ABSTRACT

Motivation: Research in the biomedical domain can have a major impact through open sharing of the data produced. For this reason, it is important to be able to identify instances of data production and deposition for potential re-use. Herein, we report on the automatic identification of data deposition statements in research articles.

Results: We apply machine learning algorithms to sentences extracted from full-text articles in PubMed Central in order to automatically determine whether a given article contains a data deposition statement, and retrieve the specific statements. With an Support Vector Machine classifier using conditional random field determined deposition features, articles containing deposition statements are correctly identified with 81% F-measure. An error analysis shows that almost half of the articles classified as containing a deposition statement by our method but not by the gold standard do indeed contain a deposition statement. In addition, our system was used to process articles in PubMed Central, predicting that a total of 52,932 articles report data deposition, many of which are not currently included in the Secondary Source Identifier [si] field for MEDLINE citations.

Availability: All annotated datasets described in this study are freely available from the NLM/NCBI website at [http://www.ncbi.nlm.nih.gov/CBBResearchFellows/Neveol/DepositionDataSets.zip](http://www.ncbi.nlm.nih.gov/CBBResearchFellows/Neveol/DepositionDataSets.zip)

Contact: aurelie.neveol@nih.gov; john.wilbur@nih.gov; zhiyong.lu@nih.gov

Supplementary Information: Supplementary data are available at Bioinformatics online.

Received on July 22, 2011; revised on September 20, 2011; accepted on October 8, 2011

1 INTRODUCTION

Research in the biomedical domain aims at furthering our knowledge of biological processes and improving human health. Major contributions toward this goal can be achieved by sharing the results of research efforts with the community, including datasets produced in the course of the research work. While such sharing behavior is encouraged by funding agencies and scientific journals, recent work has shown that the ratio of data sharing is still modest compared with actual data production. For instance, Ochsner et al. (2008) found the deposition rate of microarray data to be <50% for work published in 2007. Piwowar and Chapman (2007) show that data deposition results in increased citation of papers reporting on data production. While this should serve as an incentive to deposit data and announce it to the community, in a more recent study these same authors (Piwowar and Chapman, 2010) show that data deposition is significantly associated with high-profile journals and experienced researchers. In the course of this work, these authors have found the identification of data deposition statements to be a challenging task that can be addressed using natural language processing and machine learning methods (Piwowar and Chapman, 2008a). Information about the declaration of data deposition in research papers can be used in different ways. First, for data curation: databases such as MEDLINE® use accession numbers for certain databases as metadata that can be searched with PubMed® queries using the [si] field. Journals can benefit from such a tool to check whether their data deposition policies are enforced. This aspect is also important for researchers looking to re-use datasets and build on existing work. Second, for the analysis of emerging research trends: the type of data produced gives indications on current important research topics. In a study based on the analysis of Medical Subject Headings® (MeSH®) indexing, Moerchen et al. (2008) show that such metadata may be used to predict future research trends, including recommendations of main headings to be added to the MeSH thesaurus. Our long-term research interest is in assessing the value of using deposition statements for predicting future trends of data production. The initial step of automatically identifying deposition statements could then lead to an assessment of the need for storage space of incoming data in public repositories. In this study, we aim at identifying articles containing statements reporting the deposition of biological data.

