BMC Bioinformatics

ORAL PRESENTATION Open Access

Event extraction on PubMed scale

Filip Ginter[1*], Jari Björne[1,2], Sampo Pyysalo[3]

From Workshop on Advances in Bio Text Mining
Ghent, Belgium. 10-11 May 2010

There has been a growing interest in typed, recursively nested events as the target for information extraction in the biomedical domain. The BioNLP'09 Shared Task on Event Extraction [1] provided a standard definition of events and established the current state-of-the-art in event extraction through competitive evaluation on a standard dataset derived from the GENIA event corpus.

We have previously established the scalability of event extraction to large corpora [2] and here we present a follow-up study in which event extraction is performed from the titles and abstracts of all 17.8M citations in the 2009 release of PubMed. The extraction pipeline is composed of state-of-the-art methods: the BANNER named entity recognizer [3], the McClosky-Charniak domain-adapted parser [4], and the Turku Event Extraction System [5], the winning entry of the Shared Task.

The resulting dataset consists of over 19.2M instances of 4.5M unique events, of which 2.1M instances of 1.6M unique events recursively involve at least two different named entities. This dataset is several orders of magnitude larger than any previous event extraction effort and – having been obtained by a demonstrably state-of-the-art pipeline — represents the most accurate event extraction output achievable with presently available tools. Compiling the dataset was a technically challenging undertaking and required roughly 8,300 CPU-hours.

As the primary contribution of the study, we make the entire set of extracted events freely available at http://bionlp.utu.fi, together with the output of the individual stages of the pipeline, such as 36.5M named entity instances and syntactic analyzes for all 20M sentences containing at least one named entity. This resource will facilitate future research related to biological event networks by providing a standard, publicly available, large-scale dataset, avoiding the unnecessary duplication of efforts in executing the complex event extraction pipeline.

* Correspondence: ginter@cs.utu.fi
[1]Department of Information Technology, University of Turku, Finland
Full list of author information is available at the end of the article

Author details
[1]Department of Information Technology, University of Turku, Finland. [2]Turku Centre for Computer Science (TUCS), Finland. [3]Department of Computer Science, University of Tokyo, Japan.

Published: 6 October 2010

References

1. Kim JD, Ohta T, Pyysalo S, Kano Y, Tsujii J: **Overview of BioNLP'09 Shared Task on Event Extraction.** *In Proceedings of the BioNLP 2009 Workshop Companion Volume for Shared Task ACL* 2009, 1-9.
2. Björne J, Ginter F, Pyysalo S, Tsujii J, Salakoski T: **Complex Event Extraction at PubMed Scale.** *Proceedings of ISMB'10* 2010, **26(12)**:i382-i390.
3. Leaman R, Gonzalez G: **BANNER: an executable survey of advances in biomedical named entity recognition.** *Proceedings of Pacific Symposium on Biocomputing* 2008, 652-663.
4. McClosky D: **Any Domain Parsing: Automatic Domain Adaptation for Natural Language Parsing.** *PhD thesis* Department of Computer Science, Brown University 2009.
5. Björne J, Heimonen J, Ginter F, Airola A, Pahikkala T, Salakoski T: **Extracting Contextualized Complex Biological Events with Rich Graph-Based Feature Sets.** *Computational Intelligence* 2010.

doi:10.1186/1471-2105-11-S5-O2
Cite this article as: Ginter *et al.*: Event extraction on PubMed scale. *BMC Bioinformatics* 2010 **11**(Suppl 5):O2.