Classification from Full Text: A Comparison of Canonical Sections of Scientific Papers

Gail Sinclair **Bonnie Webber**
School of Informatics,
University of Edinburgh
{csincla1, bonnie}@inf.ed.ac.uk

Abstract

The accelerating growth in biomedical literature has stimulated activity on automated classification of and information extraction from this literature. The work described here attempts to improve on an earlier classification study associating biological articles to GO codes. It demonstrates the need, under particular assumptions, for more access to full text articles and for the use of Part-of-Speech tagging.

1 Introduction

The accelerating growth in biomedical literature is stimulating efforts both to screen individual papers quickly for useful information and to use aggregations of papers for the collective information they provide. Aggregative use may involve what one might call "binning" classification, where one decides which of N bins an entity should be slotted into (Raychaudhuri et al., 2002). Most often these tasks have been done on titles and abstracts, simply because that is what is most freely available. However the nature of titles and abstracts means that they may lack information that is relevant to the task.

The present study considers this issue, taking as its starting point work done by Raychaudhuri, Chang, Sutphin and Altman (2002). In this work: (i) articles were associated with GO codes; and then (ii) GO codes were assigned to new genes on the basis of the GO-code associations with articles about related genes. This paper reconsiders the basis for Step (i), to see if this can be done more accurately, using full text rather than titles and abstracts.

Raychaudhuri et al. (2002) investigated how statistical natural language techniques could be applied to assign GO codes to genes using the titles and abstracts of articles about related genes. GO codes are terms drawn from three controlled vocabularies (biological processes, cellular components and molecular functions) developed by the Gene Ontology Consortium (Ashburner et al., 2000). The Gene Ontology Consortium's aim is for gene products to be described in a consistent manner across independent databases and species. Each controlled vocabulary is organised as a directed acyclic graph (DAG).

The GO codes that Raychaudhuri et al. chose to assign to articles (and hence to genes) were from the biological process vocabulary, an approximate horizontal cut through the biological process DAG. However, some departures were made from the horizontal when the authors found difficulties in precisely defining the associated literature. In departing from the horizontal, parent/child dependencies were introduced. For example, GO code *transport* is a parent of *intracellular protein traffic*.

MEDLINE queries on these GO codes were manually created in order to retrieve approximately 1000 articles related to each topic. The queries contained both MeSH terms and keywords. Medical Subject Headings (MeSH) (Hutchinson, 1998) is a controlled vocabulary from the National Library of Medicine used to aid indexing and searching biological information. MEDLINE articles are indexed by MeSH headings, among other annotations. The PubMed search tool allows a user to specify desired search fields, of which Raychaudhuri et al. used title (TI), Major MeSH Heading (MAJR), MeSH Heading (MH) and date of publication (DP). The first three fields were used to specify the subject of the article while the DP field was used to limit the number of articles retrieved to approximately 1000.

Raychaudhuri et al. experimented with three machine learning approaches (Naive Bayes, K-Nearest Neighbours and Maximum Entropy) to classify articles according to the 21 GO codes. Each classifier was trained on articles retrieved using the described queries from 1999 and earlier, and tested on articles from 2000. Maximum Entropy was found to be the most successful at classifying articles, achieving 72.83% accuracy.

2 Methods

The current study was concerned with two issues - which sections of full text journal articles are most informative with regards to gene product and which Natural Language Processing techniques are most

useful in associating those products with particular articles. The scores from the Raychaudhuri et al. study are used as a baseline (see Table 1).

2.1 Data

2.1.1 Full Text Access

PubMed gives access to information in MEDLINE - the title and abstract of articles along with manual annotations such as MeSH Headings and Registry Numbers. PubMed Central[1], on the other hand, gives access to the full text (in HTML) of (currently 98) journals that are indexed in MEDLINE. Also, many other publishers are now making their journal articles available online for free on their own sites. PubMed Central will also list articles from these publishers.

BioMed Central (BMC) is another resource for full text articles. BMC, like PubMed Central, contains full text from many journals as well as having many of its own online journals. Authors can submit articles to these BMC journals and have them reviewed and published in the same month[2].

2.1.2 Full Text Retrieval

The same queries that were used in the Raychaudhuri et al. study were used to query PubMed Central in order to find full text articles relating to the 21 biological process GO codes. The DP field was omitted, in order to access as many full text articles as possible. For some of the 21 GO codes, there were not enough free full text articles available to be deemed representative and so only those codes that had 50+ full text articles associated with them were used in the rest of the study. These can be seen in Table 1.

2.1.3 Article Sections

Most journals have a format to which authors must adhere in order for an article to be considered for publication, including rules concerning the naming of sections.

With respect to the structure of scientific papers (or, more specifically, papers in biology), many people talk about them having a canonical structure consisting of a *Title, Abstract, Introduction, Materials and Methods, Results,* and *Discussion* in either this order or with *Materials and Methods* at the end.

For the experiments reported here, those articles were extracted from the full text of journals that adhere closely to this canonical structure and other sections were ignored. Sections named simply *Methods* were included with the *Materials and Methods* sections.

