Working with Russian Queries for the GIRT, Bilingual and Multilingual CLEF Tasks

Fredric C. Gey[1], Hailing Jiang[2] and Natalia Perelman[2]

[1] UC Data Archive & Technical Assistance,
[2] School of Information Management and Systems
University of California
Berkeley, CA 94720 USA

Abstract. For our activities within the CLEF 2001 evaluation, Berkeley group one participated in the bilingual, multilingual and GIRT tasks focussing on the use of Russian queries. Performance on the Russian queries⟶English documents bilingual task was excellent, comparable to performance using German queries. For the multilingual task we utilized English as a pivot language between Russian and German and the English/French/German/Italian/Spanish document collections. Performance here was merely average. The GIRT task performed Russian⟶German Cross-Language IR by comparing web-available machine translation with lookup techniques on the GIRT thesaurus.

1 Introduction

Successful cross-language information retrieval (CLIR) combines linguistic techniques (phrase discovery, machine translation, bilingual dictionary lookup) with robust monolingual information retrieval. For monolingual retrieval the Berkeley group has used the technique of logistic regression from the beginning of the TREC series of conferences. In TREC-2 [1] we derived a statistical formula for predicting probability of relevance based upon statistical clues contained within documents, queries and collections as a whole. This formula was used for document retrieval in Chinese[3] and Spanish in TREC-4 through TREC-6. We utilized the identical formula for English and German queries against the English/French/German/Italian document collections in the CLEF 2000 evaluation[10]. During the past two years, the formula has proven well-suited for Japanese and Japanese-English cross-language information retrieval as well as English-Chinese CLIR[6], even when only trained on English document collections. Participation in the NTCIR Workshops in Tokoyo
(http://research.nii.ac.jp/ñtcadm/workshop/work-en.html)
led to different techniques for cross-language retrieval, ones which utilised the power of human indexing of documents to improve retrieval. Alignments of parallel texts were used to create large-scale bilingual lexicons between English and Japanese and between English and Chinese. Such lexicons were well-suited to the technical nature of the NTCIR collections of scientific and engineering articles.

2 Logistic Regression for Document Ranking

The document ranking formula used by Berkeley in all of our CLEF retrieval runs was the TREC-2 formula [1]. The ad hoc retrieval results on the TREC test collections have shown that the formula is robust for long queries and manually reformulated queries. Applying the same formula (trained on English TREC collections) to other languages has performed well, as on the TREC-4 Spanish collections, the TREC-5 Chinese collection and the TREC-6 and TREC-7 European languages (French, German, Italian) [4, 5]. Thus the algorithm has demonstrated its robustness independent of language as long as appropriate word boundary detection (segmentation) can be achieved. The logodds of relevance of document D to query Q is given by

$$\log O(R|D,Q) = log\frac{P(R|D,Q)}{P(\overline{R}|D,Q)} \tag{1}$$

$$= -3.51 + \frac{1}{\sqrt{N}+1}\Phi + .0929 * N \tag{2}$$

$$\Phi = 37.4\sum_{i=1}^{N}\frac{qtf_i}{ql+35} + 0.330\sum_{i=1}^{N}\log\frac{dtf_i}{dl+80}$$
$$-0.1937\sum_{i=1}^{N}\log\frac{ctf_i}{cl} \tag{3}$$

where $P(R|D,Q)$ is the probability of relevance of document D with respect to query Q, $P(\overline{R}|D,Q)$ is the probability of irrelevance of document D with respect to query Q. Details about the derivation of these formulae may be found in our TREC paper [1]. It is to be emphasized that training has taken place exclusively on English documents but the matching has proven robust over seven other languages in monolingual retrieval, including Japanese and Chinese where word boundaries form an additional step in the discovery process.

3 Submissions for the CLEF main tasks

CLEF has three main tasks: monolingual (non-English) retrieval, bilingual (where non-English queries are run against the CLEF sub-collection of English language documents), and multilingual, where queries in any language are run against a multilingual collection of documents comprised of the union of subcollections in English, French, German, Italian and Spanish. We chose this year to participate in the bilingual and multilingual main tasks. In addition, where our focus last year was on English and German source queries, this year we wished to explore the interesting question of whether a less-used query language (in this case Russian) could achieve performance comparable to the more mainstream Western European languages.

