Content mining at the Spanish National Cancer Research Centre: the BioCreative experiences

Commercial use (Y/N):N
Related project (*if any – please specify funder and project name*): eTOX, under Grant Agreement nº115002, MICROME, Grant Agreement nº222886-2
Author: Martin Krallinger and Alfonso Valencia
Organization: Spanish National Cancer Research Centre
Date:

Use case description – *Describe the problem you are trying to address, the research area it affects, and the TDM application used. Please state all participants, whether and how the results were subsequently used.*

One of the key issues for the implementation, advancement and practical usefulness of TDM is the construction of suitable Gold Standard datasets to train and assess the performance of the used methodologies, which are increasingly relying on supervised and semi-supervised machine learning algorithms. In this context we are organizing an international community challenge called BioCreative (Critical Assessment of Information Extraction systems in Biology - **http://www.biocreative.org**).

This international community assessment effort has the aim of (1) evaluating the performance and limitations of text mining and information extraction systems applied to the biomedical domain as well as (2) promoting the construction of suitable training text corpora and (3) motivating the implementation of practical useful biomedical text mining applications. So far, four official BioCreative challenges have been carried out, namely BioCreative I, II, III and IV, as well as two additional efforts, BioCreative II.5 and the BioCreative 2012 workshop, with an overall participation of more than 300 developers of participating text mining and natural language processing systems.

In the organization of BioCreative several institutions and research groups have been actively involved, including the Spanish National Cancer Research Centre (Alfonso Valencia), the National Center for Biotechnology Information (John Wilbur and Zhiyong Lu), the MITRE corporation (Lynette Hirschman), Cathy H. Wu (University of Delaware), Gianni Cesareni (University of Rome Tor Vergata) or the European Bioinformatics Institute (Henning Hermjakob) among others.

The organization of BioCreative was motivated by the increasing number of groups working in the area of text mining. However, despite increased activity in this area, there were no common standards or shared evaluation criteria to enable comparison among the different approaches. The various groups were addressing different problems, often using private data sets, and as a result, it was impossible to determine how good the existing systems were, whether they would scale to real applications, and what performance could be expected. A common issue with those private datasets was restriction in terms of copyright issues to make the text annotations available to other research teams.

The main emphasis of BioCreative is on the comparison of methods and the community assessment of scientific progress, rather than on the purely competitive aspects. There is considerable difficulty in constructing suitable "gold standard" data for training and testing new information extraction

systems that handle life science literature. Thus the data sets derived from the BioCreative challenge - because biological database curators and domain experts have examined them - serve as useful resources for the development of new applications as well as helping to improve existing ones.

The resources resulting from the BioCreative challenges are described in more detail in the following special journal issues:

- BioCreative 2012 - Workshop Proceedings
- BioCreative 2012 - Database Virtual Issue
- BioCreative I - BMC Bioinformatics
- BioCreative II - Workshop Booklet
- BioCreative II - Genome Biology
- BioCreative II - Workshop Proceedings
- BioCreative II.5 - Workshop Booklet
- BioCreative II.5 - ISMB/ECCB 09 talks
- BioCreative II.5 - Nat. Biotech. & TCBB
- BioCreative III - BMC Bioinformatics
- BioCreative III - Workshop Proceedings
- BioCreative IV - Workshop Proceedings Volume 1 & 2
- BioCreative IV - Workshop Proceedings Volume 1
- BioCreative IV - CHEMDNER Proceedings Volume 2
- BioCreative IV – CHEMDNER Journal of Cheminformatics

See **http://www.biocreative.org/resources** for links to the corresponding articles.

BioCreative has encouraged the availability of the text mining systems that participated in each of the tasks either in the form of online web-applications or as modules that could be integrated into in-house text mining pipelines.

Content sources – *List the targeted content sources and types of content to used.*

- **Scientific publication abstracts** from biomedically relevant journals hosted by the PubMed database (**http://www.ncbi.nlm.nih.gov/pubmed**)
- **Scientific full text publications** from the database (**http://www.ncbi.nlm.nih.gov/pmc**)
- **Full text publications from a set of selected journals** obtained directly from a small list of publishers that could only be used during the competitions
- **Medicinal chemistry patents** (will be used for one task in BioCreative V)
- **Biomedical annotation knowledge-base records** (related to the topics of proteomics, functional annotation of genes and toxicogenomics).

