Syntactically Trained Word Vectors

...and why you should use them!

Evan Jaffe
The Ohio State University

Midwest Speech and Language Days, Indiana University
May 13-14, 2016

Problem

Many options for training, which one is best?

Word2Vec Word-window Context [Mikolov et al., 2013]

Word2VecF Syntactic Dependency Context [Levy and Goldberg, 2014]

Retrofit Semantic Ontology Context (Wordnet, FrameNet, PPDB, etc.)[Faruqui et al., 2015]

Word2Vec popular and cheap method, but not always the best choice
Some work showing adding task-specific information improves task performance

Can good annotation contribute to big data?

Should we train using syntactic contexts?

- At least for syntactic tasks, yes...
- ...but choice of syntactic context matters!
- What kind of syntactic context is best? I.e., what is the right level of representation/abstraction?

Word embeddings + NLP tasks + you!

Proposal

1. Train different sets of word embeddings on various types of syntactic (and non-syntactic) contexts

Word2Vec	baseline word-window context
Labeled directed	baseline syntactic context
Unlabeled directed	abstracts from some dependency framework-specific decisions
Unlabeled undirected	like Word2Vec with sentence-length window, constrained to only sample words connected with dependency relation

2. Evaluate on prepositional phrase attachment task [Belinkov et al., 2014], changing only pre-trained input vectors

Claims

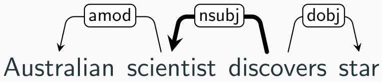
When trained on comparable data and evaluated on a downstream syntactic task,

- Labeled directed word embeddings are NOT significantly different from Word2Vec embeddings
- Unlabeled directed word embeddings ARE significantly better than Word2Vec embeddings

Syntactic dependency contexts are useful for training word embeddings IF you choose the right dependency contexts.

Approach

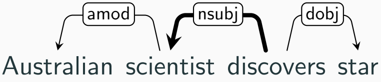
Dependency Context Training Types



Dependency Training type	Target Word	Context
Labeled directed	scientist	discovers+nsubj
	discovers	scientist-nsubj
Unlabeled directed	scientist	discovers+
	discovers	scientist-
Unlabeled undirected	scientist	discovers
	discovers	scientist

Labeled Directed

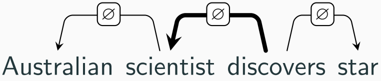
Baseline syntactic context. Similar to [Levy and Goldberg, 2014] contexts.



Target Word	Context
scientist	discovers+nsubj
discovers	scientist-nsubj

Unlabeled Directed

Retain governor-dependent information, but remove arc label. Abstracts away from dependency framework-specific labels.



Target Word	Context
scientist	discovers+
discovers	scientist-

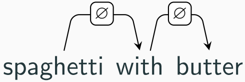
Unlabeled Undirected

Somewhat similar to Word2Vec with sentence-length window, except constrains to word-pairs connected by syntactic dependency.

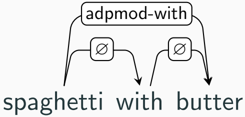
Target Word	**Context**
scientist	discovers
discovers	scientist

Higher Order Preposition Arcs

- Follows increasingly standard practice of generating arc between head and object of prepositional phrase, connecting contentful words.
- Stanford Dependencies (collapsed), Universal Dependencies, Goldberg and Levy



becomes:



Word2VecF Training Objective

$$\arg\max_{v_w, v_c} \left(\sum_{(w,c) \in D} \log \sigma(v_c \cdot v_w) + \sum_{(w,c) \in D'} \log \sigma(-v_c \cdot v_w) \right)$$

[Levy and Goldberg, 2014] Same as Mikolov et al., Skip-gram with negative sampling

Word Vector Training Data

English Wikipedia (1.6 billion tokens), parsed with version of [Goldberg and Nivre, 2012], outputting CoNLL-formatted parse with labels from [McDonald et al., 2013]

Approximately 80 million sentences.

Raw counts for most common arc types:

Label type	**Count**
adpmod	186,757,807
adpobj	183,346,238
p	183,099,676
det	152,170,759
compmod	141,968,939
nsubj	106,977,803
amod	90,965,244
ROOT	80,122,518

Evaluation

Prepositional Phrase Attachment Task

Given a prepositional phrase and a list of candidate attachment sites, choose the correct attachment. [Belinkov et al., 2014]

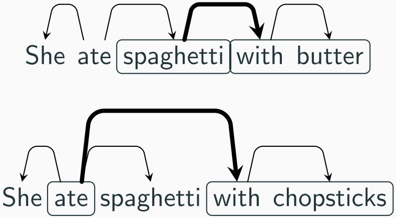


Figure 1: Example prepositional phrase attachment decision for two similar sentences. Note that the first sentence attaches the prepositional phrase to the noun *spaghetti* and the second attaches it to the verb *ate*.

