SimVerb-3500: A Large-Scale Evaluation Set of Verb Similarity Supplementary Material

1 Vector Space Models

1.1 Unsupervised Text-Based Models

These models mainly learn from co-occurrence statistics in large corpora, therefore to facilitate the generality of our results, we evaluate them on two different corpora. With **8B** we refer to the corpus produced by the `word2vec` script, consisting of 8 billion tokens from various sources (Mikolov et al., 2013a).[1] With **PW** we refer to the English Polyglot Wikipedia corpus (Al-Rfou et al., 2013).[2] d denotes the embedding dimensionality, and ws is the window size in case of bag-of-word contexts. The models we consider are as follows:

SGNS-BOW (PW, 8B) Skip-gram with negative sampling (SGNS) (Mikolov et al., 2013a; Mikolov et al., 2013b) trained with bag-of-words (BOW) contexts; $d = 500$, $ws = 2$ on 8B as in prior work (Melamud et al., ; Schwartz et al., 2016). $d = 300$, $ws = 2$ on PW as in prior work (Levy and Goldberg, 2014; Vulić and Korhonen, 2016).

SGNS-UDEP (PW) SGNS trained with universal dependency[3] (UD) contexts following the setup of (Levy and Goldberg, 2014; Vulić and Korhonen, 2016). The PW data were POS-tagged with universal POS (UPOS) tags (Petrov et al., 2012) using TurboTagger (Martins et al., 2013)[4], trained using default settings without any further parameter fine-tuning (SVM MIRA with 20 iterations) on the TRAIN+DEV portion of the UD treebank annotated with UPOS tags. The data were then parsed using the graph-based Mate parser v3.61 (Bohnet, 2010).[5] $d = 300$ as in (Vulić and Korhonen, 2016)

SGNS-DEP (8B) Another variant of a dependency-based SGNS model is taken from the recent work of Schwartz et al. (2016), based on Levy and Goldberg (2014). The 8B corpus is parsed with labeled Stanford dependencies (de Marneffe and Manning, 2008), the Stanford POS Tagger (Toutanova et al., 2003) and the stack version of the MALT parser (Goldberg and Nivre, 2012) are used; $d = 500$ as in prior work (Schwartz et al., 2016).

All other parameters of all SGNS models are set to the standard settings: the models are trained with stochastic gradient descent, global learning rate of 0.025, subsampling rate $1e-4$, 15 epochs.

SymPat (8B) A template-based approach to vector space modeling introduced by Schwartz et al. (2015). Vectors are trained based on co-occurrence of words in symmetric patterns (Davidov and Rappoport, 2006), and an antonym detection mechanism is plugged in the representations. We use pre-trained dense vectors ($d = 300$ and $d = 500$) with the antonym detector enabled, available online.[6]

Count-SVD Traditional count-based vectors using PMI weighting and SVD dimensionality reduction (*ws = 2; d = 500*). This is the best performing reduced count-based model from Baroni et al. (2014), vectors were obtained online.[7]

1.2 Models Relying on External Resources

Non-Distributional Sparse binary vectors built from a wide variety of hand-crafted linguistic resources, e.g., WordNet, Supersenses, FrameNet, Emotion and Sentiment lexicons, Connotation lexicon, among others (Faruqui and Dyer, 2015).[8]

Paragram Wieting et al. (2015) use the Paraphrase Database (PPDB) (Ganitkevitch et al., 2013) word pairs to learn word vectors which emphasise paraphrasability. They do this by fine-tuning, also known as retro-fitting (Faruqui et al., 2015), `word2vec` vectors using a SGNS inspired objective function designed to incorporate the PPDB semantic similarity constraints. Two variants are available online: $d = 25$ and $d = 300$.[9]

Paragram+CF Mrkšić et al. () suggest another variant of the retro-fitting procedure called counter-

[1] https://code.google.com/archive/p/word2vec/
[2] https://sites.google.com/site/rmyeid/projects/polyglot
[3] http://universaldependencies.org/ (version 1.2)
[4] http://www.cs.cmu.edu/ ark/TurboParser/
[5] https://code.google.com/archive/p/mate-tools/
[6] http://www.cs.huji.ac.il/˜roys02/papers/sp_embeddings/
[7] http://clic.cimec.unitn.it/composes/semantic-vectors.html
[8] https://github.com/mfaruqui/non-distributional
[9] http://ttic.uchicago.edu/∼wieting/

fitting (CF) which further improves the Paragram vectors by injecting antonymy constraints from PPDB v2.0 (Pavlick et al., 2015) into the final vector space. $d = 300$.[10]

References

Rami Al-Rfou, Bryan Perozzi, and Steven Skiena. 2013. Polyglot: Distributed word representations for multilingual nlp. In *CoNLL*, pages 183–192.

