JOURNAL OF BIOMEDICAL SEMANTICS

INTRODUCTION Open Access

Towards mature use of semantic resources for biomedical analyses

D Rebholz-Schuhmann[1*], F Rinaldi[2], S Pyysalo[3], N Collier[4], U Hahn[5]

From Fourth International Symposium on Semantic Mining in Biomedicine (SMBM) Hinxton, UK. 25-26 October 2010

[1]EMBL Outstation, European Bioinformatics Institute, Hinxton, Cambridge, CB10 1SD, UK

Use of semantic resources, such as ontologies, has been the key to the standardisation of IT solutions and their interoperability in recent years. Combining semantic resources with biomedical data analysis is developing into a dedicated research domain and stimulates related research in biomedical research fields.

The growing attention to semantics-driven analyses motivates full-fledged conferences dedicated to the latest scientific results in this domain. The conference series entitled *Symposium on Semantic Mining in Biomedicine* (*SMBM*) constitutes a platform for activities of researchers showing interest in text and data mining in biomedicine, medical, bio- and chemo-informatics, and researchers from biomedical ontology design and engineering. The prospects of these meetings are the exchange of shared resources, the integration of existing resources and training in novel approaches and solutions.

The variety of scientific events focused on the generation of semantic resources and on the exploitation of semantic resources, including the biomedical literature and the Semantic Web, is testimony of the attention attributed to the field. Additionally, in recent years the progress in the use and analysis of semantic resources has been kept under constant assessment by demanding competitive evaluations such as BioCreative, CALBC and the BioNLP shared tasks [1-3]. This has led to the standardisation of semantic resources and their use for large-scale document processing [4].

Challenges remain, as demonstrated from the presented research work at the SMBM. Such challenges are the generation of consistent semantic resources, their formalisation and exploitation for biomedical analyses. Other challenges result from analyses of the whole scientific literature requiring the identification of relations and events at best in a large-scale approach and the identification of relevant contextual information, such as speculation and negation. The research work presented at the SMBM 2010 was addressing some of these challenges.

The transformation of the textual representation in the GENIA corpus into an axiomatic OWL representation was achieved by [5,6]. This solution improves the disambiguation of the names of genes and proteins in the GENIA corpus. The formal representation also enables consistency analysis across the corpus, as well as the development of verifiable annotation guidelines and the automatic discovery of new facts through deductive inferences.

Event extraction was addressed by [7] for the identification of DNA methylation events and by [8] for gene regulatory events. [9] identify semantic relations of medical entities, whereas [10] performed a large-scale analysis of the whole of Medline to collect and characterise associations of genes and proteins through syntactic parsing.

With the help of linguistic patterns and domain knowledge, [9] is linking the correct semantic relation to any pair of medical entities that have been identified from the scientific literature. For example, the assignment of a relation denoting the treatment of a disease could be achieved at 76% precision and 61% recall.

The identification of the DNA methylation events required the preparation of a specialised corpus to evaluate the performance of a retrained state-of-the-art event extraction system (78% precision, 76% recall). For the identification of gene regulatory events, [8] made use of a semantic resource for the events to derive inference rules encoding the domain knowledge. The inference has been applied to deduce implicit events from explicitly expressed events. The system has been evaluated against different gold standard data resources showing that the inference module contributes to 53.2% of the correct extractions, but does not cause any incorrect results.

At a larger scale, [10] screened all of PubMed to study statements of gene/protein associations based on a combination of solutions including syntactic parsing. The authors suggest an estimate for known event types from existing annotated resources in comparison to the requirements for the identification of such events types from PubMed: out of all event-type associations stated in PubMed that make reference to two proteins, over 90% may be of types for which annotated resources for event extraction exist.

Apart from the identification of entities, relations and events, other research work is concerned with the analysis of the discourse structure of the scientific literature, such as the resolution of co-reference [11] and the identification of speculation, negation and hedging [12,13].

