Word Embeddings based on Fixed-Size Ordinally Forgetting Encoding

Joseph Sanu[†], **Mingbin Xu**[†], **Hui Jiang**[†] and **Quan Liu**[‡]
[†]Department of Electrical Engineering and Computer Science,
York University, 4700 Keele Street, Toronto, Ontario, M3J 1P3, Canada
[‡]Department of EEIS, University of Science and Technology of China, Hefei, China
cse83186@cse.yorku.ca, xmb@cse.yorku.ca, hj@cse.yorku.ca, quanliu@mail.ustc.edu.cn

Abstract

In this paper, we propose to learn word embeddings based on the recent fixed-size ordinally forgetting encoding (FOFE) method, which can almost uniquely encode any variable-length sequence into a fixed-size representation. We use FOFE to fully encode the left and right context of each word in a corpus to construct a novel word-context matrix, which is further weighted and factorized using truncated SVD to generate low-dimension word embedding vectors. We have evaluated this alternative method in encoding word-context statistics and show the new FOFE method has a notable effect on the resulting word embeddings. Experimental results on several popular word similarity tasks have demonstrated that the proposed method outperforms many recently popular neural prediction methods as well as the conventional SVD models that use canonical count based techniques to generate word context matrices.

1 Introduction

Low dimensional vectors as word representations are very popular in NLP tasks such as inferring semantic similarity and relatedness. Most of these representations are based on either matrix factorization or context sampling described by (Baroni et al., 2014) as count or predict models. The basis for both models is the distributional hypothesis (Harris, 1954), which states that words that appear in similar contexts have similar meaning. Traditional context representations have been obtained by capturing co-occurrences of words from a fixed-size window relative to the focus word. This representation however does not encompass the entirety of the context surrounding the focus word. Therefore, the distributional hypothesis is not being taken advantage of to the fullest extent. In this work, we seek to capture these contexts through the fixed-size ordinally forgetting encoding (FOFE) method, recently proposed in (Zhang et al., 2015b). In addition to just capturing word co-occurrences, we attempt to use the FOFE to encode the full contexts of each focus word, including the order information of the context sequences. We believe the full encoding of contexts can enhance the resulting word embedding vectors, derived by factoring the corresponding word-context matrix. As argued in (Zhang et al., 2015b), the FOFE method can almost uniquely encode discrete sequences of varying lengths into a fixed-size code, and this encoding method was used to address the challenges of a limited size window when using deep neural networks for language modeling. The resulting algorithm fulfills the needs of keeping long term dependency while being fast. The word order in a sequence is modeled by FOFE using an ordinally-forgetting mechanism which encodes each position of every word in the sequence.

In this paper, we elaborate how to use the FOFE to fully encode context information of each focus word in text corpora, and present a new method to construct the word-context matrix for word embedding, which may be weighted and factorized as in traditional vector space models (Turney and Pantel, 2010). Next, we report our experimental results on several popular word similarity tasks, which demonstrate that the proposed FOFE-based approach leads to significantly better performance in these tasks, comparing with the conventional vector space models as well as the popular neural prediction methods, such as *word2vec*, *GloVe* and more recent *Swivel*. Finally, this paper will conclude with the analysis and prospects of com-

Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing, pages 310–315
Copenhagen, Denmark, September 7–11, 2017.

bining this approach with other methods.

