BIOINFORMATICS APPLICATIONS NOTE Vol. 28 no. 17 2012, pages 2285–2287
doi:10.1093/bioinformatics/bts435

Data and text mining

Advance Access publication July 12, 2012

MyMiner: a web application for computer-assisted biocuration and text annotation

David Salgado[1,†,*], Martin Krallinger[2,†], Marc Depaule[3], Elodie Drula[3,4], Ashish V. Tendulkar[5], Florian Leitner[2], Alfonso Valencia[2] and Christophe Marcelle[1,*]

[1]Australian Regenerative Medicine Institute (ARMI), Monash University, Clayton, Victoria 3800, Australia, [2]Structural Biology and BioComputing Programme, Spanish National Cancer Research Centre (CNIO), C/ Melchor Fernández Almagro, 3, E-28029 Madrid, Spain, [3]Developmental Biology Institute of Marseille Luminy (IBDML), CNRS UMR 6216, Université de la Méditerranée, Campus de Luminy, 13288 Marseille Cedex 09, France, [4]AFMB, UMR 6098 CNRS/UI/UII, Case 932, 13288 Marseille Cedex 9, France and [5]Department of Computer Science and Engineering, Indian Institute of Technology (IIT), Madras, Chennai 600 036, India

Associate Editor: Jonathan Wren

ABSTRACT

Motivation: The exponential growth of scientific literature has resulted in a massive amount of unstructured natural language data that cannot be directly handled by means of bioinformatics tools. Such tools generally require structured data, often generated through a cumbersome process of manual literature curation. Herein, we present MyMiner, a free and user-friendly text annotation tool aimed to assist in carrying out the main biocuration tasks and to provide labelled data for the development of text mining systems. MyMiner allows easy classification and labelling of textual data according to user-specified classes as well as predefined biological entities. The usefulness and efficiency of this application have been tested for a range of real-life annotation scenarios of various research topics.

Availability: http://myminer.armi.monash.edu.au.

Contacts: david.salgado@monash.edu and christophe.marcelle@monash.edu

Supplementary Information: Supplementary data are available at *Bioinformatics* online.

Received on April 1, 2012; revised on June 29, 2012; accepted on July 3, 2012

1 INTRODUCTION

Scientific literature is one of the main sources of knowledge on which working hypotheses are built. To organize and represent biological information for researchers as well as for bioinformatics analyses, many literature biocuration efforts have been carried out (Howe *et al*., 2008). Their aim is to extract biologically relevant information from articles and transform it into structured database records. With regard to the rapid literature growth, text-mining efforts are becoming increasingly important to speed up this process. Powerful infrastructures such as GATE (General Architecture for Text Engineering) (Cunningham *et al*., 2011) usually require a good degree of technical expertise and sometimes even basic knowledge of computational linguistics, making them challenging to use for biocuration purposes. To develop text-mining systems, it is crucial to produce high-quality labelled textual data that can serve as training and test datasets. At present, most biomedical text mining applications/methods utilize a restricted number of annotated text data (corpora) prepared by biology domain experts. To help increasing this number, we have implemented MyMiner, an online annotation tool intended for expert biologists with no programming skills. Compared to other applications [e.g. BRAT (Stenetorp *et al*., 2012), PubTator (Wei *et al*., 2012), Knowtator (Ogren, 2006), NCBO Annotator (Jonquet *et al*., 2008) and DOMEO/SWAN (Ciccarese *et al*., 2012) (Supplementary Table 2)], MyMiner provides within one single web-based application user interfaces to classify documents, compare classification efficiencies, detect entities to annotate texts and link them to identifiers.

2 TOOL DESCRIPTION AND FUNCTIONALITY

MyMiner is an interactive web application based on a modular design with the purpose to assist users in biocuration and text annotation tasks. The MyMiner interface is intended to be user-friendly, not requiring installation of any local software. Each module has an export option for saving results. The time spent for processing a document is recorded in the exported file. To improve the user-friendliness, a common display layout has been adopted and conserved between application modules. The input document analysis area is located on the top of the page; options and tools are placed below the main curation zone. MyMiner combines PHP, JavaScript and AJAX to enhance user interactivity. The core of the MyMiner system covers four application modules that can be independently used or combined together following the steps of a biocuration pipeline (Supplementary Fig. S1).

MyMiner handles any plain text, including article abstracts, document sentences, ontology terms or disease descriptions.