As explained above, the study of data deposition has generated a growing interest in the past few years. In response to a Nature Methods editorial (Anonymous, 2008) describing the deposition of data such as genome sequence or microarrays as ‘routine’, Ochsner et al. (2008) used a manually built query to retrieve articles likely to report the production of microarray data in 2007 publications. They manually analyzed 398 articles reporting the production of microarray data and concluded that only ∼50% report deposition of microarray data in the Gene Expression Omnibus (GEO) or other databases. Piwowar and Chapman (2008a) further studied the links existing between microarray data deposition in public repositories (e.g. GEO and ArrayExpress) and reports of data deposition in the literature. They used machine learning to build a query suitable for retrieving research articles in PubMed Central reporting on data deposition. Piwowar and Chapman (2008b) also addressed the classification of articles (at the article level) for data sharing in five databases (GenBank, Protein DataBank, GEO, ArrayExpress, Stanford Microarray Database). The authors compared pattern matching versus machine learning. The best results were obtained with a J48 decision tree on ArrayExpress (90% F-measure),...
This task was tailored to the specific need for curation of accession number challenge (Krallinger 2011). Precision when automatically extracted passages were compared to protein interaction. The best performing team obtained consecutive full-text sentences providing evidence for protein–protein interaction as a binary classification problem. Using SVM structured abstracts. Results for unstructured abstracts and other classes studied makes the problem amenable to the use of machine learning methods. For instance, several efforts aiming at the retrieval of text passages as evidence for biological or clinical phenomena performed sentence classification. Demner-Fushman et al. (2005) addressed the classification of MEDLINE abstract sentences between seven clinical outcome categories in order to automatically identify outcome-related information in the medical text. They reported the precision of the top ranked sentence between 50% and 60% depending on category. Kim et al. (2011) used the same dataset for classifying sentences for evidence-based medicine. Their best performance for SVM classification was 81% F-measure obtained for the outcome category of sentences in structured abstracts. Results for unstructured abstracts and other categories were less successful. Polajnar et al. (2011) addressed the identification of MEDLINE abstract sentences denoting protein–protein interaction as a binary classification problem. Using SVM classifiers and protein annotations as features, they reported best F-measure performance of ~70%. While these efforts were limited to abstracts, other work used full-text articles. In the BioCreative II challenge (Krallinger et al., 2008), the interaction sentences subtask required participants to retrieve passages of up to three consecutive full-text sentences providing evidence for protein–protein interaction. The best performing team obtained ~20% precision when automatically extracted passages were compared with evidence sentences manually selected by curators. These results reflect the difficulty of the task of extracting evidence statements from full-text articles.

2 MATERIAL AND METHODS

In this study, we are interested in identifying statements declaring the deposition of biological data (such as microarray data, protein structure, gene sequences) in public repositories. In the rest of this article, we will refer to such statements as ‘deposition statements’. We take these statements as a primary method of identifying articles reporting on research that produced the kind of data deposited in public repositories. (1) and (2) show examples of such statements, with varying degrees of specificity. In (1) both the data and location are referred to in a highly specific manner [i.e. ‘the sequence of labA’ and ‘DDBJ/GenBank/EMBL databases (accession no AB281186)’], whereas in (2) data and deposition location are both very general (‘the microarray data’ and ‘MIAMEexpress’). While the mention of data, public repositories and accession numbers are strong indicators of deposition statements, (3) and (4) show that these elements can also occur when authors refer to previous work. In the remainder of this article, we will refer to statements that do not report the deposition of data in public repositories—such as (3) and (4) as ‘non-deposition statements’.

(1) The sequence of labA has been deposited in the DDBJ/GenBank/EMBL databases (accession no AB281186) (PMID 17210798).
(2) The microarray data were submitted to MIAMEexpress at the EMBL-EBI (PMID 18535205).
(3) Histone TAG Arrays are a repurposing of a microarray design originally created to represent the TAG sequences in the Yeast Knockout collection (Yuan et al., 2005; NCBI GEO Accession Number GPL1444) (PMID 18085098).
(4) Therefore, the primary sequence of native Acinetobacter CMO is identical to the gene sequence for chlb deposited under accession number ABO06902 (PMID 11352635).

Figure 1 gives an overview of the annotated datasets used in the training and test phases of this work. The various datasets shown on Figure 1 are provided as Supplementary Material and are also freely available to the research community from the NLM/NCBI website. The following sections describe details of the datasets and experiments. In Section 2.1, we describe the method used to collect the training datasets, and the analysis of deposition sentences that we carried out in order to gain an understanding of the variety and common characteristics of these statements. In Sections 2.2 and 2.3, we explain how the training datasets were used to automatically identify deposition elements and perform sentence classification. Finally, in Section 2.4 we present the test set and in Section 2.5 we describe the experiments performed on the test set.

