Bioinformatics, 32(12), 2016, 1907–1910
doi: 10.1093/bioinformatics/btv760
Advance Access Publication Date: 16 February 2016
Applications Note

Data and text mining

Beyond accuracy: creating interoperable and scalable text-mining web services

Chih-Hsuan Wei, Robert Leaman and Zhiyong Lu*

National Center for Biotechnology Information (NCBI), National Library of Medicine (NLM), Bethesda, MD 20894, USA

*To whom correspondence should be addressed.

Associate Editor: Jonathan Wren

Received on 25 August 2015; revised on 18 December 2015; accepted on 21 December 2015

Abstract

Summary: The biomedical literature is a knowledge-rich resource and an important foundation for future research. With over 24 million articles in PubMed and an increasing growth rate, research in automated text processing is becoming increasingly important. We report here our recently developed web-based text mining services for biomedical concept recognition and normalization. Unlike most text-mining software tools, our web services integrate several state-of-the-art entity tagging systems (DNorm, GNormPlus, SR4GN, tmChem and tmVar) and offer a batch-processing mode able to process arbitrary text input (e.g. scholarly publications, patents and medical records) in multiple formats (e.g. BioC). We support multiple standards to make our service interoperable and allow simpler integration with other text-processing pipelines. To maximize scalability, we have preprocessed all PubMed articles, and use a computer cluster for processing large requests of arbitrary text.

Availability and implementation: Our text-mining web service is freely available at http://www.ncbi.nlm.nih.gov/CBBresearch/Lu/Demo/tmTools/#curl

Contact: Zhiyong.Lu@nih.gov

1 Introduction

Managing the rapid growth of biomedical knowledge buried in text such as literature, medical records and patents makes the development of automated text-mining technology increasingly important. For instance, the biomedical literature contains the latest reports of scientific discoveries, but is represented in a highly unstructured format. Many text-mining systems have thus been developed in recent years to help unlock this information, both for information retrieval and for novel computational analyses. Text-mining approaches to automatically recognize and extract key biological concepts are of particular interest as this task is often considered to be a building block for many integrated and sophisticated information extraction and retrieval solutions.

Over the years, a number of biomedical named entity recognition (NER) tools have been developed. The entities targeted include genes/proteins (Hakenberg *et al.*, 2011; Tsai *et al.*, 2006; Wei *et al.*, 2015; Wermter *et al.*, 2009), chemical/drug (Leaman *et al.*, 2014; Rocktäschel *et al.*, 2012), disease (Leaman *et al.*, 2013), sequence variation (Caporaso *et al.*, 2007; Doughty *et al.*, 2011; Wei *et al.*, 2013b) and species/taxonomy(Gerner *et al.*, 2010; Wei *et al.*, 2012). To use these tools, and in particular integrate them into existing pipelines, one has to install the software and address many issues including ‘lack of modularity, operating system incompatibility, tool configuration complexity, and lack of standardization of interprocess communications’ (Wiegers *et al.*, 2014). Web-based text mining services provide an alternative solution where the details of the tool are hidden from users and no system installation or maintenance is required. Although one can access text-mining applications like Reflect (Pafilis *et al.*, 2009) and MyMiner (Salgado *et al.*, 2012) through web page visits, we are only aware of a few that offer programmatic web APIs and can therefore be integrated easily: Whatizit (Rebholz-Schuhmann *et al.*, 2008), BeCAS (Nunes *et al.*, 2013), Cocoa (http://npjoint.com/) and Acromine (Okazaki *et al.*, 2010). In comparison to these tools, our web services are unique in several aspects: (i) the entity taggers used offer highly competitive performance in benchmarks for both mention and concept level results, typically via hybrid systems, as opposed to use dictionaries in the previous systems; (ii) for system scalability, our method allows users to submit multiple documents in a single request (instead of one per request) and we process these batch requests using a computer cluster when needed. Moreover, articles in PubMed—the most common target document type—are preprocessed and handled specially so that their tagged results can be instantly retrieved and (iii) for system interoperability, we support multiple formats including BioC, a recently proposed XML format for BioNLP research (Comeau *et al.*, 2014) that complements several other existing platforms such as UIMA (Kano *et al.*, 2009).

