BIOINFORMATICS APPLICATIONS NOTE Vol. 29 no. 15 2013, pages 1915–1916
doi:10.1093/bioinformatics/btt317

Data and text mining Advance Access publication June 4, 2013

BeCAS: biomedical concept recognition services and visualization

Tiago Nunes*, David Campos, Sérgio Matos and José Luís Oliveira*

DETI/IEETA, University of Aveiro, Campus Universitário de Santiago, 3810 - 193 Aveiro, Portugal

Associate Editor: Jonathan Wren

ABSTRACT

Summary: The continuous growth of the biomedical scientific literature has been motivating the development of text-mining tools able to efficiently process all this information. Although numerous domain-specific solutions are available, there is no web-based concept-recognition system that combines the ability to select multiple concept types to annotate, to reference external databases and to automatically annotate nested and intercepted concepts. BeCAS, the Biomedical Concept Annotation System, is an API for biomedical concept identification and a web-based tool that addresses these limitations. MEDLINE abstracts or free text can be annotated directly in the web interface, where identified concepts are enriched with links to reference databases. Using its customizable widget, it can also be used to augment external web pages with concept highlighting features. Furthermore, all text-processing and annotation features are made available through an HTTP REST API, allowing integration in any text-processing pipeline.

Availability: BeCAS is freely available for non-commercial use at http://bioinformatics.ua.pt/becas.

Contacts: tiago.nunes@ua.pt or jlo@ua.pt

Received on March 13, 2013; revised on May 6, 2013; accepted on May 27, 2013

1 INTRODUCTION

Exponential growth of the biomedical literature during the past decades has prompted the development of automatic techniques and systems to ease the task of finding relevant information in unstructured documents (Lu, 2011). However, the offer of no-installation, no-maintenance and online modular solutions for concept annotation that can be easily integrated in any text-processing pipeline is still scarce. Whatizit (Rebholz-Schuhmann *et al.,* 2008), for instance, offers dictionary-based annotation of documents with a large set of vocabularies and is available both through a web service and a web page. Yet, annotation of concepts from different types is only possible by repeating the annotation process several times and combining the results generated by different pipelines. iHOP (Hoffmann and Valencia, 2004) is a web application offering programmatic access to pre-annotated abstracts from MEDLINE. It is a protein-centric system and does not allow the annotation of external documents submitted by the user. Another solution focused on genes, proteins and small-molecules is Reflect (Pafilis *et al.,* 2009), a web service that annotates these concepts on web pages and provides, through pop-ups, additional information such as synonyms, database identifiers and related literature. Cocoa is a multiple concept annotator with an online interface and an HTTP API (http://npjoint.com). It annotates entities in user submitted text, but it is limited to named entities and does not provide concept identifiers or external references.

Only few text-mining solutions for concept identification are available as web services, and most of them focus on a small number of entity types. Of those, most omit concepts that intersect other recognized concepts or that are nested within broader concepts. Moreover, to the best of our knowledge, there is no solution available that allows users to select the entity types they want to annotate on a single service invocation. BeCAS, the Biomedical Concept Annotation System, is a web-based tool for on-demand document processing and annotation that can be integrated on larger text-processing pipelines, used directly through a user-friendly and highly interactive web interface or incorporated on external web pages through a simple yet flexible widget.

2 IMPLEMENTATION

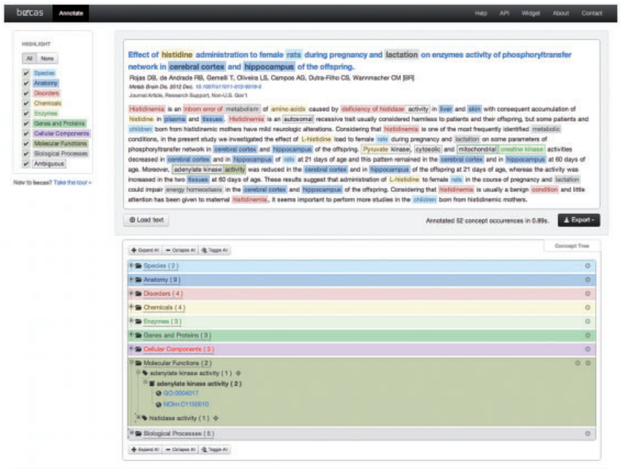
BeCAS is built on top of a modular system for biomedical concept recognition. It integrates modules for PubMed article fetching, sentence splitting, tokenization, lemmatization, POS tagging, chunking, concept identification, abbreviation resolution, external database identifier tagging and interactive visual concept highlighting. The text-processing modules were implemented in Java, the article fetching modules and web-services were built in Python and the web interface was developed using HTML, CSS and Javascript. Sentence splitting, tokenization, lemmatization, POS tagging and chunking are provided by a customized version of GDep (Sagae and Tsujii, 2007), a C++ parser wrapped by our Java modules. Concept identification modules for recognizing species, anatomical concepts, miRNAs, enzymes, chemicals, drugs, diseases, metabolic pathways, cellular components, biological processes and molecular functions apply deterministic finite automatons for dictionary matching. For this, we compiled a database of concepts from multiple meta-sources, including UMLS (Bodenreider, 2004), LexEBI (Sasaki *et al.,* 2008), Jochem (Hettne *et al.,* 2009) and NCBI BioSystems (Geer *et al.,* 2010). Genes and proteins are identified using a Conditional Random Fields tagger with entity normalization, built over Gimli (Campos *et al.,* 2013).

