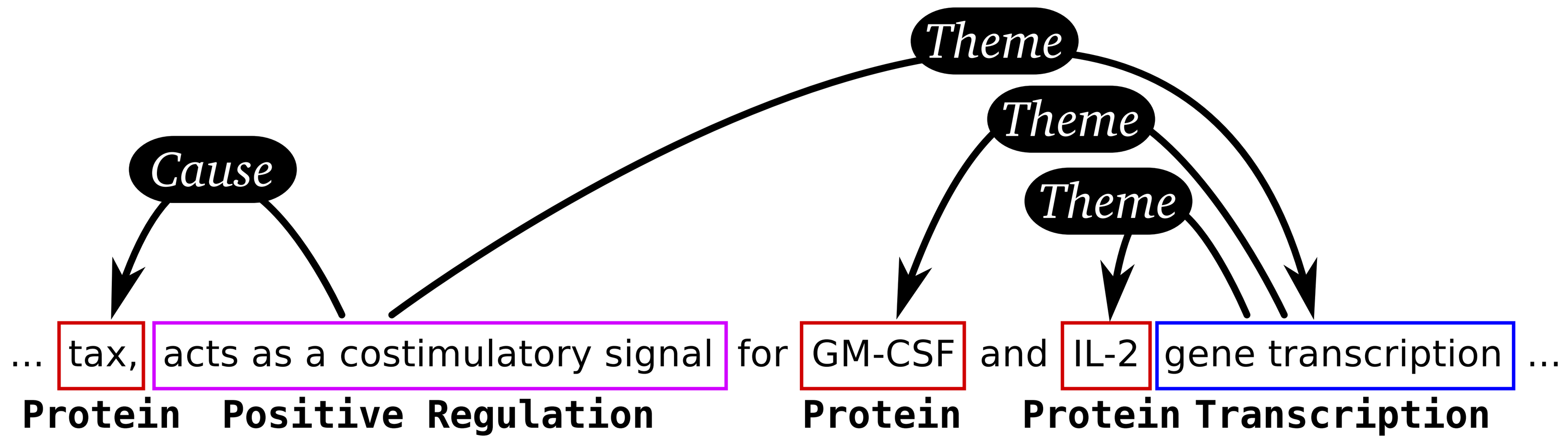
Event Extraction as Dependency Parsing

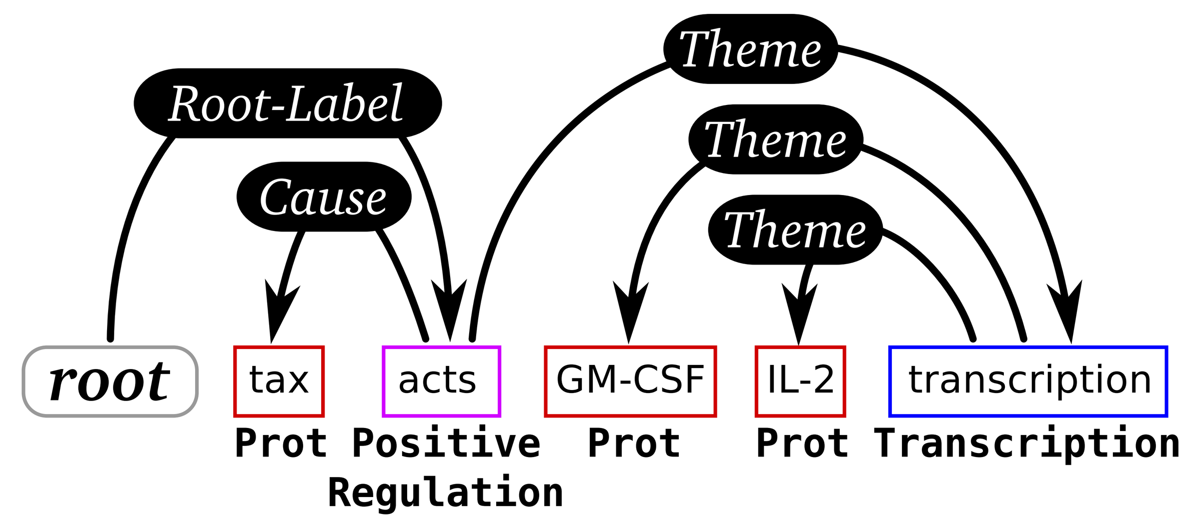
David McClosky, Mihai Surdeanu, and Christopher D. Manning
Stanford University
{mcclosky,mihais,manning}@stanford.edu

