Combining Thesauri-based Methods for Biomedical Retrieval

Edgar Meij[1] **Leonie IJzereef**[1] **Leif Azzopardi**[1*] **Jaap Kamps**[1,2] **Maarten de Rijke**[1]

[1] ISLA, University of Amsterdam
[2] Archives and Information Studies, Faculty of Humanities, University of Amsterdam
http://ilps.science.uva.nl/

Abstract: This paper describes our participation in the TREC 2005 Genomics track. We took part in the ad hoc retrieval task and aimed at integrating thesauri in the retrieval model. We developed three thesauri-based methods, two of which made use of the existing MeSH thesaurus. One method uses blind relevance feedback on MeSH terms, the second uses an index of the MeSH thesaurus for query expansion. The third method makes use of a dynamically generated lookup list, by which acronyms and synonyms could be inferred. We show that, despite the relatively minor improvements in retrieval performance of individually applied methods, a combination works best and is able to deliver significant improvements over the baseline.

1 Introduction

The main focus of our participation in the TREC 2005 Genomics track was to evaluate the impact of integrating thesauri and related expansion methods in the retrieval model. We learned from interviews with biomedical researchers that the general search strategy within this domain is geared towards achieving high recall. A biomedical researcher is typically willing to read unwanted documents, if he knows all (or most) relevant ones will be retrieved. We hypothesized that the use of a controlled vocabulary could increase retrieval performance in general and recall in particular. Our working assumption was that controlled vocabulary terms can help overcome problems with synonymy and ambiguity. Thus achieving a higher recall rate by addressing the synonymy issue, but maintaining precision by removing ambiguity. To this end we investigated the results of three thesaurus-based methods.

Firstly, we attempted to boost performance by performing blind relevance feedback using the MeSH terms associated with the topics and MEDLINE abstracts, similar to the approach used by Kraaij et al. [10]. Secondly, we attempted to exploit the textual concept descriptions within the MeSH thesaurus itself by performing query expansion using the contents of these descriptions.

Our final method comprises of the automatic extraction of synonyms and acronyms from the corpus and the Medical Subject Headings (MeSH) thesaurus. Especially gene names tend to have a large number of possible synonyms and acronyms. We posited that the use of the controlled vocabulary terms from the documents themselves and the MeSH thesaurus would minimize the negative effects of synonymy and improve recall.

After evaluation we found that two methods (blind relevance feedback and acronym expansion) provided increases in mean average precision (MAP). Blind relevance feedback on MeSH terms improved early precision as well, without the loss of recall. However, a weighted combination of these two methods delivered an even higher MAP, without the loss of early precision.

The remainder of this paper is organized as follows. In Section 2 we describe our data processing, indexing, tools and models employed for this year's edition of TREC Genomics. Then we elaborate on our proposed methods in Section 3, followed by the results of our experiments in Section 4 and a more in-depth analysis of the topics. We summarize our findings in a concluding section.

2 Experimental Setup

In this section we elaborate on the particular tools, methods and models used for indexing and retrieving. We also compare a vector-space with a language modeling approach to retrieving MEDLINE abstracts and set our baseline.

2.1 Collection Processing

The document collection consists of a 10-year subset of MEDLINE, which contains over 4.5 million abstracts (totaling 9 Gb in size). Before indexing, the corpus required some preprocessing. First we selected the fields that might be useful for retrieval, as shown in Table 1.

For the MeSH terms field, we indexed only the main MeSH descriptors. We ignored any additional qualifiers, such as the topical subheadings. Special characters, such

*Current affiliation: Department of Computer and Information Sciences, University of Strathclyde.

Field	*Description*
PMID	PubMed Unique Identifier
TI	Title
AB	Abstract
MH	MeSH Terms
OAB	Other Additional Abstract (concatenated with AB)

Table 1: Citation fields

as the asterisks used to identify a document's most important MeSH term were also ignored. In order to preserve the complex MeSH terms we translated all terms to their unique Unified Medical Language System (UMLS) id's before indexing the document collection.

2.2 Query Preprocessing

There were five generic topic templates defined for the TREC 2005 Genomics track [3]. For each template the predefined components were identified using regular expressions. These were then removed and the remaining terms were considered the free text query submitted for that topic. For example in topic 120 (shown below), the query terms are highlighted in bold-face and the remaining terms were discarded. All methods and/or runs make use of the preprocessed queries.

120. *Provide information on the role of the gene* ***nucleoside diphosphate kinase (NM23)*** *in the process of* ***tumor progression***.

