BELIEF - A semiautomatic workflow for BEL network creation

Juliane Fluck[1*], Sumit Madan[1], Sam Ansari[2], Justyna Szostak[2], Julia Hoeng[2], Marc Zimmermann[1], Martin Hofmann-Apitius[1,3], Manuel C. Peitsch[2]

[1]Fraunhofer Institute for Algorithms and Scientific Computing, Schloss Birlinghoven, Sankt Augustin, Germany.

[2]Philip Morris International R&D, Philip Morris Products S.A., Quai Jeanrenaud 5, 2000 Neuchâtel, Switzerland.

[3]Bonn-Aachen International Centre for Information Technology, Dahlmannstr. 2, Bonn, Germany

{jfluck, smadan, marc.zimmermann, mhofmann-apitius}@scai.fraunhofer.de,

{sam.ansari, julia.hoeng, manuel.peitsch}@pmi.com,

justyna.szostak@contracted.pmi.com

Abstract

In order to build networks for systems biology from the literature an UIMA based extraction workflow using various named entity recognition processes and different relation extraction methods has been composed. The Unstructured Information Management architecture (UIMA) is a Java-based framework that allows assembling complicated workflows from a set of NLP components. The new system is processing scientific articles and is writing the open-access biological expression language (BEL) as output. BEL is a machine and human readable language with defined knowledge statements that can be used for knowledge representation, causal reasoning, and hypothesis generation. In order to curate the automatically derived BEL statements, our workflow integrates a curation interface that provides access to BEL statements generated by text mining and that integrates supporting information to facilitate manual curation. By using the semi-automated curation pipeline, expert time to model relevant causal relationships in BEL could be significantly reduced. In this paper the UIMA workflow and key features of the curation interface are described.

1 Introduction

Currently, a lot of effort is invested to manually extract information from scientific articles and encode the relevant parts in machine-readable language. In order to tackle these tasks, curators must be experts in both, the biological domain and the modeling language used for the computational representation of knowledge. Although automatic relation extraction methods for biomedical entities have been developed during the last decade, the tools and the use of automatically generated networks are not widely established in the area of systems biology. This is due to many aspects, including the complexity of the underlying relations, the poor performance of the systems, missing standard output formats for systems biology and missing interfaces to support curation of automatically extracted data. The BELIEF (BEL Information Extraction workFlow) infrastructure embeds an information extraction workflow with state-of-the-art named entity recognition (NER) and relation extraction (RE) methods into an environment where the end user can start his own processes, visualize results and correct the extraction results to generate a precise knowledge base. As a modeling language we make use of the 'Biological Expression language' BEL. The OpenBEL framework is freely available[1] providing an environment for capturing, integrating, storing, and visualizing knowledge. In comparison to other formats, such as BioPax[2] or SMBL[3], BEL coding comes very close to the unstructured text (human readable) yet ensures a structured syntax (machine readable). Its closeness to the unstructured text makes it very suitable to automated extraction and knowledge coding. BEL documents are XML-based (XBEL) and by default contain citations, evidence sentences, and context annotations together with the corresponding BEL statement (cf. example in figure 1). The BEL statement itself provides the relation information in a compact, standardized way, which, after some

[1] http://www.openbel.org/
[2] http://www.biopax.org/
[3] http://sbml.org/Main_Page

training, can easily be understood by domain experts.

```
SET Citation = {"PubMed","Journal of cellular
physiology","12891700","","Michiels C",""}
SET Cell = "Endothelial Cells"
SET Evidence = "However, when exposed to thrombin,
cytokines, or LPS, endothelial cells synthesize and
express tissue factor at their surface."

p(HGNC:F2) increases surf(p(HGNC:F3))
```

Figure 1 BEL document comprising: Citation, Evidence sentence and the BEL statement itself. Context information is not limited to one reference or one evidence sentence.
p() = protein abundance, HGNC = namespace for human genes and proteins, surf = cellSurfaceExpression

2 BEL information extraction workflow (BELIEF) architecture

Several UIMA-based Analysis Engine (AE) archetypes have been developed in order to allow for an easy integration of Named Entity Recognition (NER) and Relation Extraction (RE) modules. Additional components can be added at a later stage. For the integration of new relationship types the rules in the BEL writer had to be adapted as well. The current workflow is described below in figure 2. New workflows can either be started as batch jobs from command line or from a user interface within the BELIEF dashboard. Text readers for Medline abstracts or full text in XML as well as ASCII text are available. BEL makes use of predefined namespaces and identifiers and therefore a mapping from the NER dictionaries to BEL has been integrated.

