Abeel et al. BMC Bioinformatics 2010, 11(Suppl 5):I1
http://www.biomedcentral.com/1471-2105/11/S5/I1

BMC Bioinformatics

INTRODUCTION Open Access

Highlights of the BioTM 2010 workshop on advances in bio text mining

Thomas Abeel[1*†], Sofie Van Landeghem[1†], Roser Morante[2], Vincent Van Asch[2], Yves Van de Peer[1], Walter Daelemans[2], Yvan Saeys[1]

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

This meeting report gives an overview of the keynote lectures, the panel discussion and a selection of the contributed presentations. The workshop was held in Gent, Belgium on May 10-11. It featured a tutorial aimed towards a broad audience of (computational) biologists, (computational) linguists and researchers working purely on text mining.

Introduction

Recently, the application of text mining (TM) and natural language processing (NLP) techniques to the biological and medical sciences has received increasing interest. In addition to many new workshops and conferences arising in this domain, recently also a number of community-wide tasks were conducted to benchmark text mining techniques on specific challenges (e.g. BioCreative, BioNLP Shared Task, ...)

By discussing the latest developments and potentially new applications in text mining amongst scientists in both academia and industry, this workshop aimed to provide a broad view on text mining research in biology and biomedicine. We reached out to a broad public, including researchers with an interest in text mining but with little or no experience in this domain. To this end, the workshop started with an extensive tutorial on text mining in the bio-sciences, providing sufficient background knowledge for novices.

Next, a number of keynote talks were given by leading scientists, presenting the latest advances in the field. Furthermore, participants were encouraged to submit an abstract describing their own work. Authors of accepted posters were given the opportunity to give an overview of their research in a two minute flash presentation, as well as to present a related poster during the coffee and lunch breaks.

Invited contributions

The first day consisted of a tutorial by Martin Krallinger, a renowned researcher in the field of biomedical text mining, and one of the main organizers of the BioCreative community challenge. His presentation also covered some of the latest advances in the field [1]. Sampo Pyysalo, involved in the organization of the BioNLP Shared Task, concluded the first day with a presentation focusing mainly on defining a good representation of biomolecular facts expressed in text, enabling text mining tools to extract more complex and more detailed information than ever before [2].

The second day of the workshop was more focused on practical applications. The presentations "Event extraction on PubMed scale" by Filip Ginter [3] and "Integrating text mining into high-throughput assay analysis" by Kevin Cohen [4] illustrate the very first steps in generating text mining results for the entire scientific literature and applying text mining for the analysis of high-throughput assays put experimental assays. Both presentations contribute to the goal of bringing theoretical text mining tools closer to their practical application, which was one of the main topics of the workshop.

The final two presentations "Pitfalls in applying text mining to scientific literature" by Jean-Marc Neefs [5] and "Integrating automated literature searches and text-mining in biomarker discovery" by Luc Dehaspe and Maté Ongenaert [6] gave an overview of text mining as it is being used in the pharmaceutical industry. At the same time, some important caveats for text mining algorithms were listed and discussed.

* Correspondence: thomas@abeel.be
† Contributed equally
[1]VIB Department of Plant Systems Biology, Ghent University, Gent, Belgium
Full list of author information is available at the end of the article

Contributed poster presentations

All contributed abstracts were rigorously reviewed (each by at least 2 different reviewers) and the 9 best poster abstracts were selected by the program committee for inclusion in this meeting report.

The selected set of abstracts covers some of the more traditional text mining topics such as gene normalization and extracting protein-protein interactions (PPIs), while also including exciting new topics such as identifying splice variants and extracting drug-drug interactions (DDIs). A few other abstracts further detail interesting applications of text mining in various challenges.

Covering the more traditional topics, Solt et al. presented their work on species identification for the task of gene normalization [7], while Tikk et al. applied a novel dependency graph kernel for PPI extraction [8]. Both gene normalization and PPI, together with a few other related topics, were further discussed in the presentation of Chen et al., detailing their tool created for the Biocreative II.5 challenge [9].

Broadening the view on PPI extraction in a novel way, Kafkas et al. presented their work on mining alternative splice forms to capture functional variation in PPI networks [10]. At the same time, Segura et al. apply similar, well-known supervised learning algorithms to the novel use-case of extracting DDIs from biomedical texts [11].

Sluban et al. gave a presentation on outlier detection methods developed to search for cross-context links [12]. Related to this work, Juršič et al. discussed the identification of concepts bridging diverse biomedical domains [13].

Two final papers deal with practical applications of text mining in the biomedical domain: Verslyppe et al. presented their work on semantic integration of isolation habitat and location [14], while Ohta et al. gave an update on Medie and Info-pubmed, two recently developed text mining applications [15].

Panel discussion

At the end of the second day of the workshop, a round table discussion was initiated by asking all invited speakers their opinion on current achievements and future directions for text mining research in the biomedical domain.

During this panel discussion, it was stressed that the text mining community needs to understand and address the specific needs of the users of TM tools. For example, good visualization tools are still lacking in the domain. Another example is illustrated by the biologists' interest in knowing the exact experimental technique used to determine the existence of a protein-protein interaction. Often, tools extracting PPIs do not provide this type of information.

Next, scientists working on TM algorithms should ensure the reproducibility of these tools by for example providing open-source implementations or detailed descriptions of their algorithms. Alternatively, it should be considered to make data publicly available whenever this is feasible.

