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.

Screening involves what one might call “go/no-go” classification, where one decides whether a paper is on a topic or not, or is worth curating or not, or mentions a particular gene or not, or is worth reading or not. One example of this was the KDD challenge in 2002 (Yeh et al., 2002), where participants had to develop a system that could decide whether an article contained material that should be curated for FlyBase (Flybase, 1999). Another was the first TREC Genomics challenge (Hersh and Bhupatiraju, 2003; Osborne et al., 2003), where participants had to decide whether an article described a particular gene and its function.

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). (Since any N-ary decision can be factored into N-1 binary decisions, “binning” classification can be implemented as a set or sequence of “go/no-go” classifications.)

Both these classification tasks can be used in the automatic curation of articles into existing databases (Yeh et al., 2002), before any extraction of specific information takes place. This automatic curation enables the extraction method (manual, automatic or semi-automatic) to deal with many fewer articles than it otherwise would have, since most irrelevant papers will have been discarded.

Most classification of biomedical articles has 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.

With MedLine increasing by 2,000 articles daily, most biologists do not have time to read every new article published in their field. Very often, the abstract and title will be scanned for whether it is worth reading the full text or not, i.e. whether there is any curatable information within the article. As stated above, the abstract will not always tell a biologist what he or she needs to know. Knowing which section or sections of the full text will most likely tell you if the article is curatable would be an advantage.

Of the five usual sections of a journal article (Abstract, Introduction, Materials and Methods, Results, Discussion), it seems likely that any section but Materials and Methods would contain the decision criteria being searched for. Shah et al. (2003) indeed found this to be the case, with the Materials and Methods sections mostly describing the tools of the experimentation rather than the biological concepts. Theoretically then, one could read up to four fifths of the full text before deciding that the article is, in fact, not curatable.

The present study deals with a different classification problem, based on previous 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.

1http://www.nlm.nih.gov/pubs/factsheets/medline.html
1.1 Raychaudhuri et al (2002)

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. The following 21 GO codes were finally chosen:

- autophagy
- cell adhesion
- cell death
- cell motility
- cell-cell signaling
- intracellular protein traffic
- invasive growth
- meiosis
- metabolism
- signal transduction
- stress response
- biogenesis
- cell cycle
- cell fusion
- cell proliferation
- chemimechanical coupling
- ion homeostasis
- membrane fusion
- oncogenesis
- sporulation
- transport

In departing from the horizontal, parent/child dependencies were introduced: cell cycle is a child of cell proliferation, and 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. As an example, the query for GO code GO:0006914 (Autophagy) was “(autophagy [TI] OR autophagocytosis [MAJR]) AND (Proteins [MH] OR Genes [MH]) AND 1940:1999[DP]”.

From the set of PubMed entries retrieved, Raychaudhuri et al tokenised the title and abstract, using white space and common punctuation as delimiters. After removing other punctuation and both very infrequent and frequent words, the resulting tokens were then used to represent each document as a vector of unique tokens.

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. (The best performance achieved by Naive Bayes and K-Nearest Neighbours was 66.92% and 66.9% accuracy respectively.) Accuracy refers to the percentage of articles retrieved using GO term γ that were classified as γ. If each retrieval set is taken to be the full set of articles relevant to γ, then accuracy is equivalent to recall.

2 Methods

The current study was principally 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. In February 2000, the National Library of Medicine launched PubMed Central\(^2\). This resource gives access to the full text (in HTML) of many journals that are indexed in MedLine - currently 98 journals participating in this full text access. 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

\(^2\)http://www.pubmedcentral.gov/about/intro.html
many of its own online journals. Authors can submit articles to these BMC journals and have them reviewed and published in the same month³.

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. In practice, one has to deal with the fact that many journals diverge from this canonical structure, each in their own particular way – sections may be merged or omitted, they may be presented in a different order, they may have different names, and/or additional sections may be present.

The Journal of Cell Biology⁴ and Genes and Development⁵, for example, require the sections Abstract, Introduction, Results, Discussion and Materials and Methods, while Development⁶ requires a Summary section instead of an Abstract, the Journal of Physiology⁷ renames the Materials and Methods section as just Methods, Genetics⁸ requires no Introduction (although there does seem to be an Introduction-type section in the article - it is just unnamed) and, further, Molecular and Cellular Biology⁹ and the Journal of Bacteriology¹⁰ sometimes collapse the Results and Discussion sections into one. To add to the inconsistency, some journals do not require any particular set of sections.

