TURUN YLIOPISTO
UNIVERSITY OF TURKU

Extracting Complex Biological Events with Rich Graph-Based Feature Sets

5.6.2009 BioNLP'09

Jari Björne, Juho Heimonen, Filip Ginter, Antti Airola, Tapio Pahikkala and Tapio Salakoski

Faculty of Mathematics and Natural Sciences, IT-department

Introduction

- Three-step approach to event extraction
  - Trigger detection
  - Argument detection
  - Semantic post-processing
- Graph-based representations of both syntactic and semantic data
- Machine learning with SVMs (Joachims SVM^{Multiclass})

Graph Representation

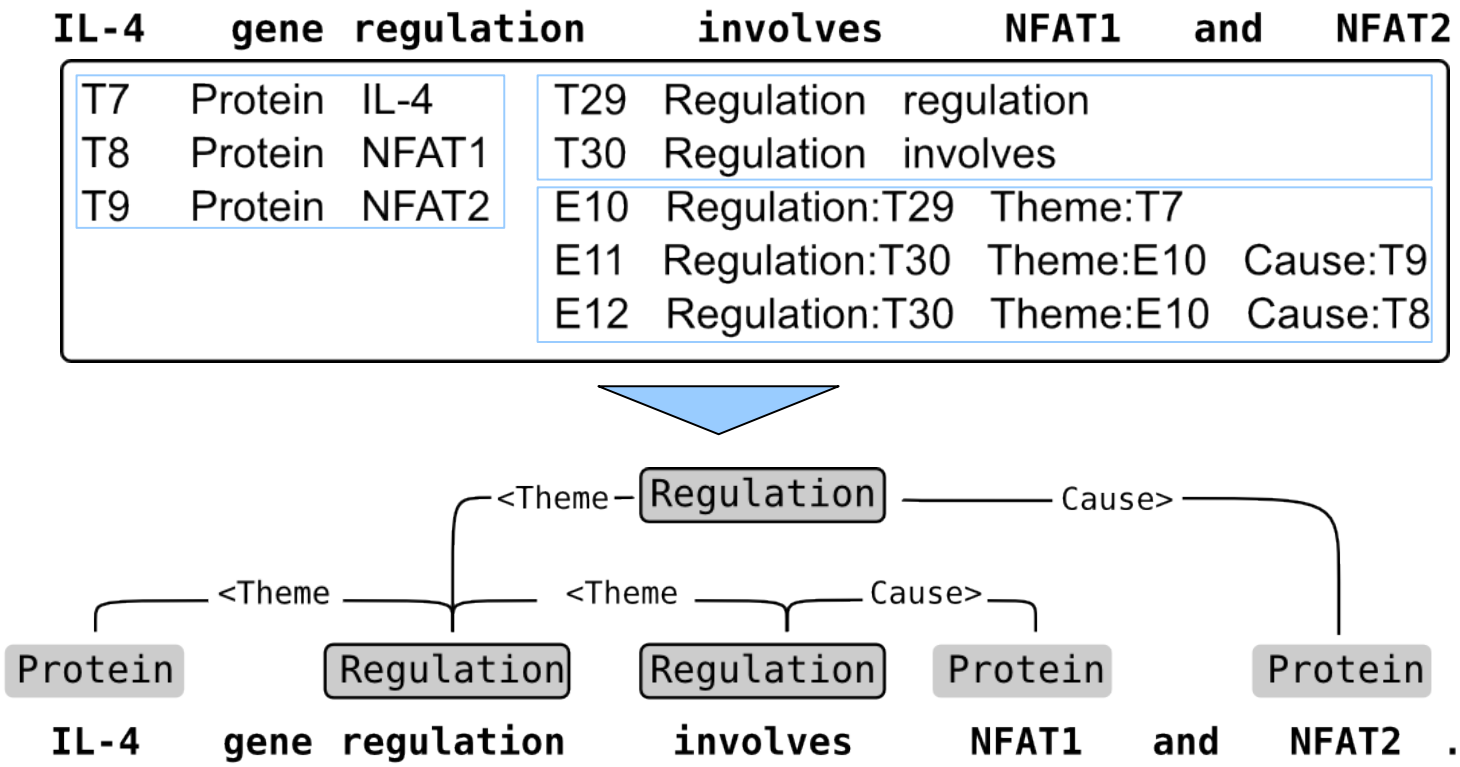
IL-4 gene regulation involves NFAT1 and NFAT2 .