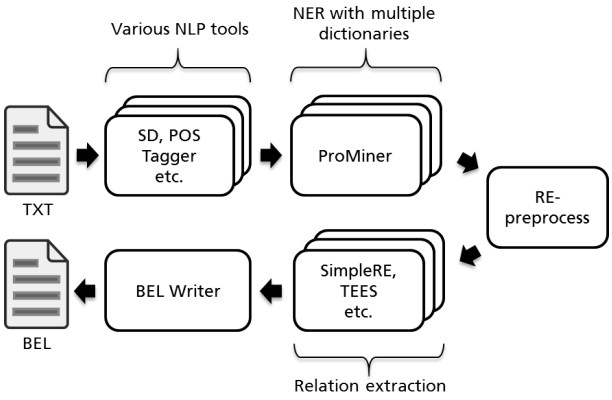

Figure 2 BELIEF extraction workflow

Two types of RE methods have been incorporated: one classification method based on sentence and NER information and one BioNLP shared tasks based method. Finally, a writer has been developed to convert the results of the RE tools into BEL syntax and ultimately into a compliant BEL document.

2.1 Named entity recognition and integrated dictionaries

In the workflow described it is simple to use different NER modules with the restriction that mappings to the established name spaces are necessary for the generation of correct BEL statements. Currently the ProMiner system is used allowing normalization and integration of different dictionaries. ProMiner is well established for NER and shows good performance for the recognition of gene and protein names (Fluck et al. 2007) or disease names (Gurulingappa et al. 2010).

Entity class	Resources	BEL namespace
Human Genes/Proteins	EntrezGene/ Uniprot	HGNC
Mouse Genes/Proteins	EntrezGene/ Uniprot	MGI
Rat Genes/Proteins	EntrezGene/ Uniprot	RGD
Protein family names	OpenBEL	PFH
Protein complex names	OpenBEL	NCH
Protein complex names	Gene Ontology	GOCC-TERM
Chemical names	OpenBEL	SCHEM
Chemical names	ChEBI	CHEBI
Chemical names	ChEMBL	SCHEM
Disease names	MeSH	MESHD
Anatomy names	MeSH	MESHA

Table 1: Integrated dictionaries for different classes.

Existing resources such as the gene/protein name dictionary or MeSH disease dictionary were mapped to corresponding OpenBEL namespace identifiers or names. Other namespace resources such as the protein family names were extended with frequent synonyms or in the case of chemical names three resources have been combined to allow for a higher coverage of concepts.
Table 1 lists all dictionaries currently in use, the corresponding entity classes, original resources, and the name space symbols used within BEL. The recognized named entities could be either used as input for relation extraction or as additional context annotations within the BEL document.

2.2 Relation extraction

The relation extraction methods build on pre-annotated named entities. Since the NER modules perform independent annotations it is necessary to unify and harmonize overlapping matches. These tasks have been combined into a separate AE called RE-preprocess. For overlapping matches with different boundaries the longest match will be taken into account. A ranking of hits is given (e.g. HGNC over MGI or RGD) for multiple hits with the same boundary. This arbitrary selection is only made for relation extraction tools. At the curation interface all detected entities are displayed to the user (cf. figure 4). Thus, the human expert can select entities that were neglected for the relation extraction task.

Currently, two different RE methods are integrated: The linear support vector machine classifier LibLINEAR was trained on five public available training corpora (AIMed, BioInfer, IEPA, HRPD50, and LLL 05) (composed by Pyysalo et al. 2008). The approach is based on lexical features such as bag-of-words (BOW) and n-grams based features. Additionally, dictionary based domain specific trigger words are taken into account as well as dependency parsing features. For details we refer to Bobic et al. 2012. As input the classifier gets sentences with co-occurring entities selected by the RE-preprocess and returns the relation information for the two entities.

As BioNLP shared task method the Turku Event Extraction System TEES (Björne et al. 2012) has been selected. It addresses nearly all BioNLP shared tasks and is one of the top scoring tools in those tasks (Björne & Salakoski 2013). In the BELIEF workflow TEES 2.1 with the models trained on the Genia Event Extraction for NFkB knowledge base (GE), Cancer Genetics (CG) and Pathway Curation (PC) has been integrated as one AE. For further details of the tasks we refer to the BioNLP shared task webpages[4]. The UIMA NER annotated text is given to TEES. Within TEES it was necessary to replace the TEES internal named entity recognizer BANNER by the RE-preprocess that selects the corresponding unified entity annotations. The GE model only gets protein annotations such as the different organism dictionaries or the protein family names. In the PC model chemical entities are required additionally. Event extraction in TEES is done according to default settings of the system and BioNLP shared task annotations are written back into the UIMA CAS object.