²http://www.biomedcentral.com/info/about/whatis
³http://www.jcb.org/
⁴http://www.genesdev.org/
⁵http://dev.biologists.org/
⁶http://jp.physoc.org/
⁷http://www.genetics.org/
⁸http://mcb.asm.org/
⁹http://jb.asm.org/
¹⁰The result is that one has to write journal-specific parsers (much like “wrappers” for different varieties of web pages). By far the most common sections are the ones initially mentioned - Abstract, Introduction, Materials and Methods, Results and Discussion along with Title. For the experiments reported here, these ones 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.

2.2 Tools
2.2.1 Classification
The current study is primarily concerned with whether NLP techniques can help to improve performance of classification and so we have postponed experimenting with different machine learning techniques. We will do so after we find which NLP techniques are the most useful.

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.

The Rainbow suite (McCallum, 1996) of document classification tools was used. This suite includes Maximum Entropy, Naive Bayes, K-Nearest Neighbours, TF-IDF options among others. We made a practical decision to use Naive Bayes since it took a fraction of the time to run as Maximum Entropy.

2.2.2 NLP techniques
Raychaudhuri et al used a bag of words approach to classification, while also removing words that occurred less than 4 and more than 10,000 times across all 20,000 retrieved articles. This option - restriction by word occurrence - was not available in Rainbow’s Naive Bayes method but is not thought to have an adverse affect on classification since the Naive Bayes’ algorithm will determine that these types of words have very low information gain and therefore have very little bearing on classification.

To these words, 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.

It can be useful to count morphologically and inflectionally related words as the same concept. This can be done with stemming, where each word is reduced to a base form or stem. So, for example, transcribed and transcribing can both be reduced to transcrib, and replication and replicate to repli-
whereas the present study used the abstracts. The earlier data was limited to post-1999 articles on word, nouns, stems and stemmed nouns.

3 Results

3.1 GO terms

As noted earlier, several of the 21 categories lacked sufficient full text articles for results on classification to be reliable, so only those nine categories that had 50 or more related articles were used in testing. These categories 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.

3.2 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.
Table 1: Comparison of individual Recall scores for previous and present studies using bag of words.

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

| 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 |

This effect can partly be explained by the sheer numbers of negative instances. With bag of words, stemming introduced an increase of approximately 2% true positives (TPs) and approximately 3-4% false positives (FPs). Since we are dealing with so many negative instances in classifying compared with positive (9:1), this 3-4% change in FPs outweighs the increase in TPs and so even though recall is increased, precision decreases.

With bag of nouns, stemming decreases TPs by approximately 10% and FPs by approximately 30%, so the decrease in FPs outweighs the decrease in TPs when it comes to calculating the precision and so precision increases, and a decrease in TPs can only decrease recall. Thus, a more balanced number of negative instances may stabilise these trends. Why stemming nouns has the opposite effect on recall than on stemming words has yet to be investigated.
Table 3: Equal-weighted and Triple-weighted Recall F-scores for Title and Abstract and Discussion sections.

| Section/Word Type        | 2PR/(P+R) | 4PR/(3P+R) |
|--------------------------|-----------|------------|
| Title and Abstract       |           |            |
| Bag of words             | 64.3      | 66.3       |
| Bag of stemmed words     | 63.3      | 66.9       |
| Bag of nouns             | 59.1      | 69.9       |
| Bag of stemmed nouns     | 60.5      | 69.8       |
| Discussion               |           |            |
| Bag of words             | 62.5      | 66.6       |
| Bag of stemmed words     | 62.0      | 66.1       |
| Bag of nouns             | 56.9      | 69.3       |
| Bag of stemmed nouns     | 58.4      | 69.4       |

3.3.1 Verbs
Classifiers were also trained using bags of verbs and bags of verbs and nouns. The bag of verbs classifiers had such dramatically increased recall (ca 99.7%), but significantly diminished precision (ca 11%) that it did not seem worth pursuing investigation into article sections. The classifiers so trained seemed to class almost anything as positive, suggesting that no matter what the biological process being discussed, the same verbs are used to describe them.

Training the classifiers on both verbs and nouns did not as significantly alter the scores compared with training on just nouns, with recall suffering (ca 80%) and precision staying about the same.

