Exploring a Probabilistic Earley Parser for Event Composition in Biomedical Texts

Mai-Vu Tran, Hoang-Quynh Le, Van-Thuy Phi, Thanh-Binh Pham, Nigel Collier

{vutm,lhquynh,thuypv,binhpt}@vnu.edu.vn, collier@ebi.ac.uk

INTRODUCTION

- Our system explored a multi-stage approach including trigger detection, edge detection and event composition
- We proposed a novel method for the composition of ambiguous events used a probabilistic variation of the Earley chart parsing algorithm (Stolcke 1995) for finding best derived trigger-argument candidates.
  - Using the event templates and named entity classes as grammar rules
  - Chart parsing approach incorporates a linear interpolation mechanism for cross-domain adaptivity between the training and testing (development) data

APPROACH

- The system consists of five main modules:
  - Pre-processing, Trigger detection, Edge detection, Simple event extraction, Complex event extraction
- We focus on the Cancer Genetic Task. CG Task have a large number of entity and event types: 18 entity classes, 40 types of event and 8 types of arguments.
- 40 events divided to two groups:
  - 36 simple events whose arguments must be entities
  - 4 complex events whose arguments may be other events

TRIGGER AND EDGE DETECTION

Trigger detection: the system classify whether a token acts as a trigger for one of the forty event types or not

- **Features:** Token feature, Neighbouring word feature, Word n-gram feature, Trigger dictionary feature, Pair n-gram feature, Parse tree shortest path feature

Edge detection: two classification models are T-E model and EV-EV model

- T-E model extract trigger-entity edges. This model classifies edge candidates to one of the 8 argument roles (*theme, cause, site, atloc, toloc, fromloc, instrument, participant*) and a negative argument class
- EV-EV model identifies relations in the sentences between 4 types of complex events and other events
- **Features:** Token feature, Neighbouring word feature, Word n-gram feature, Class feature, Pair n-gram feature, Parse tree shortest path feature

