Comparison Between Manually and Automatically Assigned Descriptors Based on a German Bibliographic Collection

Claire Fautsch, Jacques Savoy
Computer Science Department
University of Neuchatel
2009 Neuchâtel, Switzerland
{*Claire.Fautsch,Jacques.Savoy*}*@unine.ch*

***Abstract*—This paper compares and illustrates the use of manually and automatically assigned descriptors on German documents extracted from the GIRT Corpus. A second objective is to analyze the usefulness of both specialized or general thesauri to automatically enhance queries.**
To illustrate our results we use different search models such as a vector space model, a language model and two probabilistic models. We also proposed different measures to compute textual entailment between two terms allowing us to hopefully select appropriate keywords from thesauri to expand documents or queries automatically.

I. Introduction

During the last years electronic bibliographic tools gained more and more importance, partly due to the fact that electronic copies of printed media are made available on a large scale. For scientific journals, the growing printing cost especially when colors are required tends to favor electronic versions. Furthermore the distribution of electronic copies is nowadays much easier and faster than of printed media.

The information has not only to be made available, but the user must also be able to search the records easily and find pertinent information in an user-friendly way. For scientific papers, often only title and abstract are freely available in the bibliographic records database. This is mainly due to copyright issues. Hence these scientific documents often contain manually assigned keywords added to increase the matching possibilities between authors and information searchers. These keywords usually extracted from a controlled vocabulary can either be added during indexing by a person having a good knowledge in the given domain and/or by the author. An example for such an online bibliographic records database is ERIC[1], providing access to scientific literature for the educational world.

In this paper, we want to see whether manually added keywords can enhance retrieval. Moreover, we want to verify whether automatically added keywords might yield an improvement. Since domain-specific thesauri are not always available, we also use a general thesaurus for query and document expansion. We may thus see the differences between specific and general thesauri for the German language. In a second part we are interested in the impact of enhancing queries rather than documents. This is especially interesting if the searcher does not have a strong knowledge in the domain of interest and does not use domain specific terms in his/her query formulation. Expanding queries using a domain specific thesaurus might fill this gap between general and specific language more appropriate than a general thesaurus (e.g., WordNet [1]).

The rest of this paper is organized as follows. Section II presents related works, while in Section III we describe the test-collection and the thesauri used. Section IV gives a short overview of the different information retrieval (IR) models used for our evaluations and Section V explains the different lexical entailment measures used for term selection. Section VI shows the results of our different test runs. Finally in Section VII we summarize the main findings of this paper.

II. Related Work

For the various manual-indexing strategies used in the IR domain, their retrieval impact was studied and evaluated during the well-known Cranfield experiments. For example in the context of the the Cranfield II test (1,400 documents, 221 queries), Cleverdon [2] reported that single-word indexing was more effective than using terms extracted from a controlled vocabulary, where both indexing schemes were done by human beings.

Rajashekar & Croft [3] used the INSPEC test collection (12,684 documents, 84 queries) to evaluate retrieval effectiveness of various document representations. This study showed that automatic indexing based on article titles and abstracts performed better than any other single indexing schemes. While the controlled vocabulary terms by themselves were not effective representations, their presence as an additional source of evidence on document content could improve retrieval performance. Based on a corpus containing French bibliographic notices, in [4] we demonstrated that including manually assigned descriptors for title-only queries might significantly enhance MAP, compared to an approach that ignores them.

[1]Education Resources Information Center, http://www.eric.ed.gov/

In order to obtain better retrieval performance with the GIRT corpus, Petras [5] suggested adding manually assigned subject keywords in order to help make fuzzy topic descriptions less ambiguous. She later also showed that combining pseudo-relevance feedback and thesaurus-based query expansions could also improve retrieval performance.

Descriptors assigned manually represent significant cost increases for information providers and their utility must be analyzed and evaluated. In this perspective, we are concerned with the following question: Do such descriptors statistically improve the information retrieval process? The rest of this paper will try to provide answers to this question.

III. Test-Collection

The test collection we used for our different experiments is composed of the German GIRT corpora, 125 queries and two thesauri, a domain-specific thesaurus and a general thesaurus, described in the following sections.

A. GIRT Corpus

The GIRT (German Indexing and Retrieval Test database) corpus was made available through the CLEF[2] evaluation campaign. Over the years, the corpus has been enlarged to contain more than 150,000 documents, and an English translation is also available. More information about the GIRT corpora can be found in Kluck [6].

