EPE 2017:
Shared Task Results Summary

September 20, 2017
Overall results (average)

1. Stanford-Paris: 60.51
2. Szeged: 58.57
3. Paris-Stanford: 56.81

Best overall system for Event Extraction and Negation Resolution: Stanford-Paris

Best overall system for Opinion Analysis: Szeged
System comparisons

- Systems vary along several dimensions
  - parser
  - dependency representation
  - pre-processing
  - training data

- Comparisons are difficult

- Focus on comparisons of the same system with variation of only one dimension
Variety of different dependency schemes:
  - syntactic: CoNLL, SSyntS, Stanford, UD
  - semantic: CCD, DM, DSyntS, PAS, PredArg
Dependency Schemes

- Variety of different dependency schemes:
  - syntactic: CoNLL, SSyntS, Stanford, UD
  - semantic: CCD, DM, DSyntS, PAS, PredArg
- Function-based vs content-based:
  - No system contrast these
  - CoNLL overall best for Opinion Analysis subtask (even with simple parser)
  - Stanford basic overall best for Event Extraction
  - UD enhanced overall best for Negation resolution
- UD enhanced better than UD basic across all three downstream tasks, given large training set, (Stanford-Paris)
Dependency Schemes

- Semantic representation (DM) performs better than syntactic for Negation Resolution (Paris-Stanford)
- CCD better than DM for Negation Resolution (Peking)
- Intrinsic evaluation correlates with extrinsic (Peking, Paris-Stanford)

Remaining teams rely on the preprocessing supplied by the task organizers.

Prague contrast the two different types of preprocessing.

Clear difference in all three downstream tasks by varying the preprocessing strategy.
Training data vary in size and domain
  ▶ UD Treebank only: 200,000 tokens
  ▶ WSJ, Brown, Genia: 1,7 mill tokens

Stanford and Paris systems systematically vary the data sets used for the training of their parsers

best performance across all three subtasks is obtained with the larger data set
Looking ahead

- We provide all software, data, submissions, and results for public download, in the hope of continued community-driven work in this direction.
- Follow-up experimentation should seek to isolate some of the interacting factors that make interpretation of EPE 2017 results across teams challenging.