2.3 Lucene

We performed several experiments regarding the use of different fields from Table 1 in Lucene [11]. We also compared the results of Lucene's default Vector Space based settings with our in-house developed Language Modeling extension for Lucene [7]. Standard stopwords were removed in all runs, but no form of stemming was applied.

2.3.1 Vector Space Model

As our baseline we indexed the collection with off-the-shelf Lucene, using only the Abstract field. We use the vector space model, the default similarity measure in Lucene [11], i.e., for a collection D, document d and query q:

$$sim(q,d) = \sum_{t \in q} \frac{tf_{t,q} \cdot idf_t}{norm_q} \cdot \frac{tf_{t,d} \cdot idf_t}{norm_d} \cdot coord_{q,d} \cdot weight_t,$$

where

$$\begin{aligned} tf_{t,X} &= \sqrt{\text{freq}(t,X)}, \\ idf_t &= 1 + \log \frac{|D|}{\text{freq}(t,D)}, \\ norm_q &= \sqrt{\sum_{t \in q} tf_{t,q} \cdot idf_t{}^2}, \\ norm_d &= \sqrt{|d|}, \\ coord_{q,d} &= \frac{|q \cap d|}{|q|}. \end{aligned}$$

This resulted in a MAP of 0.190. When adding the Title as an extra field, MAP increased slightly to 0.198.

2.3.2 Language Model

To test whether a language model based approach could increase retrieval effectiveness, we used the standard version of Lucene with the ILPS extension [7, 11]. The extension uses a multinomial language model with tunable length prior and Jelinek-Mercer smoothing [4]. We estimated a language model for each document in the collection and for any given query we rank the documents with respect to the likelihood that the document language model generated the query. This can be viewed as estimating the probability $P(d|q)$, i.e., for a collection D, document d and query q:

$$P(d|q) = P(d) \cdot \prod_{t \in q} \left((1-\lambda) \cdot P(t|D) + \lambda \cdot P(t|d) \right),$$

We need to estimate three probabilities: the prior probability of the document, $P(d)$; the probability of observing a term in a document, $P(t|d)$; and the probability of observing the term in the collection, $P(t|D)$. We assume the query terms to be independent, and use a linear interpolation of a document model and a collection model to estimate the probability of a query term. The parameter λ is the so-called smoothing parameter, for which we use the standard value of 0.15.

The probabilities are estimated using the maximum likelihood estimate:

$$\begin{aligned} P(t|d) &= \frac{tf_{t,d}}{|d|}, \\ P(t|D) &= \frac{\text{doc\_freq}(t,D)}{\sum_{t' \in D} \text{doc\_freq}(t',D)}, \\ P(d) &= \frac{|d|}{\sum_{d' \in D} |d'|}. \end{aligned}$$

The collection model uses document frequencies rather than collection frequencies. The prior probability of a document is estimated as proportional to its length. In the implemented scoring formula the probabilities are reduced to rank-equivalent logs of probabilities.

The language model approach yielded better results than Lucene's vector-space based variant, with a MAP of 0.212 using only the Abstract field. When also selecting the title field, MAP decreased slightly to 0.204. All our subsequent retrieval runs are therefore based on this multinomial language model, using only the Abstract field.

3 Methods

In this section we provide a description of our proposed methods. Our first method uses a dynamically created look-up list of acronyms for query expansion; the other try to use the contents of the MeSH thesaurus for either blind relevance feedback or query expansion. The detailed results can be found in Section 4, particularly Table 2 and Figure 4.

3.1 Gene Name Expansion (`Ge`)

Although most gene names have several synonyms and acronyms, usually only one of these is used in either a query or a document. To be able to identify relevant documents that contain one of the alternative names, we expanded our queries with gene name and other variants. Since all topics were formulated based on only five different generic topic templates, we used the structure of these templates to identify possible abbreviations within the topics. This approach works best for topics in which there are gene names present. For example, in the following template a topic is created by filling the empty slots with respectively a gene name and a disease name:

> *Provide information about the role of the gene . . . in the disease . . .*

The query expansion was based on identifying synonyms and acronyms, which came from two different sources: the MeSH thesaurus and the MEDLINE corpus. Within the MeSH thesaurus, synonyms are defined between MeSH terms in a separate field, for example: *Vitamin C* ***see*** *Ascorbic Acid*. So whilst we could use the MeSH terms directly, we had to process the MEDLINE collection in order to extract any tacit or latent acronyms within the corpora. This was performed by extracting pairs of full terms and acronyms from the abstracts, using heuristics based on the cooccurrence of these terms, round brackets and abbreviations. For instance:

> *. . . binds* ***hepatocyte nuclear factor 4 (HNF4)*** *and COUP/TF-related proteins. . .*

This resulted in an acronym list of 33,417 combinations (13,386 unique acronyms). The acronym list and the MeSH thesaurus were used for a simple lookup procedure; if a gene name could be found in one or both, we added all its synonyms and acronyms to the query. An additional restriction was placed on this method; the original gene name (or one of its variants) has to be present in each retrieved document. Documents without the gene name or one of its variants were discarded. This results in the expanded query when applied to, for example, topic 111:

111. *Provide information about the role of the gene PRNP in the disease Mad Cow Disease.*

111. *+(PRNP "protein gene" "prp gene" "prion protein gene") Mad Cow Disease*

This approach performed quite well, with a 19% increase (as compared to the baseline) in MAP to 0.225.

