MEETING PROPOSAL

Building a cost-effective gold standard set for enriching PubAnnotation

Dongseop Kwon[1], Chih-Hsuan Wei[2], Sun Kim[2], Robert Leaman[2] and Zhiyong Lu[2*]

[*]Correspondence:
zhiyong.lu@nih.gov
[2]National Center for Biotechnology Information, National Library of Medicine, National Institutes of Health, Bethesda, MD 20894, USA

1 Motivation

While manually annotated corpora are of great importance for developing text mining systems, annotating literature is a tedious and time-consuming job for humans [1]. More importantly, current manual curation processes require too much time to complete, making it impossible to keep up with the rapid growth of biomedical literature [2]. This problem led to the development of many tools to assist humans in annotating text [3], however the main focus has been customizing annotation tools and text mining modules for specific tasks, i.e. a tool developed for annotating genes may not be compatible with identifying chemical names.

TaggerOne [4] is a machine learning model for jointly learning and predicting named entities and their normalized concepts. It was reported that TaggerOne achieved the state-of-the-art performance on diseases and chemicals, but its usage is not only limited to a few bio-entity types because it is a general learning framework for any named entities and concepts. Here, our main questions are 1) can TaggerOne (or another general-purpose NER tool) be used to assist in producing a gold standard corpus of annotated mentions for arbitrary bio-entities? 2) how many training examples are enough to build a TaggerOne model from scratch? 3) how can the output from TaggerOne be linked to PubAnnotation?

2 Approach

To address the issues described above, we propose an active learning strategy using human intervention. First, entities in a corpus are tagged using a lexicon. After this string match process, curators act as an oracle by fixing mistakes. This interactive annotation process may include selecting features as well as assigning labels [5], but in our proposal, we only correct errors made by the automatic prediction process. The next step is to train TaggerOne using the semi-supervised annotation set. This may not produce a reasonable performance in the first iteration, however the performance will improve after a few rounds of machine learning and manual annotation processes. We expect this would reduce the amount of work that curators should spend compared to full manual annotation.

Our main idea is to create a semi gold-standard set that enriches existing annotated corpora for text mining systems, but also to easily produce a high quality annotation set enough to be used for biological databases. In the hackathon, we plan to demonstrate the interactive learning process using TaggerOne and show how one can obtain a gold-standard set in a timely manner. This includes discussing what annotation types are more useful and performing a real annotation task. Since the interactive annotation process saves an annotated set in BioC [6], this resulting set

can be imported to PubAnnotation [7] via a conversion tool. Therefore, another step that should follow is to implement the conversion module, thus no extra import/export step is required for uploading annotation results to PubAnnoation.

A one-size-fits-all annotation tool was discussed in #BLAHmuc [8], however the goal was to design a common architecture or a protocol that enables interoperability. Our interest, on the other hand, is a general-purpose named entity recognition module and its interactive learning process. The proposed approach may provide insight into of how one could bridge manual and automatic annotation processes more efficiently and effectively.

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

Author details
[1]Department of Computer Engineering, Myongji University, Yongin, Gyeonggi-do 17058, South Korea. [2]National Center for Biotechnology Information, National Library of Medicine, National Institutes of Health, Bethesda, MD 20894, USA.

References

1. Verspoor, K., Cohen, K.B., Lanfranchi, A., Warner, C., Johnson, H.L., Roeder, C., Choi, J.D., Funk, C., Malenkiy, Y., Eckert, M., Xue, N., Baumgartner, W.A., Bada, M., Palmer, M., Hunter, L.E.: A corpus of full-text journal articles is a robust evaluation tool for revealing differences in performance of biomedical natural language processing tools. BMC Bioinformatics **13**(1), 207 (2012)
2. Baumgartner, W.A., Cohen, K.B., Fox, L.M., Acquaah-Mensah, G., Hunter, L.: Manual curation is not sufficient for annotation of genomic databases. Bioinformatics **23**(13), 41–48 (2007)
3. Neves, M., Leser, U.: A survey on annotation tools for the biomedical literature. Briefings in Bioinformatics **15**(2), 327–340 (2014)
4. Leaman, R., Lu, Z.: TaggerOne: joint named entity recognition and normalization with semi-Markov Models. Bioinformatics **32**(18), 2839–2846 (2016)
5. Settles, B.: Closing the loop: fast, interactive semi-supervised annotation with queries on features and instances. In: Proceedings of the Conference on Empirical Methods in Natural Language Processing, pp. 1467–1478 (2011)
6. Comeau, D.C., Islamaj Doğan, R., Ciccarese, P., Cohen, K.B., Krallinger, M., Leitner, F., Lu, Z., Peng, Y., Rinaldi, F., Torii, M., Valencia, A., Verspoor, K., Wiegers, T.C., Wu, C.H., Wilbur, W.J.: BioC: a minimalist approach to interoperability for biomedical text processing. Database **2013**, 064 (2013)
7. Kim, J.-D., Wang, Y.: PubAnnotation: a persistent and sharable corpus and annotation repository. In: Proceedings of the 2012 Workshop on Biomedical Natural Language Processing, pp. 202–205 (2012)
8. Cohen, K.B., Demner-Fushman, D., Fort, K., Grouin, C., Hunter, L.E., Leser, U., Névéol, A., Neves, M., Zweigenbaum, P.: Towards the last annotation tool. In: #BLAHmuc (2016)