Correlation-based Intrinsic Evaluation of Word Vector Representations

Yulia Tsvetkov[♠] **Manaal Faruqui**[♠] **Chris Dyer**[♣♠]
[♠]Carnegie Mellon University [♣]Google DeepMind
{ytsvetko,mfaruqui,cdyer}@cs.cmu.edu

Abstract

We introduce QVEC-CCA—an intrinsic evaluation metric for word vector representations based on correlations of learned vectors with features extracted from linguistic resources. We show that QVEC-CCA scores are an effective proxy for a range of extrinsic semantic and syntactic tasks. We also show that the proposed evaluation obtains higher and more consistent correlations with downstream tasks, compared to existing approaches to intrinsic evaluation of word vectors that are based on word similarity.

1 Introduction

Being linguistically opaque, vector-space representations of words—word embeddings—have limited practical value as standalone items. They are effective, however, in representing meaning—through individual dimensions and combinations of thereof—when used as features in downstream applications (Turian et al., 2010; Lazaridou et al., 2013; Socher et al., 2013; Bansal et al., 2014; Guo et al., 2014, *inter alia*). Thus, unless it is coupled with an extrinsic task, intrinsic evaluation of word vectors has little value in itself. The main purpose of an intrinsic evaluation is to serve as a *proxy* for the downstream task the embeddings are tailored for. This paper advocates a novel approach to constructing such a proxy.

What are the desired properties of an intrinsic evaluation measure of word embeddings? First, retraining models that use word embeddings as features is often expensive. A *computationally efficient* intrinsic evaluation that *correlates with extrinsic scores* is useful for faster prototyping. Second, an intrinsic evaluation that enables *interpretation* and analysis of properties encoded by vector dimensions is an auxiliary mechanism for analyzing how these properties affect the target downstream task. It thus facilitates refinement of word vector models and, consequently, improvement of the target task. Finally, an intrinsic evaluation that approximates a range of related downstream tasks (e.g., semantic text-classification tasks) allows to assess *generality* (or specificity) of a word vector model, without actually implementing all the tasks.

Tsvetkov et al. (2015) proposed an evaluation measure—QVEC—that was shown to correlate well with downstream semantic tasks. Additionally, it helps shed new light on how vector spaces encode meaning thus facilitating the interpretation of word vectors. The crux of the method is to correlate distributional word vectors with linguistic word vectors constructed from rich linguistic resources, annotated by domain experts. QVEC can easily be adjusted to specific downstream tasks (e.g., part-of-speech tagging) by selecting task-specific linguistic resources (e.g., part-of-speech annotations). However, QVEC suffers from two weaknesses. First, it is not invariant to linear transformations of the embeddings' basis, whereas the bases in word embeddings are generally arbitrary (Szegedy et al., 2014). Second, it produces an unnormalized score: the more dimensions in the embedding matrix the higher the score. This precludes comparison of models of different dimensionality. In this paper, we introduce QVEC-CCA, which simultaneously addresses both problems, while preserving major strengths of QVEC.[1]

2 QVEC and QVEC-CCA

We introduce QVEC-CCA—an intrinsic evaluation measure of the quality of word embeddings. Our method is a modification of QVEC—an evalua-

[1] https://github.com/ytsvetko/qvec

Proceedings of the 1st Workshop on Evaluating Vector Space Representations for NLP, pages 111–115,
Berlin, Germany, August 12, 2016.

tion based on alignment of embeddings to a matrix of features extracted from a linguistic resource (Tsvetkov et al., 2015). We review QVEC, and then describe QVEC-CCA.

