UC Berkeley at CLEF 2003 – Russian Language Experiments and Domain-Specific Cross-Language Retrieval

Vivien Petras[1], Natalia Perelman[1] and Fredric Gey[2]

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

ABSTRACT. As in the previous years, Berkeley's group 1 experimented with the domain-specific CLEF collection GIRT as well as with Russian as query and document language. The GIRT collection was substantially extended this year and we were able to improve our retrieval results for the query languages German, English and Russian. For the GIRT retrieval experiments, we utilized our previous experiences by combining different translations, thesaurus matching, decompounding for German compounds and a blind feedback algorithm. We find that our thesaurus matching technique compares to conventional machine translation for Russian and German against English retrieval and outperforms machine translation for English to German retrieval.
With the introduction in CLEF 2003 of a Russian document collection, we participated in the CLEF main task with monolingual and bilingual runs for the Russian collection. For bilingual retrieval our approaches were query translation (for German or English as topic languages) and 'fast' document translation (for English as the topic language). Document translation significantly underperformed query translation (using the PROMPT translation system).

1 Introduction

For several years, Berkeley's group 1 has experimented with domain-specific collections and investigated thesaurus-aided retrieval within the CLEF environment. We theorize that collections enhanced with subject terms from a controlled vocabulary contain more query-relevant words and phrases and, furthermore, that retrieval using a thesaurus-enhanced collection and / or queries enriched with controlled vocabulary terms will be more precise. This year's GIRT collection has been extended to contain more than 150,000 documents (as opposed to the 70,000 documents it contained in the previous years) and we investigated the usefulness of a thesaurus in a bigger document collection. The larger a document collection is, the more individual documents can be found for any chosen controlled vocabulary term. In a worst-case scenario, this effect could nullify the specificity of the thesaurus terms and have a negative outcome on the retrieval performance. However, our experiments show that incorporating the thesaurus data will achieve performance improvements. Using the multilingual GIRT thesaurus (German, English, Russian) to translate query files for bilingual retrieval has proven to be useful for performance improvement. Our so-called thesaurus matching technique is comparable to machine translation for Russian and German, but outperforms the tested machine translation systems for English to German. However, the competitiveness of thesaurus matching versus machine translation depends on the existence of controlled vocabulary terms in the query fields and the size and quality of the thesaurus.

CLEF 2003 was the first time a Russian language document collection was available in CLEF. We have worked for several years with Russian topics in both the GIRT task and the CLEF main tasks, so we welcomed the opportunity to do Russian monolingual retrieval and bilingual retrieval No unusual methodology was applied to the Russian collection, however encoding was an issue and we ended up using the KOI-8 encoding scheme for both documents and topics.

For our retrieval experiments, the Berkeley group is using the technique of logistical regression as described in [1].

2 The GIRT Retrieval Experiments

2.1 The GIRT collection

The GIRT collection (German Indexing and Retrieval Testdatabase) consists of 151,319 documents in the social science domain. The documents contain titles, abstracts and controlled vocabulary terms describing reports and papers indexed by the GESIS organization (http://www.social-science-gesis.de). The GIRT controlled vocabulary terms are based on the Thesaurus for the Social Sciences [2] and are provided in German and English. A German-Russian translation table is also provided. For the 2003 CLEF experiments, two parallel GIRT corpora were made available: (1) German GIRT 4 contains document fields with German text, and (2) English GIRT 4 contains the translations of these fields into English.

This year, we carried out the monolingual task in both the German and English corpus, testing which parts of the document (title, abstract, or thesaurus terms) will provide relevant input for retrieval. We also experimented with the bilingual task by using German, English and Russian as query languages against both corpora.

For all runs against the German collection, we used our decompounding procedure to split German compound words into individual terms. The procedure is described in [3] and [4]. All runs used only title and description fields from the topics. Additionally, we used our blind feedback algorithm for all runs to improve performance. The blind feedback algorithm assumes the top 20 documents as relevant and selects 30 terms from these documents to add to the query. From our experience, using the decompounding procedure and our blind feedback algorithm increases the performance anywhere between 10 and 30%. The run BKGRMLGG1 (Table 1) for example, which reached an average precision of 0.4965 in the official run, would have yielded only 0.3288 average precision without decompounding and blind feedback.

