Bioinformatics Advance Access published April 17, 2009

MeSH Up: Effective MeSH Text Classification for Improved Document Retrieval

Dolf Trieschnigg [2,1]*, Piotr Pezik [1], Vivian Lee [1], Franciska de Jong [2], Wessel Kraaij [3] and Dietrich Rebholz-Schuhmann [1]

[1] European Bioinformatics Institute, Hinxton, United Kingdom
[2] HMI, University of Twente, Enschede, The Netherlands
[3] TNO ICT, Delft, The Netherlands

Associate Editor Dr. Limsoon Wong

ABSTRACT

Motivation: Controlled vocabularies such as the Medical Subject Headings (MeSH) thesaurus and the Gene Ontology (GO) provide an efficient way of accessing and organizing biomedical information by reducing the ambiguity inherent to free-text data. Different methods of automating the assignment of MeSH concepts have been proposed to replace manual annotation, but they are either limited to a small subset of MeSH or have only been compared to a limited number of other systems.

Results: We compare the performance of 6 MeSH classification systems (MetaMap, EAGL, a language and a vector space model based approach, a K-Nearest Neighbor approach and MTI) in terms of reproducing and complementing manual MeSH annotations. A K-Nearest Neighbor system clearly outperforms the other published approaches and scales well with large amounts of text using the full MeSH thesaurus. Our measurements demonstrate to what extent manual MeSH annotations can be reproduced and how they can be complemented by automatic annotations. We also show that a statistically significant improvement can be obtained in information retrieval (IR) when the text of a user's query is automatically annotated with MeSH concepts, compared to using the original textual query alone.

Conclusions: The annotation of biomedical texts using controlled vocabularies such as MeSH can be automated to improve text-only IR. Furthermore, the automatic MeSH annotation system we propose is highly scalable and it generates improvements in IR comparable to those observed for manual annotations.

Contact: trieschn@ewi.utwente.nl

1 INTRODUCTION

Controlled vocabularies play an important role in the integration of large scale bioinformatics resources and applications. They are actively used to annotate scientific literature and experimental data. Probably the most well-known example is the Medical Subject Headings (MeSH) thesaurus, developed and maintained by the National Library of Medicine, which has been introduced to categorize and search MEDLINE citations. Similarly, the Gene Ontology (GO) is used for annotating genes and gene product experiments (Lu *et al.*, 2008; Gaudan *et al.*, 2008). In both cases, *concepts*, i.e. distinct entries in a controlled vocabulary, are used for the *annotation* (also called classification and categorization depending on the context) of literature and experiments.

Unsurprisingly, these controlled vocabularies are increasingly used and investigated for representing, searching and summarizing information (e.g. Ruch 2006). Using a controlled vocabulary for representing information is especially useful in the biomedical domain, where a simple text-based representation of information is too ambiguous (Nenadic *et al.*, 2004). A conceptual representation allows information from different sources, such as databases containing documented experimental data and related literature to be linked in a transparent way, facilitating further data analysis in bioinformatics. Recently, MeSH and GO concepts have been used to prioritize genes by their relevance to diseases (Yu *et al.*, 2008).

The goal of this work is twofold. Firstly, our goal is to build a system which can annotate an arbitrary piece of text with relevant MeSH terms, similar to the manual classification of MEDLINE citations with MeSH terms. In this work we extensively compare six systems in terms of their capacity to reproduce manual classification. Several classifiers have been proposed in the past (see related work), but either their usefulness or evaluation to other methods has been limited. We focus our comparison on systems which allow classification using the complete set of available MeSH terms. In addition, we evaluate if manual classifications are complemented by automatically obtained MeSH terms. Secondly, our goal is to use this automatic classification method to improve upon biomedical document retrieval. We compare the usefulness of the six classifiers to automatically annotate a textual query with MeSH concepts. The effectiveness is tested on the Text REtrieval Conference (TREC) Genomics collections (Hersh *et al.*, 2004). We show that these improvements can be traced back to classification performance.

The structure of this paper is as follows. First, we give an overview of related work, followed by an overview of the different MeSH classifiers we have tested. Next, we evaluate the text classification performance of these classifiers. After that, we try to use the classifiers for the annotation of queries from several TREC Genomics test collections to improve document retrieval. We finish with a discussion and conclusion on the results.

*to whom correspondence should be addressed

2 RELATED WORK

Quite a number of researchers have developed MeSH classification techniques (see Sohn *et al.* (2008) for more related work). The assignment of MeSH descriptors to text is a large multi-class and multi-label text classification problem: one or more of 24,000 MeSH descriptors can be assigned to a piece of text. Most out-of-the-box text classifiers, such as decision trees, rule learning, neural networks and Support Vector Machines (SVMs) are not directly suitable for this task. SVMs for example, have shown their superiority to Naive Bayes' classifiers on binary classification tasks, but without sophisticated adaptation it is not feasible to train and build a system using SVMs for 24,000 classes.

