Searching a Russian Document Collection Using English, Chinese and Japanese Queries

Fredric C. Gey
(gey@ucdata.berkeley.edu)

UC Data Archive & Technical Assistance
University of California, Berkeley, CA 94720 USA

ABSTRACT. As in CLEF 2003, Berkeley experimented with the CLEF Russian Izvestia document collection with monolingual and bilingual runs for the Russian collection. For CLEF 2004 we also experimented with Chinese and Japanese as topic languages, using English as the 'pivot' language. For bilingual retrieval our approaches were query translation (for English as a topic language) and 'fast' document translation from Russian to English (for Chinese and Japanese translated to English as the topic language). Chinese and Japanese topic retrieval significantly under-performed English → Russian retrieval because of the 'double translation' loss of effectiveness.

1 Introduction

CLEF 2003 was the first time a Russian language document collection was available in CLEF. We had worked for several years with Russian topics in both the GIRT task and the CLEF main tasks, so extension of our techniques to Russian was straightforward No unusual methodology was applied to the Russian collection, however encoding remained an issue and we ended up using the KOI-8 encoding scheme for both Russian documents and topics.

2 Document ranking

Berkeley has used a monolingual document ranking algorithm which uses statistical clues found in documents and queries to predict a dichotomous variable (relevance) based upon logistic regression fitting of prior relevance judgments. The exact formula is:

$$\log O(R \mid D,Q) = \log \frac{P(R \mid D,Q)}{1 - P(R \mid D,Q)}$$

$$= \log \frac{P(R \mid D,Q)}{P(\overline{R} \mid D,Q)}$$

$$= -3.51 + 37.4 * x_1 + 0.330 * x_2$$

$$- 0.1937 * x_3 + 0.0929 * x_4$$

where $O(R \mid D,Q), P(R \mid D,Q)$ mean, respectively, *odds* and *probability* of relevance of a document with respect to a query, and

$$x_1 = \frac{1}{\sqrt{n}+1} \sum_{i=1}^{n} \frac{qtf_i}{ql+35}$$

$$x_2 = \frac{1}{\sqrt{n}+1}\sum_{i=1}^{n}\log\frac{dtf_i}{dl+80}$$

$$x_3 = \frac{1}{\sqrt{n}+1}\sum_{i=1}^{n}\log\frac{ctf_i}{cl}$$

$$x_4 = n$$

where n is the number of matching terms between a document and a query, and
ql : query length
dl: document length
cl: collection length
qtf_i: the within-query frequency of the ith matching term
dtf_i: the within-document frequency of the ith matching term
ctf_i: the occurrence frequency of the ith matching term in the collection.

This formula has been used since the second TREC conference and for all NTCIR and CLEF cross-language evaluations [1].

3 Russian Retrieval for the CLEF main task

CLEF 2003 marked the first time a document collection was available and evaluated in the Russian language. The CLEF Russian collection consists of 16,716 articles from *Izvestia* newspaper for 1995. This is a small number of documents by most CLEF measures (the smallest other collection of CLEF 2003, Finnish, has 55,344 documents; the Spanish collection has 454,045 documents). We used the Russian and English indexes generated for CLEF 2003 for all our CLEF 2004 Russian runs. The collection is also rich in metadata, including specification of geography for news articles; this can be exploited for mapping and geotemporal querying of documents relating to place and time [2].

3.1 Encoding Issues

The Russian document collection was supplied in the UTF-8 unicode encoding, as were the Russian version of the topics. However, since the stemmer we employ is in KOI8 format, the entire collection was converted into KOI8 encoding, as with CLEF 2003 [3]. In indexing the collection, we converted upper-case letters to lower-case and applied Snowball's Russian stemmer (http://snowball.tartarus.org/russian/stemmer.html) together with Russian stopword list created by merging the Snowball list with a translation of the English stopword list. In addition the PROMPT translation system would also only work on KOI8 encoding which meant that our translations from English also would come in that encoding.

3.2 Russian Monolingual Retrieval

We submitted two Russian monolingual runs, the results of which are summarized below. As in CLEF 2003, both runs utilized blind feedback, choosing the top 30 terms from the top ranked 20 documents of an initial retrieval run For BKRUMLRR1 and BKRUMLRR2 runs we used TITLE and DESCRIPTION document fields for indexing. The results of our retrieval are summarized in Table 1. Results were reported by the CLEF organizers for 34 topics which had one or more relevant documents.

