SimiHawk: A Deep Ensemble System for Semantic Textual Similarity (SemEval-2016 Task 1)

Peter Potash, William Boag, Alexey Romanov, Vasili 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 weaknesss 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

Models

Features-based

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

TreeLSTM

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

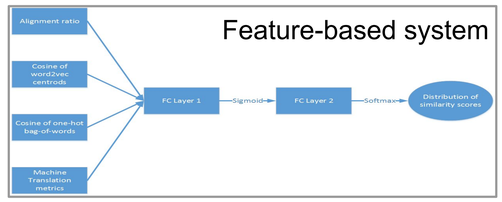
A conventional LSTM architecture. Bowman et al (2015) showed that conventional LSTM can learn tree structures.

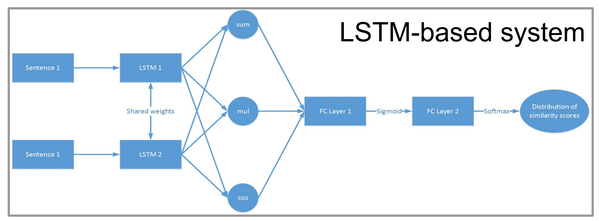
- No feature engineering
- No external parser

Ensemble

Stacking ensemble system

- Features are
  - Predictions of 3 base systems
  - Alignment ratio
- Stacking model trained on 5-fold cross-validation base predictions
- For testing, base systems trained on all training data

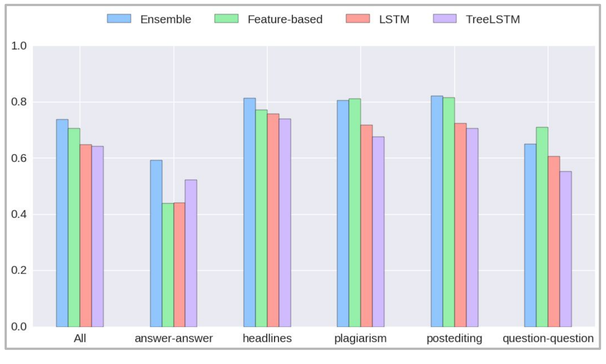




Results

	All	answer-answer	headlines	plagiarism	postediting	question-question
Ensemble	**0.73774**	**0.59237**	**0.81419**	0.80566	**0.82179**	0.65048
Feature-based	0.70647	0.44003	0.77109	**0.81105**	0.81600	**0.71035**
LSTM	0.64840	0.44177	0.75703	0.71737	0.72317	0.60691
TreeLSTM	0.64140	0.52277	0.74083	0.67628	0.70655	0.55265

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



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

Discussion

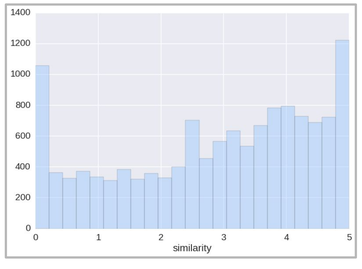
- Results suggest that three base systems have **complementary strengths and weaknesses**, and the ensemble system is able to leverage them to **eliminate noise** in the predictions.
- Feature-based system is the best-performing base system overall
- Ensemble system's predictions have the highest correlation with
  - Feature-based system in two out of five domains
  - TreeLSTM in the other three domains

Example: System predictions for a sentence pair

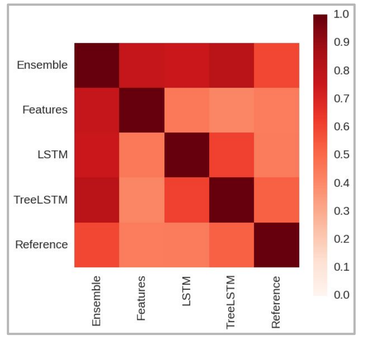
- There's not a lot you can do about that
- There's not that much that you can do with a sourdough starter.

Gold Standard	Feature-based	LSTM	TreeLSTM	Ensemble
2.0	3.96	0.31	1.39	**1.76**

- Distribution of gold similarity scores in the training data



- 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)



System	Ensemble	Features	LSTM	TreeLSTM	Reference
Ensemble	1	0.769	0.751	0.802	0.592
Feature-based	0.769	1	0.456	0.413	0.44
LSTM	0.751	0.456	1	0.608	0.442
TreeLSTM	0.802	0.413	0.608	1	0.523
Reference	0.592	0.44	0.442	0.523	1