

Journal of Bioinformatics and Computational Biology
© Imperial College Press

LARGE SCALE BIOMEDICAL CONCEPT MAPPING

PATRICK RUCH

*University Hospital of Geneva
CH-1211 Geneva 5, Switzerland
patrick.ruch@dim.hcuge.ch*

CHRISTINE CHICHESTER and FREDERIQUE LISACEK

*GeneBio SA
CH-1211 Geneva 4, Switzerland
{chichester;lisacek}@genebio.com*

ANNE-LISE VEUTHEY

*Swiss Institute of Bioinformatics
CH-1211 Geneva 4, Switzerland
anne-lise.veuthey@isb-sib.ch*

In this paper, we report on the application of simple retrieval strategies to biomedical concept mapping. We aim at evaluating the performance of a learning-free system tailored to map large collections of concepts, as they can be found in health sciences. Our system is seen as a solution in those cases where machine learning approaches cannot be applied for scalability or data unavailability reasons. For evaluation purposes, the system uses Medical Subject Headings (MeSH) as collection of concepts, and two different collections of MedLine abstracts are used for tuning and evaluation. Unlike most recent text categorization approaches, our approach relies on the combination of two data-independent classifiers. The first classification module uses a tf.idf (term frequency, inverse document frequency) weighting schema, which has been optimally selected for the task. The second classifier is based on regular variations of the concept list. Results emphasize the importance of distinct strategies for minor and major MeSH terms, and show that precision of the hybrid system is significantly improved compared to each single system. For top returned concepts, the system reaches performances comparable to machine learning systems on the OHSUMED collection, while genericity and scalability issues are clearly in favor of the learning-free approach. We draw conclusion on the importance of hybrids strategies for general key words mapping tasks, and discuss the approach in contrast with systems based on machine learning approaches.

Keywords: Text categorization; Terminological Resources; Information Retrieval.

1. Introduction

Systems for text mining are becoming increasingly important in biology because of the exponential growth of knowledge. The mass of scientific literature needs to be filtered and categorized to provide for the most efficient use of the data. The problem of accessing this increasing volume of data demands the development of systems that can extract pertinent information from unstructured text. An important step in this process is automated and optimized keyword assignment. Attributing terms serves to keep literature pointers up-to-date or to filter relevant information from

2 *P. Ruch, C. Chichester, F. Lisacek and A.-L. Veuthey*

the literature. Although this has been the general problem for linguists for decades, molecular biology texts offer the advantage that scientific language is expressed using a smaller vocabulary than common language and the terms are more accurately defined. The Medical Subject Headings (MeSH), originally developed in order to index and retrieve citations in MedLine, is probably the most widely used and largest controlled vocabulary resource available for biological literature. Therefore, the development of a system to attribute the MeSH vocabulary, which now represents a major text mining resource for biological documents (¹³ and ³²), will be an important and non-trivial tool in categorizing scientific literature.

In general the standard, for determining which information should be included in databases, is domain experts. For MedLine in particular, biomedical subject specialists analyze the subject content of articles and describe concepts that are discussed, using the MeSH controlled vocabulary. Although a human curator is clearly superior to any automated system, an automatic information extraction system will be helpful for consistency, updates, and the on-the-fly selection of information during the assignment of MeSH terms. Furthermore, MeSH terms can be used for the purpose of constructing and indexing knowledge representations of biological documents, which can be used as the basis for understanding biological text.

The identification and extraction of specific scientific information has been addressed from different perspectives; machine-selection of articles (¹² and ²⁹), automated extraction using statistical methods ³³, natural language processing techniques ⁶. Systems that do not rely on an ontology, capture specialized nomenclature such as protein and chemical names by using domain specific extracting rules ⁷ or use the matching of pre-specified templates as exemplified as a way for detecting protein-protein interactions ¹¹. Other systems have made use of the availability of molecular biological dictionaries, which have been constructed manually, to extract specific information ²⁴. In our system, we make use of the major biological terminology resource, MeSH, in a classification scheme that combines multiple strategies to optimize the precision of keyword mapping, although other terminological resources, either public or private, could be used, depending on the specific annotation needs. In conclusion, this study provides evidence that automatic keyword attribution can be effective regardless of the terminology chosen.

1.1. *General Strategies and Limits*

Concept Mapping aims at attributing known concepts to a given text. Such systems select one or more concepts into a list to be associated with a document. Typical applications use a set of key-words as concepts to be selected into a *glossary*. However, key-word assignment is only the most popular application of such systems, and the task can also be seen as a named-entity (NE) recognition task if we consider that most entities can be listed^a as it is the case in molecular biology with gene, protein and tissue entities ²⁹. Computer-based concept mapping technologies include:

- *retrieval based on word-matching*, which attributes concepts to text based on shared words between the text and the concepts;
- *empirical learning of text-concept associations* from a training set of texts and their associated concepts.

Retrieval is often presented as the weakest method ³⁹, however there are several areas of applications where training data are clearly missing and where the size of the concept list is some orders of magnitude above the capacities of current learning algorithms. In addition, retrieval-based categorization is mostly independent on the considered list of concepts to be targeted, so that they are supposed to perform similarly whatever terminology they are applied on, at least if these terminologies have similar profile. To our knowledge the largest set ever used by text classification systems is about $2 \cdot 10^4$, and they were all applied to biomedical corpora. In contrast, the May 2002 release of the UMLS (2002AB) contained 871,584 different concepts and 2.1 million

^aIt is not the case in the original NE recognition task definition, which focuses on time, location and names of persons.

terms. The SWISS-PROT Release 40.28 (September 2002) has 114033 entries, and most entries have synonyms, while the TrEMBL Release 21.12 (September 2002) has 684666 entries.

