How One Word Can Make all the Difference – Using Subject Metadata for Automatic Query Expansion and Reformulation

Vivien Petras

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

Abstract. Query enhancement with domain-specific metadata (thesaurus terms) is analyzed for monolingual and bilingual retrieval on the GIRT social science collection. We describe our technique of Entry Vocabulary Modules, which associates query words with thesaurus terms and suggest its use for monolingual as well as bilingual retrieval. Different weighting and merging schemes for adding keywords to queries as well as translation techniques are described.
Query enhancement generally improves average precision scores for both monolingual and bilingual retrieval. We take a closer look at individual queries and discuss how the query enhancements (or substitutions in bilingual retrieval) can change retrieval results quite dramatically. A query-by-query analysis provides deeper insight into strengths and weaknesses of strategies and serves as a cautionary reminder that average precision scores don't always tell the whole story.

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
Controlled vocabulary, thesauri, automatic query expansion, entry vocabulary modules

1 Introduction

Subject metadata (thesaurus terms, classification codes, library subject headings) in bibliographic databases serve several important purposes:

(i) To provide a concise topical description of the record content;
(ii) To provide a non-ambiguous term for each concept represented in the database, that is, to control the subject vocabulary;
(iii) To provide an organization scheme for the documents in the database;
(iv) To cluster all relevant documents for a concept under one term;
(v) To provide more searchable text for the user (especially in databases with sparse text in their records like library catalogs);
(vi) To aid retrieval by providing a topical access point that is unambiguous and retrieves a complete and precise document set for a given concept.

The problem with subject metadata for search is that the system vocabulary may differ from the searcher's vocabulary and so requires additional time and effort for the searcher to search / understand the controlled vocabulary of the system and to find the appropriate term for his or her information need. However, once the vocabulary has been mastered, searches are generally shorter, more precise and also more complete (finding all and only relevant documents).

In a database that contains subject metadata, it is therefore sensible to use and leverage its unique advantages. In a world that emphasizes ease of use and quick turn-around time, mastering the controlled vocabulary should not be required of the searcher. Automated query formulation support mechanisms help the searcher to find the appropriate search words by acting as an intermediary between the controlled vocabulary of the database and the natural language of the searcher.

The technique of Entry Vocabulary Modules was designed to be just that: serving as an interface between the query vocabulary of the searcher (natural language) and the controlled vocabulary entries of a database. Given any search word or phrase, it will suggest controlled vocabulary terms that represent the concept of the search. A searcher can use these terms to append to his or her query or to substitute his or her own query terms with those controlled vocabulary terms in the hope of achieving a more precise and complete retrieval.

Query expansion has been researched in the information retrieval field for a long time [1]. However, automatic query expansion has been mostly discussed in the context of blind feedback or highly evolved expert systems [e.g. 2,3]. Thesauri are mainly used for manual or interactive query expansion (for an overview, see [4]), but authors report mixed results [5-8] when comparing those techniques to free-text search.

For CLEF 2005, Berkeley's group 2 experimented with Entry Vocabulary Modules (EVMs) to automatically enhance queries with subject metadata terms or to replace query terms with them. The GIRT collection (German Indexing and Retrieval Test database) contains titles, abstracts and thesaurus terms providing an ideal test bed for monolingual and bilingual retrieval (German and English documents as well as a bilingual thesaurus).

The paper is organized as follows: first, we briefly introduce the GIRT collection and then explain Entry Vocabulary Modules and the basics of our retrieval technique. Section Five explains the runs for German and English Monolingual retrieval in detail. Section Six explains our translation techniques and how EVMs can be used for query translation. Sections 6.2 and 6.3 compare different translation techniques and discuss combinations for bilingual retrieval for English to German and German to English, respectively.

2 The GIRT Collection

The GIRT collection (German Indexing and Retrieval Test database) consists of 151,319 documents containing titles, abstracts and thesaurus terms in the social science domain. The GIRT thesaurus terms are assigned from the Thesaurus for the Social Sciences [9] and are provided in German, English and Russian. Two parallel GIRT corpora in English and German each containing 151,319 records are made available. For a detailed description of GIRT and its uses, see [10].

The English GIRT 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 entries (7064 unique phrases) and the German corpus has 1,535,582 controlled vocabulary entries (7154 unique phrases) assigned. On average, 10 controlled vocabulary terms / phrases are appended to each document.

Controlled vocabulary terms are not uniformly distributed. Most thesaurus terms occur less than a 100 times, but 307 occur more than 1,000 times and the most frequent one, "Bundesrepublik Deuschland", occurs 60,955 times.

3 Entry Vocabulary Modules

Entry Vocabulary Modules are automatically created search aids that function as intermediaries between the searcher's queries and the controlled vocabulary of a bibliographic database, in this case the GIRT thesaurus. They are referred to as Entry Vocabulary Modules because they provide a mapping from the "query vocabulary" of the searcher to the "entry vocabulary" of the database. A database's entry vocabulary consists of the subject metadata. It is this controlled vocabulary that provides an effective "entry" (access point) to the database records.

