Towards Standardisation of Named-entity Annotations in the Life Science Literature

Dietrich Rebholz-Schuhmann[1] and **Goran Nenadic**[2]
[1]European Bioinformatics Institute
Wellcome Trust Genome Campus, Hinxton, Cambridge, CB10 1SD, U.K
[2] School of Computer Science, University of Manchester
Oxford Road, Manchester, M13 9PL, U.K.
rebholz@ebi.ac.uk, G.Nenadic@manchester.ac.uk

Summary

The main aim of this proposal is to revisit and in the best case re-launch an initiative that would provide harmonised ways for representing and tagging named entities in the life science literature. We are proposing to establish common document formats that facilitate the exchange of annotation *results* contained in the literature as a complementary approach to the development of interoperable *tools*. We want to work towards (a) recommendations for a common syntax to embody entity mentions in publishers' document formats (e.g., into PMC), and (b) provision of a common way to reference semantic types. The main stakeholders (text mining users, researchers, service providers and publishers) would need to build an infrastructure that integrates literature resources with entity databases. The main benefits result from better integration of literature resources and text-mining results with data from other biomedical research groups and from the identification of the next generation challenges for novel text mining research.

1 Motivation, aims and stakeholders

Identification and annotation of entities of different semantic types is the key factor for accessing biomedical literature. While there have been numerous solutions proposed to identify entities in text (see BioCreAtIve initiative), there are very few community-wide efforts to provide harmonised annotations both for the syntactic and semantic levels, which would facilitate interoperability and re-use of processed documents (Krallinger et al., 2007). This is in contrast to widespread attempts to standardise semantic descriptions and exchange of non-textual biomedical data. Instead, text mining solutions are typically based on their own annotation schemas, making it difficult for the community to easily combine and expand different solutions. This also hinders further developments in the area, as many user and research groups need to allocate significant resources in re-developing and re-aligning existing solutions.

We would therefore like to re-launch an initiative that would result in a community-agreed way for representing and tagging named entities (NEs) in biomedical documents. A harmonised approach would provide the stakeholders with the following:

- the users would be able to use annotated results from different sites (i.e., repositories) to have efficient knowledge acquisition and exploitation (e.g., semantics-based browsing, visualisation, integration);
- the text mining research and service provision communities would profit from document annotations originated from different applications to improve the state of the art in NER, and motivate progress in other text mining tasks;
- publishers and industry would be able to provide an added value to their products, and thus facilitate data sharing, availability and interoperability.

2 Harmonising annotation of named entities: needs and obstacles

Informal discussions within the bio-text mining community (Kevin Cohen, BioNLP) have concluded that more efforts are needed to provide interoperability of tools and data, and — in partic-

ular — that named entities would make an optimal level for text annotations that would facilitate the exchange of text mining results. Recent initiatives from publishers (e.g. Elsevier, FEBS Letter experiment) have re- affirmed these conclusions: both users and data providers are interested in "changing the ways science is published" (the Elsevier Grand challenge[1] 2008), and it seems that annotating and linking NEs to databases is a minimal requirement to support this aim. Publishers already consider requesting authors to annotate key entities in their articles (at least at the document level). Although there are still issues in bio-NER, it would be useful to enable users and developers alike to move beyond named entity recognition by providing documents with pre-annotated NEs in a common format, so that they can use pre-calculated NE annotations for visualisation, browsing, indexing or further processing. Many applications need NEs recognised before any further processing, and a common way of their annotation would only improve the possibility for using and sharing results, as well as for improving research that depends on NE annotations.

The main obstacles in this process are that several research and service provision groups have already developed and used numerous in-house formats and that there is no theoretical consensus on certain annotation issues (e.g. representation of ambiguities). There have been several attempts to address representation of NEs in the community (e.g. IeXML, SciXML, Genia, TXM, Termino, etc.), but to the best of our knowledge, so far there is not a comprehensive comparative analysis between different (text-mining derived) annotation schemas. Furthermore, there have been very few attempts to integrate publisher/archiving annotation formats with text mining results (e.g. IeXML, partly SciXML, Genia) (Rebholz-Schuhmann et al., 2006; Copestake et al., 2006; Kim et al., 2003; Harkema et al., 2005)

Data representation that supports interaction with end users (both experts and non-experts) has also been identified as one of the key objectives of the recently launched EU Elixir project[2], which aims to examine the status of literature repositories throughout Europe and provide recommendations for a future information-sharing infrastructure platform that would integrate databases and literature.

3 Proposed approach

We would like to design a minimal tag set that would be *integrated* into publishers' formats and be part of meta-data used to annotate NE mentions in text and point to their semantic types and their referent IDs (if available). We would like to develop an industry-wide solution that would make interoperability much more realistic. In addition to syntactic harmonisation, we would also like to discuss semantic "normalisation" and a common way to point to (external) semantic resources. More precisely, we would like to initiate further discussions on the harmonisation of representations of bio-NEs in documents, including:

- at the syntactic level, the identification of a minimal set of NE tags and features (inline and stand-off) to be included in publishers' formats, including representation of ambiguities and multiple annotations (e.g. annotations from different groups/services);
- at the semantic level: the integration of a basic semantic type system into document formats, including the provisions for using references/pointers to external type systems (e.g. existing ontologies or purposely-built type systems[3]).

A solution would be to (a) implement a common basic/minimal syntax to annotate entity mentions in documents, and (b) provide a common way to point to (potentially external) semantic types. This way we would provide data exchange and interoperability on the level of data (in addition to potential interoperability of tools).

4 Road map

One of the results from previous discussions was a minimal annotation framework that included a single tag and number of mandatory (semantic) attributes describing entities (Rebholz-Schuhmann et al., 2006). Building on that as well as other contributions, we suggest the following road map:

1. Discuss and identify during the discussion at the SMBM 2008 the potential benefits and

[1] http://www.elseviergrandchallenge.com/

[2] http://www.elixir-europe.org

[3] E.g. a UIMA complaint type system at: http://www.u-compare.org/type_system.html

obstacles as well as issues of shared and disjoint interest.

2. Identify a working group to prepare a set of recommendations, following the consultations with interested research groups, publishers, service provides (e.g. EBI, NaCTeM, BioCreative Meta-server, etc.) and organisers of text mining challenges (e.g. BioCreative). The group will recommend a minimal annotation type system and invite for comments from the community and stakeholders.

References

A. Copestake, P. Corbett, P. Murray-Rust, C. J. Rupp, A. Siddharthan, S. Teufel, and B. Waldron. 2006. An architecture for language processing for scientific texts. In *Proc. UK e-Science All Hands Meeting*.

H. Harkema, I. Roberts, R. Gaizauskas, and M. Hepple. 2005. A web service for biomedical term lookup. *Comparative and Functional Genomics*.

J.-D. Kim, T. Ohta, Y. Tateisi, and J. Tsujii. 2003. GENIA corpus—a semantically annotated corpus for bio-textmining. *Bioinformatics*, 19:i180–182.

M. Krallinger, F. Leitner, and A. Valencia. 2007. Assessment of the second BioCreative PPI task: Automatic extraction of protein-protein interactions. In *Proc. BioCreative II*.

D. Rebholz-Schuhmann, H. Kirsch, and G. Nenadic. 2006. IeXML: towards an annotation framework for biomedical semantic types enabling interoperability of text processing modules. In *Proc. BioLink, ISMB 2006*.