T7	Protein	IL-4
T8	Protein	NFAT1
T9	Protein	NFAT2

T29	Regulation	regulation		
T30	Regulation	involves		
E10	Regulation:T29	Theme:T7		
E11	Regulation:T30	Theme:E10	Cause:T9	
E12	Regulation:T30	Theme:E10	Cause:T8	

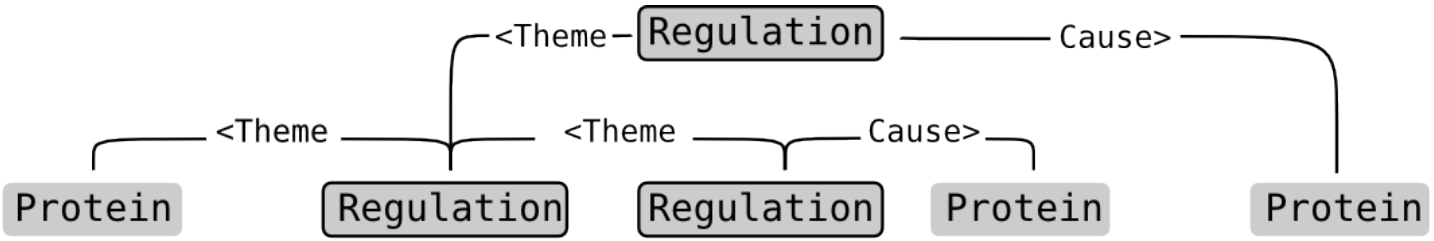
Event Annotation

Graph Representation



- Semantic network has one-to-one correspondence to task annotation

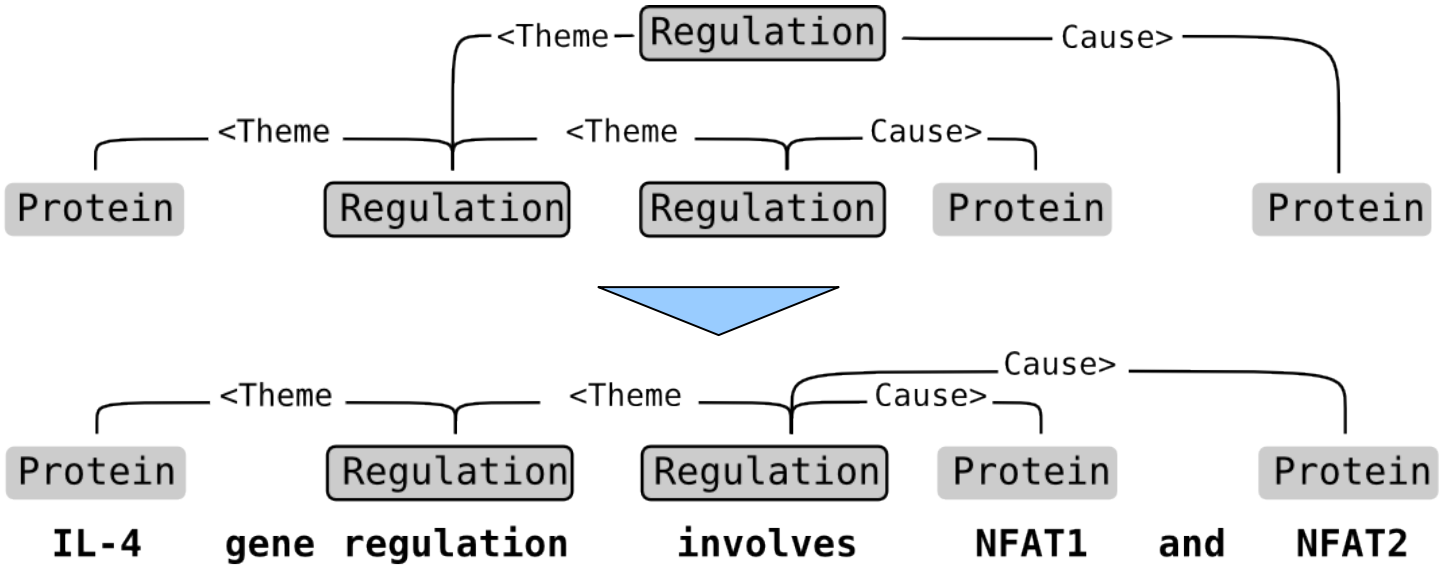
Graph Representation



Semantic Network

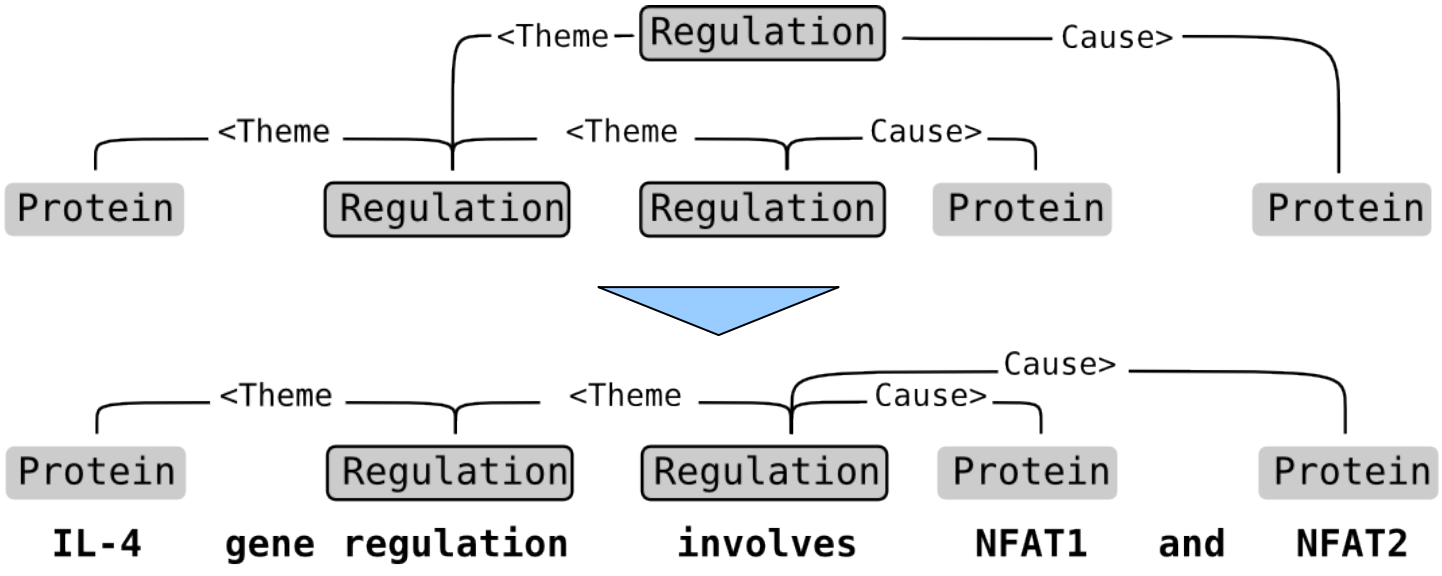