QVEC. The main idea behind QVEC is to quantify the linguistic content of word embeddings by maximizing the correlation with a manually-annotated linguistic resource. Let the number of common words in the vocabulary of the word embeddings and the linguistic resource be N. To quantify the semantic content of embeddings, a semantic/syntactic linguistic matrix $\mathbf{S} \in \mathbb{R}^{P \times N}$ is constructed from a semantic/syntactic database, with a column vector for each word. Each word vector is a distribution of the word over P linguistic properties, based on annotations of the word in the database. Let $\mathbf{X} \in \mathbb{R}^{D \times N}$ be embedding matrix with every row as a dimension vector $\mathbf{x} \in \mathbb{R}^{1 \times N}$. D denotes the dimensionality of word embeddings. Then, $\mathbf{S}$ and $\mathbf{X}$ are aligned to maximize the cumulative correlation between the aligned dimensions of the two matrices. Specifically, let $\mathbf{A} \in \{0,1\}^{D \times P}$ be a matrix of alignments such that $a_{ij} = 1$ iff $\mathbf{x}_i$ is aligned to $\mathbf{s}_j$, otherwise $a_{ij} = 0$. If $r(\mathbf{x}_i, \mathbf{s}_j)$ is the Pearson's correlation between vectors $\mathbf{x}_i$ and $\mathbf{s}_j$, then QVEC is defined as:

$$\text{QVEC} = \max_{\mathbf{A}: \sum_j a_{ij} \leq 1} \sum_{i=1}^{X} \sum_{j=1}^{S} r(\mathbf{x}_i, \mathbf{s}_j) \times a_{ij}$$

The constraint $\sum_j a_{ij} \leq 1$, warrants that one distributional dimension is aligned to at most one linguistic dimension.

QVEC-CCA. To measure correlation between the embedding matrix $\mathbf{X}$ and the linguistic matrix $\mathbf{S}$, instead of cumulative dimension-wise correlation we employ canonical correlation analysis (Hardoon et al., 2004, CCA). CCA finds two sets of basis vectors, one for $\mathbf{X}^\top$ and the other for $\mathbf{S}^\top$, such that the correlations between the projections of the matrices onto these basis vectors are maximized. Formally, CCA finds a pair of basis vectors $\mathbf{v}$ and $\mathbf{w}$ such that

$$\begin{aligned} \text{QVEC-CCA} &= \text{CCA}(\mathbf{X}^\top, \mathbf{S}^\top) \\ &= \max_{\mathbf{v},\mathbf{w}} r(\mathbf{X}^\top \mathbf{v}, \mathbf{S}^\top \mathbf{w}) \end{aligned}$$

Thus, QVEC-CCA ensures invariance to the matrices' bases' rotation, and since it is a single correlation, it produces a score in $[-1, 1]$.

3 Linguistic Dimension Word Vectors

Both QVEC and QVEC-CCA rely on a matrix of linguistic properties constructed from a manually crafted linguistic resource. Linguistic resources are invaluable as they capture generalizations made by domain experts. However, resource construction is expensive, therefore it is not always possible to find an existing resource that captures exactly the set of optimal lexical properties for a downstream task. Resources that capture more coarse-grained, general properties can be used instead, for example, WordNet for semantic evaluation (Fellbaum, 1998), or Penn Treebank (Marcus et al., 1993, PTB) for syntactic evaluation. Since these properties are not an exact match to the task, the intrinsic evaluation tests for a necessary (but possibly not sufficient) set of generalizations.

Semantic vectors. To evaluate the semantic content of word vectors, Tsvetkov et al. (2015) exploit supersense annotations in a WordNet-annotated corpus—SemCor (Miller et al., 1993). The resulting supersense-dimension matrix has 4,199 rows (supersense-annotated nouns and verbs that occur in SemCor at least 5 times[2]), and 41 columns: 26 for nouns and 15 for verbs. Example vectors are shown in table 1.

WORD	NN.ANIMAL	NN.FOOD	...	VB.MOTION
fish	0.68	0.16	...	0.00
duck	0.31	0.00	...	0.69
chicken	0.33	0.67	...	0.00

Table 1: Linguistic dimension word vector matrix with semantic vectors, constructed using SemCor.

Syntactic vectors. Similar to semantic vectors, we construct syntactic vectors for all words with 5 or more occurrences in the training part of the PTB. Vector dimensions are probabilities of the part-of-speech (POS) annotations in the corpus. This results in 10,865 word vectors with 45 interpretable columns, each column corresponds to a POS tag from the PTB; a snapshot is shown in table 2.

