Second-order contexts from lexical substitutes for few-shot learning of word representations

Qianchu Liu, Diana McCarthy, Anna Korhonen
Language Technology Lab, University of Cambridge
English Faculty Building, 9 West Road, Cambridge CB3 9DA, United Kingdom
ql261@cam.ac.uk, diana@dianamccarthy.co.uk, alk23@cam.ac.uk

Abstract

There is a growing awareness of the need to handle rare and unseen words in word representation modelling. In this paper, we focus on few-shot learning of emerging concepts that fully exploits only a few available contexts. We introduce a substitute-based context representation technique that can be applied on an existing word embedding space. Previous context-based approaches to modelling unseen words only consider bag-of-word first-order contexts, whereas our method aggregates contexts as second-order substitutes that are produced by a sequence-aware sentence completion model. We experimented with three tasks that aim to test the modelling of emerging concepts. We found that these tasks show different emphasis on first and second order contexts, and our substitute-based method achieved superior performance on naturally-occurring contexts from corpora.

1 Introduction

As language vocabulary follows the zipfian distribution, we expect to encounter a large number of rare and unseen words no matter how large the training corpus is. The effective handling of such words is thus crucial for Natural Language Processing (NLP).

Attempts to learn rare and unseen word representations can be categorized into the following three approaches: (1) constructing target word embeddings from the subword components (Pinter et al., 2017; Bojanowski et al., 2017), (2). leveraging definitions or relational structures from external resources such as Wordnet (Bahdanau et al., 2017; Pilehvar and Collier, 2017), and (3) modelling the target word from few available contexts. Our paper falls into the last approach.

We demonstrate improvements in performance by employing an alternative context representation, second-order lexical substitutes, as opposed to the traditional bag of word context representations. In line with previous research in this area, we evaluate our methodology on three tasks that measure the quality of the induced unseen word representation from contexts (Lazaridou et al., 2017; Herbelot and Baroni, 2017; Khodak et al., 2018). Our results reveal that the three tasks involve different types of contexts which put different emphasis on first or second order contexts. Our second-order substitute-based method achieves the best performance for modelling rare words in natural contexts from corpora. In the tasks in which both first order and second order contexts are important, the ensemble of these two types of contexts yields superior performance. [1]

2 Related work

2.1 First-order context

The most naive way of inducing new word representation from contexts is to simply take the average of context word embeddings that co-occur with the target word in a sentence. With stop words removed, this simple method has proven to be a strong baseline as shown in Lazaridou et al. (2017) and Herbelot and Baroni (2017). A potential improvement from the simple additive baseline model is that we weigh words with ISF (inverse sentence frequency). We follow the definition of ISF in Samardzhiev et al. (2018) and implement it as a baseline model in our study. More recently, Khodak et al. (2018) learn a transformation matrix to reconstruct pre-trained word embeddings, which essentially learns to highlight informative dimensions. Along a different line, Herbelot and Baroni (2017) take a high-risk learning rate and processing strategy for new words but would require the contexts that come at the beginning of the training to be maximally informative. Recent

[1] The experiments can be reproduced at https://github.com/qianchu/rare_we.git.

*Proceedings of the Eighth Joint Conference on Lexical and Computational Semantics (*SEM)*, pages 61–67
Minneapolis, June 6–7, 2019.

work implements a memory-augmented word embedding model (Sun et al., 2018) however our system shows comparable or superior performance on the two intrinsic tasks that they use (Table 1 below and Table 1 of their paper).

2.2 Second-order substitute-based context

An alternative to a bag-of-words representation is a second-order substitute vector generated by a language model for the target word's slot. For example, we can represent the context '*It is a __ move.*' as a substitute vector [big 0.35, good 0.28, bold 0.05, ...] with the numbers indicating fitness weights of each substitute in the context (Melamud et al., 2015; Yatbaz et al., 2012; Melamud et al., 2015). Melamud et al. (2016) later on introduced context2vec which trains both context and word embeddings in a similar setup to CBOW (Mikolov et al., 2013) except that the context is represented with a Bidirectional LSTM rather than as a bag of words. In this way, context2vec captures sequence information in the context, and is able to produce high-quality substitutes for a sentence-completion task, while overcoming the sparseness issues in the previous substitute-based approaches. Kobayashi et al. (2017) fine-tune this context2vec representation to compute entity representations in a discourse for the language modelling task.