For CLEF main tasks we submitted 7 runs, 3 for the bilingual (German/Russian-English) task and 4 for the Multilingual task. Table 1 summarizes these runs which are described in the next sections.

For the Bilingual task we submitted:			
Run Name	Language	Run type	Priority
BKBIGEM1	German	Manual	1
BKBIREM1	Russian	Manual	2
BKBIREA1	Russian	Automatic	3
For the Multilingual task we submitted:			
BKMUGAM1	German	Manual	1
BKMUEAA1	English	Automatic	2
BKMUEAA2	English	Automatic	3
BKMUREA1	Russian	Automatic	4

Table 1. Summary of seven official CLEF runs.

3.1 Bilingual Retrieval of the CLEF collections

Bilingual retrieval is performed by running queries in another language against the English collection of CLEF. We chose to focus on Russian but to do German for a baseline comparison. The run BKBIGEM1 was obtained by translating the German queries to English using the L&H Power Translator and then manually adjusting the resulting queries by searching for the untranslated terms in our own special association dictionary created from a library catalog. An example of words not found comes from Topic 88 about 'mad cow disease'. In the German version, the words Spongiformer and Enzephalopathie were not translated by the commercial system, but our association dictionary obtained the words 'hepatic encephalopathy' associated with the inquery Enzephalopathie. Further details about our methodology can be found in [8]. The run BKBIREA1 was obtained by using the PROMPT web-based translator (http://www.translate.ru/). As with the German translation, certain words were not translated. Our methodology to deal with this was twofold – first we transliterated the Russian queries to their romanized alphabetic equivalent, and then we added untranslated terms to the English query in their transliterated form. For example Topic 50 on 'uprising of Indians in Chiapas', the Russian word чиапас was not translated, It can, however, be transliterated as 'chiapas'.

3.2 Bilingual Performance

Our bilingual performance can be found in Table 2. The final line of the table, labeled "CLEF Prec" is computed as an average of each CLEF median precision among all submitted runs. The average is performed over the 47 queries for which the English collection had relevant documents. While an average of medians cannot be considered a statistic from which rigorous inference can be made, we have found it useful to average the medians of all queries as sent by CLEF organizers. Comparing our overall precision to this average of medians yields some fuzzy gauge of whether our performance is better, poorer, or about the same as the median performance.

Using this measure we can find that all our bilingual runs performed significantly better than the median for CLEF bilingual runs.

Run ID	BKBIGEM1	BKBIREM1	BKBIREA1
Retrieved	47000	47000	47000
Relevant	856	856	856
Rel. Ret	812	737	733
Precision			
at 0.00	0.7797	0.6545	0.6420
at 0.10	0.7390	0.6451	0.6303
at 0.20	0.6912	0.5877	0.5691
at 0.30	0.6306	0.5354	0.5187
at 0.40	0.5944	0.4987	0.4806
at 0.50	0.5529	0.4397	0.4167
at 0.60	0.4693	0.3695	0.3605
at 0.70	0.3952	0.3331	0.3218
at 0.80	0.3494	0.2881	0.2747
at 0.90	0.2869	0.2398	0.2339
at 1.00	0.2375	0.1762	0.1743
Brk. Prec.	0.5088	0.4204	0.4077
CLEF Prec.	0.2423	0.2423	0.2423

Table 2. Results of three official Berkeley CLEF bilingual runs.

3.3 Multilingual Retrieval of the CLEF collections

Our non-English multilingual retrieval runs were based upon our bilingual experiments, extended to French/Italian/Spanish using English as a pivot language and (again) the L&H Power Tranlator as the MT system to tranlate queries from one language to another. Run BKMURAA1 takes the English translated queries of the bilingual run BKBIREA1 and again translates them to French/German/Italian/Spanish. Run BKMUGAM1 takes the German queries of bilingual run BKBIGEM1 as well as the translation of their English equivalents into French/Italian/Spanish. For comparison we did direct translation from the English queries in runs BKMUEAA1 and BKMUEAA2. The difference between these two runs is that BKMUEAA1 used Title and Description fields only.

