GIRT and the Use of Subject Metadata for Retrieval

Vivien Petras
School of Information Management and Systems
University of California, Berkeley, CA 94720 USA
vivienp@sims.berkeley.edu

Abstract. The use of domain-specific metadata (subject keywords) is tested for monolingual and bilingual retrieval on the GIRT social science collection. A new technique, Entry Vocabulary Modules, which adds subject keywords selected from the controlled vocabulary to the query, has been tested. As in previous years, we compare our techniques of thesaurus matching and Entry Vocabulary Modules to simple machine translation techniques in bilingual retrieval. A combination of machine translation and thesaurus matching achieves better results, whereas the introduction of Entry Vocabulary Modules has negligent impact on the retrieval results. Retrieval results for the German and English GIRT collection for monolingual as well as bilingual retrieval (with English and German as query languages) will be represented.

1 INTRODUCTION

For several years now, the Berkeley group has been interested in how the use of subject metadata (additional to the full text of title and abstract of documents) can improve information retrieval and provide more precise results. For this year's CLEF evaluation, we once again focused on the GIRT collection with its thesaurus-enhanced records, giving us an experimental playing field. We believe that leveraging the high-quality keywords provided by a controlled vocabulary could help in disambiguating the fuzziness of the searcher language and aid searchers in formulating effective queries in order to match relevant documents better.

We are experimenting with a technique called Entry Vocabulary Modules, which suggests subject keywords from the thesaurus when given a natural language query. Like blind feedback terms, these subject keywords are added to the query with the goal of matching the controlled vocabulary terms added to the documents. Using the bilingual feature of the GIRT thesaurus, we substitute suggested thesaurus terms from the Entry Vocabulary Module in the query language with those in the target document language, thereby providing a crude translation mechanism for bilingual retrieval. The improvements over baseline retrieval were minimal, however. A description of the technique is provided in the next section.

Once again, we also tested thesaurus matching for bilingual retrieval against machine translation (described in section 1.2). We report positive results for a combination of thesaurus matching and machine translation.
We have used both the German and English GIRT document collection for monolingual and bilingual retrieval. English and German were used as query languages. All runs are TD (title, description) runs only.

For all retrieval experiments, the Berkeley group is using the technique of logistic regression as described in Chen et al. (1994).

1.1 Entry Vocabulary Modules

Entry Vocabulary Modules (EVMs) are intermediaries between natural language queries and the metadata "language" of a document repository. For a given query, they act as "interpreter" between the searcher and the system, (hopefully) proposing more effective query terms from the controlled vocabulary of the searched documents. The concept of Entry Vocabulary Modules is based on the idea that searching with the "correct" controlled vocabulary terms (i.e. thesaurus terms in the GIRT case) will yield better and more complete results than using any randomly chosen terms in the query. If using an EVM, the searcher is presented with a list of ranked controlled vocabulary terms that the EVM deems appropriate for the query. The searcher can then choose and add or substitute these terms in the query.

An Entry Vocabulary Module is created by building a dictionary of associations between terms and phrases of titles, authors, and / or abstracts of existing documents and the controlled vocabulary. A likelihood ratio statistic is used to measure the association between these and to predict which metadata terms best mirror the topic represented by the searcher's search vocabulary. The methodology of constructing Entry Vocabulary Indexes has been described in detail by Plaunt and Norgard (1998), and Gey et al. (1999).

As the basic technique, a lexical collocation process between document words and controlled vocabulary terms is used. If words co-occur with a higher than random frequency, there exists a likelihood that they are strongly associated. The idea is that the stronger an association between the occurrence of two or more words (document word and controlled vocabulary term), the more likely it is that the collocation is meaningful. If an Entry Vocabulary Module is used to predict metadata vocabulary terms for a document, the association weights for document term and metadata term pairs are combined by adding them. By choosing the highest value of the added weights, the probability of relevance for metadata terms for a whole document can be determined.

For the GIRT experiments, we created an EVM for each of the English and German collections using the titles and abstracts and the controlled vocabulary terms. We then automatically added the top ranked terms to the query in the same way we would add blind feedback terms to a query. This leaves out the manual selection process where a searcher selects appropriate terms counting on the prediction that an EVM will rank the "best" or most effective controlled vocabulary terms first. Although the controlled vocabulary terms seem to represent the content of the query, the retrieval results didn't improve. More analysis is necessary to find the reason.

Using EVMs to add query terms automatically carries the risk of distorting the query and misrepresenting the content by putting to much weight on more ineffective query terms. Below is an example of the top 10 suggested controlled vocabulary terms from the German EVM for GIRT query number 2. We input the title and description of the query.