1.1.1. *Concept mapping as a learning-free classification task*

General text classification has been largely studied and has led to an impressive amount of papers (see ⁴⁰ for a recent survey of the domain). A non exhaustive list of machines learning approaches to text categorization includes naive Bayes (Lewis ¹⁹, McCallum and Nigam ²²), k-nearest neighbors (Yang, ⁴⁰), SVM (Joachims ¹⁴), boosting (Shapire and Singer ²⁸), and rule-learning algorithms (Apte and al. ¹). However, most of these studies apply text classification to a small set of classes (usually a few hundreds, as in the paradigmatic Reuters' collection ⁸). In comparison, our system is designed to handle large class sets: retrieval systems are only limited by the size of the inverted file, but 10^{5-6} is still reachable.

Let us note that annotated benchmarks using the above terminologies are rare -however, recent initiatives ⁴ could change the situation- therefore the large-scale task is recasted into a more modest one, for strict evaluation purposes. The set of concepts ranges from about 20 000 -if only unique canonical MeSH terms are taken into account- up to 140 000 -if synonyms are considered in addition to their canonical class.

Last but not least, even if we assume that the availability of large and representative training data will be once solved for biomedicine, current machine learning systems still would have to face major scalability problems. The scalability issue is two folds: it concerns both the ability of the system to work with large concept sets, and its ability to learn and generalize regularities for rare events: Larkey and Croft ¹⁷ show how the frequency of the concept in the collection is a major parameter for learning systems. In this context, there is clearly a room for learning-free concept mapping tools. Word-matching methods combining retrieval strategies can logically constitute an alternative and/or a complement to knowledge- and learning-intensive systems.

1.2. *MedLine collection peculiarities*

Most text categorization studies applied to MedLine citations neglect three important aspects ^b of the MedLine's annotation by MeSH terms that will be considered in the present study:

- a. availability of thesauri: in health sciences, it is frequent that a unique concept can be expressed by several different terms. This point is of major importance for gene and protein entities^c. The MeSH is provided with an important thesaurus (120,020 synonyms), whose impact will be assessed;
- b. comprehensiveness: in the MeSH hierarchical structure ^d, the 19 632 ^e unique terms are split over 32 282 cross-references. Let's note that usually only a fraction of the complete MeSH vocabulary is used (in the OHSUMED collection, almost 30% of MeSH concepts are not represented) whereas our experiments will be conducted on the whole collection of MeSH concepts;
- c. multi-level relevance of terms: MeSH terms are provided in a *ranked* manner since a clear distinction between minor and major terms is made in MedLine citations, therefore separated evaluations are conducted.

The remainder of this paper is organized as follows: the next section presents the collection and metrics used, as well as the basic classifiers and classifiers' combination tested. Section 3 reports on the results of different strategies applied to the Cystic Fibrosis (CF) collection^f, which

^bMeSH statistics are provided for the 1997 release.

^cAs for example, *cathepsin D*, *CTSD*, *CPSD*, *lysosomal aspartyl protease* and *EC 3.4.23.5* refer to the same protein.

^dA survey of hierarchical categorization can be found in ²⁶.

^eIt increases by 13% between 1995 and 1997).

^fAvailable on Marti Hearst's pages at <http://www.sims.berkeley.edu/hearst/irbook/>

4 *P. Ruch, C. Chichester, F. Lisacek and A.-L. Veuthey*

Gene therapy has the potential to cure currently incurable conditions, including some cancers and inherited disorders. It might even be used in the womb to prevent congenital abnormalities. The potential was greeted with great excitement ten years ago, when gene therapy first appeared to be viable, but little progress is perceived. Just how close are we to solving the obstacles in the way of successful gene implantation / replacement?

MeSH Terms:

Animal; Cystic Fibrosis; Gene Therapy; Genetic Vectors*; Human; Liposomes; Retroviridae; Severe Combined Immunodeficiency.*

Fig. 1. Citation of MedLine.

is used for tuning the system, before evaluating it on the OHSUMED collection. Section 4 studies the specificities of MeSH mapping: we pay attention to distinguishing between major and minor terms, and evaluate the use of a thesaurus. Section 5 compares our approach with machine learning systems.

2. Evaluation

Classification techniques yield a ranked list of terms for each document. A purely automatic mapper would need cutoff criteria for the list of candidate terms. Lewis¹⁸ has argued that in evaluating a classification system, one should use effectiveness measures based on estimates of class membership rather than measures based on rankings, like recall-precision curves. Following¹⁷, we do not take this last step of going from a score to binary decision, partly because the correct number of terms for a document can range from 0 to 17, and partly because we do not want to draw any hypothesis on the way these classifiers will be used. They could be used in an interactive system, which would display the selected terms in an ordered list, or as fully-automatic systems, featured with a confidence threshold value. As it is usual with retrieval systems, the core measure for the evaluation is based on the 11 point average precision.