[1]http://www.pubmedcentral.gov/about/intro.html

[2]http://www.biomedcentral.com/info/about/whatis

2.2 Tools

2.2.1 Classification

Because the current study concerns whether NLP techniques can help to improve performance of classification, we have postponed experimenting with different machine learning techniques. We will do so after we find which NLP techniques are the most useful. The Rainbow[3] Naive Bayes classification tool was used.

Raychaudhuri et al. induced a single N-ary classifier, whereas this study induced 21 binary classifiers, i.e. an article was classified as either being related to a particular biological process or unrelated.

2.2.2 NLP techniques

We applied both Part-of-Speech tagging and stemming. The LT-TTT tagger (Grover et al., 2000) was used to tag the part of speech each word belonged to. This allowed us to experiment with building classifiers based only on single parts of speech as well as ones based on all words.

The most widely used stemmer among the NLP community is the Porter stemmer (Porter, 1980). A Perl version of this was used to produce stemmed sets of the articles.

We experimented with four strategies to find the best performance in classification: bag of words; bag of nouns; bag of stems; bag of stemmed nouns.

2.2.3 Training

There were too few full text articles to both train and test on, so the classifiers were trained on the original titles and abstract articles from Raychaudhuri et al. and then tested on the full text and sections thereof. The negative training instances for each category were those articles that were related to the other categories (approx 2000). Four sets of classifiers were trained: one set each for the bags of words, nouns, stems and stemmed nouns.

3 Results

3.1 GO terms

The GO terms we used are shown in Table 1, along with the baseline scores achieved in the earlier study using Maximum Entropy and the corresponding scores using Naive Bayes. It should be noted that the exact same test data were not used in this comparison, although the data were retrieved in a similar fashion (via the same MEDLINE queries). The earlier data was limited to post-1999 articles, whereas the present study used the *Titles* and *Abstracts* from any related articles that had free full text available.

[3]http://www.cs.cmu.edu/ mccallum/bow

GO Terms	No. Articles	Baseline Maxent	Naive Bayes
Cell Cycle	106	45.9	68.6
Cell Death	75	75.8	60.0
Cell Motility	62	71.4	67.2
Chemimechanical Coupling	57	79.6	51.8
Intracellular Protein Traffi c	154	68.6	77.6
Meiosis	50	77.5	91.8
Metabolism	72	67.6	58.6
Signal Transduction	84	59.9	62.2
Stress Response	57	64.8	74.6

Table 1: Comparison of individual Recall scores for previous and present studies using bag of words.

Section(s)	Words	Nouns	Stemmed Words	Stemmed Nouns
Title and Abstract	68.7 / **60.1** / **64.3**	**84.5** / 46.1 / 59.1	70.6 / 58.0 / 63.3	81.2 / 49.2 / 60.5
Full Text	70.4 / **54.9** / **60.8**	87.5 / 37.1 / 52.0	70.8 / 54.2 / 60.1	**89.8** / 27.8 / 41.7
Title	66.0 / **65.4** / **64.7**	**77.2** / 55.9 / 63.9	66.6 / 63.4 / 64.0	75.2 / 56.7 / 63.9
Abstract	68.0 / **60.2** / **62.8**	**85.9** / 45.3 / 57.9	69.3 / 58.2 / 62.2	78.3 / 48.7 / 59.2
Introduction	68.4 / **56.5** / **61.0**	**83.3** / 42.7 / 55.6	69.6 / 54.9 / 60.3	77.2 / 45.9 / 56.8
Methods	68.4 / **60.6** / **63.3**	**82.2** / 45.4 / 56.6	69.4 / 58.5 / 62.5	78.5 / 48.9 / 59.5
Results	62.5 / **56.5** / **58.0**	**81.9** / 38.1 / 51.4	61.4 / 55.3 / 56.9	78.8 / 42.4 / 54.2
Discussion	69.6 / **59.0** / **62.5**	**87.5** / 42.7 / 56.9	69.7 / 57.2 / 62.0	83.5 / 46.1 / 58.4

Table 2: Average Recall / Precision / F-score percentages of classification of full text and individual sections using the four NLP strategies.

3.2 Section Evaluation

Classification results are shown in Table 2. This table shows the recall, precision and F-score for each section of text and for each of the four word-bag types. The first line of the table corresponds to Raychaudhuri et al.'s strategy using Naive Bayes instead of Maximum Entropy. The F-score is calculated giving equal weighting to recall and precision.

Titles achieved the best F-score - this occurs because the precision was much higher than the other sections. This is not unexpected since there would be very little room for false indicators in the relatively short *Title* section. The other sections have more scope for introducing negative indicators. *Titles* consistently had lower recall in comparison with the other sections. Obviously a title can only convey the one or two main points of an article and not include every relevant topic.