Run Name	BKRUMLRR1	BKRUMLRR2
Index	Koi	Koi
Topic fields	TD	TDN
Retrieved	34000	34000

Relevant	123	123
Rel Ret	105	108
Precision		
at 0.00	0.5734	0.5856
at 0.10	0.5636	0.5688
at 0.20	0.5506	0.5394
at 0.30	0.4969	0.4871
at 0.40	0.4670	0.4465
at 0.50	0.4526	0.4459
at 0.60	0.3628	0.3619
at 0.70	0.2989	0.3175
at 0.80	0.2839	0.3175
at 0.90	0.2555	0.2573
at 1.00	0.2548	0.2555
Avg. Precision	0.4024	0. 4005

Table 1: Berkeley Monolingual Russian runs for CLEF 2004

Adding the Narrative section to the query did not significantly improve results because the Narrative section did not contribute additional content terms beyond those found in the Title and Description fields of the topics.

3.3 Bilingual Retrieval from English to Russian

We submitted eight bilingual runs against the Russian document collection, four with English as topic language and two each with Chinese and Japanese as topic languages. These runs used an index in which only the TITLE and TEXT fields of each Russian document was indexed, so are directly comparable to the monolingual runs BKMLRURR1 and BKMLRURR2 above. The four English→Russian runs utilized query translation from English topics into Russian. We compared two web-available translation systems, SYSTRAN at http://babelfish.altavista.com/ for the first two runs (BKRUBLER1, BKRUBLER2) and the PROMT system (runs BKRUBLER3, BKRUBLER4) developed in Russia and found at http://www.translate.ru.

Run Name	BKRUBLER1	BKRUBLER2	BKRUBLER3	BKRUBLER4
Translation	Babelfish	Babelfish	PROMT	PROMT
Topic fields	TD	TDN	TD	TDN
Retrieved	34000	34000	34000	34000
Relevant	123	123	123	123
Rel Ret	69	85	98	93
Precision				
at 0.00	0.2444	0.2965	0.5158	0.4575
at 0.10	0.2430	0.2965	0.5147	0.4575
at 0.20	0.2423	0.2806	0.4951	0.4493
at 0.30	0.1809	0.2269	0.4328	0.4281
at 0.40	0.1563	0.2205	0.3617	0.3239
at 0.50	0.1445	0.1976	0.3470	0.2932
at 0.60	0.0896	0.0940	0.2648	0.1990
at 0.70	0.0796	0.0813	0.2268	0.1907
at 0.80	0.0771	0.0806	0.2145	0.1782
at 0.90	0.0764	0.0802	0.1997	0.1629
at 1.00	0.0764	0.0797	0.1997	0.1629
Avg. Prec.	0.1361	0.1638	0.3291	0.2850

Table 2. Bilingual English → Russian runs.

.

The results demonstrate clearly the superiority of the PROMT system for this topic set.

3.4 Bilingual Retrieval from Chinese and Japanese to Russian

Because Chinese and Japanese were available as topic languages, we experimented with these languages by translating the topics to English (i.e. used English as a pivot language). Our approach to translation from Chinese or Japanese topics to English was to utilize a widely available software package, the SYSTRAN CJK Personal system available for less than $US100. from www.systransoft.com. However, instead of query translation a second time, we utilized a technique (also used for Russian in CLEF 2003) developed by Aitao Chen, called 'Fast Document Translation' [4]. Instead of doing complete document translation using MT software, the MT system is used to translate the entire vocabulary of the document collection on a word-by-word basis without the contextualization of position in sentence with respect to other words. Monolingual retrieval was performed by matching the English versions of the Chinese or Japanese topics against the translated English document collection. More details can be found in our CLEF-2003 final paper [3].

The results, displayed below in Table 3, show that there is considerable loss of performance when using English as a pivot language for these Asian language (we have re-displayed the best English→Russian runs for comparison). It may be that this performance was hampered by the reduced utility of the English documents translated from Russian, as was the case for our CLEF 2003 bilingual performance which used this method. We did not try merging of runs from the two methods to see if it would improvement performance.