2.2 GIRT Monolingual Retrieval

For the GIRT monolingual task, we performed two experiments for each of the German and English corpora: a monolingual run against an index containing all document fields and a monolingual run against an index without the controlled vocabulary fields. As was expected, the runs against the indexes containing all fields yielded better retrieval results than the runs against the smaller indexes. For comparison purposes, we also constructed two additional indexes containing only the controlled vocabulary terms and the controlled vocabulary terms and the titles respectively. The results for the German and English monolingual runs can be found in tables 1 and 2.

Run Name	BKGRMLGG1	BKGRMLGG2	BKGRMLGG3	BKGRMLGG4
Document Fields	All	Title, Abstract	Title, Thesaurus	Thesaurus
Retrieved	25000	25000	25000	25000
Relevant	2117	2117	2117	2117
Rel Ret	1860	1767	1624	1474
Precision				
at 0.00	0.9416	0.9270	0.8900	0.8053
at 0.10	0.8042	0.7668	0.7001	0.6194
at 0.20	0.7096	0.6938	0.6110	0.5544
at 0.30	0.6682	0.5827	0.5320	0.3800
at 0.40	0.5810	0.5140	0.4262	0.3337
at 0.50	0.5242	0.4287	0.3345	0.2797
at 0.60	0.4359	0.3385	0.2708	0.2239
at 0.70	0.3784	0.2607	0.1949	0.1555
at 0.80	0.2970	0.1527	0.1041	0.0827
at 0.90	0.1764	0.0916	0.0436	0.0301
at 1.00	0.0528	0.0362	0.0010	0.0010
Avg. Precision	0.4965	0.4199	0.3530	0.2935

Table 1. Monolingual runs on the German GIRT 4 corpus. Official runs are BKGRMLGG1 and BKGRMLGG2.

Judging from these results, the controlled vocabulary terms have a positive impact on the retrieval results, but not as big as the abstract. Runs without the thesaurus terms lose only about 16% of their average precision, whereas runs without the abstract lose about 29%. An index that only contains titles would only yield a performance of 0.1820 in average precision, which confirms the theory that most titles are not as expressive of an article's content as the controlled vocabulary terms or the abstract.
Comparing these results to last year's, the bigger collection size might have an impact. Last year, the indexes with title and abstract and title and thesaurus terms yielded about the same results. Both were about 23% worse than the general index containing all fields. This could mean that the thesaurus terms in the bigger collection do not have as much expressive power and are not as discriminating as they are in a smaller collection. However, the results can also be explained by other influences: (i) the queries contain less terms found in the thesaurus, (ii) the abstracts are more expressive, (iii) there were less controlled vocabulary terms assigned to each document.

Run Name	BKGRMLEE1	BKGRMLEE2	BKGRMLEE3	BKGRMLEE4
Document Fields	All	Title, Abstract	Title, Thesaurus	Thesaurus
Retrieved	25000	25000	25000	25000
Relevant	1332	1332	1332	1332
Rel Ret	1214	763	1160	1092
Precision				
at 0.00	0.9373	0.7762	0.9464	0.6722
at 0.10	0.7739	0.5993	0.8086	0.5373
at 0.20	0.7220	0.4648	0.6701	0.4557
at 0.30	0.6462	0.3700	0.5998	0.4111
at 0.40	0.5863	0.2684	0.5449	0.3733
at 0.50	0.5324	0.1909	0.4751	0.3376
at 0.60	0.4734	0.1343	0.4225	0.2822
at 0.70	0.4127	0.0774	0.3714	0.2320
at 0.80	0.3437	0.0477	0.3062	0.1820
at 0.90	0.2508	0.0415	0.2033	0.0877
at 1.00	0.0846	0.0414	0.0570	0.0458
Avg. Precision	0.5192	0.2484	0.4853	0.3207

Table 2. Monolingual runs against the English GIRT 4 corpus. Official runs are BKGRMLEE1 and BKGRMLEE2.

For the English GIRT corpus, the results seem to be quite different. Here the index with only title and thesaurus term fields yields almost as good a result as the general index. The index without the thesaurus terms shows a performance only half as good as the general index. However, this result can probably be explained by the fact that there are far fewer abstracts in the English GIRT corpus than there are controlled vocabulary terms. The title and thesaurus terms seem to bear the brunt of the retrieval effort in this collection.

