Attentive Mimicking: Better Word Embeddings by Attending to Informative Contexts

Timo Schick
Sulzer GmbH
Munich, Germany
timo.schick@sulzer.de

Hinrich Schütze
Center for Information and Language Processing
LMU Munich, Germany
inquiries@cislmu.org

Abstract

Learning high-quality embeddings for rare words is a hard problem because of sparse context information. Mimicking (Pinter et al., 2017) has been proposed as a solution: given embeddings learned by a standard algorithm, a model is first trained to reproduce embeddings of frequent words from their surface form and then used to compute embeddings for rare words. In this paper, we introduce *attentive mimicking*: the mimicking model is given access not only to a word's surface form, but also to all available contexts and learns to attend to the most informative and reliable contexts for computing an embedding. In an evaluation on four tasks, we show that attentive mimicking outperforms previous work for both rare and medium-frequency words. Thus, compared to previous work, attentive mimicking improves embeddings for a much larger part of the vocabulary, including the medium-frequency range.

1 Introduction

Word embeddings have led to large performance gains in natural language processing (NLP). However, embedding methods generally need many observations of a word to learn a good representation for it.

One way to overcome this limitation and improve embeddings of infrequent words is to incorporate surface-form information into learning. This can either be done directly (Wieting et al., 2016; Bojanowski et al., 2017; Salle and Villavicencio, 2018), or a two-step process is employed: first, an embedding model is trained on the word level and then, surface-form information is used either to fine-tune embeddings (Cotterell et al., 2016; Vulić et al., 2017) or to completely recompute them. The latter can be achieved using a model trained to reproduce (or *mimic*) the original embeddings (Pinter et al., 2017). However, these methods only work if a word's meaning can at least partially be predicted from its form.

A closely related line of research is embedding learning for *novel words*, where the goal is to obtain embeddings for previously unseen words from at most a handful of observations. While most contemporary approaches exclusively use context information for this task (e.g. Herbelot and Baroni, 2017; Khodak et al., 2018), Schick and Schütze (2019) recently introduced the *form-context model* and showed that joint learning from both surface form and context leads to better performance.

The problem we address in this paper is that often, only few of a word's contexts provide valuable information about its meaning. Nonetheless, the current state of the art treats all contexts the same. We address this issue by introducing a more intelligent mechanism of incorporating context into mimicking: instead of using all contexts, we learn – by way of self-attention – to pick a subset of especially informative and reliable contexts. This mechanism is based on the observation that in many cases, reliable contexts for a given word tend to resemble each other. We call our proposed architecture *attentive mimicking* (AM).

Our contributions are as follows: (i) We introduce the attentive mimicking model. It produces high-quality embeddings for rare and medium-frequency words by attending to the most informative contexts. (ii) We propose a novel evaluation method based on VecMap (Artetxe et al., 2018) that allows us to easily evaluate the embedding quality of low- and medium-frequency words. (iii) We show that attentive mimicking improves word embeddings on various datasets.

2 Related Work

Methods to train surface-form models to mimic word embeddings include those of Luong et al.

Proceedings of NAACL-HLT 2019, pages 489–494
Minneapolis, Minnesota, June 2 - June 7, 2019.

(2013) (morpheme-based) and Pinter et al. (2017) (character-level). In the area of fine-tuning methods, Cotterell et al. (2016) introduce a Gaussian graphical model that incorporates morphological information into word embeddings. Vulić et al. (2017) retrofit embeddings using a set of language-specific rules. Models that directly incorporate surface-form information into embedding learning include fastText (Bojanowski et al., 2017), LexVec (Salle and Villavicencio, 2018) and Charagram (Wieting et al., 2016).

While many approaches to learning embeddings for novel words exclusively make use of context information (Lazaridou et al., 2017; Herbelot and Baroni, 2017; Khodak et al., 2018), Schick and Schütze (2019)'s form-context model combines surface-form and context information.

Ling et al. (2015) also use attention in embedding learning, but their attention is *within* a context (picking words), not *across* contexts (picking contexts). Also, their attention is based only on word type and distance, not on the more complex factors available in our attentive mimicking model, e.g., the interaction with the word's surface form.

3 Attentive Mimicking

3.1 Form-Context Model

We briefly review the architecture of the form-context model (FCM), see Schick and Schütze (2019) for more details.

FCM requires an embedding space of dimensionality d that assigns high-quality embeddings $v \in \mathbb{R}^d$ to frequent words. Given an infrequent or novel word w and a set of contexts $\mathcal{C}$ in which it occurs, FCM can then be used to infer an embedding $v_{(w,\mathcal{C})}$ for w that is appropriate for the given embedding space. This is achieved by first computing two distinct embeddings, one of which exclusively uses surface-form information and the other context information. The surface-form embedding, denoted $v^{\text{form}}_{(w,\mathcal{C})}$, is obtained from averaging over a set of n-gram embeddings learned by the model; the context embedding $v^{\text{context}}_{(w,\mathcal{C})}$ is obtained from averaging over all embeddings of context words in $\mathcal{C}$.

