# Literature Review on Information Extraction for Biomedical NLP

Amitha Supragna Sandur

MS Bioinformatics, Dept. of Computer Science, UIUC

NetId: asandur2

## Introduction

This review will focus on brief overview of the state of the art in biomedical text mining across different biological, medical and clinical domains, in terms of methods of extraction of information in general. It will also provide an overview of the frontiers of biomedical text mining and lab groups conducting research in the area of Biomedical Information Extraction.

Electronic Medical Health Records (EMHR) data and all other kinds of biomedical literature data are generated in huge amounts which carry important information of patients such as demographics, medications, laboratory tests, diagnosis codes and procedures[1]. But the research to solve the problems in this area are not as extensively done as it is in other fields such as Computer Vision, etc. Some common problems include reproducibility, very few free datasets availability. So the goal of this review is to look at recent methodological advances for text mining in Biomedical NLP.

## Body

Papers published in biomedical journals contain important information for biologists, pharmaceutical industries. Figure 1 shows the amount of biomedical literature published over these years[3]. So there is a need for automated tools to read and extract knowledge from these papers. Donaldson et al. paper used Support Vector Machine for mining protein-protein interactions. They built PreBIND and Textomy which identify articles with biomolecular interactions [2]. The problem that arises in mining these kinds of biological interactions is that there are no good datasets. And if there is some dataset available to train on, if we use the model to test on another dataset, it will result in improper numbers and results. To give an example in real life, if we give a couple of sentences to different biologists and ask them to annotate it, they will give back different annotations. So the fields of open information extraction, deep learning are a part of this review as well.

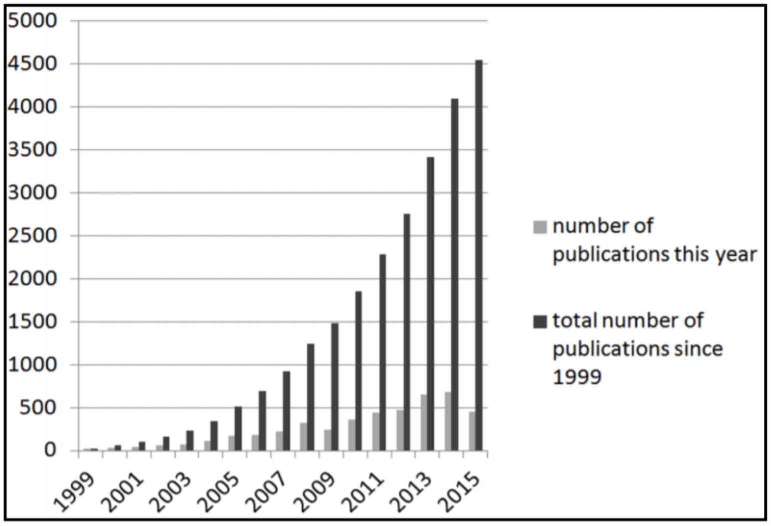

Figure 1. The number of publications in biomedical information extraction each year (grey bar) and cumulative number since 1999 (dark bar), using the PubMed query of "(biomedical OR clinical OR medical) AND (text+mining OR text+processing OR natural+language+processing) AND (information+extraction OR named+entity+detection OR named+entity+recognition OR relation+extraction OR event+extraction)".

Some groups or research institutes working actively in this field are: BioNLP group at NCBI, Boston Children's Hospital Natural Language Processing Laboratory, National Centre for Text Mining (NaCTeM), Mayo Clinic's clinical natural language processing program, UTHealth Houston Biomedical Natural Language Processing Lab, Columbia University Department of Biomedical Informatics and many others.

Frontiers of biomedical text mining [5] are:

(1) Pierre Zweigenbaum from the Language, Information and Representation Group, Computer Sciences Laboratory for Mechanics and Engineering Sciences (LIMSI), National Center for Scientific Research (CNRS). His work is focused on NLP in the biological domain
(2) Dina Demner-Fushman from the Communications Engineering Branch, Lister Hill National Center for Biomedical Communications, U.S. National Library of Medicine. Her work is in the area of clinical question answering
(3) Hong Yu from University of Wisconsin-Milwaukee, Departments of Computer Science and Health Sciences. Her interests are discourse analysis and question answering
(4) Kevin B. Cohen from University of Colorado School of Medicine's Center for Computational Pharmacology

Brief overview of the methodologies used in this field are: rule-based method, knowledge-based method, statistics-based method, learning-based method and hybrid method. Learning methods for biological information extraction include conditional random fields, structured support vector machines and deep neural networks [3]. They all fall into the either category of supervised or unsupervised learning.

Conclusion

Data driven approaches will continue to be the main technique used in text mining of biological data. Open Information Extraction method has been known to work well on the heterogenous biomedical data. Deep Learning has also shown to have important applications in biomedical information extraction, although it is predicted that combination of biological experimental knowledge along with deep learning techniques can lead to greater accuracy of results [3].

The automated annotated biomedical text data is needed to contribute towards advances in the field of Biomedical Information Extraction. The systems that fulfil this need has applications in concrete clinical and biomedical research contexts [4].

References

[1] Yadav, P., Steinbach, M., Kumar, V., & Simon, G. (2018). Mining Electronic Health Records (EHRs) A Survey. ACM Computing Surveys (CSUR), 50(6), 1-40.

[2] Donaldson, I., Martin, J., De Bruijn, B., Wolting, C., Lay, V., Tuekam, B., ... & Pawson, T. (2003). PreBIND and Textomy–mining the biomedical literature for protein-protein interactions using a support vector machine. BMC bioinformatics, 4(1), 1-13.

[3] Liu, F., Chen, J., Jagannatha, A., & Yu, H. (2016). Learning for biomedical information extraction: Methodological review of recent advances. arXiv preprint arXiv:1606.07993.

[4] Meystre, S. M., Savova, G. K., Kipper-Schuler, K. C., & Hurdle, J. F. (2008). Extracting information from textual documents in the electronic health record: a review of recent research. Yearbook of medical informatics, 17(01), 128-144.

[5] Zweigenbaum, P., Demner-Fushman, D., Yu, H., & Cohen, K. B. (2007). Frontiers of biomedical text mining: current progress. Briefings in bioinformatics, 8(5), 358-375.