2 Related Work

There has been some debate as to what the optimal length of a text should be for measuring word similarity. Word occurrences from a fixed context window of words can be used to represent a context (Lund and Burgess, 1996). The word co-occurrence frequencies are based on fixed windows spanning in both directions from the focus word. This is then used to create a word-context matrix from which row vectors can be used to measure word similarity. A weighting step is usually applied to highlight words with close association in the co-occurrence matrix, and the truncated SVD is used to factorize the weighted matrix to generate low-dimension word vectors. Recently, (Mikolov et al., 2013a) has introduced an alternative way to generate word embeddings using the skipgram model trained with stochastic gradient descent and negative sampling, named as SGNS. SGNS tries to maximize the dot product between $w \cdot c$ where both a word w and a context c are obtained from observed word-context pairs, and meanwhile it also tries to minimize the dot product between $w \cdot c'$ where c' is a negative sample representing some contexts that are not observed in the corpus. More recently, (Levy and Goldberg, 2014) has showed that the objective function of SGNS is essentially seeking to minimize the difference between the models estimate and the log of co-occurrence count. Their finding has shown that the optimal solution is a weighted factorization of a pointwise mutual information matrix shifted by the log of the number of negative samples.

SGNS and GloVe (Pennington et al., 2014) select a fixed window of usually 5 words or less around a focus word to encode its context and the word order information within the window is completely ignored. Other attempts to fully capture the contexts have been successful with the use of recurrent neural networks (RNNs) but these methods are much more expensive to run over large corpora when comparing with the proposed FOFE method in this paper. Some previous approaches to encode order information, such as such as BEAGLE (Jones and Mewhort, 2007) and Random Permutations (Sahlgren et al., 2008), typically require the use of expensive operations such as convolution and permutation to process all n-grams within a context window to memorize order information for a given word. On the contrary, the FOFE methods only use a simple recursion to process a sentence once to memorize both context and order information for all words in the sentence.

3 FOFE based Embedding

To capture the full essence of the distributional hypothesis, we need to fully encode the left and right context of each focus word in the text, and further take into accounts that words closer to the focus word should play a bigger role in representing the context relevant to the focus word than other words locating much farther away. Traditional co-occurrence word-context matrixes fail to address these concerns of context representation.

In this work, we propose to make use of the fixed-size ordinally-forgetting encoding (FOFE) method, proposed in (Zhang et al., 2015b) as a unique encoding method for any variable-length sequence of discrete words.

Given a vocabulary of size K, FOFE uses 1-of-K one-hot representation to represent each word. To encode any variable-length sequence of words, FOFE generates the code using a simple recursive formula from the first word (w_1) to the last one (w_T) of the sequence: (assume $z_0 = 0$)

$$z_t = \alpha \cdot z_{t-1} + e_t \; (1 \leq t \leq T) \qquad (1)$$

where z_t denotes the FOFE code for the partial sequence up to word w_t, α is a constant forgetting factor, and e_t denotes the one-hot vector representation of word w_t. In this case, the code z_T may be viewed as a fixed-size representation of any sequence of $\{w_1, w_2, \cdots, w_T\}$. For example, assume we have three symbols in vocabulary, e.g., *A*, *B*, *C*, whose 1-of-K codes are $[1, 0, 0]$, $[0, 1, 0]$ and $[0, 0, 1]$ respectively. When calculating from left to right, the FOFE code for the sequence $\{ABC\}$ is $[\alpha^2, \alpha, 1]$, and that of $\{ABCBC\}$ is $[\alpha^4, \alpha + \alpha^3, 1 + \alpha^2]$.

The uniqueness of the FOFE code is made evident if the original sequence can be unequivocally recovered from the given FOFE code. According to (Zhang et al., 2015b), FOFE codes have some nice theoretical properties to ensure the uniqueness, as exemplified by the following two theorems [1]:

Theorem 1 *If the forgetting factor α satisfies $0 < \alpha \leq 0.5$,* FOFE *is unique for any K and T.*

[1] See (Zhang et al., 2015a) for the proof of these two theorems.

Theorem 2 *For* $0.5 < \alpha < 1$*, given any finite values of* K *and* T*,* FOFE *is almost unique everywhere for* $\alpha \in (0.5, 1.0)$*, except only a finite set of countable choices of* α*.*

Finally, for alpha values less than or equal to 0.5 and greater than 0, the FOFE is unique for any sequence. For alpha values greater than 0.5, the chance of collision is extremely low and the FOFE is unique in almost all cases. Too find more about the theoretical correctness of FOFE, please refer to (Zhang et al., 2015b). In other words, the FOFE codes can almost uniquely encode any sequences, serving as a fixed-size but theoretically lossless representation for any variable-length sequences.

