection while the English und German subtask work on the GIRT corpus
- In the bilingual subtasks queries in one language are used with a collection in a different language. The domain-specific track includes 6 bilingual subtasks for all combinations of English, German, and Russian.
- The multilingual task uses queries in one language against the combination of the content of all collection.

Queries in the CLEF are provided as topics. Each topic is the representation of an information need form the social science domain. Topics as well as documents are structured with the topic including a title, a short description of the information need and a longer narrative for further clarification. The documents, among other information, include fields such as author, title and year of publication, terms from a thesaurus and an abstract:

Title	Childlessness in Germany
Description	Find information on the factors for childlessness in Germany.
Narrative	All documents examining causes and consequences of childlessness among couples in Germany are of interest. This also includes documents covering the topic's macroeconomic aspects.

The participating groups submits one or more runs for each subtask. A run includes the 1000 highest ranked document for each of the 25 topics. The runs from all groups for a given subtask are pooled, the highest ranking document are then assessed by domain experts in order to obtain a reference on the relevance of the documents.

2.1 The GIRT4 and INION Collections

GIRT, the German Indexing and Retrieval Testdatabase is a pseudo-parallel corpus with 302.638 documents in English and German. The documents cover primarily the domain of the German Social Sciences, in particular scientific publications. Details about the Girt corpus, especially about the development of the current fourth version, can be found in [1,2]. The INION corpus covers the Russian social sciences and economics. The collection includes 145.802 documents and replaces the Russian Social Science Corpus used in 2005.

2.2 Participants and submitted Runs

The University of California in Berkeley, the University of Hagen in Germany, the University of Technology in Chemnitz and the University of Neuchatel took part in the 2006 domain specific track, submitting 36 runs in total: 12 bilingual runs and 2 multilingual runs. For detailed figures see the following table:

Sub-task	# Participants	# Runs	Topic Language
Multi-lingual	1	2	DE:1 EN:1
Bilingual X → DE	2	6	EN:6
Bilingual X → EN	1	3	DE:2 RU:2
Bilingual X → RU	1	3	EN:2 DE:1
Monolingual DE	4	13	
Monolingual EN	3	8	
Monolingual RU	1	1	
Total	4	36	

Table 1: Submitted runs

Overview of the 2006 Domain Specific Track

2.3 Methods and Results Overview

Details on the results and employed methods of all groups are given in the corresponding chapters of this volume. This chapter gives only a brief overview of the methods employed. The University of Hagen used a re-ranking approach based on antagonistic terms, improving the mean average precision. The Chemnitz Technical University used Apache Lucene as the basis for the retrieval and achieved the best results with a combination of suffix stripping, stemming, and decompounding. The University of California, Berkeley, did not use methods for thesaurus-based query expansion and de-compounding as they were employed before. The University of Neuchatel used DFR GL2 and Okapi with several combinations of the fields included in a topic against the different fields of GIRT4 documents.

2.4 Topics

The team responsible for the topic-creation process changed in 2006. This has lead to changes in the topic-creation process itself as well. In order to produce a large number of potential information needs covering a wide area of the social sciences, a number of domain experts familiar with the GIRT corpus were asked for contributions to the topics. 25 topics were selected from 42 submissions. These topics were compared to the topics from the years 2001-2005 in order to remove topics too similar to topics already used. Removed topics were replaced from the pool of previously unused topics. The definition of topics by domain experts with knowledge about the corpus resulted in many topics of interest but also caused more effort in the selection of the actual topic-set. This was due to the fact that not all submitters of topics had detailed knowledge of the CLEF topic-creation rules and corrections had to be made afterwards to provide consistent structure and wording. For future CLEF-campaigns, this process will be further optimized.

The definition of topics suitable for the GIRT corpus as well as the Russian INION corpus remains challenging. Lack of Russian language skills in the topic-creation team increases the difficulty of creating a set of topics well suited for both corpora. An additional factor in 2006 was lack of experience with the INION corpus. Also in this regard, we hop to improve the topic cretation process in the next campaign as the multilingual corpus is now stable again.

