The Domain-Specific Track at CLEF 2008

Vivien Petras, Stefan Baerisch
GESIS Social Science Information Centre, Lennéstr. 30, 53113 Bonn, Germany
{vivien.petras | stefan.baerisch@gesis.org}

Abstract

The domain-specific track evaluates retrieval models for structured scientific bibliographic collections in English, German and Russian. Documents contain textual elements (title, abstracts) as well as subject keywords from controlled vocabularies, which can be used in query expansion and bilingual translation. Mappings between the different controlled vocabularies are provided. This year, new Russian language resources were provided, among them Russian-English and Russian-German terminology lists as well as a mapping table between the Russian and German controlled vocabularies. Six participants experimented with different retrieval systems and query expansion schemes. Compared to previous years, the queries were more discriminating, so that fewer relevant documents were found per query.

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, Controlled Vocabularies

1 Introduction

The domain-specific track has been running continuously since the inception of CLEF in 2000 (Kluck & Gey, 2001; Kluck, 2004). The collections, topics and assessments efforts are provided by the GESIS Social Science Information Centre in Bonn, Germany in cooperation with its partners INION (Russia), Cambridge Scientific Abstracts (USA) and the University of Padova (Italy) as the developers and operators of the DIRECT system.

The track focuses on mono- and cross-language information retrieval in structured social science bibliographic data collections. The focus is the leveraging of controlled vocabularies and other structured metadata entities to improve information retrieval and translation.

The participants are provided with four collections for retrieval (1 German, 2 English, and 1 Russian) as well as a number of supplemental mapping and terminology tables for the controlled vocabularies. Each year, 25 new topics are created in German and then translated into English and Russian.

2 The Domain-Specific Task

The domain-specific track includes three subtasks:

- *Monolingual retrieval* against the German GIRT collection, the English GIRT and CSA Sociological Abstract collections, or the Russian INION ISISS collection;
- *Bilingual retrieval* from any of the source languages to any of the target languages;
- *Multilingual retrieval* from any source language to all collections / languages.

2.1 The Test Collections

The GIRT databases (currently in version 4) contain extracts from the German Social Science Information Centre's SOLIS (Social Science Literature) and SOFIS (Social Science Research Projects) databases from 1990-2000. The INION ISISS corpus covers social sciences and economics in Russian. The second English collection is an extract from CSA's Sociological abstracts.

German
The German GIRT collection (the social science **G**erman **I**ndexing and **R**etrieval **T**estdatabase) contains with 151,319 documents covering the years 1990-2000 using the German version of the Thesaurus for the Social Sciences (GIRT-description, 2007). Almost all documents contain an abstract (145,941).

English
The English GIRT collection is a pseudo-parallel corpus to the German GIRT collection, providing translated versions of the German documents. It also contains 151,319 documents using the English version of the Thesaurus for the Social Sciences but only 17% (26,058) documents contain an abstract.

The Sociological Abstracts database from Cambridge Scientific Abstracts (CSA) holds 20,000 documents, 94% of which contain an abstract. The documents were taken from the SA database covering the years 1994, 1995, and 1996. Additional to title and abstract, each document contains subject-describing keywords from the CSA Thesaurus of Sociological Indexing Terms and classification codes from the Sociological Abstracts classification.

Russian
For the retrieval of Russian collections, the INION corpus ISISS with bibliographic data from the social sciences and economics with 145,802 documents was once again used. ISISS documents contain authors, titles, abstracts (for 27% of the test collection or 39,404 documents) and keywords from the Inion Thesaurus.

2.2 Controlled Vocabularies

The GIRT collections have descriptors from the GESIS Thesaurus for the Social Sciences in German and English depending on the collection language. The CSA Sociological Abstracts documents contain descriptors from the CSA Thesaurus of Sociological Indexing Terms and the Russian ISISS documents are provided with Russian INION Thesaurus terms. GIRT documents also contain classification codes from the GESIS classification and CSA SA documents from the Sociological Abstracts classification. Table 1 shows the distribution of subject-describing terms per document in each collection.