A related application of second-order substitutes is word sense induction. Baskaya et al. (2013) represent contexts as second-order substitutes and apply co-occurrence modelling on top of the instance id - substitute pairs. Alagić et al. (2018) propose a similar method to our paper and showed that second-order lexical substitutes and first-order contexts complement each other in word sense induction. Our paper provides alternative evidence for the use of lexical substitutes in the setting of rare word modelling with analysis on the effect from different contexts.

.

3 Proposed Method

In this paper, we make a simple modification from the previous work by representing the context of an unseen word as the weighted sum of the lexical substitute vectors in a continuous embedding space such as the word2vec space. This can be seen as a post-processing technique applied on an existing embedding space. The substitutes and their fitness scores are generated from context2vec. Compared with the context2vec representation itself, our method isolates the effect of the second-order substitutes and can be applied on top of an existing pre-trained embedding space. For each context, we generate the top N most likely substitutes at the slot of the unseen word by computing the nearest neighbours from the context2vec context representation. [2] We then compute the centroid of these substitutes from our base word embedding space, weighted by each substitute's fitness, cosine similarity, to the context representation. Let $\mathbf{ContextVec}$ [3] be the context representation produced by context2vec, S' be the set of the top 20 substitute target word vectors produced by context2vec, S be the same 20 substitutes that we look up in our base word embedding space, and $f(\mathbf{S'_i})$ be the normalized fitness score of $\mathbf{S'_i}$ as defined in equation 1. The substitute-based context ($\mathbf{SC}$), and thus the unseen word representation for this context, is defined in equation 2. If the unseen word occurs multiple times, we average the unseen word representations across the multiple contexts.

$$f(\mathbf{S'_i}) = \frac{cosine(\mathbf{ContextVec}, \mathbf{S'_i})}{\sum_{j=1}^{20} cosine(\mathbf{ContextVec}, \mathbf{S'_j})} \quad (1)$$

$$\mathbf{SC} = \sum_{i=1}^{20} f(\mathbf{S'_i}) * \mathbf{S_i} \quad (2)$$

To directly compare with the previous studies, we take the word2vec embedding model and the 1.6B Wikipedia training corpus provided by Herbelot and Baroni (2017) for our substitute-based method and for training Context2vec. Model parameters for training Context2vec, as listed in Appendix A, are fine-tuned on the training sets of the intrinsic tasks as there are no development sets.

4 The definitional Nonce dataset (Nonce)

Nonce is introduced in Herbelot and Baroni (2017) as a task that challenges the models to reconstruct target word embeddings from single wikipedia definitions. The quality of the representations is evaluated by measuring how close they are to the original word embeddings trained from the whole

[2] From experiments on the training sets of the tasks (Notice that there are no development sets), we found that N=20 is optimal.

[3] Symbols in bold indicate vectors

Methods	Nonce		Chimera		
	MRR	Med. Rank	2 Sent.	4 Sent.	6 Sent.
word2vec (Lazaridou et al., 2017)	0.00007	111012	0.1459	0.2457	0.2498
Additive (Lazaridou et al., 2017)	0.03686	861	0.3376	0.3624	0.4080
Additive ISF	0.04493	531	0.3964	**0.4016**	0.4107
nonce2vec (Herbelot and Baroni, 2017)	0.04907	623	0.3320	0.3668	0.3890
a la carte (Khodak et al., 2018)	**0.07058**	**166**	0.3634	0.3844	0.3941
mem2vec (Sun et al., 2018)	0.05416	518	0.3301	0.3717	0.3897
context2vec(Melamud et al., 2016)	0.04577	536	0.3574	0.3376	0.3692
substitutes	0.05152	1442	0.3946	0.3662	0.4424
substitutes + additive ISF	0.06074	577	**0.4167**	0.3879	**0.4469**

Table 1: Comparison with baselines and the previously-reported state-of-the-art results on the Chimera and Nonce datasets. The Chimera dataset is evaluated with Spearman Rank coefficients. The top half of the table contains first-order context methods and the bottom half has methods using second-order context or ensemble methods using first and second order.