We performed the developments and tuning of the system using the Cystic Fibrosis (CF) collection³⁰, before applying the system on a fragment of the OHSUMED⁹ collection[§] (the first 10 000 abstracts of OHSUMED 1991). The CF collection is a collection of 1239 citations, which are strictly related to cystic fibrosis topics. For each citation, we used the content of the abstract field as input in the system^h. Using other fields, such as the title or the publication's source may have provided interesting additional evidences for classification, but we decided to work only with the abstract in order to minimize the number of variables to be controlled. Although both collections are very similar, the average number of concepts per abstract is higher in OHSUMED than in the CF collection: respectively 14.9 vs. 12.3. Most of the following measures were done considering the top-15 terms returned (TR) by each system.

3. Method

One of the most comprehensive study of MeSH classification based on simple word-matching has been carried at the National Library of Medicine and has led to the development of the MetaMap tool. For developing MetaMap, different methods and combination of methods were compared³, including retrieval strategies (based on INQUERY distance metrics), syntactic and statistical phrase chunking, and MeSH cooccurrencesⁱ. Unfortunately the system has been evaluated on the "UMLS collection", which is not available. We used the UMLS distribution of the MetaMap system with default settings as a blackbox baseline measure for comparison with our system. Table

[§]<http://trec.nist.gov/data.html>

^hBecause we did not want to use the titles we did not select citations which were provided without abstracts. However, in the CF collection we accepted citations provided with extracts.

ⁱThe idea behind cooccurrences' tables, shares some common points with clustering (as it is achieved in k-nearest neighbors), so that MetaMap is not a pure learning-free system.

1 shows the results of MetaMap, we see that it outperforms any basic classifier on the complete Cystic Fibrosis (CF) collection ^j.

3.1. Basic classifiers

Two main modules constitute the skeleton of our system: the regular expression component, and the vector space component. The former component uses tokens as indexing units and can be merged with a thesaurus, while the latter uses stems (Porter-like). Each of the basic classifiers uses known approaches to document retrieval. The first tool is based on a regular expression pattern matcher. Although such approach is less used in modern information retrieval systems ^k, it is expected to perform well when applied on very short documents such as key words: MeSH terms do not contain more than 5 words. The second classifier is based on a traditional vector-space model. This second tool is expected to provide high recall in contrast with the regular expression-based tool, which should privilege precision.

3.1.1. Regular expressions and MeSH thesaurus

The regular expression (RegEx) pattern matcher can be either applied on the simple canonical MeSH (19 936 concepts) collection or on the canonic list augmented with its thesaurus (RegEx+T, 139 956 terms). In this system, text normalization is mainly processed by removing punctuation or by the MeSH terminological resources when the thesaurus is used. Indeed, the MeSH thesaurus provides a large set of “synonyms”, which are mapped to a unique MeSH representative in the canonic collection. Instead of synonyms, this set gathers morpho-syntactic variants (mainly for plural forms), noun phrase reformulations, strict synonyms, and a last class of related terms, which mixes up generic terms, specific terms (for example, *Inhibition* is mapped to *Inhibition (Psychology)*) and some other kinds of less obvious semantic relations. The manually crafted transition network of the finite-state transducer is very simple, as it allows some insertions or deletions within a MeSH term, and ranks the proposed candidate terms based on these basic edit operations following a completion principle: the more terms are matched, the more the term is relevant. The system hashes the abstract into 5 token phrases and moves the window through the abstract. The same type of operations is allowed at the token level, so that the system is able to handle minor string variations, as for instance between *diarrhea* and *diarrhoea*. Table 1 shows that the single RegEx system performs better than any single tf.idf system, what confirm that mapping tasks can be properly performed by simple pattern matching.

3.1.2. Vector space system

The vector space (VS) module is based on a general IR system with *tf.idf* weighting schema^l. In this study, it uses stems (Porter-like, with minor modifications) as indexing terms, and a stop word list. The number of unique strings in the stemmed collection is 12 847. While stemming can be an important parameter, whose impact is a matter of discussion (cf. ¹⁵ and ¹⁰), we did not notice any significant differences regarding mapping effectiveness between the use of tokens and the use of stems, while the index’s size is larger (14 914) when tokens are chosen as indexing units. The graceful behavior of stemming is probably due to the fact that tokens of the biomedical vocabulary are usually longer than in regular English, so that word conflation creates only few confusing stems. However, we noticed that a significant set of semantically related stems should have been conflated in the same indexing units: for example, the morpheme *immun* is found in 48 different stems, and using a morpheme-based word conflation system could have improved the system. Finally, let us note that MeSH terms contain 1 to 5 words, so that, we could have used

^jAvailable at <http://www.sims.berkeley.edu/hearst/irbook/>

^kWith a notable exception, the GLIMPSE system ²¹.

^lWeighting schema and normalization factors uses the *de facto* SMART ²⁷ standard representation, cf. ²⁵ for a short introduction.

6 *P. Ruch, C. Chichester, F. Lisacek and A.-L. Veuthey*

phrases instead of token-based units: thus, Tan and al. ³⁴ used statistical n-grams to enhance text categorization based on support vector machines (SVM), while Tolle and Chen ³⁵ and Aronson ² rely on linguistically-motivated phrases. However, we believe that part of the improvement that could have been brought by using phrases is probably achieved by the RegEx module.

