# INTEGRATING ENTITY RELATIONS INTO THE BIOMEDICAL EVENT EXTRACTION PIPELINE

*Jessie Cheuk Ling Leung*
*Supervisor: Dr. Josiah Poon*
School of Information Technologies
FACULTY OF ENGINEERING & INFORMATION TECHNOLOGIES

## BACKGROUND

### 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. [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 [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.

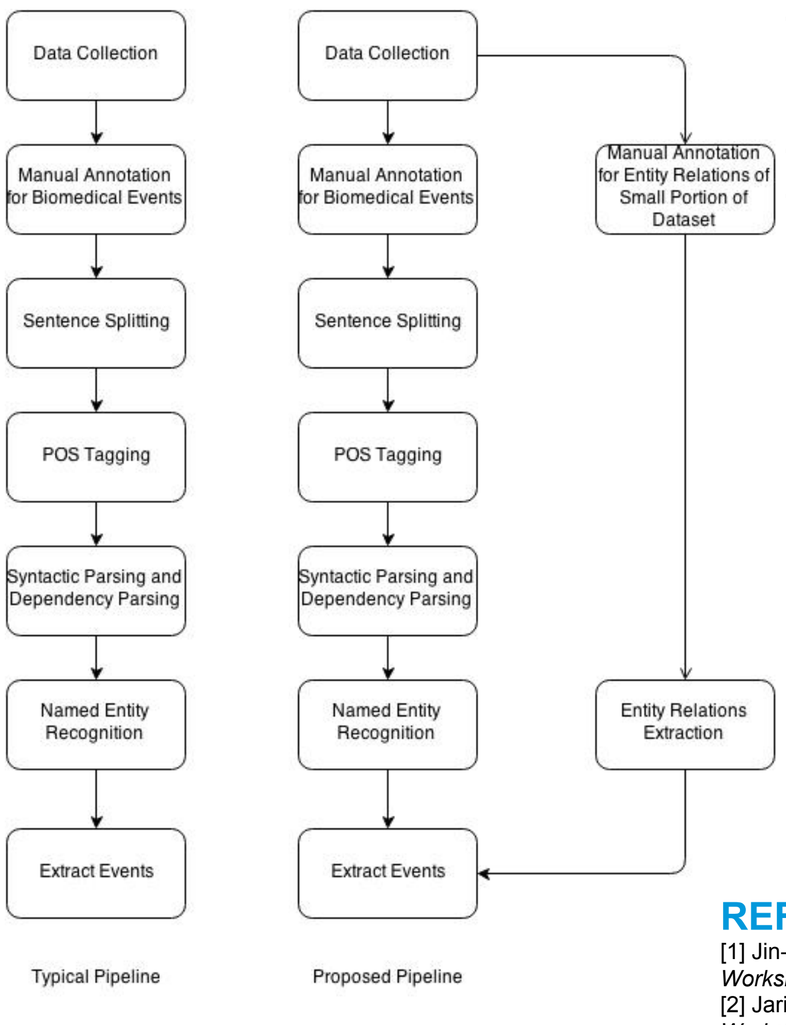


***Figure 2.*** *Traditional event extraction pipeline compared with proposed pipeline*

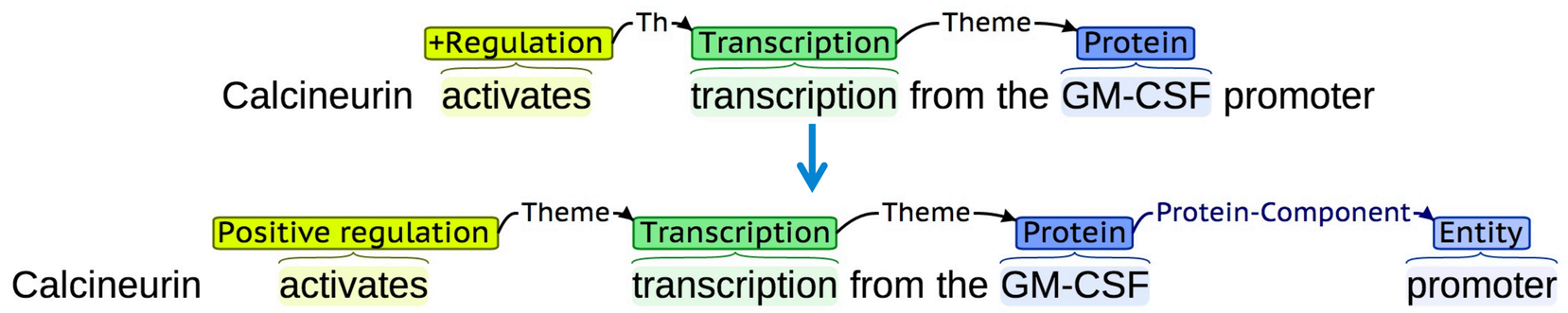


***Figure 1.*** *Top: standard event extraction model, consisting of event arguments (blue), event triggers (green). Bottom: modified event extraction model, with entity relations Subunit-Complex and Protein-Component relations added.*

- 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.

| Relation Type | F1-Score |
|---|---|
| **All Relations** | 63.12 |

***Table 1.*** *Baseline performance entity relations extraction*

### 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.

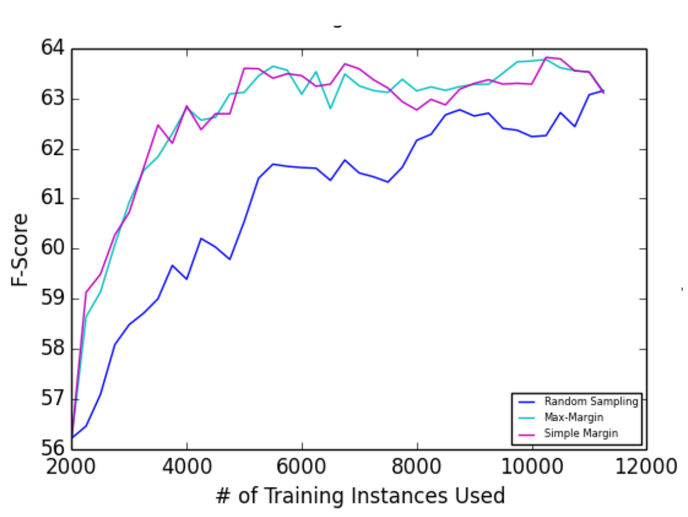


***Figure 3.*** *A comparison of 3 different uncertainty strategies. We see that for the F1 score, active learning is superior to random sampling (used in passive learning)*

## 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.
- Event extraction pipeline extensions implemented on the TEES-2.1 system [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.

| Event Type | F1-Score | | % Ratio |
|---|---|---|---|