2.1 Training corpus collection and analysis

Corpus collection: to gain a better understanding of the variety of deposition statements based on previous work by Pruwor and Chapman (2008a) and Ochsnert et al. (2008) that we extended. Specifically, 112 microarray deposition statements from 105 articles were obtained using the existing corpora. After a manual review of these statements, two strategies were devised to collect additional statements. Our regular expression strategy consisted in two steps. First, the Ochsnert et al. query[1] was used to retrieve 2008 articles in PubMed Central. Second, articles were segmented into sentences and sentences likely to report data deposition were retrieved if they met the three following criteria: (i) sentence length was between 50 and 500 characters to avoid section titles and sentence segmentation errors; (ii) sentence contained a mention of GEO

[1] microarray[All Fields] OR genome-wide[All Fields] OR microarrays[All Fields] OR ‘expression profile’[All Fields] OR ‘expression profiles’[All Fields] OR ‘transcription profiling’[All Fields] OR ‘transcriptional profiling’[All Fields] AND (Endocrinology[jour] OR Mol Endocrinol[jour] OR J Biol Chem[jour] OR Proc Natl Acad Sci U S A[jour] OR Mol Cell Biol[jour] OR Nature[jour] OR Nat Med[jour] OR Nat Cell Biol[jour] OR Nat Genet[jour] OR Nat Struct Mol Biol[jour] OR Science[jour] OR Cancer Res[jour] OR FASEB J[jour] OR Cell[jour] OR Nat Methods[jour] OR Mol Cell[jour] OR J Immunol[jour] OR Immunol[jour] OR EMBO J[jour] OR Blood[jour]).
and (4) were kept as examples of non-deposition statements, and used in the pool of deposition statements. The remaining 110 sentences (such as (3) assigned, archived. submit, submission, available, access, uploaded, entered, posted, provided, contained one deposition action seed from the following: deposit, found, 3308

pattern [micro−array … data

et al. or ArrayExpress, which are the largest databases for microarray data (Stokes et al., 2008) or a mention of a GEO or ArrayExpress accession number or the pattern [micro]array … data/experiment/analyz; and (iii) sentence contained one deposition action seed from the following: deposit, found, submit, submission, available, access, uploaded, entered, posted, provided, assigned, archived.

After manual review, 133 of the 243 candidate sentences were added to the pool of deposition statements. The remaining 110 sentences (such as (3) and (4)) were kept as examples of non-deposition statements, and used in our machine learning strategy to retrieve deposition statements for data other than microarray.

In the machine-learning strategy, we aimed at enriching our training corpus, as proposed by Yegana et al. (2011). A simple (i.e. only using sentence tokens as features) Naïve Bayes (NB) model was built using the 243 microarray data deposition statements as positive examples and ∼33,000 sentences (the 110 above non-deposition statements, plus sentences from MEDLINE abstracts that contained the word ‘deposit’ or ‘deposited’) as negative examples.

In spite of our blanket assumption that the sentences extracted from MEDLINE abstracts were non-deposition statements, we did expect a small number of them to be actual deposition statements. Our reasoning was that the proportion of true non-deposition statements would be high enough to train an efficient model, while applying the model on the set of so-called negatives, it would rank the deposition statements high enough to collect them and adjust our training sets. By iterating on this method recursively, we finally obtained a training set of 586 positive or deposition statements (including the initial 243 microarray deposition statements) and 578 negative or non-deposition statements that scored high with the model (including the initial 110 non-deposition statements). This set was used as training data for building NB and SVM data deposition models, and will be referred to as Train-D (Fig. 1).

Analysis of deposition elements: to better characterize deposition statements, sentences were tagged for components referring to data, deposition action and deposition location using the following guidelines:

‘Data’: a phrase referring to biological data that can be found in public repositories. Patent data and data relevant to ClinicalTrials.gov were not considered. However, generic references to data were marked, when used in the context of biological data. This included expressions such as ‘the data’, ‘the protein’, ‘RNA’, ‘DNA’. In addition, specific references to data such as ‘5p53 conditional knockout mouse aCGH data’ were also marked.