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 then had approx 850 negative test instances. If, then, 1 out of every 10 negative instances were incorrectly classified as positives, and 1 out of every 5 positive instances were incorrectly classified as negatives, 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 be 66.7%, with Recall staying the same.

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 in calculating the F-score results in the figures shown in Table 3. These scores are more representative of the relative benefit of stemming and POS-tagging.

The selection of GO codes is also an issue as some codes are parents of others and so the more specific articles could also be said to be about the more general. One could also expect the more general to be mentioned in the specific-related article and vice versa. This could introduce more false positives in classification than if the codes were hierarchically at the same level.

We noted earlier that Naive Bayes classification was used as opposed to Maximum Entropy which was used in the initial study. It would be hoped that Maximum Entropy would probably perform better than Naive Bayes as it did in Raychaudhuri et al, and that this would not alter the relative rankings of the sections. An initial investigation into Maximum entropy shows average Recall, Precision and F-scores for nouns with Title and Abstract to be 61.3%, 72.1% and 64.7% respectively and with Discussion to be 46.0%, 70.9% and 54.7%, indicating a decrease in recall, increase in precision and marginal variation in F-score over Naive Bayes. However, more work needs to be done on this to optimise Maximum Entropy’s features.

We also noted in Section 3.2 that evidence
against a particular category in the Results section could be influencing the performance of classification with this section. A brief investigation of the disposition of negative indicators across sections of negative indicators, (e.g. “not”, “none”, “absence”, “inactivity”) provides initial substantiation. Comparing the counts of these negative indicator words in each section, we found that 50% of the total occurrences occur in the Results sections.

Surprisingly, simply stemming the tokens only marginally increased recall on the whole. A lemmatiser was also experimented with at the beginning of this study. Using this sort of normalisation only increased recall by less than one percent over stemming. A stemmer or lemmatiser more adept at handling the morphology and inflexion of biological terms may prove more beneficial, however, regardless of how a word is normalised, it would seem that the normalisation only slightly increases performance as compared to using the properties of a word, such as part of speech tags. This may suggest that further use of shallow parsing of text would benefit classification even further.

The comparison of performance between this study and the previous one may perhaps seem unbalanced since the previous was a “binning” type classification and the current is a “go/no go” type classification. However biological articles can often be about more than one process and so it may not be prudent to rule out one category in favour of another. The current study can easily be transformed into a binning task since Naive Bayes gives probabilities of category and so the category with the highest probability across all classifiers could be taken as the chosen “bin”. Similarly, each section of an article could be taken as independent evidence for classification using a “committee” type voting scheme. This is currently being pursued.

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. Gene ontology: Tool for the unification of biology. the gene ontology consortium. Nat. Genet., 25:25–29.
Flybase. 1999. The flybase database of drosophila genome projects and community literature. Nucleic Acids Research, 27(1):85–88.

Claire Grover, Colin Matheson, Andrei Mikheev, and Marc Moens. 2000. LT TTT - a flexible tokenisation tool. In Proceedings of Second International Conference on Language Resources and Evaluation (LREC 2000).

W. Hersh and R.T. Bhupatiraju. 2003. Trec genomics track overview. In Proceedings of the Twelfth Text REtrieval Conference (TREC-12).

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.

Andrew McCallum. 1996. Bow: A toolkit for statistical language modeling, text retrieval, classification and clustering. http://www.cs.cmu.edu/~mccallum/bow.

M. Osborne, M. Cumiskey, G. Sinclair, M. Smillie, B. Webber, J. Chang, N. Mehra, V. Rotemberg, and R.B. Altman. 2003. Edinburgh-stanford trec-2003 genomics track. In Proceedings of the Twelfth Text REtrieval Conference (TREC-12).

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 gene ontology codes using a maximum entropy analysis of biomedical literature. Genome Research, 1:203–214.

P. K. Shah, C. Perez-Iratxeta, P. Bork, and M. A. Andrade. 2003. Information extraction from full text scientific articles: Where are the keywords? BMC Bioinformatics, 4(20).

Alexander S. Yeh, Lynette Hirschman, and Alexander A. Morgan. 2002. Background and overview for kdd cup 2002 task 1: Information extraction from biomedical articles. In Proceedings of the Eighth International Conference on Knowledge Discovery and Data Mining (KDD-2002), pages 87–89.