An Entry Vocabulary Module is in fact a dictionary of associations between terms in titles and abstracts in documents and the controlled vocabulary terms associated with the document. If title/abstract words and thesaurus terms co-occur with a higher than random frequency, there exists a likelihood that they are associated. A likelihood ratio statistic is used to measure the association between any natural language term and a controlled vocabulary term. Each pair is assigned an association weight (rank) representing the

strength of their association. The higher the rank, the more a thesaurus term represents the concept represented by the document word. The methodology of constructing Entry Vocabulary Modules has been described in detail in [11] and [12].

Once an Entry Vocabulary Module is constructed and a table of associations and their weights exist, we can look up a word in the dictionary and find its most highly associated thesaurus term. This is how we find thesaurus terms to associate with the GIRT queries. After experimenting with looking up query title and description words, we found that query title words are sufficient to find relevant thesaurus terms. For all CLEF 2005 experiments, only query title words (after stopword removal) were used for thesaurus term look-up. If more than one word appears in the query title, we need to merge the results from the thesaurus term look-ups to receive a list of terms for the query as a whole. We experimented with two merging strategies discussed below.

3.1 Absolute Rank Merging

For absolute rank merging, an absolute rank for each thesaurus term is calculated by adding the association weights if it is associated with several title words. The five thesaurus terms with the highest rank are then added to the query. We will use the English GIRT query 132 to illustrate this:

Title 132: Sexual Abuse of Children

	Sexual		*Abuse*		*Children*	*Absolute rank*	
3365.05	sexuality	1014.61	sexual abuse	19711.75	child	20468.45	child
1233.47	sexual abuse	767.84	abuse	2778.81	family	3640.36	sexuality
936.22	sex offense	431.38	child	2605.75	parents	2836.85	family
650.17	sexual harassment	307.05	sex offense	2344	parents-child relationship	2741.82	parents
471.52	homosexuality	275.07	maltreatment	2178.56	adolescent	2569.02	sexual abuse

This table shows a sample of the thesaurus terms associated with each individual title word and the absolute rank order for thesaurus terms after adding the weights for each thesaurus term and ranking again. For child, the association rank of the word “sexual” with the thesaurus term *child* is looked up (325.31 not shown in table), then added to the association rank of the title word “abuse” with *child* (431.38) and then added to the association for “children” (19711.75). The resulting 20468.45 is the absolute rank for the thesaurus term *child* and makes it the top-ranking thesaurus term for this query.

3.2 Round Robin Merging

The above example for absolute rank merging also shows the pitfall of this merging strategy: some association pairs (like “children” - *child* in query 132) have such high weights that other important query word – thesaurus term combinations will be ranked lower no matter what. To avoid this problem, we also tested a round robin merging strategy: for each query word, we looked up the two highest ranked thesaurus terms and added them to the query. The English GIRT query 138 will serve as an example:

Title 138: Insolvent Companies

Absolute rank merging	*Round robin merging*
enterprise	liquidity
firm	indebtedness
medium-sized firm	enterprise
small-scale business	firm
flotation	

The first two thesaurus terms in the round robin strategy are highly associated with “insolvent”, the second two with “companies”. As one can see in the absolute rank strategy, the thesaurus terms for “companies” seem to ‘overpower’ the ones for “insolvent”.

Sometimes, this strategy is prone to errors as topic 143 proves. The words looked up in the EVM are “smoking” and “giving”, which is misleading. The absolute rank strategy performs better in this case.

Title 143: Giving up Smoking

Absolute rank merging	*Round robin merging*
smoking	donation
tobacco consumption	social relations
tobacco	smoking
behavior modification	tobacco consumption
behavior therapy	

For German with its compounds ("Unternehmensinsolvenzen" instead of "Insolvent Companies" for topic 138), the round robin strategy sometimes only adds two instead of five thesaurus terms to the query, the ranking otherwise being equal to the absolute rank strategy.

4 Retrieval Technique

4.1 Document Ranking

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 [13]. The logodds of relevance of document D to query Q is given by

$$\log O(R \mid D,Q) = \frac{\log P(R \mid D,Q)}{\log P(\overline{R} \mid D,Q)} = -3.51 + 37.4 * x_1 + 0.33 * x_2 - 0.1937 * x_3 + 0.0929 * x_4$$

where $\log P(R \mid D,Q)$ is the probability of relevance of document D with respect to query Q and $\log P(\overline{R} \mid D,Q)$ is the probability of non-relevance of document D with respect to query Q. The regression variables are defined as follows:

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

$$x_2 = \frac{1}{\sqrt{n}+1} \sum_{i=1}^{n} \log \frac{dtf_i}{dl+80} \qquad (2)$$

$$x_3 = \frac{1}{\sqrt{n}+1} \sum_{i=1}^{n} \log \frac{ctf_i}{cl} \qquad (3)$$

$$x_4 = n \qquad (4)$$

where n is the number of terms common to both a document and a query, qtf_i/ dtf_i represent the frequency of term i within the query and document respectively, ctf_i is the frequency of term i in the collection, ql / dl represent the number of terms in the query and document respectively and cl is the collection length, i.e. the number of terms in the collection.