A large part of this study was dedicated to tuning the VS engine, and reporting on the whole list of tf.idf weighting parameters, which was tested, goes beyond the scope of this paper, but the conclusion is that cosine normalization was especially effective for our task. This is not surprising, considering the fact that cosine normalization performs well when all documents are short as is the case of MeSH terms ^m. Thus, in table 1, the top-4 weighting function uses cosine as normalization factor. We also observed that the *idf* factor, which was calculated on the MeSH collection performed well, it means that the canonical MeSH terminology is large enough to effectively underweight non-content words (such as *disease* and *syndrome*). Calculating the idf factor on a large collection of abstracts could have been beneficial regarding classification effectiveness, but such solution may have resulted in making the system more collection-dependent.

3.2. Classifiers' fusion

The hybrid system combines the regular expression classifier with the vector-space classifier. Unlike ¹⁷ we do not merge our classifiers by linear combination, because the RegEx module does not return a scoring consistent with the vector space system. Therefore the combination does not use the RegEx's score, and instead it uses the list returned by the vector space module as a *reference* list (*RL*), while the list returned by the regular expression module is used as *boosting* list (*BL*), and serves in order to improve the ranking of terms listed in *RL*. A third factor takes into account the length of terms: both the character's length (L_1) and the token's length (L_2 , with $L_2 > 3$) are computed, so that long and/or compound terms appearing in both lists are favored over single word terms. We assume that the reference list has exhaustive coverage, and we do not set any threshold on it. For each term t listed in the *RL*, the combined Retrieval Status Value (RSV) is:

$$RSV_{Hybrid} = \begin{cases} RSV_{VS}(t) \cdot Ln(L_1(t)) \cdot L_2(t) \cdot k & \text{if } t \in BL, \\ RSV_{VS}(t) & \text{otherwise.} \end{cases} \quad (1)$$

The value of the k constant (1.6) have been set manually, however more optimal tunings based on parameter's selection could have provided better settings (as proposed in ³⁶). The hybrid system is evaluated with and without the addition of the thesaurus (+T). Table 2 shows that the optimal tf.idf parameters *lnc.atn* for the basic VS classifier does not provide the optimal combination with RegEx. The optimal combination is obtained with *lnc.lnn* settings ⁿ.

The *atn.ntn* setting maximizes the top candidate (TR = 1, i.e. *Precision*_{at Recall=0}) measure, but for a general purpose system, we prefer to maximize average precision, therefore the combination *RegEx + T + lnc.lnn* is applied in the detailed evaluation, since this is the only measure that summarizes the performance of the full ordering of concepts. However, in the context of a fully automatic system, the top-ranked concepts (1 or 2) are clearly of major importance, therefore we also provide this measure.

4. Results

Results of the comparative evaluation of major, minor and both MeSH terms are given for the OHSUMED collection, for the optimal hybrid system (RegEx + VS + T) selected above, with *lnc.lnn* weighting function for the VS system. Figure 2 shows precision and recall curves calculated using both major and minor MeSH terms. Figure 3 provides the same evaluation for minor MeSH

^mAs for more advanced schema, we tested the combination of RegEx with pivoted normalization (Lnu.ltu ³¹, the slope parameter was set to 0.2) and it did not outperform the combination RegEx + lnc.lnn.

ⁿFor the augmented term frequency factor (noted a , which is defined by the function $\alpha + \beta \times (tf/\max(tf))$), the value of the parameters is $\alpha = \beta = 0.5$.

Table 1. Results for MetaMap, RegEx, and (tf.idf) classifiers. weighting schemas. For the VS engine, tf.idf parameters are provided: the first triplet indicates the weighting applied to the “document collection”, i.e. the concepts, while the second is for the “query collection”, i.e. the abstracts. The total of relevant terms is 15193.

System or parameters	Relevant retrieved	Top precision	11pt Average precision
MetaMap	4075	.7425	.1790
RegEx	3986	.7128	.1601
tf.idf			
lnc.atn	3838	.7733	.1421
anc.atn	3813	.7733	.1418
ltc.atn	3788	.7198	.1341
ltc.lnn	2946	.7074	.111
lnc.lnn	2798	.6552	.1055

Table 2. Combining VS with RegEx

Weighting function concepts.abstracts	Relevant retrieved	Top Precision	Average Precision
Hybrids: tf.idf + RegEx			
ltc.lnn	4308	.8884	.1818
lnc.lnn	4301	.8784	.1813
anc.ntn	4184	.8746	.1806
anc.ntn	4184	.8669	.1795
atn.ntn	3763	.9143	.1794

terms, while figure 4 shows curves for major MeSH terms only. In table 4 (discussed in the comparison section), results of the hybrid system with *atn.ntn* are also provided for TR = 1.

As for the general shape of the curves, we see that they are not very smooth, and if precision is high for the low values of recall, it decreases rapidly. As expected the RegEx+VS combination outperforms both the RegEx, the VS and the baseline. Improvement is observed almost at any point of recall. Except for major terms, all curves show the same performances for recall above 30% (i.e after approximately 5 terms).

It is also interesting to notice how the addition of a thesaurus (+T) to the RegEx+VS system results in a moderate degradation of precision for low recall values (for recall = 0 to 0.4), and in general the overall improvement is poor. The inflexion point is located between *Recall* = 0.3 and *Recall* = 0.4 (on figure 2), with precision improving for higher recall values. This result is consistent with most expansion strategy, which tends to improve recall together with reducing precision.

