Supplementary Materials: Learning Word Representations with Hierarchical Sparse Coding

Dani Yogatama DYOGATAMA@CS.CMU.EDU
Manaal Faruqui MFARUQUI@CS.CMU.EDU
Chris Dyer CDYER@CS.CMU.EDU
Noah A. Smith NASMITH@CS.CMU.EDU
Language Technologies Institute, School of Computer Science, Carnegie Mellon University, Pittsburgh, PA 15213, USA

1. Additional Results

In Table 1, we compare FOREST with three additional baselines:

- Murphy *et al.* (2012): a word representation trained using non-negative sparse embedding (NNSE) on our corpus. Similar to the authors, we use an NNSE implementation from `http://spams-devel.gforge.inria.fr/` (Mairal *et al.*, 2010).
- Mikolov *et al.* (2013): a log bilinear model that predicts a word given its context, trained using hierarchical softmax with a binary Huffman tree (continuous bag of words, CBOW-HS). We use an implementation from `https://code.google.com/p/word2vec/`.
- Mikolov *et al.* (2013): a log bilinear model that predicts context words given a target word, trained using hierarchical softmax with a binary Huffman tree (skip gram, SG-HS). We use an implementation from `https://code.google.com/p/word2vec/`.

We train these models on our corpus using the same setup as experiments in our paper.

2. Additional Two-Dimensional Projections

For FOREST, SG, and NCE with $M = 520$, we project the learned word representations into two dimensions using the t-SNE tool (van der Maaten and Hinton, 2008) from `http://homepage.tudelft.nl/19j49/t-SNE.html`. We show projections of words related to the concept "good" vs. "bad" and "man" vs. "woman" in Figure 1.

3. List of Word Similarity Datasets

We use the following word similarity datasets in our experiments:

- Finkelstein *et al.* (2002): WordSimilarity dataset (353 pairs).
- Agirre *et al.* (2009): a subset of WordSimilarity dataset for evaluating similarity (203 pairs).
- Agirre *et al.* (2009): a subset of WordSimilarity dataset for evaluating relatedness (252 pairs).
- Miller and Charles (1991): semantic similarity dataset (30 pairs)
- Rubenstein and Goodenough (1965): contains only nouns (65 pairs)
- Luong *et al.* (2013): rare words (2,034 pairs)
- Bruni *et al.* (2012): frequent words (3,000 pairs)
- Radinsky *et al.* (2011): MTurk-287 dataset (287 pairs)
- Halawi and Dror (2014): MTurk-771 dataset (771 pairs)
- Yang and Powers (2006): contains only verbs (130 pairs)

References

Agirre, E., Alfonseca, E., Hall, K., Kravalova, J., Pasca, M., and Soroa, A. (2009). A study on similarity and relatedness using distributional and wordnet-based approaches. In *Proc. of NAACL-HLT*.

Bruni, E., Boleda, G., Baroni, M., and Tran, N.-K. (2012). Distributional semantics in technicolor. In *Proc. of ACL*.

Finkelstein, L., Gabrilovich, E., Matias, Y., Rivlin, E., Solan, Z., Wolfman, G., and Ruppin, E. (2002). Placing search in context: The concept revisited. *ACM Transactions on Information Systems*, **20**(1), 116–131.

Halawi, G. and Dror, G. (2014). The word relatedness mturk-771 test collection.

Table 1. Summary of results for non-negative sparse embedding (NNSE), continuous bag-of-words and skip gram models trained with hierarchical softmax (CBOW-HS and SG-HS). Higher number is better (higher correlation coefficient or higher accuracy).

M	Task	NNSE	CBOW-HS	SG-HS	FOREST
52	Word similarity	0.04	0.38	0.47	**0.52**
	Syntactic analogies	0.10	19.50	**24.87**	24.38
	Semantic analogies	0.01	5.31	**14.77**	9.86
	Sentence completion	0.01	22.51	28.78	**28.88**
	Sentiment analysis	61.12	68.92	71.72	**75.83**
520	Word similarity	0.05	0.50	0.57	**0.66**
	Syntactic analogies	0.81	46.00	**50.40**	48.00
	Semantic analogies	0.57	8.00	31.05	**41.33**
	Sentence completion	22.81	25.80	27.79	**35.86**
	Sentiment analysis	67.05	78.50	79.57	**81.90**