The related work on MeSH classification shows a clear separation between research on sophisticated techniques limited to a subset of the problem and more straightforward techniques which do offer a complete solution.

Several researchers have used the OHSUMED collection and investigated the performance of their classifiers on a subset of MeSH descriptors, for example those in the Heart Disease branch (Lam and Ho, 1998; Ruiz and Srinivasan, 2002), or by only considering generalized descriptors (Rak *et al.*, 2007). Recently, Sohn *et al.* (2008) investigated optimal training sets for Naive Bayes' classifiers on a small set of 20 MeSH descriptors. Despite the reported improvements over the K-Nearest Neighbors approach, so far such a classifier has not been proven feasible for all 24,000 MeSH terms.

The systems which do classification on all descriptors are usually inspired by information retrieval techniques and return a ranked list of the most appropriate MeSH terms (e.g. Ruch (2006); Lam *et al.* (1999)). The actual classification, i.e. the binary assignment of a particular term to a piece of text, is achieved by cutting off the list at a particular rank or score.

The well-known Medical Text Indexer (MTI) introduced by Aronson *et al.* (2004) is further discussed in section 3.4.

We focus this work on systems which can be used for classifying the full set of MeSH terms.

A fair amount of research has also been carried out to incorporate MeSH terms in a retrieval system (see for example Camous *et al.* (2006) for an overview). During TREC Genomics[1] a large effort was spent to improve document retrieval using knowledge sources such as UMLS (including MeSH), Entrez Gene and Uniprot. Extending a user query with appropriate concepts from such knowledge sources showed mixed results. In fact, Hersh *et al.* (2004) note that in comparison to an out-of-the-box full-text search system "approaches that attempted to map to controlled vocabulary terms [such as MeSH] did not fare as well". In this article we show the relationship between the quality of this mapping and improvements observed in IR. Next to mapping query text to concepts using string matching, a common method to obtain a MeSH based representation of the query is to use relevance feedback: The original query is used to retrieve a set of documents and based on the MeSH terms assigned to these documents, a MeSH-based query is obtained. Srinivasan (1996) observes improvements in document retrieval using MeSH terms and despite the limited size of the collection concludes that MeSH terms are important for retrieval. Our work reinvestigates this conclusion in the context of improved document retrieval methods on larger document and query collections.

[1] `http://ir.ohsu.edu/genomics/`

3 SYSTEM AND METHODS

Four types of approaches are investigated. Firstly, two classifiers which only use the information in the MeSH thesaurus itself (referred to as "Thesaurus-oriented" classifiers). Secondly, two systems which use training data to build explicit models for each MeSH concept ("Concept-oriented" classifiers). Thirdly, a system which uses the manual annotations of documents similar to the text to classify, to determine suitable concepts ("K-Nearest Neighbor" classifier). Finally, a hybrid and manually refined system which combines different approaches ("Hybrid" classifier).

One of the Concept-oriented classifiers and the K-Nearest Neighbor classifier are based on information retrieval based on language models, a commonly used retrieval framework which is briefly explained in the online supplement. In sections 3.1 to 3.4 the investigated systems are explained. Example output of the different systems can be found in the online supplement. In the last section we discuss the evaluation methodology for the two tasks.

3.1 Thesaurus-oriented classifiers

The first two investigated classifiers both rely on information in the MeSH thesaurus only. The assignment of MeSH terms is based on the match between the information about a particular MeSH term, such as its synonyms and short description, and the text to classify.

3.1.1 MetaMap MetaMap is a major component of the NLM's Medical Text Indexer (see 3.4). Thesaurus concepts are found by first parsing the text into simple noun phrases and then by matching a large number of generated variants to the entries in the Unified Medical Language System (UMLS) metathesaurus. In our experiments, we filtered the output to concepts which occur in the MeSH thesaurus. Aronson (2001) describes MetaMap in more detail. MetaMap assigns a confidence score to each concept found. These scores are used to rank the list of MeSH terms in descending confidence order.

3.1.2 EAGL Ruch (2006) introduced a retrieval based system for MeSH classification. For each MeSH term, its synonyms and description are indexed as a single document in a retrieval index. A piece of text, the query to the retrieval system, is classified with the best ranked MeSH "documents". The advantages of this approach are high speed and small index size. One drawback is that it may return MeSH terms which only share a single word with the text to classify. The phrase "Breast cancer" could, for example, yield the MeSH term 'Breast cancer', but also other MeSH terms containing the word 'cancer', such as 'Testicular cancer' and 'Stomach cancer'.

3.2 Concept-oriented classifiers

The MeSH thesaurus has already been used extensively to classify MEDLINE citations, so it seems obvious to use the available manual assignments of MeSH terms to citations as training data.

