BIOINFORMATICS ORIGINAL PAPER Vol. 22 no. 6 2006, pages 658–664
doi:10.1093/bioinformatics/bti783

Data and text mining

Automatic assignment of biomedical categories: toward a generic approach

Patrick Ruch*

University Hospitals of Geneva, Medical Informatics Service, CH-1201, Geneva

Received on April 18, 2005; revised on November 11, 2005; accepted on November 13, 2005
Advance Access publication November 15, 2005
Associate Editor: Alfonso Valencia

ABSTRACT

Motivation: We report on the development of a generic text categorization system designed to automatically assign biomedical categories to any input text. Unlike usual automatic text categorization systems, which rely on data-intensive models extracted from large sets of training data, our categorizer is largely data-independent.

Methods: In order to evaluate the robustness of our approach we test the system on two different biomedical terminologies: the Medical Subject Headings (MeSH) and the Gene Ontology (GO). Our light-weight categorizer, based on two ranking modules, combines a pattern matcher and a vector space retrieval engine, and uses both stems and linguistically-motivated indexing units.

Results and Conclusion: Results show the effectiveness of phrase indexing for both GO and MeSH categorization, but we observe the categorization power of the tool depends on the controlled vocabulary: precision at high ranks ranges from above 90% for MeSH to <20% for GO, establishing a new baseline for categorizers based on retrieval methods.

Contact: Patrick.Ruch@sim.hcuge.ch

1 INTRODUCTION

Automatic text categorization (ATC) aims at assigning a set of concepts to an input text. Typical applications use a set of keywords as concepts to be selected from a glossary. Database annotation in genomics and proteomics is also an important application field for categorization tools, which can help curators to select some appropriate categories.

1.1 Retrieval versus learning

From a methodological perspective, computer-based text categorization technologies include:

- retrieval based on string matching, which assign concepts to texts based on shared features (words, stems, phrases...);
- empirical learning of text-concept associations from a training set of texts and their associated concepts.

In the former approach, the targeted concepts are indexed and each indexing unit receives a specific weight, while for the latter, a more complex model of the data is built-up in order to provide text-concept associations beyond strict features sharing. Word-based matching approaches, which include vector-space (Singhal, 2001) and pattern matching engines (Manber and Wu, 1994), are often presented as weak categorization methods (Yang, 1996b) (Yang and Chute, 1992) (Wilbur and Yang, 1996), because associations between text and categories are based on simple string matching strategies, but in several situations learning approaches cannot be applied. With the explosion of concepts in molecular biology and life sciences in general, we believe the use of ranking-based methods, and their combinations, which are computationally cheaper and simpler than binary classifiers (Amini *et al*., 2005) should be revisited.

1.2 Categorization by ranking

Designing the categorization as a retrieval task means that the engine has to index the collection of terms of the vocabulary as if they were documents and then it treats each input document as if it was a query. Then, the tool uses the score (called retrieval status value) attributed to each term to rank them. So, unlike for binary categorization, which tries to decide whether a concept is relevant or not, we do not try to replace the judgement of the curator and instead, in our definition of the task, concepts are simply ranked by order of relevance. Like for document retrieval, the curator can screen through the returned categories to decide whether they are of interest or not.

Because the document collection is made of terminological entities that are clearly shorter than usual documents, the study aims at exploring the behavior of retrieval statistical models. The use of a vector space engine and its combination with a search tool based on pattern matching are investigated. The outline of the paper is as follows: after presenting the research background in Section 2, we describe the architecture of the system in Section 3 together with results measured for each combination of our system; related results are discussed in section 4; conclusions are presented in Section 5.

2 BACKGROUND

To our knowledge the largest set of categories ever used by text classification systems has an order of magnitude of 10^4. Thus, Yang and Chute (1992) work with the International Classification of Diseases (~12 000 concepts), while Yang (1999) and Wilbur and Yang (1996) report on experiments conducted with a search space of less than 18 000 Medical Subject Headings (MeSH). To evaluate our system, it is tested using two different benchmarks: (1) the OHSUGEN (Hersh, 2005) collection for the MeSH terminology and (2) the BioCreative data for Gene Ontology (GO). GO is currently the main controlled-vocabulary for molecular biology. MeSH

*To whom correspondence should be addressed.

is a more general glossary as it covers also medical and clinical fields, but has been acknowledged as an important resource for text mining in the domain (Shah *et al*., 2003).

