AZMAT: Sentence Similarity Using Associative Matrices

Evan Jaffe, Lifeng Jin, David King, Marten Van Schijndel

Overview
- SemEval-2015 Task 2: English STS
  - Model human judgements of sentential similarity (0-5)
- SVM with linear kernel
- Unfolding Recursive AutoEncoder (Socher et al., 2011)
- Associative Matrices (Anderson et al., 1977; Howard and Kahana, 2002)
- GloVe global vectors (Pennington et al., 2014)
- Surface lexical overlap
- Training input is previous SemEval tasks 2012-2014
- Ranks 69th out of 74 systems
- Question: Does phrasal cosine similarity help?

Subsystem Combination
- Generate an embedding for every node of a binarized phrase-structure tree
- Measure cosine similarity of every node in sentence A to every node in sentence B
- Generate a fixed-length feature vector that concatenates ordered similarity scores, repeating out to required space, sorting

Unfolding Recursive Autoencoders
- Used in paraphrase detection
- Composes embeddings for each node in a binary phrase-structure tree, given leaf embeddings
- Learns to encode and decode, with objective of minimizing reconstruction error
- Uses Stanford parser, not GCG tree
- Current work replaces dynamic pooling with depth-sensitive vector expansion in order to avoid lossy operations while retaining global structural similarity

GloVe Global Vectors
- 300-dimensional vectors trained on 42 billion tokens
- Composition just percolates up the head word from composition with associative matrices does not seem to work well
- Possible matrix saturation with too few dependency labels
- Finer-grained syntactic info (beyond depth) when grouping cosine similarities
- SVM regularization tuning
- Using phrasal nodes to do similarity is an open challenge

Conclusions
- Leaf features are better predictors than phrasal comparison features
- Overfitting evident from development analysis
- Surface features are complementary to cosine similarity features (SUGA vs. UGA model performance)
- Composition with associative matrices does not seem to work well

Discussion

Acknowledgements
This material is based upon work supported by the National Science Foundation Graduate Research Fellowship under Grant No. DGE-1343012. Any opinion, findings, and conclusions or recommendations expressed in this material are those of the authors and do not necessarily reflect the views of the National Science Foundation. We would also like to thank the anonymous reviewers for their helpful suggestions and comments.

Contact Information
If you have any questions/comments/concerns, you may reach us at the following:
Evan Jaffe: jaffe.59@osu.edu
Lifeng Jin: jin.544@osu.edu
David King: king.2136@osu.edu
Marten Van Schijndel: vanschm@ling.osu.edu

The Ohio State University
linguistics.osu.edu