The various concept recognition modules were tested on the CRAFT (Bada *et al.,* 2012), AnEM (Ohta *et al.,* 2012) and NCBI Diseases (Doğan and Lu, 2012) corpora, achieving f-measure results for overlap matching of 76% for genes and proteins, 95% for species, 65% for chemicals, 83% for cellular components, 92% for cells, 63% for molecular functions and biological processes, 83% for anatomical entities and 85% for diseases. These results are on par with current state-of-the-art biomedical annotation systems.

3 ANNOTATING BIOMEDICAL TEXT

BeCAS exposes its functionalities through three interfaces: an HTTP REST API, a widget embeddable in web pages and an interactive web application. It provides annotations both for user-supplied texts

*To whom correspondence should be addressed.



Effect of histidine administration to female rats during pregnancy and lactation on enzymes activity of phosphoryltransfer network in cerebral cortex and hippocampus of the offspring.

Rojas DB, de Andrade RB, Gemelli T, Oliveira LS, Campos AG, Dutra-Filho CS, Wannmacher CM [BR]

Metab Brain Dis. 2012 Dec. 10.1007/s11011-012-9319-5

Journal Article, Research Support, Non-U.S. Gov't

Histidinemia is an inborn error of metabolism of amino acids caused by deficiency of histidase activity in liver and skin with consequent accumulation of histidine in plasma and tissues. Histidinemia is an autosomal recessive trait usually considered harmless to patients and their offspring, but some patients and children born from histidinemic mothers have mild neurologic alterations. Considering that histidinemia is one of the most frequently identified metabolic conditions, in the present study we investigated the effect of L-histidine load to female rats during pregnancy and lactation on some parameters of phosphoryltransfer network in cerebral cortex and hippocampus of the offspring. Pyruvate kinase, cytosolic and mitochondrial creatine kinase activities decreased in cerebral cortex and in hippocampus of rats at 21 days of age and this pattern remained in the cerebral cortex and in hippocampus at 60 days of age. Moreover, adenylate kinase activity was reduced in the cerebral cortex and in hippocampus of the offspring at 21 days of age, whereas the activity was increased in the two tissues at 60 days of age. These results suggest that administration of L-histidine to female rats in the course of pregnancy and lactation could impair energy homeostasis in the cerebral cortex and hippocampus of the offspring. Considering that histidinemia is usually a benign condition and little attention has been given to maternal histidinemia, it seems important to perform more studies in the children born from histidinemic mothers.

Annotated 52 concept occurrences in 0.89s.

Fig. 1. BeCAS web interface showing an annotated PubMed abstract. Each annotated entity type can be highlighted separately (left). The concept tree (bottom) displays all annotations along with the associated concepts and external references

and for MEDLINE abstracts, which are automatically fetched from PubMed.

3.1 Exploring annotations on a web interface

BeCAS web interface was built with a strong focus on usability. Specific entity types can be highlighted or muted in real-time by using simple toggle controls, and nested and intersected annotations are also easily identified by the colour-coding scheme used. An info-box with links to external databases is displayed by placing the mouse over highlighted entities, and users can explore this same information, grouped by concept type, through the concept tree (Fig. 1). Annotated text can be exported in several formats such as JSON and A1. Users and other websites can link to annotated PubMed publications by using direct links (e.g. http://bioinformatics.ua.pt/becas/pmid/22957306).