Test Set

- Prepositional phrases with gold attachments from Penn Treebank
- Belinkov et al: PTB 2-21 Training, 23 Test
- This work: 10-fold cross validation gives about 30,000 items

Belinkov et al. System - Head Prep Child Distance (HPCD)

Neural network learner that composes and scores word vectors for candidate head, preposition and prepositional object words.

- Compose word vectors to generate phrasal embedding.
- Score phrasal embedding.
- Choose head by taking argmax of scored candidate phrasal embeddings.

Belinkov et al. System

Learn θ :

- composition matrix $\mathbf{W} \in \mathbb{R}^{n \times 2n}$
- weight vector $\mathbf{w} \in \mathbb{R}^{n}$
- bias term $\mathbf{b} \in \mathbb{R}^{n}$

Choose $\hat{h} \in H$ for sentence x, preposition z, model parameters θ:

$$\hat{h} = \underset{h \in H}{argmax} \;\; \text{score}(x, z, h; \theta) \tag{1}$$

Scoring function is the dot product of the weight vector $\mathbf{w}$ and the phrasal embedding for a given head:

$$\hat{h} = \underset{h \in H}{argmax} \;\; \mathbf{p}_h \cdot \mathbf{w} \tag{2}$$

To generate a phrasal embedding from any two vectors:

$$\mathbf{p} = \tanh(\mathbf{W}[\mathbf{u}; \mathbf{v}] + \mathbf{b}) \tag{3}$$

Results

Baseline syntactic contexts not different from Word2Vec

Model		Accuracy	P-value
UD	Unlabeled directed	.8535	
LD	Labeled directed	.8448	
W2V	Word2Vec	.8434	*0.26*
UU	Unlabeled undirected	.8362	

McNemar's Chi-Square Test. With Bonferroni comparison for multiple comparisons, significance threshold is $p<0.002$. Word2Vec default negative sampling rate used for all models.

Unlabeled directed contexts are better than Word2Vec

Model		Accuracy	P-value
UD	Unlabeled directed	.8535	
LD	Labeled directed	.8448	
W2V	Word2Vec	.8434	*2.2e-16*
UU	Unlabeled undirected	.8362	

McNemar's Chi-Square Test. With Bonferroni comparison for multiple comparisons, significance threshold is $p<0.002$. Word2Vec default negative sampling rate used for all models.

Arc direction matters

Model		Accuracy	P-value
UD	Unlabeled directed	.8535	
LD	Labeled directed	.8448	
W2V	Word2Vec	.8434	
UU	Unlabeled undirected	.8362	*4.9e-09*

McNemar's Chi-Square Test. With Bonferroni comparison for multiple comparisons, significance threshold is $p<0.002$. Word2Vec default negative sampling rate used for all models.

Conclusion

Summary

- Baseline syntactic dependency contexts (labeled arc tuples) not better than W2V, even for a syntactic task!
- Unlabeled dependency contexts do improve this syntactic task performance.
- Consider using unlabeled dependencies when training word embeddings for a syntactic task.

Thanks to: William Schuler, Mike White, Micha Elsner, Marie-Catherine de Marneffe, Lifeng Jin

This work was supported by NSF Grant: DGE-1343012.

Questions?

Appendix

Model		Accuracy by negative sampling rate	
		5	15
HO	Higher Order PP	.8552	.8535
UD	Unlabeled directed	.8535	.8496
LD	Labeled directed	.8448	.8464
W2V	Word2Vec	.8434	.8453
UU	Unlabeled undirected	.8362	.8412

References I

Belinkov, Y., Lei, T., Barzilay, R., and Globerson, A. (2014).
Exploring compositional architectures and word vector representations for prepositional phrase attachment.
Transactions of the Association for Computational Linguistics, 2:561–572.

Faruqui, M., Dodge, J., Jauhar, S. K., Dyer, C., Hovy, E., and Smith, N. A. (2015).
Retrofitting word vectors to semantic lexicons.
In *Proceedings of NAACL*.

Goldberg, Y. and Nivre, J. (2012).
A dynamic oracle for arc-eager dependency parsing.
In *Proceedings of COLING*.

References II

Levy, O. and Goldberg, Y. (2014).
Dependency-based word embeddings.
In *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics, ACL 2014, June 22-27, 2014, Baltimore, MD, USA, Volume 2: Short Papers*, pages 302–308.

McDonald, R. T., Nivre, J., Quirmbach-Brundage, Y., Goldberg, Y., Das, D., Ganchev, K., Hall, K. B., Petrov, S., Zhang, H., Täckström, O., et al. (2013).
Universal dependency annotation for multilingual parsing.
In *ACL (2)*, pages 92–97. Citeseer.

Mikolov, T., Chen, K., Corrado, G., and Dean, J. (2013).
Efficient estimation of word representations in vector space.
CoRR, abs/1301.3781:1–12.