4.2 Collection and Query Processing

For all runs, we used a stopword list to remove very common words from the English and German collections and queries as well as an implementation of the Muscat stemmer for both English and German.

For German runs, we used a decompounding procedure developed and described by Aitao Chen [14,15], which has been shown to improve retrieval results. The decompounding procedure looks up document and query words in a base dictionary and splits compounds when found.

As a general procedure, we also use Aitao Chen's blind feedback algorithm [14,15] in 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.

query → stopword removal → (decompounding) → stemming → ranking → blind feedback

All query expansion and reformulation experiments described apply to the original query before submission to those processing steps and remain the same otherwise.

5 Monolingual Retrieval

For monolingual retrieval, we experimented with three query expansion strategies:

- adding five thesaurus terms retrieved with the EVM absolute rank merging from query title words;
- adding five thesaurus terms from the absolute rank merging strategy (using only query title words) but removing all thesaurus terms from the dictionary that occurred more than a 1,000 times in the document collection, thereby hoping to remove thesaurus terms that would not discriminate effectively;
- adding two thesaurus terms retrieved from the EVM for each query title word using the round robin merging strategy.

Last year, we experienced an improvement in precision when we weighted the expanded part of the query (the thesaurus terms) half as much as the original query words. This is also true for our other expansion mechanism (blind feedback), where new terms are added with half the weight as compared to the original query terms. For every expansion strategy, we analyze one run where the thesaurus terms are downweighted and one where they are treated as equally important part of the query.

We also experimented with submitting only the title of the query to the retrieval system, assuming that the shortness of the queries will simulate real user queries better than a title+description query. Since the EVMs don't need more information than the title words, we can also use the technique for these sparse queries.

For every run, we not only compared the overall average precision but also the precision scores on a query-by-query basis. This shows more clearly where the strengths and weaknesses of the individual strategies are but also reveals that sometimes just one word can influence precision scores dramatically.

5.1 German

5.1.1 Title + Description Runs

As the following table 1 shows, query expansion always improves over the baseline run of title+description if the expanded part is downweighted. If the thesaurus terms are not downweighted, only the round robin strategy improves over the baseline run. However, this case is also the dominating strategy, not only improving the baseline by 13% but also improving on the downweighted strategy and on the other merging strategies.

run	TD baseline	ABS HW	ABS	ABS -1000 HW	ABS -1000	RR HW	RR
official run	BK2G MLGG1	BK2G MLGG2		BK2G MLGG3		BK2G MLGG4	
average precision	0.4547	0.4733	0.4369	0.4595	0.3866	0.4936	*0.5144*

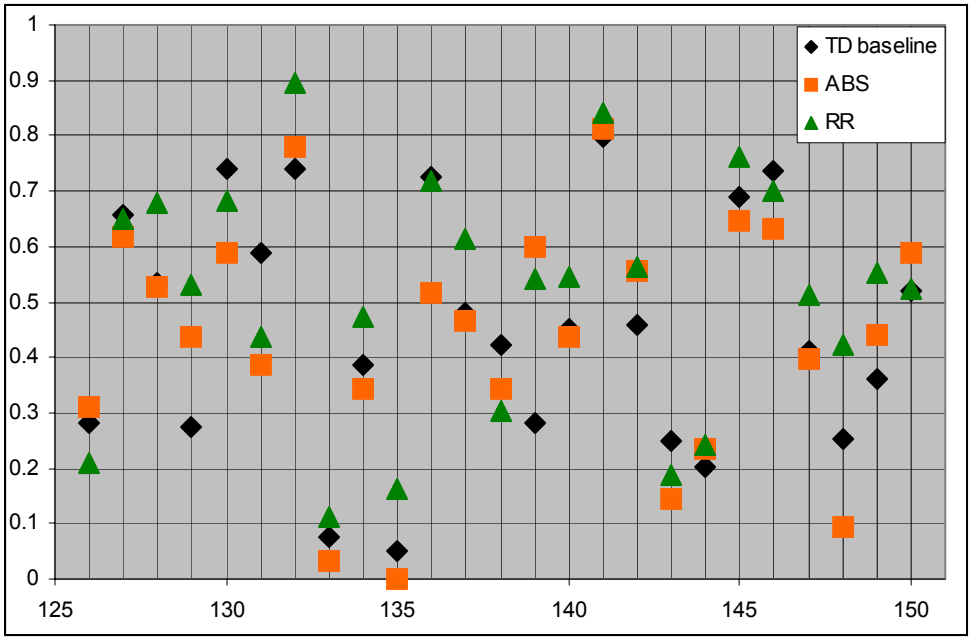
ABS	absolute rank strategy
ABS -1000	absolute rank strategy omitting thesaurus terms that occur more than 1000times in the collection
RR	round robin merging
HW	expanded thesaurus terms are downweighted by half in this run