4.1. Minor MeSH terms

Figure 3, in comparison with figure 2, does not demonstrate any additional information except that scores are generally lower for minor terms, as compared to major ones (figure 4). Fortunately, the more relevant terms for human indexers are also more easily extracted by automatic systems. We also can observe that the addition of a thesaurus tends to moderately decrease the precision: at 10% and 20 recall, precision is respectively affected by 1% and 2%.

8 P. Ruch, C. Chichester, F. Lisacek and A.-L. Veuthey

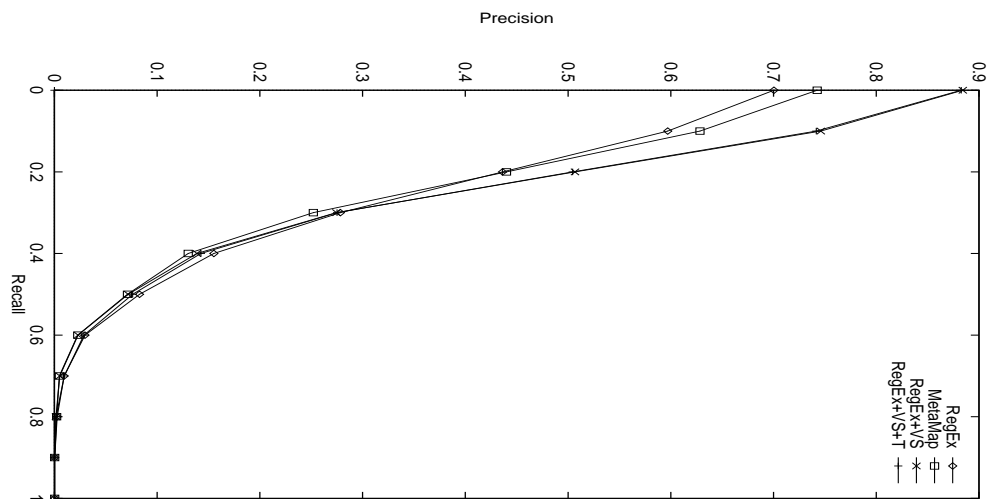


Fig. 2. Comparative recall/precision curves for minor and major MeSH mapping.

MedLine 2003	Cystic Fibrosis Collection
1 CASE REPORT	CASE-REPORT
2 CHLORIDES	CHLORIDES
3 CYSTIC FIBROSIS*	CYSTIC-FIBROSIS*
4 HUMAN	HUMAN
5 INFANT	INFANT
6 KARTAGENER SYNDROME*	KARTAGENER-TRIAD*
7 LUNG	LUNG
8 MALE	MALE
9 SITUS INVERSUS	SITUS-INVERSUS
10 SODIUM	SODIUM
11 SWEAT	SWEAT

Table 3. Diachronism in MeSH terminology for MedLine ID=4545192.

4.2. Major MeSH terms

Surprisingly, the addition of a thesaurus improves term categorization for major MeSH terms. At any point of recall, we observe a modest improvement. It is interesting because in general, the addition of a thesaurus does not improve radically the performance of systems on *ad hoc* retrieval tasks, however under certain conditions the impact of a thesaurus has proven to be beneficial². One possible explanation for such an improvement is that major terms are often explicitly present in the abstract, while, on the opposite, minor terms are less obviously present in the text of the abstracts. The improvement is significant (about 5%), but is hidden when both major and minor terms are considered together in the evaluation: in average, there are at least 4 minor terms for 1 major.

5. Discussion

As mentioned in the introduction, the MeSH terminology is not completely static, and as most terminologies, it sometimes changes over the years. This natural adaptation of the MeSH, which

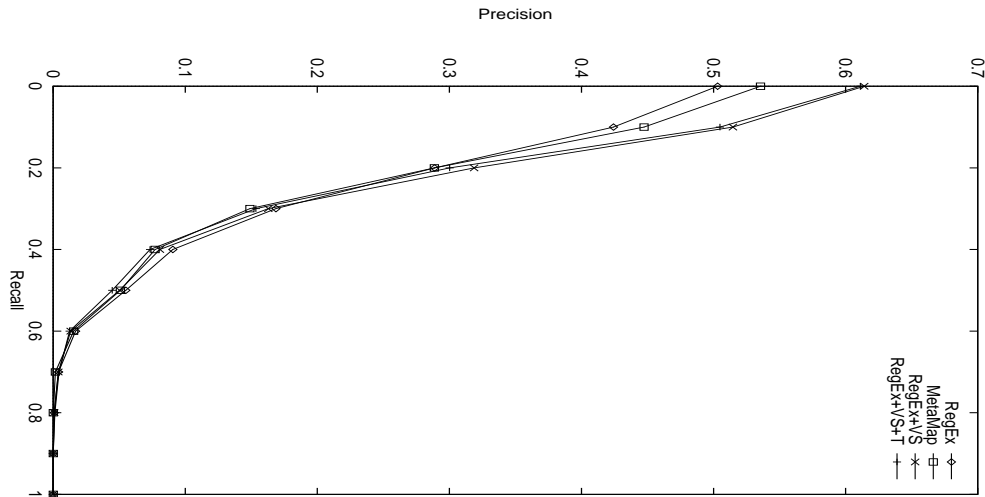


Fig. 3. Comparative recall/precision curves for minor MeSH mapping.