2.1 Scalability issues

General purpose machine learning methods might be inappropriate for some ATC tasks in biomedical terminologies because reliable training data are often not available (Camon *et al*., 2003). To some extent, this statement can be applied to MeSH as well: between 2004 and 2005, 487 new headings were introduced, while 60 were deleted and 129 were modified, so about two concepts are added every day.

In contrast, our approach is data-economic, because it only demands a small collection of annotated texts for fine tuning the statistical model.

2.2 Features normalization

In information retrieval, as well as in ATC, the basic feature is the word, or a normalized variant of the word, such as the stem. However, various phrase indexing methods have been proposed in the past to go beyond the so-called bag-of-words representation, which assumes that the order of words in a document can be neglected. In the language of probability theory, this is an assumption of exchangeability for the words in a document (Aldous, 1985), which is intuitively wrong. Unfortunately, retrieval or categorization performance conclusions on the use of phrases as indexing units are inconsistent (Rasolofo and Savoy, 2003). Thus, for the 2003 Trec Retrieval Conference TREC genomics *ad hoc* retrieval task, de Bruijn and Martin (2003) reported lower retrieval effectiveness when word bigrams were used, while Kim *et al*. (2001) and Aronson *et al*. (2005) report that recognizing MeSH phrases does help retrieval in MEDLINE. As for our present concerns, we test the use of noun phrases rather than statistical phrases. Indeed, usually inspired by mutual information measures (Stolz, 1965), statistical extraction of phrases requires important volumes of training data, while we aim at designing a data-independent system[1].

2.3 Collection and metrics

The majority of experiments made with machine learning approaches in a standard computational environment, applies text classification to a small set of classes; usually a few hundreds. In contrast, our system is designed to handle large class sets: retrieval tools are only limited by the size of the inverted file, but 10^6 is still a modest range. Because there is no benchmark with such a large set of categories, our evaluations are conducted on smaller scales[2]. The search space of our system ranges from 19 936 MeSH categories, if only unique canonical MeSH terms are taken into account, up to 139 956, if synonyms are considered in addition to their preferred representatives. The three other sets of concepts are provided by GO, which gathers three different subvocabularies. Each GO classifier corresponds to the mutually exclusive axes of the GO (Table 1): cellular components (1711 items with synonyms), molecular functions (18 106 items with synonyms) and biological processes (9604 items with synonyms). As usual for retrieval systems, the main evaluation measure is based on the mean average precision (MAP), since this is the only measure that summarizes the performance of the full ordering of concepts. However, top ranked concepts are clearly of major importance, therefore we also provide the $\text{Precision}_{\text{at Recall}=0}$ (P0), which measure the precision of the top returned category; see Cooper (1971) for an introduction on retrieval metrics.

Table 1. MesH terms and GO categories for an abstract (PMID = 9506968) describing the Cyclin-dependent kinase 2-associated protein 1

MeSH Terms	Amino Acid Sequence; Animals; Catalysis; Cells, Cultured; Chromosome Mapping; Chromosomes, Human, Pair 12; Cloning, Molecular; DNA Polymerase I; DNA Primase; DNA Replication; DNA, Complementary; Genes, Tumor Suppressor*; Hamsters; Humans; Molecular Sequence Data; Mutation; Proteins*; Sequence Homology, Amino Acid; Tumor Cells, Cultured; Tumor Suppressor Proteins*
GO Annotation	
Functions	DNA binding
Processes	S phase of mitotic cell cycle; DNA dependent DNA replication; protein amino acid phosphorylation
Components	Nucleus; cytoplasm

Major MeSH are marked with *; check tags and subheadings are removed.

2.3.1 MeSH assignment The OHSUGEN collection contains 4 591 015 MEDLINE citations. We extracted two randomly-selected sets of citations: set A (500 items) is used for tuning the system, set B (1000 items) is used to evaluate the system. Only citations provided with an abstract were selected. For each citation, we merge the content of the abstract field with the content of the title field. We do not distinguish between major and minor MeSH terms (cf. Table 1). Experiments were done using the top 15 terms returned by the engine, which is the average number of keywords in MEDLINE citations.

2.3.2 GO assignment For assessing the GO categorizer, we rely on the BioCreative benchmark (Hirschman *et al*., 2005). An initial set of 640 articles (called ALL-GO) from the Journal of Biological Chemistry, was provided by the organizers, 320 articles were used for tuning our tools (A-GO) and the other half was used for our evaluations (B-GO). Only abstracts and titles of the articles are used. An example of the GO annotation is given in Table 1. The number of GO terms per protein in BioCreative data, which are a sample of Swiss-Prot, is extremely variable and ranges from 1 to 33 (Ehrler *et al*., 2005), but following the experimental design of the BioCreative competition for the GO categorization we assume that the number of expected categories per axis is a priori known in our experiments.