Table 1. Average precision scores for title + description German Monolingual Runs

Comparing precision on a query-by-query basis, it becomes clear that downweighting clearly dominates for the absolute rank strategies, whereas not downweighting equally dominates for the round robin strategy although the average precision scores are much closer. In 18 of 25 queries, absolute rank merging with downweighting had a better precision than the not downweighted absolute rank strategy, for the absolute rank –1000 strategy, downweighting achieved a better result in 20 cases. For round robin, not downweighting turned out to be better in 17 of 25 cases compared to downweighting.

Comparing all seven runs with each other shows that the best run (RR) dominates in 11 cases, the baseline run in 6 cases, ABS HW in 3 cases, RR HW in 3 cases and ABS –1000 HW in 2 cases, changing the ranking order compared to average precision scores.

However, it makes more sense to compare strategies pair wise to see which one is stronger. We will look at the absolute rank and round robin strategies more closely to see how expanding a query by just a few words can change the results. Although downweighting works better for absolute rank merging (16 queries better than baseline) than not downweighting (9 queries better than baseline), we will use the not downweighted strategy to control for the effects of the weighting schemes.



Graph 1. Comparing precision scores per query for German Monolingual Retrieval

Graph 1 shows that results can vary for each strategy and query, the most dramatic change being the improvement from 0.2812 in the baseline to 0.6003 for ABS in query 139 (an improvement of 113%!). Even more amazing, looking at individual queries shows how little it takes to improve or degrade.

Query 131 serves as example where the baseline is better than query expansion:
<DE-title> Zweisprachige Erziehung </DE-title>
<DE-desc> Finde Dokumente, die die bilinguale Erziehung diskutieren. </DE-desc>

ABS	*RR*
Erziehung	Mehrsprachigkeit
Pädagogik	interkulturelle Erziehung
Schule	Erziehung
Bildung	Pädagogik
interkulturelle Erziehung	

The table shows the thesaurus terms that were appended to the query. Even though all of them seem relevant, the double occurrence of the word “Erziehung” in the thesaurus terms might skew the results too much towards documents dealing with education (Erziehung) alone and less with the bilingual aspect of it. Indeed, deleting the word “Erziehung” from the thesaurus terms in the RR strategy raises the precision from 0.43 to 0.55 (+28%), proving that sometimes one word can cause a huge improvement.

Query 139 serves as example where the expansion strategy works much better than the baseline:
<DE-title> Gesundheitsökonomie </DE-title>
<DE-desc> Finde Dokumente, die die Versorgung der Bevölkerung mit medizinischen und ärztlichen Dienstleistungen aus ökonomischer Sicht diskutieren. </DE-desc>

ABS	*RR*
Gesundheitswesen	Gesundheitswesen
Ökonomie	Ökonomie
Kostentheorie	
Wirtschaftskreislauf	
Gesundheitspolitik	

In this case, the words "Gesundheitswesen" and "Ökonomie" help most in improving the precision, but even leaving these terms out, the other three suggested thesaurus terms from the ABS strategy still raise the precision from 0.2812 to 0.5049 (+79%)!

Finally, query 148 is an interesting case showing how query expansion can be both advantageous and disadvantageous – depending on the terms expanded.
<DE-title> Russlanddeutsche und Sprache </DE-title>
<DE-desc> Finde Dokumente, die die sprachliche Integrität von Russlanddeutschen der ehemaligen Sowjetunion in Deutschland oder Russland diskutieren. </DE-desc>

ABS	*RR*
Sprache	Auswanderung
Sprachgebrauch	Spätaussiedler
Linguistik	Sprache
Fachsprache	Sprachgebrauch
Kommunikation	

The absolute rank strategy adds thesaurus terms that are too general for the query, decreasing precision by 62%. However, just adding the term "Spätaussiedler" from the round robin strategy improves precision by 44%.

5.1.2 Title only Runs
For title only runs we only experimented with the best strategy: round robin merging. As table 2 shows, queries expanded with thesaurus terms clearly improve precision over the baseline run (19%). For title only runs, downweighting thesaurus terms works better, improving the precision over the baseline by 30% and even more so, slightly improving on the baseline of the title+description run!

run	T baseline	RR	RR HW
average precision	0.3643	0.4339	*0.4748*

Table 2. Title only runs for German Monolingual retrieval

Comparing these runs on a query-by-query basis shows the dominance of the query expansion strategy even clearer: in 18 of 25 cases, RR is better than the baseline, and in 22 out of 25 cases RR HW is better than the baseline. RR HW is better than a title+description run in 14 cases.