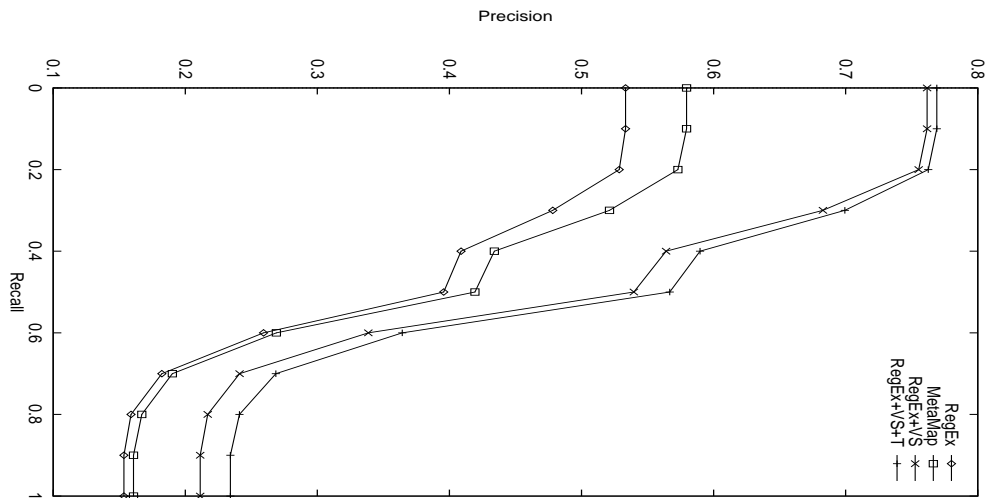


Fig. 4. Comparative recall/precision curves for major MeSH mapping.

follows the emerging of new research fields, may result in some mismatches when working with a collection indexed with a given release of the MeSH terminology. This issue is specific to concept mapping, and does not concern traditional evaluations of text categorization, where categories are controlled and are necessarily relevant for at least one document in the collection. We could

10 *P. Ruch, C. Chichester, F. Lisacek and A.-L. Veuthey*

have tried to automatically check each MeSH term^o, however the validation could have been a labor-intensive step, and instead we decided to assess the miscorrespondence rate. We randomly selected 2 sets: 100 citations from the CF collection (1099 terms), and 100 from OHSUMED (1235) and counted manually the number of terms that are different in MedLine 2003. We found one inconsistency in the latter set, while in the former we counted 3 mismatches: an example of such a phenomenon is illustrated by the 6th term in table 3. Considering such a low level of mismatching (found in 0.2% of citations), we decided not to take this parameter into account, therefore we assume that it does not significantly affect our measures.

5.1. Related results

While several works have concentrated on applying machine learning methods to text categorization, it is often difficult to compare and synthesize the wide quantity of results provided in these studies. One of the main reason is probably that there is no strict definition of the task, which we believe must be seen as a subtask^p rather than a task in itself. Indeed, apart from the central classification problem and the common textual material, which are shared by all these subtasks, there are few common points between them. The gap is well exemplified if we consider on the one side categorization applied to sentence extraction (for example for detecting protein interactions²³) and on the other side concept mapping: while the former work with binary classes, the latter uses virtually infinite sets of classes. Between the two edges, a continuous span of text classification experiments can be identified, whose the most studied -which can also be seen as the paradigmatic ones- are centrally located from some hundreds up to some thousands of classes.

5.1.1. OHSUMED

To the best of our knowledge only two studies have used the entire set of 14,000 MeSH categories^{38 16} used in OHSUMED, and no one ever used the complete 20000-items MeSH terminology that we used, therefore comparison is difficult. The main reason for this is that many categorization methods cannot process such large sets. Yang³⁸, Lewis et al.¹⁸, and Lam and Ho¹⁶ have published results using the subset of categories from the “Heart Diseases” sub-tree (HD-119, so-called because it uses only 119 concepts). In²⁰, 42 categories of the HD sub-tree were excluded, because these categories had a frequency inferior to 15. Yang³⁸ reduces the collection to only those documents that are positive examples of the categories of the HD-119. The final profile of the test collection is very different as the number of terms per abstract is 1.4. She explains that the reason for such reduction is the scalability of the Linear Least Square Fit (LLSF) method, which computes a singular value decomposition, a procedure that cannot be performed efficiently on large matrices. Like us, she uses only citations provided with abstracts, but she also used the title field of citations. Joachims¹⁴ has also published results for the OHSUMED collection using SVM. His work uses the first 20,000 documents of the year 1991 divided into two subsets of 10,000 documents each that are respectively used for training and testing. He reports on very impressive results but his classification task is very different: he assumes that if a category in the MeSH tree is assigned then its more general category in the hierarchy is also present, so that he uses only the high level disease categories. This simplifies the task considerably and may partially explain the good results obtained in these experiments. Chai *et al.*⁵ works with a similar data (63 target concepts !), using neural networks, and reports on similar performances.

Nevertheless, we still attempt to provide some elements for comparing our system with previous studies. The most similar experiment was probably conducted by⁴¹ (noted YC in the following). The authors use a classifier based on singular value decomposition (LLSF) for text categorization. They use the international Classification of Diseases (ICD) as concept list, and full-text diagnosis as instances to be classified. ICD -like the MeSH- contains a large number of

^oEvery replacement of a term by another one is reported in the thesaurus.

