Pacific Symposium on Biocomputing 15:477-480(2010)

EXTRACTION OF GENOTYPE-PHENOTYPE-DRUG RELATIONSHIPS FROM TEXT: FROM ENTITY RECOGNITION TO BIOINFORMATICS APPLICATION

ADRIEN COULET[1,2], NIGAM SHAH[2], LAWRENCE HUNTER[4], CHITTA BARRAL[5], RUSS B. ALTMAN[1,3]

1.Department of Genetics, 2.Department of Medicine, 3.Department of Bioengineering, Stanford University, Stanford, CA 94305, USA

4.Center for Computational Pharmacology, University of Colorado Denver School of Medicine, Aurora, CO 80045, USA

5. Department of Computer Science and Engineering, Arizona State University, Tempe, AZ 85287, USA

E-mail: coulet@stanford.edu

Advances in concept recognition and natural language parsing have led to the development of various tools that enable the identification of biomedical entities and relationships between them in text. The aim of the *Genotype-Phenotype-Drug Relationship Extraction from Text* workshop (or *GPD-Rx* workshop) is to examine the current state of art and discuss the next steps for making the extraction of relationships between biomedical entities integral to the curation and knowledge management workflow in Pharmacogenomics. The workshop will focus particularly on the extraction of Genotype-Phenotype, Genotype-Drug, and Phenotype-Drug relationships that are of interest to Pharmacogenomics. Extracting and structuring such text-mined relationships is a key to support the evaluation and the validation of multiple hypotheses that emerge from high throughput translational studies spanning multiple measurement modalities. In order to advance this agenda, it is essential that existing relationship extraction methods be compared to one another and that a community wide benchmark corpus emerges; against which future methods can be compared. The workshop aims to bring together researchers working on the automatic or semi-automatic extraction of relationships between biomedical entities from research literature in order to identify the key groups interested in creating such a benchmark.

Keywords: NLP; Pharmacogenomics; Entity Recognition; Event Extraction; Genotype-Phenotype-Drug Relationships

1. Introduction

Research in the BioNLP community, such as BioCreative II[1] and the BioNLP Shared Task'09,[2] have led to the development of efficient BioNLP methods for entity recognition and event extraction. The aim of the *GPD-Rx* workshop is to discuss how results of previous shared tasks can be adapted and improved in order to efficiently provide a detailed representation of complex pharmacogenomic processes described in the literature. The extraction of a structured and fine-grained representation is a key to evaluate and validate hypothesis that emerge from translational studies. The objective of the *GPD-Rx* workshop is thus to advance in this direction by identifying key groups and propose corpus, standard vocabularies, knowledge representation language and evaluation methods that would enable the comparison and the interoperability of future results.

2. Entity Recognition

Entity recognition or named entity recognition is the task of identifying, in free text, words that mention a known entity. Most of the efforts aimed at extracting relationships between entities start with this fundamental task in order to identify entities to be related.

Entity recognition has been extensively studied in the biomedical domain with varying results. Some of the proposed methods are generic and can identify any kind of entity that is part of a dictionary provided as a reference to the system.[3,4] Other methods are specialized in the recognition of specific kinds of entities such as genes/proteins,[5] genomic variations,[6,7] diseases,[8] or drugs.[9] Machine learning approaches are commonly integrated with entity recognition methods to improve their results.[10]

The first goal of the *GPD-Rx* workshop is to discuss issues in the recognition of entities relevant to pharmacogenomics.

Pacific Symposium on Biocomputing 15:477-480(2010)

3. Extraction of Relationships between Entities

The second goal of the workshop is to discuss the application, in pharmacogenomics, of methods that extract relationships between relevant entities (*e.g.* genomic variation, phenotype, drug).

One simple approach is based on the hypothesis that two entities which are frequently mentioned together are associated. Entity recognition methods have been applied to search for the co-occurrences of entities with the goal of discovering associated ones.[11] This approach has been applied for the construction of gene networks[12] or the guidance of biomedical curation.[13] In such co-occurance driven approaches, associations have a higher chance to be true when the co-occurrence of entities is observed in a small amount of text (*e.g.* a sentence), and a lower chance to be true when observed in larger amounts (*e.g.* a full section).

The development of natural language parsers have led to a second approach that enables, by providing the grammatical structure of sentences, the extraction of relationships (or events) mentioned in the text. The importance of learning protein-protein interactions in biology has motivated many researchers to use parsers to extract such relations with a high accuracy. The work of Fundel *et al.*,[14] of Rebholz-Schuhmann *et al.*,[15] of Hunter *et al.*,[16] and of Miyao *et al.*[17] illustrate the latest research in extracting biomedical relationships from text.

Similar approaches have already been developed for the extraction of Genotype-Phenotype-Drug relationships.[18–21] The *GPD-Rx* workshop aims at identifying issues specific to this task and to using the output of such efforts. For example, the comparison of extracted relationships, to determine agreement or to point out a contradiction, is a key to make extracted relationships actionable.

4. Standards

We belive that BioNLP groups focused on relationship extraction tasks would have a mutual interest in using shared standards to facilitate the comparison and the interoperability of their results. The main ones are:

- the use of unique identifier for entities involved in relationships,
- the use of a common knowledge representation language for the description of relationships,
- evaluation methods for the extraction of relationships,
- shared text corpora and vocabularies of entity names and vocabularies of relationship type,
- set of gold standard relationships.

The workshop aims to stimulate discussion for identifying, sharing and wide-spread use of such standards when applying text-mining in the realm of pharmacogenomics.

Acknowledgments

We would like to thank the PSB 2010 organizers and particularly Tiffany Murray for helping us in the organization of the *GPD-Rx* workshop.