Concept highlighting with external references can easily be integrated in any website through the use of the BeCAS Javascript widget. Host pages only need to include a <script> tag linking to the plugin and a few configuration parameters. Every feature implemented in the main web interface is exposed by the widget, apart from the concept tree.

3.2 Processing text and MEDLINE publications

Text can be annotated programmatically using one of BeCAS HTTP REST endpoints. Clients should make HTTP POST requests to one of the endpoints with a JSON encoded payload, specifying the text to annotate, the desired output format and types of entities that should be annotated. Because of inherent representation constraints, the available output formats support different levels of granularity in the results. CoNLL format is the most comprehensive, providing sentence splitting, tokenization, lemmatization, POS tagging, chunking and identification of isolated, nested and intersected concepts. JSON format includes sentence splitting and concept identification. IeXML formatted results contain the same information as JSON, but nested and intersected annotations are limited to a depth of one level, with deeper annotations resolved to the largest span. Results in A1 format provide concept identifiers, including nested and intersected annotations.

Apart from supplying text directly to the API, BeCAS is capable of fetching and annotating PubMed articles. A client can issue an HTTP POST request to one of the abstract annotation endpoints, optionally providing a JSON-encoded payload of entity types for annotation. As publications have multiple fields, such as the title, abstract, authors, MeSH terms and others, results are provided exclusively as PubMed annotated IeXML or JSON. The service returns XML documents delivered by the Entrez eFetch Utility, with the ‘ArticleTitle’ and ‘AbstractText’ fields enriched with IeXML annotations.

Comprehensive documentation of all API methods and parameters, along with usage examples, is available online.

4 CONCLUSION

BeCAS provides three distinct user interfaces for biomedical concept identification, presenting state-of-the-art performance, as evaluated on various corpora. It currently recognizes and annotates 1.2 million concepts and enriches them with 1.6 million external references to 30 online resources. The REST API is suitable for integration in custom text-processing pipelines, whereas the widget can be easily integrated in any web page. Finally, users can also use BeCAS annotation services as a stand-alone web application. In the future, we plan to add support for more entity types and continue to improve annotation performance, with focus on concept disambiguation.

Funding: This work was supported by Fundação para a Ciência e Tecnologia [PTDC/EIA-CCO/100541/2008 (FCOMP-01-0124-FEDER-010029)] and FCT Ciência2007.

Conflict of Interest: none declared.

REFERENCES

Bada,M. *et al.* (2012) Concept annotation in the CRAFT corpus. *BMC Bioinformatics*, **13**, 161.

Bodenreider,O. (2004) The unified medical language system (UMLS): integrating biomedical terminology. *Nucleic Acids Res.*, **32** (**Suppl. 1**), D267–D270.

Campos,D. *et al.* (2013) Gimli: open source and high-performance biomedical name recognition. *BMC Bioinformatics*, **14**, 54.

Doğan,R.I. and Lu,Z. (2012) An improved corpus of disease mentions in PubMed citations. In *Proceedings of BioNLP’12*. Association for Computational Linguistics.

Geer,L. *et al.* (2010) The NCBI biosystems database. *Nucleic Acids Res.*, **38** (**Suppl. 1**), D492–D496.

Hettne,K. *et al.* (2009) A dictionary to identify small molecules and drugs in free text. *Bioinformatics*, **25**, 2983–2991.

Hoffmann,R. and Valencia,A. (2004) A gene network for navigating the literature. *Nat. Genet.*, **36**, 664–664.

Lu,Z. (2011) Pubmed and beyond: a survey of web tools for searching biomedical literature. *Database*, **2011**, baq036.

Ohta,T. *et al.* (2012) Open-domain anatomical entity mention detection. In *Proceedings of DSSD* 2012, ACL’12, pp. 27–36.

Pafilis,E. *et al.* (2009) Reflect: augmented browsing for the life scientist. *Nat. Biotechnol.*, **27**, 508–510.

Rebholz-Schuhmann,D. *et al.* (2008) Text processing through web services: calling Whatizit. *Bioinformatics*, **24**, 296–298.

Sagae,K. and Tsujii,J. (2007) Dependency parsing and domain adaptation with LR models and parser ensembles. In *Proceedings of EMNLP-CoNLL’07 shared task*, Vol. 7, pp. 1044–1050.

Sasaki,Y. *et al.* (2008) BioLexicon: a lexical resource for the biology domain. In *Proceedings of SMBM* 2008, Vol. 3, pp. 109–116.