- Overlapping nodes are discarded → one potential node per word token

Graph Representation



- Overlapping nodes are discarded → one potential node per word token

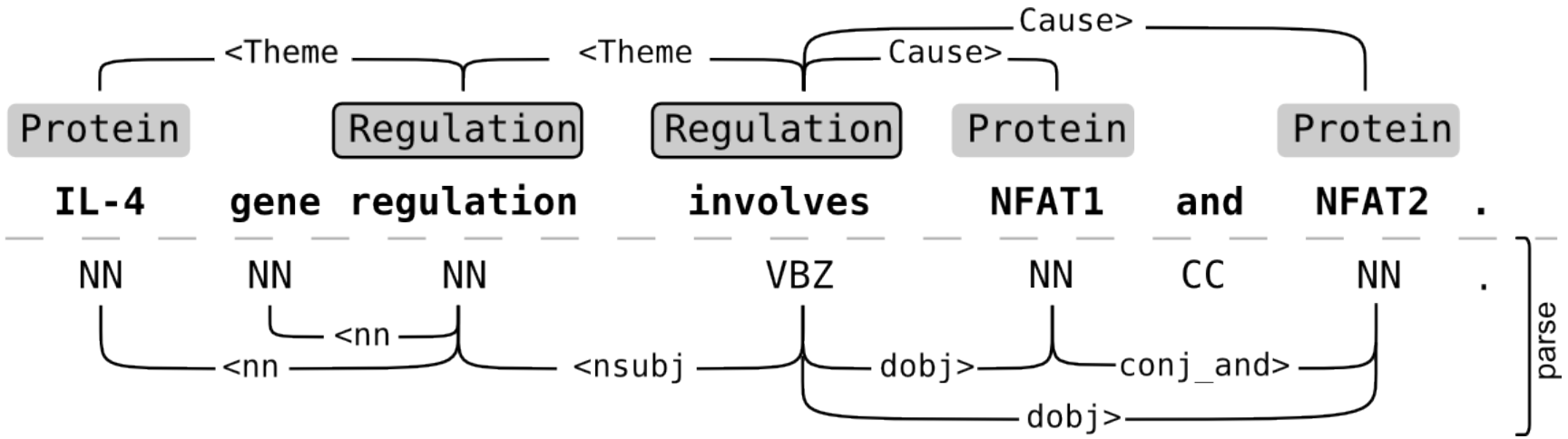
Graph Representation



- Overlapping nodes are discarded → one potential node per word token
- Flat graph is *extraction target*
- Semantic post-processing reduplicates nodes

Dependency Parses

- Collapsed Stanford format, McClosky-Charniak parser
- >45% of event arguments are separated by a single dependency (*shortest path*)