‘Action’: a phrase describing the action undertaken by authors regarding depositing data. This includes phrases such as: deposit, submit, upload/download, is available, can be found, etc.

‘General Location’: reference to the location of data deposition, e.g. public repository name or website URL (e.g. http://www.ncbi.nlm.nih.gov/genbank/). This also includes a reference to an organization hosting a public repository in the context of data deposition.

‘Detailed Location’: detailed reference to the location of data deposition. This includes accession numbers and specific URLs allowing direct access to the data deposited (e.g. http://www.ncbi.nlm.nih.gov/nuccore/GQ386843).

(1t-4t) show how the statements exemplified in (1–4) were tagged.

(1) <data>The sequence of labA</data>
  <action>has been deposited</action>
  <location="general">in the DDBJ/GenBank/EMBL databases</location>
(2) <data>The microarray data</data>
  <action>were submitted</action>
  <location="general">to MIAMExpress at the EMBL-EBI</location>
(3) <data>Hi-cHistone TAG Arrays</data> are a repurposing of a microarray design originally created to represent <data>the TAG sequences</data> in the Yeast Knockout collection (Yuan et al., 2005 <location="general">NCBI GEO</location>
(4) Therefore, <data>the primary sequence of native Acinetobacter CMO</data> is identical to <data>the gene sequence for

A.Névéol et al.
Overview of component occurrences in data deposition statements

| Component     | Unique occurrences | Total occurrences | Variability (%) |
|---------------|--------------------|-------------------|-----------------|
| Data          | 468                | 645               | 73              |
| Action        | 77                 | 611               | 13              |
| Location (general) | 387           | 584               | 66              |
| Location (detailed)  | 521          | 534               | 98              |

*When accession numbers were unified, the variability lessened considerably with 71 unique occurrences only.*

Table 1.

Based on this tagging effort, Table 1 shows a summary of component occurrences over the corpus of 586 deposition statements. Only 16% of sentences contain information that is not included in one of the four tags (7% for full-text sentences, 24% for abstract sentences).

Data' is a category with high variability. While general references to data such as 'the data reported in this paper' (25 occurrences), 'the microarray data' (22 occurrences) and 'the sequences' (20 occurrences) are the most frequent phrases used, they are not prevalent overall.

'Action' is the category with the least variability. It is expressed by verbs in most cases. In other (rare) cases, nominalization expresses the action, e.g. 'the deposition' or 'deposited action'.

General Location' is also of high variability, in spite of the fact that there are a limited number of locations referenced, such as GenBank or GEO. Variation factors are as follows: (i) preposition introducing the location (e.g. at/from/in/into/through/onto/in/via/with/within); (ii) URL used (e.g. http://www.ncbi.nlm.nih.gov/); (iii) use of full name and/or abbreviation for institutes (NCBI, EBI) and database (GEO); (iv) types, spelling errors and other variation (e.g. database versus data bank).

Detailed Location' is a category with relatively low variability if we consider accession numbers as one token type. Variation factors are as follows: (i) preposition introducing the location through/under/with; (ii) reference to 'accession number' - code/number/no/(super)series; and (iii) list of numbers versus only one number. In the case of a list, a specific data description may be embedded within the list.

Automatic identification of deposition components in sentences

Based on the analysis above, the identification of the four data deposition components (data, deposition action, general location and specific location) in deposition statements appeared to be important for extracting specific deposition information. To provide a complete description of the sentences, any part not covered by the four tags was considered belonging to a default tag, 'nil'. In addition, we anticipated that these components might be useful features for the classification of deposition statements. For this reason, in addition to the 586 data deposition sentences tagged, another 697 non-deposition sentences were also tagged manually. The negative sentences tagged here are different from the 586 sentences used to train the SVM classifier in order to provide a good balance of sentences that might be useful for the classification of deposition statements. For this reason, in addition to the 586 data deposition sentences tagged, another 697 non-deposition sentences were also tagged manually. The negative sentences tagged here are different from the 586 sentences used to train the SVM classifier in order to provide a good balance of sentences that might be useful features for the classification of deposition statements.
3 RESULTS

3.1 Sentence-level classification

Table 3 shows the performance of selected NB and SVM models for sentence classification on the two sentence test sets (the NB models were applied to Test-SB while the SVM models were applied to Test-SS). While differences between the models were very small for cross-validation on the training set, some of them are emphasized on the test set, in particular between the different feature sets. The overall performance obtained was 89% F-measure—which corresponds to 87% accuracy. This accuracy compares favorably to the inter-annotator agreement computed on a subset of 100 sentences that was found to be 95%. The classification results from the best model comprised 39 sentences misclassified as negative and 56 sentences misclassified as positive.