The results show that with Russian queries we are about one third lower in average precision than with either English or German queries. We are currently studying why this is so. In addition our overall performance seems only slightly above the CLEF-2001 median performance. This seemed puzzling when compared to our excellent bilingual performance and our above average performance at CLEF-2000. For comparison we also inserted the average of median precisons for last year (Row 2000 Prec. at the bottom of Table 3). As can be seen, the median performance in terms of query precision for CLEF-2001 of 0.2749 is about 50 percent better than the median multilingual performance of 0.1843 of CLEF-2000. This argues that significant progress has been made by the CLEF community in terms of European cross-language retrieval performance.

Run ID	BKMUEAA2	BKMUEAA1	BKMUGAM1	BKMURAA1
Retrieved	50000	50000	50000	50000
Relevant	8138	8138	8138	8138
Rel. Ret	5520	5190	5223	4202
Precision				
at 0.00	0.8890	0.8198	0.8522	0.7698
at 0.10	0.6315	0.5708	0.6058	0.4525
at 0.20	0.5141	0.4703	0.5143	0.3381
at 0.30	0.4441	0.3892	0.4137	0.2643
at 0.40	0.3653	0.3061	0.3324	0.1796
at 0.50	0.2950	0.2476	0.2697	0.1443
at 0.60	0.2244	0.1736	0.2033	0.0933
at 0.70	0.1502	0.1110	0.1281	0.0556
at 0.80	0.0894	0.0620	0.0806	0.0319
at 0.90	0.0457	0.0440	0.0315	0.0058
at 1.00	0.0022	0.0029	0.0026	0.0005
Brk. Prec.	0.3101	0.2674	0.2902	0.1838
CLEF Prec.	0.2749	0.2749	0.2749	0.2749
2000 Prec.	0.1843	0.1843	0.1843	0.1843

Table 3. Results of four official CLEF-2001 multilingual runs.

4 GIRT retrieval

The special emphasis of our current funding has focussed upon retrieval of specialized domain documents which have been assigned individual classification identifiers by human indexers. These classification identifiers can come from thesauri. Since many millions of dollars are expended on developing such classification schemes and using them to index documents, it is natural to attempt to exploit the resources to the fullest extent possible to improve retrieval. In some cases such thesauri are developed with identifiers translated (or provided) in multiple languages, and can thus be used to transfer words across the language barrier.

The GIRT collection consists of reports and papers (grey literature) in the social science domain. The collection is managed and indexed by the GESIS organization (http://www.social-science-gesis.de). GIRT is an excellent example of a collection indexed by a multilingual thesaurus, originally German-English, recently translated into Russian. The GIRT multilingual thesaurus (German-English), which is based on the Thesaurus for the Social Sciences [2], provides the vocabulary source for the indexing terms within the GIRT collection of CLEF. Further information about GIRT can be found in [7] There are 76,128 German documents in GIRT subtask collection. Almost all the documents contain manually assigned thesaurus terms. On average, there are about 10 thesaurus terms assigned to each document. Figure 1 is an example of a thesaurus entry. Since transliteration of the Cyrillic alphabet is a key part of our retrieval strategy, we have transliterated all Russian thesaurus entries.

```
- <entry>
    <german>regionale Wirtschaftspolitik</german>
    <russian>региональная экономическая политика</russian>
    <translit>regional'naia ekonomicheskaia politika</translit>
  </entry>
```

Fig. 1. GIRT German-Russian Thesaurus Entry with Transliteration

In our experiments, we indexed the TITLE and TEXT sections in each document (not the E-TITLE or E-TEXT). The CLEF rules specified that indexing any other field would need to be declared a manual run. For our CLEF runs this year we again used the Muscat stemmer, which is similar to the Porter stemmer but for the German language.