For the concept-oriented classifiers, we build a model for each MeSH concept offline, i.e. before the actual classification. Similar to EAGL, an index is created in which each MeSH term is represented by a special "MeSH document". This MeSH document is simply created by merging the titles and abstracts of a number of documents assigned with that MeSH term. Two common retrieval methods are used for retrieving the most relevant MeSH documents, one based on language models (described in the online supplement) and

the other using a vector based representation, which are described below.

3.2.1 Concept Language Models For the classification system based on language models, a concept language model (CLM) is created for each MeSH term based on the MeSH document introduced before. This CLM is a probability distribution over words which are associated to a MeSH term. The parameters of the CLM are a maximum likelihood estimate based on the relative occurrence frequencies of words in the MeSH document.

Formally, the probability of a term t in a CLM is defined as:

$$P(t|M) = \sum_{D \in D_M} P(t, D|M) \approx \sum_{D \in D_M} P(t|D)P(D|M)$$

where D_M is a set of documents assigned to the MeSH term M, $P(D|M)$ is the probability a document language model is picked to describe this term (which is assumed uniformly distributed over D_M) and $P(t|D)$ is the smoothed document language model of D.

A piece of text is classified by creating a query language model $P(t|Q)$ for this text and ranking the concept language models using the negative cross entropy $-H$:

$$-H(Q, D) = -\sum_t P(t|Q) \log P(t|D)$$

This system shows close resemblances to a Naive Bayes classifier, commonly used for text classification (Lewis, 1998).

3.2.2 BM25 The Okapi BM25 is a vector space retrieval model which is commonly used as a baseline for retrieval experiments (Robertson *et al.*, 1996). The MeSH document is indexed as a TF.IDF vector and the text to classify is used as a query Q.

Given a query Q, the BM25 score of a MeSH document is:

$$score(D, Q) = \sum_{q \in Q} IDF(q) \frac{f(q, D) \cdot (k_1 + 1)}{f(q, D) + k_1 \cdot (1 - b + b \cdot \frac{|D|}{avgdl})},$$

where $IDF(q)$, is the inverse document frequency of the term q. k_1 and b are tuning parameters, and $avgdl$ is the average document length.

3.3 K-Nearest Neighbors classifier

The K-Nearest Neighbors (KNN) classifier investigated here is similar to the PubMed Related Citations Algorithm (Lin and Wilbur, 2007). A piece of text is classified by looking at the manual classification of similar or neighboring documents. We consider KNN for three reasons. Firstly, it can be easily scaled up to such a large classification task. Secondly, it gracefully integrates documents as a link between text and groups of related concepts. For document classification and retrieval such an integration may be preferred over approaches which model separate (rules for) concepts. The last reason is practical: in many research environments a full text search system on MEDLINE is already available, making KNN straightforward to implement.

Our KNN classifier relies on a retrieval system based on language models. Similar to CLM, the parameters of the query language model are estimated on the text to classify. Next, citations most similar to this query language model are retrieved. The classification is based on the MeSH terms assigned to the top K retrieved documents (based on preceding experiments $K = 10$ was used). The relevance of a MeSH term is determined by summing the retrieval score of the top documents that have been assigned that term.

3.4 Hybrid classifier: Medical Text Indexer

The Medical Text Indexer (MTI), provided to registered users by the NLM, incorporates different classifiers, including MetaMap, the 'Pubmed Related Citations algorithm' and 'Restrict to MeSH'. Different processing steps including clustering and applying (manually defined) rule-based filtering are used in this hybrid system. Parts of the systems have been evaluated using user questionnaires and in a 'machine learning setting' (Kim *et al.*, 2001; Aronson *et al.*, 2004). An evaluation against other classification systems or an assessment of its usefulness for information retrieval is missing however. Details of the system can be found on the Semantic Knowledge Representation website[2]. We treat MTI as a black box system, using the default settings to obtain MeSH classifications which favors the MeSH term suggested by MetaMap (with weight 7) over the ones from the related citations component (weight 2).

3.5 Evaluation methods

3.5.1 Evaluating text classification A commonly used method to evaluate MeSH text classification is to see how well a classifier reproduces the manual annotations of MEDLINE citations. Selected citations of the OHSUMED collection (Hersh *et al.*, 1994) have been used as training and test data, but as Ruiz and Srinivasan (2002) note, different test collections and variable numbers of categories have been used, making comparisons difficult. Moreover, the OHSUMED collection is not up-to-date anymore. At the time of its creation, the MeSH thesaurus consisted of around 14,000 MeSH terms. Currently, the thesaurus contains around 24,000 terms, making an evaluation using OHSUMED not representative for the current state of MeSH. Similar to Ruch (2006), we therefore take a random sample of a 1000 citations from the MEDLINE 2008 baseline distribution.