A typical record of the GIRT corpus consists of author name, title, document language, publication year and abstract and may as well contain manually added keyword terms. The document parts relevant for our experiments can be separated into two categories, on the one hand we have the title and abstract and on the other manually added keywords. The remaining fields (such as publication year) are not considered important for our experiments and will thus be ignored.

B. Topics

For our test runs, we used the queries deployed in the domain-specific track in the CLEF campaigns from 2004 to 2008. This gives us a total of 125 queries (i.e. 25 per year). Each topic is structured into three logical sections. The first part of a topic is a short title (T) followed by a brief description (D) of what the user is looking for, generally consisting of one short sentence. While these two sections represent the real user's needs, the last part (N) is a longer characterization of the user's needs indicating relevance assessment criteria. All topics have been judged on the same GIRT corpus.

[2]Cross Language Evaluation Forum, http://www.clef-campaign.org/

C. Thesauri

One of our objectives in this paper is to analyze the improvements in retrieval if additional keywords are added either manually or automatically to the documents or the queries. As a second objective we want to see if automatically added keywords, extracted from a thesaurus, add the same benefit to documents as those manually added.

1) Domain Specific Thesaurus: For the domain specific track in CLEF, a machine readable version of the German-English thesaurus for social science [7] was made available. The manually added controlled vocabulary terms were extracted from this thesaurus. We use this thesaurus as a domain specific thesaurus for automatically expanding documents or queries with keywords. The machine readable version is formatted in XML and contains 10,624 entries. Each entry represents a German descriptor, given with narrower and/or broader terms as well as with related terms. Other attributes that might also be given for a descriptor are *use-instead*, *use-combination* and *scope note*.

2) General Thesaurus: As a second, general thesaurus we use OpenThesaurus, freely available from http://www.openthesaurus.de/[3]. This thesaurus contains 17,619 entries, but on the contrary to the social science thesaurus each entry is just a set of words with a similar meaning. As the name implies, this thesaurus is "open" and regularly enlarged from different users through a collaborative effort. More information can be found in [8].

IV. IR Models

For indexing the documents and queries, we first normalize each indexing unit by transforming it to lowercase letters and removing diacritics (e.g., "Überraschung" would be normalized to "uberraschung"). We then apply our light stemmer[4], a decompounding algorithm for the German language [9] and remove words occurring in a stopword list (603 words, e.g., "der", "in", "ist").

To give a solid base to our empirical studies, we used different models to retrieve relevant information. As a baseline approach, we use a standard *tf idf* weighting scheme with a cosine normalization. As a second approach we used the Okapi (BM25) model proposed by Robertson *et al.* [10], evaluating the document D_i score for the query Q by applying the following formula:

$$Score(D_i, Q) = \sum_{t_j \in Q} qtf_j \cdot \log(\frac{n - df_j}{df_j}) \cdot \frac{(k_1 + 1) \cdot tf_{ij}}{K + tf_{ij}} \quad (1)$$

with $K = k_1 \cdot [(1 - b) + b \cdot \frac{l_i}{avdl}]$ where qtf_j denotes the frequency of term t_j in the query Q, n the number of documents in the collection, df_j the number of documents in which the term t_j appears and l_i the length of the document

[3]We use the image from November 19th 2008, 00:47

[4]Freely available at http://www.unine.ch/info/clef/

D_i. The constant b was set to 0.55 and k_1 to 1.2. $avdl$ represents the average document length.

As a third model we used $InB2$ derived from the *Divergence of Randomness* paradigm [11]. In the *Divergence of Randomness* framework two information measures are combined to obtain the weight w_{ij} of the term t_j in the document D_i. We then obtain following formula for the document score:

$$Score(D_i, Q) = \sum_{t_j \in Q} qtf_j \cdot w_{ij} \tag{2}$$

where

$$w_{ij} = Inf^1_{ij} \cdot Inf^2_{ij} = -\log_2(Prob^1_{ij}(tf_{ij})) \cdot (1 - Prob^2_{ij}(tf_{ij}))$$

For $InB2$, the two information measures are defined as follows:

$$Inf^1_{ij} = tfn_{ij} \cdot \log_2\left((n+1)/(df_j + 0.5)\right) \tag{3}$$

$$Prob^2_{ij} = 1 - [(tc_j + 1)/(df_j \cdot (tfn_{ij} + 1))] \tag{4}$$

with $tfn_{ij} = tf_{ij} \cdot \log_2\left(1 + ((c \cdot mean\_dl)/l_i)\right)$ where tc_j represents the number of occurrences of the term t_j in the collection. Moreover, c is a constant, fixed at 1.5 for our test cases and $mean\_dl$ is the mean document length.