4 Experiments

Experimental setup. We replicate the experimental setup of Tsvetkov et al. (2015):

[2]We exclude sparser word types to avoid skewed probability estimates of senses of polysemous words.

WORD	PTB.NN	PTB.VB	...	PTB.JJ
spring	0.94	0.02	...	0.00
fall	0.49	0.43	...	0.00
light	0.52	0.02	...	0.41

Table 2: Linguistic dimension word vector matrix with syntactic vectors, constructed using PTB.

- We first train 21 word vector models: variants of CBOW and Skip-Gram models (Mikolov et al., 2013); their modifications CWindow, Structured Skip-Gram, and CBOW with Attention (Ling et al., 2015b; Ling et al., 2015a); GloVe vectors (Pennington et al., 2014); Latent Semantic Analysis (LSA) based vectors (Church and Hanks, 1990); and retrofitted GloVe and LSA vectors (Faruqui et al., 2015).
- We then evaluate these word vector models using existing *intrinsic* evaluation methods: QVEC and the proposed QVEC-CCA, and also word similarity tasks using the WordSim353 dataset (Finkelstein et al., 2001, WS-353), MEN dataset (Bruni et al., 2012), and SimLex-999 dataset (Hill et al., 2014, SimLex).[3]
- In addition, the same vectors are evaluated using *extrinsic* text classification tasks. Our semantic benchmarks are four binary categorization tasks from the 20 Newsgroups (20NG); sentiment analysis task (Socher et al., 2013, Senti); and the metaphor detection (Tsvetkov et al., 2014, Metaphor).
- Finally, we compute the Pearson's correlation coefficient r to quantify the linear relationship between the intrinsic and extrinsic scorings. The higher the correlation, the better suited the intrinsic evaluation to be used as a proxy to the extrinsic task.

We extend the setup of Tsvetkov et al. (2015) with two syntactic benchmarks, and evaluate QVEC-CCA with the syntactic matrix. The first task is POS tagging; we use the LSTM-CRF model (Lample et al., 2016), and the second is dependency parsing (Parse), using the stack-LSTM model of Dyer et al. (2015).

Results. To test the efficiency of QVEC-CCA in capturing the semantic content of word vectors, we evaluate how well the scores correspond to the scores of word vector models on semantic benchmarks. QVEC and QVEC-CCA employ the semantic supersense-dimension vectors described in §3.

[3]We employ an implementation of a suite of word similarity tasks at `wordvectors.org` (Faruqui and Dyer, 2014).

In table 3, we show correlations between intrinsic scores (word similarity/QVEC/QVEC-CCA) and extrinsic scores across semantic benchmarks for 300-dimensional vectors. QVEC-CCA obtains high positive correlation with all the semantic tasks, and outperforms QVEC on two tasks.

	20NG	Metaphor	Senti
WS-353	0.55	0.25	0.46
MEN	0.76	0.49	0.55
SimLex	0.56	0.44	0.51
QVEC	0.74	0.75	0.88
QVEC-CCA	0.77	0.73	0.93

Table 3: Pearson's correlations between word similarity/QVEC/QVEC-CCA scores and the downstream text classification tasks.

In table 4, we evaluate QVEC and QVEC-CCA on syntactic benchmarks. We first use linguistic vectors with dimensions corresponding to part-of-speech tags (denoted as PTB). Then, we use linguistic vectors which are a concatenation of the semantic and syntactic matrices described in §3 for words that occur in both matrices; this setup is denoted as PTB+SST.

		POS	Parse
PTB	**WS-353**	-0.38	0.68
	MEN	-0.32	0.51
	SimLex	0.20	-0.21
	QVEC	0.23	0.39
	QVEC-CCA	0.23	0.50
PTB+SST	**QVEC**	0.28	0.37
	QVEC-CCA	0.23	0.63

Table 4: Pearson's correlations between word similarity/QVEC/QVEC-CCA scores and the downstream syntactic tasks.

Although some word similarity tasks obtain high correlations with syntactic applications, these results are inconsistent, and vary from a high negative to a high positive correlation. Conversely, QVEC and QVEC-CCA consistently obtain moderate-to-high positive correlations with the downstream tasks.