Targeted users – *Describe end users, their number and expertise.*

- **Text mining and NLP** developers working on the processing of unstructured data in the domains of life sciences, biomedicine or chemistry. Also developers of building block technologies such as information retrieval, text categorization, named entity recognition, named entity grounding, information extraction, relation mining, development of interactive text mining systems or visualization of text mining results.
- **Database curators** carrying out literature curation, including model organism databases (MOD) like TAIR, MGI, RGD, WormBase, FlyBase, MaizeGDB; Functional genome annotation databases (e.g. GOA), proteomics databases (BioGRID, IntAct, MINT), comparative toxicogenomics (CTD).
- **Experimental biomedical and basic science researches:** (1) to improve the interpretation and design of experimental research by improving the access to previously published information on the studied bio-entities; (2) using literature mining and knowledge discovery software for the generation of new hypotheses that will be experimentally validated.
- **Clinicians**: improve the information access for evidence-based clinical practice using text mining technologies and biomedical semantic search engines
- **Chemists:** systematic access to chemical information (structure associated chemical entities) described in the literature and patents.
- **Bioinformaticians:** Text-mining assisted curation results are useful as Gold Standard validation sets for predictive bioinformatics results.
- **Pharma Industry:** Drug discovery and target selection, identifying adverse drug effects, competitive intelligence and knowledge management
- **Publishers:** semantic annotation of online publications, structured digital abstracts
- **Scientific papers authors:** author derived annotations (assisted completion of structured digital abstracts)
- **Patients:** improved search engines, especially important for the detection of cases of similar rare disease cases and personalized medicine (automatic detection of mutations descriptions published in the literature)
- **Computer scientists:** useful training data provided by BioCreative to improve the performance of cutting edge statistical machine learning algorithms and feature selection/exploration.

Impact – *Describe all possible impacts of the use case. Specify any cross-border, societal and economic effects (if possible state any monetary benefits and market advantages).*

- **Improving performance of TDM methodologies.** The BioCreative challenges were able to monitor progress over time of text mining addressing particular tasks. This was done by comparing the performance of previously existing systems with those developed using BioCreative datasets. In the case of some key tasks, these were repeated across several BioCreative challenges to study the potential improvement in performance depending on the used training data (size and sampling) as well as the adaptation of new methodologies (especially cutting edge machine-learning algorithms).
- **Promote the development of accessible real world TDM applications.** BioCreative has promoted the implementation of several text mining systems which were made accessible during and after the challenges as interactive online web applications evaluated by domain experts. In the case of some tasks, automatic annotation servers provided by participating teams have been tested and integrated. This was the case with the BioCreative Metaserver prototype.
- **Explore strategies to integrate TDM results with knowledgebase**. Several tasks at BioCreative were posed that required the integration of text mining results to existing knowledge bases of widespread use in the life sciences field.
- **Define formal and practical solution of TDM interoperability.** Improvement of text mining systems (both in terms of performance as well as in terms of technical data integration) required harmonization of text annotations of the same kind on one side and on annotation alignment of results addressing different aspects relevant to the text mining extraction pipeline on the other side. BioCreative promoted the development of a minimal annotation format standard called BioC to improve the exchange and interoperability of text annotations.

Constraints – *Describe any legal, technical, economic, societal, organizational, cultural, multi-lingual or other limitations and you have encountered.*

- *Legal:* There are serious legal hurdles in terms of using full text scientific literature for text mining purposes, staring with a clearly underspecified metadata associated to articles to be processed automatically. There are also no clear guidelines on how much content from full text papers can be redistributed either to carry out manual searches or additional text mining analysis. It is moreover a clear challenge to identify unambiguously the necessary published information and contact person for requesting text mining usage exemptions.
- *Data accessibility:* In general there is no specific way to systematically access full text papers for text mining usage. Such an access should not be provided through conventional online display or papers but through some sort of specific repository and download of full text articles (FTP or web-service).
- *Data format: A more standardized distribution of full text article in terms of format (XML) and metadata would be crucial to avoid serious bottleneck currently existing in terms of document standardization.*

Comments – *Other comments and any recommendations you may have.*