The BEL Information Extraction Workflow (BELIEF): Updates and Evaluation

Sam Ansari1, Sumit Madan2, Justyna Smokta1, Philipp Senger2, Marja Talikka2, Juliane Fluck2, and Julia Hoeng1

1Philip Morris International R&D, Philip Morris Products S.A., Quai Jeanneaud 5, 2000 Neuchâtel, Switzerland (part of Philip Morris international group of companies)
2Fraunhofer Institute for Algorithms and Scientific Computing, Schloss Birlinghoven, Sankt Augustin, Germany

Introduction

The pace in which knowledge is published in the scientific domain is much higher than its application in the interpretation of biological data. In order to reduce this gap, methods are required to convert literature knowledge into a more applicable format, here network models in the biological domain [1]. In 2014, we had introduced the BEL Information Extraction workflow (BELIEF) [2], a semi-automated workflow featuring a text mining pipeline as well as a curation interface. Based on natural language processing (NLP), BELIEF automatically extracts biological entities as well as their relationships, the way they have with each other. These triples are coded in BEL and used for the interpretation of mainly high-throughput data such as transcriptomic data [3].

In this poster, we present the new version with an improved text mining pipeline as well as a new curation interface and show performance indicators that were collected from the BioCreative V Track 4 setup [4,5] and IAT.

BEL (Biological Expression Language)

BEL [6] is a machine and human-readable language that represents molecular relationships and events as semantic tripes where context can include information about the biological and experimental system in which the relationships were observed as well as the publications cited. Unlike other knowledge representation standards such as BioPAX and SBML, BEL comes very close to natural language and proved suitable as the exchange format between text mining and human curation.

BELIEF Improvements

1. Integrated dictionaries for different classes with normalization

2. The Relation Extraction (RE) preprocess selects unified entity annotations for relation extraction.

   The relation extraction tools only receive their own unified annotation, all other matches are provided by the curation.

   The relation extraction is performed with rule based extraction and relation classification as well as an additional deep parsing via BioNLP tools.

   - LiUINEAR [7]
   - TESSY2.1 [8]

   The relation extraction process uses a unified entity annotation for relation extraction.

   The system is very easy to learn for a user who is already familiar with BEL.

   In particular, the annotated protein identifiers are invariably correct which fully stands out when I start to deal with those.

Table 1: Types of BEL Information Extraction

<table>
<thead>
<tr>
<th>Task</th>
<th>Tool</th>
<th>Model usage</th>
<th>Uncontrolled results</th>
<th>Controlled results</th>
</tr>
</thead>
<tbody>
<tr>
<td>Text-mining</td>
<td>BELIEF</td>
<td>High-throughput</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Curation</td>
<td>BELIEF</td>
<td>High-throughput</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Table 2: Text mining results for several classes of BioCreative V NLP task (10-fold cross-validation)

<table>
<thead>
<tr>
<th>Task</th>
<th>Tool</th>
<th>Model usage</th>
<th>Uncontrolled results</th>
<th>Controlled results</th>
</tr>
</thead>
<tbody>
<tr>
<td>Text-mining</td>
<td>BELIEF</td>
<td>High-throughput</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Curation</td>
<td>BELIEF</td>
<td>High-throughput</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Table 3: Curation results for several classes of BioCreative V NLP task (10-fold cross-validation)

<table>
<thead>
<tr>
<th>Task</th>
<th>Tool</th>
<th>Model usage</th>
<th>Uncontrolled results</th>
<th>Controlled results</th>
</tr>
</thead>
<tbody>
<tr>
<td>Text-mining</td>
<td>BELIEF</td>
<td>High-throughput</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Curation</td>
<td>BELIEF</td>
<td>High-throughput</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Summary

BELIEF in its current version better supports domain experts in different stages of knowledge acquisition and network model creation. The results certify that BELIEF shows an improved performance in both, accuracy and recall, as well as a good system stability. The overall impression of all untrained testers was that BELIEF speeds up and further simplifies the creation of BEL statements.

The new and impactful features are:

- Single point of entry including document and task management
- Reduced BEL coding effort due to full and partial BEL statement generation and validation on modifications
- Automatic citation from the Pubmed ID
- Two curation views to facilitate curation (evidence and statement centric view)
- Possibility to use custom dictionaries and re-running the text mining pipeline with these
- Show adjacent sentences to support curation

The key learnings from the user acceptance testing are:

- The success of curation tools lies in providing all relevant information to the curator and limit the curation task to the actual limitations of the automated system
- The preparation of annotation guidelines is critical for consistent annotations across several users
- Comprehensive supporting material is required to facilitate BEL coding
- Collaborative curation is becoming more and more common and should be supported

References

1. Philip Morris International Research & Development, Geneva, Switzerland
2. Fraunhofer Institute for Algorithms and Scientific Computing, Schloss Birlinghoven, Sankt Augustin, Germany
3. The research described in the poster was funded by PMI Research & Development.