2 Materials and Methods

Figure 1 describes the overall architecture of our web services, which use standard HTTP method calls (often known as RESTful services) and allow two access modes: (i) a batch-oriented processing function for any raw text input (abstract, full text, patent, etc), submitted via HTTP POST and (ii) instant retrieval of pre-tagged results of PubMed abstracts via HTTP GET. For the batch-processing function, users may submit one or multiple documents per batch, and large requests will be sent to a computer cluster for parallel processing.

When retrieving pre-tagged results of PubMed abstracts, the request only requires the PMIDs of the requested abstracts. This option is provided because annotating biomedical literature is the most common use case for such a text-mining service. From a technical standpoint, the preprocessing is made possible by our previous system PubTator (Wei *et al.*, 2013a), which stores text-mined annotations for every article in PubMed and keeps in sync with PubMed via nightly updates. We show in Table 1 the five entity types we currently support, along with their associated tagger and respective benchmarking performance (tmChem for chemicals, SR4GN for species, DNorm for diseases, tmVar for mutation/variations, GNormPlus for gene/proteins). Figure 1 shows one example for each access mode. For instance, the disease tagger (DNorm) is being requested to process a text via the RESTful API using our JSON format. Once the request is submitted, our web service responds immediately with a unique session ID, which can be used to check the processing status. Once finished, the user can use the same session ID to retrieve the result, as shown in Figure 1. The text-mining output can be directly visualized using PubTator as shown in Figure 2 where computer-tagged entities are highlighted in various colors throughout the document.

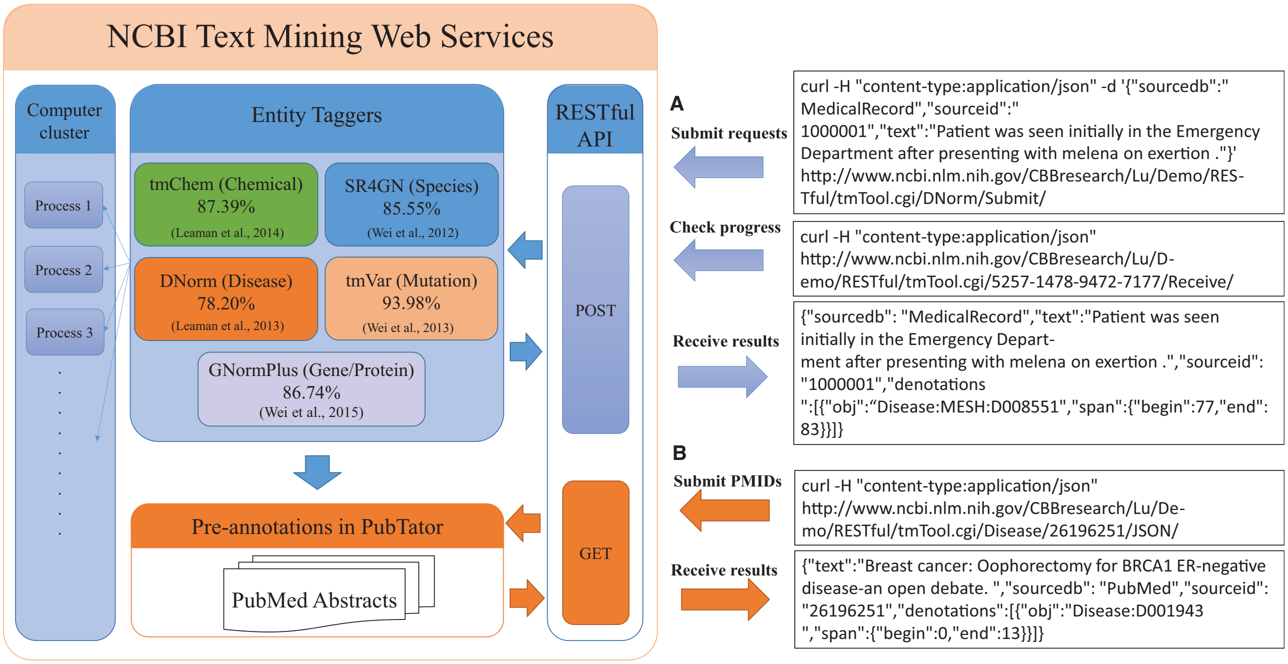
To improve system interoperability, we support multiple formats including BioC/XML (Comeau *et al.*, 2014), PubTator/TXT (Wei *et al.*, 2013a) and PubAnnotation/JSON (Kim and Wang, 2012). By doing so, our service becomes interoperable for different applications. To simplify programmatic access to our web services, we also provide sample client code in Perl, Python and Java.