In this work, we propose to use FOFE to encode the full context where each focus word appears in text. As shown in Figure 1, the left context of a focus word, i.e., *bank*, may be viewed as a sequence and encoded as a FOFE code L from the left to right while its right context is encoded as another FOFE code R from right to left. When a proper forgetter factor α is chosen, the two FOFE codes can almost fully represent the context of the focus word. If the focus word appears multiple times in text, a pair of FOFE codes $[L, R]$ is generated for each occurrence. Next, a mean vector is calculated for each word from all of its occurrences in text. Finally, as shown in Figure 1, we may line up these mean vectors (one word per row) to form a new word-context matrix, called the FOFE matrix here.

4 PMI-based Weighting and SVD-based Matrix Factorization

We further weight the above FOFE matrix using the standard positive pointwise mutual information (PMI) (Church and Hanks, 1990) which has been shown to be of benefit for regular word-context matrices (Pantel and Lin, 2002). PMI is used as a measure of association between a word and a context. PMI tries to compute the association probabilities based on co-occurrence frequencies. Positive pointwise mutual information is a commonly adopted approach where all negative values in the PMI matrix are replaced with zero. The PMI-based weighting function is critical here since it helps to highlight the more surprising events in original word-context matrix.

There are significant benefits in working with low-dimensional dense vectors, as noted by (Deerwester et al., 1990) with the use of truncated singular value decomposition (SVD). Here, we also use truncated SVD to factorize the above weighted FOFE matrix as the product of three dense matrices U, Σ, V^T, where U and V^T have orthonormal columns and Σ is a diagonal matrix consisting of singular values. If we select Σ to be of rank d, its diagonal values represent the top d singular values, and U_d can be used to represent all word embeddings with d dimensions where each row represents a word vector.

5 Experiments

We conducted experiments on several popular word similarity data sets and compare our FOFE method with other existing word embedding models in these tasks. In this work, we opt to use five data sets: *WordSim353* (Finkelstein et al., 2001), *MEN* (Bruni et al., 2012), *Mechanical Turk* (Radinsky et al., 2011), *Rare Words* (Luong et al., 2013) and *SimLex-999* (Hill et al., 2015). The word similarity performance is evaluated based on the Spearman rank correlation coefficient obtained by comparing cosine distance between word vectors and human assigned similarity scores.

For our training data, we use the standard *enwiki9* corpus which contains 130 million words. The pre-processing stage includes discarding extremely long sentences, tokenizing, lowercasing and splitting each sentence as a context. Our vocabulary size is chosen to be 80,000 for the most frequent words in the corpus. All words not in the vocabulary are replaced with the token <*unk*>. In this work, we use a python-based library, called *scipy* [2], to perform truncated SVD to factorize all word-context matrices.

5.1 Experimental Setup

Our first baseline is the conventional vector space model (VSM) (Turney and Pantel, 2010), relying on the PMI-weighted co-occurrence matrix with dimensionality reduction performed using truncated SVD. The dimension of word vectors is chosen to be 300 and this number is kept the same for all models examined in this paper. Our main goal is to outperform VSM as the model proposed in this paper also uses SVD based matrix factorization. This allows for appropriate comparisons between the different word encoding methods.

[2] See `http://docs.scipy.org/doc/scipy/reference/`.

