

Domain-Specific Russian Retrieval: A Baseline Approach

Fredric C. Gey

UC Data Archive & Technical Assistance (UC DATA),
University of California, Berkeley, CA 94720-5100 USA
gey@berkeley.edu

Abstract. Berkeley group 2 chose to perform some very straightforward experiments in retrieval of Russian documents using queries derived from topics in all three languages. Thus we performed two runs with monolingual Russian retrieval and one cross-lingual run each with German topics and English topics. Query translation was done using the online PROMT translator (www.translate.ru). Monolingual results were substantially better than the overall median performance of all Russian runs, and cross-language results were encouraging with German→Russian retrieval doing substantially better than English→Russian.

Categories and Subject Descriptors

H.3 [Information Storage and Retrieval]: H.3.3 Information Search and Retrieval -- Query Formulation; H.3.1 Content Analysis and Indexing -- Thesauruses; H.3.4 Systems and Software -- Performance evaluation (efficiency and effectiveness); H.3.7 Digital Libraries

General Terms

Measurement, Performance, Experimentation

Keywords

Cross-language information retrieval, Russian retrieval

1 Introduction

Domain-specific retrieval has been a track in CLEF since the beginning with the GIRT collections [4]. For the CLEF 2005 campaign, the domain-specific included a Russian social science abstract collection and the topics were available in German, English and Russian for experiments with all DS collections. Berkeley group 2 performed some very straightforward experiments in retrieval of Russian documents using queries derived from topics in all three languages. Thus we performed two runs with monolingual Russian retrieval and one cross-lingual run each with German topics and English topics. Query translation was done using the online PROMT translator (www.translate.ru) which prior experience had shown to produce more useful translations than the SYSTRAN translation system (<http://babelfish.altavista.com>).

The Russian collection of CLEF 2005 domain specific track consists of 94,581 documents containing titles (for all documents) abstracts (for 47,130 documents or 50% of the collection). Unfortunately for this collection, only 12% of the collection (11,403 documents) have controlled-vocabulary thesaurus terms assigned. The GIRT thesaurus terms are assigned from the Thesaurus for the Social Sciences [7] which has been made available in German, English and Russian.

2 Document Ranking, Collection and Query Processing and Translation

In all its CLEF submissions, the Berkeley 2 group used a document ranking algorithm based on logistic regression first used in the TREC-2 conference [1]. For all runs, we used a 256 word Russian stopword list developed for the CLEF 2003 Izvestia collection to remove very common words [6]. For stemming we utilized the Russian SNOWBALL Stemmer available from www.tartarus.org/snowball. As a general procedure, we also use Aitao Chen's blind feedback algorithm [2,3] every run. It selects the top 30 ranked terms from the top 20 ranked documents from the initial search to merge with the original query. Thus the sequence of processing for retrieval is: query → stopword removal → (decompounding) → stemming → ranking → blind feedback

The only translation done was query translation from English and German to Russian using the PROMT translator found online at www.translate.ru.

3 Runs and Results

Our results are summarized by topic in the following table with comparison to overall precision. The highlighted columns are the median performances for monolingual and cross-language IR while the final row is precision averaged over all 25 topics:

Topic	BestMono	MedMono	BK2MLRU1	BK2MLRU2	BestCLIR	MedCLIR	BK2BLER1	BK2BLGR1
126	0.5437	0.2004	0.5437	0.2083	0.5182	0.4119	0.421	0.5182
127	0.9036	0.8295	0.9036	0.8789	0.8691	0.6872	0.8691	0.7559
128	0.7085	0.2613	0.2783	0.1973	0.3793	0.2374	0.2594	0.3793
129	0.0596	0.0279	0.0596	0.0095	0.0021	0	0.0021	0.0011
130	0.1227	0.0143	0.0801	0.026	0.0597	0.0061	0.0025	0.0061
131	1	0.0005	1	0.5089	0.5294	0.0976	0.5294	0.2976
132	0.125	0.027	0.125	0.0312	0.304	0.125	0.125	0.1
133	0.1791	0.0606	0.1716	0.1152	0.4643	0.1071	0.3915	0.4643
134	0.3917	0.0992	0.1024	0.0959	0.0913	0.02	0.0913	0.0607
135	0.534	0.1463	0.1419	0.534	0.1876	0.0801	0.1876	0.0257
136	0.6905	0.5087	0.585	0.4324	0.1109	0.022	0.1109	0.1002
137	0.287	0.1797	0.287	0.1855	0.191	0.1114	0.1555	0.191
138	0.5313	0.4702	0.4727	0.3337	0.177	0.0432	0.0432	0.177
139	0.616	0.4282	0.3966	0.4223	0.5145	0.2241	0.2294	0.5145
140	0.0503	0.0368	0.0292	0.0342	0.0358	0.0271	0.0255	0.0271
141	0.2847	0.0454	0.0539	0.2847	0.2086	0.1933	0.1933	0.1344
142	0.7698	0.3085	0.3731	0.2439	0.2886	0.0678	0.0136	0.2886
143	1	0.2667	1	0.45	1	0.7381	0.0094	1
144	0.0402	0.0089	0.0056	0.0091	0.027	0.0137	0.0065	0.0137
145	0.6553	0.5809	0.5335	0.2058	0.6821	0.5949	0.5949	0.6821
146	0.0435	0.0197	0.004	0.0091	0	0	0	0
147	0.125	0	0	0.125	0.0016	0	0.0011	0
148	0.3939	0.2492	0.2405	0.3587	0.1618	0.0639	0.1618	0.0551
149	0.2066	0.0111	0.2066	0.1734	0.088	0.0257	0.088	0.0257
150	0	0	0	0	0.0178	0.0139	0.0139	0.0102
Avg	0.3887	0.1832	0.3038	0.2349	0.2557	0.14	0.181	0.2331