Motivation



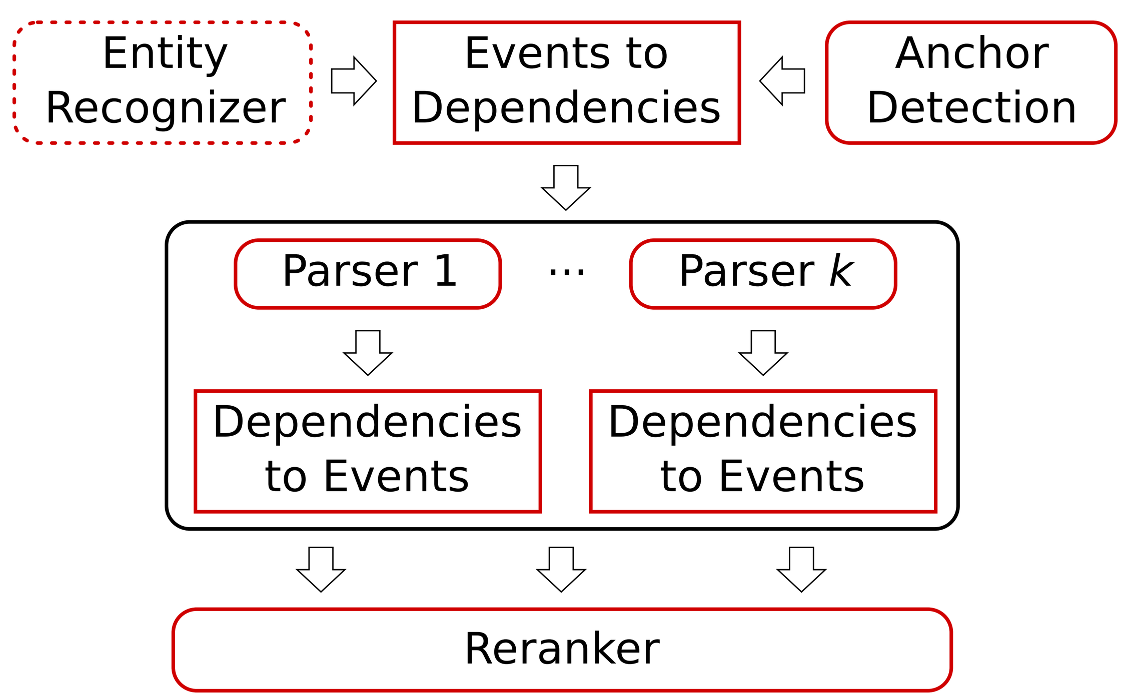
Leading event extraction systems use pipelines which ignore joint information. Casting event extraction as a parsing problem allows use of standard parsing tools which can balance decisions more globally.

Approach



Above event structures converted to dependencies

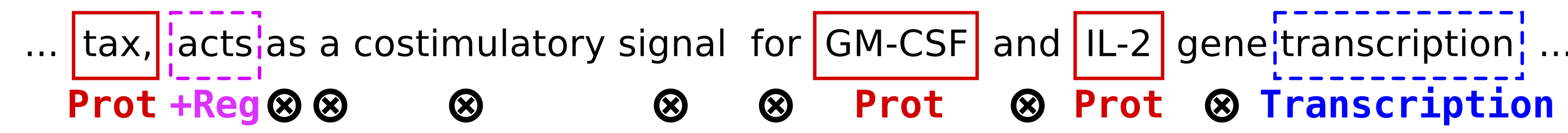
Convert event structures to dependency graphs, use reranking parser to predict event structure.



System architecture for training the event extractor

Multiple n-best parsers are combined inside a reranking framework. Dashed components are given as input, square borders indicate domain dependence.

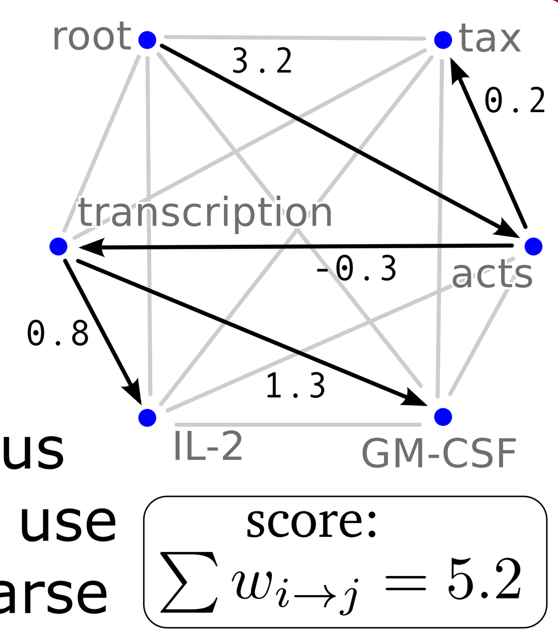
Anchor Detection



- Cast as token classification problem (9 event types)
- Multiword anchors reduced to their syntactic head (shown as dashed boxes)
- Features include word and lemma of nearby words, syntactic context, gazetteer, and nearby entities
- Token-level logistic regression applied
- Skewed towards high recall (parser can ignore extraneous event anchors)