3 Usage

Since the inception of our web services on March 31, 2015, millions of requests have been made, primarily through the HTTP GET access mode. From interactions with some of our users, we learned that the results of our text-mining services are being used in many different research areas from biocuration, to crowdsourcing, to translational bioinformatics. For instance, our web services are used to provide initial annotations for the mark2cure crowdsourcing project (https://mark2cure.org/) and our gene tagger results are used in assisting the daily curation of HuGE navigator (Yu *et al.*, 2008) a knowledge base for human genome epidemiology.

4 Discussion and Conclusion

We previously developed a number of high performance NER tools and made them open source for public use. In this work, we provide a new way to access these tools in an interoperable and scalable manner, making it simpler to integrate them into complex customized systems. Our format can be converted to new formats like Open Annotation (Pyysalo *et al.*, 2015) via existing converter. Since providing instant access to the tagged results of PubMed abstracts is



```
curl -H "content-type:application/json" -d '{"sourcedb":"
MedicalRecord","sourceid":"
1000001","text":"Patient was seen initially in the Emergency
Department after presenting with melena on exertion ."}'
http://www.ncbi.nlm.nih.gov/CBBresearch/Lu/Demo/RES-
Tful/tmTool.cgi/DNorm/Submit/
```

```
curl -H "content-type:application/json"
http://www.ncbi.nlm.nih.gov/CBBresearch/Lu/D-
emo/RESTful/tmTool.cgi/5257-1478-9472-7177/Receive/
```

```
{"sourcedb": "MedicalRecord","text":"Patient was seen
initially in the Emergency Depart-
ment after presenting with melena on exertion .","sourceid":
"1000001","denotations
":[{"obj":“Disease:MESH:D008551","span":{"begin":77,"end":
83}}]}
```

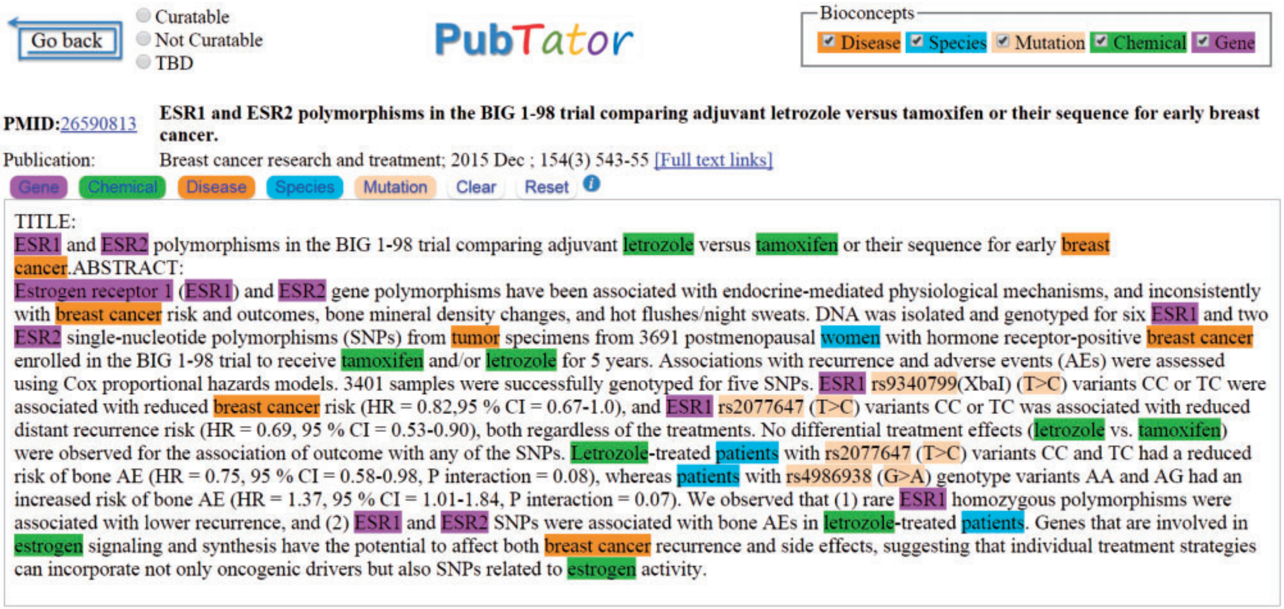
```
curl -H "content-type:application/json"
http://www.ncbi.nlm.nih.gov/CBBresearch/Lu/De-
mo/RESTful/tmTool.cgi/Disease/26196251/JSON/
```

```
{"text":"Breast cancer: Oophorectomy for BRCA1 ER-negative
disease-an open debate. ","sourcedb": "PubMed","sourceid":
"26196251","denotations":[{"obj":"Disease:D001943
","span":{"begin":0,"end":13}}]}
```