4.1 Query translation from Russian to German

In order to prepare for query translation, we first extracted all the single words and bigrams from the Russian topic fields. Since we do not have a Russian POS tagger, we took any two adjacent words (overlapping word bigram) to be considered a potential phrase. The single words and bigrams in each Russian query were then compared against the Russian-German thesaurus. If a word or bigram was found in the thesaurus, its German translation was added to the new German query being created. The resulting German query was then run against the German collection to retrieve relevant documents.

For comparison, we also used an online MT system
(Promt-Reverso: http://translation2.paralink.com/)
to translate the Russian queries to German.

Fuzzy Matching for the Thesaurus The first approach to thesaurus-based translation was exact matching for thesaurus lookup. From the 25 GIRT topics we obtain about 1300 Russian query terms (words and bigrams). Only 50 of them were directly found in the thesaurus, and these were all single words. Two problems contribute to the low matching rate:

First, a Russian word may have several forms or variations. Usually only the base form or general form appears in the thesaurus. For example, "evropa" (Europe in English) is in the thesaurus, but "evrope" and "evropu" are not. In this case, a Russian morphological analyzer would be helpful. Since we do not have a Russian morphological analyzer, we used fuzzy matching to address this problem.

There are different kinds of algorithms for fuzzy matching, such as Levenshtein distance, common n-grams, longest common subsequence, etc [9]. We found that the simple common bigram algorithm to be very efficient and effective for matching different word forms. The two strings are divided into their constituent bigrams and Dice's coefficient is used to compute the similarity between the two strings.

Original Russian word	Russian word in the thesaurus	German translation
migratsiiu	migratsiia	wanderung
migratsii	migratsiia	wanderung
bezrabotitsei	bezrabotitsa	arbeitslosigkeit
televideniia	televidenie	fernsehen
kul'turu	kul'tura	kultur
kul'turoi	kul'tura	kultur
tekhnologiei	tekhnologiia	technologie
tekhnologii	tekhnologiia	technologie

Above are some examples of Russian words that do not occur in the thesaurus but whose different forms were found in the thesaurus by fuzzy matching (the Russian characters in the examples are transliterated for easy reading).

The second problem lies in finding query bigrams which do not match exactly to thesaurus entries. Fuzzy matching was also useful for finding different forms of bigrams, even in cases where word order is changed, examples are:

Original Russian bigram	Bigram found in the thesaurus	German translation
tekhnologicheskogo razvitiia	tekhnologicheskoe razvitie	technologische entwicklung
razvitie i organizatsiia	organizatsionnoe razvitie	organisationsentwicklung
upravlenie organizatsiia	organizatsiia upravleniia	verwaltungsorganisation
rabochem meste	rabochee mesto	arbeitsplatz
rukovodiashchikh rabotnikov	rukovodiashchie rabotniki	führungskraft

The way bigram 'phrases' were created had two problems: first, many of the bigrams were simply not meaningful; second, even though most genuine phrases contain two words (bigrams), approximately 25 percent of Russian terms in the thesaurus contain 3 or more words. A Russian POS tagger would be very helpful for finding meaningful or long phrases.

4.2 GIRT results and analysis

Our GIRT results are summarized in Table 4. The runs can be described as follows: BKGRGGA is a monolingual run using the German version of the topics run against the German GIRT document collection. BKGRRGA1 is a Russian-German bilingual run using MT system for query translation. BKGRRGA2 is a Russian-German bilingual run using thesaurus lookup and fuzzy matching. BKGRRGA3 is a Russian-German bilingual run which is identical in methodology to BKGRRGA2 except that only title and description sections of the topic were used for matching. As we can see, our Russian runs achieve only about 1/3 of the precision of the German-German monolingual run, with a significant edge to the machine translation version. While the full narrative run BKGRRGA3