2.5 Assessment Process

As in 2005 the assessment for English and German were done using a Java Swing program while our partner from the Research Computing Center of the M.V.Lomonosov Moscow State University used the CLEF assessment tool for the assessment of the Russian documents. The decision to use this software for English and German was made based on the fact that the assessments of English and German documents from the pseudo-parallel corpus should be compared to each other. An interesting finding was the statement of both assessors that they liked the ability to choose between keyboard-shortcuts and a mouse driven interface since it allowed for different working styles in the otherwise monotonous assessment task. Figures 1a and 1b show the number of documents in the pool for each topic and the percentage of relevant documents. It should be noted that, while the number of documents found for each topic is relatively stable, the percentage of relevant documents shows outliers with the topics number 167 and 169. Investigation on this outliers and discussion with the assessors showed that the outliers were caused by complex exclusion criteria in case of topic 167 and the use of the complex concepts of gender specific education and the specifics of the German primary education system in case of topic 169.

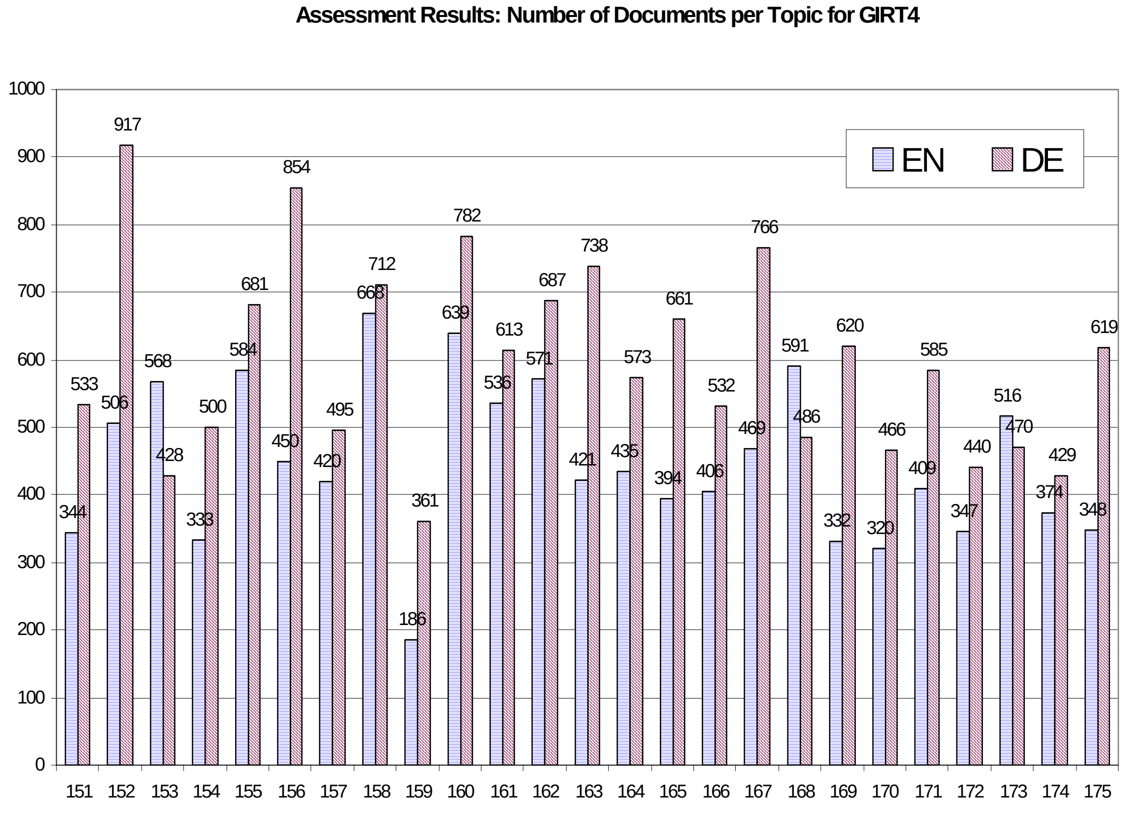

Figure 1a: Number of documents per topic

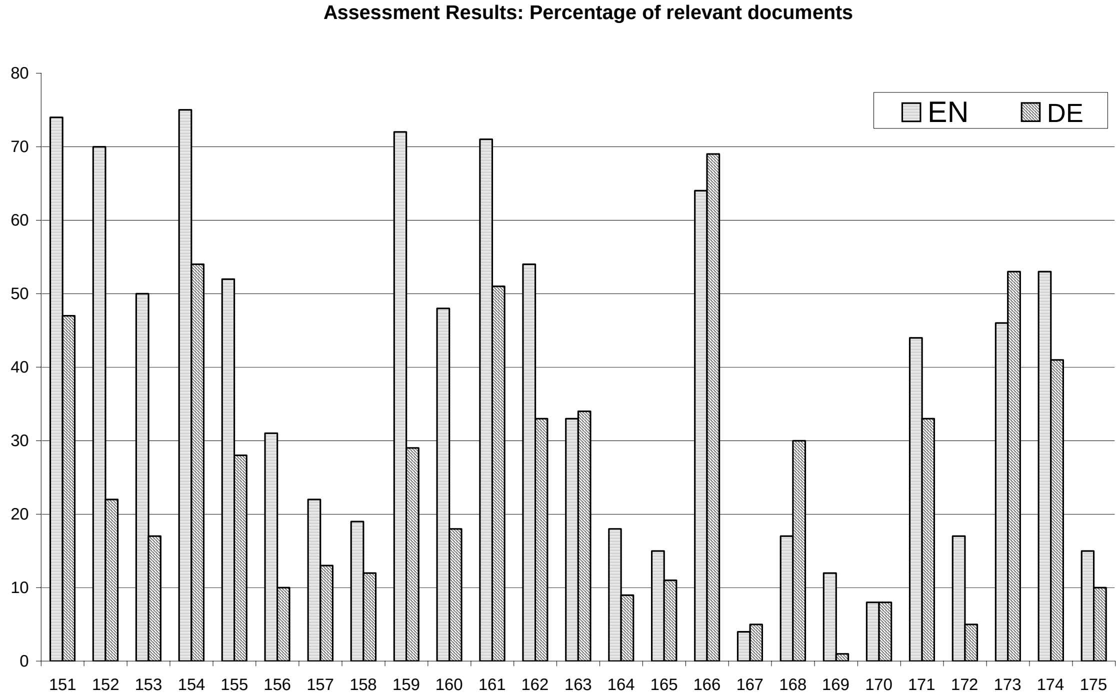

Figure 1b: Percentage of relevant documents per topic

The reassessment of ”pair” documents present in the English and the German pools of the parallel corpus - first introduced in the 2005 campaign – was repeated in 2006. When the two document of a pair had different relevance judgement after the first round of the assessment, a second round of assessment was done after an discussion between the assessors on their individual criteria for the assessment.

In addition to the reassessment of pairs a second round of assessment on the topic-level was conducted in 2006. This reassessment focused on the topics with the most significant difference in the percentage of the relevant documents between German and English.

3 Results

The detailed information on recall and precision for the participants and their runs can be found in the groups chapters in this volume [2,3,4,5] and shall not be repeated here in full. The Recall-Precision graph for the tracks which attracted the most participants can be found in the figures 2, 3 and 4.

