BIOINFORMATICS

Vol. 19 Suppl. 1 2003, pages i91–i94
DOI: 10.1093/bioinformatics/btg1011

Combining NLP and probabilistic categorisation for document and term selection for Swiss-Prot medical annotation

Pavel B. Dobrokhotov[1,], Cyril Goutte[2,*], Anne-Lise Veuthey[1] and Eric Gaussier[2]*

[1]Swiss Institute of Bioinformatics, CMU, 1 Michel-Servet - CH-1211 Genève 4, Switzerland and [2]Xerox Research Centre Europe, 6 ch. de Maupertuis - F-38240 Meylan, France

Received on January 6, 2003; accepted on February 20, 2003

ABSTRACT

Motivation: Searching relevant publications for manual database annotation is a tedious task. In this paper, we apply a combination of Natural Language Processing (NLP) and probabilistic classification to re-rank documents returned by PubMed according to their relevance to Swiss-Prot annotation, and to identify significant terms in the documents.

Results: With a Probabilistic Latent Categoriser (PLC) we obtained 69% recall and 59% precision for relevant documents in a representative query. As the PLC technique provides the relative contribution of each term to the final document score, we used the Kullback-Leibler symmetric divergence to determine the most discriminating words for Swiss-Prot medical annotation. This information should allow curators to understand classification results better. It also has great value for fine-tuning the linguistic preprocessing of documents, which in turn can improve the overall classifier performance.

Availability: The medical annotation dataset is available from the authors upon request

Contact: Pavel.Dobrokhotov@isb-sib.ch; Cyril.Goutte@xrce.xerox.com

INTRODUCTION

A major challenge faced by curators of biological knowledge bases is to search manually for relevant information through a vast number of publications. Probabilistic classifiers, such as Naïve Bayes, have been shown to tackle such problems successfully (Marcotte *et al.*, 2001; Wilbur, 2000). We applied a similar approach to the Swiss-Prot (Boeckmann *et al.*, 2003) medical annotation that deals with all genetic variants of a human protein, with the exception of nonsense and frameshift amino-acid changes. It presents some distinct advantages as a target for bio-text mining: the search space is small and clearly defined—only human proteins—, sufficient background knowledge is available—such as official gene names and synonyms—, and documents belong to only one class at a time. However, the main difficulty is that the collection of information should be very comprehensive. Hence, we consider all references returned by PubMed for a given protein and re-rank them in a way that is more relevant to curators, by pushing important documents to the top of the list. As soon as curators feel they have sufficient information, they can stop processing the list. We performed such re-ranking using a probabilistic classifier that assigned documents to one of three categories, relevant, irrelevant and unsure documents, and ordered them according to their significance. We also identified the most informative and discriminating words associated with each category.

SYSTEMS AND METHODS

Dataset

32 human genes were chosen from a list scheduled for medical annotation. The corresponding 2188 abstracts were retrieved from PubMed using queries with the gene name and keywords: mutation, variant and polymorphism. All abstracts were then manually classified by Swiss-Prot curators: 15% were assigned to class 'Good' (relevant for medical annotation), 70% to 'Bad' (irrelevant) and 15% to 'Unclear' (not enough information to judge relevance). The number of retrieved documents per gene ranged from 2 to 258, and the proportion of 'Good' ranged from 1% to 82%. This variability accurately reflects the diversity of the data encountered by curators.

Document processing

All textual parts of the documents (title and abstract) were analysed using Xerox NLP tools[†]. Texts were

*To whom correspondences should be addressed.

[†] http://www.xrce.xerox.com/competencies/content-analysis

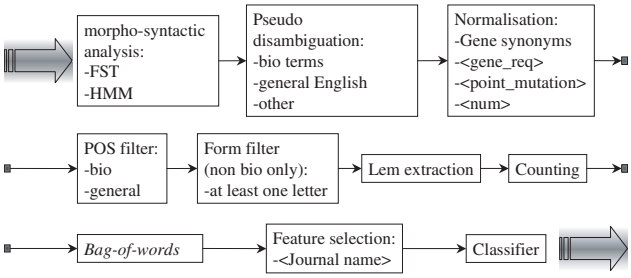

Fig. 1. Document processing flow.

segmented, lemmatised and tagged. Furthermore, remaining ambiguities were resolved in the following way. If a word has a general and a biological meaning, the biological reading was enforced; for other types of ambiguity (e.g. between English words, abbreviations), the first returned meaning was taken. Gene synonyms or protein names were reduced to a canonical form, using the biological resources described in Hagège *et al.* (2002). Gene names used for queries were replaced by a generic token 'gene_req'. Mutations indicated by a pattern *AaaNBbb* (where *Aaa* and *Bbb* are the 1- or 3-letter IUPAC-IUB amino-acid codes and *N* is an integer) were replaced by a token 'point_mutation', and all numbers which are not part of a word were replaced by a token 'num'. Finally, grammatical words (prepositions, determiners, etc.) were filtered out, as were English words shorter than 3 characters or without any letter. The resulting *bag-of-words* contains the textual features associated with each document, together with their frequency. As an extra feature, the journal name was added to serve as an indication of the information source. Figure 1 shows a schematic representation of this document processing.