The first monolingual Russian run (BK2MLRU1) and the two bilingual runs (BK2BLER1, BK2BLER2) were made using the required Title and Description (T-D) fields. The second monolingual run (BK2MLRU2) used the Title, Description and Narrative (T-D-N) fields. The T-D run (BK2MLRU1) achieved overall mean average precision of 0.304 with 9 best-of-topic results out of the 25 topics. Interestingly, the T-D run performed 30 percent higher than the T-D-N monolingual run (BK2MLRU2) which had an average precision of only 0.235. We speculate that this is because over half the documents in the collection only have a <TITLE> field and not a <TEXT> field. Topic 150 Поведение во время телепередач (Television Behaviour) retrieved zero relevant documents from all DS monolingual run, while bilingual runs to the Russian found only two relevant document with best average precision of 0.0178.

The German-Russian bilingual run BK2BLGR1 (MAP of 0.233) performed twenty nine percent better than the English-German run BK2BLER1 (MAP of 0.181). Much of this difference can be attributed to topic 143 Отказ от курения (Giving up Smoking) where the German translation seems to have been more accurate than the English one. The G->R precision for topic 143 was 1.0 while the E->R precision was 0.0094.

4 Conclusion

We believe we achieved our goal of providing a baseline performance for the Russian Domain Specific collection of CLEF. We believe our results provide a foundation from which more sophisticated experiments can be developed which leverage the controlled vocabulary indexing of the CLEF DS collections. For the future of CLEF domain specific Russian to be interesting and successful, substantially more documents will need to have indexing keywords assigned to the documents – 12 percent is simply not enough to perform meaningful experiments on the utility of controlled vocabulary. In addition, document abstracts provide a richer set of textual clues from which to mine associations to controlled vocabulary terms as the work by Petras shows [5].

5 Acknowledgement

Thanks to Aitao Chen for implementing and permitting the use of the logistic regression formula for probabilistic information retrieval as well as German decompounding and blind feedback in his MULIR retrieval system. Thanks also to Vivien Petras who performed the actual indexing and running of the experiments.

6 References

1. Chen, A, W Cooper, and F Gey. 1994. Full text retrieval based on probabilistic equations with coefficients fitted by logistic regression. In *The Second Text Retrieval Conference (TREC-2)*, edited by D. K. Harman.
2. Chen, Aitao. 2003. *Cross-Language Retrieval Experiments at CLEF 2002. Lecture Notes in Computer Science 2785*, Springer 2003.
3. Chen, A, and F Gey. 2004. Multilingual Information Retrieval Using Machine Translation, Relevance Feedback and Decompounding. *Information Retrieval 7* (1-2):149-182.
4. Kluck, Michael. 2003. The GIRT Data in the Evaluation of CLIR Systems - from 1997 Until 2003. In *Comparative Evaluation of Multilingual Information Access Systems: 4th Workshop of the Cross-Language Evaluation Forum, CLEF 2003*, edited by C. A. Peters. Trondheim, Norway, August 21-22, 2003: Lecture Notes in Computer Science 3237, Springer 2004.
5. Petras, V, How One Word Can Make all the Difference – Using Subject Metadata for Automatic Query Expansion and Reformulation, in *this volume, 2005*.
6. Petras, V., N. Perelman, and F Gey. 2003. UC Berkeley at CLEF 2003 -- Russian Language Experiments and Domain-Specific Cross-Language Retrieval. In *Comparative Evaluation of Multilingual Information Access Systems: 4th Workshop of the Cross-Language Evaluation Forum, CLEF 2003*. Trondheim, Norway, August 21-22, 2003: Lecture Notes in Computer Science 3237, Springer 2004.
7. Schott, Hannelore. 2000. *Thesaurus for the Social Sciences*. 2 vols. Vol. 1. German - English, 2. English - German. Bonn: Informations-Zentrum Socialwissenschaften.