Wikipedia corpora. Following Herbelot and Baroni (2017), we report in the Nonce columns of Table 1 the mean reciprocal rank (MRR) and median rank (Med. Rank) of the gold-vector (trained from the whole Wikipedia) in the ranked list of nearest neighbours from the induced representation in the 300 test cases.

We see strong performance from first-order context representation especially the a la carte method. Manual observations show that definitions are designed to be maximally informative with many synonyms, hypernyms or words semantically related to the target word in the context, and the first-order context models can easily exploit this information. Also, the sequential context around the target word in a definition may not reflect the context in which a target word will be typically used in a corpus. The good performance of first-order context models is therefore to be expected. Furthermore, the Nonce task tests how well the model reconstructs the original embedding but does not probe into the semantic properties or relations captured in the induced word representations. A la carte is thus especially suitable for this task as it has been explicitly trained to match the original embedding. However, we demonstrate in the following experiments that the superior performance from a la carte may not always be transferred to other tasks.

5 The Chimera dataset (Chimera)

In the Chimera dataset, Lazaridou et al. (2017) introduce unseen novel concepts (chimeras), each of which is formed by combining two related nouns

Additive ISF	substitutes
drowning	civet
drown	tapir
drowns	langur
shoos	crocodile
undresses	opossum

Table 2: Nearest neighbours produced by additive ISF and substitutes approaches for the Chimera concept elephant_bison in the context '*but his pleasure soon turns to distress when he sees that a baby ___ is stuck in the mud and drowning* .' (from the Chimeras dataset)

(For example, buffalo and elephant). Each novel concept is accompanied by 2, 4 or 6 natural contexts that originally belong to the related nouns. The model needs to induce representation for these novel concepts from the contexts. The quality of the representations is evaluated by similarity judgment with probe words. Following Herbelot and Baroni (2017) and Lazaridou et al. (2017), we report in the Chimera columns of Table 1 the average Spearman Rank coefficients against human annotations for 110 test cases in each sentence condition .

We observe that the additive ISF model turns out to be the strongest of the first-order context models, outperforming all the other previously-reported results. We see immediate improvement when we represent the context as substitutes in the 6 sentence condition. We see further improvement when combining both additive ISF (first order) and substitutes (second order contexts), which yields the best performance in 2 sentence and 6 sentence conditions. The positive effect of the