One more experiment gives food for thought: instead of submitting the original query text, we only submitted the suggested EVM thesaurus terms from the round robin strategy, therefore reformulating the query instead of expanding it. Although the precision compared to the baseline decreases to 0.3075, substituting the thesaurus terms for the original query text works better in 12 of the 25 cases, showing that free-text does not dominate a controlled vocabulary strategy.

5.2 English

5.2.1 Title + Description Runs
As table 3 shows, query expansion with EVM suggested thesaurus terms is not as successful for English monolingual retrieval. However, the trend remains the same as in German monolingual retrieval. The round robin strategy without downweighting is still the dominating strategy, improving on the baseline by 6%. For the absolute rank strategies, downweighting works better, although they don't improve on the baseline.

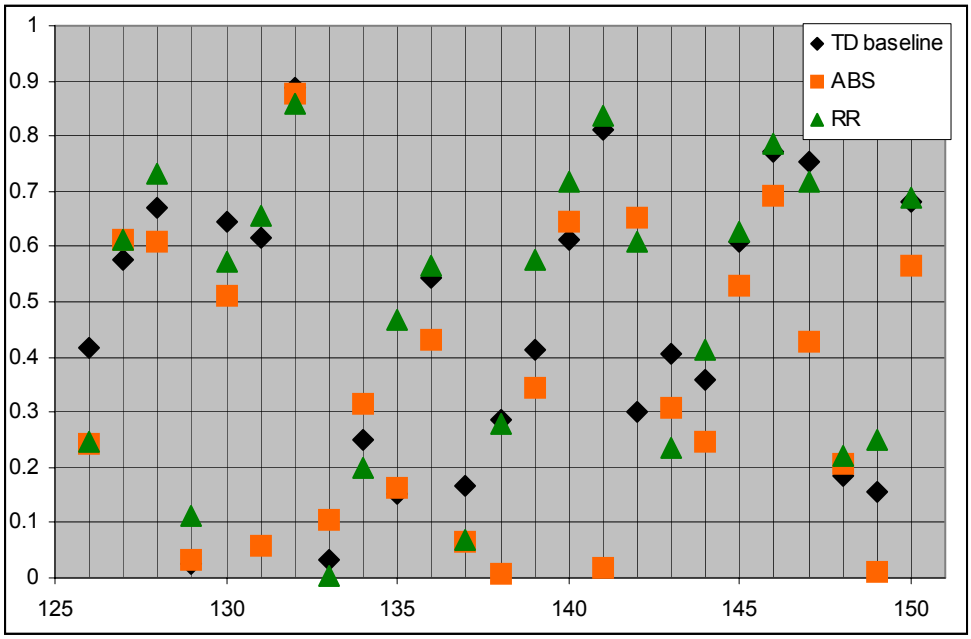
run	TD baseline	ABS HW	ABS	ABS - 1000 HW	ABS - 1000	RR HW	RR
official run	BK2G MLEE1	BK2G MLEE2		BK2G MLEE3			
average precision	0.4531	0.4149	0.3462	0.4125	0.3092	0.4697	*0.4818*

Table 3. Average precision scores for title + description English Monolingual Runs

The difference between downweighting or not is more pronounced when looking at the results on a query-by-query basis: in 21 out of 25 cases downweighting is better for the absolute rank strategy and in 20 of 25 cases for the absolute rank –1000 strategy. Not downweighting works better for round robin merging in 14 out of the 25 cases.

Comparing all seven runs shows that the best run (RR) only dominates in 9 cases, the baseline in 5, RR HW in 4, ABS HW in 3, ABS –1000 HW in 2 cases and ABS in 1 case demonstrating a weaker trend than in German monolingual retrieval.

Once again, graph 2 shows a comparison of precision scores for the baseline, the absolute rank and the round robin strategy. The absolute strategy works better than the baseline in 8 cases, but round robin is clearly better in 16 cases.



Graph 2. Comparing precision scores per query for English Monolingual Retrieval

Looking at graph 2 reveals two things: First, the absolute strategy seems to make things much much worse in some cases (131, 138, 141). This is because it adds thesaurus terms that are too general. But even the round robin strategy doesn't seem to improve precision as much as in German monolingual retrieval. Ironically, it seems that the unique characteristics of the German language (compounds) help in suggesting thesaurus terms that are not only more on the mark but are also compounds themselves retrieving more relevant documents. For example, the thesaurus term *way of life* translates to *Lebensweise* in German. Whereas for English, the retrieval system will look for documents containing "way" and "life" (very general!), the retrieval system will look for "Lebensweise" in German, which is much more precise.