3.2 MeSH Lookup (`Ml`)

There are over 22,000 descriptors in the 2004 MeSH. The MeSH thesaurus itself consists of records containing individual descriptions of the MeSH concepts. These descriptions include not only synonyms but also scope notes, information about semantic types, previous indexing names, and so on.

We hypothesized that the contents of the MeSH thesaurus could also be used for query expansion. Each descriptive record for a MeSH term is essentially equivalent to a document about that term. Hence, we considered all the textual information about a MeSH term as a document, to which a topic can be compared. This was performed by indexing the contents of the MeSH thesaurus with Lucene. We tried to identify the MeSH terms that are most related to a topic by querying this index with the query terms extracted from the topic. When querying the index, we allowed for some fuzziness to account for spelling variances in terms. A maximal edit distance of 1 was found to be the optimal fuzziness setting, based on the training data. We then selected the 5 top-ranked MeSH terms and these were subsequently used for query expansion and added to the MeSH field of the original query. The results on the final topics are detrimental, with a MAP of 0.029.

3.3 MeSH Based Feedback (`Fb`)

The rationale behind relevance feedback is that augmenting the original query with relevant terms from a retrieved set of documents can improve results considerably [1]. But since this is a non-interactive track, we need to rely on blind (or pseudo-) relevance feedback. Rocchio [13] proved the value of blind relevance feedback using statistical methods and Salton and Buckley [14] elaborated on this. Ponte [12] suggested a language modeling approach to relevance feedback. He adds additional terms to the original query based on the log ratio of the probability of occurrence in the model for relevant documents to the probability in the whole collection. Based on the 2004 TREC Genomics data, IJzereef et al. [6] have shown that blind relevance feedback on MeSH terms led to an improvement of retrieval effectiveness.

So, for our second method we performed an initial retrieval run using the same specifications as our Abstract-only language model run. We then follow Ponte's approach and identified the top *n* significant MeSH terms of the top *m* retrieved documents for every topic, using the ILPS extension for Lucene [7]. We then added these to the MeSH field of our original query and performed another retrieval run using this expanded query.

We decided to quantitatively measure the effects of different values of *m* and *n* on the retrieval results. Figure 1 shows the MAP plotted as a function of number of terms *n*, for different values of *m*. Figure 2 shows the same effect on precision at the first 10 retrieved documents (P@10). The horizontal lines show the MAP and P@10 of the initial run respectively, and represent a basis for comparison.

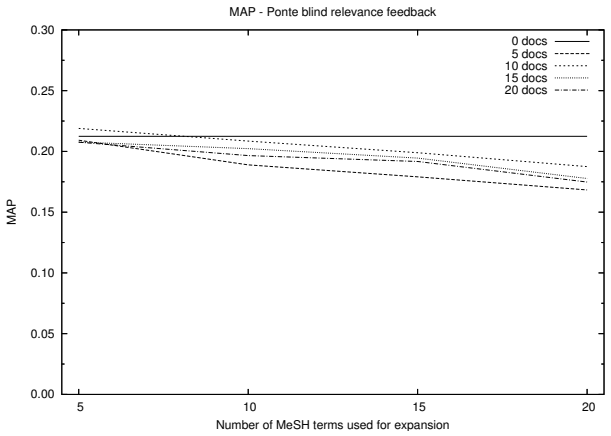

Figure 1: Effects of varying the number of terms and documents with Ponte feedback on MAP.

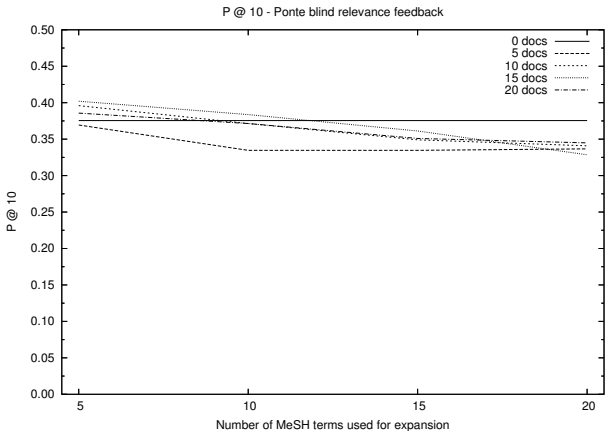

Figure 2: Effects of varying the number of terms and documents with Ponte feedback on P@10.