2.3 GIRT Bilingual Retrieval

We submitted 5 official runs for the GIRT bilingual task and used all query languages (German, English and Russian) available. Generally, the runs against the English GIRT collection (with translated query files from German and Russian) yielded better results than the runs against the German GIRT collection. This can be most probably attributed to the better quality of machine translation systems for the English language as opposed to the German language. However, there does not seem to be a high variation in the results between the Russian and German / English query languages, which points to a rapid improvement in the machine translation for Russian, which can be seen in the definite increase of precision figures as compared to the detrimental results of last year.

We used two machine translation systems for each query language: L & H Power Translator and Systran for German and English; and Promt and Systran for the Russian language. We also used our thesaurus matching as one translation technique [5], which will be further discussed in part 2.4.

For thesaurus matching, we identify phrases and terms from the topics files and search them against the thesaurus. Once we find an appropriate thesaurus term, we substitute the query term or phrase with the thesaurus term in the language used for retrieval.

The results for the bilingual runs against German and English and a comparison of the different translation techniques can be found in tables 3 & 4 for Russian to German and English to German respectively and table 5 & 6 for Russian and German to English respectively. All runs are against the full indexes containing all document fields.

Run Name	BKGRBLRG3	BKGRBLRG4	BKGRBLRG1	BKGRBLRG5	BKGRBLRG2
Transl. Technique	Systran	Promt	Sys + Promt	Thes. Matching	Sys + Promt + Thes.
Retrieved	25000	25000	25000	25000	25000
Relevant	2117	2117	2117	2117	2117
Rel Ret	1264	1555	1547	1343	1577
Precision					
at 0.00	0.5301	0.6674	0.7035	0.5067	0.7281
at 0.10	0.3381	0.5086	0.5756	0.3816	0.5994
at 0.20	0.2698	0.4256	0.4796	0.3239	0.5121
at 0.30	0.2472	0.3750	0.4261	0.2873	0.4591
at 0.40	0.2271	0.3275	0.3711	0.2480	0.3966
at 0.50	0.1960	0.2880	0.3301	0.1950	0.3395
at 0.60	0.1598	0.2257	0.2570	0.1487	0.2690
at 0.70	0.1262	0.1864	0.1946	0.0997	0.2072
at 0.80	0.0987	0.1380	0.1383	0.0769	0.1346
at 0.90	0.0657	0.0654	0.0713	0.0393	0.0741
at 1.00	0.0154	0.0116	0.0103	0.0027	0.0009
Avg. Precision	0.1925	0.2798	0.3117	0.1983	0.3269

Table 3. Bilingual Russian runs against the German GIRT 4 corpus. Official runs are BKGRBLRG1 and BKGRBLRG2.

From the Russian runs against the German GIRT corpus, one can see the superior quality of the Promt translator (about 30% better results than the Systran Babelfish translating system). The Systran system is also handicapped in that it has no direct translation from Russian to German. English was used as a Pivot language and could have introduced additional errors or ambiguities. Nevertheless, a combination of both translating systems reaches an improvement in overall precision, but not in recall.

Our thesaurus matching technique – although with a much more restricted vocabulary – compares with the Systran translator in precision and reaches a better recall. This can be explained with the superior quality (in terms of relevance for retrieval) of the thesaurus terms in a search statement. Whereas in last year's experiment the combination of translation and thesaurus matching achieved a performance improvement of 30%, this year the combination achieves only marginal improvements in precision and recall. This can mostly be explained with the improved quality of the machine translation system Promt, so that our thesaurus matching technique does not add as many high-quality terms to the query as it did last year.

Run Name	BKGRBLEG2	BKGRBLEG3	BKGRBLEG1	BKGRBLEG4	BKGRBLEG5
Transl. Technique	L+H Power	Systran	Sys + L+H	Thes. Matching	L+H + Thes.
Retrieved	25000	25000	25000	25000	25000
Relevant	2117	2117	2117	2117	2117
Rel Ret	1656	1488	1672	1712	1803
Precision					
at 0.00	0.7719	0.6633	0.7184	0.8823	0.9257
at 0.10	0.6607	0.5201	0.6425	0.7505	0.8118
at 0.20	0.5757	0.4348	0.5727	0.6004	0.6898
at 0.30	0.5114	0.3841	0.5300	0.5510	0.6132
at 0.40	0.4569	0.3548	0.4568	0.5089	0.5606
at 0.50	0.3996	0.3084	0.3747	0.4475	0.4787
at 0.60	0.3471	0.2608	0.3109	0.3869	0.3832
at 0.70	0.3002	0.2082	0.2363	0.3132	0.3197
at 0.80	0.2159	0.1686	0.1781	0.2477	0.2421
at 0.90	0.1141	0.0985	0.0918	0.1259	0.1374
at 1.00	0.0189	0.0117	0.0099	0.0196	0.0121
Avg. Precision	0.3886	0.3001	0.3669	0.4299	0.4606

Table 4. Bilingual English runs against the German GIRT 4 corpus. Official run is BKGRBLEG1.