However, it also cannot be overlooked that the English collection contains less text (fewer abstracts) than the German collection to search on. It might be that the added thesaurus terms skew search results in that they take away weight from the free-text search terms ranking documents containing the thesaurus terms (more likely) higher than ones containing the free-text search terms. This would explain the greater improvement of the downweighting strategies for absolute rank merging as compared to German (precision increases by 20% and

33% for ABS and ABS –1000 in English, whereas only by 8% and 19% in German) and the smaller improvement of not downweighting for round robin (2.5% in English vs. 4% in German).

Nevertheless, one query can serve as an example that one word can make a difference in English also: just adding the EVM suggested thesaurus term *morals* to query 142 (Advertising and Ethics) will improve precision by 31%.

5.2.2 Title only Runs

For title only runs, query expansion seems to improve on the baseline (+7%), although not as much as in German (19%). Downweighting again works better, improving the baseline by 14%.

run	T baseline	RR	RR HW
average precision	0.3972	0.4242	*0.4542*

Table 4. Title only runs for English Monolingual retrieval

Looking at the results on a query-by-query basis shows the dominance of the expansion strategies a little better: in 16 cases out of 25 RR dominates over the baseline, whereas RR HW is better in 18 cases. The best strategy for title only runs can compete with the baseline title+description run, with similar average precision and a better performance in 12 out of 25 cases.

However, replacing the title words with EVM suggested thesaurus terms works less well than in German: in only 5 cases this strategy performs better, decreasing the overall average precision to 0.2983 (-25%).

5.3 EVM Query Expansion vs. Blind Feedback

Although it has been shown that query expansion with EVM suggested thesaurus terms will improve monolingual retrieval in general, it might be of interest to compare this automatic technique of query expansion to another one – blind feedback. We have used blind feedback with success in previous years and now use it in all our retrieval experiments. Although EVM and blind feedback query expansion are quite different in nature – EVM works from the query title text, blind feedback from the result set document text – they are used to enhance the query to achieve better results. Table 5 gives a quick overview of runs using either strategy, both or none.

	Without query expansion	blind feedback	EVM suggested terms	blind feedback + EVM suggested terms
German				
avg precision	0.4622	0.4547	0.4902	0.5144
# of best queries		7	18	
English				
avg precision	0.4175	0.4531	0.4517	0.4818
# of best queries		15	10	

Table 5. Comparing blind feedback and EVM query expansion with pair-wise comparison for the blind feedback and EVM technique. The numbers represent the numbers of queries where this strategy achieved a higher precision score than the other (e.g. for German, the EVM technique achieved a higher precision in 18 cases).

The combination of both techniques outperforms the baseline and the individual query expansion techniques. For German monolingual retrieval, only EVM suggested terms improve over the baseline (in 16 out of 25 cases). For English, however, EVM terms improve only slightly over the baseline (13 cases), whereas blind feedback improves over the baseline (16 cases) and outperforms EVM expansion (better in 15 cases).

6 Bilingual Retrieval

6.1 Translation Methods

For bilingual retrieval, we experimented with query expansion and query reformulation using EVMs in addition to query translation. Three translation techniques are compared:

1. Machine translation. We used a combination of the Systran translator (http://babelfish.altavista.com/) and the L & H Power Translator.
2. Thesaurus matching. Words and phrases from the query are looked up in the thesaurus with a fuzzy-matching algorithm and if a matching thesaurus term in the query language is found, the equivalent thesaurus term in the target language is used. See [16] for a more detailed description.
3. EVM. The query title words were submitted to the query language EVM and the round robin merging technique was used to retrieve thesaurus terms. The thesaurus terms in the query language were then replaced by the thesaurus terms in the target language. The query was then reformulated using only thesaurus terms.

Query 144 serves as example for the different output of the translation strategies.

German query title:	Radio und Internet
English query title:	Radio and Internet
Machine translation:	radio and internet (L & H) radio and InterNet (Systran)
Thesaurus matching:	tradition (inaccurate due to fuzzy matching) internet
EVM suggestions:	radio / radio program Internet / online service

For bilingual retrieval, we will first compare these translation techniques separately and then in combination. In previous years, a combination of machine translation and thesaurus matching achieved the best results. For machine translation and thesaurus matching, both title and description of the query were submitted, for EVM only the suggested thesaurus terms were submitted.

6.2. Translation

Table 6 shows the average precision scores for the three translation methods in comparison for both bilingual tracks from German to English and English to German. For more comparison, the table also shows the number of queries with the better precision in a pair-wise comparison.