As can be seen from these figures, the optimal values for *m* and *n* are 10 and 5 respectively, with a resulting MAP of 0.219. Increasing the number of feedback terms and/or documents decreases MAP considerably. Additional experiments, which are not visualized here, have shown that further decreasing the number of MeSH terms has a negative effect on MAP as well as P@10.

3.4 Combining Runs

Intuitively speaking would a high early precision with the initial run imply better results from our subsequent MeSH based feedback method. We therefore expanded an initial Gene name expansion run (P@10 = 0.411) using 10 documents and 5 terms. This resulted in a MAP of 0.238.

This increase provided ground for further exploration. We started evaluating the results of combinations of methods, using the CombSUM method to combine each pair of methods Fox and Shaw [2], Kamps and de Rijke [9]. The rationale was that doing so would boost the relevant documents that are found with either method. Precision would thus increase because relevant documents get lower ranks and recall would increase because more relevant documents would end up in the top 1000 retrieved documents.

We computed MAP for varying weighting factors between methods, and every possible combination. Figure 3 shows the most interesting results with, from left to right, varying percentages of combining runs. Each smoothed line ends at the MAP of the best performing individual run, at 100% Gene name expansion (0.225). They represent the effect on MAP, when blending in other runs by reranking the results. The horizontal line is again included as a reference, representing the MAP of the Gene name expansion run.

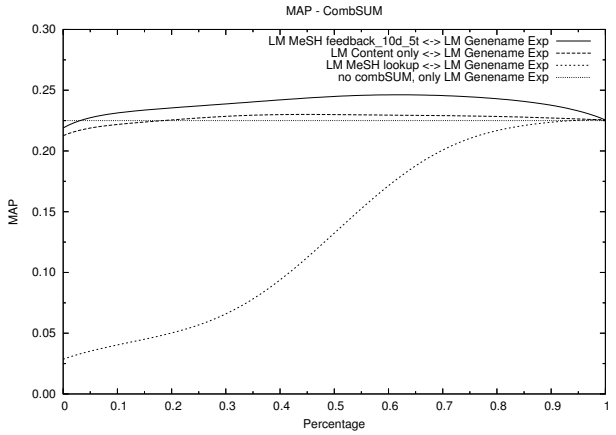

Figure 3: Effects of varying the weight on combining methods on MAP.

Earlier experiments based on the training data already showed that combining Gene name expansion and MeSH based feedback (`GeFb`) delivered promising results. The graph shows that combining the MeSH based feedback run with the Gene name expansion run can indeed boost MAP. The best results on the final topics are achieved when combining MeSH based feedback (5 terms, 10 documents) with Gene name expansion, in a ratio of 60% to 40% respectively. This combination results in a MAP of 0.25. It is interesting that the increase in MAP does not aversively affect early precision (P@10 = 0.42). Combining Gene name expansion

and MeSH lookup (GeMl) also performed well on the training data, but fails when applied to the final topics.

3.5 Official Runs

We computed the best weight factor for every combination, based on the results of the training data. These earlier experiments showed that combining Gene name expansion and MeSH based feedback (GeFb) and Gene name expansion and MeSH lookup (GeMl) delivered the best overall performance. With these findings in mind we had devised our TREC submissions accordingly.[1]

We submitted two runs for evaluation: UAmscombGeFb and UAmscombGeMl. For both runs the Gene name expansion was applied as described in subsection 3.1. The submitted runs both use different forms of MeSH based query expansion. The weights by which the individual methods were combined differed as well, based on the results from evaluations with the training data.

UAmscombGeFb

- **Gene name expansion (weight 0.60).**
- **MeSH based feedback (weight 0.40).** Our feedback method has been applied to the baseline run as described in subsection 3.3. Experiments performed on the training data showed that a selection of 15 feedback MeSH terms based on the 10 top-ranking documents yielded optimal results.

UAmscombGeMl

- **Gene name expansion (weight 0.85).**
- **MeSH lookup (weight 0.15).** The five best matching MeSH terms were selected per topic and added to the original query as described in subsection 3.2.