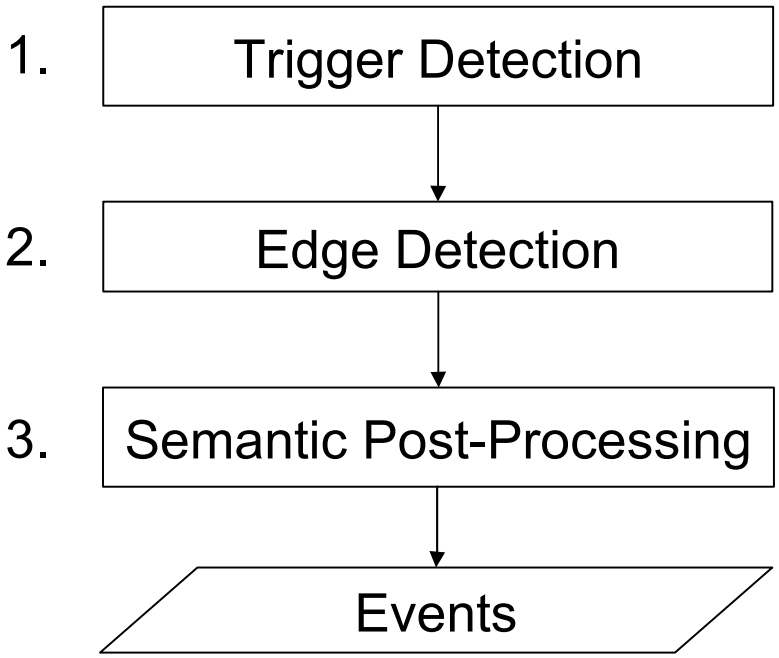


TURUN YLIOPISTO
UNIVERSITY OF TURKU

Preparing the Data

- We process one sentence at a time
- Events between sentences are discarded
- 95 % of all annotated events are within one sentence

Extraction Process



Trigger Detection

- Trigger type is predicted per token

Protein					Protein		Protein	
IL-4	gene	regulation	involves		NFAT1	and	NFAT2	.

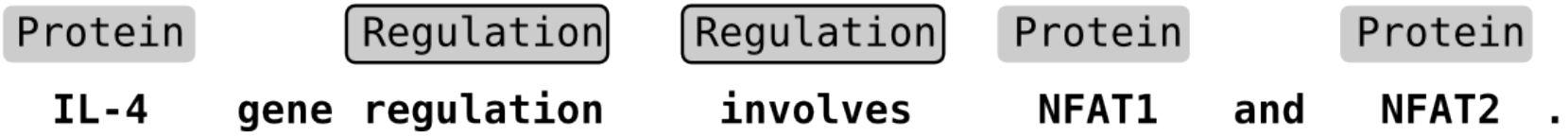
Trigger Detection

- Trigger type is predicted per token

Protein	Neg	Regulation	Regulation	Protein	Neg	Protein	
IL-4	gene	regulation	involves	NFAT1	and	NFAT2	.

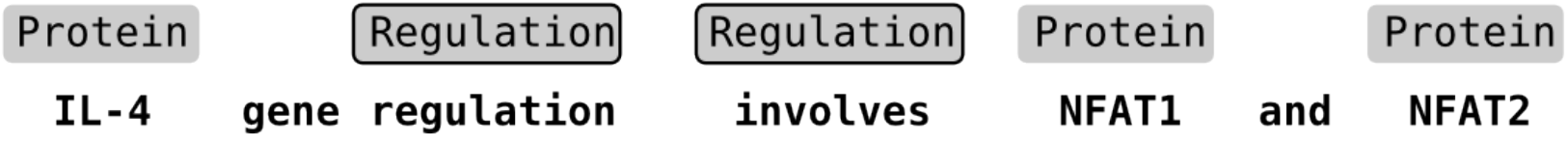
Trigger Detection

- Trigger type is predicted per token
- Trigger nodes are formed based on token predictions



Trigger Detection (details)

- Adjacent triggers with same type are merged, if merged string has been seen in training data (not in the example shown)
- Overlapping triggers of *different types* can be predicted with merged type classes
- 9 trigger types → multi-class classification

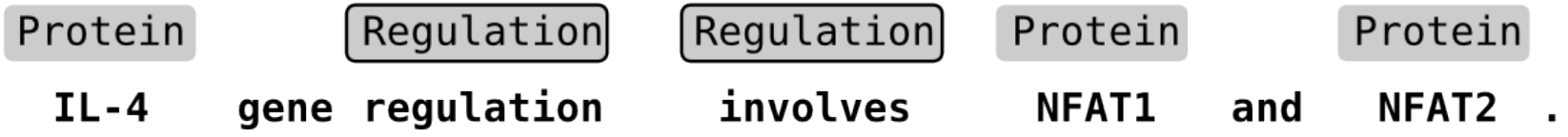


Trigger Detection Features

- Token features
  - Character *n*-grams, stem, heuristics
- Frequency features
  - Number of entities, bag-of-word counts
- Dependency *N*-grams
  - Undirected chain of dependencies and tokens
  - Up to depth of three