	German → English			*English → German*		
	Machine Translation	Thesaurus Matching	EVM thesaurus terms	Machine Translation	Thesaurus Matching	EVM thesaurus terms
avg. precision	*0.3917*	0.3052	0.3339	0.3532	*0.3558*	0.3236
# of best queries (out of 25)	14	11		10	15	
	14		11	15		10
		11	14		15	10

Table 6. Comparing 3 translation techniques for bilingual retrieval with pair wise comparison of strategy. The last 3 rows compare 2 strategies with each other, first machine translation vs. thesaurus matching, then machine translation vs. EVM terms and then thesaurus matching vs. EVM terms.

This table demonstrates once again that although average precision scores might differ significantly, a query-by-query analysis shows differently. Although thesaurus matching seems to perform worse in German-English retrieval (-23%), machine translation is better in only little over half of the cases. And although machine translation and thesaurus matching seem to perform equally well in English-German retrieval, thesaurus matching performs better in 3/5th of the cases. The performance of the EVM suggested thesaurus terms compared to machine translation is astonishing: an automatically associated list of controlled vocabulary terms

performs almost as well as the combined textual-based translations of two commercial machine translation programs!

6.3 Combining translation techniques

Combining translation techniques means submitting the translated output from the different methods in one and the same run. This increases the number of query words and the danger of introducing more non-discriminating search terms as well as favoring easy to translate terms (they most likely to occur in all methods), but for CLEF, this strategy has worked successfully in previous years. Combining translation methods helps with hard to translate words (higher chance of one method getting it right) and reduces the risk of mis-translation.

6.3.1 Official runs

The official runs for bilingual retrieval used the absolute rank merging technique when EVM suggested thesaurus terms were used. However, later experiments showed that round robin merging is also dominant for bilingual retrieval and so we report results for round robin merging. For documentation purposes we briefly state which official runs used which translation combinations below. Later sections will report combination runs with EVM round robin merging in more detail.

BK2GBLEG1 / GE1 machine translation + thesaurus matching
BK2GBLEG2 / GE2 machine translation + thesaurus matching + EVM absolute rank
BK2GBLEG3 / GE3 thesaurus matching + EVM absolute rank
BK2GBLEG4 machine translation + thesaurus matching + EVM absolute rank (downweighted)

6.3.2 German-English Bilingual Retrieval

Table 7 compares combination runs for German-English retrieval.

	Machine Translation + Thesaurus Matching	Machine Translation + EVM thesaurus terms	Thesaurus Matching + EVM thesaurus terms	Machine Translation + Thesaurus Matching + EVM thesaurus terms
avg. precision	0.4514	0.4566	0.4346	*0.4803*
# of best queries	13	12		
	12		13	
	7			18
		12	13	
		9		16
			7	18

Table 7. Combinations of translation techniques for German-English bilingual retrieval with pair-wise comparison of strategy

As one can see, a combination of all three techniques is clearly the dominating strategy – it seems that adding more words describing the same concept generally improves the precision instead of adding too many non-discriminating terms. It is also worth mentioning that all combination runs perform better than machine translation alone, even if one combines thesaurus matching and EVM terms only. In fact, even though lower in precision, this combination performs better in 13 out of 25 cases compared to both the machine translation – thesaurus matching and the machine translation – EVM pairs; a worthy competitor to the commercial translation solutions.

6.3.3 English-German Bilingual Retrieval

	Machine Translation + Thesaurus Matching	Machine Translation + EVM thesaurus terms	Thesaurus Matching + EVM thesaurus terms	Machine Translation + Thesaurus Matching + EVM thesaurus terms
avg. precision	0.4201	0.4059	0.4254	*0.4374*
# of best queries	14	11		
	13		12	
	13			12
		13	12	
		8		17
			12	13

Table 8. Combinations of translation techniques for English-German bilingual retrieval with pair wise comparison of strategy

For English-German retrieval, all combination runs seem to perform similarly. However, once again, they clearly outperform machine translation alone. Of course, not all combinations work equally well for each query and, sometimes, one translation technique alone works much better. Query 136 serves as example:

English query title:	Ecological waste economics
German query title:	Ökologische Abfallwirtschaft
Machine translation:	Ökologische Überflüssige Wirtschaftswissenschaft
	Ökologische Überschüssige Volkswirtschaft
Thesaurus matching:	Ökologische Partei
	Kaste
	Volkswirtschaftslehre
EVM suggestions:	Ökologie / Umweltpolitik
	Abfallwirtschaft / Abfall
	Wirtschaft / Volkswirtschaftslehre

Only the EVM round robin strategy manages to suggest the important word "Abfall" (waste) – whereas the other strategies either mistranslate "waste" or select the wrong thesaurus term due to incorrect fuzzy matching. The EVM words alone achieve a precision score of 0.6558, whereas the highest combination strategy achieves only 0.414 (thesaurus matching + EVM) – still better than the combination of machine translation and thesaurus matching (0.136), which is still better than machine translation (0.0295) or thesaurus matching (0.0236) alone.