Lam *et al.* (1999) describe three quality metrics which can be used in this context: document, category and decision perspective metrics. The document perspective metric evaluates the assignment of MeSH terms at the document level. Since all our classification systems rank the suggested MeSH terms, a summary measure can be used to indicate which system has the ability to rank manually assigned categories higher than others: 10 or 11-point average precision, more commonly known as *Mean Average Precision* in IR can be used to indicate the performance at a document level. A more intuitive document perspective metric is Precision at 10 (P10), which indicates how many of the first 10 suggested terms correspond to manual annotations. The category perspective metric calculates the F-measure, Precision and Recall for each MeSH term. Finally, the decision perspective metric (micro recall, precision and F-measure) looks at the number of correct and incorrect decisions a classification system makes, where each possible document and category pair form a decision. Both the F-measure and micro F-measure require a discrete number of classifications per instance and our classifiers return a ranked list of classes. Similar to Lam *et al.* (1999) we report the measures using the optimal cutoff value (additional

[2] `http://skr.nlm.nih.gov/`

measurements are provided in the online supplement) [3]. For more information about these measures, see Lam *et al.* (1999).

Despite the fact that manual annotations of MEDLINE are carefully created and on average the most important terms are assigned, we note that using these manual annotations for evaluation is an idealization. Manual annotators do accidentally assign irrelevant MesH terms or miss relevant terms. To investigate this issue, an experienced annotator judged some of the false positives, i.e. automatic annotations which are not in the set of manually assigned terms. For 50 of the 1000 citations in the test, the annotator judged the three highest ranking false positives from MetaMap, CLM and KNN[4] on a 5-point scale. To test the reliability of our annotator three manual annotations were added to each citation as well. For each of the 50 citations, the title and abstract were presented with 12 (9 false positives and 3 true positives) randomly ordered MeSH terms. Each MeSH term was then judged on a 5-point scale ranging from "Strongly irrelevant/Incorrect" to "Strongly relevant" (the scale is discussed in detail in the online supplement). This analysis provides additional insights into the performance of the different classification systems. Some of the automatically identified terms may have been judged as irrelevant (false positives), because they were not included in the original MeSH annotations. By taking a closer look, however, we may actually find them to be highly relevant, i.e. appropriate to represent the text to classify.

3.5.2 Evaluating document retrieval To determine the added value of using MeSH terms for document retrieval we carry out a TREC-style evaluation[5]. Given a fixed document collection and a number of queries for information, systems are evaluated for their ability to improve document retrieval. As a baseline we only use the textual representation of the queries and documents. Using the evaluated systems, for each textual query we generate a set of MeSH terms, which serve as conceptual queries. In a first experiment, the conceptual queries are matched against the original MeSH annotations provided by MEDLINE (Table 4). In a second set of experiments, the conceptual queries are matched against document annotations generated automatically using the KNN system (Table 3).

The retrieval model is again based on unigram language models (described in the online supplement), which proved to be successful during previous TREC evaluations (Hiemstra and Kraaij, 1999). To differentiate between the added value of the conceptual and textual representations, separate text and concept indices are created. In contrast to for example Srinivasan (1996), who indexes the words in the MeSH terms, unique identifiers are used in the concept index. The MeSH term 'Adult' for example, is indexed as 'D000328'. This prevents matching MeSH terms with overlapping surface forms (e.g. 'Mad hatter disease' with 'Mad cow disease') and makes concept matching as unambiguous as possible. Similarly, two query models are created, one textual and one conceptual. The parameters of the textual query model are based on a maximum likelihood estimate on the query text. The conceptual query model is based on output of one of the six classifiers on the query text. The parameters of the model are based on relevance scores of suggested MeSH terms:

$$P(c|Q_C) \propto \frac{s(c, Q_C)}{\sum_{c' \in Q_C} s(c', Q)},$$

where $s(c, Q)$ is the classification score assigned to MeSH concept c for the query and Q_C is the set of terms suggested by the system.

As a matching model, we interpolate the query likelihood of both representations:

$$P(Q|D) = \alpha P(Q_C|D_C) + (1 - \alpha)P(Q_T|D_T), \quad (1)$$

where α defines the mix between text and concepts.

For the baseline, in which we only use the textual representations for retrieval, α is set to 0. Since we do not know the optimal value of α we vary this value between 0 and 1 with steps of 0.05, to find the optimal mix.

Following the commonly used TREC evaluation criteria, we use mean average precision and precision at 10 as performance indicators. As suggested by Smucker *et al.* (2007), Fisher's randomization test is used to determine the statistical significance of the results.

4 MESH DOCUMENT CLASSIFICATION

As an initial test of our selected MeSH classifiers, we look at their capability to reproduce manual MEDLINE annotation.

4.1 Experimental setup

A thousand random MEDLINE citations are selected as a test set from the MEDLINE 2008 baseline distribution, with the only requirement that they should have at least one MeSH term assigned to them. The list of citations can be downloaded for followup research[6]. The test set covers 3951 distinct MeSH terms (9596 assignments). The remaining citations in the 2008 baseline distribution are used for training: to build an index for the KNN approach and for sampling citations (at most 1000 citations per MeSH term) to assemble the MeSH document for the BM25 and CLM approach.