Run ID	BKGRGGA	BKGRRGA1	BKGRRGA2	BKGRRGA3
Retrieved	25000	25000	25000	25000
Relevant	1111	1111	1111	1111
Rel. Ret	1054	774	781	813
Precision				
at 0.00	0.9390	0.4408	0.4032	0.3793
at 0.10	0.8225	0.3461	0.3018	0.2640
at 0.20	0.7501	0.2993	0.2441	0.2381
at 0.30	0.6282	0.2671	0.2257	0.1984
at 0.40	0.5676	0.2381	0.1790	0.1782
at 0.50	0.5166	0.1999	0.1341	0.1609
at 0.60	0.4604	0.1400	0.0993	0.1323
at 0.70	0.4002	0.1045	0.0712	0.1125
at 0.80	0.3038	0.0693	0.0502	0.0765
at 0.90	0.2097	0.0454	0.0232	0.0273
at 1.00	0.0620	0.0051	0.0013	0.0000
Brk. Prec.	0.5002	0.1845	0.1448	0.1461

Table 4. Results of official GIRT Russian-German runs.

retrieved more relevant documents, this did not translate into higher overall precision.

5 Summary and Acknowlegments

The participation of Berkeley's Group One in CLEF-2001 has enabled us to explore the difficulties in extending cross-language information retrieval to a non-Roman alphabet language, Russian, for which limited resources are available. Specifically we have explored a comparison of bilingual and multilingual retrieval where original queries were in Russian when compared against German as a query language or English as a query language for multilingual retrieval. For the GIRT task we compared various forms of Russian⟶German retrieval. We have determined that there is significant work to be done before cross-language information retrieval from the Russian language will become competitive to other European languages.

This research was supported by DARPA (Department of Defense Advanced Research Projects Agency) under contract N66001-97-8541; AO# F477: Search Support for Unfamiliar Metadata Vocabularies within the DARPA Information Technology Office. We thank Aitao Chen for indexing the main CLEF collections.

References

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

2. Hannelore Schott (ed.). *Thesaurus for the Social Sciences. [Vol. 1:] German-English. [Vol. 2:] English-German. [Edition] 1999.* InformationsZentrum Sozialwissenschaften Bonn, 2000.
3. A Chen J He L Xu F Gey and J Meggs. Chinese text retrieval without using a dictionary. In A. Desai Narasimhalu Nicholas J. Belkin and Peter Willett, editors, *Proceedings of the 20th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval, Philadelphia*, pages 42–49, 1997.
4. F. C. Gey and A. Chen. Phrase discovery for english and cross-language retrieval at trec-6. In D. K. Harman and Ellen Voorhees, editors, *The Sixth Text REtrieval Conference (TREC-6), NIST Special Publication 500-240*, pages 637–647, August 1998.
5. F. C. Gey and H. Jiang. English-german cross-language retrieval for the girt collection – exploiting a multilingual thesaurus. In Ellen Voorhees, editor, *The Eighth Text REtrieval Conference (TREC-8), draft notebook proceedings*, pages 219–234, November 1999.
6. A Chen H Jiang and F Gey. Berkeley at ntcir-2: Chinese, japanese and english ir experiments. In N. Kando, editor, *Proceedings of the Second NTCIR Workshop on Evaluation of Chinese and Japanese Text Retrieval and Summarization, Tokoyo Japan*, pages 32–39, March 2001.
7. Michael Kluck and Fredric Gey. The domain-specific task of clef - specific evaluation strategies incross-language information retrieval. In *Cross Language Retrieval Evaluation, Proceedings of the CLEF 2000 Workshop*. Springer, 2001.
8. F Gey M Buckland R Larson and A Chen. Entry vocabulary – a technology to enhance digital search. In *Proceedings of the First International Conference on Human Language Technology*, March 2001.
9. Michael P. Oakes. *Statistics for Corpus Linguistics*. Edinburgh University Press, 1998.
10. F Gey H Jiang V Petras and A Chen. Cross-language retrieval for the clef collections – comparing multiple methods of retrieval. In Carol Peters, editor, *Cross Language Retrieval Evaluation, Proceedings of the CLEF 2000 Workshop*. Springer, 2001.