3 METHODS AND RESULTS

In this section, we present the basic modules and the strategies, which were chosen to merge these basic modules. Results reported in this section were computed on the evaluation sets (sets B-MeSH and B-GO). Tuning experiments, which include varying the

[1]However, data needed to extract statistical phrases are not of the same kind as those needed for training a classifier: the former approach requires only large corpora, while the latter needs manual annotation, so both tasks are data-dependent but statistical phrase extraction is much cheaper than supervised text categorization.

[2]In the following, statistics are given for September 2003 releases.

Table 2. Results of REx and VS classifiers for automatic assignment of MeSH terms

System or parameters	Relevant retrieved	Prec. at Rec. = 0	Av. Prec.
REx	2842	0.7168	0.1655
VS			
ltc.atn	2736	0.5752	0.0653
ltc.lnn	2701	0.5862	0.0557

The total number of relevant terms is 12591.

Table 3. Results of REx and VS classifiers, averaged for each GO subgraph

System or parameters	Relevant retrieved	Prec. at Rec. = 0	Av. Prec.
REx	104	0.1469	0.0691
VS	100	0.1523	0.0595

The total number of relevant terms is 1607. The VS system applies the following weighting profiles: *ltc.atn* for molecular functions, *ltc.atn* for cellular components and *atc.atn* for biological processes.

different weighting schema of the vector space ranker to compute the optimal combination factors[3], were conducted on the tuning sets (A-MeSH and A-GO). Table 2 shows results for the MeSH categorizer. Table 3 reports averaged results for the three axes of the GO.

Two main modules constitute the skeleton of our system: the regular expression (REx) component, and the vector space (VS) component. The former component uses tokens as indexing units, while the latter uses stems (Porter). The first tool is based on a regular expression pattern matcher. it is expected to perform well when applied on very short textual segments such as MeSH keywords or GO categories. This second tool, based on a vector space model, is expected to provide high recall in contrast with the regular expression-based tool, which should favor precision.

For the VS module, different combination of the weighting factors were tested to obtain the best schema for the task. We used the SMART notation to represent our statistical model (Ruch and Baud, 2002): the first triplet letter 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 first parameter of the triplet refers to the term frequency (n: real, l: logarithmic or a: augmented), the second parameter refers to the inverse document frequency (n: no inverse document frequency factor; t: inverse document frequency) and the third parameter refers to the length normalization (n: no normalization; c: cosine). We observe that the term frequency applied on the collection of concepts can be regarded as constant, since in general an indexing term appears only once in a given category[4].

[3]See (Ruch and Baud, 2002) and (Singhal, 2001) for a formal description.
[4]There are a few exceptions, like in DNA dependent DNA replication, where DNA appears twice.

3.1 Ranking based on pattern matching

This module does not use any specific string normalization and settings are similar for MeSH and GO categorization. The system extracts every contiguous sequence of N tokens by moving a window through the abstract. The value of N is empirically set to 5, which is the maximum number of tokens in a MeSH terms. This number can be higher for GO terms, but 80% of GO terms contain four words or less than four words (Ehrler *et al*., 2005). Pentagrams are then matched against the collection of terms. Basically, the manually crafted finite-state automata allow two insertions or one deletion within a term, and ranks the proposed candidate terms based on these basic edit operations: insertion costs 1, while deletion costs 2. The same type of operations are allowed at the string level, so that the system is able to handle minor string variations, as between diarrhea and diarrhoea. String variations are only computed on tokens that have more than 8 characters to avoid string confusion. A description of the string edit distance algorithm can be found in Ruch (2002). The resulting pattern matcher behaves like a term proximity scoring system (Rasolofo and Savoy, 2003), but with a 5 token matching window.

3.2 Ranking based on retrieval

The engine uses stems as indexing units, and a stop word list (544 items). As for setting the weighting factors, we observe that cosine normalization (expressed by the *c* letter) was especially effective for our task, which is consistent with the fact that cosine normalization tends to perform well when all documents have similar length (Singhal *et al*., 1996).