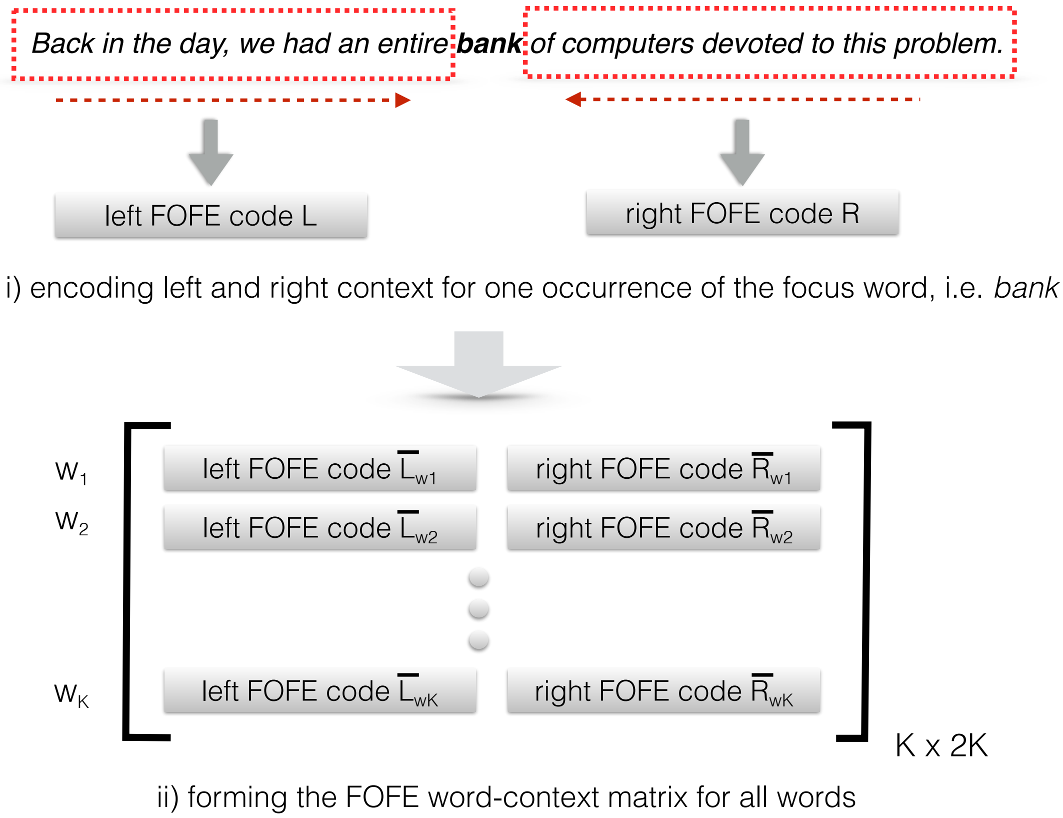

Figure 1: i) encoding left and right contexts of each focus word with FOFE and ii) forming the FOFE word-context matrix.

For the purpose of completeness, the other non-SVD based embedding models, mainly the more recent neural prediction methods, are also compared in our experiments. As a result, we build the second baseline using the skip-gram model provided by the `word2vec` software package (Mikolov et al., 2013a), denoted as SGNS. The word embeddings are generated using the recommended hyper-parameters from (Levy et al., 2015). Their findings show a larger number of negative samples is preferable and increments on the window size have minimal improvements on word similarity tasks. In our experiments the number of negative samples is set to 5 and the window size is set to 5. In addition, we set the subsampling rate to 10^{-4} and run 3 iterations for training. In adition to SGNS, we also obtained results for CBOW, GloVe (Pennington et al., 2014) and Swivel (Shazeer et al., 2016) models using similar recommended settings. While the window size has a fixed limit in the baseline models, our model does not have a window size parameter as the entire sentence is fully captured as well as distinctions between left and right contexts when generating the FOFE codes. The impact of closer context words is further highlighted by the use of the forgetting factor which is unique to the FOFE based word embedding.

Finally, we use the FOFE codes to construct the word-context matrix and generate word embedding as described in sections 3 and 4. Throughout our experiments, we have chosen to use a constant forgetting factor $\alpha = 0.7$. There is no significant difference in word similarity scores after experimenting with different α values between $[0.6, 0.9]$ when generating FOFE codes.