<num> 102 </num>
<DE-title> Deregulierung des Strommarktes </DE-title>
<DE-desc> Finde Dokumente, die über die Deregulierung in der Elektrizitätswirtschaft berichten. </DE-desc>

<cv>Deregulierung </cv>
<cv>Flexibilität </cv>
<cv>Elektrizitätswirtschaft </cv>
<cv>Arbeitsmarkt </cv>
<cv>Telekommunikation </cv>
<cv>Wettbewerb </cv>
<cv>Ordnungspolitik </cv>
<cv>Privatisierung </cv>
<cv>Wirtschaftspolitik </cv>
<cv>Elektrizität </cv>

Although some controlled vocabulary terms are wrongly suggested (e.g. Arbeitsmarkt), these terms could be specific enough to add more information to the query and not distort the original sense of the query. Following however is an example from the English EVM for GIRT where the EVM doesn't necessarily suggest "wrong" controlled vocabulary terms but also doesn't seem to add much valuable content to the query.

<num> 114 </num>
<EN-title> Illegal Employment in Germany </EN-title>
<EN-desc> Find documents reporting on illicit work in the Federal Republic of Germany.
</EN-desc>

<cv>labor market </cv>
<cv>federal republic of germany </cv>
<cv>labor market policy </cv>
<cv>unemployment </cv>
<cv>employment policy </cv>
<cv>new bundeslaender </cv>
<cv>employment trend </cv>

<cv>employment </cv>
<cv>effect on employment </cv>
<cv>old bundeslaender </cv>

The controlled vocabulary term "Federal Republic of Germany" occurs over 60,000 times in the collection and "Labor Market" and "Unemployment" over 4,000 times respectively. Adding these words is not discriminating for the search at all, just the opposite.

More analysis is necessary to find a more selective way of adding controlled vocabulary terms, maybe based on distribution measures within the document collection and appropriate fit with the query. It might be possible that EVMs cannot be used in a completely automatic manner (adding terms without manual pre-selection).

1.2 Thesaurus Matching

We have been experimenting with thesaurus matching for three years and yielded astonishingly good results. Thesaurus matching is a translation technique where the query is first split into words and phrases (the longest possible phrase is chosen). Secondly, these words and phrases are looked up in the thesaurus that is provided with the GIRT collection and, if found, substituted with the target language terms from the thesaurus. Words and phrases that cannot be translated (not found in the thesaurus) are kept in the original language. For a more detailed description of the technique, see Petras et al. (2002) and for a discussion of efficiency and advantages and disadvantages, see our paper from last year (Petras et al., 2003).

Thesaurus matching is in essence leveraging the high-quality translations of controlled vocabulary terms in multilingual thesauri. The GIRT thesaurus provides a controlled vocabulary in English, German and Russian. We experimented with thesaurus matching from German to English and from English to German and achieved comparable results to machine translation.

Although thesaurus matching relies only on the exact terms and phrases as they appear in the query, enough seem to be found to achieve a reasonable representation of the query content in controlled vocabulary terms. Even though Entry Vocabulary Modules also represent the query content in controlled vocabulary terms, adding them to the query instead of substituting query terms with them doesn't yield as noticeable results in bilingual retrieval. This might have several reasons, among them the number of added terms, the preciseness and distinctiveness of the chosen terms and the size of the controlled vocabulary (how many records contain the same controlled vocabulary term and how effective is adding a controlled vocabulary term).

1.3 The GIRT collection

The GIRT collection (German Indexing and Retrieval Test database) consists of 151,319 documents containing titles, abstracts and controlled vocabulary terms in the social science domain. The GIRT controlled vocabulary terms are based on the Thesaurus for the Social Sciences (Schott, 2000) and are provided in German, English and Russian.

In 2003, 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. Although these corpora are described as parallel, they are not identical.

Both collections contain 151,319 records, but the English collection contains only 26,058 abstracts (ca. one out of six records) whereas the German collection contains 145,941 - providing an abstract for almost all documents. Consequently, the German collection contains more terms per record to search on. The English corpus has 1,535,445 controlled vocabulary terms (7064 unique phrases) and 301,257 classification codes (159 unique phrases) assigned. The German corpus has 1,535,582 controlled vocabulary terms (7154 unique phrases) and 300,115 classification codes (158 unique phrases) assigned. On average, 10 controlled vocabulary terms and 2 classification codes have been assigned to each document.

