# Large-scale Semantic Indexing with Biomedical Ontologies

**Chih-Hsuan Wei**
National Center for Biotechnology Information (NCBI), National Library of Medicine (NLM)
Bethesda, Maryland, USA
chih-hsuan.wei@nih.gov

**Robert Leaman**
National Center for Biotechnology Information (NCBI), National Library of Medicine (NLM)
Bethesda, Maryland, USA
robert.leaman@nih.gov

**Zhiyong Lu**
National Center for Biotechnology Information (NCBI), National Library of Medicine (NLM)
Bethesda, Maryland, USA
zhiyong.lu@nih.gov

***Abstract*— We introduce PubTator, a web-based application that enables large-scale semantic indexing and automatic concept recognition in biomedical ontologies. Not only was PubTator formally evaluated and top-rated in BioCreative, it also has been widely adopted and used by the scientific community from around the world, supporting both research projects and real-world applications in biocuration, crowdsourcing and translational bioinformatics.**

***Keywords—PubTator; TaggerOne; Text Mining; Biomedical Ontologies***

## I. INTRODUCTION

With over 26 million articles in PubMed, the biomedical literature is a knowledge-rich resource and forms an important foundation for future research. However, the rapid expansion of the scientific literature and the increasingly cross-disciplinary nature of biomedical research are making it difficult than ever for individual researchers to find and assimilate all of the relevant information from the literature. Research in automated text processing is of a growing importance to relieve today's information overload problem. Hence, processing the biomedical literature with automated tools becomes more important as its growth accelerates.

We present PubTator [1], a web-based application that indexes the ever-growing biomedical literature with ontological concepts in biomedicine. PubTator features a PubMed-like interface and is equipped with multiple high-performing text mining algorithms (e.g. DNorm for disease concepts in MeSH or SNOMED-CT) to ensure the quality of its text-mined results over the entire set of articles in PubMed. PubTator was first developed as an interactive text mining system through our participation in BioCreative (see [2] for more details and related work). More recently, we created RESTful Web Services [3] for PubTator to further increase its scalability and ease its use by non-experts of text mining, allowing its users to focus on results rather than technical methodology.

## II. SYSTEM DESCRIPTION

### A. *Concept Recognition using PubTator*

PubTator currently utilizes five state-of-the-art named entity recognition and normalization tools to locate and identify important biomedical entities. Specifically, the entity types currently supported and their respective systems with F-scores are: genes and proteins (GNormPlus [4] - 86.74%), diseases (DNorm [5] - 80.90%), chemicals (tmChem [6] - 87.51%), species (SR4GN [7] - 85.42%) and genetic variants (tmVar [8] - 91.39%).

While the entity types currently covered includes those most commonly searched [9], our most recent work, TaggerOne, is trainable to identify arbitrary entity types, requiring only annotated training data and a corresponding lexicon [10]. TaggerOne employs a novel machine learning model to address named entity recognition and normalization jointly, reducing cascading errors and enabling the NER (name entity recognition) task to directly exploit the lexical information provided by the normalization. TaggerOne achieves state of the art performance on diseases (NCBI Disease corpus [11]) and chemicals (BioCreative 5 CDR corpus [12]) and is being used to tag anatomy terms (including organs, tissues, cellular components) in PubMed articles so they can be mapped to the corresponding concept identifiers in multiple biomedical ontologies in http://www.obofoundry.org/.

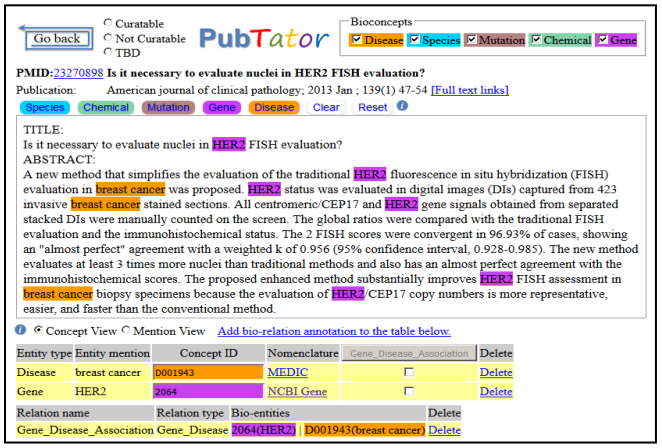


Figure 1. The screenshot of a PubMed article in PubTator with concepts and relations highlighted in color.

Recognizing ontological concepts requires the creation of a lexical resource identifying the concepts desired, their terms and relevant variations. We recently proposed a modification of TaggerOne to automatically identify inconsistencies that arise when creating a single lexical resource from multiple knowledge resources, including ontologies, and then address the inconsistency semi-automatically. The proposed method actively learns a model to identify identical concepts from separate resources, with preliminary results showing the model successfully identifies both synonymous tokens (e.g. "kidney" and "renal") and contrastive terms ("dominant" vs. "recessive").

### B. *Scalability and interoperability*

Large scale use of PubTator or open-source tools requires a significant investment in infrastructure and maintenance time. These barriers to entry reduce the ability of individual researchers to explore applying text mining to problems in their research area and consequently impair the continued adoption of text mining tools. Web services provide on-demand access to software tools through the Internet using straightforward interfaces and data formats. Providing text mining tools as web services therefore lowers the bar to use for end users and bioinformatics researchers not working specifically in text mining, allowing free exploration and the ability to focus on results rather than methodology.

