

# Analysis and Repair of Name Tagger Errors

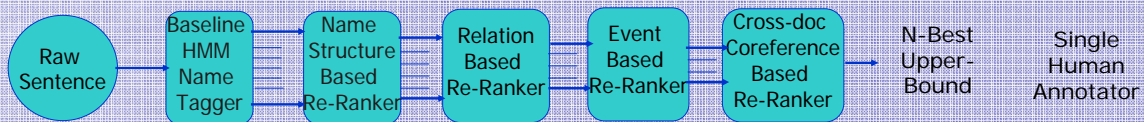
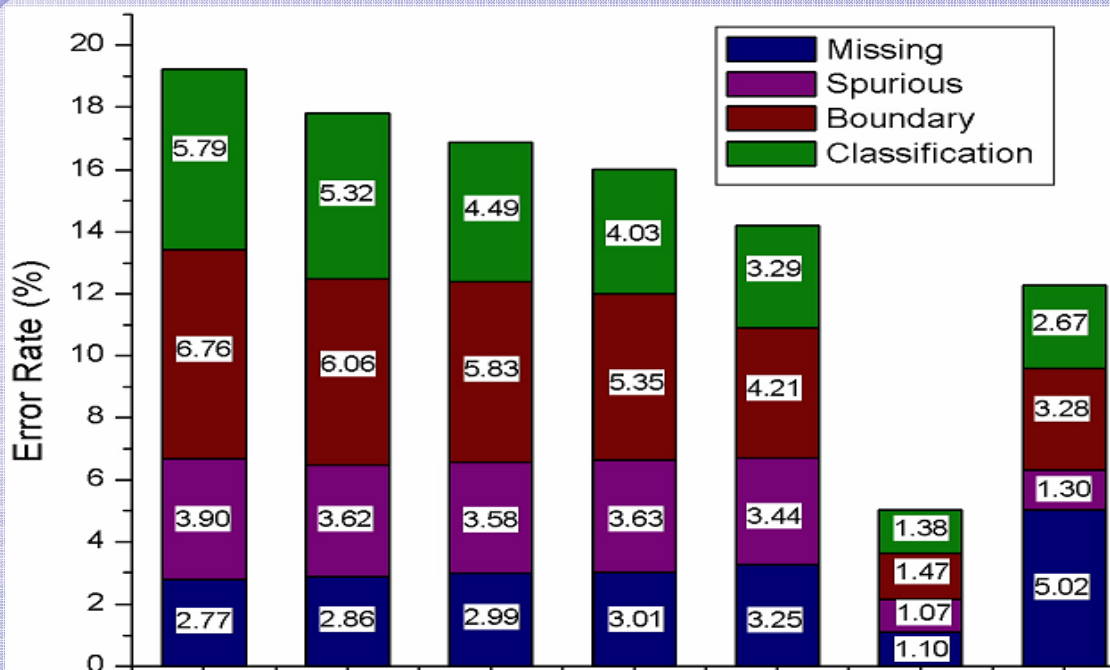
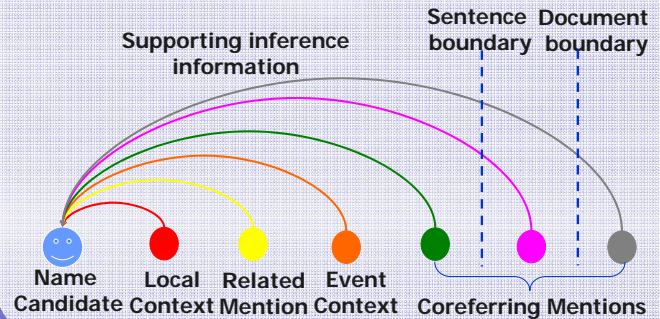
Heng Ji and Ralph Grishman (New York University)

*This work is supported by DARPA and NSF*

## Outline

- What is behind "F-Measure"? – Decompose name errors into identification error (missing, spurious, boundary) and classification error
- Repair errors by global re-ranking of N-Best Hypotheses
- Check remaining errors by comparing with human

## Re-Ranking Features from Joint Inference



Test on 100 ACE texts (2813 names: 1126 Persons, 712 GPEs, 785 Organizations and 190 Locations)

## Performance on Chinese Name Tagging

Model	Identification			Classification	Overall		
	P	R	F	Accuracy	P	R	F
Baseline	93.2	93.4	93.3	93.8	87.4	87.6	87.5
+name structure	94.0	93.5	93.7	94.3	88.7	88.2	88.4
+relation	93.9	93.7	93.8	95.2	89.4	89.2	89.3
+event	94.1	93.8	93.9	95.7	90.1	89.8	89.9
+cross-doc coreference	95.1	93.9	94.5	96.5	91.8	90.6	91.2

## Conclusion and Remaining Error Types

- Global re-ranking features help to reduce classification and boundary errors
- Main human annotation errors were the names overlooked by one annotator
- Gain from events was limited by event pattern frequency → extend event
- 70% of remaining spurious errors were 'other' name types → explicit model
- Unsupervised learning to boost cross-doc coref re-ranking