For English to German retrieval, the L+H Power Translator system reaches much better results in retrieval than Systran, so that the combination of both translations actually degraded the retrieval performance of the overall run (although recall increased slightly).

Two queries negatively impacted the retrieval results using machine translation: 94 (Homosexuality and Coming-Out) and 98 (Canadian Foreign Policy). Both were caused by wrong translations of critical search words. “Coming-Out” for query 94 was translated into “Herauskommen” (a direct translation of the English phrases), although the phrase remains as is in German as a borrowed construct. Query 98 contains the phrase “foreign policy”, which was translated into “fremde Politik”, a common mistake in word-for-word translation systems. Although “foreign” is most commonly translated with “fremd”, in the phrase “foreign policy” it should become the compound “Aussenpolitik” – an error that dropped this query’s precision to 0.0039. However, the phrase “foreign policy” is a controlled vocabulary term and was therefore correctly translated using our thesaurus matching technique. Using thesaurus matching improved this query’s average precision to 0.3798.

For English to German retrieval, thesaurus matching proved to be most effective; this run outperformed the best machine translation run by roughly 10%. Combining machine translations and translations using our thesaurus matching improves performance even more: the BKGRBLEG5 run outperformed the best machine translation run by 18%.

Run Name	BKGRBLRE2	BKGRBLRE3	BKGRBLRE1	BKGRBLRE4	BKGRBLRE5
Transl. Technique	Systran	Promt	Sys + Promt	Thes. Matching	Promt + Thes.
Retrieved	25000	25000	25000	25000	25000
Relevant	1332	1332	1332	1332	1332
Rel Ret	997	1084	1042	935	1077
Precision					
at 0.00	0.7193	0.7922	0.7696	0.7309	0.8350
at 0.10	0.5466	0.6579	0.6251	0.5206	0.7103
at 0.20	0.4931	0.6165	0.5856	0.4443	0.6281
at 0.30	0.4559	0.5671	0.5522	0.4120	0.5638

at 0.40	0.3802	0.4980	0.4770	0.3632	0.5289
at 0.50	0.3451	0.4479	0.4459	0.3146	0.4808
at 0.60	0.2927	0.3776	0.3867	0.2691	0.4194
at 0.70	0.2450	0.3338	0.3269	0.2211	0.3633
at 0.80	0.2059	0.2683	0.2637	0.1785	0.2981
at 0.90	0.1562	0.2020	0.1750	0.1191	0.1937
at 1.00	0.0604	0.0638	0.0573	0.0572	0.0769
Avg. Precision	0.3420	0.4258	0.4111	0.3107	0.4524

Table 5. Bilingual Russian runs against the English GIRT 4 corpus. Official run is BKGRBLRE1.

Also for Russian to English retrieval, the Promt translator shows superior quality – even better than for Russian to German. It outperforms the Systran translator in a way that a combination of the translations actually proves to be disadvantageous to the retrieval outcome.
Our thesaurus matching run yields the worst results of all runs – this is partly due to the fact that there is no direct mapping table between the Russian and English thesaurus version so that German had to be used as a pivot language. In the process of mapping the Russian queries to the German and then English thesaurus versions, information was lost and consequently two queries (93 & 95) could not be effectively translated and no documents were retrieved from the English collection.
Nevertheless, a translation using thesaurus matching adds new and relevant search terms to some queries so that a combination of machine translation plus thesaurus matching translation slightly outperformed the best machine translation run by 6%.