Controlled vocabulary terms and classification codes are not uniformly distributed. For example, the top 12 most often assigned controlled vocabulary terms for both corpora make up about half of the number of assigned terms. Whereas the distribution of controlled vocabulary terms has no impact on the thesaurus matching technique, it influences the performance of the statistical association technique for Entry Vocabulary Modules, i.e. skews towards more often assigned terms. For this year's experiments, we haven't made efforts to normalize the data to ensure optimal training of the EVMs, which is a next step.

2 GIRT RETRIEVAL EXPERIMENTS

2.1 GIRT Monolingual

For GIRT monolingual retrieval, six runs for each language are presented, five of which were official runs. We compared two ways of using controlled vocabulary terms provided by the EVMs and submitted one official run for each.

We submitted the required run against a GIRT document index without the added thesaurus terms. For both languages, this was the run with the lowest average precision. However, the English run is much worse than the German (both in the first column of tables 1 and 2), demonstrating the effect of added keywords to documents when a lot of the abstracts are missing (see section 1.3 for a small analysis of the GIRT collections).

As a baseline, a run against the full document collection (including thesaurus and classification terms) without additional query keywords was used (second column of both tables 1 and 2). This baseline run was only minimally surpassed by the EVM-enhanced runs, yielding an average precision of 0.4150 for German and 0.3834 for English respectively.

The first method of adding controlled vocabulary terms to the query was used in official runs BKGRMLGG2 and BKGRMLEE2 for German and English respectively. The top three ranked suggested thesaurus terms from the Entry Vocabulary Modules (one for German and one for English) were added to the title and description of the query. The added terms were then down weighted by half as compared to title and description terms in retrieval. In columns 3-5 of tables 1 and 2, retrieval runs adding one, three and five controlled vocabulary terms suggested by an EVM are compared.

The second method of utilizing EVMs was used in official runs BKGRMLGG1 and BKGRMLEE1. Whereas the terms from the title and description of the query were run against a full document index, the added thesaurus terms were run against a special index consisting of the controlled vocabulary terms added to the documents only. The results of these two runs were then merged by comparing values of the probability rank provided by our logistic regression retrieval algorithm. For both German and English, this merging yielded worse results than the baseline run indicating that the run against the index with thesaurus terms only distorted results. The thesaurus terms alone might not have enough distinctive power to discriminate against irrelevant documents.

2.1.1 German Monolingual

For all runs against the German GIRT collection, we used our decompounding procedure to split German compound words into individual terms in both the documents and the queries. The procedure is described in Chen & Gey (2004). We also used a German stopword list and a stemmer in retrieval.

Additionally, we used our blind feedback algorithm for all runs except BKGRMLGG1 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. Using the decompounding procedure and our blind feedback algorithm usually increases the performance anywhere between 10 and 30%.

Table 1 summarizes the results for the German monolingual runs. The best run was adding 5 EVM-suggested thesaurus terms and then down weighting them in retrieval.

	BKGRMLGG0			BKGRMLGG2		BKGRMLGG1
	document index w/o thesaurus terms	baseline run	TD terms are weighted double	TD terms are weighted double	TD terms are weighted double	CV terms against separate CV index
Recall at		TD only	TD + 1 CV term	TD + 3 CV terms	TD + 5 CV terms	TD & 3 CV terms
0.00	0.7878	0.7273	0.7442	0.7843	0.8021	0.7290
0.10	0.6154	0.6587	0.6436	0.6725	0.6995	0.6666
0.20	0.5695	0.6025	0.5995	0.6268	0.6510	0.6101
0.30	0.5124	0.5584	0.5557	0.5703	0.5815	0.5631
0.40	0.4070	0.5033	0.5021	0.4921	0.4943	0.5038
0.50	0.3631	0.4457	0.4418	0.4505	0.4588	0.4206
0.60	0.3049	0.3841	0.3714	0.3835	0.3790	0.3728
0.70	0.2554	0.3093	0.2924	0.2960	0.2968	0.2958
0.80	0.2003	0.2509	0.2360	0.2287	0.2350	0.2324
0.90	0.1450	0.1723	0.1640	0.1614	0.1523	0.1579
1.00	0.0424	0.0525	0.0500	0.0678	0.0631	0.0604
Average	**0.3706**	**0.4150**	**0.4079**	**0.4177**	***0.4280***	**0.4102**

Table 1. GIRT German Monolingual

2.1.2 English Monolingual

For all runs against the English GIRT collection, an English stopword list and stemmer were used. We also used our blind feedback algorithm for all runs except BKGRMLEE1.

The best run in this series was adding one EVM-suggested thesaurus term and down weighting it in retrieval. It is still unclear how many added thesaurus terms might be best, especially since this seems to differ between the German and English collection.