4.2 Results

Table 1. MeSH classification performance on 1000 random MEDLINE citations, using title and abstract as input. All differences in MAP and P10 are significant with a p-value < 0.005, based on Fisher's randomization test.

	Document				Category		Decision	
Method	MAP		P10		F_1		micro F_1	
MTI	0.2536		0.3200		0.4503		0.4415	
BM25	0.0912	-64%	0.1021	-68%	0.2251	-50%	0.1972	-55%
MetaMap	0.1623	-36%	0.1910	-40%	0.3187	-29%	0.2968	-33%
CLM	0.1783	-30%	0.1748	-45%	0.3429	-24%	0.2982	-32%
EAGL	0.1976	-22%	0.2119	-34%	0.2987	-34%	0.2977	-33%
KNN	0.5052	+99%	0.4515	+41%	0.4074	-10%	0.4963	+12%

[3] by assuming the number of top classes which gives the highest score

[4] Restricted to these systems because of resource limitations.

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

[6] http://www.ebi.ac.uk/~triesch/meshup/testset_v1.xml

Table 1 shows the classification results of the different systems when presented with the title and abstract of a 1000 random MEDLINE citations.

MTI serves as the baseline to compare the other systems to and shows to perform quite well on the classification task. Both thesaurus-oriented classifiers (MetaMap and EAGL) and concept-oriented classifiers (CLM and BM25) perform worse than MTI on all metrics. KNN forms a notable exception: it shows 99% improvement in terms of MAP, 41% improved precision at 10 (P10) and 12% improvement in micro F_1. On average, more than three of the top 10 returned terms from MTI correspond to manual annotation, whereas KNN returns more than four matching terms. In terms of Category F_1 KNN performs 10% worse than MTI: when considering one MeSH term, MTI is better in choosing whether to assign it to a citation or not. Considering the performance from a document perspective, KNN outperforms MTI: given the title and abstract of a citation, KNN finds more correct/manual MeSH terms and ranks them higher.

MTI shows to be very sensitive to the amount of input provided. When presented with only the title (and the PMID) of the citation (tables are available in the online supplement), it performs much worse on all measures (loss between 35 and 43%). In contrast, KNN only shows a moderate decrease (drops between 4 and 9%), which indicates that it is more robust when less information is presented. Also the other four systems (BM25, EAGL, MetaMap and CLM) are less sensitive to the length of the input (dropping at most 30%).

Additional investigation (see online supplement for metrics) shows that MTI and KNN are capable of reproducing both general and specific MeSH terms. The other four systems perform relatively well on reproducing specific MeSH terms, i.e. terms which are not frequently used for annotation in MEDLINE.

Table 2 shows the results of the annotation process described in section 3.5.1. The first column of table 2 shows that in 88% of the cases our annotator judged the original MeSH annotations as (very) relevant. Using more common inter-annotator agreement measures, such as Cohen's Kappa is not applicable in this case, since we do not know the explicitly negative judgments of the MeSH annotators.

Despite MetaMap's relatively poor performance on reproducing manual annotations, the results show in many cases its terms are useful for representing the text (58% of its false positives are judged as "Relevant" or better). Only few false positives (3%) are indicated as totally incorrect. Compared to CLM and KNN, only few terms (14,7%) get labeled "Undecided". This is because MetaMap requires an almost direct link between words in the text to classify and the MeSH terms it suggests. As expected, quite a few terms are suggested of which only part can be related to the text to classify.

The largest part of the false positives from the CLM system are judged as "Undecided" (35.5%). The system returns too many specific terms and some of the suggestions cannot be directly linked to the text to classify. For KNN, most of the false positives (31%) are indicated as "irrelevant". This value can be explained because KNN often returns general terms which are found in similar documents, but are not appropriate to this specific piece of text.

In general we notice that a fair share of the false positives is judged "relevant" or better (58% for MetaMap, 37% for CLM and 34% for KNN), indicating automatic annotations do contribute relevant terms in addition to manual annotations.

Table 2. Results from the analysis of false positives.

	True		False positives					
Judgment	positives		MetaMap		CLM		KNN	
Very relevant	94	75%	40	29%	44	24%	37	20%
Relevant	17	13%	39	29%	26	14%	27	14%
Undecided	12	10%	20	15%	66	35%	49	26%
Irrelevant	1	1%	33	24%	35	19%	58	31%
Incorrect	2	2%	4	3%	16	9%	17	9%

5 IMPROVING DOCUMENT RETRIEVAL USING MESH TERMS

Our second series of experiments investigates if any of the automatic classification systems described above can improve IR.