(1) The 'File-labelling' module is a simple to use manual text classification interface that allows classifying documents, abstracts, sentences or terms, offering the possibility to enter user-specified class labels. This module could be

†The authors wish it to be known that, in their opinion, the first two authors should be regarded as joint First Authors.
*To whom correspondence should be addressed.

used for instance to classify documents as either relevant or not to a specific topic from a PubMed query. Its purpose is to cover the triage task (article selection) carried out by database annotators but it can also be used for any manual classification recording. The labelled data that result from this classification can serve as training and test sets for text categorization systems. To reduce the manual classification time, it provides the option of dynamically setting positive and negative text highlights. These are expressions that users can set at any time during the tagging process to highlight text relevant (marked in yellow) or non-relevant (marked in red) to the topic of interest. The system offers the possibility to upload the classification guidelines so that the annotator can refer to them when necessary. Users can pause and resume the curation process at any time by saving the classified document. To resume classification, the saved file is uploaded as input file. The time spent by a user to select the corresponding label is recorded. This may be useful to estimate the efficiency of annotators and the difficulty of the task.

(2) The 'Compare File' module facilitates the direct comparison of collections of labelled items generated by several approaches or persons. In addition, it is possible to create subsets from these collections based on the agreement or disagreement of annotation labels. This module could be used to compare and evaluate document classification methods between various persons or softwares. It displays a global summary with information covering: (i) the number of documents within each class; (ii) the average time needed to classify the text; (iii) the correlation between classification time and text length or (iv) the number of items tagged differently between annotators. This module allows extracting a collection of texts (known as a Gold Standard corpus) that have been labelled consistently by all annotators. Alternatively, the module can also be used to extract the borderline cases tagged differently. Hasty/inaccurate annotations can be detected from inter-annotator disagreements and/or a poor correlation between document size versus classification time. These cases can then be used to refine and improve the classification guidelines. The Compare File module has been used to estimate consistency of manual annotations between various individuals and methods (Supplementary Data Section 2).

(3) The purpose of the 'Entity Tagging' (entity mention recognition) module is to manually detect important conceptual objects within a document, a first step for further identification of annotation events and relations to populate knowledge databases. This module could be used to create a corpus of gene and protein mentions to test and train a Named Entity Recognition tool. This module offers an interactive interface allowing users to semi-automatically identify various kinds of entities within documents. It has been designed as a WYSIWYG (What You See Is What You Get) online editor that allows the addition of user-specified labels for new entity types. For the detection of important bio-entities, this module provides the automatic recognition of proteins, DNA, RNA, cell lines and cell types by integrating the ABNER tagger (Settles, 2005). The LINNAEUS system is incorporated into MyMiner to identify species and organisms (Gerner *et al*., 2010). Additionally, user-defined entities can be detected if terms-tags dictionaries are provided. To improve the accuracy of the annotations, tags can be edited and wrongly generated labels can be removed. To define simple relationships between entities and terms, a matrix check box display was added to this module (Supplementary Fig. S14).

(4) The 'Entity Linking' module facilitates the manual annotation of bioentities mentioned in a document with standardized identifiers. This module could be used to manually link articles to disease and protein identifiers to create a catalogue of proteins involved in pathologies. Gene/protein names are automatically recognized and displayed as a list that can be manually edited, and new entities can be added and incorrectly identified ones can be removed. For each gene/protein name, MyMiner suggests a ranked list of UniProt identifiers that utilize the UniProt search scoring mechanism (Arighi *et al*., 2011). Species mentions are normalized to NCBI taxon identifiers; OMIM identifiers are associated to diseases and ontology terms are linked to identifiers from submitted ontology files. For this purpose, MyMiner launches asynchronous queries to respective databases (UniProt, NCBI taxonomy, OMIM and user provided ontology file) using AJAX requests. For organisms, proteins, diseases and ontology terms, a short description is displayed to help validate potential candidate hits and to assist during the manual disambiguation of potential databases identifiers. Check boxes allow the selection of the most appropriate identifiers from the candidate list. If species are specified prior to a protein identifier search, species-specific constraints are applied to reduce the number of potential candidates from UniProt.

3 EVALUATION AND USER CASES

3.1 Evaluation

MyMiner was compared to manual text annotations generated by unassisted or assisted human annotators (through a command line script). This showed a decrease of annotation time (up to 90%, average 70%), with no change in the quality of annotation (Supplementary Data Section 2 and Figs 2–13). MyMiner has been evaluated, by the BioCreative User Advisory Group in Biocreative III, Interactive Task (IAT) (Arighi *et al*., 2011).