Fig. 1. Overview of the NCBI text-mining web services. The overall architecture is provided on the left while system input/output is shown on the right

Table 1. Results of our individual taggers when benchmarked on public test collections

Taggers	Bioconcepts	Evaluation corpus	Precision (%)	Recall (%)	F-score (%)
GNormPlus (Wei et al., 2015)	Gene	BioCreative II–GN corpus (Morgan et al., 2008)	87.08	86.41	86.74
tmChem (Leaman et al., 2014)	Chemical	CHEMDNER corpus (Krallinger et al., 2015)	89.09	85.75	87.39
DNorm (Leaman et al., 2013)	Disease	NCBI Disease corpus (Doğan et al., 2014)	80.30	76.30	78.20
tmVar (Wei et al., 2013a)	Mutation	MutationFinder corpus (Caporaso et al., 2007)	98.80	89.62	93.98
SR4GN (Wei et al., 2012)	Species	Linnaeus corpus (Gerner et al., 2010)	85.82	85.28	85.55



PMID:26590813 **ESR1 and ESR2 polymorphisms in the BIG 1-98 trial comparing adjuvant letrozole versus tamoxifen or their sequence for early breast cancer.**

Publication: Breast cancer research and treatment; 2015 Dec ; 154(3) 543-55 [Full text links]

TITLE:
ESR1 and ESR2 polymorphisms in the BIG 1-98 trial comparing adjuvant letrozole versus tamoxifen or their sequence for early breast cancer.ABSTRACT:
Estrogen receptor 1 (ESR1) and ESR2 gene polymorphisms have been associated with endocrine-mediated physiological mechanisms, and inconsistently with breast cancer risk and outcomes, bone mineral density changes, and hot flushes/night sweats. DNA was isolated and genotyped for six ESR1 and two ESR2 single-nucleotide polymorphisms (SNPs) from tumor specimens from 3691 postmenopausal women with hormone receptor-positive breast cancer enrolled in the BIG 1-98 trial to receive tamoxifen and/or letrozole for 5 years. Associations with recurrence and adverse events (AEs) were assessed using Cox proportional hazards models. 3401 samples were successfully genotyped for five SNPs. ESR1 rs9340799(XbaI) (T>C) variants CC or TC were associated with reduced breast cancer risk (HR = 0.82,95 % CI = 0.67-1.0), and ESR1 rs2077647 (T>C) variants CC or TC was associated with reduced distant recurrence risk (HR = 0.69, 95 % CI = 0.53-0.90), both regardless of the treatments. No differential treatment effects (letrozole vs. tamoxifen) were observed for the association of outcome with any of the SNPs. Letrozole-treated patients with rs2077647 (T>C) variants CC and TC had a reduced risk of bone AE (HR = 0.75, 95 % CI = 0.58-0.98, P interaction = 0.08), whereas patients with rs4986938 (G>A) genotype variants AA and AG had an increased risk of bone AE (HR = 1.37, 95 % CI = 1.01-1.84, P interaction = 0.07). We observed that (1) rare ESR1 homozygous polymorphisms were associated with lower recurrence, and (2) ESR1 and ESR2 SNPs were associated with bone AEs in letrozole-treated patients. Genes that are involved in estrogen signaling and synthesis have the potential to affect both breast cancer recurrence and side effects, suggesting that individual treatment strategies can incorporate not only oncogenic drivers but also SNPs related to estrogen activity.