Collection	*GIRT-4 (German or English)*	*CSA Sociological Abstracts*	*INION ISISS*
Thesaurus descriptors / document	10	6.4	3.9
Classification codes / document	2	1.3	n/a

Table 1. Distribution of subject-describing terms per collection

Vocabulary mappings
Vocabulary mappings are one-directional, intellectually created term transformations between two controlled vocabularies. They can be used to switch from the subject metadata terms of one knowledge system to the other, enabling a retrieval system to treat the subject descriptions of two or more different collections as one and the same.

For the English and German collections, mappings between the GESIS Thesaurus for the Social Sciences and the English CSA Thesaurus of Sociological Indexing Terms are provided. The mapping from the English Thesaurus for the Social Sciences to the English CSA Thesaurus of Sociological Indexing Terms is supplied for monolingual retrieval. Additionally, there is also a translation table with the German and English terms from the GESIS Thesaurus for the Social Sciences.

Three new Russian resources were developed in 2008: two translation tables as well as a mapping. One translation table contains translation between the German and Russian terms from the GESIS Thesaurus for the Social Sciences), which can also be used in conjunction with the German-English translation table. The second translation table lists Russian and English translation (11694 term pairs) for the INION ISISS descriptor list. Finally, mappings from the Russian INION ISISS descriptor list to the GESUS Thesaurus Sozialwissenschaften were made available.

An example of a mapping from the English Thesaurus for the Social Sciences to the English CSA Thesaurus of Sociological Indexing Terms is given below:

<mapping>
<original-term> *counseling for the aged* </original-term>
<mapped-term> *Counseling + Elderly*</mapped-term>
</mapping>

This example shows that a mapping can overcome differences in technical language and the treatment of singular and plural in different controlled vocabularies.

2.3 Topic Preparation

For topic preparation, colleagues from the GESIS Social Science Information Centre suggested 2-5 topics related to specialized subject areas and potentially relevant in the years 1990-2000 (the coverage of our test collections). Specialized subject areas are based on the 28 subject categories utilized for the GESIS bibliographic service sofid, which biannually publishes updates on new entries in the SOLIS and SOFIS databases (from which the GIRT collections were generated). Topics range from general sociology, family research, women and gender studies, international relations, research on Eastern Europe to social psychology and environmental research. An overview of the service including the 28 topics can be found at the following URL:
http://www.gesis.org/en/information/soFid/index.htm.

The suggestions are then checked for their breadth, variance from previous years and coverage in the test collections and edited for style and format. In 2008, topics 201-225 for the domain-specific collections were created. Figure 1 shows topic 207 as an example.

<top>
<num>207</num>
<EN-title>*Economic growth and environmental destruction*</EN-title>
<EN-desc>*Find documents on the topic of the connection between economic growth and environmental destruction*.</EN-desc>
<EN-narr>*Relevant documents address the connection between economic growth and environmental destruction, particularly the question of whether continued economic growth generally leads to environmental destruction or if the concept of qualitative growth can prevent this*.</EN-narr>
</top>

Figure 1. Example topic in English

All topics were initially created in German and then translated into English and Russian. The method works well for German and English, because the German and English collections are virtually equivalent. However, Russian topic preparation is somewhat more difficult as the collection is different in scope, contains shorter documents and a large and non-controlled vocabulary. Consequently, not all Russian topic translations will retrieve relevant documents in the database.

Table 2 lists all 25 topic titles.

201 Health risks at work	213 Migrant organizations
202 Political culture and European integration	214 Violence in old age
203 Democratic transformation in Eastern Europe	215 Tobacco advertising
204 Child and youth welfare in the Russian Federation	216 Islamist parallel societies in Western Europe
205 Minority policy in the Baltic states	217 Poverty and social exclusion
206 Environmental justice	218 Generational differences on the Internet
207 Economic growth and environmental destruction	219 (Intellectually) Gifted
208 Leisure time mobility	220 Healthcare for prostitutes
209 Doping and sports	221 Violence in schools
210 Establishment of new businesses after the reunification	222 Commuting and labor mobility
211 Shrinking cities	223 Media in the preschool age
212 Labor market and migration	224 Employment service
	225 Chronic illnesses

Table 2. English topic titles for the domain-specific track 2008

3 Overview of the 2008 Domain-Specific Track

Details of the individual runs and methods tested can be found in appendix C of the working notes and in the corresponding articles by the participating groups.