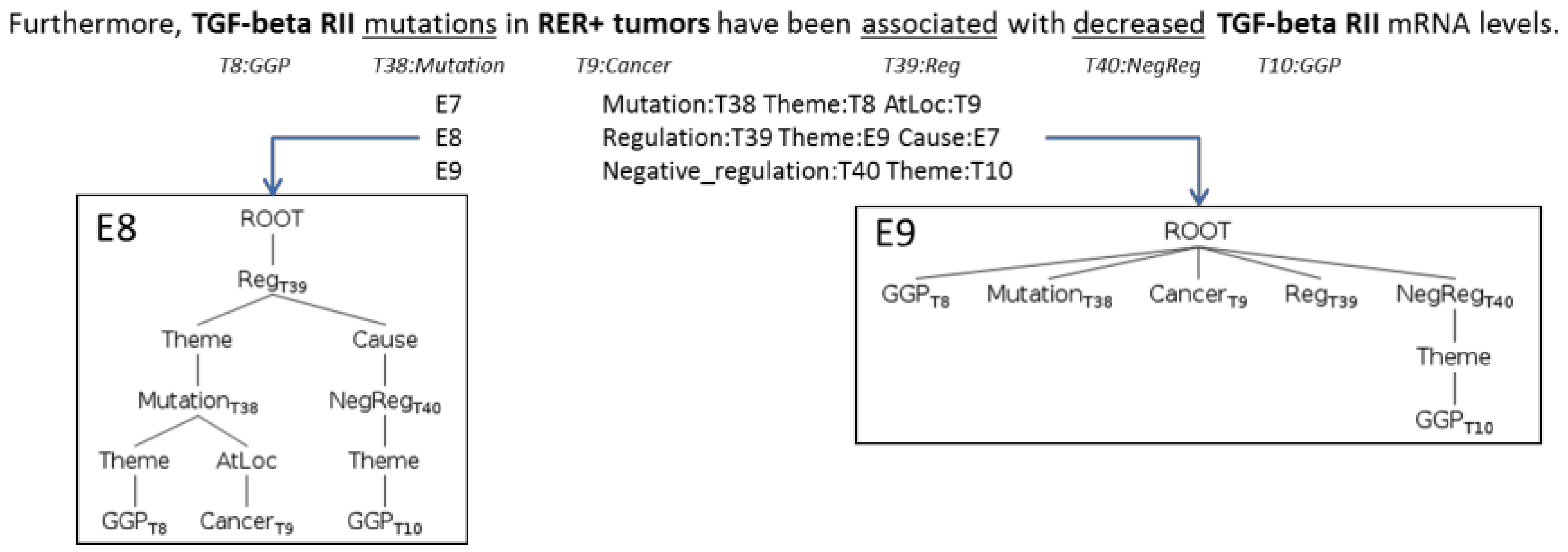
SIMPLE EVENT EXTRACTION

- Combine edge candidates identified in the T-E model into complete simple events. We had the results which belong to the 36 simple event types and relations between 4 complex events and entities
- Selecting the edge candidates use event-argument pattern based probabilities derived from the training set. An example of a *Development* event-arguments pattern:

Development* → *Theme(Gene_expression) + AtLoc(Cancer)

COMPLEX EVENT EXTRACTION

An example of two complex events as two event trees



- Build a tree for each complex event. Labels of entity classes and event types are retained while terms of triggers and entities are removed
- Using the Earley parsing algorithm (Jay Earley, 1970) to find alternative structures. To choose the best event tree candidates, we built a probabilistic Earley parser which developed from the idea of Stolcke (1995)
- The scoring function for each node is:

$$Score(\text{node}) = \frac{\sum_{edges \in node} P(\text{edge} \mid \text{argument})}{num(\text{edges})} + P_{\text{Occurrence}}(\text{arguments} \mid \text{node})$$

- num(edge): number of edges that have a link to the node
- $P_{Occurence}$(arguments|node): a distribution which represents the co-occurrence of entity/trigger labels in the arguments of an event type.
- λ is a linear interpolation parameter in the range of [0,1]
- $P_{Classifier}$(edge|argument): the probability obtained from the edge classifier.
- P_{Prior}(edge|argument): the training set's prior probability for the edge.

→The final score of an event tree candidate was calculated as ROOT's value

RESULTS AND DISCUSSION

Baseline results for event composition on the CG task development data

Event	F1	Event	F1
Development	86.67	Phosphorylation	68.45
Blood vessel development	84.15	Dephosphorylation	66.67
Growth	76.77	DNA methylation	85.71
Death	61.95	DNA demethylation	-
Cell death	53.06	Pathway	61.81
Breakdown	77.68	Localization	66.11
Cell proliferation	59.82	Binding	70.68
Cell division	100.00	Dissociation	100.00
Remodeling	60.00	Regulation	69.55
Reproduction	-	Positive regulation	68.13
Mutation	78.74	Negative regulation	68.57
Carcinogenesis	60.67	Planned process	49.99
Metastasis	74.39	Acetylation	100.00
Metabolism	62.50	Glycolysis	69.89
Synthesis	52.63	Glycosylation	-
Catabolism	59.27	Cell transformation	66.67
Gene expression	79.18	Cell differentiation	71.18
Transcription	75.00	Ubiquitination	75.00
Translation	80.00	Amino acid catabolism	100.00
Protein processing	100.00	Infection	75.86
		Total	73.67

Error classification of 50 missing false negatives

Cause	Trigger	Event
Ambiguity in event class	9	
Co-reference	6	
Do not match with any event argument patterns	7	
No training instance	7	4
Choose best argument entity in simple event extraction		5
No argument		4
No Earley parser rule		8
Total	29	21

- Shared task testing set was overall disappointing with an F-score of 29.94 (Recall = 19.66, Precision = 62.73) indicating low coverage caused by severe over-fitting issues.

CONCLUSIONS & FUTURE WORK

- Built a system based on supervised machine learning with rich features, semantic post-processing rules and the dynamic programming Earley parser
- The system achieved an F-score of 29.94 on the CG task with high precision of 62.73
- **Future work**: Focus on extending recall for complex events and looking at how we can avoid over-fitting to benefit cross-domain adaptivity