Run Name	BKGRBLGE2	BKGRBLGE3	BKGRBLGE1	BKGRBLGE4	BKGRBLGE5
Transl. Technique	L+H Power	Systran	Sys + L+H	Thes. Matching	L+H + Thes.
Retrieved	25000	25000	25000	25000	25000
Relevant	1332	1332	1332	1332	1332
Rel Ret	1067	1116	1121	1074	1197
Precision					
at 0.00	0.7505	0.8080	0.8338	0.7401	0.8298
at 0.10	0.6486	0.6089	0.6630	0.6054	0.7388
at 0.20	0.5921	0.5145	0.5707	0.5267	0.6933
at 0.30	0.4958	0.4549	0.5185	0.4921	0.6020
at 0.40	0.4354	0.3921	0.4777	0.4466	0.5465
at 0.50	0.3924	0.3573	0.4001	0.4204	0.4810
at 0.60	0.3482	0.3179	0.3591	0.3702	0.4306
at 0.70	0.2965	0.2837	0.2971	0.3155	0.3701
at 0.80	0.2561	0.2451	0.2401	0.2738	0.2995
at 0.90	0.2036	0.2076	0.1816	0.2091	0.2267
at 1.00	0.0830	0.0722	0.0808	0.0917	0.0954
Avg. Precision	0.4022	0.3748	0.4068	0.3977	0.4731

Table 6. Bilingual German runs against the English GIRT 4. Official run is BKGRBLGE1.

Once again, the L+H Power translator outperforms the Systran translator also when it comes to the opposite direction of English to German retrieval. However, a combination of the two MT systems marginally outperforms L+H in precision and makes an impact on recall.
Thesaurus matching from German to English reaches a result similar to any of the machine translations systems but the combination of the L+H Power translation and our translation from thesaurus matching achieves a performance improvement of 17%.

2.4 The Effectiveness of Thesaurus Matching

Thesaurus matching is a translation technique where the system relies exclusively on the vocabulary of the thesaurus to provide a translation. The topic files are searched for terms and phrases that occur in the thesaurus and are then substituted by their foreign language counterparts. A more detailed description can be found in [5]. Due to this process, the translated query consists of controlled vocabulary terms in the appropriate language and untranslated words that were not found in the thesaurus.
This has the advantage of emphasizing highly relevant search terms (which will occur in the thesaurus term fields of the relevant documents) but also has a major drawback. The technique will only work when the queries contain enough words and phrases that occur in the multilingual thesaurus and when those terms and phrases represent the meaning of the search statement. Fortunately, almost all queries contain more than one term that can be found in the thesaurus and therefore translated.
Nevertheless, most of the variation in our retrieval results (comparing query by query to the machine translation results) can be accounted for by looking at which queries contain the most thesaurus terms and how many good phrases our algorithm can detect. A large general thesaurus should be able to provide a good translation approximation but specialized thesauri with highly technical vocabulary might not fare as well. However, depending on the nature of the query, specialized thesauri could help in identifying important search terms from a search statement. Additionally, our thesaurus matching technique might be able to improve: (i) by allowing a better fuzzy match between query terms and thesaurus terms, (ii) by incorporating partial matching of query terms to thesaurus terms, (iii) by exploiting narrower and broader term relationships in the thesaurus when expanding the query, or (iv) by exploiting the use-instead and used-for relationships in the thesaurus (which we have ignored so far).
Further experiments should show whether our thesaurus matching technique can improve and – considering that its competitive advantage over the three investigated MT systems lies in its ability to translate phrases - whether it can compete against phrase dictionaries as well.

3 Russian Retrieval for the CLEF main task

CLEF 2003 marked the first time a document collection has been available and evaluated in the Russian language. The CLEF Russian collection consisted of 16,716 articles from *Izvestia* newspaper from 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). There were 37 Russian topics, which were chosen by the organizers from the 60 topics of the CLEF main multilingual task. In our bilingual retrieval we worked with English and German versions of these topics.

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. 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 and German also would come in that encoding.

3.2 Russian Monolingual Retrieval

We submitted four Russian monolingual runs, the results of which are summarized below. All runs utilized blind feedback, choosing the top 30 terms from the top ranked 20 documents of an initial retrieval run. This was the same methodology used above in the GIRT retrieval. For BKRUMLRR1 and BKRUMLRR2 runs we used TITLE and TEXT document fields for indexing. BKRUMLRR3 and BKRUMLRR4 were run against an index containing TITLE, TEXT, SUBJECT, GEOGRAPHY, and RETRO fields.

The results of our retrieval are summarized in Table 7. Results were reported by the CLEF organizers for 28 topics which had one or more relevant documents.

