Automatic extraction of data deposition sentences: where do the research results go?

Aurélie Névéol, W. John Wilbur, Zhiyong Lu
National Center for Biotechnology Information
U.S. National Library of Medicine
Bethesda, MD 20894, USA
{Aurelie.Neveol,John.Wilbur,zhiyong.lu}@nih.gov

Abstract

Research in the biomedical domain can have a major impact through open sharing of data produced. In this study, we use machine learning for the automatic identification of data deposition sentences in research articles. Articles containing deposition sentences are correctly identified with 73% f-measure. These results show the potential impact of our method for literature curation.

1 Background

Research in the biomedical domain aims at furthering the knowledge of biological processes and improving human health. Major contributions towards 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 to actual data production. For instance, Ochsner et al. (2008) found the deposition rate of microarray data to be less than 50% for work published in 2007.

Information about the declaration of data deposition in research papers can be used both for data curation and for the analysis of emerging research trends. Our long-term research interest is in assessing the value of deposition sentences for predicting future trends of data production. The initial step of automatically identifying deposition sentences would then lead to an assessment of the need for storage space of incoming data in public repositories.

2 Objective

In this study, we aim at automatically performing a fine-grained identification of biological data deposition sentences in biomedical text. That is, we aim at identifying articles containing deposition sentences, extracting the specific sentences and characterizing the information contained in the sentences in terms of data type and deposition location (e.g. database, accession numbers).

3 Material and Methods

Data deposition sentences. A collection of sentences reporting the deposition of biological data (such as microarray data, protein structure, gene sequences) in public repositories was compiled based on previous work that we extended. We take these sentences 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 sentences. In contrast, (3) and (4) contain elements related to data deposition while focusing on other topics.

(1) The sequences reported in this paper have been deposited in the GenBank database (accession numbers AF034483 for susceptible strain RC688s and AF034484 for resistant strain HD198r).

(2) The microarray data were submitted to MIAMExpress at the EMBL-EBI.

(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).

(4) The primary sequence of native Acinetobacter CMO is identical to the gene sequence for chnB deposited under accession number AB006902.

Proceedings of the 2011 Workshop on Biomedical Natural Language Processing, ACL-HLT 2011, pages 103–104,
Portland, Oregon, USA, June 23-24, 2011. ©2011 Association for Computational Linguistics

Sentence classification. A Support Vector Machine (SVM) classifier was built using a corpus of 583 positive data deposition sentences and 578 other negative sentences. Several sets of features were tested, including the following: sentence tokens, associated part-of-speech tags obtained using MEDPOST[1], relative position of the sentence in the article, identification of elements related to data deposition (data, deposition action, database, accession number) obtained using a CRF model[2].

Article classification. The automatic classification of articles relied on sentence analysis. The full text of articles was segmented into sentences, which were then scored by the sentence-level SVM classifier described above. An article is classified as positive if its top-scored sentence is scored higher than a threshold, which is predetermined as the 25th percentile score for positive sentences in the training set.

Evaluation corpus. A corpus composed of 670 PubMed Central articles was used to evaluate article classification. 200 articles were considered as “positive” for data deposition based on MEDLINE gold standard annotations in the [si] field used to curate newly reported accession numbers.

4 Results

Table 1 shows the performance of selected SVM models for article classification on the test set. While differences were very small for cross-validation on the training set, they are emphasized on the test set.

Features	**P**	**R**	**F**
Tokens, position, part-of-speech tags	52%	56%	54%
Token, position, CRF+, part-of-speech tags	65%	58%	62%
Tokens, position, CRF+/-, part-of-speech tags	**69%**	**78%**	**73%**

Table 1: Precision, Recall and F-measure of SVM models for article classification on test set.

5 Discussion and Conclusion

Portability of the method. Although trained mainly on microarray data deposition sentences, the method adapts well to the identification of other data deposition sentences, e.g. gene sequences, protein coordinates.

Comparison to other work. Our approach is not directly comparable to any of the previous studies. At the article level, we perform an automatic classification of articles containing data deposition sentences, in contrast with Oshner et al. who performed a one-time manual classification. Piwowar et al used machine learning and rule-based algorithms for article classification. However, they relied on identifying the names of five predetermined databases in the full text of articles. Our approach is generic and aiming at the automatic identification of any biological data deposition in any public repository. Furthermore, our approach also retrieves specific data deposition sentences where data and deposition location are identified. At the sentence level, this is also different from the classification of databank accession number sentences performed by Kim et al. (2010) in two ways: first, we focus on retrieving sentences containing accession numbers if they are deposition sentences (vs. data re-use, etc.) and second, we are also interested in retrieving data deposition sentences that do not contain accession numbers.

Error analysis. Almost half of the articles classified as containing a deposition sentence by our method but not by the gold standard were found to indeed contain a deposition sentence.

Conclusion. These results show the potential impact of our method for literature curation. In addition, it provides a robust tool for future work assessing the need for storage space of incoming data in public repositories.

Acknowledgments

This research was supported by the Intramural Research Program of the NIH, NLM.

References

Jongwoo Kim, Daniel Le, Georges R. Thoma. Naïve bayes and SVM classifiers for classifying databank accession number sentences from online biomedical articles. Proc. SPIE 2010 (7534): 7534OU-OU8

Scott A. Ochsner, Davd L Steffen, Christian J Stoeckert Jr, Neil J. McKenna. Much room for improvement in deposition rates of expression microarray datasets. Nat Methods. 2008 Dec;5(12):991.

Heather A. Piwowar, Wendy W. Chapman. Identifying data sharing in biomedical literature.AMIA Annu Symp Proc. 2008 Nov 6:596-600.

[1] http://www.ncbi.nlm.nih.gov/staff/lsmith/MedPost.html

[2] http://mallet.cs.umass.edu/