The extraction of event-argument relations from the discourse in biomedical documents requires co-reference resolution to improve the recall of existing approaches. In the presented work by [11], an implementation based on Support Vector Machine (SVM) classifiers has been compared against a joint Markov Logic Network (MLN) and led to the finding that the latter outperforms the former.

For the identification of negation and speculation, [12] acquired hedge cues for the labelling of sentences according to certainty and as part of their solution applied random indexing to achieve dimensionality reduction necessary for large classifiers. [13] exploited linguistic cues from the BioScope corpus to categorize biological events in the GENIA corpus for uncertainty and negation.

Other research work presented at the SMBM is concerned with the processing of data streams [14,15] for the prediction of disease outbreaks and the generation of large-scale data resources [16].

[14] has trained a classifier to attribute class labels for avoidance behaviour, increased sanitation, seeking pharmaceutical intervention, wearing a mask and self-reported diagnosis to 5,283 Twitter messages. SVM classification performed better than a Naives Bayes classifier delivering evidence for a high degree of correlation between pre-diagnostic social media signals and gold standard laboratory data. The analysis of multilingual news for the prediction of disease outbreak leads to the result that sources in

different languages improve the sensitivity of the prediction in comparison to English only news [15].

Another large-scale approach led to the generation of an annotated biomedical corpus (CALBC project) with four different semantic groups through the harmonisation of annotations from automatic text mining solutions, called the Silver Standard Corpus (SSC-I) [16]. This corpus has been used for the First CALBC Challenge asking the participants to annotate the corpus with their automatic annotators (named entity taggers). The best performance over all semantic groups was achieved from two annotation solutions that were trained on the SSC-I.

Altogether, the SMBM2010 symposium covered a large number of questions and tasks that reside on the edge of semantic resources, literature analysis and large-scale data processing. Growing all resources together could significantly mature the field of semantics in biomedicine. The key steps forward are determined by the standardisation, formalisation and reuse of semantic resources, i.e. ontologies in the first place, the stringent evaluation of analytical solutions and the interoperability of resources and analytical solutions on a large scale. Certainly SMBM will continue to contribute to these objectives in the future and will help to shape and judge on the progress.

Acknowledgements

This article has been published as part of *Journal of Biomedical Semantics* Volume 2 Supplement 5, 2011: Proceedings of the Fourth International Symposium on Semantic Mining in Biomedicine (SMBM). The full contents of the supplement are available online at http://www.jbiomedsem.com/supplements/2/S5.

Author details

[1]EMBL Outstation, European Bioinformatics Institute, Hinxton, Cambridge, CB10 1SD, UK. [2]University of Zürich, Zürich, Switzerland. [3]Department of Computer Science, University of Tokyo, Tokyo, Japan. [4]National Institute of Informatics, 2-1-2 Hitotsubashi, Chiyoda-ku,Tokyo, Japan. [5]Language & Information Engineering (JULIE) Lab, Friedrich-Schiller-Universität, Jena, Germany.