Probabilistic classifier

For classification, we use the Probabilistic Latent Categoriser (PLC) described by Gaussier *et al.* (2002). It models a collection of pre-processed documents using a generative mixture model of co-occurrences of terms t and documents d:

$$P(t,d) = \sum_{\alpha} P(\alpha)\, P(d|\alpha)\, P(t|\alpha)$$

The class variable α runs over the class labels, e.g. from 1 to N for N-class classification. As each class α has its own class-conditional term distribution $P(t|\alpha)$, it is possible to find terms that have different importance for different classes by comparing these distributions. One way to measure the difference between two distributions $P(t|\alpha_0)$ and $P(t|\alpha_1)$ is the (symmetrised) Kullback-Leibler (KL) divergence:

$$D(\alpha_0,\alpha_1) = \sum_t \underbrace{(P(t|\alpha_0) - P(t|\alpha_1)) \log\left(\frac{P(t|\alpha_0)}{P(t|\alpha_1)}\right)}_{\varepsilon_t}$$

with $D(\alpha_0,\alpha_1) = 0$ iff $P(t|\alpha_0)$ and $P(t|\alpha_1)$ are identical.

Hence, the importance of a term for one class with respect to another is estimated by its contribution ε_t to the KL divergence. The higher the ε_t, the more important t is to differentiate between classes α_0 and α_1.

We evaluated a three-class and various types of two-class classifiers, depending on the handling of 'Unclear' documents. We also tested a cascade of the 'Good or Unclear' versus 'Bad' and the 'Good' versus 'Bad or Unclear' two-class classifiers. Assuming $P_{GU} = P(\text{'Good or Unclear'}|\text{d})$ and $P_B = P(\text{'Bad'}|\text{d})$ with the first classifier; $P_G = P(\text{'Good'}|\text{d})$ and $P_{UB} = P(\text{'Unclear or Bad'}|\text{d})$ with the second classifier, the assignment rule becomes:

> If $P_{GU} < P_B$, assign to 'Bad' (with score P_B);
> else, if $P_G > P_{UB}$, assign to 'Good' (with score P_G);
> else, assign to 'Unclear' (with score P_G).

The first two rules ensure high precision in the 'Bad' and 'Good' zone, while the remaining default assignment aim at enforcing high recall in the 'Unclear' zone.

Table 1. Performance of different classifiers

	Two-level classifier		Three-class classifier	
Class:	**p**	**r**	**p**	**r**
Good	**58.89**	69.28	54.07	73.86
Good or Unclear	48.95	**83.99**	47.94	83.66
Bad	**96.26**	82.46	96.15	81.81

Performance evaluation

In order to provide an unbiased evaluation, the collection was first split into 5 roughly identical blocks. Models were estimated on four blocks and evaluated on the left-out block, and performance was averaged over 5 splits, in a 'cross-validation' fashion. To assess performance of the classifier, we used traditional Information Retrieval measures: precision (p) and recall (r). For re-ranking, documents were ordered in each class according to probabilities given by PLC, so that the most relevant (or least irrelevant) appear at the top and the resulting lists were concatenated in the following sequence: 'Good', 'Unclear' and 'Bad'. Re-ranking techniques are typically evaluated using precision-recall curves, giving the precision at various levels of recall. The higher the curve, the better.

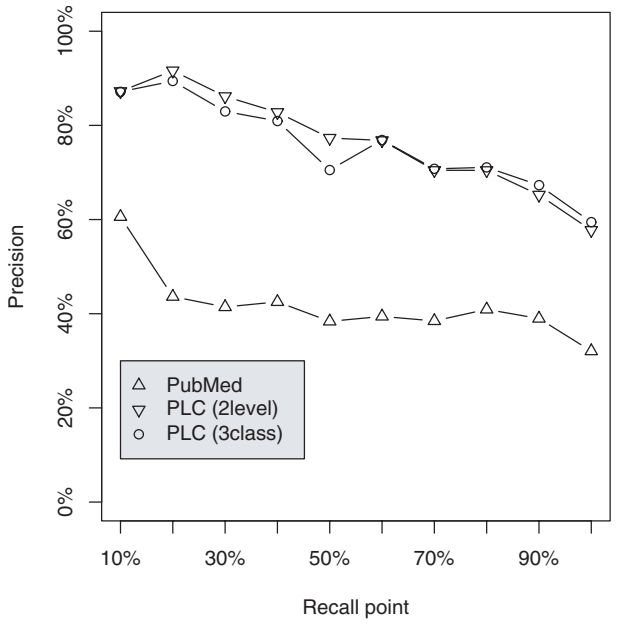

Fig. 2. Re-ranking performance of different classifiers.