Comparing performance of QVEC-CCA in PTB and PTB+SST setups sheds light on the importance of linguistic signals captured by the linguistic matrices. Appending supersense-annotated columns to the linguistic matrix which already contains POS-annotated columns does not affect correlations of QVEC-CCA with the POS tagging task,

since the additional linguistic information is not relevant for approximating how well dimensions of word embeddings encode POS-related properties. In the case of dependency parsing—the task which encodes not only syntactic, but also semantic information (e.g., captured by subject-verb-object relations)—supersenses introduce relevant linguistic signals that are not present in POS-annotated columns. Thus, appending supersense-annotated columns to the linguistic matrix improves correlation of QVEC-CCA with the dependency parsing task.

5 Conclusion

We introduced QVEC-CCA—an approach to intrinsic evaluation of word embeddings. We also showed that both QVEC and QVEC-CCA are not limited to semantic evaluation, but are general approaches, that can evaluate word vector content with respect to desired linguistic properties. Semantic and syntactic linguistic features that we use to construct linguistic dimension matrices are rather coarse, thus the proposed evaluation can approximate a range of downstream tasks, but may not be sufficient to evaluate finer-grained features. In the future work we propose to exploit existing semantic, syntactic, morphological, and typological resources (e.g., universal dependencies treebank (Agić et al., 2015) and WALS (Dryer and Haspelmath, 2013)), and also multilingual resources (e.g., Danish supersenses (Martínez Alonso et al., 2015)) to construct better linguistic matrices, suited for evaluating vectors used in additional NLP tasks.

Acknowledgments

This work was supported by the National Science Foundation through award IIS-1526745. We thank Benjamin Wilson for helpful comments.

References

Željko Agić, Maria Jesus Aranzabe, Aitziber Atutxa, Cristina Bosco, Jinho Choi, Marie-Catherine de Marneffe, Timothy Dozat, Richárd Farkas, Jennifer Foster, Filip Ginter, Iakes Goenaga, Koldo Gojenola, Yoav Goldberg, Jan Hajič, Anders Trærup Johannsen, Jenna Kanerva, Juha Kuokkala, Veronika Laippala, Alessandro Lenci, Krister Lindén, Nikola Ljubešić, Teresa Lynn, Christopher Manning, Héctor Alonso Martínez, Ryan McDonald, Anna Missilä, Simonetta Montemagni, Joakim Nivre, Hanna Nurmi, Petya Osenova, Slav Petrov, Jussi Piitulainen, Barbara Plank, Prokopis Prokopidis, Sampo Pyysalo, Wolfgang Seeker, Mojgan Seraji, Natalia Silveira, Maria Simi, Kiril Simov, Aaron Smith, Reut Tsarfaty, Veronika Vincze, and Daniel Zeman. 2015. Universal dependencies 1.1. LINDAT/CLARIN digital library at Institute of Formal and Applied Linguistics, Charles University in Prague.

Mohit Bansal, Kevin Gimpel, and Karen Livescu. 2014. Tailoring continuous word representations for dependency parsing. In *Proc. of ACL*.

Elia Bruni, Gemma Boleda, Marco Baroni, and Nam-Khanh Tran. 2012. Distributional semantics in technicolor. In *Proc. of ACL*.

Kenneth Ward Church and Patrick Hanks. 1990. Word association norms, mutual information, and lexicography. *Computational Linguistics*, 16(1):22–29.

Matthew S. Dryer and Martin Haspelmath, editors. 2013. *WALS Online*. Max Planck Institute for Evolutionary Anthropology. `http://wals.info/`.

Chris Dyer, Miguel Ballesteros, Wang Ling, Austin Matthews, and Noah A. Smith. 2015. Transition-based dependency parsing with stack long short-term memory. In *Proc. of ACL*.

Manaal Faruqui and Chris Dyer. 2014. Community evaluation and exchange of word vectors at wordvectors.org. In *Proc. of ACL (Demonstrations)*.

Manaal Faruqui, Jesse Dodge, Sujay Kumar Jauhar, Chris Dyer, Noah A. Smith, and Eduard Hovy. 2015. Retrofitting word vectors to semantic lexicons. In *Proc. of NAACL*.