Published: 6 October 2011

References

1. Colosimo M, Morgan A, Yeh A, Colombe J, Hirschman L: **Data preparation and interannotator agreement: BioCreAtIvE task 1B.** *BMC bioinformatics* 2005, **6(Suppl 1)**:S12.
2. Rebholz-Schuhmann D, Jimeno Yepes A, Van Mulligen E, Kang N, Kors J, Milward D, Corbett P, Buyko E, Beisswanger E, Hahn U: **"CALBC Silver Standard Corpus.".** *J Bioinform Comput Biol* 2010, **8(1)**:163-79.
3. Kim J, Ohta T, Tsuruoka Y, Tateisi Y, Collier N: **Introduction to the bio-entity recognition task at JNLPBA.** *Proceedings of the International Joint Workshop on Natural Language Processing in Biomedicine and its Applications* Association for Computational Linguistics; 2004, 70-75.
4. Rebholz-Schuhmann D, Kirsch H, Nenadic G: **IeXML: towards a framework for interoperability of text processing modules to improve annotation of semantic types in biomedical text.** *Proc. of BioLINK, ISMB 2006, Fortaleza, Brazil* 2006.
5. Kim J, Ohta T, Tateisi Y, Tsujii J: **GENIA corpus-a semantically annotated corpus for bio-textmining.** *Bioinformatics* 2003, **19(1)**:180-182.
6. Hoehndorf R, Ngonga Ngomo AC, Pyysalo S, Ohta T, Oellrich A, Rebholz-Schuhmann D: **Ontology design patterns to disambiguate relations between genes and gene products in GENIA.** *Journal of Biomedical Semantics* 2011, **2(Suppl 5)**:S1.
7. Ohta T, Pyysalo S, Miwa M, Tsujii J: **Event Extraction for DNA Methylation.** *Journal of Biomedical Semantics* 2011, **2(Suppl 5)**:S2.
8. Kim JJ, Rebholz-Schuhmann D: **Improving the extraction of complex regulatory events from scientific text by using ontology-based inference.** *Journal of Biomedical Semantics* 2011, **2(Suppl 5)**:S3.
9. Abacha AB, Zweigenbaum P: **Automatic Extraction of Semantic Relations between Medical Entities: a Rule Based Approach.** *Journal of Biomedical Semantics* 2011, **2(Suppl 5)**:S4.
10. Pyysalo S, Ohta T, Tsujii J: **An Analysis of Gene/Protein Associations at PubMed Scale.** *Journal of Biomedical Semantics* 2011, **2(Suppl 5)**:S5.
11. Yoshikawa K, Riedel S, Hirao T, Asahara M, Matsumoto Y: **Coreference Based Event-Argument Relation Extraction on Biomedical Text.** *Journal of Biomedical Semantics* 2011, **2(Suppl 5)**:S6.
12. Velldal E: **Predicting speculation: A simple disambiguation approach to hedge detection in biomedical literature.** *Journal of Biomedical Semantics* 2011, **2(Suppl 5)**:S7.
13. Vincze V, Szarvas G, Mora G, Ohta T, Farkas R: **Linguistic scope-based and biological event-based speculation and negation annotations in the BioScope and Genia Event corpora.** *Journal of Biomedical Semantics* 2011, **2(Suppl 5)**:S8.

14. Collier N, Son NT, Nguyen NM: **OMG U got flu? Analysis of shared health messages for bio-surveillance.** *Journal of Biomedical Semantics* 2011, **2(Suppl 5)**:S9.
15. Collier N: **Towards Cross-lingual Alerting for Bursty Epidemic Events.** *Journal of Biomedical Semantics* 2011, **2(Suppl 5)**:S10.
16. Rebholz-Schuhmann D, Yepes AJ, Li C, Kafkas S, Lewin I, Kang N, Corbett P, Milward D, Buyko E, Beisswanger E, Hornbostel K, Kouznetsov A, Witte R, Laurila JB, Baker CJO, Kuo CJ, Clematide S, Rinaldi F, Farkas R, Móra G, Hara K, Furlong L, Rautschka M, Neves ML, Pascual-Montano A, Wei Q, Collier N, Faisal M, Chowdhury M, Lavelli A, Berlanga R, Morante R, Van Asch V, Daelemans W, Marina JL, van Mulligen E, Kors J, Hahn U: **Assessment of NER solutions against the first and second CALBC Silver Standard Corpus.** *Journal of Biomedical Semantics* 2011, **2(Suppl 5)**:S11.

doi:10.1186/2041-1480-2-S5-I1
Cite this article as: Rebholz-Schuhmann *et al.*: Towards mature use of semantic resources for biomedical analyses. *Journal of Biomedical Semantics* 2011 **2**(Suppl 5):I1.