3.1 Participants

Six of the nine registered groups (listed in table 3) have submitted runs and descriptions of their experiments (Fautsch, Dolamic & Savoy, 2008; Gobeill & Ruch, 2008; Kürsten, Wilhelm & Eibl, 2008; Larson, 2008; Meij & de Rijke, 2008; Müller & Gurevych, 2008).

Abbreviation	*Group Institution*	*Country*
Amsterdam	University of Amsterdam	The Netherlands
Chemnitz	Chemnitz University of Technology	Germany
Cheshire	School of Information, UC Berkeley	USA
Darmstadt	TU Darmstadt	Germany
Hug	University Hospitals Geneva	Switzerland
UniNE	Computer Science Department, University of Neuchatel	Switzerland

Table 3. Domain-specific track 2008 - participants

3.2 Submitted Runs

The total number of submitted runs decreased slightly compared to last year, although one more group submitted runs. Table 4 shows the number of runs (numbers from 2007 in brackets).

Task	Runs
Monolingual	
- against German	10 (13)
- against English	12 (15)
- against Russian	9 (11)
Bilingual	
- against German	12 (14)
- against English	9 (15)
- against Russian	8 (9)
Multilingual	9 (9)

Table 4. Submitted runs per task in the domain-specific track 2008

English is the most popular language for monolingual retrieval as well as a starting language for bilingual retrieval. All groups participated in the monolingual English task, and four groups took part in the German and Russian monolingual tasks respectively. Three groups experimented with bilingual against German or English, whereas only 2 groups tackled the bilingual against Russian and multilingual tasks respectively.

3.3 Relevance Assessments

As last year, all relevance assessments were processed using the the DIRECT system (Distributed Information Retrieval Evaluation Campaign Tool) provided by Giorgio M. Di Nunzio and Nicola Ferro from the Information Management Systems (IMS) Research Group at the University of Padova, Italy.

Documents were pooled using the top 100 ranked documents from each submission. Table 5 shows pool sizes and the number of assessed documents per topic for the three different languages.

	Pool size	Documents assessed per topic
German	14793	592
English	14835	593
Russian	13930	557

Table 5. Pool sizes in the domain-specific track 2008

Because of a late submission, the runs by the Hug group were not included in the pooling process but were analyzed with the existing pools. One assessor was assigned for each language to avoid as many interpersonal assessment differences as possible.

Both the feedback from the assessors as well as the precision numbers show that this year's topics were somewhat more difficult or more discriminating. The average number of relevant topics per task and language (table 6) also corroborate this impression. The average number of relevant documents decreased for all three languages with Russian seeing the largest drop. As in previous years, however, the German and English averages are similar.

	German	English	Russian
2008	15%	14%	2%
2007	22%	25%	10%
2006	39%	26%	n/a
2005	20%	21%	9% (RSSC)

Table 6. Relevant documents per language pool

The next three images show the number of relevant documents per individual topics for the three languages.