4 Experimental Results

The retrieval performance of each method from the previous section was thoroughly evaluated using the final adhoc topics. The results are shown in Table 2, with the results of the baseline run included as reference. The best scores are in bold-face. Figure 4 provides a visual overview of the proposed methods. The significance of the found results has been determined using Student's t-test.[2]

[1] Shortly after submitting these runs we discovered a flaw in the used term extractor. Due to this fact the results were slightly worse than could be expected when the proper tokenizer would have been used. For the remainder of this paper we will therefore be using the results of runs using the corrected term extractor instead of the actually submitted runs.

[2] There have been extensive discussions as to whether this particular test can be applied in this context, because of the assumption of normality of the distribution [5]. However, recent work has shown that it in fact it is just as reliable as non-parametric tests [15].

Clearly, applying the MeSH lookup method significantly degraded retrieval performance and this also has an effect on the performance of UAmscombGeMl. It is not as was expected; the proposed method retrieves many non-relevant documents. Gene name expansion gives a significant increase in MAP, without degrading early precision. The approach of using a dynamically created lookup list therefore has its definite merits. UAmscombGeFb gives statistically significant improvements for both recall and mean average precision as compared to the baseline. It also improves early precision, but to a lesser extent. The improvement in recall does not, contrary to common practice, aversively effect early precision, which is quite remarkable.

Figure 4 shows the precision-recall curves for all runs. Since this graph is quite cluttered, we have also included a graph of only our baseline, worst and best performing run. It is obvious to see from these graphs that UAmscombGeFb improves retrieval performance considerably, when compared to off-the-shelf Lucene.

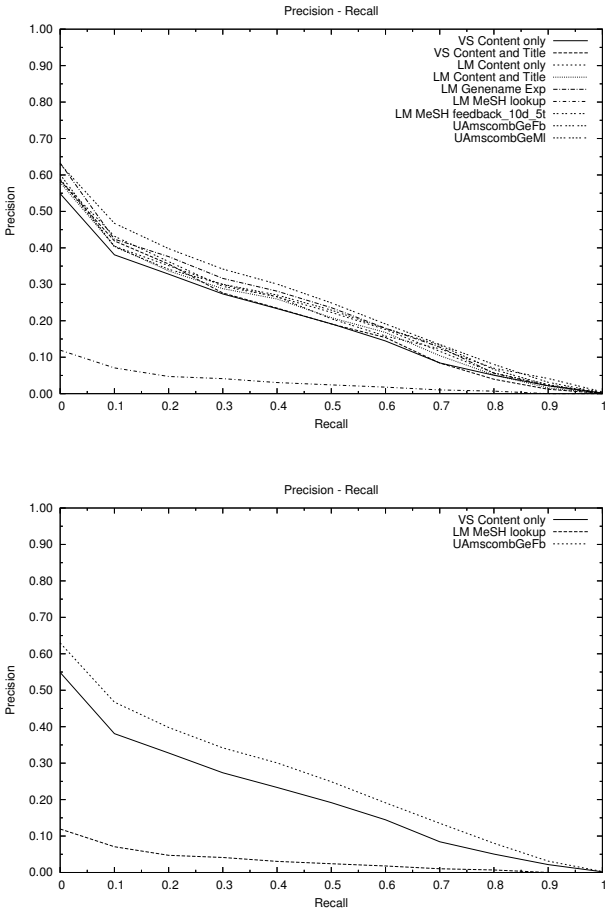

Figure 4: Precision-Recall curves.

