BeCalm and keep BLAHing

Lars Juhl Jensen[1]*

[1] Disease Systems Biology Program, Novo Nordisk Foundation Center for Protein Research, Faculty of Health and Medical Sciences, University of Copenhagen, Copenhagen, Denmark.

* Contact lars.juhl.jensen@cpr.ku.dk

BioCreative V.5 will feature a new task in which technical aspects of text-mining servers. To participate, the servers must support the new BeCalm API. I propose the development a BeCalm layer that builds on top of PubAnnotation and/or Open Annotation at BLAH3, which will enable databases that already implement one of these to participate in the upcoming BioCreative challenge.

Introduction

The BioCreative challenges (http://biocreative.org/) have played a key role in pushing forward the development of better biomedical text-mining software, in particular related to named entity recognition (NER) of genes/proteins and other biomedical concepts. Whereas first BioCreative challenges focused purely on evaluating performance metrics such as precision and recall, the tasks have since diversified to evaluate also other important aspects of biomedical text-mining software, such as its utility and usability.

The upcoming BioCreative V.5 challenge features a new task, the Technical Interoperability and Performance of annotation Servers (TIPS) task, which aims to evaluate technical aspects such as stability and response time as well as data formats and metadata. To participate in the task, participants must support the BeCalm API (http://www.becalm.eu/api). The specification for this new API has only very recently been released, and the BeCalm online servers have not gone live at the time of writing, making it difficult to start implementing support for it. However, the online servers should be available before end of 2016. In light of this, and the deadline for TIPS evaluations already end of January 2017, participants will thus have a very short time window for implementing robust support for the API. The timing of BLAH3 is ideal for a joint effort on coding BeCalm API support.

Despite being intended as an API for running NER tools, the BeCalm API does not facilitate the submission of arbitrary text. Rather, it allows users to request NER results from three sources of text, namely PubMed, PubMed Central (PMC0, and patents (http://www.becalm.eu/api). This, combined with the aims of speed and robustness in the TIPS task, makes hosting precomputed NER results behind a BeCalm API an attractive, if unintended, alternative to real-time tagging. Rather than coding this for each individual resource, I propose to construct a thin BeCalm layer that operates on top of the PubAnnotation (1) and/or RESTful Open Annotation (2) interfaces already implemented for a number of resources. This would enable the developers to easily support the API required for participation in the TIPS task (i.e. BeCalm) while keeping the focus on implementing support for biomedical linked annotation (i.e. keep BLAHing).

Implementation

As mentioned, it is currently impossible to develop — or at least to test — an implementation of the BeCalm API. At present no implementation thus currently exists. I see three alternative, but not mutually exclusive, options for how to provide a BeCalm API for existing databases/corpora of precomputed NER results.

PubAnnotation. The first option is to import all the results into PubAnnotation and provide the API on top of this. This is that it would result in a single, self-contained resource, but would also rely entirely on this resource scaling to handle large corpora such as the entire PubMed and entire open-access subset of PMC. It would also have to be able to continuously accept updates to such corpora as more documents are added.

Open Annotation hub. The second option is to build a single proxy that provides a BeCalm API for a number of existing databases that support Open Annotation. Like the first option, this has the advantage of requiring only a single resource to be built. However, it differs in that the server will not be self-contained (i.e. it queries other servers for the actual NER results).

Distributed Open Annotation layer. The third option also implies writing a proxy that bridges the BeCalm API and Open Annotation. However, the big difference is that each site would run their own proxy. This requires the code to be easy to install and configure at each site. However, it has the advantage that there is no single point of failure and that each site is solely responsible for their availability and performance.

Conclusions

Depending on the implementation chosen, this project would enable BLAH3 participants to participate in the BioCreative V.5 TIPS task, either individually or as a jointly as a single team. The TIPS deadline shortly after BLAH3 would strongly encourage finishing the work at BLAH3.

Acknowledgments

This work was funded by the Novo Nordisk Foundation (NNF14CC0001).

References

1. Kim JD and Wang Y (2012) PubAnnotation: a persistent and sharable corpus and annotation repository. *Proceedings of BioNLP 2012*, 202–205.
2. Pyysalo S, Campos J, Cejuela JM, *et al.* (2015). Sharing annotations better: RESTful Open Annotation. *Proceedings of ACL-IJCNLP 2015*, 91–96.