Run Name	BKRUMLRR1	BKRUMLRR2	BKRUMLRR3	BKRUMLRR4
Index	Koi	Koi	Koi-all	Koi-all
Topic fields	TD	TDN	TD	TDN
Retrieved	28000	28000	28000	28000
Relevant	151	151	151	151
Rel Ret	125	127	146	148
Precision				
at 0.00	0.5201	0.6626	0.5311	0.6503
at 0.10	0.5201	0.6626	0.5311	0.6503
at 0.20	0.4844	0.5750	0.5278	0.6208
at 0.30	0.4777	0.5409	0.5047	0.5597
at 0.40	0.4309	0.4370	0.4554	0.5009
at 0.50	0.4007	0.3992	0.4375	0.4522
at 0.60	0.3087	0.2873	0.3448	0.3699
at 0.70	0.2382	0.2368	0.3107	0.3401
at 0.80	0.1637	0.1612	0.2965	0.3093
at 0.90	0.1210	0.1206	0.2535	0.2641
at 1.00	0.1210	0.1206	0.2392	0.2471
Avg. Precision	0.3338	0.3655	0.3878	0.4395

Table 7: Berkeley Monolingual Russian runs for CLEF 2003.

3.3 Russian Bilingual Retrieval

We submitted six bilingual runs against the Russian document collection. These runs only indexed the TITLE and TEXT fields of each Russian document, so are directly comparable only to the monolingual runs BKMLRURR1 and BKMLRURR2 above. Four of these runs (BKRUBLGR1, BKRUBLGR2, BKRUBLER1, BKRUBLER2) utilized query translation from either German or English topics into Russian.

Run Name	BKRUBLGR1	BKRUBLGR2	BKRUBLER1	BKRUBLER2	BKRUMLEE1	BKRUMLEE2
Language	German	German	English	English	En	En
Topic fields	TD	TDN	TD	TDN	TD	TDN
Retrieved	28000	28000	28000	28000	28000	28000
Relevant	151	151	151	151	151	151
Rel Ret	121	122	125	126	119	121
Precision						
at 0.00	0.4672	0.5219	0.5119	0.5953	0.2821	0.3809
at 0.10	0.4613	0.5058	0.5119	0.5953	0.2761	0.3764
at 0.20	0.4209	0.4531	0.4560	0.5224	0.2567	0.3645
at 0.30	0.4066	0.4348	0.4408	0.4978	0.2468	0.3535
at 0.40	0.3603	0.4025	0.3416	0.4833	0.1811	0.2881
at 0.50	0.3258	0.3687	0.3031	0.4428	0.1719	0.2644
at 0.60	0.2458	0.2755	0.1978	0.3155	0.1516	0.2049
at 0.70	0.1637	0.1886	0.1440	0.1618	0.1038	0.1513
at 0.80	0.1365	0.1632	0.1155	0.1336	0.0735	0.0949
at 0.90	0.0953	0.1161	0.0726	0.0825	0.0491	0.0756
at 1.00	0.0953	0.1161	0.0726	0.0825	0.0491	0.0756
Avg. Prec.	0.2809	0.3125	0.2766	0.3478	0.1604	0.2227

Table 8. Bilingual Russian runs.

Translation to Russian was done using the PROMPT online translation facility at http://www.translate.ru The only difference between runs numbered one and two was the addition of the narrative field in topic indexing.

Two final runs (BKRUMLEE1 and BKRUMLEE2) utilized a technique developed by Aitao Chen, called 'Fast Document Translation' [6]. 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. Using this technique will choose only one translation for a polysemous word, but this defect is compensated by extremely fast translations of the all the documents into the target language. We submitted 246.252 unique Russian words from the Izvestia collection to the PROMPT translation system (this was done 5,000 words at a time) for translation to English and then used this to translate all the documents into English. Monolingual retrieval was performed by matching the English versions of the topics against the translated English document collection.

3.5. Brief Analysis of Russian Retrieval Performance

Bilingual retrieval was in all cases worse than monolingual (Russian-Russian) retrieval in terms of overall precision. German→Russian retrieval was comparable to English→Russian retrieval for TD runs, but the English→Russian TDN run was substantially better than its German→Russian counterpart. Speculation (without evidence) is that de-compounding the German narrative before translation would have improved the performance. Fast document translation runs significantly underperformed query translation runs, which differs from experiments with other languages; we are investigating why this is the case.