^pHowever, document filtering as in TREC-9 is a real task.

categories (about 12 000), and is also provided with an important thesaurus. Both collections are lexically related: we can notice that most of the 6000 diseases listed in the MeSH subtree for diseases have an equivalent in ICD codes, so that ICD can be seen as a more specific partition of the MeSH categories restricted to the “disease” subtree. So assigning ICD codes and MeSH terms are quite similar tasks and supports a possible comparison. Unfortunately only comparison with $Precision_{atRecall=0}$ is available in their study ⁹. We also indicates the results obtained by the SMART system as reported in ³⁷ (noted YY in the following). Even if she works with about 4000 MeSH terms only, this result is useful in order to provide a common baseline measure.

Method/Collection/Paper	Av. Prec.	$Prec_{atRec=0}$
SMART/OHSU/YY	.15 (0)	.61 (0)
LLSF/ICD	-	.840 (+37.7)
ltc.lnn/CFC	.1818 (+20.0)	.8884 (+45.0)
atn.ntn/CFC	-	.9143 (+49.9)

Table 4. Comparison: data-poor vs. learning systems.

Comparison measures are reported in table 4. For top precision (i.e. $Prec_{atRec=0}$), we observe that our hybrid system (+45.9 for *ltc.lnn* and +50.9 for *atn.ntn*) is more efficient than LLSF (+37.7%). Now, regarding average precision, our method outperforms SMART by 25%.

Finally, these results are opposite to what is concluded in ³⁸: simple word-based strategies behaves gracefully when conceptual granularity is fine, i.e. the more concepts there are in a collection, the more effective retrieval strategies are. We can assume that retrieval approaches perform well when categories are numerous, not only because training becomes a major issue for learning systems [†], but because the high granularity may help the retrieval system to cover every dimension of the conceptual space. On the opposite, learning systems are able to infer and cluster categories (generic or specific concepts) that are not explicitly present in the source document, so high granularity does not really help them.

6. Conclusion and future work

We have reported on the development and evaluation of a mapping tool for biomedical purposes. In contrast to traditional approaches, which cannot scale up to the number of concepts defined health sciences terminologies, our method is largely domain-independent, and is expected to perform well when target concepts are numerous. The evaluation has been conducted on MedLine samples, using the MeSH ontology. The systems combines a pattern matcher, based on regular expressions of MeSH terms, and a vector space retrieval engines that uses stems as indexing terms, a traditional *tf.idf* weighting schema, and cosine as normalization factor. The hybrid system showed results similar or better to machine learning tools for the top returned candidate terms, while scalability of our data-poor (if not -independent) approach is also an advantage as compared to data-driven system. The system provides a new baseline for text categorization systems, improving average precision by 20% in comparison to standard retrieval engines (SMART). Finally, combining learning and learning-free systems could be beneficial in order to design general broad-coverage concept mapping systems. While the evaluation has been conducted on a MedLine sample, using the MeSH concepts, we intend to evaluate the system on other ontologies, such as the Gene Ontology.

⁹Again, there are some differences: YC uses a collection of very short texts: the average text had only nine words, and needed to be associated with a unique code, in addition there were many duplicates texts.

[†]Again, this problem is avoided in studies conducted with learning systems by filtering out concepts with low frequencies.

