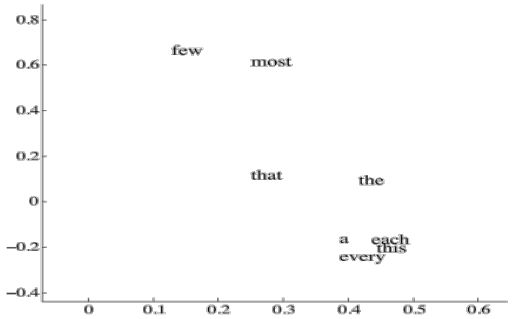
How much do word embeddings encode about syntax?

Presented by: Satyam Kumar Shivam
Guide: Prof. Amitabha Mukerjee
Course: CS671A
Paper authors: Jacob Andreas and Prof. Dan Klein

Introduction

- Investigates ways to augment a constituency parser with a discrete space state by using word embeddings.
- Hypothesis: Features, like clusters and embeddings, can improve dependency parsers by providing syntactic abstractions.



[1]

Hypothesis on embeddings-parser interaction

1. **Vocabulary expansion hypothesis**
   a. Word embeddings are useful for handling out-of-vocabulary words
   b. Treat unknown words in same way as known words with similar representations
2. **Statistic sharing hypothesis**
   a. Word embeddings are useful for handling in-vocabulary words
   b. Pool statistics for related words
3. **Embedding structure hypothesis**
   a. The space structure directly encodes syntactic information in its coordinate axes
   b. Feature corresponding to a word's position can be useful in a feature-based lexicon

Parser extensions

1. **Vocabulary expansion: OOV model**
   a. A simple but targeted out-of-vocabulary (OOV) model
   b. Every unknown word is simply replaced by its nearest neighbor in the training set
2. **Statistic sharing: Lexicon pooling model**
   a. A smoothing technique is used.
   b. The probability P(w|t) is stored using a smoothed, kernelized lexicon

$$p(w|t) = \frac{1}{Z} \sum_{w'} \alpha_{t,w'} e^{-\beta ||\phi(w) - \phi(w')||^2}$$

Parser extensions

Embedding structures: Embedding features

1. Maryland featured parser is used.
2. Set of lexical template features is replaced by set of indicator features.
3. For each dimension *i*, an indicator feature is created corresponding to the linearly-bucketed value of feature at that index.
4. Morphological features are removed from the parser in order to focus on the effect of word embeddings.

OOV, lexican pooling and featured models

Model		300	3000	Full
Baseline		71.88	84.70	91.13
OOV	(C&W)	72.20	84.77	91.22
OOV	(CBOW)	72.20	84.78	91.22
Pooling	(C&W)	72.21	84.55	91.11
Pooling	(CBOW)	71.61	84.73	91.15
Features	(ident)	67.27	82.77	90.65
Features	(C&W)	70.32	83.78	91.08
Features	(CBOW)	69.87	84.46	90.86

[1]

Results

1. OOV model achieves small gains over the baseline for a 300-word training corpus, but become statistically insignificant with more training data.
2. This behaviour of OOV model is insensitive to choice of embedding.
3. For lexicon pooling model, performance decreased with small set of beta values and performance increased with increased set of beta values.
4. For a combination of lexicon pooling model with beta = 0.3 and OOV, there is small gain observed with 300 and 3000 train sentences, but a decrease in performance in observed on the full corpus.

Conclusion

1. These modified parser have slight gains on extremely small training sets, which quickly vanishes as training set size increases.
2. Unsupervised word embeddings do contain some syntactically useful information.
3. This information is redundant with what the model is able to determine for itself from only a small amount of labeled training data.
4. Gains over baseline, by these modified parser, is extremely sensitive to training conditions.

References

1. Andreas, Jacob, and Dan Klein. "How much do word embeddings encode about syntax." *Proceedings of ACL*. 2014.
2. Bansal, Mohit, Kevin Gimpel, and Karen Livescu. "Tailoring continuous word representations for dependency parsing." *Proceedings of the Annual Meeting of the Association for Computational Linguistics.* 2014.
3. Petrov, Slav, and Dan Klein. "Improved Inference for Unlexicalized Parsing." *HLT-NAACL.* Vol. 7. 2007.
4. Collobert, Ronan, et al. "Natural language processing (almost) from scratch." *The Journal of Machine Learning Research* 12 (2011): 2493-2537.
5. Mikolov, Tomas, et al. "Distributed representations of words and phrases and their compositionality." *Advances in neural information processing systems*. 2013.