To complete our models, we use a language model. Contrary to the Okapi and $InB2$ model, the language model approach is a non-parametric probabilistic model. We adopt a model proposed by Hiemstra [12] and described in Equation 5

$$P(D_i|Q) = P(D_i) \prod_{t_j \in Q} (\lambda_j \cdot P(t_j|D_i) + (1 - \lambda_j) \cdot P(t_j|C)) \tag{5}$$

with $P(t_j|D_i) = tf_{ij}/l_i$ and $P(t_j|C) = df_j/lc$ with $lc = \sum_k df_k$, where λ_j is a smoothing factor fixed at 0.35 for our experiments, and lc an estimate of the size of the corpus C.

V. Textual Entailment and Similarity Measures

In natural language processing, different measures are used to calculate textual entailment between two terms. We retain three measures.

As a first and simple measure we use the Jaccard similarity. Let u and v be the two terms for which we want to calculate similarity, and U and V the set of documents where they occur. We denote by $|U|$ (resp $|V|$) the cardinal of these sets. The Jaccard similarity between u and v is defined by following equation:

$$Jaccard(u, v) = \frac{|U \cap V|}{|U \cup V|} \tag{6}$$

The advantage of this similarity measure is that it is easy to calculate. As drawback it is known that this measure does not take into account the frequencies of the terms u and v in a document or in the collection. Under this consideration we use two other measures to compute lexical entailment. The first is a simple probability, defined as

$$P(v|u) = \sum_{d \in D} P(v|d)P(d|u) \tag{7}$$

where D is the set of documents in the collection and $P(v|u)$ is the probability of finding v in a document knowing this document contains u. $P(d|u)$ cannot be calculated easily, but we can assume that $P(d)$ is uniform (constant) and that if $u \notin d$, $P(d|u) = 0$. We can also assume that the length of d does not play any role. With these assumptions Equation 7 can be rewritten as

$$P(v|u) \propto \sum_{d \in D : u \in d} P(v|d)P(u|d) \tag{8}$$

As third and last measure we will use an average mutual information (MI) between two terms, defined as follows

$$I(u, v) = \sum_{X \in \{u, \widetilde{u}\}} \sum_{Y \in \{v, \widetilde{v}\}} P(X, Y) \log_2 \frac{P(X, Y)}{P(X)P(Y)} \tag{9}$$

where $\widetilde{u}$ (respectively $\widetilde{v}$) stands for the absence of u (respectively v). We note that if u and v are independent, $I(u, v) = 0$.

VI. Results

First we want to analyze the impact of manually or automatically assigned descriptors and see the difference in efficiency of human selected keywords versus automatically selected keywords. In a second step we automatically expand queries using a thesaurus with the intention of improving retrieval effectiveness. We used the four IR models described in Section IV and to measure the retrieval performance, we used MAP (Mean Average Precision) values computed on the basis of 1000 retrieved documents per query using TREC_EVAL[5].

A. Manually Indexing Evaluation

As a baseline we first evaluate a simple run searching only in the title and abstract part of the documents and using short (title only, T) query formulations (MAP depicted in second column of Table I).

To analyze the effect of manually added keywords, we then perform retrieval over the complete document, i.e searching for relevant information not only in the title and abstract part, but also in the keywords. In the third column of Table I (label "+Manual") we depicted the MAP when searching in title, abstract and keywords. The last column shows the performance difference before and after considering manually assigned descriptors.

This table shows that the inclusion of manually added keywords from a controlled vocabulary considerably improves retrieval results. This is a first indication showing

[5]http://trec.nist.gov/trec_eval/

	MAP		
Model	Title & Abstract	+Manual	%Change
tf idf	0.1929	0.2275	17.94
LM2	0.2865	0.3215	12.22
InB2	**0.3157**	0.3493	10.64
Okapi	0.3042	**0.3494**	14.86

Table I
MAP WITH AND WITHOUT MANUALLY ASSIGNED DESCRIPTORS FOR SHORT QUERIES (T-ONLY)

	MAP		
Model	Title & Abstract	+Automatic	%Change
tf idf	0.1929	0.1404	-27.22
LM2	0.2865	0.1992	-30.47
InB2	**0.3157**	**0.2496**	-20.94
Okapi	0.3042	0.2151	-29.29

Table II
DOCUMENT EXPANSION USING GIRT-THESAURUS

	MAP		
Model	Title & Abstract	+Automatic	%Change
tf idf	0.1929	0.1874	-2.85
LM2	0.2865	0.238	-16.93
InB2	**0.3157**	0.2406	-23.79
Okapi	0.3042	**0.2654**	-12.75

