BIOINFORMATICS APPLICATIONS NOTE
Vol. 24 no. 2 2008, pages 296–298
doi:10.1093/bioinformatics/btm557

Data and text mining

Text processing through Web services: calling Whatizit

Dietrich Rebholz-Schuhmann*, Miguel Arregui, Sylvain Gaudan, Harald Kirsch and Antonio Jimeno

European Bioinformatics Institute, Wellcome Trust Genome Campus, Hinxton, Cambridge, CB10 1SD, UK

Received on July 10, 2007; revised on October 2, 2007; accepted on November 4, 2007

Advance Access publication November 15, 2007

Associate Editor: Jonathan Wren

ABSTRACT

Motivation: Text-mining (TM) solutions are developing into efficient services to researchers in the biomedical research community. Such solutions have to scale with the growing number and size of resources (e.g. available controlled vocabularies), with the amount of literature to be processed (e.g. about 17 million documents in PubMed) and with the demands of the user community (e.g. different methods for fact extraction). These demands motivated the development of a server-based solution for literature analysis.

Whatizit is a suite of modules that analyse text for contained information, e.g. any scientific publication or Medline abstracts. Special modules identify terms and then link them to the corresponding entries in bioinformatics databases such as UniProtKb/Swiss-Prot data entries and gene ontology concepts. Other modules identify a set of selected annotation types like the set produced by the EBIMed analysis pipeline for proteins. In the case of Medline abstracts, Whatizit offers access to EBI's in-house installation via PMID or term query. For large quantities of the user's own text, the server can be operated in a streaming mode (http://www.ebi.ac.uk/webservices/whatizit).

Contact: rebholz@ebi.ac.uk

1 INTRODUCTION

Text-mining (TM) is a complex task and requires integration of various resources (e.g. terminologies) and reuse of specialized technologies (e.g. solutions for the syntactical analysis of sentences). Integration of these resources has led to the development of standalone solutions that are delivered to scientists (Friedman *et al*., 2001) or to complex modular solutions that require installation of components like programming language libraries (Gaizauskas *et al*., 1996). Furthermore, such solutions have an architecture that does not support well integration of bioinformatics services similar to open IT solutions such as Taverna (Hull *et al*., 2006). Other systems such as iHOP provide special interfaces for programmatic access but iHOP offers precompiled data instead of text processing services and thus does not allow to process other documents than Medline abstracts with new means (Hoffmann and Valencia, 2005).

A Web service-based TM solution centralizes and harmonizes crucial tasks and thus solves a number of difficulties reducing maintenance for users. A server based solution can incorporate large terminology sets from biomedical data resources: updates to these resources are efficiently propagated through the server. The end user profits from a harmonized schema including coupling of text processing services to bioinformatics data resources.

Only a limited number of TM solutions have been made available as Web services, e.g. e-Utils at the National Library of Medicine (NLM). Certainly, the limited access to the scientific literature still lowers the benefits from the use of TM Web services. The majority of the publishers refrain from making their content available to the public for automatic analysis and NLM restricts the set of abstracts that can be downloaded at a time. Furthermore, such services would have to integrate different TM components such as a sentenciser, POS tagger, named entity recognizer, acronym resolver, chunker, shallow parser and others. This increases the overhead for making such services available.

The Whatizit Web services provide access to an IT infrastructure that analyses text delivered by the user or retrieved from EBI's Medline installation. The user profits from modules for named entity recognition of selected semantic types or from combinations of such modules. These services satisfy the need for terminology-driven feature extraction from text for document classification and relation extraction. Furthermore, the modules automatically integrate links to database concepts to offer additional support to readers (e.g. like CiteXplore or www.hubmed.org) and to authors (e.g. as part of authoring tools) and will be extended in the future with novel modules for information extraction.