Parser

- Dependency graph: nodes are words, edges are semantic relations (Theme, Cause, etc.)
- Use MSTParser to handle n-best non-projective parses
- Features: original MST features plus new event-specific features which use "full" sentence and its syntactic parse
- Multiple decoders available, capture different views (e.g. 2P = second order edge features, projective)

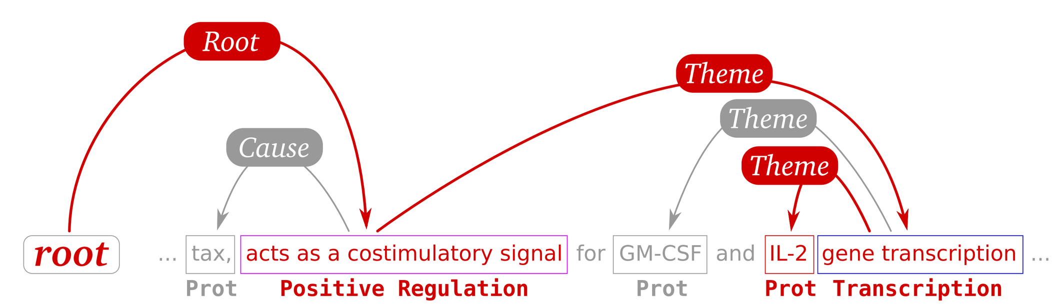


$$\sum w_{i \to j} = 5.2$$

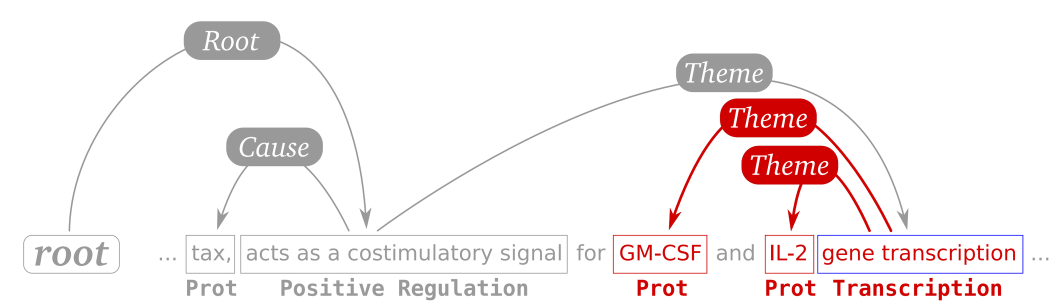
Reranker

- Rerank n-best list by rescoring each event structure
- Global features: **Decoder score**, **# decoders**

Path to root (length, words, types, relations, etc.)



Event frames (children types, # children, relations, etc.)



Results

AD	Parse	RR	Conv	R	P	F1
√	√		√	45.9	61.8	52.7
√	√	√	√	48.7	59.3	53.5
G	√		√	68.9	77.1	72.7
G	√	√	√	68.5	78.2	73.1
G	G	G	√	81.6	93.4	87.1

Ablation study between system components

G indicates gold, √ indicates component was used. Current system, Parse and Reranker (RR), perform well with predicted triggers (53.5%) and gold (73.1%).

		n-best parses considered			
Anchors	Decoder(s)	1	2	10	All
Gold	1P	70.7	76.6	84.0	85.7
	2P	71.8	77.5	84.8	86.2
	1P, 2P, 2N	—	—	—	**86.7**
Predicted	1P	52.0	60.3	69.9	72.5
	2P	52.7	60.7	70.1	72.5
	1P, 2P, 2N	—	—	—	**73.4**

Oracle reranking scores for different decoder(s)

n-best lists for gold anchors contain events almost as good as possible given our conversion limits (87.1%), multiple decoders get even closer.

Conclusions

- **Final system performs competitively** (48.6% BioNLP f-score on final test set)
- Would have been **2nd place** in BioNLP 2009. Best current results are 53.3% (Miwa et al., 2010) and 52.0% (Björne et al., 2009).
- **Joint inference** and **global features** in reranker can capture more structure than standard pipeline approaches.
- **Minimal domain-specific tuning:** almost no domain dependent features were used thus easily adaptable.