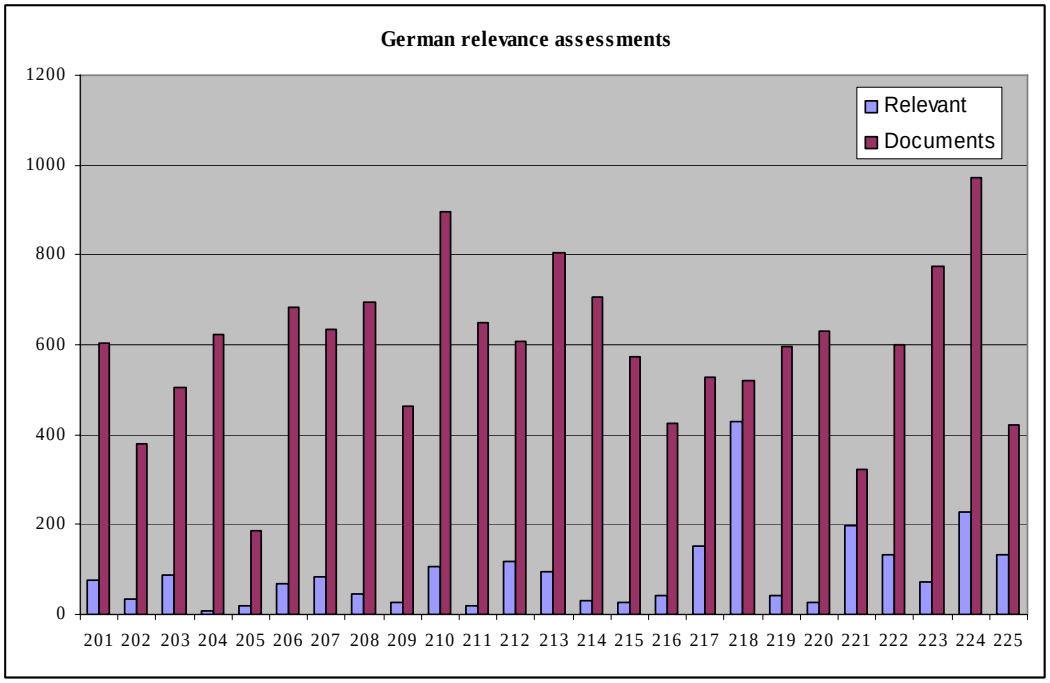

Figure 2. German assessments per topic 2008

For German, six topics stand out as having more than 20% relevant documents in their pool: 217, 218, 221, 222, 224 and 225.

For English, seven topics retrieved more than 20% relevant documents (201, 202, 211, 212, 217, 221, 225). Three of these topics (217, 221, 225) overlap with the German results, surprisingly however, topic 218, which retrieved the greatest number of relevant documents in German, retrieved the least (percentage-wise) in English. This might be due to different interpretations and assessments of the content of the topic (Generational differences on the Internet).

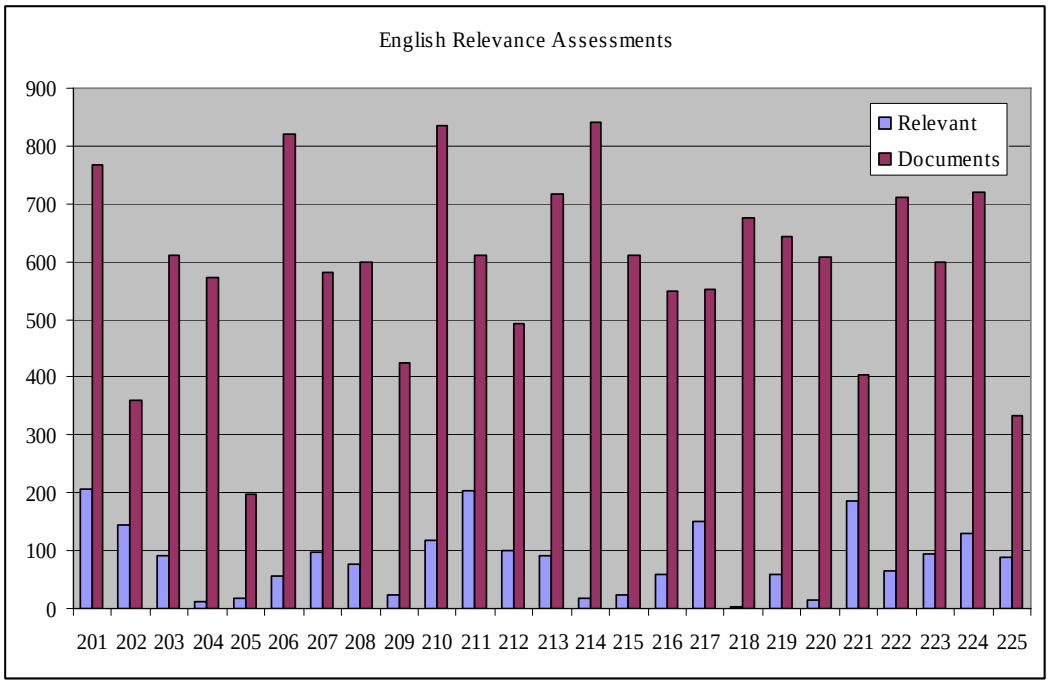

Figure 3. English assessments per topic 2008

One topic (209) did not retrieve any relevant documents in the Russian collection.