Depending on the interfaces of the named entity annotations it is possible to integrate further relation extraction modules into the BELIEF workflow.

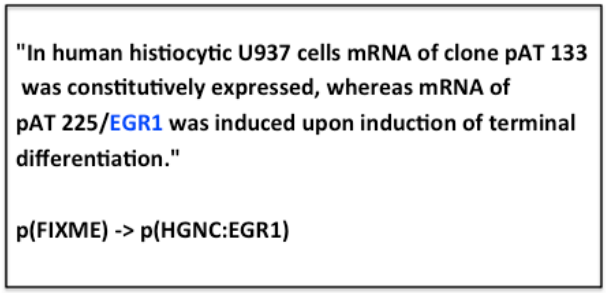

Figure 3 BioNLP to BEL conversion example. For incomplete relations FIXME was introduced as subject to generate a valid BEL statement. The user must correct these FIXMEs during manual curation (in this case to bp(GO: monocyte differentiation); bp = biological process).

2.3 BEL writer

For the translation of the BioNLP shared task systems output to BEL statements, a rule set was generated. The conversion process was described for GE task in detail in Fluck et al. 2013. Standard output for the interaction partners are preferred names together with the name space information and abundance information (cf. examples figure 1 *p(HGNC:F2)*). By default, protein abundance (*p()*) is chosen for proteins, but is converted to RNA abundance (*r()*) for gene expression and transcription events. Protein modification events such as phosphorylation can be directly converted to BEL terms *p(namespace:protein, pmod…)*. All 'Positive Regulation' events in the Shared Task annotations are converted to 'increase' statements of BEL. Similarly, all 'Negative Regulation' events are converted to a 'decrease' statement. Figure 3 displays an example of an incomplete statement and gives an impression of the restrictions of the current automated workflow and the necessity of manual curation. The entity FIXME is always introduced when no CAUSE is found. By definition BEL statements without CAUSE or subject are invalid. Therefore, these statements require manual curation.

[4] http://2013.bionlp-st.org/

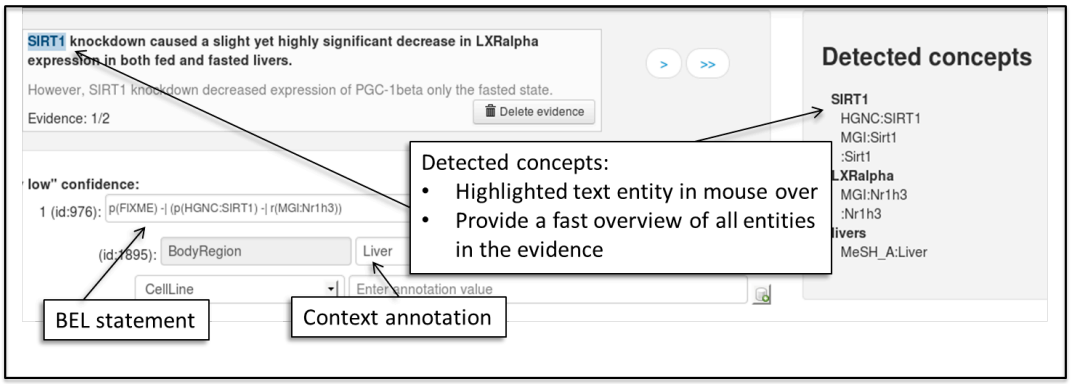

Figure 4 Screenshot of the web based BELIEF Curation Editor.

3 BELIEF Dashboard

The BELIEF Dashboard is a web-based tool allowing for relation extraction and subsequent manual curation of resulting BEL statements. The user can create and manage projects where documents can be uploaded and processed through the BELIEF text mining pipeline. Results of the pipeline are stored in a database and can be displayed in the dashboard. The Curation editor visualizes the extracted BEL statements together with the evidence sentence shown in bold letters (c.f. Figure 4 upper text box). To enable a better understanding of the context the sentences surrounding the evidence text are displayed as well. The corresponding concepts found in the text are shown on the right hand side of the editor. While the mouse is held over the concepts the annotated text in the evidence sentence is highlighted.

The dashboard allows for editing or deleting of found statements (shown in the lower level) or the introduction of new statements based on the evidence sentence. In addition to the BEL statements context information such as organism, anatomy or disease context is shown as annotation and can be edited as well.

The curated statements can be stored and are automatically validated for correct format, valid name spaces and valid reference citation. Based on PMIDs the system is able to retrieve correct citation automatically through a web based PubMed search.

