EVENT EXTRACTION PIPELINE

Biomedical Text Mining

- Due to the emergence of new biomedical research, there has been an exponential increase in biomedical literature.
- Researchers are increasingly unable to keep up-to-date with relevant literature, slowing down research and scientific discovery.
- Aim: automatically extract textual information.

Biomedical Event Extraction

- Task: to extract causal relationships between biomolecules from textual data.
- Aids in the discovery and understanding of the roles played by biomolecules and in turn, phenotypic outcomes such as diseases.
- In BioNLP Shared Task (BioNLP-ST) 2013, top performing systems obtained F-scores of 51% \[1,2\].
- Fundamentally difficult task; recursive events and multiple themes, causes, sites to extract.

PROBLEM & MOTIVATION

- In event extraction, the protein or gene/gene product (GGP) is the target for extraction, as well as the event type.
- Studies by Ohta et al. \cite{3} indicates that the true target for extraction is often the domain term of the GGP and not the GGP itself.
- Thus, current systems are producing potentially unusable information.
- Entity relations i.e. static relations between bio-entities, rectifies and extends the current model for biomedical event extraction (Fig 1).
- Preliminary work \cite{4} highlights the potential of entity relations in improving event extraction performance (currently low 50s F-score).
- However, integrating entity relations implies additional annotation efforts, a bottleneck in the pipeline. We investigate the potential for active learning to speed up this process.

EVENT EXTRACTION PIPELINE

- Typical pipeline approaches employ a sequence of classifiers to extract events.
- 2 focus areas:
  - Extraction of entity relations
  - Integration of entity relations into event extraction
- See Fig 2 for a comparison of the typical biomedical event extraction pipeline and the proposed pipeline.

ENTITY RELATION EXTRACTION: AN ACTIVE LEARNING APPROACH

Experimental Setup

- BioNLP-ST REL corpus, a collection of 1210 PubMed abstracts, was used.
- REL corpus contains 2 protein-entity relation types, subunit-complex and protein-component.

Baseline (Passive Learning)

- Linear SVM was trained on the whole gold-annotated training set with features extracted from all protein-entity pairs at sentence level, labeled with its relation type (none, subunit-complex or protein-component).
- The trained model was then evaluated against the development test set.

Active Learning

3 uncertainty measures were compared: random sampling, simple margin and max margin.

Observations:

- Random sampling (used in passive learning) performance increases monotonically whereas the other two active uncertainty measures’ performance increases sharply.
- 51% reduction in annotation efforts to achieve same performance as standard machine learning.
- Only 5000 instances used out of 11244.
- Peak F-score exceeds that of passive learning, reaching 63.60 compared to 63.12.

INTEGRATING ENTITY RELATIONS INTO EVENT EXTRACTION

Experimental Setup

- The GENIA Event corpus is a collection of 1224 PubMed abstracts and full text documents, annotated with the event types shown in Table 2 and entity relations extended model.
- Event extraction pipeline extensions implemented on the TEES-2.1 system \cite{2}.

Extended Features Representation

- The original TEES feature set was extended to include features from entity relations.
- Examples of features added are: string of the domain term in entity relation, type of entity relation, whether the event trigger is equal to the entity related to the protein.
- The addition of the type of entity relation to the original feature set boosted F-score by ~2 points to 56%.
- Table 2 shows a complete dissection of the improvements across the event types.

Table 2. Comparison of performance between baseline and entity relations extended model

<table>
<thead>
<tr>
<th>Event Type</th>
<th>TEES F-Score</th>
<th>Ours F-Score</th>
<th>% Ratio</th>
</tr>
</thead>
<tbody>
<tr>
<td>Gene Expression</td>
<td>77.88</td>
<td>77.90</td>
<td>100.0</td>
</tr>
<tr>
<td>Transcription</td>
<td>59.18</td>
<td>59.18</td>
<td>100.0</td>
</tr>
<tr>
<td>Protein Catabolism</td>
<td>89.80</td>
<td>91.67</td>
<td>102.1</td>
</tr>
<tr>
<td>Phosphorylation</td>
<td>77.33</td>
<td>88.50</td>
<td>114.4</td>
</tr>
<tr>
<td>Localization</td>
<td>69.70</td>
<td>77.27</td>
<td>110.8</td>
</tr>
<tr>
<td>Simple Events</td>
<td>75.32</td>
<td>76.98</td>
<td>102.2</td>
</tr>
<tr>
<td>Binding</td>
<td>43.90</td>
<td>50.00</td>
<td>113.9</td>
</tr>
<tr>
<td>Non-regulation Events</td>
<td>67.81</td>
<td>70.53</td>
<td>104.0</td>
</tr>
<tr>
<td>Regulation</td>
<td>36.36</td>
<td>37.55</td>
<td>103.3</td>
</tr>
<tr>
<td>Positive Regulation</td>
<td>44.91</td>
<td>45.81</td>
<td>102.0</td>
</tr>
<tr>
<td>Negative Regulation</td>
<td>39.51</td>
<td>39.85</td>
<td>100.9</td>
</tr>
<tr>
<td>Regulation Events</td>
<td>42.10</td>
<td>42.90</td>
<td>102.0</td>
</tr>
<tr>
<td>All Events</td>
<td>54.28</td>
<td>56.00</td>
<td>103.2</td>
</tr>
</tbody>
</table>

Table 2. Comparison of performance between baseline and entity relations extended model

KEY CONTRIBUTIONS

- We have presented the first study to integrate entity relations into the event extraction pipeline approach holistically.
- Active learning allows a 51% reduction in annotation efforts, making entity relation extraction more feasible.
- Integrating features related to entity relations increases overall event extraction performance by approximately 2% in F-score.
- Most promising is the 6% increase in Binding, which previously suffered from poor performance as it is a complex event and is difficult to extract.

REFERENCES


\cite{4} Sofie Van Landeghem, Sampo Pyysalo, Tomoko Ohta, and Yves Van de Peer. 2010. Integration of static relations to enhance event extraction from text. In Proceedings of the 2010 Workshop on Biomedical Natural Language Processing, pages 144–152. Association for Computational Linguistics.