We have applied the same hyperparameters to both VSM and FOFE methods and fine-tune them based on the recommended settings provided in (Levy et al., 2015). Although it has been previously reported that context distribution smoothing (Mikolov et al., 2013b) can provide a net positive effect, it did not yield significant gains in our experiments. On the other hand, the eigenvalue

Table 1: The best achieved performance of various word embedding models on all five examined word similarity tasks.

Method	WordSim353	MEN	Mech Turk	Rare Words	SimLex-999
VSM+SVD	0.7109	0.7130	0.6258	0.4813	**0.3866**
CBOW	0.6763	0.6768	0.6621	0.4280	0.3549
GloVe	0.5873	0.6350	0.5831	0.3934	0.2883
SGNS	0.7028	0.6689	0.6187	0.4360	0.3709
Swivel	0.7303	0.7246	**0.7024**	0.4430	0.3323
FOFE+SVD	**0.7580**	**0.7637**	0.6525	**0.5002**	**0.3866**

weighting parameter tuning (Caron, 2001) proved to be incredibly effective for some datasets but ineffectual in others. The net benefit however is palpable and we include it for both VSM and FOFE methods.

5.2 Results and Discussion

The best results of all word embedding models are summarized in Table 1 for all five examined data sets, which include the the traditional count based VSM with SVD alongside SGNS using `word2vec` and our proposed FOFE word embeddings. The most discernible piece of information from the table is that the FOFE method significantly outperforms the traditional count based VSM method on most of these word similarity tasks. The results in Table 1 show that substantial gains are obtained by FOFE in *WordSim353*, *MEN* and *Rare Words* data sets. The *MEN* dataset shows a 7% relative improvement over the conventional VSM.

Among all of these five data sets, the proposed FOFE word embedding significantly outperforms VSM in four tasks while yielding similar performance as VSM in the last data set, i.e. *SimLex-999*. FOFE also outperforms all the other models except Swivel in the *Mech Turk* dataset. It is important to note that this paper does not state that SVD is obligatory to obtain the best model. The FOFE method can be complemented with other models such as Swivel in place of count based encoding methods. It is also theoretically guaranteed that the original sentence is perfectly recoverable from this FOFE code. This theoretical guarantee is clearly missing in previous methods to encode word order information, such as both BEAGLE and Random Permutations. It is evident that overall the FOFE encoding method does achieve significant gains in performance in these word similarity tests over the traditional VSM method that applies the same factorization method. This is substantial as (Levy et al., 2015) demonstrates that larger window sizes when using SVD does not payoff and the optimal context window is 2. We establish that we can indeed encode more information into our embedding with the FOFE codes.

In summary, our experimental results show great promise in using the FOFE encoding to represent word contexts for traditional matrix factorization methods. As for future work, the FOFE encoding method may be combined with other popular algorithms, such as Swivel, to replace the co-occurrence statistics based on a fixed window size.

6 Conclusion

The ability to capture the full context without restriction can play a crucial factor in generating superior word embeddings that excel in NLP tasks. The fixed-size ordinally forgetting encoding (FOFE) has the ability to seize large contexts while discriminating contexts that are farther away as being less significant. Conventional embeddings are derived from ambiguous co-occurrence statistics that fail to adequately discriminate contexts words even within the fixed-size window. The FOFE encoding technique trumps other approaches in its ability to procure the state of the art results in several word similarity tasks when combined with prominent factorization practices.

Acknowledgments

This work is partially supported by a Discovery Grant from Natural Sciences and Engineering Research Council (NSERC) of Canada, and a research donation from iFLYTEK Co., Hefei, China.

References

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 (1)*, pages 238–247.

Elia Bruni, Gemma Boleda, Marco Baroni, and Nam-Khanh Tran. 2012. Distributional semantics in technicolor. In *Proceedings of the 50th Annual Meeting of the Association for Computational Linguistics*, pages 136–145. Association for Computational Linguistics.