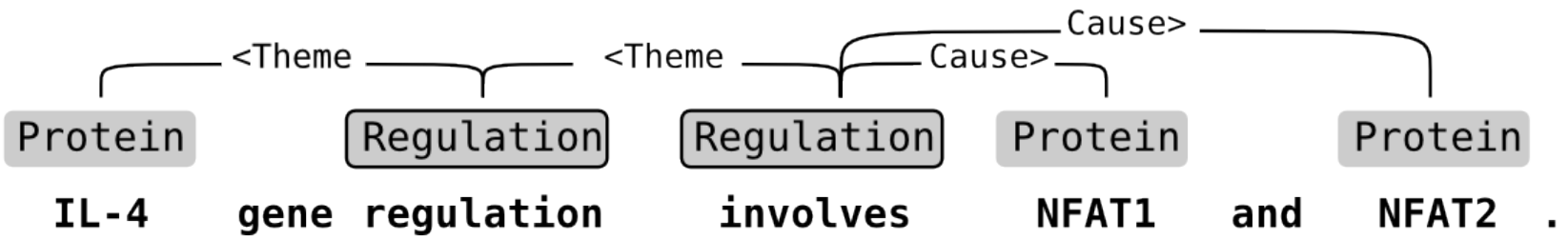
Edge Detection

- Edges are predicted between named entities and predicted triggers



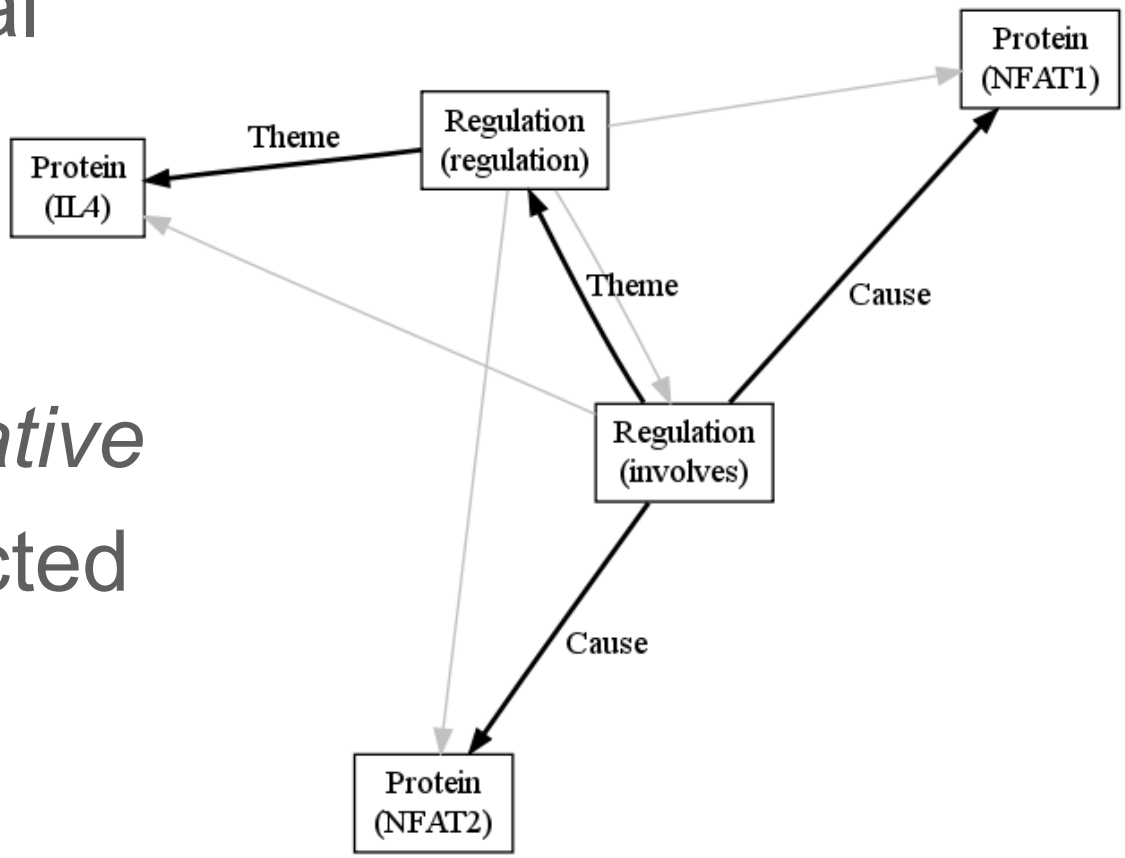
Edge Detection

- Edges are predicted between named entities and predicted triggers
- Result is a flattened event graph