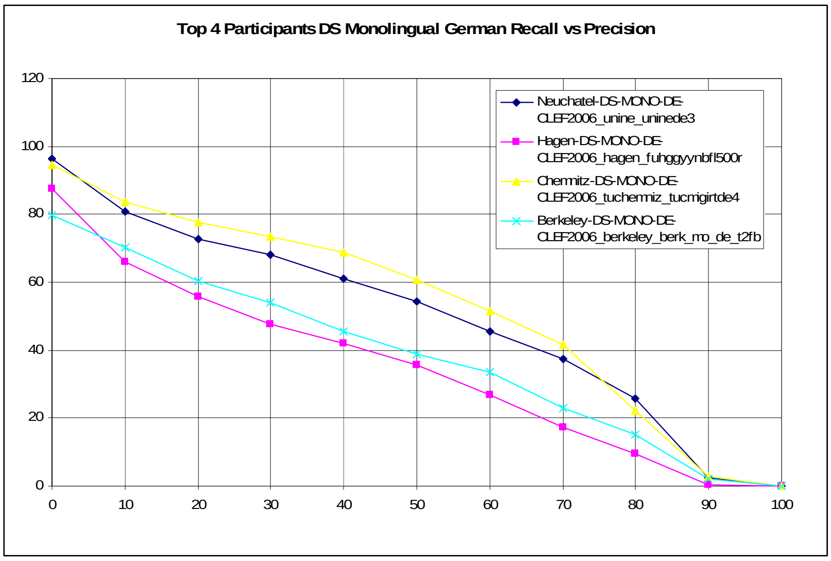

Figure 2: DS Monolingual Task German

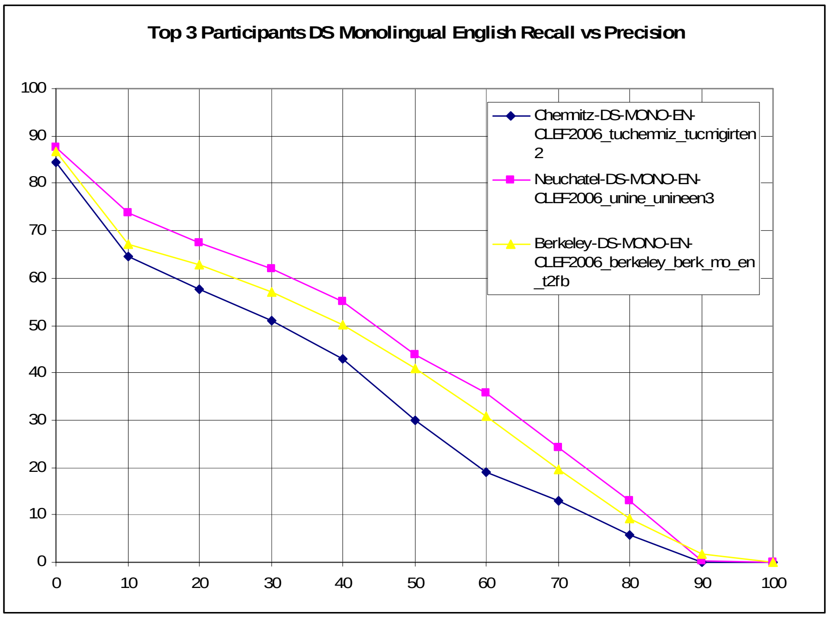

Figure 3: DS Monolingual Task English

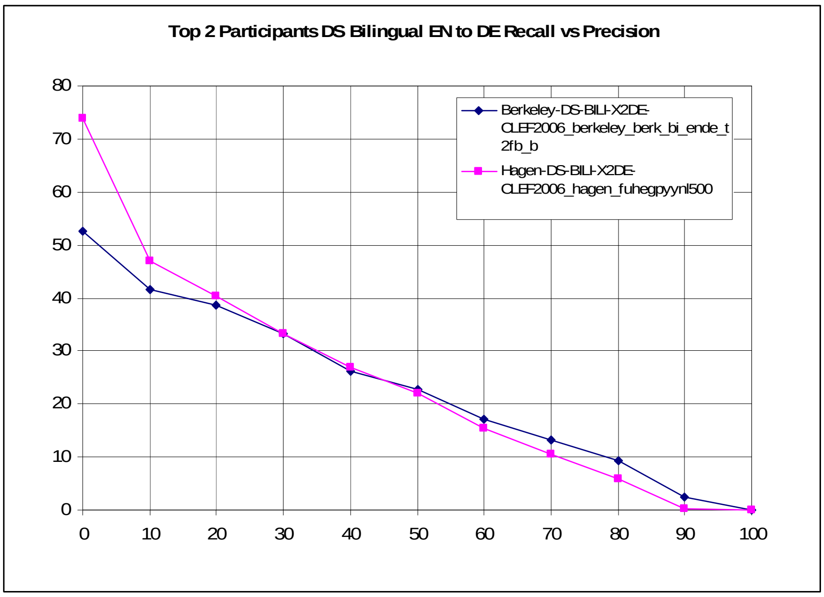

Figure 4: DS Bilingual Task English to German

4 Outlook

Several methods are being investigated to improve the quality of the topic-creation and assessment process. In Order to create topics covering a broader scope of the social sciences as well as realistic information needs it is planned to involve more domain experts into the topic creation process. These experts will be asked for topic proposals in the form of narratives. The most suitable topics from the resulting pool will then be selected and produced as topics conforming to the topic creation guidelines. In order to better verify the topics against the INION and GIRT4 corpora, a new retrieval system will be used which will contain both corpora (deployment planned for autumn of 2006) and will allow also our Russian partner from the Research Computing Center of the M.V.Lomonosov Moscow State University to better participate in the topic creation process.

Concerning the assessment process several methods are evaluated in order to allow the assessors to better coordinate their assessment criteria. Proposed methods include breaking the assessment und reassessment processes into smaller batches of 5 topics or the notification of the assessors in case of notable differences in the results for a topics.

Acknowledgements