For the more difficult Russian collection, the highest percentage of relevant documents retrieved was found for topic 204 (12%), followed by 224 (9%) and 203 (7%). The pool for topic 224 (Employment service) contains also more than 20% relevant documents in the German collection and more than 17% in the English collection.

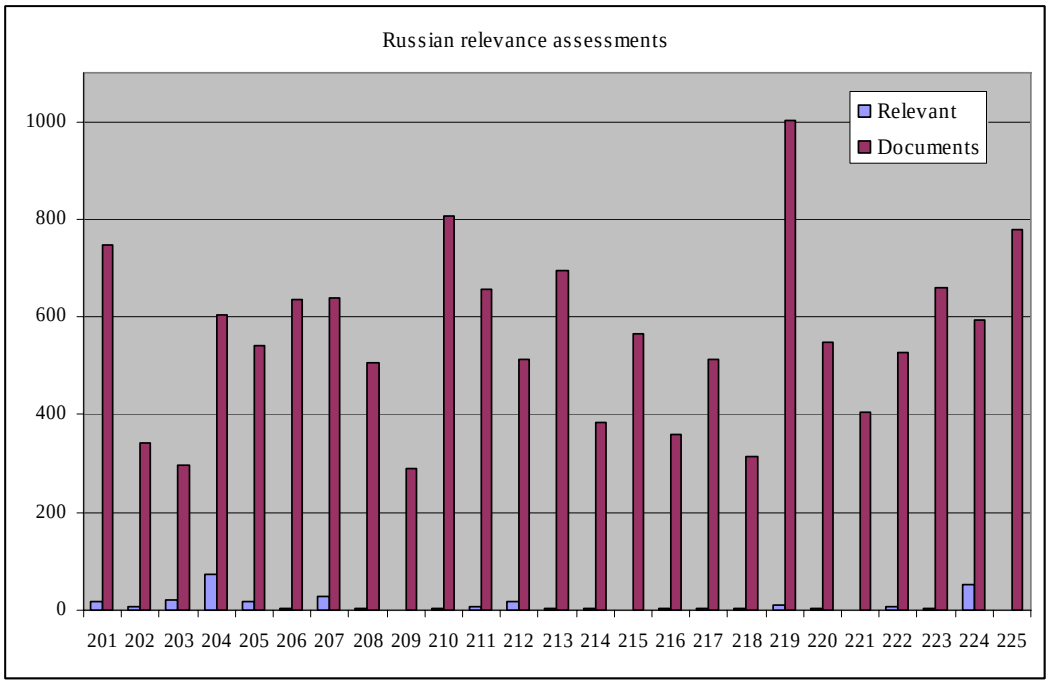

Figure 4. Russian assessments per topic 2008

A closer look at the correlation between the number of relevant documents per topics and precision and recall might reveal more insight. One interesting question is whether the topics with the most relevant documents available are also the "easiest" for retrieval systems to find in terms of precision and recall measures.

3.4 Results

In the Appendix of this volume, varied evaluation measures for each run per task and recall-precision graphs for the top-performing runs for each task can be looked up.

4 Domain-Specific Experiments

This year's track saw the use of a broad range of retrieval models, language processing, translation, and query expansion approaches. Statistical language models, probabilistic and vector-space models were employed with translation approaches that leverage thesaurus mappings as well as machine translation systems or web-based translation services. Two of the six participants employed concept models based on semantic relatedness both for translation and query expansion.

4.1 Retrieval Models

The participants of the 2008 domain-specific track utilized a number of different retrieval models. Statistical language models were used as well as different implementations of the probabilistic model and vector-space schemes. The structure of the collection documents, the topics and the controlled vocabularies and the associated mappings were used to different degrees.

The Chemnitz group (Kürsten, Wilhelm & Eibl, 2008) used their Apache Lucene-based Xtrieval framework for the experiments and utilized the Z-score Operator (Savoy, 2005) to combine the results of runs with different language processing and translation approaches.