John Caron. 2001. Experiments with lsa scoring: Optimal rank and basis. In *Proceedings of the SIAM Computational Information Retrieval Workshop*, pages 157–169.

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

Scott Deerwester, Susan T Dumais, George W Furnas, Thomas K Landauer, and Richard Harshman. 1990. Indexing by latent semantic analysis. *Journal of the American society for information science*, 41(6):391.

Lev Finkelstein, Evgeniy Gabrilovich, Yossi Matias, Ehud Rivlin, Zach Solan, Gadi Wolfman, and Eytan Ruppin. 2001. Placing search in context: The concept revisited. In *Proceedings of the 10th international conference on World Wide Web*, pages 406–414. ACM.

Zellig S Harris. 1954. Distributional structure. *Word*, 10(2-3):146–162.

Felix Hill, Roi Reichart, and Anna Korhonen. 2015. Simlex-999: Evaluating semantic models with (genuine) similarity estimation. *Computational Linguistics*.

M.N Jones and D.J.K Mewhort. 2007. Representing word meaning and order information in a composite holographic lexicon. *Psychological Review*, 114(1):1–37.

Omer Levy and Yoav Goldberg. 2014. Neural word embedding as implicit matrix factorization. In *Advances in Neural Information Processing Systems*, pages 2177–2185.

Omer Levy, Yoav Goldberg, and Ido Dagan. 2015. Improving distributional similarity with lessons learned from word embeddings. *Transactions of the Association for Computational Linguistics*, 3:211–225.

Kevin Lund and Curt Burgess. 1996. Producing high-dimensional semantic spaces from lexical co-occurrence. *Behavior Research Methods, Instruments, and Computers*, 28(2):203–208.

Thang Luong, Richard Socher, and Christopher D Manning. 2013. Better word representations with recursive neural networks for morphology. In *CoNLL*, pages 104–113. Citeseer.

Tomas Mikolov, Kai Chen, Greg Corrado, and Jeffrey Dean. 2013a. Efficient estimation of word representations in vector space. *arXiv preprint arXiv:1301.3781*.

Tomas Mikolov, Ilya Sutskever, Kai Chen, Greg S Corrado, and Jeff Dean. 2013b. Distributed representations of words and phrases and their compositionality. In *Advances in neural information processing systems*, pages 3111–3119.

Patrick Pantel and Dekang Lin. 2002. Discovering word senses from text. In *Proceedings of the eighth ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 613–619.

Jeffrey Pennington, Richard Socher, and Christopher Manning. 2014. Glove: Global vectors for word representation. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 1532–1543, Doha, Qatar. Association for Computational Linguistics.

Kira Radinsky, Eugene Agichtein, Evgeniy Gabrilovich, and Shaul Markovitch. 2011. A word at a time: computing word relatedness using temporal semantic analysis. In *Proceedings of the 20th international conference on World wide web*, pages 337–346. ACM.

Magnus Sahlgren, Anders Holst, and Kanerva Pentti. 2008. Permutations as a means to encode order in word space. In *In Proceedings of the 30th Annual Conference of the Cognitive Science Society*, pages 1300–1305.

Noam Shazeer, Ryan Doherty, Colin Evans, and Chris Waterson. 2016. Swivel: Improving embeddings by noticing whats missing. *arXiv preprint arXiv:1602.02215*.

Peter D. Turney and Patrick Pantel. 2010. From frequency to meaning: Vector space models of semantics. *Journal of Artificial Intelligence Research*, 37:141–188.

Shiliang Zhang, Hui Jiang, Mingbin Xu, Junfeng Hou, and Lirong Dai. 2015a. A fixed-size encoding method for variable-length sequences with its application to neural network language models. *arXiv preprint arXiv:1505.01504*.

Shiliang Zhang, Hui Jiang, Mingbin Xu, Junfeng Hou, and Lirong Dai. 2015b. The fixed-size ordinally-forgetting encoding method for neural network language models. In *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing*, pages 495–500, Beijing, China. Association for Computational Linguistics.