We greatly acknowledge the partial funding by the European Commission through the DELOS Network of Excellence on Digital Libraries. Natalia Lookachevitch from the Research Computing Center of the M.V.Lomonosov Moscow State University helped in acquiring the Russion corpus and organized the translation of topics into Russian and the assessment process for the Russian corpus. Claudia Henning and Jeof Spiro did the German and English assessments.

References

1. Kluck, M.: The GIRT Data in the Evaluation of CLIR Systems – from 1997 until 2003. In: Peters, C., Gonzalo, J., Braschler, M., Kluck, M. (eds.): Comparative Evaluation of Multilingual Information Access Systems. 4th Workshop of the Cross-Language Evaluation Forum, CLEF 2003, Trondheim, Norway, August 21-22, 2003, Revised Selected Papers. Lecture Notes in Computer Science, Vol. 3237. Springer-Verlag Berlin Heidelberg New York, 379-393, (2004)

2. Kluck, M.: The Domain-Specific track in CLEF 2004: Overview of the Results and Remarks on the Assessment Process. In: Peters, C., Clough, P., Gonzalo, J., Jones, G.J.F., Kluck, M., Magnini, B. (eds.): Multilingual Information Access for Text, Speech and Images. 5th Workshop of the Cross-Language Evaluation Forum, CLEF 2004, Bath, UK, September 2004, Revised Selected Papers. Lecture Notes in Computer Science, Vol. 3491. Springer-Verlag Berlin Heidelberg New York, 260-270 (2005)

3. Jens Kürsten, Maximilian Eibl: Monolingual Retrieval Experiments with a Domain Specific Document Corpus at the Chemnitz Technical University

4. Ray R. Larson: Domain Specic Retrieval: Back to Basics (this volume) (2006)

5. Johannes Leveling: University of Hagen at CLEF2006: Reranking documents for the domain-specific task (this volume) (2006)

6. Jacques Savoy, Samir Abdou: UniNE at CLEF 2006: Experiments with Monolingual, Bilingual, Domain-Specific and Robust Retrieval (this volume) (2006)