CLEF 2005: Domain-Specific Track Overview

Michael Kluck[1], Maximilian Stempfhuber[2]

[1]Stiftung Wissenschaft und Politik (SWP), German Institute for International and Security Affairs, Berlin, Germany
michael.kluck@swp-berlin.org
[2]Informationszentrum Sozialwissenschaften (IZ). Bonn, Germany
stempfhuber@iz-soz.de

The domain-specific track aims at mono- and cross-language information retrieval on structured scientific data. This track studies retrieval in a domain-specific context using two social science databases: The German Indexing and Retrieval Testdatabase (GIRT) (forth version GIRT-4: German/English pseudo-parallel corpus with identical documents) with 302,638 documents in total, and the Russian Social Science Corpus (RSSC) with 94,581 documents.

Different sub-tasks have been available:

1. Monolingual task: German topics against German data GIRT4-DE, English topics against English data GIRT4-EN, Russian topics against Russian data RSSC;

2. Bilingual task: German topics against English data GIRT4-EN, German topics against Russian data RSSC, English topics against German data GIRT4-DE, English topics against Russian data RSSC, Russian topics against German data GIRT4-DE, Russian topics against English data GIRT4-EN;

3. Multilingual task: German topics against all data GIRT4-DE, GIRT4-EN, RSSC, English topics against all data GIRT4-DE, GIRT4-EN, RSSC, Russian topics against all data GIRT4-DE, GIRT4-EN, RSSC.

The domain-specific task attracted 8 participating groups (three of them from Berkeley), which delivered a total of 76 runs: 40 monolingual runs, 33 bilingual runs and 3 multilingual runs. For detailed figures see the following table:

Sub-task	**# Participants**	**# Runs**	**Topic Language**
Multi-lingual	1	3	DE 1; EN 1; RU 1
Bilingual X → DE	5	15	EN 14; RU 1
Bilingual X → EN	4	13	DE 7; RU 6
Bilingual X → RU	3	5	DE 2; EN 3
Monolingual DE	6	17	
Monolingual EN	6	15	
Monolingual RU	5	8	
Sum	8	76	

The retrieval systems explicitly mentioned by the participants are based on logistic regression or OKAPI formula. For translation purposes several MT systems have been used: L+H, SYSTRAN, PROMT, WorldLingo, IMTranslator, FreeTranslation, Eurodictautom. Some groups concentrated on data fusion aspects. Most groups used the thesaurus information provided with GIRT to translate queries or to produce a translation vocabulary. Linguistic treatment reached from the use of stemmers, POS, de-compounding to the extraction of semantically related concepts or WordNet concepts. Some groups did experiments with blind relevance feedback.

Concerning the results some groups were emphasizing the importance of robustness of the used methodology and of high-quality results on a per query basis rather than high average precision computed over all queries.

Remarks and figures on the assessment process will conclude the overview of the domain-specific task.