We performed an error analysis in order to assess the underlying cause of these errors, and manually reviewed all misclassified sentences. We found that error causes and distribution were similar for NB and SVM. The breakdown of errors by cause (for SVM) is shown in Table 4.

The major sources of error are top sentences that score low in spite of being deposition sentences and sentences that report data reuse and not data deposition. The error analysis also brought to attention eight sentences (marked as ‘GS dispute’ in Table 4) that proved difficult to judge and/or were examples of genuine errors in the gold standard. These problematic sentences seem to be within the limits of 95% annotation consistency determined on the 100 sentences set.

### Table 3. Overall precision (P), recall (R), F-measure (F) and accuracy (A) of NB and SVM models for sentence classification

| Model       | Features                                    | P  | R  | F  | A  |
|-------------|---------------------------------------------|----|----|----|----|
| NB Tokens, position, POS tags | 60 84 70 75 |  | | |
| SVM Tokens, position, POS tags | 74 81 77 84 |  | | |
| SVM Tokens, position, POS tags | 78 83 80 87 |  | | |

Threshold is set at the 25th percentile of model scores on the training set Train-D.

### Table 4. Error analysis for SVM sentence classification

| Classification error | Error type          | Cases |
|----------------------|---------------------|-------|
| False negative       | Low score           | 34    |
|                      | GS dispute          | 2     |
|                      | Ambiguous sentence  | 3     |
|                      | Total               | 39    |
| False positive       | Data reuse          | 32    |
|                      | Database mention    | 7     |
|                      | Ambiguous sentence  | 7     |
|                      | GS dispute          | 6     |
|                      | Non biological data | 4     |
|                      | Total               | 56    |
Table 5. Positive precision (P), recall (R) and F-measure (F) of SVM models for article classification on test set

| Model | Features                        | P | R | F  |
|-------|---------------------------------|---|---|----|
| NB    | Tokens, position, POS tags      | 67| 82| 74 |
|       | Above features plus component tags | 83| 78| 81 |
| SVM   | Tokens, position, POS tags      | 82| 75| 79 |
|       | Above features plus component tags | 86| 76| 81 |

Threshold is set at the 25th percentile of model scores on the training set Train-D. The best performance is shown in bold characters.

Table 6. Error analysis for article classification with NB model

| Error type                        | Cases |
|-----------------------------------|-------|
| Low score                         | 49    |
| Ranked in top 5                   |       |
| Ranked in top 10                  | 2     |
| Other rank                        | 2     |
| No deposition sentence found in article | 6     |
| Sentence not scored (length >500) | 2     |
| Total                             | 61    |

3.2 Article-level classification

Table 5 shows the performance of NB and SVM models for article classification on the test set (both models were applied to Test-A, based on their respective results obtained from Test-SB for NB and Test-SS for SVM). As could be expected, the best results are obtained with the models including component tags as features, which also perform best for sentence classification.

The classification results from the best NB model comprised 61 articles misclassified as negative and 44 articles misclassified as positive. For this part of the study, we focused on positive articles that were classified as negative by our system, in order to assess the underlying cause of the errors. Each case was manually reviewed. The breakdown by cause is shown in Table 6.

The major cause for article misclassification is the direct result of a sentence classification error: the relevant deposition sentence was assigned a score below the threshold. Nonetheless, in these cases, relevant sentences appear in the top five scored sentences in 49 out of 53 low scoring cases. Another interesting result from the error analysis is the fact that six articles did not contain a deposition sentence in the full text, and therefore could not be classified properly by our system.