2 IMPLEMENTATION

Whatizit is a modular infrastructure that delivers TM services to the public. Each module processes and annotates text, for example identifies named entities and introduces links to database entries. Individual modules can be composed of a number of internal modules (see Section 3.2). All modules are implemented in Java partly based on special libraries for the matching of large terminology sets (Kirsch *et al*., 2006). All terminologies are based on publicly available resources (e.g. UniProtKb/Swiss-Prot, gene ontology, DrugBank,

*To whom correspondence should be addressed.

see below). Terms are matched to the text taking morphological variability into consideration (Kirsch *et al*., 2006). Part of speech tagging for syntactical analysis is based on the TreeTagger (Schmid, 1994).

All services are preloaded to avoid startup overhead (on a 2 dual core Opteron compute engine, 16 GB main memory). All documents are processed in a single stream right upon submission, i.e. no preprocessing is performed to store intermediary information. Any contained query is passed to the index service (Lucene based) for document retrieval (Hatcher and Gospodnetic, 2004). As soon as a request reaches the central Whatizit server, the requested modules are activated and the stream of text is passed through all required components in a single stream.

3 PROCESSING TEXT INPUT

3.1 Accessing Whatizit SOAP Web services

For access to Whatizit Web services, the calling process has to specify the name of the service (i.e. Whatizit at the EBI) and has to state that the submission represents text only (Unicode encoded) or a list of PMIDS to retrieve Medline abstracts. In addition, the module for the information extraction pipeline has to be specified (e.g. WhatizitGO for gene ontology annotation), which is then applied to the submitted text or to the retrieved Medline abstracts, respectively. The description of each pipeline's functionality can be found in the online documentation of Whatizit.

The following methods are supported from the server process and are specified in the Web services definition language file (WSDL file). First, getPipelineStatus returns a list of available pipelines together with their current status and a description of their task. Second, contact receives the name of a pipeline and the text to be annotated and returns the text with all the annotations contained. Third, queryPmid receives a pipeline name and a list of PubMed IDs. It retrieves all Medline abstracts and returns them annotated. Finally, search requires name of a pipeline and a term query, again retrieves all Medline abstracts and annotates them.

3.2 Available modules

Different types of modules are available through the Whatizit infrastructure. One set of modules annotates named entities. **whatizitChemical** searches for chemical entities based on terminology from ChEBI and the identification of chemical terms by OSCAR3 (Corbett and Murray-Rust, 2006). **whatizitDisease** identifies disease terms using a controlled vocabulary (CV) extracted from MedlinePlus, whereas **whatizitDiseaseUMLS** allows access to MetaMap (Aronson, 2001). For **whatizitDrugs** the CV has been extracted from DrugBank (http://redpoll.pharmacy.ualberta.ca/drugbank/). **whatizitGO** is a pipeline searches for gene ontology terms using exact matching and considering morphological variability (Ashburner *et al*., 2000). Finally, **whatizitOrganism** identifies species names extracted from the NCBI taxonomy (NLM).

Other annotation pipelines represent solutions that are more complex. They identify combinations of semantic types without imposing other restrictions, i.e. identification of relations. **whatizitSwissprotGo** is the pipeline for the annotation of proteins contained in UniProtKb/Swiss-Prot in conjunction with GO annotations (see above whatizitGO). The annotation of proteins is based on the identification of their names in the text considering morphological variability (Kirsch *et al*., 2006). Ambiguous acronyms representing proteins and general English terms are assigned to a protein, if the long form of the acronym is mentioned in the text or on frequency parameters of the term in general English based on the British National Corpus (Rebholz-Schuhmann *et al*., 2006a). For a better understanding of protein name identification refer to BioCreAtIve (Hirschman *et al*., 2005). **whatizitSwissprotPOS** includes the annotation part-of-speech information in addition to the proteins from UniProtKb/Swiss-Prot. **whatizitEbiMed** is the access point to the annotation pipeline from EbiMed (Rebholz-Schuhmann *et al*., 2007). It incorporates whatizitSwiss-protGo, whatizitDrug and whatizitOrganism. **whatizitEbiMedDiseaseChemical** comprises whatizitEbiMed, whatizitDisease and whatizitChemical. The retrieval engine for Medline abstracts is accessible via the module **whatizitQbmarsdf**. For the retrieval, the user has to submit query terms or PubMed IDs.