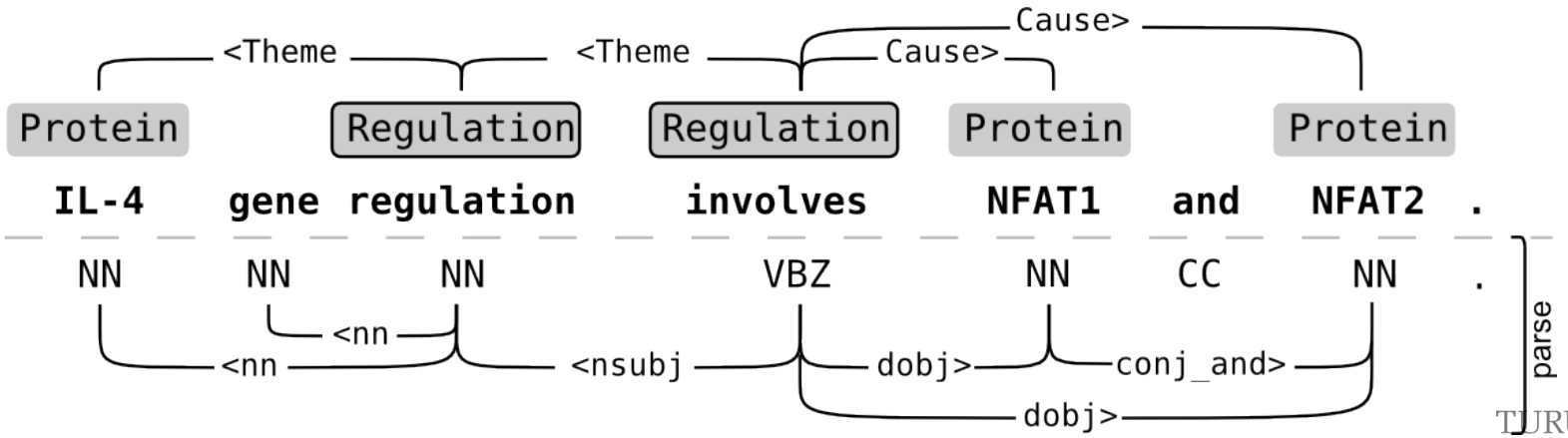
Edge Detection

- Several potential edges between entities
- Classes *theme*, *cause* and *negative*
- All edges predicted independently



Edge Detection Features

- Mostly based on the *shortest path of dependencies*
- Training data for edge detector
  - 31 792 examples
  - 295 034 unique features



TURUN YLIOPISTO
UNIVERSITY OF TURKU

Edge Detection Features

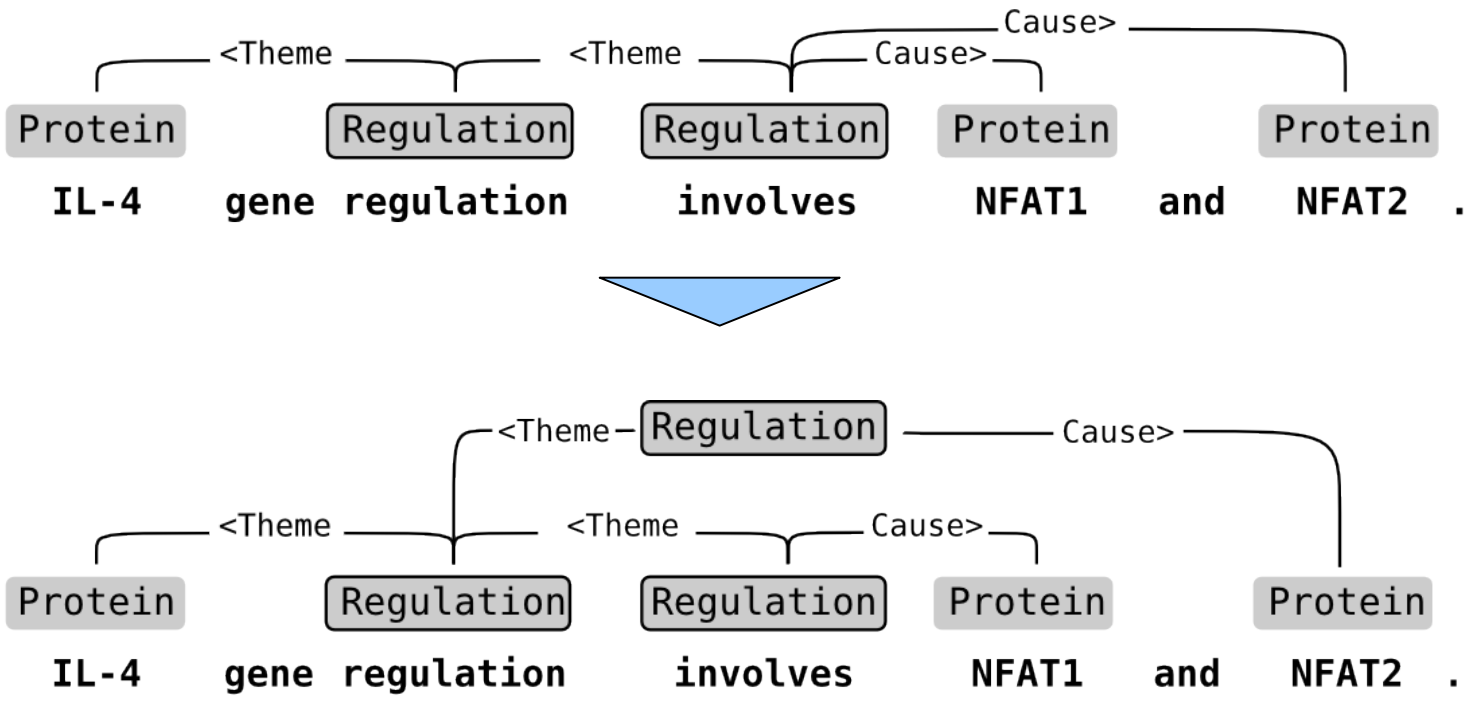
- Dependency *N*-grams
  - 2-4 consecutive dependencies and tokens
- Semantic node features
  - Built from the end nodes of the potential edge
- Frequency features
  - Length of shortest path, number of entities and triggers in sentence