Table III
DOCUMENT EXPANSION USING OPENTHESAURUS

	MAP			MAP	
Model	No Exp.	GIRT	%Change	OpenThes.	%Change
tf-idf	0.2275	0.2285	0.44	0.2289	0.62
LM2	0.3215	0.324	0.78	0.3233	0.56
InB2	**0.3493**	0.3485	-0.23	0.3483	-0.29
Okapi	0.3494	**0.3503**	0.26	0.3510	0.46

Table IV
MAP AFTER QUERY EXPANSION WITH SHORT QUERIES (T)

us that adding keywords to bibliographic resources might be helpful. If we have a closer look at our results, we observe that for the *InB2* model for example, we have an improvement for 78 queries, but also a decrease for 45 queries. The question that then comes up, is if it is worth to spend human resources to add this keywords. Manually added keywords require time and people qualified in the given domain. Therefore we want to analyze if automatically added keywords based on a thesaurus might yield the same performance improvement.

B. Automatic Document Expansion

In this section we presented the results obtained when extending documents automatically with keywords. For manual expansion, an expert selects appropriate keywords from the thesaurus based on its knowledge of the domain and the context of the documents. With a computer we need an algorithm to select the controlled terms to be added. Our expansion procedure is mainly based on the textual entailment measures proposed in Section V and can be divided into four steps. First we select the part of the document (or query) to be extended. Then for each term t_i, we do a search in the thesaurus. For each retrieved thesaurus entry for the term t_i we retain all the terms w_j^i contained in the entry and compute their similarity score $score_{ij}$ with their related term t_i using one of the similarity measures described in Section V. Once we have finished this step for all terms t_i, we have a set of couples $(w_j^i, score_{ij})$. The terms w_j^i are candidates for expansion. Finally, since the number of potential candidates might be elevated, we select the N_{Best} terms with the highest score to extend the documents. For some documents there might be less than N_{Best} terms available. In this case all the candidate terms are added. Since the number of documents to expand is quite high, after some empirical analysis we selected the $Jaccard$ similarity for the expansion procedure, and fixed N_{Best} at 50 (which also equals the mean number of controlled vocabulary terms per documents). Table II shows the results of the retrieval using documents expanded with GIRT-thesaurus, and Table III using OpenThesaurus for expansion. We observe that automatically enhancing documents does not improve retrieval. Compared to manually added keywords, we only have improvement for 22 queries (vs. 78). However for 22 queries automatic document expansion performs better than manual expansion. For example with Query #153 ("Kinderlosigkeit in Deutschland"), the MAP is 0.6030 with manual expansion and 0.0839 with our suggested automatic expansion, while for Query #204 ("Kinder- und Jugendhilfe in der russischen Föderation") we have a MAP of 0.0494 after manual and 0.1930 after automatic expansion.

Compared to the GIRT-thesaurus the results are slightly better for OpenThesaurus, but we still have an important decrease compared to the retrieval results without keywords.

C. Query Expansion

In this part we present our results obtained when extending queries. After several tests, we decided to fix N_{Best} at 5, i.e. to each query are added at most 5 terms extracted form the thesaurus. We used the three measures presented in Section V to measure textual entailment between terms and chose expansion terms, as well as four retrieval models and the two thesauri and two query formulations, a short one using only the title part (T) and a longer using title and description (TD). We search in the complete document (title, abstract and keyword). Since our test runs show that all textual entailment measures perform the same, we only present results for the Jaccard measure.

Table IV shows a recapitulation of query expansion using Jaccard similarity for short query formulations (T) for both thesauri and the comparison to the baseline. We observe that query expansion does not bring any significant improvement.

	MAP			MAP	
Model	No Exp.	GIRT	%Change	OpenThes.	%Change
tf-idf	0.2428	0.243	0.08	0.2431	0.12
LM2	0.3606	0.3621	0.42	0.3616	0.28
InB2	0.379	0.3795	0.13	0.3793	0.08
Okapi	**0.3856**	**0.3861**	0.13	**0.3865**	0.23

Table V
MAP AFTER QUERY EXPANSION WITH LONG QUERIES (TD)

The small variations in the MAP are due to minor changes in the order of the retrieved documents rather than in the expected better retrieval of relevant documents for expanded queries. We make the same observations for longer query formulations as seen in Table V.