7 Conclusion

Query expansion techniques have been a topic of research in the IR field for decades. Automatic query expansion has been analyzed mostly in terms of blind feedback mechanisms based on a preliminary ranked list of documents. Query expansion based on thesauri or other controlled vocabularies is mostly a topic for manual query expansion or interactive modes of query expansion. This paper discusses an automatic query expansion strategy using controlled vocabulary terms.

Expanding a query with terms from a thesaurus is like asking an information expert to translate your search strategy into the search language of the database, hopefully providing better search terms than the original search statement. The information expert for this set of experiments is an association dictionary of thesaurus terms and free-text words from titles and abstracts from the collection. Based on title words from the query, thesaurus terms that are highly associated with those words are suggested. Two merging strategies have been tested: absolute rank merging, based on all title words as a set and round robin merging, which suggests two thesaurus terms for each individual query word.

For monolingual retrieval, query expansion with EVM suggested thesaurus terms improves over the baseline of title + description submission by 13% (German) and 6% (English), respectively. Downweighting the added terms performs better for absolute rank but not for the round robin merging. For German, submitting only thesaurus terms (replacing the original query) decreases the average precision over 25 cases, but achieves better precision in 12 individual cases.

Comparing EVM query expansion to blind feedback (terms are taken from ranked result set documents and downweighted when added to query) shows that EVM query expansion improves over blind feedback in German and similar in performance in English, and a combination of both dominates either strategy and the baseline.

For bilingual retrieval, using the thesaurus for translation works surprisingly well. Just using thesaurus terms for the query submission works almost as well as machine translation. Although average precision decreases (9% for English-German and 15% for German-English), EVM suggested thesaurus terms perform better in one third of the queries. A combination of two thesaurus techniques (EVM and thesaurus matching) outperforms machine translation. The combination of machine translation, thesaurus matching and EVM suggested terms outperforms all other strategies.

It has been shown that EVM suggested terms can provide the impact to raise precision for a query – if they are high quality search terms. High quality search terms are those that provide discriminating search power (they

occur mostly in relevant documents), describe the information need exactly and, ideally, add new terms to the query. Added terms that are too vague will almost always degrade the performance. One word is all it takes to make the difference – now if we could only figure out which one!

8 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.

9 References

1. Efthimiadis, Efthimis N. 1996. Query Expansion. In *Annual Review of Information Systems and Technology (ARIST)*, edited by M. E. Williams. Medford, NJ: Information Today.
2. Gauch, S., and J. B. Smith. 1993. An expert system for automatic query reformation. *Journal of the American Society for Information Science* 44 (3):124-36.
3. Doszkocs, T.E., and R.K. Sass. 1992. An Associative Semantic Network for Machine-Aided Indexing, Classification and Searching. In *Advances in Classification Research, Vol. 3. Proceedings of the 3rd ASIS SIG/CR Classification Research Workshop*: Medford, NJ: Learned Information.
4. Shiri, A. A., C. Revie, and G. Chowdhury. 2002. Thesaurus-enhanced search interfaces. *Journal of Information Science* 28 (2):111-22.
5. Jones, S. 1995. Interactive thesaurus navigation: intelligence rules OK? *Journal of the American Society for Information Science* 46 (1):52-9.
6. Sihvonen, Anne, and Pertti Vakkari. 2004. Subject knowledge improves interactive query expansion assisted by a thesaurus. *Journal of Documentation* 60 (6):673-690.
7. Joho, Hideo, Mark Sanderson, and M. Beaulieu. 2004. A study of user interaction with a concept-based interactive query expansion support tool. In *ECIR 2004*, edited by S. McDonald and J. Tait. Berlin Heidelberg: Springer.
8. Suomela, Sari, and Jaana Kekäläinen. 2005. Ontology as a search-tool: a study of real users' query formulation with and without conceptual support. In *ECIR 2005*, edited by D. E. Losada and J. M. Fernández-Luna. Berlin Heidelberg: Springer.
9. Schott, Hannelore. 2000. *Thesaurus for the Social Sciences*. 2 vols. Vol. 1. German - English, 2. English - German. Bonn: Informations-Zentrum Socialwissenschaften.
10. 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.
11. Plaunt, C., and B. A. Norgard. 1998. An association-based method for automatic indexing with a controlled vocabulary. *Journal of the American Society for Information Science* 49 (10): 888-902.
12. Gey, Fred et al. 1999. Advanced Search Technology for Unfamiliar Metadata. *Third IEEE Metadata Conference, April 1999*. Bethesda, Maryland.
13. 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.
14. Chen, Aitao. 2003. *Cross-Language Retrieval Experiments at CLEF 2002*. 2785 ed, *Lecture Notes in Computer Science*.
15. Chen, A, and F Gey. 2004. Multilingual Information Retrieval Using Machine Translation, Relevance Feedback and Decompounding. *Information Retrieval* 7 (1-2):149-182.
16. 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.