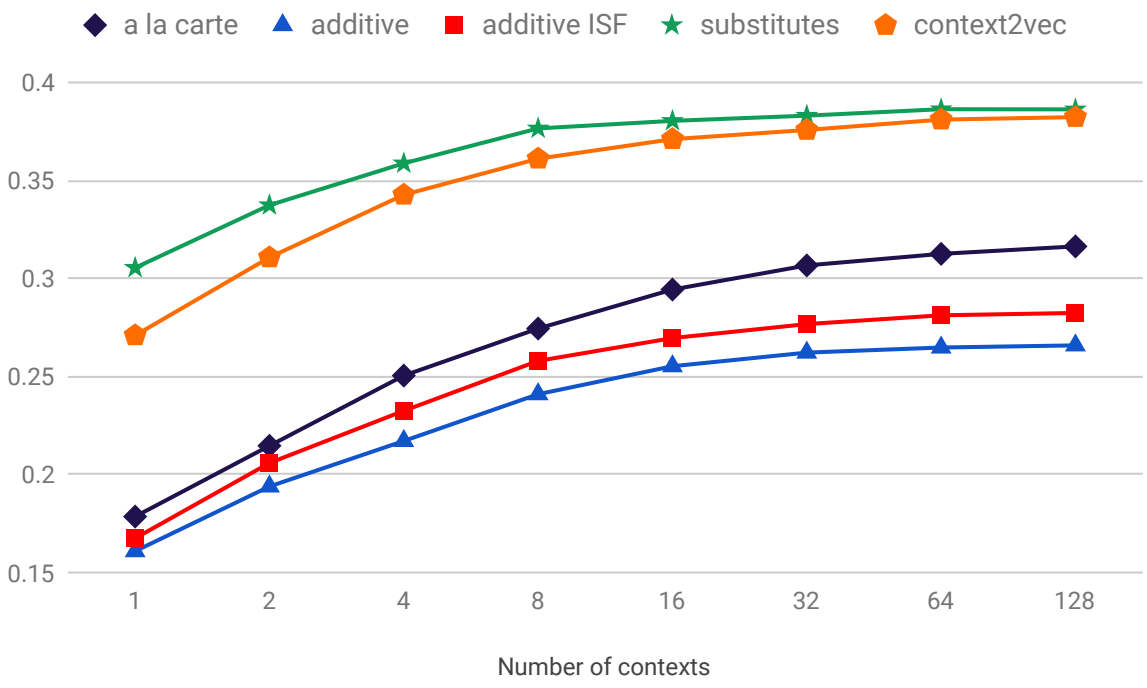

Figure 1: Spearman Rank coefficients averaged across 100 trials on CRW in various context conditions

ensemble method from combining first-order and second-order contexts shows that the two different contexts capture complementary information in this task. This is especially due to the fact that the contexts were controlled for informativeness so as to have different degrees of overlap with feature norms. Therefore at least some, but not all, contexts will have a high bag-of-word overlap with features that are semantically related to the concepts (Lazaridou et al., 2017). These contexts will easily benefit from first-order contexts alone. However, for the other contexts where there is few or even no overlap with feature norms in the context words, it is the contextual sequence, and thus second-order context, that will give the maximum information about the target word. We show such an example with the nearest neighbours of the representations induced by our substitutes model and additive ISF in Table 2. We can see that while the additive ISF representation is easily affected by unrelated words in the sentence, the substitutes approach clearly has at least identified that the target word is likely to be a kind of animal.

6 The Contextual Rare Words dataset (CRW)

The Contextual Rare Words dataset (CRW) was introduced by Khodak et al. (2018). It consists of a subset of 562 word pairs from the original Rare Word (RW) Dataset (Luong et al., 2013). For each pair, the second word is the rare word and is accompanied by 255 contexts. We follow the experiment setup in Khodak et al. (2018) and use their pre-trained vectors on the subcorpus that does not contain any of the rare words from the dataset. This subcorpus is also used to train the context2vec model that generates substitutes. As in Khodak et al. (2018), we randomly choose 2, 4, 6..128 number of contexts as separate conditions for 100 trials, and use these contexts to predict the rare word representations. Cosine similarity is computed between the rare word representation from the given rare word contexts in the trial (2,4..128) and the embedding of the other word in the pair from the pre-trained vectors. The cosine-similarity of each pair is compared against similarity judgments from human annotations. The average Spearman Rank coefficients against human annotations across the trials are reported in Figure 1. Standard deviations are reported in Appendix B.

We see dramatic improvement from the substitutes method over all the other methods including the previous state-of-the-art a la carte in this datasets which come from corpora-based natural contexts of rare words. The result here suggests that, in natural contexts, the sequence information rather than bag of words plays a more important role in predicting a target word's meaning.