				BKGRMLEE2		BKGRMLEE1
	document index w/o thesaurus terms	baseline run	TD terms are weighted double	TD terms are weighted double	TD terms are weighted double	CV terms against separate CV index
Recall at		TD only	TD + 1 CV term	TD + 3 CV terms	TD + 5 CV terms	TD & 3 CV terms
0.00	0.6794	0.7610	0.7660	0.7767	0.7757	0.7644
0.10	0.4263	0.5943	0.6368	0.6488	0.6017	0.6066
0.20	0.3664	0.5029	0.5319	0.5271	0.4868	0.5131
0.30	0.2979	0.4660	0.4895	0.4882	0.4348	0.4577
0.40	0.2429	0.4400	0.4705	0.4516	0.3907	0.4205
0.50	0.2160	0.3858	0.4045	0.3936	0.3396	0.3830
0.60	0.1687	0.3487	0.3599	0.3486	0.2882	0.3415
0.70	0.1136	0.2972	0.3078	0.2933	0.2256	0.2752
0.80	0.0381	0.2423	0.2548	0.2275	0.1779	0.2173
0.90	0.0085	0.1788	0.1753	0.1592	0.1383	0.1619
1.00	0.0013	0.0630	0.0584	0.0593	0.0629	0.0495
Average	**0.2131**	**0.3834**	***0.3985***	**0.3908**	**0.3445**	**0.3732**

Table 2. GIRT English Monolingual

2.2 GIRT Bilingual

For GIRT bilingual retrieval, 8 runs for each language are presented, 10 of which were official runs (5 for each language). For bilingual retrieval, we compared the behavior of machine translation, thesaurus matching, EVMs (suggesting controlled vocabulary terms and substituting them with their target language equivalent) and any combination of these.

The best bilingual runs rival the monolingual runs in average precision with one German → English run (BKGRBLGE1) marginally outperforming all English monolingual runs.

Last year, we compared the Systran and L & H Power Translator against each other with L & H alone performing better on both English → German and German → English translations than Systran or the combination of both. All translations of the query title and description were therefore undertaken with the L & H Power Translator only.

Both machine translation (L & H Power Translator) and thesaurus matching performed equally well. However, the combination of machine translation and thesaurus matching (coupling the translated title and description from machine translation and thesaurus matching and then down weighting terms that are duplicates) achieved even better results. All three runs can be compared in the first 3 column of tables 3 and 4. The combination runs were official runs (BKGRBLEG1 and BKGRBLGE1). The combined run outperforms all other runs in the German → English series and is second best in the English → German series.

Thesaurus matching outperforms a run composed of 5 translated thesaurus terms suggested by an EVM. This is not surprising since 5 terms or phrases seem not enough for effective retrieval. It remains to be seen whether a higher number of suggested terms could achieve comparable results or deteriorate because of increasing impreciseness of query words.

Official runs BKGRBLEG2, BKGRBLEG5, BKGRBLGE2 and BKGRBLGE5 combined machine translation provided by L & H and 5 or 3 EVM-suggested thesaurus terms respectively.

Runs BKGRBLEG4 and BKGRBLGE4 combined thesaurus matching and 5 EVM-suggested thesaurus terms.

The last 2 columns of tables 3 and 4 show combination runs of machine translation, thesaurus matching and EVM-suggested thesaurus terms, BKGRBLEG3 and BKGRBLGE3 were official runs.

2.2.1 Bilingual English → German

			BKGRBLEG1	BKGRBLEG5	BKGRBLEG2	BKGRBLEG4		BKGRBLEG3
Recall at	MT	Thes. Match	MT + Thes. Match	MT + 3 CV terms	MT + 5 CV terms	Thes. Match + 5 CV terms	MT + Thes. Match + 3 CV terms	MT + Thes. Match + 5 CV terms
0.00	0.6825	0.6238	0.7751	0.6956	0.7021	0.7012	0.7787	0.7912
0.10	0.5517	0.5167	0.6637	0.5552	0.5792	0.5362	0.6620	0.6590
0.20	0.4848	0.4659	0.5711	0.5033	0.5259	0.4752	0.5969	0.5735
0.30	0.4025	0.4234	0.5137	0.4612	0.4606	0.4178	0.5384	0.5126
0.40	0.3531	0.3952	0.4597	0.3961	0.3593	0.3141	0.4568	0.4028
0.50	0.3182	0.3685	0.4100	0.3435	0.2995	0.2869	0.3990	0.3601
0.60	0.2727	0.3114	0.3404	0.2635	0.2516	0.2372	0.3330	0.2998
0.70	0.2309	0.2522	0.2693	0.2055	0.2010	0.1945	0.2673	0.2430
0.80	0.1659	0.1991	0.1962	0.1541	0.1352	0.1484	0.1990	0.1757
0.90	0.1069	0.1209	0.1296	0.0719	0.0775	0.0833	0.1254	0.1138
1.00	0.0220	0.0177	0.0482	0.0219	0.0218	0.0167	0.0519	0.0307
Average	**0.3146**	**0.3287**	**0.3868**	**0.3224**	**0.3176**	**0.2964**	***0.3871***	**0.3641**