References

1. Lynette Hirschman, Martin Krallinger, John Wilbur and Alfonso Valencia, Editors. 'The BioCreative II - Critical Assessment for Information Extraction in Biology Challenge', *Genome Biology*, **9**(S2), (2008).
2. Jun'ichi Tsujii, Editor. *Proceedings of the BioNLP 2009 Workshop Companion Volume for Shared Task*, (2009).
3. Alan R. Aronson, 'Effective mapping of biomedical text to the UMLS Metathesaurus: the MetaMap program'. *Proceedings of the AMIA Symposium*, pp. 17–21, (2001).
4. Manhong Dai, Nigam H. Shah, Wei Xuan, Mark A. Musen, Stanley J. Watson, Brian D. Athey and Fan Meng, 'An Efficient Solution for Mapping Free Text to Ontology Terms', *Proceedings of the AMIA Summit on Translational Bioinformatics*, (2008).
5. Larry Smith, Lorraine K Tanabe, Rie Johnson nee Ando, Cheng-Ju Kuo, I-Fang Chung, Chun-Nan Hsu, Yu-Shi Lin, *et al.*, 'Overview of BioCreative II gene mention recognition', *Genome Biology*, **9**(S2), (2008).
6. J. Gregory Caporaso, William A. Baumgartner Jr., David A. Randolph, K. Bretonnel Cohen and Lawrence Hunter, 'MutationFinder: a high-performance system for extracting point mutation mentions from text', *Bioinformatics*, **23**(14), pp. 1862–1865, (2007).

Pacific Symposium on Biocomputing 15:477-480(2010)

7. Christopher J.O. Baker and Dietrich Rebholz-Schuhmann, Editors. 'Proceedings of the European Conference on Computational Biology (ECCB) 2008 Workshop: Annotations, interpretation and management of mutations (AIMM)', *Bioinformatics*, **10**(S8), (2009).
8. Rong Xu, Kaustubh Supekar, Alex Morgan, Amar Das and Alan M. Garber, 'Unsupervised Method for Automatic Construction of a Disease Dictionary from a Large Free Text Collection', *Proceedings of the AMIA Symposium*, (2008).
9. Isabel Segura-Bedmar, Paloma Martnez and Mara Segura-Bedmar, 'Drug name recognition and classification in biomedical texts: A case study outlining approaches underpinning automated systems', *Drug Discovery Today*,**13**(17-18), pp. 816–823 (2008).
10. Robert Leaman and Graciela Gonzalez, 'Banner: An executable survey of advances in biomedical named entity recognition', in *Pacific Symposium on Biocomputing*, pp. 652–663, (2008).
11. Graciela Gonzalez, Juan C. Uribe, Luis Tari, Colleen Brophy and Chitta Baral, 'Mining gene-disease relationships from biomedical literature: Weighting proteinprotein interactions and connectivity', in *Pacific Symposium on Biocomputing*, pp. 28–39, (2007).
12. Tor-Kristian Jenssen, Astrid Laegreid, Jan Komorowski and Eivind Hovig, 'A literature network of human genes for high-throughput analysis of gene expression', *Nature Genetics*, **28**(1), pp. 21–8, (2001).
13. Yael Garten and Russ B. Altman, 'Pharmspresso: a text mining tool for extraction of pharmacogenomic concepts and relationships from full text', *BMC Bioinformatics*, **10**(Suppl 2), S6, (2009).
14. Katrin Fundel, Robert Küffner and Ralf Zimmer, 'Relex - relation extraction using dependency parse trees', *Bioinformatics*, **23**(3), 365–371, (2007).
15. Dietrich Rebholz-Schuhmann, Harald Kirsch, Miguel Arregui, Sylvain Gaudan, Mark Riethoven and Peter Stoehr, 'Ebimed - text crunching to gather facts for proteins from medline', *Bioinformatics*, **23**(2), 237–244, (2007).
16. Lawrence Hunter, Zhiyong Lu, James Firby, William A. Baumgartner Jr, Helen L. Johnson, Philip V. Ogren and K. Bretonnel Cohen, 'Opendmap: An open-source, ontology-driven concept analysis engine, with applications to capturing knowledge regarding protein transport, protein interactions and cell-type-specific gene expression', *BMC Bioinformatics*, **9**(78), (2009).
17. Yusuke Miyao, Kenji Sagae, Rune Saetre, Takuya Matsuzaki and Jun'ichi Tsujii, 'Evaluating contributions of natural language parsers to proteinprotein interaction extraction', *Bioinformatics*, **25**(3), 394–400, (2009).
18. Thomas C. Rindflesch, Lorraine Tanabe, John N. Weinstein and Lawrence Hunter, ' EDGAR: extraction of drugs, genes and relations from the biomedical literature', in *Pacific Symposium on Biocomputing*, pp. 517-528, (2000).
19. Caroline B. Ahlers, Marcelo Fiszman, Dina Demner-Fushman, François-Michel Lang and Thomas C. Rindflesch, 'Extracting semantic predications from medline citations for pharmacogenomics', in *Pacific Symposium on Biocomputing*, pp. 209–220, (2007).
20. Luis Tari, Jörg Hakenberg, Graciela Gonzalez and Chitta Baral, 'Querying parse tree database of medline text to synthesize user-specific biomolecular networks', in *Pacific Symposium on Biocomputing*, pp. 87–98, (2009).
21. Luis Tari, Saadat Anwar, Shanshan Liang, Jörg Hakenberg and Chitta Baral 'Synthesis of Pharmacokinetic Pathways Through Knowledge Acquisition and Automated Reasoning', in *Pacific Symposium on Biocomputing*, (2010).