| | TEES | Ours | |
| Gene Expression | 77.88 | 77.90 | 100.0 |
| Transcription | 59.18 | 59.18 | 100.0 |
| Protein Catabolism | 89.80 | 91.67 | 102.1 |
| Phosphorylation | 77.33 | 88.50 | 114.4 |
| Localization | 69.70 | 77.27 | 110.8 |
| **Simple Events** | **75.32** | **76.98** | **102.2** |
| Binding | 43.90 | 50.00 | 113.9 |
| **Non-regulation Events** | **67.81** | **70.53** | **104.0** |
| Regulation | 36.36 | 37.55 | 103.3 |
| Positive Regulation | 44.91 | 45.81 | 102.0 |
| Negative Regulation | 39.51 | 39.85 | 100.9 |
| **Regulation Events** | **42.10** | **42.90** | **102.0** |
| **All Events** | **54.28** | **56.00** | **103.2** |

***Table 2.*** *Comparison of performance between baseline event extraction model 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

[1] Jin-Dong Kim, Yue Wang and Yasamoto Yasunori. 2013. The Genia Event Extraction Shared Task, 2013 Edition – Overview. In *Proceedings of BioNLP Shared Task 2013 Workshop*, pages 8–14. Association for Computational Linguistics, Sofia, Bulgaria.
[2] Jari Bjorne and Tapio Salakoski. 2013. TEES 2.1: Automated annotation scheme learning in the BioNLP 2013 shared task. In *Proceedings of the BioNLP Shared Task 2013 Workshop*, pages 16–25. Association for Computational Linguistics, Sofia, Bulgaria.
[3] Tomoko Ohta, Sampo Pyysalo, Jin-Dong Kim, and Jun'ichi Tsujii. 2010. A re-evaluation of biomedical named entity-term relations. *Journal of Bioinformatics and Computational Biology*, 8(5):917–928.
[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.