Run Name	BKRUBLER3	BKRUBLER4	BKRUMLZE1	BKRUMLZE2	BKRUMLJR1	BKRUMLJR2
Language	English	English	Chinese	Chinese	Japanese	Japanese
Translation	PROMT	PROMT	Systran CJK	Systran CJK	Systran CJK	Systran CJK
Topic fields	TD	TDN	TD	TDN	TD	TDN
Retrieved	34000	34000	34000	34000	34000	34000
Relevant	123	123	123	123	123	123
Rel Ret	98	93	57	68	64	67
Precision						
at 0.00	0.5158	0.4575	0.1659	0.1924	0.2036	0.1709
at 0.10	0.5147	0.4575	0.1659	0.1924	0.2036	0.1709
at 0.20	0.4951	0.4493	0.1559	0.1822	0.1888	0.1699
at 0.30	0.4328	0.4281	0.1167	0.1417	0.1689	0.1249
at 0.40	0.3617	0.3239	0.1137	0.1414	0.1607	0.1185
at 0.50	0.3470	0.2932	0.1051	0.1215	0.1288	0.1130
at 0.60	0.2648	0.1990	0.0844	0.1077	0.0858	0.0898
at 0.70	0.2268	0.1907	0.0704	0.0921	0.0808	0.0846
at 0.80	0.2145	0.1782	0.0551	0.0782	0.0653	0.0726
at 0.90	0.1997	0.1629	0.0540	0.0776	0.0611	0.0701
at 1.00	0.1997	0.1629	0.0540	0.0776	0.0611	0.0701
Avg. Prec.	0.3291	0.2850	0.0956	0.1197	0.1166	0.1050

Table 3. Bilingual Chinese/Japanese → Russian runs

3.5. Brief Analysis of Retrieval Performance

Our monolingual Russian performance was acceptable but certainly not outstanding. For many topics, Title-Description runs out-performed Title-Description-Narrative runs, because the Narrative section added no new information and might sometimes add noise terms.

For all our runs our bilingual retrieval results were worse than monolingual (Russian-Russian) retrieval in terms of overall precision. However the translation of English to Russian by the PROMT system achieved 82% of monolingual for the TD runs. One puzzling and interesting topic was number 202 ("Nick Leeson's Arrest") where our bilingual retrieval out-performed our monolingual runs – it seems that the PROMT translation and transliteration “Арест Ника Лизона” came up with a better spelling of the last name than the Russian topic creator who used “Арест Ника Леесон”, which did not seem to match any relevant documents. According to the summary results for Russian monolingual, at least one run achieved 1.00 precision for this topic; it would be most interesting to see how they modified the topic to match to the three relevant documents.

A cautionary note must be made about the CLEF-2004 Russian topic set. The total number of relevant documents was only 123 for the entire topic set, with a mean of 3.6 relevant documents per topic. Because of the nature of the retrieval results by query from the Russian collection (22 of the 34 topics have 2 or fewer relevant documents) one has to be careful about drawing conclusions from any submitted results.

4 Summary and Acknowledgments

For CLEF 2004, we experimented with the CLEF Russian document collection with both monolingual Russian and bilingual to Russian from English, Chinese and Japanese topics In addition to query translation methodology for bilingual retrieval, we tried a fast document translation method of the Russian collection to English and performed English-English monolingual retrieval with the translated topics from Chinese and English to Japanese. Chinese→Russian and Japanese→Russian bilingual performance results were significantly worse than query translation from English to Russian.

We would like to thank Aitao Chen for supplying writing the logistic regression ranking software and for performing the fast document translation from Russian to English.

5 References

[1] A. Chen, W. Cooper and F. Gey. Full text retrieval based on probabilistic equations with coefficients fitted by logistic regression. In: D.K. Harman (Ed.), *The Second Text Retrieval Conference (TREC-2)*, pages 57-66, March 1994

[2] F. Gey and K. Carl, Geotemporal Querying of Multilingual Documents, Proceedings of the Workshop on Geographic Information Retrieval, available at: http://www.geo.unizh.ch/~rsp/gir/abstracts/gey.pdf.

[3] V. Petras, N. Perelman and F. Gey. UC Berkeley at CLEF-2003 – Russian Language Experiments and Domain-Specific Retrieval. To appear in: *Proceedings of the CLEF 2003 Workshop*, Springer Computer Science Series.

[4] A. Chen and F. Gey. Multilingual Information Retrieval Using Machine Translation, Relevance Feedback, and Decompounding, *Information Retrieval Journal: Special Issue on CLEF*, V7 No 1-2, pp 149-182, Jan-Apr 2004