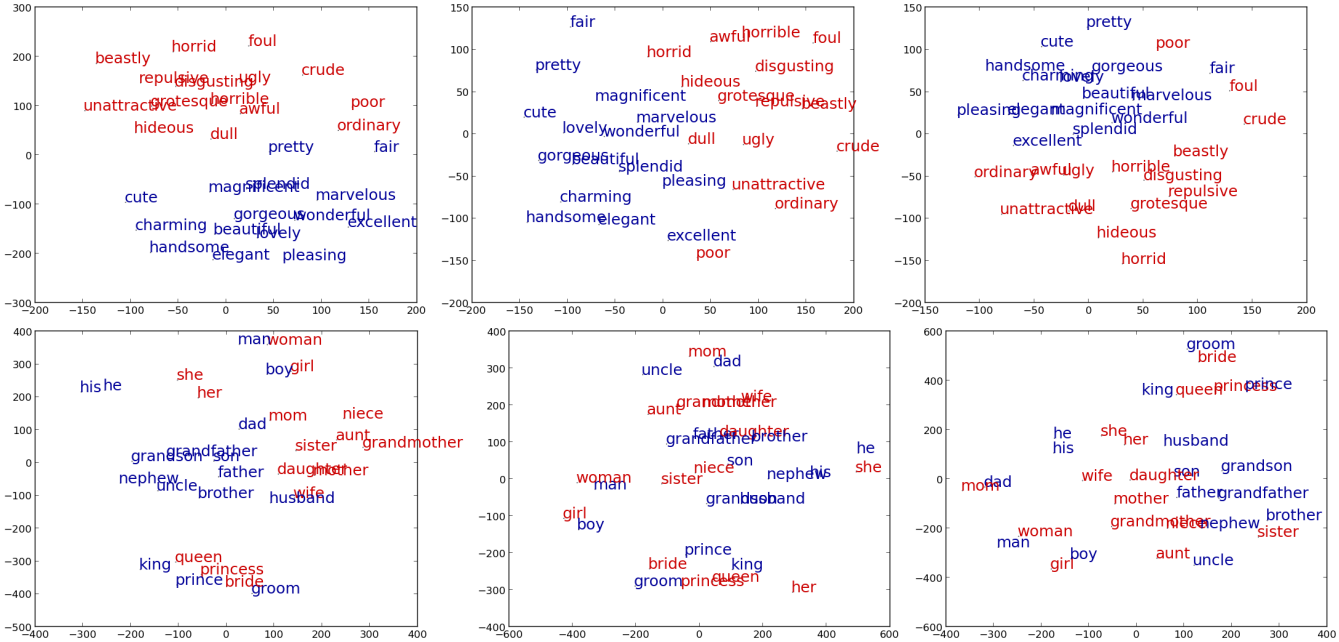

Figure 1. Two dimensional projections of the FOREST (left), SG (middle), and NCE (right) word representations using the t-SNE tool (van der Maaten and Hinton, 2008). Words associated with "good" (top) and "man" (bottom) are colored in blue, words associated with "bad" (top) and "woman" (bottom) are colored in red. The two plots on the top left are the same plots shown in the paper.

Luong, M.-T., Socher, R., and Manning, C. D. (2013). Better word representations with recursive neural networks for morphology. In *Proc. of CONLL*.

Mairal, J., Bach, F., Ponce, J., and Sapiro, G. (2010). Online learning for matrix factorization and sparse coding. *Journal of Machine Learning Research*, **11**, 19–60.

Mikolov, T., Chen, K., Corrado, G., and Dean, J. (2013). Efficient estimation of word representations in vector space. In *Proc. of Workshop at ICLR*.

Miller, G. A. and Charles, W. G. (1991). Contextual correlates of semantic similarity. *Language and Cognitive Processes*, **6**(1), 1–28.

Murphy, B., Talukdar, P., and Mitchell, T. (2012). Learning effective and interpretable semantic models using non-negative sparse embedding. In *Proc. of COLING*.

Radinsky, K., Agichtein, E., Gabrilovich, E., and Markovitch, S. (2011). A word at a time: Computing word relatedness using temporal semantic analysis. In *Proc. of WWW*.

Rubenstein, H. and Goodenough, J. B. (1965). Contextual

correlates of synonymy. *Communications of the ACM*, **8**(10), 627–633.

van der Maaten, L. and Hinton, G. (2008). Visualizing data using t-sne. *Journal of Machine Learning Research*, **9**, 2579–2605.

Yang, D. and Powers, D. M. W. (2006). Verb similarity on the taxonomy of wordnet. In *Proc. of GWC*.