Using Discourse Structure Improves Machine Translation Evaluation

Francisco Guzmán, Shafiq Joty, Lluís Màrquez and **Preslav Nakov**
Qatar Computing Research Institute

معهد قطر لبحوث الحوسبة
Qatar Computing Research Institute
Member of Qatar Foundation عضو في مؤسسة قطر

Discourse for MT Evaluation

- Hypothesis**: discourse structure can help MT evaluation**
- Discourse is an important information source:
  - Complements lexical, POS, syntax, SRL, etc., info
  - Improves many existing metrics

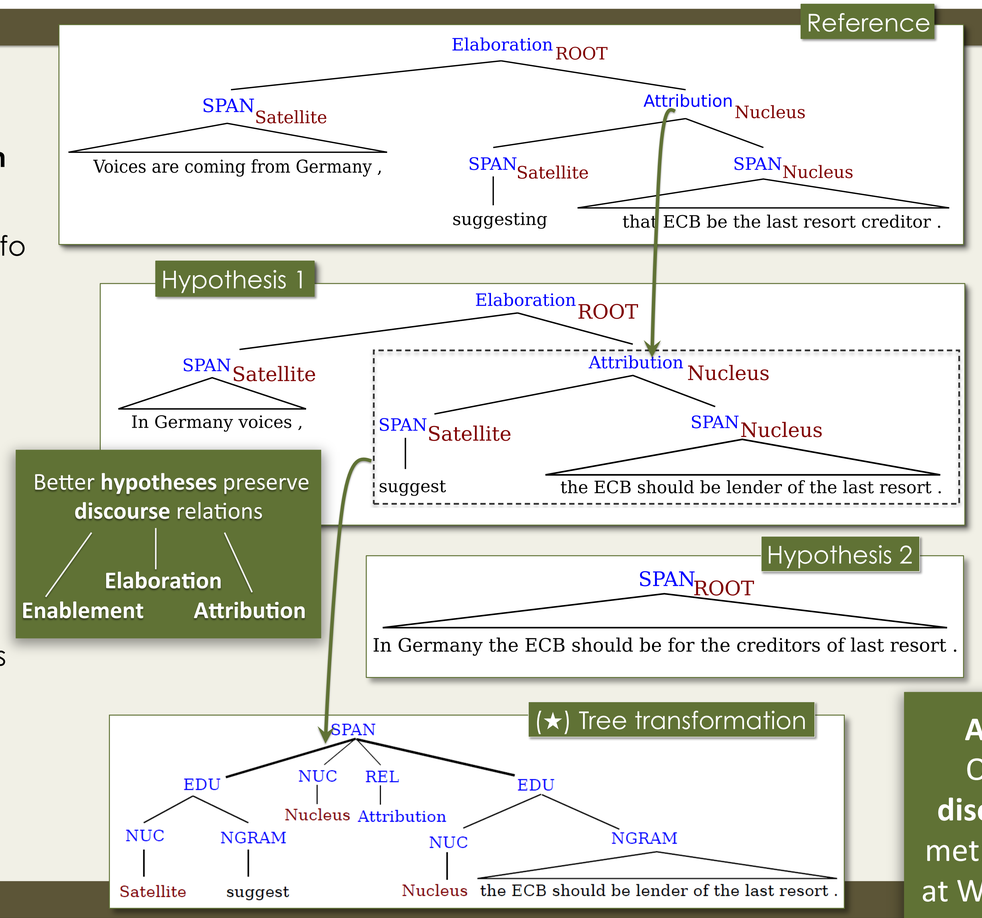
Method

Compute discourse similarity between *Hyp* and *Ref*

- RST-parse *Hyp* and *Ref* (Joty et al., 2012)
- Transform the discourse trees (★)
- Compute similarity with a syntactic tree kernel (Collins & Duffy, 2002)
  - Use this similarity as a segment-level score
  - For system-level, average segment level scores

Combine discourse similarity with existing metrics

- Uniform linear interpolation
- Tuned (MaxEnt pairwise learning)



Average Improvements

System-level	Segment-level		
WMT12	WMT12	WMT12 (cv-tuned)	WMT11 (WMT12 tuned)
+.035	**+.026**	**+.057**	**+.061**

Conclusion

- Using discourse improves MT evaluation
- Extension of this work yielded the **best scoring metrics at WMT14!**

- Future work:
  - Go beyond the sentence-level
  - Use discourse-based measures for machine translation

Application: Our DiscoTK **discourse**-based metrics **ranked 1st** at WMT14 Metrics Evaluation task!

Results