The *Methods* section was expected to fare worse than other sections, since it contains more technical data, such as investigative techniques, chemicals and measurements, than information about biological processes. However, performance on the *Methods* section was on a par with the *Abstract* and *Introduction*, suggesting that the *Methods* sections may give the reasoning behind certain experiments. This is in contrast to Shah et al.'s (2003) conclusion that the *Methods* section was not valuable for the extraction of keywords relating to biological concepts compared with the other sections.

The *Introduction* section can conceivably contain any type of information, including similar/opposite studies, ultimate goal of the present study, other processes related to the gene(s)/protein(s) in question and so can have many positive and negative indicators of category. Thus, a similar performance to Abstracts and Methods is not to be unexpected.

The *Results* section generally produced the worst performance. This could be considered surprising since here is where one would expect the proof of biological processes occurring in experimentation. This outcome may be because no reasoning is made about the results at this point. Also there are frequent indicators *against* a category, when a biological process is found not to be affected in the experiment and is so stated, e.g. "**Biogenesis** of the vacuole is **not** obviously disturbed in aut9 cells" (Lang et al., 2000). Explicit negative information is im-

portant for biologists, so that they do not to waste resources by repeating work that has already been investigated. On the other hand, it has an adverse affect on classification.

The whole full text achieved both the best recall and the worst precision. The full text has maximum potential for including positive indicators of biological process just as it has maximum potential for including misleading indicators.

All individual sections except *Titles* underperformed in comparison with the baseline of *Title and Abstract* with regard to equally-weighted F-score. (See Section 4 for discussion of alternatives to equally-weighted F-score.) Similarly, nouns, stemmed words and stemmed nouns all produced a lower equally-weighted F-score than did the baseline of bag of words. *Discussion* was the only section dataset to outperform the *Title and Abstract* with regards to recall, while no section significantly bettered *Title and Abstract* on precision except *Titles* alone.

3.3 Evaluation of stemming and POS-tagging

While both training and testing on nouns and stems increased the performance compared to simply using a bag of words, combining these two techniques seemed to interfere with their individual usefulness. The combination - first retrieving the nouns and then stemming them - achieved an increase in recall compared with just stemming, however recall was decreased compared with just using nouns.

The trend between classifying with words and nouns differed depending on whether they are stemmed or not. Recall generally increased and precision generally decreased when going from classifying with whole words to classifying with stemmed words. In contrast, recall *decreased* and precision *increased* when going from classifying with nouns to classifying with stemmed nouns.

4 Discussion

The increase in performance using the *Discussion* sections as compared with the *Title and Abstract* does not perhaps seem significant enough to warrant the effort involved in retrieving and processing the HTML of the full text. However, this study was based on the classifiers being trained on the titles and abstracts, and so further studies are currently ongoing with full text and sections thereof being used to both train and test the classifiers.

The nature of the data is such that, for any class, the number of negative instances far exceeds the number of positive instances. Thus, the low precision scores were influenced by the amount of negative instances in the test data. For example, if a category had 100 positive instances, it also had approx 850 negative test instances. If 1 out of every 10 negative instances were incorrectly classified as positive, and 1 out of every 5 positive instances were incorrectly classified as negative, recall would be 80% and precision would be 48.5%. However, if the test data included 400 negative instances with the same error rates, precision would increase to 66.7%.

If the role of this classification task is as a first-pass filter, then recall is more important than precision, as we want to minimise the loss to false negatives. As such, it would perhaps be more indicative of the performance of the classifiers to actually calculate an F-score that gives more weight to recall. For example, triple-weighting recall promotes nouns as classifiers, with *Title and Abstract* F-scores becoming 66.3% (words), 69.9% (nouns), 66.9% (stemmed words) and 69.8% (stemmed nouns). A triple-weighted recall F-score also promotes the *Discussion* section as the basis for classification, with F-scores of 66.6%, 69.9%, 66.1%, 69.4% respectively. These scores may be more representative of the relative benefit of stemming and POS-tagging.

5 Acknowledgements

This work has been supported by Scottish Enterprise, through the Stanford-Edinburgh LINK project.

References

M. Ashburner, Ball C.A., Blake J.A., Botstein D., Butler H., Cherry J.M., Davis A.P., Dolinski K., Dwight S.S., and Eppig J.T. 2000. The Gene Ontology Consortium. *Nat. Genet.*, 25:25–29.

Claire Grover, Colin Matheson, Andrei Mikheev, and Marc Moens. 2000. LT TTT - a flexible tokenisation tool. In *Proc. of LREC 2000*.

D. Hutchinson. 1998. *MEDLINE for health professionals: How to search PubMed on the Internet*. New Wind, Sacramento, CA.

T. Lang, S. Reiche, M. Straub, M. Bredschneider, and M. Thumm. 2000. Autophagy and the Cvt pathway both depend on AUT9. *J. Bacteriol*, 182:2125–2133.

M. F. Porter. 1980. An algorithm for suffix stripping. *Program*, 14(3):130–137.

S. Raychaudhuri, J.T. Chang, P.F. Sutphin, and R. Altman. 2002. Associating genes with GO Codes using a maxent analysis of biomedical literature. *Genome Research*, 1:203–214.