The two embeddings are then combined using a weighting coefficient α and a $d \times d$ matrix A, resulting in the form-context embedding

$$v_{(w,\mathcal{C})} = \alpha \cdot A v^{\text{context}}_{(w,\mathcal{C})} + (1-\alpha) \cdot v^{\text{form}}_{(w,\mathcal{C})}\,.$$

The weighing coefficient α is a function of both embeddings, modeled as

$$\alpha = \sigma(u^\top [v^{\text{context}}_{(w,\mathcal{C})}; v^{\text{form}}_{(w,\mathcal{C})}] + b)$$

with $u \in \mathbb{R}^{2d}$, $b \in \mathbb{R}$ being learnable parameters and σ denoting the sigmoid function.

3.2 Context Attention

FCM pays equal attention to all contexts of a word but often, only few contexts are actually suitable for inferring the word's meaning. We introduce *attentive mimicking* (AM) to address this problem: we allow our model to assign different weights to contexts based on some measure of their "reliability". To this end, let $\mathcal{C} = \{C_1, \ldots, C_m\}$ where each C_i is a multiset of words. We replace the context-embedding of FCM with a weighted embedding

$$v^{\text{context}}_{(w,\mathcal{C})} = \sum_{i=1}^{m} \rho(C_i, \mathcal{C}) \cdot v_{C_i}$$

where v_{C_i} is the average of the embeddings of words in C_i and ρ measures context reliability.

To obtain a meaningful measure of reliability, our key observation is that reliable contexts typically agree with many other contexts. Consider a word w for which six out of ten contexts contain words referring to sports. Due to this high inter-context agreement, it is then reasonable to assume that w is from the same domain and, consequently, that the four contexts not related to sports are less informative. To formalize this idea, we first define the similarity between two contexts as

$$s(C_1, C_2) = \frac{(M v_{C_1}) \cdot (M v_{C_2})^\top}{\sqrt{d}}$$

with $M \in \mathbb{R}^{d \times d}$ a learnable parameter, inspired by Vaswani et al. (2017)'s scaled dot-product attention. We then define the reliability of a context as

$$\rho(C, \mathcal{C}) = \frac{1}{Z} \sum_{i=1}^{m} s(C, C_i)$$

where $Z = \sum_{i=1}^{m} \sum_{j=1}^{m} s(C_i, C_j)$ is a normalization constant, ensuring that all weights sum to one.

The model is trained by randomly sampling words w and contexts $\mathcal{C}$ from a large corpus and mimicking the original embedding of w, i.e., minimizing the squared distance between the original embedding and $v_{(w,\mathcal{C})}$.

4 Experiments

For our experiments, we follow the setup of Schick and Schütze (2019) and use the Westbury Wikipedia Corpus (WWC) (Shaoul and Westbury, 2010) for training of all embedding models. To obtain training instances $(w, \mathcal{C})$ for both FCM and AM, we sample words and contexts from the WWC based on their frequency, using only words that occur at least 100 times. We always train FCM and AM on skipgram embeddings (Mikolov et al., 2013) obtained using Gensim (Řehůřek and Sojka, 2010).

Our experimental setup differs from that of Schick and Schütze (2019) in two respects: (i) Instead of using a fixed number of contexts for $\mathcal{C}$, we randomly sample between 1 and 64 contexts and (ii) we fix the number of training epochs to 5. The rationale behind our first modification is that we want our model to produce high-quality embeddings both when we only have a few contexts available and when there is a large number of contexts to pick from. We fix the number of epochs simply because our evaluation tasks come without development sets on which it may be optimized.

To evaluate our model, we apply a novel, intrinsic evaluation method that compares embedding spaces by transforming them into a common space (§4.1). We also test our model on three word-level downstream tasks (§4.2, §4.3, §4.4) to demonstrate its versatile applicability.

4.1 VecMap

We introduce a novel evaluation method that explicitly evaluates embeddings for rare and medium-frequency words by downsampling frequent words from the WWC to a fixed number of occurrences.[1] We then compare "gold" skipgram embeddings obtained from the original corpus with embeddings learned by some model trained on the downsampled corpus. To this end, we transform the two embedding spaces into a common space using VecMap (Artetxe et al., 2018), where we provide all but the downsampled words as a mapping dictionary. Intuitively, the better a model is at inferring an embedding from few observations, the more similar its embeddings must be to the gold embeddings in this common space. We thus measure the quality of a model by computing the average cosine similarity between its embeddings and the gold embeddings.