4 Applications

The BELIEF workflow is currently in use for the generation of BEL disease models. ProMiner has been evaluated in the BioCreative assessments for the recognition of gene and protein names and has recall and precision values of approx. 80 percent for human and mouse gene/protein name recognition (Fluck et al. 2007; Hanisch et al. 2005). For chemical names a recall of 90 percent via combination of three dictionaries could be reached in a test set mainly focusing on relations between proteins and protein inhibitors. Similarly, a combination of GO_complex names and BEL complex names lead to a recall rate around 80 percent. For evaluation of relationship extraction only sentences with correctly annotated protein were considered. TEES extracted 42 % correct protein pairs the LibLinerar classification 60%. The combination of both methods reached an overall recall of 74 %. The LibLinerar classification has a higher recall but more curation effort is necessary for the generation of complete BEL statements.

Compared to manual curation, our assisted approach led to significant reduction (40 %) of curation time (Ansari et al. 2014). Additionally, in the Improver Network Verification Challenge[5], a web service was set up supporting participants in writing BEL statements. Users could send text of interest to a BELIEF web service and resulting BEL statements were sent back via e-mail.

5 Summary and Outlook

The first version of BELIEF is in production mode and is already suitable for semi-automatic curation. A demo web server is available under http://www.scaiview.com/belief. In this setting text can be delivered to the BELIEF workflow and results are sent back via e-mail. The release of the BELIEF workflow with the integrated relation extraction modules is planed for the near future.

[5] https://sbvimprover.com/challenge-3/challenge

References

Ansari, S. et al., 2014. A Semi-automated Curation Process for Causal Knowledge Extraction. In *The 7th International Biocuration Conference, Abstract Booklet*. Toronto, p. Poster Abstract 31. Available at: http://biocuration2014.events.oicr.on.ca/files/abstractbooklet.pdf [Accessed June 27, 2014].

Björne, J., Ginter, F. & Salakoski, T., 2012. University of Turku in the BioNLP'11 Shared Task. *BMC bioinformatics*, 13 Suppl 1(Suppl 11), p.S4. Available at: http://www.biomedcentral.com/1471-2105/13/S11/S4 [Accessed June 12, 2014].

Björne, J. & Salakoski, T., 2013. TEES 2.1: Automated annotation scheme learning in the BioNLP 2013 Shared Task. In *Proceedings of the BioNLP Shared Task 2013 Workshop*. Association for Computational Linguistics, pp. 16–25. Available at: http://scholar.google.fi/citations?view_op=view_citation&hl=en&user=geKJpscAAAAJ&citation_for_view=geKJpscAAAAJ:8k81kl-MbHgC [Accessed June 24, 2014].

Bobic, T. et al., 2012. Improving Distantly Supervised Extraction of Drug-Drug and Protein-Protein Interactions. In *Proceedings of ROBUS-UNSUP 2012: Joint Workshop on Unsupervised and Semi-Supervised Learning in NLP*. pp. 35 – 43.

Fluck, J. et al., 2013. BEL networks derived from qualitative translations of BioNLP Shared Task annotations. In *Proceedings of the BioNLP Shared Task 2013 Workshop*. Association for Computational Linguistics (ACL), pp. 80–88. Available at: http://anthology.aclweb.org//W/W13/W13-1910.pdf [Accessed June 26, 2014].

Fluck, J. et al., 2007. ProMiner: Recognition of Human Gene and Protein Names using regularly updated Dictionaries. In L. Hirschmann, M. Krallinger, & A. Valencia, eds. *Proceedings of the Second BioCreative Challenge Evaluation Workshop*. pp. 149–151. Available at: https://www.scai.fraunhofer.de/fileadmin/prominer/ProMinerBioCreative2.pdf [Accessed May 20, 2014].

Gurulingappa, H. et al., 2010. An Empirical Evaluation of Resources for the Identification of Diseases and Adverse Effects in Biomedical Literature. In S. Ananiadou, ed. *International Conference on Language Resources and Evaluation (LREC) : Workshop on Building and Evaluating Resources for Biomedical Text Mining (BioTxtM)*. Valetta, pp. 15–22. Available at: http://www.nactem.ac.uk/biotxtm/papers/Gurulingappa.pdf [Accessed June 24, 2014].

Hanisch, D. et al., 2005. ProMiner: rule-based protein and gene entity recognition. *BMC bioinformatics*, 6 Suppl 1, p.S14.

Pyysalo, S. et al., 2008. Comparative analysis of five protein-protein interaction corpora. *BMC bioinformatics*, 9 Suppl 3(Suppl 3), p.S6. Available at: http://www.biomedcentral.com/1471-2105/9/S3/S6 [Accessed May 2, 2014].