3.2 User cases/usefulness

In the context of the BioCreative III competition, MyMiner was utilized to label over 20 000 abstracts to prepare datasets for protein interaction relevance (Krallinger *et al*., 2011), producing the necessary datasets for the implementation of the protein interaction information extraction (PIE) abstract retrieval system (Kim *et al*., 2012). It can thus be applied to prepare

training data for document categorization and ranking applications such as MedlineRanker (Fontaine *et al.*, 2009) (Supplementary Fig. S15). MyMiner was used to validate ranked lists of documents, terms and bio-entities for creating reference sets for systems biology (Krallinger *et al.*, 2009). MyMiner was used in the muscle biology field to create labelled data for the extraction of mutations and gene regulation sentences as well as to annotate experimental evidence cues (Krallinger, 2010).

4 CONCLUSION

MyMiner is an online application designed to allow biologists and non-IT experts to semi-automatically classify and annotate (biomedical) information in text with the help of text-mining tools. It allows validating these automatically generated results, as it provides a simple interface to manipulate the recognized bio-entities and permits the addition of new labels for manual tagging. This system can be exploited to efficiently label text for document classification and prioritization tasks, for entity recognition purposes and for manual annotation of relationships between entities. It has been used during various annotation efforts by diverse groups of biomedical scientists. During the past 2 years, MyMiner has been used by 240 unique visitors.

On the MyMiner webpage we provide a comprehensive set of related systems to point visitors to complementary software. The MyMiner repository offers the possibility to host and share corpora provided by the user community.

ACKNOWLEDGEMENTS

We acknowledge the help of Stephen Dart and colleagues from Monash University e-Research; feedback from Astrid Laegreid and Miguel Vazquez and help in the classification evaluation by Charlène Guillot and Camille Gueniot.

Funding: MyMiner was funded by grants from the French Association against Myopathies (AFM) and by the EU 6th Framework Programme Network of Excellence MYORES. CNIO participation was funded by MICROME (Grant Agreement Number 222886-2). The Australian Regenerative Medicine Institute is supported by grants from the State Government of Victoria and the Australian Government.

Conflict of Interest: none declared.

REFERENCES

Arighi,C.N. *et al.* (2011) BioCreative III interactive task: an overview. *BMC Bioinformatics*, **12** (**Suppl 8**), S4.

Ciccarese,P. *et al.* (2012) Open semantic annotation of scientific publications using DOMEO. *J Biomed Sem*, **3**, S1.

Cunningham,H. *et al.* (2011) Text Processing with GATE (Version 6), University of Sheffield, Department of Computer Science, 2011.

Fontaine,J.-F. *et al.* (2009) MedlineRanker: flexible ranking of biomedical literature. *Nucleic Acids Res.*, **37**, W141–W146.

Gerner,M. *et al.* (2010) LINNAEUS: a species name identification system for biomedical literature. *BMC Bioinformatics*, **11**, 85.

Howe,D. *et al.* (2008) Big data: the future of biocuration. *Nature*, **455**, 47–50.

Jonquet,C. *et al.* (2008) The open biomedical annotator. *Summit on Translat Bioinforma*, **2009**, 56–60.

Kim,S. *et al.* (2012) PIE the search: searching PubMed literature for protein interaction information. *Bioinformatics*, **28**, 597–598.

Krallinger,M. (2010) Importance of negations and experimental qualifiers in biomedical literature. In *Proceedings of the Workshop on Negation and Speculation in Natural Language Processing*. Association for Computational Linguistics, pp. 46–49.

Krallinger,M. *et al.* (2009) Creating reference datasets for systems biology applications using text mining. *Ann. N Y Acad. Sci.*, **1158**, 14–28.

Krallinger,M. *et al.* (2011) The Protein-Protein Interaction tasks of BioCreative III: classification/ranking of articles and linking bio-ontology concepts to full text. *BMC Bioinformatics*, **12** (**Suppl 8**), S3.

Ogren,P. (2006) Knowtator: a Protégé plug-in for annotated corpus construction. In In *The Proceedings of the 2006 Conference of the North American Chapter of the Association for Computational Linguistics on Human Language Technology*. pp. 273–275.

Settles,B. (2005) ABNER: an open source tool for automatically tagging genes, proteins and other entity names in text. *Bioinformatics*, **21**, 3191–3192.

Stenetorp,P. *et al.* (2012) BRAT: A Web-based Tool for NLP-Assisted Text Annotation. In *Proceedings of the Demonstrations Session at EACL 2012*. (to appear) 2012.

Wei,C. *et al.* (2012) PubTator: a PubMed-like interactive curation system for document triage and literature curation. In *Proceedings of the BioCreative 2012 workshop*. pp. 145–150.