Ensemble: 7 out of 115
Feature-based: 37, LSTM: 73, TreeLSTM: 77

Results

Features-based
An extended version of the winner in last year’s STS challenge:
- Alignment ratio (Sultan et al., 2015)
- Cosine of word2vec centroids
- Cosine of one-hot bag-of-words
- Machine Translation metrics
  - BLUE, METEOR, BADGER, TER, TERp, NIST

Generalization of LSTMs to tree-structured network topologies that performed well on semantic relatedness (Tai et al., 2015)
- No feature engineering
- Requires an external parser

LSTM
- No feature engineering
- No external parser

Discussion

- Base systems have pairwise low correlation: they capture different views of the data
- Correlation with ensemble system for all base systems is high (>0.7)

Example: System predictions for a sentence pair
- There’s not a lot you can do about that
- There’s not that much that you can do with a sourdough starter.

Distribution of gold similarity scores in the training data

<table>
<thead>
<tr>
<th>System</th>
<th>Ensemble</th>
<th>Feature-based</th>
<th>LSTM</th>
<th>TreeLSTM</th>
<th>Reference</th>
</tr>
</thead>
<tbody>
<tr>
<td>Ensemble</td>
<td>2.0</td>
<td>3.96</td>
<td>0.31</td>
<td>1.39</td>
<td>1.76</td>
</tr>
<tr>
<td>Feature-based</td>
<td>1.0</td>
<td>0.769</td>
<td>0.751</td>
<td>0.802</td>
<td>0.592</td>
</tr>
<tr>
<td>LSTM</td>
<td>0.751</td>
<td>0.456</td>
<td>1.0</td>
<td>0.608</td>
<td>0.442</td>
</tr>
<tr>
<td>TreeLSTM</td>
<td>0.802</td>
<td>0.413</td>
<td>0.608</td>
<td>1.0</td>
<td>0.523</td>
</tr>
<tr>
<td>Reference</td>
<td>0.592</td>
<td>0.44</td>
<td>0.442</td>
<td>0.523</td>
<td>1</td>
</tr>
</tbody>
</table>