Learning to Differentiate Better from Worse Translations
Francisco Guzmán, Shafiq Joty, Lluís Márquez, Alessandro Moschitti, Preslav Nakov and Massimo Nicosia
Qatar Computing Research Institute

1. Task Formulation
• Decide which of two alternative translations \( t_1 \) and \( t_2 \) is better given the reference \( r \)
• Motivation: Higher correlation with human judgments compared to absolute scores (Duh, 2008; Song & Cohn, 2011)

2. Proposed Solution
• Use the framework of structured kernel learning (Severyn & Moschitti, 2012)
  • Pairwise learning to rank formulation with kernels
  • Is more powerful than kernel similarity (Guzmán et al., 2014)
  • Learns features (structure fragments) automatically
  • Allows integrating several information sources
• Integrate lexical, syntactic, and discourse information in a single structural representation
• Use both reference and system output simultaneously
• Learning object: \( \langle t_1, t_2, r \rangle \)

3. Enriched Structural Representation

4. Structured Kernel Learning

5. Experimental Settings
• Train: 10K judgments per language (WMT-11)
• Eval: Kendall’s Tau as a measure of correlation on WMT-12 data (official)
• Results are compared with direct kernel similarity

6. Evaluation Results
Train & Test for each language pair separately on different structures

<table>
<thead>
<tr>
<th>Structure</th>
<th>Similarity</th>
<th>Structured Kernel Learning</th>
</tr>
</thead>
<tbody>
<tr>
<td>1 SYN</td>
<td>0.169</td>
<td>0.188</td>
</tr>
<tr>
<td>2 DIS</td>
<td>0.130</td>
<td>0.174</td>
</tr>
<tr>
<td>3 DIS+POS</td>
<td>0.135</td>
<td>0.186</td>
</tr>
<tr>
<td>4 DIS+SYN</td>
<td>0.116</td>
<td>0.205</td>
</tr>
</tbody>
</table>

SYN (syntactic parse), DIS (RST discourse parse relations), POS (part of speech)

Observations
• Learning with structural kernels works better than using simple kernel similarity ⇒ new features are learned
• Shallow syntax and discourse yield similar improvement individually
• Combining them yields further improvement
• We outperform popular metrics like TER (0.217), NIST (0.214) and BLEU (0.185)

7. Conclusion
• Unified framework for integrating layers of linguistic information for MT evaluation
• Pairwise learning-to-rank with structural kernels
• Competitive performance

8. Future Work
• More linguistic information: SRL, Brown clusters, etc.
• Integrate scores from other MT evaluation metrics
• Use of more relations between \( t \) and \( r \)

Acknowledgments
This research is part of the Interactive Systems for Answer Search (Iyass) project, conducted by the Arabic Language Technologies (ALT) group at Qatar Computing Research Institute (QCRI) within the Qatar Foundation