Interesting topics for future work have been suggested to include the integration of database information with experimental data. Also, despite recent efforts, the processing of full texts continues to be a largely unsolved challenge, in particular mining tables and figures poses significant problems, though the results will be very rewarding as a lot of information is (only) provided in tables.

One important last issue of discussion involves the need for manual curation of biomedical texts, enabling the support of supervised learning methods, which have shown to produce state-of-the-art performance. However, several questions arise: who should do the actual annotation? And who can pay for these efforts?

During a lively discussion with the audience, it was agreed upon that annotation should still be done by trained curators, as authors are generally not good or consistent enough to perform this complex task. However, TM tools could help curators by offering suggestions for annotations which can then be evaluated and adjusted by the human expert. Regarding funding, it was suggested that people should be creative with project money, as it is often too hard to receive grants for doing purely annotation work.

Considering post translational modifications as a use-case for annotation in biomolecular texts, it should be noted that about 300 distinct types exist. It is obvious that it is practically impossible to cover all these types by only one research group or project. To be able to share the "annotation burden", the community should thus consider adopting one standard format of annotation, e.g. following the event annotation format of the recent BioNLP Shared Task. It should then become possible for different groups to provide new types of information, by simply adding their annotations to the dataset publicly available for the whole community.

Acknowledgements

The organizers would like to acknowledge the Fund for Scientific Research Flanders (FWO-Vlaanderen) for supporting this workshop through the following research communities (FWO-WOG): "Machine Learning for Data Mining and its Applications" and "Computational Linguistics in Flanders". Additional funding was obtained from the Biograph project, the universities of Ghent and Antwerp, and the Flemish Interuniversitary Institute for Biotechnology (VIB).

Author details

[1]VIB Department of Plant Systems Biology, Ghent University, Gent, Belgium. [2]CLiPS - Computational Linguistics, University of Antwerp, Antwerp, Belgium.

Published: 6 October 2010

References

1. Krallinger M, Tendulkar AV, Leitner F, Chatr-aryamontri A, Valencia A: **The PPI affix dictionary (PPIAD) and BioMethod Lexicon: importance of affixes and tags for recognition of entity mentions and experimental protein interactions.** *BMC Bioinformatics* 2010, **11(Suppl 5)**:O1.
2. Pyysalo S: **Entities, relations, events, representing biomolecular semantics.** *BMC Bioinformatics* 2010, **11(Suppl 5)**:O6.
3. Ginter F, Björne J, Pyysalo S: **Event extraction on PubMed scale.** *BMC Bioinformatics* 2010, **11(Suppl 5)**:O2.
4. Cohen KB: **Integrating text mining into high-throughput assay analysis.** *BMC Bioinformatics* 2010, **11(Suppl 5)**:O3.
5. Neefs J.-M: **Pitfalls in applying text mining to scientific literature.** *BMC Bioinformatics* 2010, **11(Suppl 5)**:O4.
6. Ongenaert M, Dehaspe L: **Integrating automated literature searches and text mining in biomarker discovery.** *BMC Bioinformatics* 2010, **11(Suppl 5)**: O5.
7. Solt I, Tikk D, Leser U: **Species identification for gene name normalization.** *BMC Bioinformatics* 2010, **11(Suppl 5)**:P5.
8. Tikk D, Palaga P, Leser U: **A fast and effective dependency graph kernel for PPI relation extraction.** *BMC Bioinformatics* 2010, **11(Suppl 5)**:P8.
9. Chen Y, Liu F, Manderick B: **BioLMiner and the BioCreative II.5 challenge.** *BMC Bioinformatics* 2010, **11(Suppl 5)**:P6.
10. Kafkas Ş, Varoğlu E, Rebholz-Schuhmann D, Taneri B: **Functional variation of alternative splice forms in their protein interaction networks: A literature mining approach.** *BMC Bioinformatics* 2010, **11(Suppl 5)**:P1.
11. Segura-Bedmar I, Martínez P, de Pablo-Sánchez C: **Extracting drug-drug interactions from biomedical texts.** *BMC Bioinformatics* 2010, **11(Suppl 5)**: P9.
12. Sluban B, Lavrač N: **Supporting the search for cross-context links by outlier detection methods.** *BMC Bioinformatics* 2010, **11(Suppl 5)**:P2.
13. Juršič M, Mozetič I, Grčar M, Cestnik B, Lavrač N: **Identification of concepts bridging diverse biomedical domains.** *BMC Bioinformatics* 2010, **11(Suppl 5)**:P4.
14. Verslyppe B, De Smet W, De Vos P, De Baets B, Dawyndt P: **Semantic integration of isolation habitat and location in StrainInfo.** *BMC Bioinformatics* 2010, **11(Suppl 5)**:P3.
15. Ohta T, Matsuzaki T, Okazaki N, Miwa M, Sætre R, Pyysalo S, Tsujii J: **Medie and Info-pubmed: 2010 update.** *BMC Bioinformatics* 2010, **11(Suppl 5)**:P7.

doi:10.1186/1471-2105-11-S5-I1
Cite this article as: Abeel *et al.*: Highlights of the BioTM 2010 workshop on advances in bio text mining. *BMC Bioinformatics* 2010 **11**(Suppl 5):I1.