SimiHawk: A Deep Ensemble System for Semantic Textual Similarity (SemEval-2016 Task 1)
Peter Potash, William Boag, Alexey Romanov, Vasilii Ramanishka, Anna Rumshisky
Text Machine Lab for Natural Language Processing, Dept. of Computer Science, University of Massachusetts Lowell

Problem - SemEval Task 1

Semantic Textual Similarity (STS) measures the degree of equivalence in the underlying semantics of paired snippets of text.

Range from 0 to 5:
0 - the sentences are completely independent
5 - the sentences are semantically equivalent

Example*:
Sentence 1: A Pyrrhic victory
Sentence 2: Cutting off your nose to spite your face
* Human annotation: 3.0174

Approach

Goal: evaluate strengths and weaknesses of different approaches:
- A classifier with heavily hand-engineered features
  - Performed best at last year’s challenge
- Two deep neural network architectures (learned representation)
  - Conventional LSTM: recurrent neural network
  - TreeLSTM: Recursive neural network
    → composes the current state from many child units

Results

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

All systems train on all available data from previous shared tasks -- a total of 13,061 pairs.

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

Models

Features-based
- Alignment ratio
- 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 at al, 2015)
- No feature engineering
- Requires an external parser

LSTM
- No feature engineering
- No external parser

Ensemble
- Stacking model trained on 5-fold cross-validation base predictions
- For testing, base systems trained on all training data