12 *P. Ruch, C. Chichester, F. Lisacek and A.-L. Veuthey*

References

1. C. Apté, F. Damerou, and S. Weiss. Automated learning of decision rules for text categorization. *ACM Transactions on Information Systems (TOIS)*, 12(3):233–251, 1994.
2. A. Aronson. Exploiting a large thesaurus for information retrieval. *Proceedings of RIAO*, pages 197–216, 1994.
3. A. Aronson, O. Bodenreider, H. Chang, S. Humphrey, J. Mork, S. Nelson, T. Rindfleisch, and W. Wilbur. The indexing initiative. A report to the board of scientific counselors of the lister hill national center for biomedical communications. Technical report, NLM, 1999.
4. E. Camon, D. Barrell, C. Brooksbank, M. Magrane, and R. Apweiler. The Gene ontology Annotation (GOA) project - application of GO in SWISS-PROT, TrEMBL and InterPro. *Comp Funct Genom*, pages 71–74, 2003.
5. K. Chai, H. Ng, and H. Chieu. Bayesian Online Classifiers for Text Classification and Filtering. In *ACM-SIGIR*, pages 97–104, 2002.
6. C. Friedman, P. Kra, M. Krauthammer, H. Yu, and A. Rzhetsky. Genies: a natural-language processing system for the extraction of molecular pathways from journal articles. *Bioinformatics*, 17 suppl 1:74–82, 2001.
7. K. Fukuda, A. Tamura, T. Tsunada, and T. Takagi. Toward information extraction: identifying protein names from biological papers. *Pac Symp Biocomput*, pages 707–18, 1998.
8. P. Hayes and S. Weinstein. A system for content-based indexing of a database of news stories. *Proceedings of the Second Annual Conference on Innovative Applications of Intelligence*, 1990.
9. W. Hersh, C. Buckley, T. Leone, and D. Hickam. OHSUMED: An interactive retrieval evaluation and new large test collection for research. In *SIGIR*, pages 192–201, 1994.
10. D. Hull. Stemming algorithms: A case study for detailed evaluation. *Journal of the American Society of Information Science*, 47(1):70–84, 1996.
11. K. Humphreys, G. Demetriou, and R. Gaizauskas. Two applications of information extraction to biological science journal article: enzymes interactions and protein structures. *Pac Symp Biocomput*, pages 505–16, 2000.
12. I. Iliopoulos, A. Enright, and C. Ouzounis. TextQuest: document clustering of MedLine abstracts for concept discovery in molecular biology. *Pac Symp Biocomput*, pages 384–95, 2001.
13. T. Jenssen, A. Laegreid, J. Komorowski, and E. Hoving. A literature network of human genes for high-throughput analysis of gene expression. *Nature Genetics*, 28, 2001.
14. T. Joachims. Making large-scale svm learning practical. *Advances in Kernel Methods - Support Vector Learning*, 1999.
15. R. Krovetz. Viewing morphology as an inference process. *Proc of ACM-SIGIR*, 1993.
16. W. Lam and C. Ho. Using a generalized instance set for automatic text categorization. In *SIGIR*, pages 81–89, 1998.
17. L. Larkey and W. Croft. Combining classifiers in text categorization. In *SIGIR*, pages 289–297. ACM Press, New York, US, 1996.
18. D. Lewis. Evaluating and Optimizing Autonomous Text Classification Systems. In *SIGIR*, pages 246–254. ACM Press, 1995.
19. D. Lewis and M. Ringuette. A comparison of two learning algorithms for text categorization. In *SDAIR*, pages 81–93, 1994.
20. D. Lewis, R. Shapire, J. Callan, and R. Papka. Training algorithms for linear text classifiers. In *SIGIR*, pages 298–303, 1996.
21. U. Manber and S. Wu. GLIMPSE: A tool to search through entire file systems. In *Proceedings of the USENIX Winter 1994 Technical Conference*, pages 23–32, San Francisco

- CA USA, 17-21 1994.
22. A. McCallum and K. Nigam. A comparison of event models for naive bayes text classification, 1998.
 23. C. Nédellec, M. Vetah, and P. Bessières. Sentence Filtering for Information Extraction in Genomics, a Classification Problem. In *PKDD*, pages 326–337, 2001.
 24. T. Ono, H. Hishigaki, A. Tanigami, and T. Takagi. Automated extraction of information on protein-protein interactions from biological literature. *Bioinformatics*, 17:155–61, 2001.
 25. P. Ruch and R. Baud. Evaluating and reducing the effect of data corruption when applying bag of words approaches to medical records. *Int J Med Inf*, 67(1-3):75–83, 2002.
 26. M. Ruiz and P. Srinivasan. Hierarchical text categorization using neural networks. *Information Retrieval*, 5(1):87–118, 2002.
 27. G. Salton. *The SMART Retrieval System - Experiment in Automatic Document Retrieval*. Prentice Hall, 1971.
 28. R. Schapire and Y. Singer. BoosTexter: A boosting-based system for text categorization. *Machine Learning*, 39(2/3):135–168, 2000.
 29. H. Shatkay, S. Edwards, W. Wilbur, and M. Boguski. Genes, themes and microarrays: using information retrieval for large-scale gene analysis. *Proc Int Conf Intell Syst Mol Biol*, 8:317–28, 2000.
 30. W. Shaw, J. Wood, R. Wood, and H. Tibbo. The cystic fibrosis database: Content and research opportunities. *LSIR*, 13:347–366, 1991.
 31. A. Singhal, C. Buckley, and M. Mitra. Pivoted document length normalization. *ACM-SIGIR*, pages 21–29, 1996.
 32. P. Srinivasan. MeSHmap: A text mining tool for MEDLINE. *Proc of AMIA 2001*, pages 103–130, 2001.
 33. M. Stephens, M. Palakal, S. Mukhopadhyay, R. Raje, and J. Mostafa. Detecting gene relations from MedLine abstracts. *Pac Symp Biocomput*, pages 383–95, 2001.
 34. C. Tan, Y. Wang, and C. Lee. The use of bigrams to enhance text categorization. *Information Processing and Management*, 38(4):529–546, 2002.
 35. K. Tolle and H. Chen. Comparing noun phrasing techniques for use with medical digital library tools. *Journal of the American Society for Information Science*, 51(4):352–370, 2000.
 36. J. Vivaldi, L. Marques, and H. Rodriguez. Improving term extraction by system combination using boosting. *ECML*, pages 515–526, 2001.
 37. Y. Yang. Expert Network: Effective and Efficient Learning from Human Decisions in Text Categorization and Retrieval. In W. Croft and C. van Rijsbergen, editors, *SIGIR*, pages 13–22. ACM/Springer, 1994.
 38. Y. Yang. An evaluation of statistical approaches to medline indexing. *AMIA*, pages 358–362, 1996.
 39. Y. Yang. Sampling strategies and learning efficiency in text categorization. In *Proceedings of the AAAI Spring Symposium on Machine Learning in Information Access*, 1996.
 40. Y. Yang. An evaluation of statistical approaches to text categorization. *Journal of Information Retrieval*, 1:67–88, 1999.
 41. Y. Yang and C. Chute. A linear least squares fit mapping method for information retrieval from natural language texts. *COLING*, pages 447–453, 1992.