**Pairwise MT Evaluation**

- Learn to differentiate better from worse translations
- State-of-the-art: structured input and preference-kernel learning ([Guzmán et al., EMNLP 2014](#))
- Inspired by human ranking-based MT evaluation.

**Why Neural Networks?**

- State-of-the-art uses computationally expensive tree kernels (esp. at test time). NNS provide fast inference
- NNS can learn effectively from compact semantic and syntactic distributed representations
- They are highly competitive

**Learning Task**

- Binary classification: \( y = \begin{cases} 1 & \text{if } t_1 \text{ is better than } t_2 \text{ given } r \\ 0 & \text{if } t_1 \text{ is worse than } t_2 \text{ given } r \end{cases} \)
- Model:

  \[ p(y|t_1, t_2, r) = \text{Ber}(y|f(t_1, t_2, r)) \]

  \[ \hat{y} \theta = f(t_1, t_2, r) = \text{sig}(w^T \phi(t_1, t_2, r) + b) \]

- Cost function:
  - Negative log-likelihood: \( J_\theta = - \sum_{i} y_i \log \hat{y}_\theta + (1 - y_i) \log (1 - \hat{y}_\theta) \)
  - Kendall’s-tau:

  \[ J_\theta = - \sum_{i} y_i \text{sig}(\gamma \Delta_i) + (1 - y_i) \text{sig}(\gamma \Delta_i) \]

  \[ \Delta = f(t_1, t_2, r) - f(t_2, t_1, r) \]

**Features**

- Pairwise lexical features: BLEU, METEOR, NIST, TER
- Word embeddings:
  - Syntactic embeddings from an RNN parser ([Socher et al. 2013](#))
  - Semantic embeddings from word2vec, GloVe, COMPOSES

**Experimental Setup**

- Data (human pairwise judgments):
  - Train: WMT11 (11,160 pairs)
  - Dev: WMT13 (5,000 pairs)
  - Test: WMT12 (3,798 pairs)

  Features were normalized using min-max

- Training:
  - Optimization: SGD+adagrad for 10k epochs with early stopping and L2 regularization
  - Learning rate: 0.01
  - Mini batch size: 30
  - Hidden layer size: 4 with tanh activations

- Evaluation: WMT12 version of Kendall’s tau

**Results (Kendall Tau)**

<table>
<thead>
<tr>
<th>BLEU Components + Embeddings</th>
<th>18.46</th>
</tr>
</thead>
<tbody>
<tr>
<td>BLEU</td>
<td>18.46</td>
</tr>
<tr>
<td>BLEU+COMP</td>
<td>19.75</td>
</tr>
<tr>
<td>+SYN25</td>
<td>23.70</td>
</tr>
<tr>
<td>+GW25</td>
<td>24.92</td>
</tr>
<tr>
<td>+SYN25+GW25</td>
<td>26.15</td>
</tr>
</tbody>
</table>

**Deep vs. Flat NN**

| Single-layer                  | 29.10 |
| Multi-layer                   | 29.70 |

**Different Semantic Embeddings**

<table>
<thead>
<tr>
<th>Source</th>
<th>Alone</th>
<th>Comb.</th>
</tr>
</thead>
<tbody>
<tr>
<td>GW25</td>
<td>10.01</td>
<td>29.70</td>
</tr>
<tr>
<td>GW300</td>
<td>9.66</td>
<td>29.90</td>
</tr>
<tr>
<td>CC-300-42B</td>
<td>12.16</td>
<td>29.68</td>
</tr>
<tr>
<td>CC-300-840B</td>
<td>11.41</td>
<td>29.88</td>
</tr>
<tr>
<td>Word2Vec300</td>
<td>7.72</td>
<td>29.13</td>
</tr>
<tr>
<td>COMPOSES400</td>
<td>12.35</td>
<td>28.54</td>
</tr>
</tbody>
</table>

**Conclusion and Future Work**

- Proposed a novel NN framework for MT evaluation:
  - Flexible in incorporating different sources of information
  - Results are additive w.r.t. the sources of information
  - Enables fast inference
  - Achieves state-of-the-art results

- Future work:
  - Add source-sentence information
  - Use the NN framework for:
    - re-ranking
    - quality estimation
    - system combination