We also notice that applying second order information on word2vec space consistently outperforms Context2vec alone which generates the second order substitutes. We suspect that this is because the context representation induced by context2vec is more syntactically-oriented whereas the tasks in our study mainly test semantic relations. We confirm this assumption by following Herbelot and Baroni (2017) to test the target word embeddings produced by context2vec on the MEN dataset (Bruni et al., 2014). We find that context2vec (Spearman $\rho = 0.65$) correlates less with human's semantic relatedness judgment than word2vec (Spearman $\rho = 0.75$) on this dataset. Isolating the second order information from Context2vec and applying it on the word2vec space as an external constraint effectively preserves the semantic relations present in word2vec and at the same time provides a paradigmatic view which finds a both syntactically and semantically appropriate position for the rare word.

7 Conclusion

To conclude, our paper teases apart the effect of second-order context by proposing a simple second-order substitute-based method that can post-process and improve over an existing embedding space. Our substitute-based method achieves the state-of-the-art performance when modelling emerging concepts in natural contexts from corpora. This is not surprising as the substitutes contain rich linguistic constraints from their surrounding contextual sequences to inform the word representation. We plan to investigate whether the second order information is also the key element in the success of the recently-proposed language model embeddings (Peters et al., 2018; Devlin et al., 2018), for example, by testing whether the performance of these contextualized embeddings correlate more with first-order context representation or the second-order substitute context across the different tasks in this study. However, we need further research to find ways to bring type-level and token-level representations of these contextualized embeddings into the same space for these tasks.

Also, as we found that definitions seem to exhibit different properties from natural contexts in corpora, it may be advisable to model definitions and corpora contexts differently. An aspect that we did not cover in this paper is the morphological information from target words. As contexts, definitions and subword information can provide complementary information (Schick and Schütze, 2019), in future work, we plan to leverage subwords, contexts and definitions together in modelling rare or unseen words.

Acknowledgments

We acknowledge Peterhouse College at University of Cambridge for funding Qianchu Liu's PhD research. We also appreciate the helpful discussion and feedback from Dr Ivan Vulić, Dr Nigel H. Collier, Dr Taher Pilehvar and Dr Angeliki Lazaridou. We also would like to thank Dr Aurélie Herbelot for sharing the training corpora and insightful thoughts.

References

Domagoj Alagić, Jan Šnajder, and Sebastian Padó. 2018. Leveraging lexical substitutes for unsupervised word sense induction. In *Thirty-Second AAAI Conference on Artificial Intelligence*.

Dzmitry Bahdanau, Tom Bosc, Stanislaw Jastrzebski, Edward Grefenstette, Pascal Vincent, and Yoshua Bengio. 2017. Learning to compute word embeddings on the fly. *arXiv preprint arXiv:1706.00286*.

Osman Baskaya, Enis Sert, Volkan Cirik, and Deniz Yuret. 2013. Ai-ku: Using substitute vectors and co-occurrence modeling for word sense induction and disambiguation. In *Second Joint Conference on Lexical and Computational Semantics (*SEM), Volume 2: Proceedings of the Seventh International Workshop on Semantic Evaluation (SemEval 2013)*, pages 300–306, Atlanta, Georgia, USA. Association for Computational Linguistics.

Piotr Bojanowski, Edouard Grave, Armand Joulin, and Tomas Mikolov. 2017. Enriching word vectors with subword information. *Transactions of the Association for Computational Linguistics*, 5:135–146.

Elia Bruni, Nam-Khanh Tran, and Marco Baroni. 2014. Multimodal distributional semantics. *Journal of Artificial Intelligence Research*, 49:1–47.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2018. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*.

Aurélie Herbelot and Marco Baroni. 2017. High-risk learning: acquiring new word vectors from tiny data. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pages 304–309. Association for Computational Linguistics.

Mikhail Khodak, Nikunj Saunshi, Yingyu Liang, Tengyu Ma, Brandon Stewart, and Sanjeev Arora. 2018. A la carte embedding: Cheap but effective induction of semantic feature vectors. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 12–22. Association for Computational Linguistics.