Figure 2 presents a more specific comparison between NB and SVM models built using all feature types.

3.3 Overall estimation of data deposition statement prevalence in the biomedical literature

To estimate the prevalence of data deposition statements in the biomedical literature, we processed all the PubMed Central articles available to us in full-text XML as of 2 June 2011 \((N = 827762)\) and used the NB model (with component tags as features) to classify them according to data deposition status. In total, \(~6\%\) of articles \((N = 52932)\) were found to contain a data deposition statement according to our method. For the subset of articles that are part of the MEDLINE database \((N = 774442)\), we also processed abstracts with our method and found that \(<0.1\%\) contained a deposition sentence. About 4.2% of the PubMed Central articles included in MEDLINE \((N = 32651)\) had a curated [si] field. Most of these PubMed Central articles were also classified as positive for data deposition by our method \((N = 22428)\). This is consistent with the results of our evaluation.

4 DISCUSSION

Choice of features: interestingly, the difference in performance with and without component tags observed in the cross-validation was greatly magnified on the test sets both for sentence-level and article-level classification. We think this is due to the more challenging nature of the test data. In previous work (e.g. Kim et al., 2011) on MEDLINE abstract sentence classification, structure information has proved to be a useful feature when it is available. Our positive feature was intended to serve as a substitute for structure information, but had little impact. Structure information could be considered for future improvement of the sentence classifier; however, this information is not trivial to extract from abstracts (McKnight and Srinivasan, 2003; Ripple et al., 2011); similar issues with added complexity can be anticipated for full text.

Portability of the method: although trained mainly on microarray data deposition statements, the method adapts well to the identification of statements reporting on the deposition of other data such as gene sequences or protein coordinates. This is evidenced...
by the database breakdown of articles in our test corpus according to the MEDLINE [si] field: 110 articles reported data deposition in GenBank, 50 in PDB and only 47 in GEO.

Comparison to other work: while similar to other work mentioned in the related work section, our approach is not directly comparable to any of the previous studies on data deposition. At the article level, we perform an automatic classification of articles containing data deposition statements, in contrast with Ochsner et al. who performed a one-time manual classification in order to assess the rate of data deposition in 2008. Piwowar et al. assessed machine learning and rule-based algorithms for article classification. However, their approach focused on five databases and relied on the identification of predetermined database names in the full text of articles. In contrast, our approach is general and aiming at the automatic identification of any biological data deposition in any public repository. Nonetheless, as an indicative comparison, it can be noted that our overall performance for article-level classification is 81% F-measure, compared with overall 69% for Piwowar et al. (on a different evaluation corpus).

Furthermore, in addition to article classification, our approach also retrieves specific data deposition elements allowing for a finer-grained characterization of both data and deposition location. At the statement level, this is also different from the classification of databank accession number sentences performed by Kim et al. (2010) in two ways: first, we are only interested in retrieving sentences containing accession numbers if these sentences are data deposition statements (versus comment on the data, or data reuse) and second, we are also interested in retrieving data deposition statements that do not contain accession numbers.

Interest of this study: one general interest of this study is our application of the method proposed by Yeganova et al. in order to build a training corpus when large amounts of unlabeled data are available. We showed that the method of Yeganova et al. could be easily and successfully adapted to our specific classification scenario. More specific to data deposition statement classification, the method presented in this article can be used as a curation aid in MEDLINE or other databases indexing articles with accession numbers; this tool can also be used to help updating current databases. For instance, as announced in York (2006) GEO accession numbers; this tool can also be used to help updating current databases indexing articles with accession numbers.

Limitations of the study: our study addressed the identification of data deposition statements in full-text articles. While the availability of full-text is definitely a limitation, our overall study of the prevalence of data deposition statements (Section 3.3) indicates that data deposition statements are significantly more often found in the full-text of articles, compared with abstract text. While our method is not limited to data deposition in databases specifically curated in MEDLINE, it is focused on the deposition of biological data as opposed to clinical data as might be recorded in ClinicalTrials.gov, one of the [si] curated databases. Finally, our classification results are obtained based on a threshold of sentence score, which was empirically established at the 25th percentile of model scores on the training data. Other methods of determining the threshold could be investigated in future work.