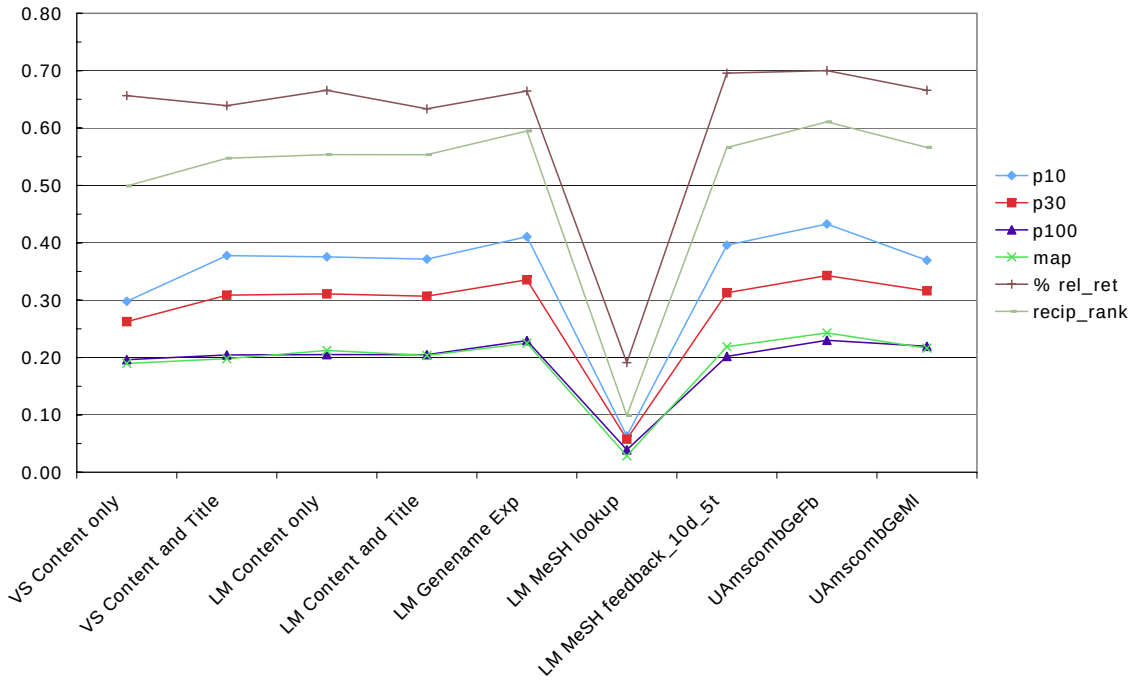

Figure 5: Overview of the results of all evaluated methods.

4.1 Topic Analysis

The results from Table 2 can be broken down into the scores of the individual topics. A graphical representation of the comparison of our best performing run (`UAmscombGeFb`) with baseline can be found in Figure 6. As can be seen in this figure, our combined thesaurus-based approach does, in general, improve recall as well as mean average precision.

There are still some topics however, where Lucene with its default settings performs better. These topics typically contain gene names for which incorrect acronyms or synonyms are found. Another cause for a drop in recall are the occurrences of multi-term gene names, such as *Insulin receptor gene*. In our baseline run each of these terms is considered individually as query terms. During the application of our Gene name expansion method, these separate terms are taken together and as such considered as a single, combined query term which, in turn, leads to reduced recall. For example, the application of Gene name expansion resulted in a drop in recall of 0.182 on topic 134. This topic is shown below (in original as well as expanded form).

134 *Provide information about the genes CFTR and Sec61 in degradation of CFTR which leads to cystic fibrosis.*

134 *+((CFTR "cl" "cystic fibrosis gene"*
"conductance regulator"
"cystic fibrosis transmembrane conductance regulator"
"conductance regulator protein"
"cystic fibrosis transmembrane regulator"
"conductance regulator gene") (Sec61))
degradation of CFTR which leads to cystic fibrosis

As can be seen from this example, the term *conductance regulator protein* was included, but is not indicative of the information need of the topic. In biomedical research it is uncommon to speak about proteins when referring to the gene that encodes for them, thus the drop in performance. This is also reflected in the fact that this particular term does not appear in any of the relevant documents for this topic.

Finally, there are many topics which benefit from both the proposed strategies. The next example returned no relevant documents during our baseline run, as opposed to a recall of 0.6316 using `UAmscombGeFb`. This improvement can be attributed mostly to the accurate Gene name expansion:

129 *Provide information on the role of the gene Interferon-beta in the process of viral entry into host cell.*

129 *+(Interferon-beta "beta-interferon"*
"fibroblast interferon" "interferon beta"
"beta 1 interferon" "interferon beta1"
"beta interferon" "beta-1 interferon"
"interferon beta 1" "interferon-beta1"
"ifn-beta" "fiblaferon" "interferon beta-1"
"interferon fibroblast" "ifnbeta")
viral entry into host cell

	P@10	% imp.	P@100	% imp.	MAP	% imp.	R@1000	% imp.
VS Content only	0.298	-	0.196	-	0.190	-	0.656	-
VS Content and Title	0.378**	26.7%	0.205	4.2%	0.198	4.3%	0.639	-2.7%
LM Content only	0.376**	26.0%	0.205	4.5%	0.212	11.9%	0.666	1.4%
LM Content and Title	0.371**	24.6%	0.205	4.4%	0.204	7.3%	0.634	-3.5%
LM Genename Exp	0.411**	37.8%	0.230	17.0%	0.225	18.5%	0.664	1.2%
LM MeSH lookup	0.063**	-78.8%	0.039**	-80.0%	0.029**	-84.9%	0.191**	-70.9%
LM MeSH feedback 10d 5t	0.396**	32.9%	0.202	2.9%	0.219	15.3%	0.696	6.0%
UAmscombGeFb	**0.433****	45.2%	**0.230***	17.2%	**0.243****	28.0%	**0.700****	6.6%
UAmscombGeMl	0.369*	24.0%	0.220	11.9%	0.216	14.0%	0.666	1.4%

Table 2: Tabular overview, with improvement over baseline. Best scores are in bold-face; significance *: $p < 0.05$, **: $p < 0.01$.