Table 3. GIRT English → German Bilingual

For English to German bilingual retrieval, the combination of machine translation and suggested EVM terms marginally outperforms machine translation alone but not the combination of machine translation and thesaurus matching. The combination of thesaurus matching and EVM suggested terms performs worse than thesaurus terms alone suggesting a deteriorating effect of the added terms. The combination of all three methods doesn’t achieve better results than the combination of thesaurus matching and machine translation alone.

2.2.2 Bilingual German → English

			BKGRBLGE1	BKGRBLGE5	BKGRBLGE2	BKGRBLGE4		BKGRBLGE3
Recall at	MT	Thes. Match	MT + Thes. Match	MT + 3 CV terms	MT + 5 CV terms	Thes. Match + 5 CV terms	MT + Thes. Match + 3 CV terms	MT + Thes. Match + 5 CV terms
0.00	0.6559	0.6326	0.7434	0.6312	0.6386	0.6348	0.6990	0.7184
0.10	0.5371	0.5450	0.6626	0.5184	0.5398	0.5394	0.5992	0.5957
0.20	0.4891	0.4843	0.5636	0.4916	0.4737	0.4894	0.5362	0.5407
0.30	0.4470	0.4507	0.5173	0.4567	0.4260	0.4300	0.4876	0.4875
0.40	0.4186	0.4120	0.4845	0.4035	0.3748	0.3948	0.4422	0.4454
0.50	0.3710	0.3499	0.4106	0.3691	0.3218	0.3609	0.3955	0.3903
0.60	0.3047	0.3096	0.3675	0.3095	0.2733	0.3172	0.3325	0.3200
0.70	0.2423	0.2534	0.3074	0.2533	0.2156	0.2471	0.2889	0.2618
0.80	0.2060	0.1915	0.2421	0.1959	0.1280	0.1868	0.2322	0.2178
0.90	0.1368	0.1169	0.1835	0.1468	0.0818	0.1083	0.1684	0.1499
1.00	0.0250	0.0498	0.0762	0.0442	0.0203	0.0280	0.0775	0.0446
Average	**0.3431**	**0.3370**	***0.4053***	**0.3370**	**0.3054**	**0.3340**	**0.3748**	**0.3668**

Table 4. GIRT German → English Bilingual

For German to English bilingual retrieval, the addition of EVM suggested thesaurus terms generally seems to deteriorate results probably by adding “noise” words to the query instead of relevant discriminative terms. Looking at the suggested EVM terms, however, doesn’t yet confirm this hypothesis. Most EVM suggestions seem quite sensible. It should be interesting to find out how much a manual selection of terms could improve results and how much “wrongly” suggested thesaurus terms worsen it.

3 References

Chen, A. and F. Gey (2004). Multilingual Information Retrieval Using Machine Translation, Relevance Feedback and Decompounding In: Information Retrieval, Volume 7, Issue 1-2, Jan. – Apr. 2004. pp. 149-182.

Chen, A.; Cooper, W. and F. Gey (1994). 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), pp 57-66, March 1994.

Gey, F. et al. (1999). Advanced Search Technology for Unfamiliar Metadata. In: Proceedings of the Third IEEE Metadata Conference, April 1999, Bethesda, Maryland 1999.

Petras, V.; Perelman, N. and F. Gey (2003). UC Berkeley at CLEF-2003 – Russian Language Experiments and Domain-Specific Retrieval. In: Proceedings of the CLEF 2003 Workshop, Springer Computer Science Series.

Petras, V.; Perelman, N. and F. Gey (2002). Using Thesauri in Cross-Language Retrieval of German and French Indexed Collections. In: Proceedings of the CLEF 2002 Workshop, Springer Computer Science Series.

Plaunt, C., and B. A. Norgard (1998). An Association-Based Method for Automatic Indexing with Controlled Vocabulary. Journal of the American Society for Information Science 49, no. 10 (1998), pp. 888-902.

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