In table 2, we report results obtained by two of the best schemas: *ltc.atn* and *ltc.lnn*. *ltc.atn* performs better regarding the average precision, but *ltc.lnn* is slightly better for precision at high ranks. As for the average precision of each basic module, Tables 2 and 3 show that the REx system performs better than any *tf.idf* schema used by the VS engine, so regular pattern-matchers provide better average precision than VS engines: for MeSH, REx = 0.1655 versus *ltc.atn* = 0.0653; for GO, REx = 0.0691 versus VS = 0.0595). Regarding the number of relevant categories proposed by each system (column Relevant retrieved), which provides an estimate of the recall, we observe that for MeSH categories, the best VS schema retrieves 2701 relevant terms, while REx retrieves 2842 relevant terms. For GO categorization, the REx modules performs better regarding average precision (REx = 0.0691 versus VS = 0.0595) and global recall (104 relevant categories for REx versus 100 for the VS module), but not regarding precision at high ranks (VS = 0.1523 and REx = 0.1469). This differences suggest that the two retrieval methods might be complementary, and so combining the two approaches might result in a better system.

Looking at the respective performances on the two different vocabularies, these tables show that assigning MeSH keywords is easier than assigning GO categories. The precision of the top proposed GO category is only 15% (P0 = 0.1523 for VS) versus 70% for MeSH concepts (P0 = 0.7168 for *ltc.atn*). Obviously, returning a relevant category out of 15 possible MeSH keywords is easier than out of two or three GO categories, but this statement is also consistent with the nature of the two terminological systems: MeSH terms are intended to express textual contents, while GO concepts express biological descriptions.

3.3 Terminological resources

Both the MeSH and the GO vocabulary provide a large set of synonyms, which are linked to a unique representative (the preferred term) in the vocabulary. Synonyms provided in the GO are of good quality (for example: protoplasm/lintracellular; cell division/lcytokinesis) and can be used to expand the matching power of our tools without introducing any additional noise. In contrast, we remark that the MeSH thesaurus gathers morpho-syntactic variants, real synonyms and a last class of related terms, which mixes up generic and specific terms. For instance, Inhibition is mapped to Inhibition (Psychology). To solve this issue, a dozen of obvious confusing synonyms were manually removed from the MeSH thesaurus during the tuning procedure.

When synonyms are used, they are indexed as if they were different concepts. The normalization step removes synonyms from the proposed ranked list of terms. Indexing synonyms implies that a unique concept can be found at different ranks in the list of retrieved terms, so to ensure the uniqueness of the concept, only the first occurrence of the concept is kept and the following occurrences are deleted.

3.4 Linguistically-motivated phrases

GO terms have a more variable length—between 1 and 28 tokens—than MeSH terms but each terminology contains almost verb-free noun phrases (NP), if we ignore some rare participle forms; therefore NP indexing is expected to be beneficial for both vocabularies.

Our shallow parser combines statistical and manually written patterns. Patterns are applied at the syntactic level (part-of-speech) of each sentence (Ruch *et al*., 2000). The parser concentrates on adjective (A) and noun (N) sequences, such as: [A*][N*], i.e. N, AN, NN, ANN, NNN, AANN, ANNN, NNNN... Adjectives as well as prepositions such as of or with are optional. Apart from adjectives and nouns, we counted 1376 conjunctions (mainly and and or) in the MeSH, including the thesaurus, i.e. 1% out of 139 956 items. The GO vocabulary is syntactically more complex and the proportion of conjunctions increases to 2%. Nevertheless, unlike in some technical vocabularies (Park *et al*., 2002), which may need more advanced linguistic methods (Gaizauskas, 2003), this proportion means that patterns with conjunctions are rare both in MeSH and GO items, so we decided to simply ignore them and we assume that long distance term dependencies will be handled by the bag-of-words model of the VS module.

The identification of phrases is based on the input query, which merges together the title and the abstract. Our working hypothesis is a weak variant of the Phrase Retrieval Hypothesis (Arampatzis *et al*., 2000): we assume that NP recognition can help reducing *noisy mapping* for *subterms*. We call noisy subterm mapping the erroneous behavior of the mapping process, when it selects some erroneous terms that are subpart of a relevant one. Thus, a text dealing with cystic fibrosis is relevantly indexed by the term cystic fibrosis, while fibrosis is irrelevant. However, discarding all subterms from the candidate list may have negative effects, therefore subterm removal must be based on additional evidence. The category is removed only if it does not occur in the set of NPs extracted from the abstract. The way this index of NPs fuses with the index of stems is described in the next paragraph.

Table 4. Combining VS with REx for MeSH categorization

Weighting function concepts.abstracts	Relevant retrieved	Prec. at Rec. = 0	Av. Prec.
VS + REx			
ltc.atn	3073	0.9202	0.2073
ltc.lnn	2856	0.9110	0.1991

3.5 Fusion of basic modules

The first combination merges the REx and the VS module. This new list of candidates is then compared with the NP index to produce a final ranked list of categories.