Fig. 2. The results of our RESTful API can be readily visualized in PubTator (Color version of this figure is available at *Bioinformatics* online.)

an extremely useful feature of the current system, we plan to include preprocessed results of PMC full text articles in the future.

Acknowledgement

This research was supported by the NIH Intramural Research Program, National Library of Medicine.

Conflict of Interest: none declared.

References

Caporaso,J.G. *et al.* (2007) MutationFinder: a high-performance system for extracting point mutation mentions from text. *Bioinformatics*, **23**, 1862–1865.

Comeau,D.C. *et al.* (2014) BioC interoperability track overview. *Database*, **2014**, bau053.

Doğan,R.I. *et al.* (2014) NCBI disease corpus: a resource for disease name recognition and concept normalization. *J. Biomed. Inform.*, **47**, 1–10.

Doughty,E. *et al.* (2011) Toward an automatic method for extracting cancer- and other disease-related point mutations from the biomedical literature. *Bioinformatics*, **27**, 408–415.

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

Hakenberg,J. *et al.* (2011) The GNAT library for local and remote gene mention normalization. *Bioinformatics*, **27**, 2769–2771.

Kano,Y. *et al.* (2009) U-Compare: share and compare text mining tools with UIMA. *Bioinformatics*, **25**, 1997–1998.

Kim,J.D. and Wang,Y. (2012) PubAnnotation: a persistent and sharable corpus and annotation repository. In: *Proceedings of the 2012 Workshop on Biomedical Natural Language Processing*. Association for Computational Linguistics, Montreal, Canada, pp. 202–205.

Krallinger,M. *et al.* (2015) The CHEMDNER corpus of chemicals and drugs and its annotation principles. *J. Cheminform.*, **7**, S2.

Leaman,R. *et al.* (2013) DNorm: disease name normalization with pairwise learning to rank. *Bioinformatics*, **29**, 2909–2917.

Leaman,R. *et al.* (2014) tmChem: a high performance approach for chemical named entity recognition and normalization. *J. Cheminform.*, (**Suppl. 1**), S3.

Morgan,A.A. *et al.* (2008) Overview of BioCreative II gene normalization. *Genome Biol.*, **9**, S3.

Nunes,T. *et al.* (2013) BeCAS: biomedical concept recognition services and visualization. *Bioinformatics*, **29**, 1915–1916.

Okazaki,N. *et al.* (2010) Building a high-quality sense inventory for improved abbreviation disambiguation. *Bioinformatics*, **26**, 1246–1253.

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

Pyysalo,S. *et al.* (2015) Sharing annotations better: RESTful open annotation. In: *Proceedings of ACL-IJCNLP 2015 System Demonstrations*. Association for Computational Linguistics, Beijing, China, pp. 91–96.

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

Rocktäschel,T. *et al.* (2012) ChemSpot: a hybrid system for chemical named entity recognition. *Bioinformatics*, **28**, 1633–1640.

Salgado,D. *et al.* (2012) MyMiner: a web application for computer-assisted biocuration and text annotation. *Bioinformatics*, **28**, 2285–2287.

Tsai,R.T.H. *et al.* (2006) NERBio: using selected word conjunctions, term normalization, and global patterns to improve biome dical named entity recognition. *BMC Bioinformatics*, **7**, S11.

Wei,C.H. *et al.* (2012) SR4GN: a species recognition software tool for gene normalization. *PLoS One*, **7**, e38460.

Wei,C.H. *et al*. (2013a) PubTator: a web-based text mining tool for assisting biocuration. *Nucleic Acids Res*., W518–W522.

Wei,C.H. *et al*. (2013b) tmVar: a text mining approach for extracting sequence variants in biomedical literature. *Bioinformatics*, **29**, 1433–1439.

Wei,C.H. *et al*. (2015) GNormPlus: an integrative approach for tagging genes, gene families, and protein domains. *BioMed. Res. Int*., 918710.

Wermter,J. *et al*. (2009) High-performance gene name normalization with GeNo. *Bioinformatics*, **25**, 815–821.

Wiegers,T.C. *et al*. (2014) Web services-based text-mining demonstrates broad impacts for interoperability and process simplification. *Database*, **2014**, bau050.

Yu,W. *et al*. (2008) A navigator for human genome epidemiology. *Nat. Genet*., **40**, 124–125.