Christiane Fellbaum, editor. 1998. *WordNet: an electronic lexical database*. MIT Press.

Lev Finkelstein, Evgeniy Gabrilovich, Yossi Matias, Ehud Rivlin, Zach Solan, Gadi Wolfman, and Eytan Ruppin. 2001. Placing search in context: the concept revisited. In *Proc. of WWW*.

Jiang Guo, Wanxiang Che, Haifeng Wang, and Ting Liu. 2014. Revisiting embedding features for simple semi-supervised learning. In *Proc. of EMNLP*.

David R Hardoon, Sandor Szedmak, and John Shawe-Taylor. 2004. Canonical correlation analysis: An overview with application to learning methods. *Neural Computation*, 16(12):2639–2664.

Felix Hill, Roi Reichart, and Anna Korhonen. 2014. SimLex-999: Evaluating semantic models with (genuine) similarity estimation. *CoRR*, abs/1408.3456.

Guillaume Lample, Miguel Ballesteros, Sandeep Subramanian, Kazuya Kawakami, and Chris Dyer. 2016. Neural architectures for named entity recognition. In *Proc. of NAACL*.

Angeliki Lazaridou, Eva Maria Vecchi, and Marco Baroni. 2013. Fish transporters and miracle homes: How compositional distributional semantics can help NP parsing. In *Proc. of EMNLP*.

Wang Ling, Lin Chu-Cheng, Yulia Tsvetkov, Silvio Amir, Ramon Fermandez, Chris Dyer, Alan W Black, and Isabel Trancoso. 2015a. Not all contexts are created equal: Better word representations with variable attention. In *Proc. of EMNLP*.

Wang Ling, Chris Dyer, Alan Black, and Isabel Trancoso. 2015b. Two/too simple adaptations of `word2vec` for syntax problems. In *Proc. of NAACL*.

Mitchell P Marcus, Mary Ann Marcinkiewicz, and Beatrice Santorini. 1993. Building a large annotated corpus of English: The penn treebank. *Computational Linguistics*, 19(2):313–330.

Héctor Martínez Alonso, Anders Johannsen, Sussi Olsen, Sanni Nimb, Nicolai Hartvig Sørensen, Anna Braasch, Anders Søgaard, and Bolette Sandford Pedersen. 2015. Supersense tagging for Danish. In *Proc. of NODALIDA*, page 21.

Tomas Mikolov, Kai Chen, Greg Corrado, and Jeffrey Dean. 2013. Efficient estimation of word representations in vector space. In *Proc. of ICLR*.

George A. Miller, Claudia Leacock, Randee Tengi, and Ross T. Bunker. 1993. A semantic concordance. In *Proc. of HLT*, pages 303–308.

Jeffrey Pennington, Richard Socher, and Christopher D. Manning. 2014. GloVe: Global vectors for word representation. In *Proc. of EMNLP*.

Richard Socher, Alex Perelygin, Jean Wu, Jason Chuang, Christopher D. Manning, Andrew Y. Ng, and Christopher Potts. 2013. Recursive deep models for semantic compositionality over a sentiment treebank. In *Proc. of EMNLP*.

Christian Szegedy, Wojciech Zaremba, Ilya Sutskever, Joan Bruna, Dumitru Erhan, Ian Goodfellow, and Rob Fergus. 2014. Intriguing properties of neural networks. In *Proc. of ICLR*.

Yulia Tsvetkov, Leonid Boytsov, Anatole Gershman, Eric Nyberg, and Chris Dyer. 2014. Metaphor detection with cross-lingual model transfer. In *Proc. of ACL*, pages 248–258.

Yulia Tsvetkov, Manaal Faruqui, Wang Ling, Guillaume Lample, and Chris Dyer. 2015. Evaluation of word vector representations by subspace alignment. In *Proc. of EMNLP*, pages 2049–2054.

Joseph Turian, Lev Ratinov, and Yoshua Bengio. 2010. Word representations: a simple and general method for semi-supervised learning. In *Proc. of ACL*.