3.5.1 Combination of rankers The hybrid system combines the regular expression classifier with the vector-space classifier. Because the REx module does not return a scoring consistent with the vector space system, we do not merge our classifiers by linear combination Larkey and Croft (1996). The combination 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), which serves to improve the ranking of terms listed in RL. A third factor takes into account the length of terms: both the number of characters (L_1) and the number of tokens (L_2, with $L_2 > 3$) are computed, so that long and compound terms, which appear in both lists, are favored over single and short terms. We assume that the reference list has good recall, and we do not set any threshold on it. For each concept t listed in the RL, the combined Retrieval Status Value (cRSV, Equation 1) is:

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

The value of the k parameter is set empirically by direct search on the tuning sets. The objective function we maximize is the mean average precision. The combined system is evaluated with and without the thesaurus (+T). For MeSH the simple combination of VS and REx significantly[5] improves (with $P < 10^{-6}$) the average precision of the tool: from 0.1655 (Table 2) for the REx module alone to 0.1991 (+20%, Table 4) for the combination *ltc.atn* + REx. For GO, the VS + REx combination achieves a MAP = 0.0753 (Table 5) versus 0.0691 (Table 3) for REx alone, i.e. + 9%. This confirms that REx and VS are complementary. In Table 5, we can see that the impact of synonyms for the MeSH categorization is rather modest (+0.2%). The impact of the GO thesaurus is more significant (+3.45%, Table 5). A possible explanation for these differences can be that GO synonyms are more focused than MeSH synonyms, which may introduce misleading associations between concepts. Indeed, several abbreviations proposed as MeSH synonyms are likely to have a particular meaning in genomics. Thus, *ret* is used as abbreviation for retired, while it also refers to the ret proto-oncogene. For several of these acronyms, contextual disambiguation (Pustejovsky *et al*., 2001) may be necessary.

[5]Tests are computed using a non-parametric signed test, cf. (Zobel, 1998) for more details.

Table 5. Comparison of different combinations on the evaluation sets

Combination	Relevant retrieved	Prec. at Rec. = 0	Av. Prec.
MeSH			
Baseline + T + NP	3075	0.9118	0.2117 (+2.1%)
Baseline + NP	3068	0.9205	**0.2130 (+2.8%)**
Baseline + T	3075	0.9051	0.2079 (+0.2%)
Baseline	3073	0.9202	0.2073 (100%)
Gene Ontology			
Baseline + T + NP	112	0.1711	**0.0802 (+6.51%)**
Baseline + T	110	0.1711	0.0779 (+3.45%)
Baseline	105	0.1696	0.0753 (100%)

The baseline is given by the combination VS + REx. Top performing combinations are in bold.

3.5.2 Using noun phrases The index of phrases is used to reorder the set of terms returned by the engine. The strategy is the following: when a given term is found in the list of terms (TL) returned by the hybrid system (REx + VS), and this term is not found alone in the phrase list (PL) generated from this abstract, then the RSV of this concept is downscored. The shorter the subterm, the more its RSV is affected, as expressed in Equation (2), which gives the final RSV (fRSV; m is the maximal number of tokens per term in the vocabulary):

$$fRSV = \begin{cases} \dfrac{cRSV}{m - L_2(t)} & \text{if } t \in TL \text{ and } t \notin PL \\ cRSV_1 & \text{otherwise} \end{cases} . \quad (2)$$

For MeSH, in Table 5, we observe that the NP index improves the average precision by up to 2.8%. The improvement is statistically significant ($P < 10^{-6}$). The use of the thesaurus brings no significant improvement (+0.2%), while it degrades the categorization effectiveness when used with NP indexing. As expected with query expansion in general and thesaurus in particular, using a thesaurus means that we trade recall for precision (Hirschman *et al*., 2005), it is particularly true at P0 when phrase indexing is not used: from 0.9202 to 0.9051 (−1.7%). These contrasts validate our architectural choices regarding the integration of the NP index, since the proposed combination is effective. For GO categorization, the impact of NPs and synonyms is even stronger than for MeSH terms: Table 5 shows that the overall improvement goes up to 6.51%.

4 DISCUSSION

Comparison with the state-of-the-art is difficult because information retrieval methods have rarely been used for text categorization and also because studies based on supervised learning cannot be directly compared to our approach.