5 Conclusions and Future Work

Our main focus while participating in this year's TREC Genomics has been to evaluate the integration of thesauri in the retrieval model. We posited that the use of a controlled vocabulary would help the system overcome synonymy and ambiguity issues and come closer towards the information need of an end-user. To this end we have developed three thesauri-based methods. One method uses automatically extracted synonym/acronym pairs from the corpus and MeSH thesaurus. The other two, MeSH based feedback and MeSH lookup, use the contents and structure of the assigned MeSH terms respectively. Of these, Gene name expansion performs best. It improves MAP significantly, without losing early precision. MeSH lookup did not perform well. One can argue that searching in the descriptions of a thesaurus might not be a recommended approach. Based on the results on the training data however, it seemed a promising method. Perhaps the introduction of a cut-off point on the similarity measure instead of the number of found terms will improve effectiveness. Further research is needed however, to justify this assumption.

Based on the training data, as well as the final topics, we arrived at the conclusion that in fact combinations of methods work best. When applied individually, the proposed methods do not achieve significant improvements over our baseline run in terms of retrieval effectiveness. However, the combination of Gene name expansion with MeSH based feedback, with the proper weights, is able to deliver significant improvements over baseline. When examining the results of the individual topics, we found however that some topics benefited more from our proposed strategies than others. This can be attributed mostly to an incorrect matching of acronyms in the final topics. We wanted to test whether the proposed methods, with the additional effort involved, could significantly outperform off-the-shelf Lucene, which is the case.

Based on our interviews with biomedical researchers, we gained further insight in their search behavior and strategies. Besides using a strictly keyword-based search, they also use additional metadata. After an initial keyword-based retrieval run, they continue their search based not only on MeSH terms, but also on citations or authors of one or more top-ranked documents. If this kind of metadata is suitable for integration within the retrieval model remains to be seen.

Additionally, we would like to investigate the use of thesauri and/or ontologies within the retrieval model further.

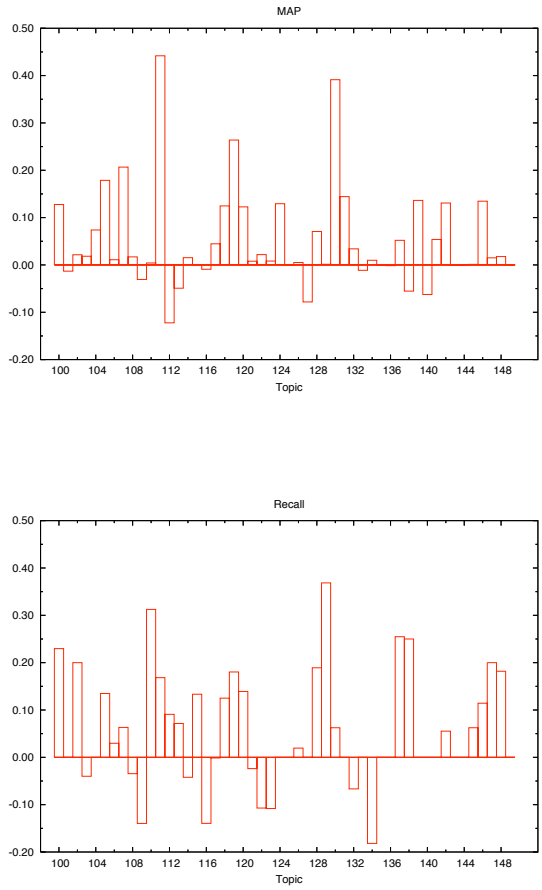

Figure 6: Per-topic breakdown of the results of UAmscombGeFb, as compared to baseline: MAP (top) and recall (bottom).

For example using not only MeSH, but also the contents and structure of the UMLS Metathesaurus for query expansion and/or blind relevance feedback. Wollersheim and Rahayu [16] have developed a framework for the implementation and evaluation of expanding queries on the conceptual, instead of the lexical level using ontological relations.

Reranking documents using a controlled vocabulary can improve retrieval effectiveness in domain-specific collections, such as the Cross-Language Evaluation Form (CLEF) [8]. We believe this might be the same when applied within the biomedical domain. Further research is needed to justify this hypothesis however. These are all issues we intend to address during our participation in the TREC Genomics track next year.

