Using Discourse Structure Improves Machine Translation Evaluation
Francisco Guzmán, Shafiq Joty, Lluís Màrquez and Preslav Nakov
Qatar Computing Research Institute

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 (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)

Results

Average Improvements
<table>
<thead>
<tr>
<th></th>
<th>WMT12</th>
<th>WMT12 (cv-tuned)</th>
<th>WMT11 (WMT12 tuned)</th>
</tr>
</thead>
<tbody>
<tr>
<td>System-level</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>(+DR−LEX)</td>
<td>.035</td>
<td>.026</td>
<td>.057</td>
</tr>
<tr>
<td>Segment-level</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>(+DR−LEX)</td>
<td>.035</td>
<td>.026</td>
<td>.057</td>
</tr>
</tbody>
</table>

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

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

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

System-level
- WMT11 system-level results
- WMT12 system-level results
- WMT12 segment-level results
- WMT11 segment-level results

Segment-level
- XEnErrCats
- WordBlockEC
- BlockErrCats
- posF
- BLEU
- TER
- SIMPBLEU
- TerrorCat
- Meteor
- AMBER

Metric
- Spearman’s Rho
- Kendall’s Tau

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

Reference
Hypothesis 1
Better hypotheses preserve discourse relations
Elaboration
Enablement
Attribution

Hypothesis 2
In Germany voices, the ECB should be the last resort creditor.

[Diagram showing discourse structure]