Therefore, we recently developed NCBI text-mining web services on top of PubTator by using standard HTTP method calls (often known as RESTful services), which allows instant retrieval of pre-annotated PubTator results via HTTP GET. To improve system interoperability, we support multiple data formats including BioC/XML [13], PubTator/TXT [1] and PubAnnotation/JSON [14]. To simplify programmatic access to our web services, we also provide sample client code in Perl, Python and Java.

### C. *Evaluation & Usage*

PubTator was formally assessed by a group of external evaluators during the BioCreative Interactive Text Mining challenge task where it was top-rated in all categories from system design to learnability to usability [15].

More recently, through collaboration with curation groups, PubTator has been successfully integrated into the production pipeline of multiple curation databases including SwissProt [16] and the CDC's human genome epidemiology knowledge base called HuGE navigator [17].

Furthermore, since the inception of PubTator Web Services, millions of requests have been made by the scientific community from around the world. 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 in bioinformatics. For instance, our web services are used to provide initial annotations for the mark2cure crowdsourcing project (https://mark2cure.org/).

## III. Conclusions & Future Work

In the future, we plan to expand the automatic concept recognition to additional biomedical ontologies and include their results in PubTator. Text mining open-access full-length articles in PMC for key ontological concepts in real-world applications (e.g. computer-assisted biocuration) would be another exciting opportunity to pursue.

## Acknowledgment

This research is supported by the Intramural Research Program of the National Institutes of Health, National Library of Medicine.

## References

[1] C.-H. Wei, H.-Y. Kao, and Z. Lu, "PubTator: a web-based text mining tool for assisting biocuration," Nucleic Acids Research, vol. 41, 2013, pp. W518-W522.

[2] C. N. Arighi, B. Carterette, K. B. Cohen, M. Krallinger, W. J. Wilbur, P. Fey, et al., "An overview of the BioCreative 2012 Workshop Track III: interactive text mining task," Database, 2013, pp. bas056.

[3] C.-H. Wei, R. Leaman, and Z. Lu, "Beyond accuracy: Creating interoperable and scalable text mining web services," Bioinformatics, 2016.

[4] C.-H. Wei, H.-Y. Kao, and Z. Lu, "GNormPlus: An Integrative Approach for Tagging Genes, Gene Families, and Protein Domains," BioMed Research International, 2015, pp. 918710.

[5] R. Leaman, R. Islamaj Doğan, and Z. Lu, "DNorm: Disease name normalization with pairwise learning-to-rank," Bioinformatics, vol. 29, 2013, pp. 2909-2917.

[6] R. Leaman, C.-H. Wei, and Z. Lu, "tmChem: a high performance approach for chemical named entity recognition and normalization," Journal of Cheminformatics, vol. 7, 2015, pp. S3.

[7] C.-H. Wei, H.-Y. Kao, and Z. Lu, "SR4GN: a species recognition software tool for gene normalization," PLoS One, vol. 7, 2012, pp. e38460.

[8] C.-H. Wei, B. R. Harris, H.-Y. Kao, and Z. Lu, "tmVar: A text mining approach for extracting sequence variants in biomedical literature," Bioinformatics, vol. 29, 2013, pp. 1433-1439.

[9] R. I. Dogan, G. C. Murray, A. Névéol, and Z. Lu, "Understanding PubMed user search behavior through log analysis," Database, 2009, pp. bap018.

[10] R. Leaman and Z. Lu, "TaggerOne: Joint Named Entity Recognition and Normalization with Semi-Markov Models," Bioinformatics, vol. In Press, 2016.

[11] R. I. Doğana, R. Leaman, and Z. Lu, "NCBI disease corpus: A resource for disease name recognition and concept normalization," Journal of Biomedical Informatics, vol. 47, 2014, pp. 1-10.

[12] J. Li, Y. Sun, R. J. Johnson, D. Sciaky, C.-H. W. R. Leaman, A. P. Davis, et al., "BioCreative V CDR task corpus: a resource for chemical disease relation extraction," Database, 2016, pp. baw068.

[13] D. C. Comeau, R. I. Doğan, P. Ciccarese, K. B. Cohen, M. Krallinger, F. Leitner, et al., "BioC: a minimalist approach to interoperability for biomedical text processing," Database, 2013, pp. bat064.

[14] J.-D. Kim, K. B. Cohen, and J.-J. Kim, "PubAnnotation-query: a search tool for corpora with multi-layers of annotation," BMC Proceedings, vol. 9, 2015, pp. A3.

[15] C.-H. Wei, B. R. Harris, D. Li, T. Z. Berardini, E. Huala, H.-Y. Kao, et al., "Accelerating literature curation with text-mining tools: a case study of using PubTator to curate genes in PubMed abstracts," Database, 2012, pp. bas041.

[16] The UniProt Consortium, "UniProt: a hub for protein information," Nucleic Acids Research, vol. 43, 2015, pp. D204-D212.

[17] W. Yu, M. Gwinn, M. Clyne, A. Yesupriya, and M. J. Khoury, "A navigator for human genome epidemiology," Nature Genetics, vol. 40, 2008, pp. 124-125.