Sosuke Kobayashi, Naoaki Okazaki, and Kentaro Inui. 2017. A neural language model for dynamically representing the meanings of unknown words and entities in a discourse. In *Proceedings of the Eighth International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 473–483. Asian Federation of Natural Language Processing.

Angeliki Lazaridou, Marco Marelli, and Marco Baroni. 2017. Multimodal word meaning induction from minimal exposure to natural text. *Cognitive science*, 41:677–705.

Thang Luong, Richard Socher, and Christopher Manning. 2013. Better word representations with recursive neural networks for morphology. In *Proceedings of the Seventeenth Conference on Computational Natural Language Learning*, pages 104–113. Association for Computational Linguistics.

Oren Melamud, Ido Dagan, and Jacob Goldberger. 2015. Modeling word meaning in context with substitute vectors. In *Proceedings of the 2015 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 472–482. Association for Computational Linguistics.

Oren Melamud, Jacob Goldberger, and Ido Dagan. 2016. context2vec: Learning generic context embedding with bidirectional lstm. In *Proceedings of The 20th SIGNLL Conference on Computational Natural Language Learning*, pages 51–61. Association for Computational Linguistics.

Tomas Mikolov, Kai Chen, Greg Corrado, and Jeffrey Dean. 2013. Efficient estimation of word representations in vector space. *CoRR*, abs/1301.3781.

Matthew Peters, Mark Neumann, Mohit Iyyer, Matt Gardner, Christopher Clark, Kenton Lee, and Luke Zettlemoyer. 2018. Deep contextualized word representations. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, pages 2227–2237. Association for Computational Linguistics.

Mohammad Taher Pilehvar and Nigel Collier. 2017. Inducing embeddings for rare and unseen words by leveraging lexical resources. In *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics: Volume 2, Short Papers*, pages 388–393. Association for Computational Linguistics.

Yuval Pinter, Robert Guthrie, and Jacob Eisenstein. 2017. Mimicking word embeddings using subword rnns. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pages 102–112. Association for Computational Linguistics.

Krasen Samardzhiev, Andrew Gargett, and Danushka Bollegala. 2018. Learning neural word salience scores. In *Proceedings of the Seventh Joint Conference on Lexical and Computational Semantics*, pages 33–42. Association for Computational Linguistics.

Timo Schick and Hinrich Schütze. 2019. Learning semantic representations for novel words: Leveraging both form and context. In *Thirty-Third AAAI Conference on Artificial Intelligence*.

Jingyuan Sun, Shaonan Wang, and Chengqing Zong. 2018. Memory, show the way: Memory based few shot word representation learning. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 1435–1444. Association for Computational Linguistics.

Mehmet Ali Yatbaz, Enis Sert, and Deniz Yuret. 2012. Learning syntactic categories using paradigmatic representations of word context. In *Proceedings of the 2012 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning*, pages 940–951. Association for Computational Linguistics.

A Context2vec model parameters for reproducing the experiments in the paper

1. Nonce:

   minimum word freq: 52;
   dimension units 800;
   batchsize: 800;
   learning rate: 0.0001;
   iteration: 12

2. Chimera:

   minimum word freq: 100;
   dimension units 800;

batchsize: 800;
learning rate: 0.0001;
iteration: 14

3. CRW

   minimum word freq: 100;
   dimension units 800;
   batchsize: 600;
   learning rate: 0.0005;
   iteration: 8

B Standard deviations in the CRW experiment in the main paper

number of contexts	a la carte	additive ISF	additive	substitutes	context2vec
1	0.0274	0.0318	0.0357	0.0281	0.0276
2	0.0272	0.0278	0.0314	0.0229	0.0242
4	0.0184	0.0215	0.0218	0.0168	0.0193
8	0.0158	0.0157	0.0193	0.0108	0.0149
16	0.0114	0.0116	0.0123	0.0082	0.0099
32	0.0070	0.0080	0.0099	0.0054	0.0062
64	0.0051	0.0055	0.0062	0.0035	0.0046
128	0.0032	0.0031	0.0038	0.0022	0.0026