4.1 MeSH assignment

As for MEDLINE collections and MeSH categorization, they have been used by very few researchers for text categorization. Most of these studies were carried on a tiny fraction of the MeSH, using the OHSUMED collection (Hersh *et al*., 1994). Lewis (1995) has published results using the subset of categories from the 'Heart Diseases' sub-tree of the MeSH (so-called HD-119, because the search space is then reduced to 119 categories). In Lewis *et al*. (1996), 42 categories of the HD sub-tree were excluded because they occurred only 15 times in the training set. Yang (1996a) reduces the collection to only those documents that are positive examples of the categories of the HD-119. The resulting test collection has 1.4 concepts per abstract, versus about 15 in our experiments. Joachims (1999) has also published results for the OHSUMED collection using support vector machines, but he uses only the high level disease categories, i.e. 20 concepts. These studies achieve a precision up to 65%. More comparable regarding the scales, Yang and Chute (1992) and Wilbur and Yang (1996) report results ranging from 0.34 to 0.40 for the average precision and about 0.85 for the top precision, which makes our simpler approaches competitive with trained systems for precision at high ranks. But the only directly comparable result concerns their baseline method, which uses the SMART retrieval engine, and which achieves ∼30% of the average precision of our best combination.

In any case, direct comparison should go beyond classification performances: while sufficiently trained systems would in principle outperform any simple retrieval system, other important aspects such as availability of training data, overfitting and complexity[6] should be considered.

4.2 GO assignment

Although direct comparison with other participants of the BioCreative challenge is difficult because: (1) official evaluations were based on utility measures and human judgements and (2) a short segment of text was also to be provided to support the assignment of the category, our lightweight methods achieved competitive recall and precision ratio in this competition; see Ehrler *et al*. (2005) for a detailed presentation and Couto *et al*. (2004) and Hirschman *et al*. (2005) for a synthetic comparison of the different methods and results.

4.3 Qualitative evaluation: questioning metrics

An example of the automatic MeSH and GO assignment as proposed by the tool is given in Table 6. The expected categories are provided in Table 1. MeSH categories proposed on the top of the list were expected and marked as major: tumor suppressor proteins; genes, tumor suppressor. Some other relevant concepts, such as species (hamsters) or more specific like DNA Polymerase I, DNA Replication, Mutation are provided, but we also observe that suggested categories are often too generic. Thus, chromosome is proposed instead of Chromosome Mapping; Chromosomes, Human, Pair 12. These observations apply to GO categorization as well: cellular components (cytoplasm and nucleus) are good categories in this example. While nucleus does not occur in the abstract, the stemming is able to associate nuclei to nucleus. Like for MeSH, genericity is problematic: DNA replication is proposed

[6]Although some Bayesian classifiers (Domingos and Pazzani, 1997) (Ruch *et al*., 2005) have linear complexity, most algorithms have a quadratic complexity. Thus, the so-called scalable implementation of Support Vector Machines proposed by Joachims (1999) needed three weeks to train a classifier tailored to discriminate categories listed in a subset of the Cardiovascular Diseases sub-tree, i.e. less than 100 concepts. Transporting such an approach to the GO and disregarding the fact that annotated data are not available would need approximately 20 years!

Table 6. Ranked results for an example of automatic assignment: in parenthesis the retrieval status value (RSV) of the term expresses a similarity

Automatic GO Assignment
Components: (0004643) nucleus; (0006575) cytoplasm
Functions: (0050669) tumor suppressor
Processes: (0008697) phosphorylation; (0008722) regulation of dna replication; (0028143) dna replication
Automatic MeSH Assignment
(0001528) mutation; (0001685) keratinocytes; (0001699) cell cycle; (0001762) chromosomes; (0001839) cytoplasm; (0002456) dna, complementary;
(0002580) phosphorylation; (0002769) dna polymerase i; (0003794) hamsters; (0004008) long interspersed nucleotide elements; (0008590) dna replication;
(0009172) tumor suppressor proteins; (0010780) sequence analysis; (0018097) polymerase chain reaction; (0019233) genes, tumor suppressor.

instead of DNA dependent DNA replication, and phosphorylation is proposed while protein amino acid phosphorylation is expected.

From a user perspective, the reported precision at high ranks means that more than nine MeSH categories out of ten are relevant, while only one GO category out of five is relevant. Finally, it would be interesting to question the quality of our benchmarks. Inter-annotator studies on the subject are rare, but Funk and Reid (1983) report that a 40% agreement should be regarded as a good score, suggesting that any precision above that score might be the result of some overfitting phenomena. In the same vein, current categorization metrics are not able to use the hierachical information embedded in terminologies, although some errors are less irrelevant than others: thus, if the term Rats is proposed instead of Rattus Norvegicus, it is still a better match than Mammals, which is better than Animals!

