SimiHawk: A Deep Ensemble System for Semantic Textual Similarity

Peter Potash, William Boag, Alexey Romanov, Vasili Ramanishka, Anna Rumshisky
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

Approach

We were interested in comparing three approaches

- Heavily **hand-engineered features**
  - Require a lot of human labor
- Deep **neural-network** architectures
  - Can learn the features by themselves

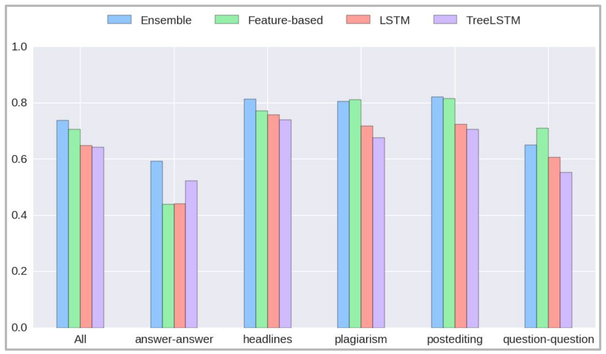
Moreover, we wanted to compare two NN approaches

- Conventional LSTM: standard **recurrent** neural network
- TreeLSTM: **Recursive** neural network
  - Composes the current state from many child units

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



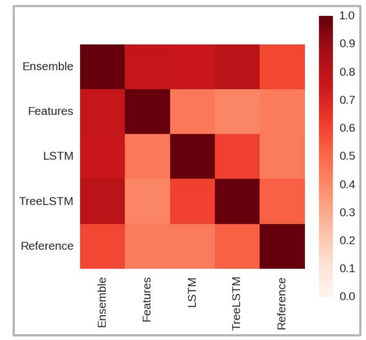
All systems train on all available data from previous shared tasks -- a total of 13,061 pairs.

Discussion

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

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)



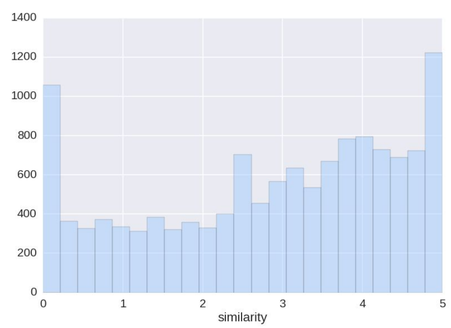
The ensemble system has an ability to **form a consensus** among the base systems and **eliminate noise** in the predictions.

Feature-based system is best-performing base system; ensemble system only correlated with it the highest 2 out of 5 domains. Other 3 were TreeLSTM.

Example predictions for 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**



Models

Features-based

Such system had an impressive success in the previous STS

- 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 (Tai at al, 2015)

- No need in feature engineering
- Requires an external parser

LSTM

Is a tree really necessarily to achieve good results? (Bowman et al, 2015b) showed that just conventional LSTM can lear tree structures

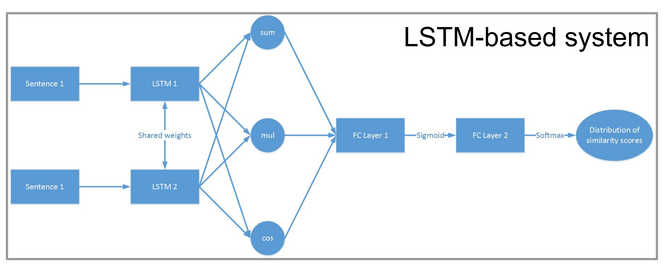
- No need in feature engineering
- No need in an external parser

Ensemble

Combines all three systems -- predictions of 3 base systems act as features.

Based on empirical results, alignment ratio is added as fourth feature.

Model is trained on 5-fold cross-validation predictions from base systems. For final predictions, base systems train on all training data.



* Human annotation: 3.0174