5.1 Experimental setup

The TREC Genomics collections from 2004 to 2007 are used for retrieval experiments (Hersh *et al.* 2004 and onwards). The 2004 and 2005 tasks use a document collection of 4.5 million MEDLINE citations, consisting of a title and optionally an abstract. The 2006 and 2007 tasks use a collection of 160,000 full-text articles from Highwire Press. We only consider document retrieval performance for the 2006 and 2007 tasks, which are originally passage retrieval tasks: documents containing a relevant passage are considered relevant. For the 2004 task, we use the 'title' and 'narrative' of the topic descriptions as queries. In total we have 4 query sets, consisting of 164 queries, on two document collections.

As explained in section 3.5.2 for the second set of experiments we use an automatically obtained MeSH representation of the documents. Since classifying the whole document collection takes rather long, we only used the KNN classifier on the smaller TREC 2006 collection to obtain an automatic conceptual representation.

5.2 Results

Table 3 shows the retrieval performance when using the conceptual query representation obtained from the tested classifiers. *Baseline* indicates the performance of the retrieval system only using the textual representation. The percentages in the table indicate the differences from this baseline. When only the MeSH representation is used (table available in the online supplement), all classification systems perform worse than this baseline (varying from a drop of 32% to 96% in terms of MAP). The KNN classifier performs closest to the baseline, but performance is still poor compared to text-only retrieval (between -32% and -53% MAP). When the textual and conceptual representations are optimally[7] mixed, most of the classifiers don't show significant improvements. KNN forms the notable exception here, where significant improvements (up to 15% MAP) are observed for all query sets. Despite MTI's hybrid approach, it performs slightly better than its major component MetaMap but worse than KNN.

Although the MeSH thesaurus is not the most appropriate choice for improving Genomics retrieval, we do notice that in some cases searching with MeSH terms only improves searching with only text.

[7] using the best-performing α, see online supplement for values

Table 3. Retrieval performance on TREC Genomics collections. † and ‡ indicate a significant difference from the baseline ($p < 0.05$ or 0.005 respectively).

	2004		2005		2006		2007	
Method	MAP	P10	MAP	P10	MAP	P10	MAP	P10
KNN+	.379 +12% ‡	.584 +11% †	.224 +15% ‡	.351 +10% †	.405 +11% †	.465 +2%	.291 +10% †	.469 +4%
MTI+	.352 +4% †	.542 +3%	.208 +7%	.333 +4%	.381 +5%	.442 -3%	.274 +4%	.475 +6%
CLM+	.345 +2%	.520 -1%	.199 +2%	.298 -6%	.364 0%	.458 0%	.267 +1%	.461 +2%
BM25+	.342 +1%	.512 -3%	.195 0%	.318 0%	.363 0%	.458 0%	.268 +1%	.467 +4%
MetaMap+	.341 +1%	.526 0%	.200 +2% †	.318 0%	.364 0%	.442 -3%	.265 0%	.450 0%
EAGL+	.341 +1%	.520 -1%	.198 +2%	.312 -2%	.365 +1%	.477 +4%	.268 +2%	.453 +1%
baseline	.339	.526	.195	.318	.363	.458	.264	.450

Table 4. Retrieval performance on document index based on KNN.

	2006		2007	
Method	MAP	P10	MAP	P10
KNN+	.411 +13%	.504 +10% †	.280 +6% ‡	.472 +5%
EAGL+	.374 +3%	.462 +1%	.273 +3%	.458 +2%
MetaMap+	.372 +2%	.458 0%	.268 +2%	.456 +1%
MTI+	.367 +1%	.454 -1%	.277 +5%	.464 +3%
baseline	.363	.458	.264	.450
CLM+	.363 0%	.458 0%	.264 0%	.464 +3%
BM25+	.363 0%	.458 0%	.264 -0%	.453 +1%

For some topics, the queries can be easily mapped to concepts. For example, "What is the role of Transforming growth factor-beta1 (TGF-beta1) in cerebral amyloid angiopathy (CAA)?" (topic 166), mentions concepts "Transforming Growth Factor beta" and "Cerebral Amyloid Angiopathy". But in many cases the lack of gene and protein name coverage in MeSH hurts retrieval performance. The query text specifically mentions a gene and the representation in MeSH concepts simply misses this key aspect of the query.

The results show that a mixed textual and conceptual representation only improves retrieval if the classification is of high quality. The KNN system clearly outperformed the other systems in the text classification evaluation. In this retrieval setting it is the only system which shows the added value of using the conceptual representation. In only a few cases, topics show a modest drop in performance. In these cases, important words from the query are not represented by concepts from the MeSH thesaurus, leading to query drift.

Table 4 shows the results of using automatic annotation (based on KNN) for the documents in the collection as well. Again, the query representation based on KNN mixed with the textual representation yields optimal performance, and although not all the improvements are significant, using the automatically assigned MeSH terms generated results similar to the ones obtained for the manual MeSH annotations.

6 DISCUSSION

The MeSH classification experiments clearly show the limitations and advantages of using different methods.