5 CONCLUSION

We have reported on the development of a generic categorization system. The systems combines a pattern matcher and a vector space retrieval engine, which uses both stems and NPs. The addition of synonyms to handle polysemy had minor effect on the MeSH categorization task but higher effect on GO categorization. The use of phrases significantly improve the categorization's average precision, both for MeSH and GO assignment. From a comparative perspective, the MeSH categorizer shows results competitive with machine learning tools for top returned concepts and establish a new baseline for retrieval methods. For GO categories, precision is generally lower than for MeSH categories.

ACKNOWLEDGEMENTS

The study has been supported by the EU (SemanticMining IST 507505 - OFES 03.0399) for the Gene Ontology categorizer and by the SNF (3252BO-105755/1) for the MeSH categorizer. Part of the study has been supported by the NIH, via an ORISE visiting faculty grant at the National Library of Medicine. I would like to thank Alan 'anti-B' Aronson, Ananta Bangalore, Cliff Gay, Will Rogers for their daily support, as well as Dina Demner-Fushman, Hong-Fang Liu, Miguel Ruiz, Larry Smith, Lorrie Tanabe and John Wilbur, for the fruitful discussions during our weekly LHC/NCBI meetings.

Conflict of Interest: none declared.

REFERENCES

Aldous,D. (1985) Exchangeability and related topics. In *École d'été de probabilités de Saint-Flour, XIII—1983, Volume 1117 of Lecture Notes in Mathematics*. Springer, Berlin, pp. 1–198.

Amini,M., Usunier,N. and Gallinari,P. (2005) Automatic text summarization based on word-clusters and ranking algorithms. In *Proceedings of ECIR '05*, Santiago de Compostela, Spain, pp. 142–156.

Arampatzis,A. *et al.* (2000) Linguistically motivated information retrieval. In *Encyclopedia of Library and Information Science*. Marcel Dekker, Inc., Basel, NY, **69**.

Aronson,A. *et al.* (2005) Fusion of knowledge-intensive and statistical approaches for retrieving and annotating textual genomics documents. In *TREC Notebook Proceedings of the Conference of TREC*, November 15–18, Gaithersburg, MD, USA, pp. 36–45.

Camon,E. *et al.* (2003) The Gene Ontology Annotation (GOA) project: implementation of GO in Swiss-Prot, TrEMBL and InterPro. *Genome Res.*, **13**, 562–672.

Cooper,W. (1971) A definition of relevance for information retrieval. *Inf. Storage Retr.*, **7**, 19–37.

Couto,F. *et al.* (2004) FIGO: findings GO terms in unStructured text. *BioCreative Notebook Papers, CNB 2004*.

de Bruijn,B. and Martin,J. (2003) Finding gene functions using litminer. In *Proceedings of the 12th Text Retrieval Conference (TREC), NIST*, Gaithersburg, MD, pp. 451–459.

Domingos,P. and Pazzani,M. (1997) On the optimality of the simple bayesian classifier under zero-one loss. *Mach. Learn.*, **29**, 103–130.

Ehrler,F. *et al.* (2005) Data-poor Categorization and Passage Retrieval for Gene Ontology Annotation in Swiss-Prot. *BMC Bioinformatics*, **6**(Suppl. 1).

Funk,M. and Reid,C. (1983) Indexing consistency in medline. *Bull Med. Libr. Assoc.*, **71**, 176–83.

Gaizauskas,R. (2003) Recent advances in computational terminology. *Comput. Linguist.*, **29**, 328–332.

Hersh,W. (2005) Report on the TREC 2004 Genomics track. *SIGIR Forum*, **39**, 21–24.

Hersh,W. (1994) OHSUMED: an interactive retrieval evaluation and new large test collection for research. In *Proceedings of the 17th Annual International ACM-SIGIR Conference on Research and Development in Information Retrieval,* Dublin, Ireland, pp. 192–201.

Hirschman,L. *et al.* (2005) Overview of BioCreAtIvE: critical assessment of information extraction for biology. *BMC Bioinformatics*, **6**(suppl. 1).

Joachims,T. (1999) Making large-scale SVM learning practical. In Schölkopf,B., Smola,B. and Alex,J. (eds), *Advances in Kernel Methods—Support Vector Learning*. MIT Press, pp. 1–16.

Kim,W. *et al.* (2001) Automatic MeSH term assignment and quality assessment. *Proc. AMIA Symp.*, 319–23.