Acknowledgements

This work was carried out in the context of the Virtual Laboratory for e-Science project (`http://www.vl-e.nl`). This project is supported by a BSIK grant from the Dutch Ministry of Education, Culture and Science (OC&W) and is part of the ICT innovation program of the Ministry of Economic Affairs (EZ).

Leif Azzopardi was supported by grants from the Netherlands Organization for Scientific Research (NWO) under project number 612.000.106. Jaap Kamps was supported by grants from NWO under project numbers 612.066.302 and 640.001.501. Maarten de Rijke was supported by grants from NWO under project numbers 017.001.190, 220-80-001, 264-70-050, 354-20-005, 612-13-001, 612.000.106, 612.000.207, 612.066.302, 612.069.006, 640.001.501, and 640.002.501.

6 References

[1] N. J. Belkin, C. Cool, J. Head, J. Jeng, D. Kelly, S. jeng Lin, L. Lobash, S. Park, P. A. Savage-Knepshield, and C. Sikora. Relevance feedback *versus* local context analysis as term suggestion devices: Rutgers' trec-8 interactive track experience. In D. Harman and E. Voorhees, editors, *TREC-8, Proceedings of the Eighth Text Retrieval Conference.*, pages 565–574, Washington, D. C., 1999. NIST.

[2] E. A. Fox and J. A. Shaw. Combination of multiple searches. In D. K. Harman, editor, *The Second Text REtrieval Conference (TREC-2)*, pages 243–252. National Institute for Standards and Technology. NIST Special Publication 500-215, 1994.

[3] W. Hersh, A. Cohen, J. Yang, R. Bhupatiraju, P. Roberts, and M. Herst. Trec 2005 genomics track overview. In *TREC 2005 notebook*, pages 14–25, 2005.

[4] D. Hiemstra. *Using Language Models for Information Retrieval*. PhD thesis, University of Twente, 2001.

[5] D. Hull. Using statistical testing in the evaluation of retrieval experiments. In *SIGIR '93: Proceedings of the 16th annual international ACM SIGIR conference on Research and development in information retrieval*, pages 329–338. ACM Press, New York, NY, USA, 1993.

[6] L. IJzereef, J. Kamps, and M. de Rijke. Biomedical retrieval: How can a thesaurus help? In R. Meersman and Z. Tari, editors, *CoopIS/DOA/ODBASE*, pages 1432–1448. LNCS 3761, 2005.

[7] ILPS. The ILPS extension of the Lucene search engine, 2005. `http://ilps.science.uva.nl/Resources/`.

[8] J. Kamps. Improving retrieval effectiveness by reranking documents based on controlled vocabulary. In S. McDonald and J. Tait, editors, *ECIR*, volume 2997 of *Lecture Notes in Computer Science*, pages 283–295. Springer, 2004.

[9] J. Kamps and M. de Rijke. The effectiveness of combining information retrieval strategies for european languages. In *SAC '04: Proceedings of the 2004 ACM symposium on Applied computing*, pages 1073–1077, New York, NY, USA, 2004. ACM Press.

[10] W. Kraaij, M. Weeber, S. Raaijmakers, and R. Jelier. Mesh based feedback, concept recognition and stacked classification for curation tasks. In *Proceedings of TREC 2004*. NIST, 2005.

[11] Lucene. The Lucene search engine, 2005. `http://jakarta.apache.org/lucene/`.

[12] J. M. Ponte. Language models for relevance feedback. In W. B. Croft, editor, *Advances in Information Retrieval*, The Kluwer International Series in Information Retrieval, chapter 3, pages 73–95. Kluwer Academic Publishers, Boston, 2000.

[13] J. Rocchio. Relevance feedback in information retrieval. In *The Smart System – experiments in automatic document processing*, Englewood Cliffs, NJ, 1971. Prentice Hall.

[14] G. Salton and C. Buckley. Improving retrieval performance by relevance feedback. In K. S. Jones and P. Willett, editors, *Readings in information retrieval*, pages 355–364, San Francisco, CA, USA, 1997. Morgan Kaufmann Publishers Inc. ISBN 1-55860-454-5.

[15] M. Sanderson and J. Zobel. Information retrieval system evaluation: effort, sensitivity, and reliability. In *SIGIR '05: Proceedings of the 28th annual international ACM SIGIR conference on Research and development in information retrieval*, pages 162–169, New York, NY, USA, 2005. ACM Press.

[16] D. Wollersheim and J. W. Rahayu. Using medical test collection relevance judgements to identify ontological relationships useful for query expansion. In *Proceedings of the 21st International Conference on Data Engineering, ICDE 2005*, page 1160, Tokyo, Japan, 2005. IEEE Computer Society.