Darmstadt (Müller & Gurevych, 2008) applied a statistical model implemented in Lucene in addition to two semantic models, SR-Text and SR-Word. The semantic models utilize both Wikipedia and Wiktionary as sources for terms to form concepts that facilitate the use of semantic relatedness in the retrieval process. The CombSUM method by Fox and Shaw (Fox & Shaw, 1994) was used for the merging of results from the multiple retrieval models

The Geneva group (Gobeill & Ruch, 2008) used their EasyIR system, which supports both regular expression searches and retrieval based on the vector space model.

Berkeley (Larson, 2008) implemented a probabilistic logistic regression model with the Cheshire II system that was also employed for the Adhoc and GeoCLEF tracks.

UniNE (Fautsch, Dolamic & Savoy, 2008) employed and evaluated multiple retrieval models. A tf-idf based statistical model was compared with two probabilistic models, the BM25 scheme and four implementations of the Divergence from Randomness model. Additionally, an approach based on a statistical language model was utilized.

The Amsterdam (Meij & de Rijke, 2008) group used a language model approach to map between query terms, controlled vocabulary concepts and document terms. Parsimonization was used to increase the probability weights of specific terms compared to more general terms in the corpus.

4.2 Language Processing

A number of different combinations of stemming, lemmatization and decompounding techniques were utilized by the participants, often in combination with stopword lists.

Chemnitz used combinations of the Porter and the Krovetz stemmers for English and the Snowball stemmer and an N-Gram based decompounding approach for German. The group used a stemmer developed by UniNE for Russian.

The UniNE group used stopword lists of between 430 and 603 words for the three different corpora languages. Stemming for English was done using the SMART stemmer. 52 stemming rules that removed inflections due to gender, number and case were defined for Russian. German words were treated with a lightweight stemmer and decompounding algorithm developed by the group.

Darmstadt used the probabilistic part-of-speech tagging system TreeTagger (Schmid, 1994) for lemmatization. Decompounding was employed for German words. For retrieval, both a compound word and its elements were used in combination.

Geneva used an implementation of a Porter stemmer.

Berkeley did employ a stopword list for common words in all languages, but did not use decompounding for German.

Amsterdam did not do any preprocessing on the document collections.

4.3 Translation

Different approaches to translation and the treatment of different languages were used by the groups. Besides the use of machine translations software, the language mappings of the provided controlled vocabularies were used in addition to the use of concepts models from external sources (Wikipedia) for cross-language retrieval.

Darmstadt used the Systran machine translation system and utilized cross-language links in the Wikipedia in order to map between concept vectors for different languages in the SR-Text system.

Berkeley used the commercial LEC Power translator with good results but intends to undertake further evaluation to compare the translator with systems like PROMT or Babelfish.

Chemnitz made use of the Google AJAX language API. In addition to pure translation, a combination of automatic translation and language mappings as provided by the bilingual translation tables was employed.

Geneva did not use translation, but employed the bilingual thesaurus for query expansion as described below.

Amsterdam used a combined approach that leveraged concept models for both translation and query expansion.

4.4 Query Expansion

All participants used query expansion. The techniques employed include the expansion by terms from the top-k documents as well the utilization of concept models, idf-based approaches and the use of Google and the Wikipedia.

Chemnitz used a blind feedback approach that was combined for some runs with query expansion based on thesaurus terms. It was found that such use of the controlled vocabulary did not benefit the retrieval effectiveness.

The UniNE group tested four different blind feedback approaches. The classic Rocchio blind feedback method is compared to two variants of an approach that extends a query with terms selected based on their pseudo document frequency, which are considered for inclusion in the query if they are within 10 words of the search term in the document. Finally, Google and Wikipedia were used for query expansion where the terms included in text snippets were used for query expansion.

Geneva used the bilingual thesaurus for query expansion. The descriptors in the top 10 documents for a German query were collected and transfered into English using the bilingual thesaurus, the resulting terms were used for query expansion.

Amsterdam used a blind relevance feedback approach based on concept models of the thesauri provided for the track that used the concepts defined in the thesauri as a pivot language.

Berkeley used a probabilistic blind feedback approach based on the work by Robertson and Sparck Jones (Robertson, 1976).

Darmstadt implemented a query expansion method based on concept models derived from Wikipedia and Wiktionary.