Marco Baroni, Georgiana Dinu, and Germán Kruszewski. 2014. Don't count, predict! a systematic comparison of context-counting vs. context-predicting semantic vectors. In *ACL*, pages 238–247.

Bernd Bohnet. 2010. Top accuracy and fast dependency parsing is not a contradiction. In *COLING*, pages 89–97.

Dmitry Davidov and Ari Rappoport. 2006. Efficient unsupervised discovery of word categories using symmetric patterns and high frequency words. In *ACL*, pages 297–304.

Marie-Catherine de Marneffe and Christopher D. Manning. 2008. The Stanford typed dependencies representation. In *Proceedings of the Workshop on Cross-Framework and Cross-Domain Parser Evaluation*, pages 1–8.

Manaal Faruqui and Chris Dyer. 2015. Non-distributional word vector representations. In *ACL*, pages 464–469.

Manaal Faruqui, Jesse Dodge, Sujay Kumar Jauhar, Chris Dyer, Eduard H. Hovy, and Noah A. Smith. 2015. Retrofitting word vectors to semantic lexicons. In *NAACL-HLT*, pages 1606–1615.

Juri Ganitkevitch, Benjamin Van Durme, and Chris Callison-Burch. 2013. Ppdb: The paraphrase database. In *NAACL-HLT*, pages 758–764.

Yoav Goldberg and Joakim Nivre. 2012. A dynamic oracle for arc-eager dependency parsing. In *COLING*, pages 959–976.

Omer Levy and Yoav Goldberg. 2014. Dependency-based word embeddings. In *ACL*, pages 302–308.

André F. T. Martins, Miguel B. Almeida, and Noah A. Smith. 2013. Turning on the turbo: Fast third-order non-projective turbo parsers. In *ACL*, pages 617–622.

Oren Melamud, David McClosky, Siddharth Patwardhan, and Mohit Bansal. The role of context types and dimensionality in learning word embeddings. pages 1030–1040.

Tomas Mikolov, Kai Chen, Gregory S. Corrado, and Jeffrey Dean. 2013a. Efficient estimation of word representations in vector space. In *ICLR: Workshop Papers*.

Tomas Mikolov, Ilya Sutskever, Kai Chen, Gregory S. Corrado, and Jeffrey Dean. 2013b. Distributed representations of words and phrases and their compositionality. In *NIPS*, pages 3111–3119.

Nikola Mrkšić, Diarmuid Ó Séaghdha, Blaise Thomson, Milica Gašić, Lina Maria Rojas-Barahona, Pei-Hao Su, David Vandyke, Tsung-Hsien Wen, and Steve J. Young. Counter-fitting word vectors to linguistic constraints. In *NAACL-HLT*, pages 142–148.

Ellie Pavlick, Pushpendre Rastogi, Juri Ganitkevitch, Benjamin Van Durme, and Chris Callison-Burch. 2015. PPDB 2.0: Better paraphrase ranking, fine-grained entailment relations, word embeddings, and style classification. In *ACL*, pages 425–430.

Slav Petrov, Dipanjan Das, and Ryan T. McDonald. 2012. A universal part-of-speech tagset. In *LREC*, pages 2089–2096.

Roy Schwartz, Roi Reichart, and Ari Rappoport. 2015. Symmetric pattern based word embeddings for improved word similarity prediction. In *CoNLL*, pages 258–267.

Roy Schwartz, Roi Reichart, and Ari Rappoport. 2016. Symmetric patterns and coordinations: Fast and enhanced representations of verbs and adjectives. In *NAACL-HLT*, pages 499–505.

Kristina Toutanova, Dan Klein, Christopher D. Manning, and Yoram Singer. 2003. Feature-rich part-of-speech tagging with a cyclic dependency network. In *NAACL-HLT*, pages 173–180.

Ivan Vulić and Anna Korhonen. 2016. Is "universal syntax" universally useful for learning distributed word representations? In *ACL*.

John Wieting, Mohit Bansal, Kevin Gimpel, and Karen Livescu. 2015. From paraphrase database to compositional paraphrase model and back. *Transactions of the ACL*, 3:345–358.

[10]https://github.com/nmrksic/counter-fitting