Because of the nature of the retrieval results by query for the Russian collection (eleven of the 28 topics have 2 or fewer relevant documents) one has to be cautious about drawing conclusions from the results. In general, monolingual retrieval substantially outperformed bilingual retrieval over almost all topics. However, for Topic 169 the bilingual retrieval is much better (best precision 1.0 for German-to-Russian) than the monolingual, with the best run being German-to-Russian where the German topic contains the words CD-Brennern which translates to laser disc (лазерного диска) and music industry (Musikindustrie → музыкальной индустрии) instead of the use, in the Russian version of topic 169, of the words компакт-дисков (compact disk) and аудио-промышленности (audio industry) which aren't very discriminating. The German→Russian retrieval for Topic 187 (with one relevant document) fell victim to translation problems: "Radioactive waste" in English is expressed in German as " radioaktivem Müll". The English "waste" is translated correctly as "отходы" while the German "Müll" is translated as "мусору," or "garbage". This and other differences in translation lead to a decrease from 1.0 precision for English bilingual to 0.25 for German bilingual for topic 187. Several other topics have the same disparity of translation. We merged the document rankings from German and English bilingual runs using un-normalized retrieval status value – the resulting ranked list showed no significant improvement in performance. It would be useful to try merging the topic translations (adjusting for word count weights) before retrieval.

4 Summary and Acknowledgments

Berkeley's group 1 participated in the CLEF GIRT tasks and CLEF Main tasks for Russian mono- and bilingual retrieval. We experimented with German, English and Russian as collection and query languages.

Within the GIRT domain-specific collection, we investigated the use of thesauri in document retrieval, document index enhancement and query translation. Documents that have controlled vocabulary terms added to the usual title and abstract information prove advantageous in retrieval because the thesaurus terms add valuable search terms to the index. An index containing titles, abstracts and thesaurus terms will always outperform an index only containing title and abstract. However, the theory that thesaurus terms might be able to substitute abstracts because of their specific nature was premature. Retrieval involving thesauri can be influenced by several factors: the size of the collection, the size of the controlled vocabulary and the nature of the queries.

For topic translations, we found that although a combination of different machine translation systems might not always outperform an individual machine translation system, a combination of a machine translation system and our thesaurus matching technique does. Thesaurus matching outperformed machine translation in English to German retrieval and added new and relevant search terms for all other query languages. For German and Russian queries, thesaurus matching yielded comparable results to machine translation.

We experimented with the CLEF 2003 Russian document collection with both monolingual Russian and bilingual to Russian from German and English topics. In addition to query translation methodology for bilingual retrieval, we tried a fast document translation method to English and performed English-English monolingual retrieval, which did not perform as well as query translation.

We would like to thank Aitao Chen for supplying his German decompounding software and for performing the fast document translation from Russian to English. This research was supported in part by DARPA (Department of Defense Advanced Research Projects Agency) under research grant N66001-00-1-8911: Translingual Information Detection Extraction and Summarization (TIDES) within the DARPA Information Technology Office.

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] H. Schott (Ed.). **Thesaurus for the Social Science**. [Vol. 1:] German-English. [Vol. 2:] English-German. Informations-Zentrum Sozialwissenschaften Bonn, 2000

[3] A. Chen. Multilingual information retrieval using english and chinese queries. In C. Peters, M. Braschler, J. Gonzalo, and M. Kluck, editors, *Evaluation of Cross-Language Information Retrieval Systems: Second Workshop of the Cross-Language Evaluation Forum, CLEF-2001*, Darmstadt, Germany, September 2001, pages 44–58. Springer Computer Science Series LNCS 2406, 2002.

[4] A. Chen. Cross-language Retrieval Experiments at CLEF 2002. C. Peters (Ed.). *Working Notes for the CLEF 2002 Workshop* 19-20 September, Rome, Italy, 2002.

[5] V. Petras, N. Perelman and F. Gey. Using Thesauri in Cross-Language Retrieval of German and French Indexed Collections. To appear in: *Proceedings of the CLEF 2002 Workshop*, Springer Computer Science Series.

[6] A. Chen and F. Gey. Multilingual Information Retrieval Using Machine Translation, Relevance Feedback, and Decompounding, To appear in: *Information Retrieval Journal: Special Issue on CLEF*, 2003.