5 Outlook

The results and group papers show that query expansion with blind feedback mechanisms using document, controlled vocabulary terms or external resources is still a major experimentation area for domain-specific retrieval.

This year, new language resources for Russian were provided but the collections remained the same. Nevertheless, due to more difficult queries, the number of relevant documents per topic as well as the precision values have gone down compared to previous years.

Pending availability of resources and permissions, the following different tasks and options might be offered in 2009:

- Potentially additional corpus data
- Full topic run: 125 topics from the years 2003-2008 span the same GIRT corpora – we can offer some experimental runs to compare retrieval results over a small traditional run of 25 topics and the complete topic set
- Change in task: for a given topic, find the most relevant subject headings / keywords (by either cumulating from the relevant documents or other means)
- Adding to the robust track: taking the most difficult topics from the last 5 years and devising a task of 25 topics for a robust domain-specific track
- Proof-of-concept for potential track extension in 2010: small experimental full-text corpus of social science articles (scientific publications)

Acknowledgements

We would like to thank Cambridge Scientific Abstracts for providing the documents for the Sociological Abstracts test collection and INION for providing the documents for the ISISS collection.

We greatly acknowledge the support of Natalia Loukachevitch and her colleagues from the Research Computing Center of M.V. Lomonosov Moscow State University in translating the topics into Russian.

Very special thanks also to Giorgio Di Nunzio and Nicola Ferro from the Information Management Systems (IMS) Research Group at the University of Padova for providing the DIRECT system and all their help in the assessments process and for providing the graphs and numbers for the results analysis.

Claudia Henning did the German assessments. Jeof Spiro translated and assessed the English topics. Oksana Schäfer provided the Russian assessments.

References

GIRT Description (2007). GIRT - Mono- and Cross-language Domain-Specific Information Retrieval (GIRT4). http://www.gesis.org/en/research/information_technology/girt4.htm

Claire Fautsch & Ljiljana Dolamic, Jacques Savoy (2008). UniNE at Domain-Specific IR - CLEF 2008: Scientific Data Retrieval: Various Query Expansion Approaches. This volume.

E. Fox & J. Shaw (1994). Combination of Multiple Searches. Proceedings of the 2nd Text REtrieval Conference (Trec-2), pages 243–252.

Julien Gobeill & Patrick Ruch (2008). First Participation of University and Hospitals of Geneva to Domain-Specific Track in CLEF 2008. This volume.

Michael Kluck & Frederik C. Gey (2001). The Domain-Specific Task of CLEF - Specific Evaluation Strategies in Cross-Language Information Retrieval . In: Carol Peters (ed.): Cross-Language Information Retrieval and Evaluation. Workshop of the Cross-Language Information

Evaluation Forum, CLEF 2000, Lisbon, Portugal, September 21-22, 2000, Revised Papers. Berlin/Heidelberg/New York: Springer 48-56 (Lecture Notes in Computer Science, 2069)

Michael Kluck (2004). 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. Berlin/Heidelberg/New York: Springer 2004, 379-393 (Lecture Notes in Computer Science, 3237)

Jens Kürsten, Thomas Wilhelm & Maximilian Eibl (2008). The Xtrieval Framework at CLEF 2008: Domain-Specific Track. This volume.

Ray R. Larson (2008). Back to Basics - Again - for Domain Specific Retrieval. This volume.

Edgar Meij & Maarten de Rijke (2008). The University of Amsterdam at the CLEF 2008 Domain Specific Track: Parsimonious Relevance and Concept Models. This volume.

Christof Müller & Iryna Gurevych (2008). Using Wikipedia and Wiktionary in Domain-Specific Information Retrieval. This volume.

Jacques Savoy (2005). Data Fusion for Effective European Monolingual Information Retrieval. Multilingual Information Access for Text, Speech and Images: 5th Workshop of the Cross-Language Evaluation Forum, CLEF 2004, Bath, UK, September 15-17, 2004: Revised Selected Papers, 2005.

H. Schmid (1994). Probabilistic part-of-speech tagging using decision trees. Proceedings of International Conference on New Methods in Language Processing, 12, 1994.

S. Robertson, S. Jones, et al. (1976). Relevance Weighting of Search Terms. Journal of the American Society for Information Science, 27(3):129–46, 1976.