Larkey,L. and Croft,W. (1996) Combining classifiers in text categorization. In *SIGIR '96: Proceedings of the Nineteenth Annual International ACM SIGIR Conference on Research and Development in Internation Retrieval*. ACM Press, NY, pp. 289–297.

Lewis,D. (1995) Evaluating and optimizing autonomous text classification systems. In *Proceedings of SIGIR-95, 18th ACM International Conference on Research and Development in Information Retrieval*. ACM Press, NY, pp. 246–254.

Lewis,D., Schapire,R., Callan,J. and Papka,R. (1996) Training algorithms for linear text classifiers. *In Proceedings of SIGIR-96, 19th ACM International Conference on Research and Development in Information Retrieval*. ACM Press, NY, pp. 298–303.

Manber,U. and Wu,S. (1994) GLIMPSE: a tool to search through entire file systems. In *Proceedings of the USENIX Conference*. San Fransisco, CA, pp. 23–32.

Park,Y., Byrd,R. and Boguraev,B. (2002) Automatic glossary extraction: beyond terminology identification. *In 19th International Conference on Computational Linguistics*, Taipei, Taiwan.

Pustejovsky,J., Castano,J., Cochran,B., Kotecki,M., Morrell,M. and Rumshisky,A. (2001) Extraction and disambiguation of acronym–meaning pairs in medline. In *Proceedings of MedInfo '2001*, London, UK, September 2–5, 2001.

Rasolofo,Y. and Savoy,J. (2003) Term proximity scoring for keyword-based retrieval systems. In *25th European Conference on IR Research (ECIR)*, Pisa, Italy, April 14–16, LNCS 2633. Springer, pp. 101–1160.

Ruch,P. (2002) Information retrieval and spelling errors: improving effectiveness by lexical disambiguation. *ACM-SAC Information Access and Retrieval Track*.

Ruch,P. and Baud,R. (2002) Evaluating and reducing the effect of data corruption when applying bag of words approaches to medical records. *Int. J. Med. Inf.*, **67**, 75–83.

Ruch,P., Baud,R., bouillon,P. and Robert,G. (2000) Minimal commitment and full lexical disambiguation: balancing rules and hidden Markov models. In *Proceedings of the Fourth Conference on Computational Natural learning and of the Second Learning Language in Logic Workshop*, Lisbon, ACL, NY, pp. 111–114.

Ruch,P., Perret,L. and Savoy,J. (2005) Features combination for extracting gene functions from medline. In *27th European Conference on IR Research (ECIR, 2005)*, Santiago de Compostela, Spain, March 21–23, LNCS 3408, Springer.

Shah,P. *et al*. (2003) Information extraction from full text scientific articles: where are the keywords? *BMC Bioinformatics*, **4**.

Singhal,A. (2001) Modern information retrieval: a brief overview. *IEEE Data Eng. Bull.*, **24**, 35–43.

Singhal,A., Buckley,C. and Mitra,M. (1996) Pivoted document length normalization. In *Proceedings of the 1996 ACMSIGIR Conference on Research and Development in Information Retrieval (SIGIR'96)*, August 18–22, 1996. ACM, Zurich, Switzarland, pp. 21–29.

Stolz,W. (1965) A probabilistic procedure for grouping words into phrases. *Lang. Speech*, **8**.

Wilbur,J. and Yang,Y. (1996) An analysis of statistical term strength and its use in the indexing and retrieval of molecular biology texts. *Comput. Biol. Med.*, **26**, 209–222.

Yang,Y. (1996a) An evaluation of statistical approaches to medline indexing. *In Proceedings of AMIA 1996. Fall Symposium of the American Medical Informatics Association*, Washington, pp. 358–362.

Yang,Y. (1996b) Sampling strategies and learning efficiency in text categorization. In *Proceedings of the AAAI Spring Symposium on Machine Learning in Information Access*. AAAI press, pp. 88–95.

Yang,Y. (1999) An evaluation of statistical approaches to text categorization. *J. Inf. Ret.*, **1**, 67–88.

Yang,Y. and Chute,C. (1992) A linear least squares fit mapping method for information retrieval from natural language texts. In *Proceedings of the 14th International Conference on Computational Linguistics (COLING 92)*, August 23–28, 1992Nantes, France, pp. 447–453.

Zobel,J. (1998) How reliable are large-scale information retrieval experiments? In *Proceedings of the the 21st Annual Internation ACM SIGIR Conference on Research and Development in Information Retrieval*. August 24–28, 1998, ACM, Melbourne, Australia, pp. 307–314.