If we have a closer look at the results query-by-query for the *InB2* model and short queries (T), we see that for GIRT-thesaurus we have an improvement for 52 queries and decrease for 72. For OpenThesaurus, the use of thesaurus improves retrieval for 36 queries and decreases for 38. Query #44 ("Radio und Internet") for example has MAP 0.3986 if we do not use any query expansion. Using GIRT-thesaurus boosts MAP to 0.4509 (added words are "Rundfunk", "Datennetz", "Datenaustausch" and "Welle"), while OpenThesaurus even performs a MAP of 0.4846 ("Hörfunk", "Rundfunk", "Netze", "Netz" and "Funk"). For Query #118 ("Generationsunterschiede im Internet") however, the use of the GIRT-thesaurus drops MAP from 0.4789 to 0.4408 (added "Datennetz", "Datenaustausch" and "Intranet") while OpenThesaurus improves MAP to 0.4827 ("Netze", "Netz", "Web", "WWW" and "World"). This last example also shows that the choice of terms to expand the query is important, some expansion terms might hurt performance while some others might improve.

VII. CONCLUSION

In this paper we present the use of manually added keywords for searching relevant information in bibliographic records database written in the German language. While the manually assigned descriptors extracted from a controlled vocabulary considerably improve retrieval performance ($+13.9\%$ in mean), automatically added terms either from the same controlled vocabulary or from a general thesaurus hurt the retrieval performance. In a second part we tried to enhance queries rather than documents. The inclusion of keywords to the query however does not improve retrieval results.

We can conclude that adding terms extracted from a controlled vocabulary may improve retrieval performance. The problem however is to choose the right keyword terms to add to the documents. We tried different techniques to select expansion terms, but all show the same performance. Human specialists seem to be more accurate in selecting the appropriate keywords to enhance retrieval performance. In contrary to machines, a human person having a good knowledge in the given domain can take into account the semantics and pragmatics as well as the importance of a keyword term in the underlying corpus. Although if for some queries even manual indexing does not help to improve retrieval, it seems to be worth to invest time and human resources to gain in the overall performance for finding relevant information.

Acknowledgments: This research was supported in part by the Swiss NSF under Grant #200021-113273.

REFERENCES

[1] E. Voorhees, "Using WordNet™ to disambiguate word senses for text retrieval," in *Proceedings ACM-SIGIR'93*, 1993, pp. 171–180.

[2] C. W. Cleverdon, "The Cranfield Tests on Index Language Devices," *Aslib Proceedings*, vol. 19, pp. 173–194, 1967.

[3] T. B. Rajashekar and W. B. Croft, "Combining automatic and manual index representations in probabilistic retrieval," *Journal of the American Society for Information Science*, vol. 46, pp. 272–283, 1995.

[4] J. Savoy, "Bibliographic database access using free-text and controlled vocabulary: an evaluation," *Information Processing & Management*, vol. 41, pp. 873–890, 2005.

[5] V. Petras, "How one word can make all the difference - using subject metadata for automatic query expansion and reformulation," in *Working Notes for the CLEF 2005 Workshop, 21-23 September, Vienna, Austria.*, 2005.

[6] M. Kluck, "Die GIRT-Testdatenbank als Gegenstand informationswissenschaftlicher Evaluation," in *ISI*, ser. Schriften zur Informationswissenschaft, B. Bekavac, J. Herget, and M. Rittberger, Eds., vol. 42. Hochschulverband für Informationswissenschaft, 2004, pp. 247–268.

[7] H. Schott, Ed., *Thesaurus Sozialwissenschaften*. Informationszentrum Sozialwissenschaften, Bonn, 2002.

[8] D. Naber, "OpenThesaurus: ein offenes deutsches Wortnetz," in *Sprachtechnologie, mobile Kommunikation und linguistische Ressourcen: Beiträge zur GLDV-Tagung 2005 in Bonn*. Peter-Lang-Verlag, Frankfurt, 2005, pp. 422–433.

[9] J. Savoy, "Combining multiple strategies for effective monolingual and cross-lingual retrieval," *IR Journal*, vol. 7, pp. 121–148, 2004.

[10] S. E. Robertson, S. Walker, and M. Beaulieu, "Experimentation as a way of life: Okapi at TREC," *Information Processing & Management*, vol. 36, pp. 95–108, 2000.

[11] G. Amati and C. J. V. Rijsbergen, "Probabilistic models of information retrieval based on measuring the divergence from randomness," *ACM Trans. Inf. Syst.*, vol. 20, pp. 357–389, 2002.

[12] D. Hiemstra, "Term-specific smoothing for the language modeling approach to information retrieval: The importance of a query term," in *25th ACM Conference on Research and Development in Information Retrieval (SIGIR'02)*, 2002, pp. 35–41.