3.3 Accessing Whatizit through the Web interface

A Web interface has been set up to offer access to Whatizit services without building a client application. On the Web interface, Whatizit provides a text input area where user can submit any kind of Unicode-encoded text or retrieve Medline abstracts for subsequent analysis.

4 DISCUSSION

TM services in the biomedical domain have to cope with large terminological resources (see Introduction section) and should keep up with updates from the primary resource. In the best case, such services are available through a centralized service that scales with the amount of integrated resources, with the demands of different extraction methods and with the amount of literature processed over time. Whatizit is such a service.

In the future, the annotation pipelines will be adapted to a common annotation scheme that would allow better exchange of annotated documents and interoperability of TM components (Rebholz-Schuhmann *et al*., 2006b). In the best case, all providers of such solutions will comply with such standards leading to benefits in the research community due to the free exchange of annotated documents and TM services, all integrated into a pipeline of processing modules (Kirsch *et al*., 2006).

ACKNOWLEDGEMENTS

This work is funded by the Network of Excellence 'Semantic Mining' (NoE 507505). Medline abstracts are provided from the National Library of Medicine (NLM, Bethesda, MD, USA.) Sylvain Gaudan is supported by an 'E-STAR' fellowship (EC's FP6 Marie Curie Host fellowship, MESTCT-2004-504640).

Conflict of Interest: none declared.

REFERENCES

Aronson,A.R. (2001) Effective mapping of biomedical text to the UMLS Metathesaurus: the MetaMap program. *Proc. AMIA Symp.*, **5**, 17–21.

Ashburner,M. *et al.* (2000) Gene ontology: tool for the unification of biology. The gene ontology consortium. *Nat. Genet.*, **25**, 25–9.

Corbett,P. and Murray-Rust,P. (2006) High-throughput identification of chemistry in life science texts. *CompLife, LNBI*, **4216**, 107–118.

Friedman,C. *et al.* (2001) GENIES: a natural-language processing system for the extraction of molecular pathways from journal articles. *Bioinformatics*, **17** (Suppl. 1), S74–S82.

Gaizauskas,R. *et al.* (1996) GATE: an environment to support research and development in natural language engineering. In *Proceedings of the Eighth IEEE International Conference on Tools with Artificial Intelligence*. pp. 58–66, ISBN 0-8186-7686-8.

Hatcher,E. and Gospodnetic,O. (2004) Lucene in Action. Manning, USA.

Hirschman,L. *et al.* (2005) Overview of BioCreAtIvE task 1B: normalized gene lists. *BMC Bioinformatics*, **6** (Suppl. 1), S11.

Hoffmann,R. and Valencia,A. (2005) Implementing the iHOP concept for navigation of biomedical literature. *Bioinformatics*, **21** (Suppl. 2), ii252–ii258.

Hull,D. *et al.* (2006) Taverna: a tool for building and running workflows of services. *Nucleic Acids Res.*, **34**, W729–W732.

Kirsch,H. *et al.* (2006) Distributed modules for text annotation and IE applied to the biomedical domain. *Int. J. Med. Inform.*, **75**, 496–500.

Rebholz-Schuhmann,D. *et al.* (2006a) Annotation and disambiguation of semantic types in biomedical text: a cascaded approach to named entity recognition. In *Workshop on ''Multi-Dimensional Markup in NLP'', EACL 2006*. ACL, USA.

Rebholz-Schuhmann,D. *et al.* (2006b) IeXML: towards an annotation framework for biomedical semantic types enabling interoperability of text processing modules. *SIG BioLink, ISMB 2006*, Fortaleza, Brasil.

Rebholz-Schuhmann,D. *et al.* (2007) EBIMed – text crunching to gather facts for proteins from Medline. *Bioinformatics*, **23**, e237–e244.

Schmid,H. (1994) Probabilistic part-of-speech tagging using decision trees. In *Proceedings of International Conference on New Methods in Language Processing*. Vol. 12, Manchester, UK.