5 Conclusion
We presented a method to automatically identify biological data deposition statements in biomedical text. The method, an SVM (or, equivalently, NB) classifier using CRF-determined features characterizing data deposition components, was able to correctly identify articles containing data deposition statements with 81% F-measure. Our analysis shows that deposition statements are scored high for all types of databases and biological data types, even those not currently curated in MEDLINE. This shows the potential impact of our method for literature curation. In addition, we believe it provides a robust tool for future work assessing the need for storage space of incoming data in public repositories.

Acknowledgements
The authors would like to thank Dr Won Kim for his assistance in implementing the NB and SVM classification tools.

Funding: Intramural Research Program of the National Institutes of Health, National Library of Medicine. Grant number LM000002-01.

Conflict of interest: None declared.

References
Anonymous (2008) Thou shall share your data. Nat. Methods, 5, 209.
Denner-Fushman,D. et al. (2006) Automatically identifying health outcome information in MEDLINE records. J. Am. Med. Inform. Assoc., 13, 52-60.
Hauswieser,M. et al. (2011) Annotationing genes and genomes with DNA sequences extracted from biomedical articles. Bioinformatics, 27, 980-986.
Kim,J. et al. (2010) Naïve Bayes and SVM classifiers for classifying databank accession number sentences from online biomedical articles. In IS&T/SPIE’s 22nd Annual Symposium on Electronic Imaging, San Jose, CA, 7534-75340C-1-8.
Kim,S. et al. (2011) Automatic classification of sentences to support Evidence Based Medicine. BMC Bioinformatics, 12 (Suppl. 1), S5.
Krippinger,M. et al. (2008) Overview of the protein-protein interaction annotation extraction task of BioCreative II. Genome Biol., 9 (Suppl. 2), S4.
McKennai,L. and Steniusa,P. (2001) Categorization of sentence types in medical abstracts. AMIA Annu. Sympos. Proc., 2008, 440-444.
Mienchen,F. et al. (2008) Emerging trend prediction in biomedical literature. AMIA Annu. Sympos. Proc., 485-489.
Ochsner,S. et al. (2008) Much room for improvement in deposition rates of expression microarray datasets. Nat. Methods, 5, 991.
Piwowar,H.A. et al. (2007) Sharing detailed research data is associated with increased citation rate. PLoS One, 2, e108.
Piwowar,H.A. and Chapman,W.W. (2008a) Linking database submissions to primary citations with PubMed Central. In Proceedings of the BioLINK workshop at ISMB. Toronto, Canada.
Piwowar,H.A. and Chapman,W.W. (2008b) Identifying data sharing in biomedical literature. AMIA Annu. Sympos. Proc., 2008, 596-600.
Piwowar,H.A. and Chapman,W.W. (2010) Public sharing of research datasets: a pilot study of associations. J. Informatics, 4, 148-156.
Polajnar,T. et al. (2011) Protein interaction sentence detection using multiple semantic kernels. J. Biomed. Semantics, 2, 7.
Rippe,A.M. et al. (2011) A retrospective cohort study of structured abstracts in MEDLINE, 1992–2006. J. Med. Libr. Assoc., 99, 160-163.
Sebastiani,F. (2002) Machine learning in automated text categorization. ACM Comput. Surv., 34(1), 1-47.
Smith,D. et al. (2004) MedPost: a part-of-speech tagger for biomedical text. Bioinformatics, 20, 2320-2321.
Stokes,TH. et al. (2008) ArrayWiki: an enabling technology for sharing public microarray data repositories and meta-analytics. BMC Bioinformatics, 9, (Suppl. 6), S18.
Yeganova,L. et al. (2011) Text mining techniques for leveraging positively labeled data. In Proceedings of the ACL Workshop BioNLP, pp. 155-161.
Yorks,M. (2006) GEO accession numbers in MEDLINE. NLM Tech. Bull., 349, e5.