The tested thesaurus-only systems (EAGL and MetaMap) are limited in their capability to produce general MeSH terms or terms which are indirectly related. The false-positive analysis underlines that it is easy for the user to link the suggested concepts to the text through the words that they share. Advantages of EAGL are its classification speed and moderate index size. Unfortunately, many general terms are missed and incorrect terms are suggested only on the basis of a partial match with the text to classify.

The concept-oriented classifiers (CLM and BM25) require a large amount of training data but are straightforward to train. The BM25 method performs poorly, probably caused by ineffective parameter settings and its limitations to cope with MeSH documents of different lengths. The CLM system performs on a par with the EAGL system and returns very specific classifications. The false positive analysis confirms that the CLM and BM25 methods return relevant classifications which can only be related indirectly to the text to classify. Again these methods fail to produce general MeSH terms. We expect that a better trade-off between general and specific MeSH terms can be accomplished by adding a prior to the CLM system.

The classification system based on similar documents (KNN) shows the best trade-off between general and specific MeSH terms. It strongly outperforms the other classifiers in reproducing manual annotations. Documents related to the text to classify, yield not only relevant specific MeSH terms, but also very potentially relevant general MeSH terms. In addition, relevant terms are returned which are not explicitly mentioned in the text. Some of the drawbacks include its classification speed (around a second per abstract on a desktop system) and the required index size. Moreover, the classifier will fail to return MeSH terms which are rarely used. Finally, quite a few of the false positives are either irrelevant or incorrect, due to general MeSH terms which are appropriate for related documents, but not for a document in particular.

The false positive analysis might be biased in favor of the thesaurus-oriented classifiers. For both KNN and CLM, it was more difficult to judge a false positive if part of the suggested MeSH term did not occur in the text. This would favor the thesaurus-oriented approaches, since they rely on more explicit overlap. Moreover, we should note that our annotator did not have access to the same information as the annotators responsible for the MEDLINE annotations; the latter are provided with the full-text of the citation under annotation as well.

The second set of experiments shows a clear relationship between classification performance and usefulness for improving information retrieval. Despite the fact that the MeSH thesaurus was not

built specific for Genomics retrieval, it can still be incorporated to improve state-of-the-art text retrieval if the classification performance is of acceptable level. In our experiments this was only the case for the KNN classifier, which by far outperformed the other four classifiers during the classification evaluation.

Using only a conceptual representation results in poor performance, simply because all query aspects cannot be represented in the conceptual language. In case a query can be accurately represented in MeSH terms, improved retrieval performance was observed compared to a text-only representation. This corresponds to earlier results (Schuemie *et al.*, 2007) where a Genomics specific thesaurus was used. A mix between text and concepts however improved retrieval even in cases where the conceptual representation of a query is not complete. The ranking component based on MeSH terms preselects a large group of documents which are more likely to be relevant. The ranking based on the text makes sure that the truly relevant documents are favored, resulting in a higher precision.

Surprisingly, MTI, which includes (a variant of) KNN, performed worse than our implementation of KNN on the document retrieval task. We have three explanations for this. Firstly, MTI has been built to classify new citations rather than old citations, favoring recently introduced MeSH terms. Therefore, using its classifications to find older citations might yield poor results. Secondly, MTI suggests fewer, but likely conforming better to the NLM's indexing practice, MeSH terms than KNN. Thus it might be more useful for suggesting index terms rather than complete search terms. Finally, the poor classification performance of MTI on short input, i.e. only the title of a citation, might explain why its output on the short Genomics queries could not be used to improve IR.

The KNN classifier can be viewed as a form of pseudo relevance feedback in which the top retrieved documents for a query are used for query refinement. In the language modeling framework this has been modeled as relevance models. In this case different representations (text and MeSH) are used, this relates to cross-lingual relevance models in which query and documents are formulated in different languages (Lavrenko *et al.*, 2002). The difference with ordinary cross-lingual retrieval is that both representations are available and they can jointly be used to improve retrieval.

7 CONCLUSION

In this work we tested several MeSH classifiers to do text classification and its use to improve document retrieval.

Classifiers based on only information in the meta thesaurus show to perform comparably to a system which models MeSH terms based on a selection of documents assigned to it. However, a system which automatically annotates text based on the manual annotations of similar documents, strongly outperforms all other approaches. In fact it is the only system which is both highly scalable and capable of improving biomedical information retrieval to the degree observed for manual MeSH annotations.

Further experiments are required to find out whether having a complete MeSH classifier can be applicable to other tasks, such as generating relevance feedback to users of retrieval systems. We are also currently applying our approach to enable MeSH-based phenotypic categorization of micro-array experiments available in Gene Atlas (Parkinson *et al.*, 2009).

FUNDING & ACKNOWLEDGMENTS

This work was supported by a fellowship granted by the Netherlands Genomics Initiative; the BioRange programme of the Netherlands Bioinformatics Centre (supported by a BSIK grant through the NGI); and the EC STREP project BOOTStrep [FP6-028099].

We would like to thank Stephen Robertson for his insightful comments and suggestions.