Semantic Post-processing

- Shared task restricts event arguments
  - Remove invalid edges from graph
- Predicted graph contains max one node per word token, per event type
  - Duplicate trigger nodes for overlapping events
- Convert graph to shared task format
- Rule-based system

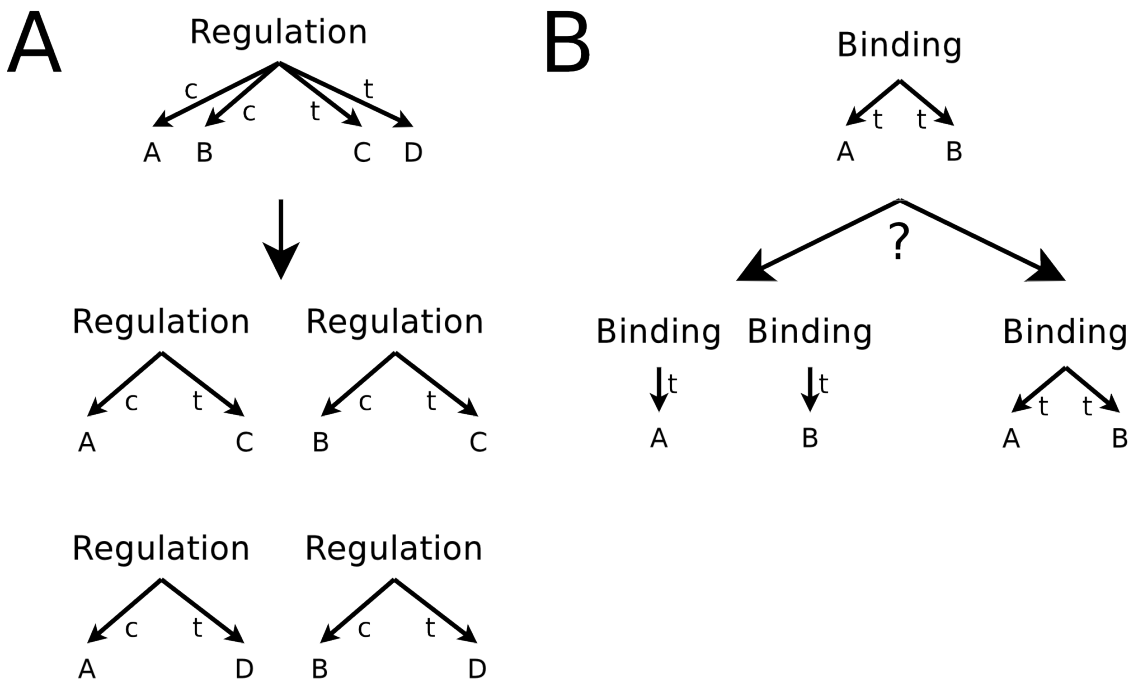
Semantic Post-processing

- To recover events, some semantic network nodes need to be duplicated



Semantic Post-processing

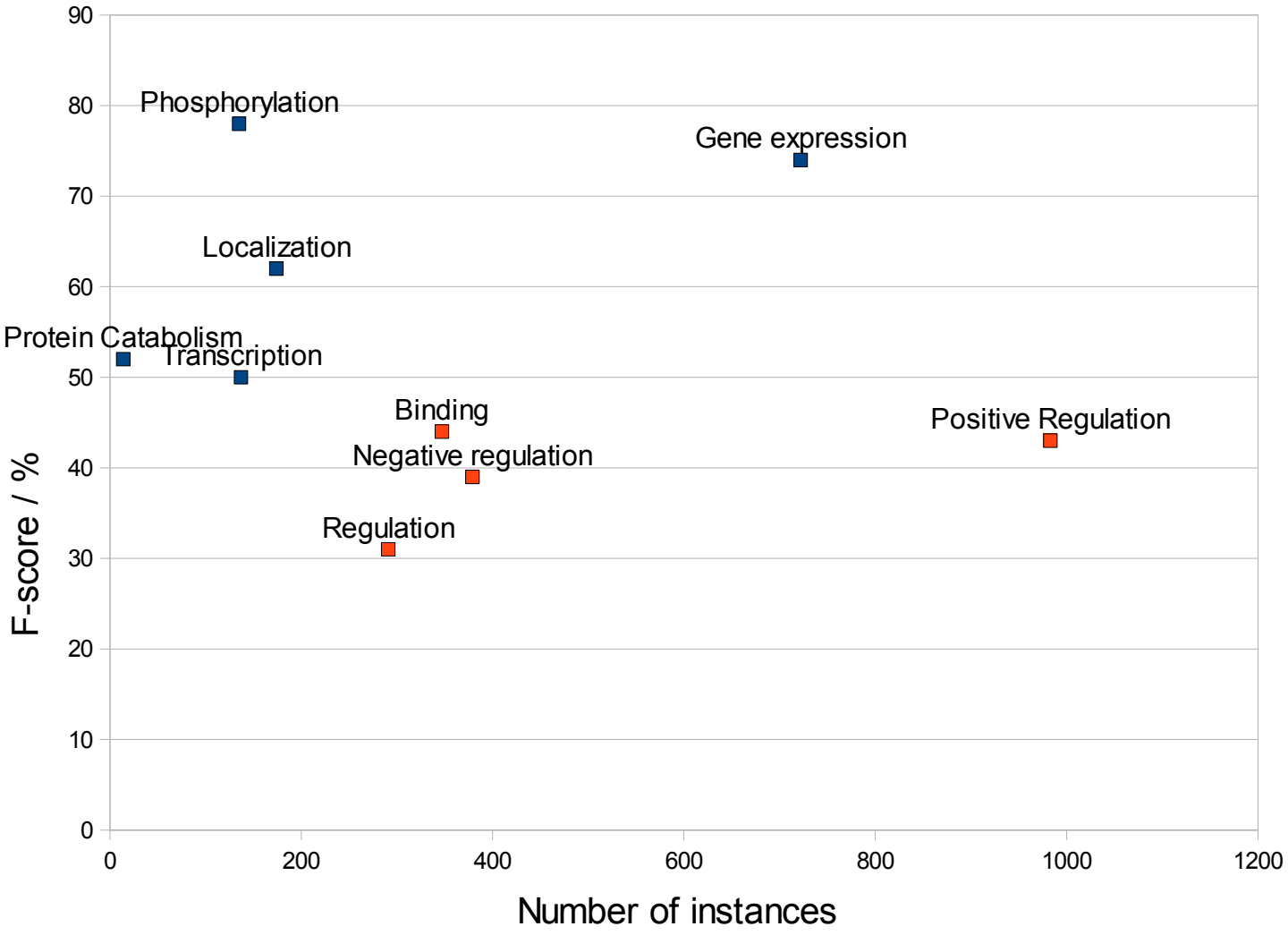
- Graph processing based on trigger node type



Results

- Approximate Span & Recursive **51.95** % (primary measure of task 1)
- Approximate Span **51.72** %
  - Only a few nesting events
- Strict **47.41** %
  - Trigger spans explain most of the difference vs. the primary measure

Per-class Results



Effect of Components

- Predictions (pred) of a single component at a time are replaced with gold-standard (GS) data
- Shows impact of component on overall performance

Triggers	Edges	Post-processing	F	ΔF
pred	pred	pred	53.50	
GS	pred	pred	72.08	18.58
GS	GS	pred	94.69	22.61
GS	GS	GS	100	5.31

Alternative Directions

- Several attempts to relax independence assumptions
  - Graph reranking for argument edges
  - Structural SVM with Hidden Markov models for trigger detection
- Coreference detection for 4,8% of events crossing sentence boundaries (machine learning)

Conclusions

- Splitting the task into subproblems
- Careful feature engineering
- Thorough optimization of parameters for each subtask
- Program to be published under open source license

Thank You!

- BioNLP'09 Shared Task team
- BioNLP'09 and NAACL organizers
- Academy of Finland
- CSC – IT Center for Science Ltd.