RESULTS AND DISCUSSION

Table 1 summarises results achieved by a two-level and a three-class classifier. The performance in the 'Good' and 'Good or Unclear' classes are for relevant documents, while the performance for the 'Bad' class is for the irrelevant references. Both classifiers performed very closely, except for the 'Good' class, where a two-level classifier showed a higher precision, at the cost of some recall points. As the former is more important for this class, the two-level classifier was considered as being the best in this comparison. It actually achieved the best overall performance among all tested classifiers and document pre-treatments (for details see Dobrokhotov *et al.*, 2003).

We also compared the re-ranked list returned by the probabilistic classifier to the default order returned by PubMed. Figure 2 shows that the PLC provides an improvement from 25% to 45% depending on the recall point.

Using the Kullback-Leibler divergence, we estimated the most discriminating words for each class. Table 2 shows the top 10 items for three different comparisons: 'Good or Unclear' versus 'Bad' (recall favoured), or 'Good' versus 'Bad or Unclear' and 'Good' versus 'Bad' (precision favoured). As expected, we find words such as 'mutation' and 'missense_mutation' or 'patient' and 'carrier' that are often encountered in sentences describing mutations. In addition, the normalised 'point_mutation' token is first or second in these lists, proving that this generalisation step was helpful.

Table 2. Most discriminating words with different classifier strategies

'Good or Unclear' versus 'Bad'	'Good' versus 'Bad or Unclear'	'Good' versus 'Bad' (3 class classifier)
mutation	point_mutation	mutation
point_mutation	mutation	point_mutation
patient	missense_mutation	missense_mutation
family	exon	patient
missense_mutation	num	family
num	family	exon
gene_req	patient	num
exon	carrier	carrier
disease	FH	substitution
carrier	substitution	porphyria

By favouring precision with the second classifier, words have a different ranking, however eight of them remain the same (middle column in Table 2). Interestingly, the token 'gene_req' is downgraded to the 25th position, suggesting that while the frequency at which the query gene appears in the title/abstract is generally important to filter out irrelevant documents, it may be too broad to ascertain the relevance of a paper. A more in-depth study will be necessary in order to validate this hypothesis. The three-class classifier (right column in Table 2) also ranks the same 8 words at the top ('gene_req' is on the 19th position) and shares a new word 'substitution' with the 'Good' versus 'Bad or Unclear' classifier. Another observation is that the general word 'disease' (#9 in left column) is shifted down to the 34th and 13th place in the two other rankings, and its place is taken by FH (familial hypercholesterolemia) and porphyria—specific disease names associated with some highly represented genes in our dataset. This exemplifies the difficulties encountered when working with relatively small corpora.

Finally, we asked Swiss-Prot curators to validate lists from these classifiers, and they confirmed that these words, especially the top-ranking ones we show, were indeed used in their manual selection process.

CONCLUSION AND FURTHER DEVELOPMENT

The results we obtained suggest that natural language processing techniques, combined with Probabilistic Latent Categoriser, can be successfully applied to document ranking problems in the biomedical field. After re-ranking, all the good documents are found in the upper 40% of the list and bad ones are identified with a 96% accuracy. PLC also allows one to identify the most informative words of the text and their impact on document classification automatically. This is not only useful for tuning the classifier and identifying possible sources of errors, but will also be helpful for curators. These words can be highlighted in the text, thus helping to evaluate the

relevance of each document and speeding the review of the re-ranked list.

Further research on the classifier chain will continue, particularly on the disambiguation and normalisation parts and the assessment of their impact upon final classification. This includes the usage of higher quality biological dictionaries and term recognition, the use of more complex mutation patterns, etc. Furthermore, a better weighting scheme for the different components of the final *bag-of-words* (title versus abstract versus journal name) will be devised.

ACKNOWLEDGEMENTS

We are grateful to Livia Famiglietti, Raffaella Gatto and Arnaud Gos of the Swiss-Prot medical annotation team for providing the dataset of abstracts.

This work was supported in part by a grant from the Sibelius project (INRIA/SIB).

REFERENCES

Boeckmann,B. *et al.* (2003) The SWISS-PROT protein knowledgebase and its supplement TrEMBL in 2003. *Nucleic Acids Res.*, **31**, 365–370.

Dobrokhotov,P.B. *et al.* (2003) A probabilistic information retrieval approach to medical annotation in Swiss-Prot. *Proc. MIE2003*, (to appear).

Gaussier,E., Goutte,C., Popat,K. and Chen,F. (2002) A hierarchical model for clustering and categorising documents. In Crestani,F., Girolami,M. and van Rijsbergen,C.J. (eds), *Advances in Information Retrieval*. Springer, Berlin, pp. 229–247.

Hagège,C., Sandor,A. and Schiller,A. (2002) Linguistic processing of biomedical texts. In Ranchod,E. and Mamede,N.J. (eds), *Advances in Natural Language Processing*. Springer, Berlin, pp. 197–208.

Marcotte,E.M. *et al.* (2001) Mining literature for protein-protein interactions. *Bioinformatics*, **17**, 359–363.

Wilbur,J.W. (2000) Boosting Naïve Bayesian Learning on a Large Subset of MEDLINE. *Proc. AMIA Symp*, 918–922.