REFERENCES

Aronson, A. R. (2001). Effective mapping of biomedical text to the UMLS metathesaurus: the MetaMap program. *Proc AMIA Symp*, pages 17–21.

Aronson, A. R., Mork, J. G., Mork, J. G., Gay, C. W., Humphrey, S. M., and Rogers, W. J. (2004). The NLM Indexing Initiative's Medical Text Indexer. In *Proceedings of MEDINFO 2004*, pages 268–272.

Camous, F., Blott, S., and Smeaton, A. F. (2006). On combining MeSH and text searches to improve the retrieval of Medline documents. In *Proceedings of the Third Conference en Recherche d'Informations et Applications (CORIA)*.

Gaudan, S., Yepes, A. J., Lee, V., and Rebholz-Schuhmann, D. (2008). Combining evidence, specificity, and proximity towards the normalization of gene ontology terms in text. *EURASIP J. Bioinformatics Syst. Biol.*, **8**(1), 1–9.

Hersh, W., Buckley, C., Leone, T. J., and Hickam, D. (1994). OHSUMED: an interactive retrieval evaluation and new large test collection for research. In *SIGIR '94*, pages 192–201, New York, NY, USA. Springer-Verlag New York, Inc.

Hersh, W., Bhuptiraju, R., Ross, L., Cohen, A., and Kraemer, D. (2004). TREC 2004 genomics track overview. In *TREC 2004 proceedings*.

Hiemstra, D. and Kraaij, W. (1999). Twenty-One at TREC-7: ad-hoc and cross-language track. In *TREC 7*, pages 227–238.

Kim, W., Aronson, A. R., and Wilbur, W. J. (2001). Automatic MeSH term assignment and quality assessment. *Proc AMIA Symp*, pages 319–323.

Lam, W. and Ho, C. Y. (1998). Using a generalized instance set for automatic text categorization. In *SIGIR '98*, pages 81–89, New York, NY, USA. ACM.

Lam, W., Ruiz, M., Ruiz, M., and Srinivasan, P. (1999). Automatic text categorization and its application to text retrieval. *IEEE Trans. Knowl. Data Eng.*, **11**, 865–879.

Lavrenko, V., Choquette, M., and Croft, W. B. (2002). Cross-lingual relevance models. In *SIGIR '02*, pages 175–182, New York, NY, USA. ACM.

Lewis, D. D. (1998). Naive (bayes) at forty: The independence assumption in information retrieval. In *Proceedings of ECML '98*, pages 4–15.

Lin, J. and Wilbur, W. J. (2007). Pubmed related articles: a probabilistic topic-based model for content similarity. *BMC Bioinformatics*, **8**(1), 423.

Lu, Y., Rosenfeld, R., Simon, I., Nau, G. J., and Bar-Joseph, Z. (2008). A probabilistic generative model for go enrichment analysis. *Nucleic Acids Res*, **36**(17), e109.

Nenadic, G., Spasic, I., and Ananiadou, S. (2004). Mining biomedical abstracts: What is in a term? In *IJCNLP*, pages 247–54., Sanya, China.

Parkinson, H. *et al.* (2009). Arrayexpress update–from an archive of functional genomics experiments to the atlas of gene expression. *Nucleic Acids Res*, **37**(Database issue), D868–72.

Rak, R., Kurgan, L. A., and Reformat, M. (2007). Multilabel associative classification categorization of MEDLINE articles into MeSH keywords. *IEEE Eng Med Biol Mag*, **26**(2), 47–55.

Robertson, S. E., Walker, S., Hancock-Beaulieu, M., Gull, A., and Lau, M. (1996). Okapi at TREC-4. *Proceedings of TREC 1995*.

Ruch, P. (2006). Automatic assignment of biomedical categories: toward a generic approach. *Bioinformatics*, **22**(6), 658–664.

Ruiz, M. E. and Srinivasan, P. (2002). Hierarchical text categorization using neural networks. *Information Retrieval*, **5**(1), 87–118.

Schuemie, M., Trieschnigg, D., and Kraaij, W. (2007). Cross language information retrieval for biomedical literature. In *TREC 2007*.

Smucker, M. D., Allan, J., and Carterette, B. (2007). A comparison of statistical significance tests for information retrieval evaluation. In *CIKM '07*, pages 623–632, New York, NY, USA. ACM.

Sohn, S., Kim, W., Comeau, D. C., and Wilbur, W. J. (2008). Optimal training sets for bayesian prediction of MeSH assignment. *J Am Med Inform Assoc*, **15**(4), 546–553.

Srinivasan, P. (1996). Retrieval feedback in medline. *J Am Med Inform Assoc*, **3**(2), 157–167.

Yu, S., Van Vooren, S., Tranchevent, L.-C., De Moor, B., and Moreau, Y. (2008). Comparison of vocabularies, representations and ranking algorithms for gene prioritization by text mining. *Bioinformatics*, **24**(16), i119–i125.