[1]The VecMap dataset is publicly available at https://github.com/timoschick/form-context-model

	number of occurrences							
model	1	2	4	8	16	32	64	128
skipgram	8.7	18.2	30.9	42.3	52.3	59.5	66.7	**71.2**
fastText	**45.4**	44.3	45.7	50.0	55.9	56.7	62.6	67.7
Mimick	10.7	11.7	12.1	11.0	12.5	11.0	10.6	9.2
FCM	37.9	**45.3**	49.1	53.4	**58.3**	55.4	59.9	58.8
AM	38.0	45.1	**49.6**	**53.7**	**58.3**	55.6	60.2	58.9
FCM†	32.3	36.9	41.9	49.1	57.4	59.9	67.3	70.1
AM†	32.8	37.8	42.8	49.8	57.7	**60.5**	**67.6**	70.4

Table 1: Average cosine similarities for the VecMap evaluation, scaled by a factor of 100. †: Downsampled words were included in the training set.

	maximum word frequency				
model	10	50	100	500	1000
skipgram	–0.16	0.21	0.33	0.55	**0.66**
fastText	–0.20	0.10	0.23	0.50	0.61
Mimick	0.00	0.01	–0.03	0.40	0.56
FCM	0.21	0.37	0.37	0.55	0.63
AM	**0.27**	**0.39**	**0.40**	**0.56**	0.64

Table 2: Spearman's ρ for various approaches on SemEval2015 Task 10E

As baselines, we train skipgram and fastText on the downsampled corpus. We then train Mimick (Pinter et al., 2017) as well as both FCM and AM on the skipgram embeddings. We also try a variant where the downsampled words are included in the training set (i.e., the mimicking models explicitly learn to reproduce their skipgram embeddings). This allows the model to learn representations of those words not completely from scratch, but to also make use of their original embeddings. Accordingly, we expect this variant to only be helpful if a word is not too rare, i.e. its original embedding is already of decent quality. Table 1 shows that for words with a frequency below 32, FCM and AM infer much better embeddings than all baselines. The comparably poor performance of Mimick is consistent with the observation of Pinter et al. (2017) that this method captures mostly syntactic information. Given four or more contexts, AM leads to consistent improvements over FCM. The variants that include downsampled words during training (†) still outperform skipgram for 32 and more observations, but perform worse than the default models for less frequent words.

4.2 Sentiment Dictionary

We follow the experimental setup of Rothe et al. (2016) and fuse Opinion lexicon (Hu and Liu,

	$f=1$		$f\in[2,4)$		$f\in[4,8)$		$f\in[8,16)$		$f\in[16,32)$		$f\in[32,64)$		$f\in[1,100]$	
model	acc	F1	acc	F1	acc	F1	acc	F1	acc	F1	acc	F1	acc	F1
skipgram	0.0	2.6	2.2	7.8	11.5	30.7	44.7	64.5	37.8	59.4	**35.0**	**59.7**	33.5	58.3
fastText	44.6	51.1	50.5	65.1	48.4	62.9	44.3	59.6	34.1	53.5	29.8	55.7	31.4	56.4
Mimick	0.0	0.0	0.0	0.0	0.0	0.0	1.0	4.0	1.0	1.0	3.9	14.4	4.2	14.8
FCM	86.5	88.9	76.9	85.1	**72.0**	**81.8**	57.7	68.5	36.0	54.2	27.7	52.5	30.1	53.4
AM	**87.8**	**90.7**	**79.1**	**86.5**	**72.0**	80.9	59.5	**70.9**	37.8	56.1	28.9	53.4	31.1	54.5
AM+skip	**87.8**	**90.7**	**79.1**	**86.5**	**72.0**	81.6	**60.1**	**70.9**	**40.7**	**59.9**	**35.0**	**59.7**	**36.8**	**60.5**

Table 3: Results on the Name Typing dataset for various word frequencies f. The model that uses a linear combination of AM embeddings with skipgram is denoted AM+skip.

2004) and the NRC Emotion lexicons (Mohammad and Turney, 2013) to obtain a training set of words with binary sentiment labels. On that data, we train a logistic regression model to classify words based on their embeddings. For our evaluation, we then use SemEval2015 Task 10E where words are assigned a sentiment rating between 0 (completely negative) and 1 (completely positive) and use Spearman's ρ as a measure of similarity between gold and predicted ratings.

We train logistic regression models on both skipgram and fastText embeddings and, for testing, replace skipgram embeddings by embeddings inferred from the mimicking models. Table 2 shows that for rare and medium-frequency words, AM again outperforms all other models.

4.3 Name Typing

We use Yaghoobzadeh et al. (2018)'s name typing dataset for the task of predicting the fine-grained named entity types of a word, e.g., PRESIDENT and LOCATION for "Washington". We train a logistic regression model using the same setup as in §4.2 and evaluate on all words from the test set that occur $\leq$100 times in WWC. Based on results in §4.1, where AM only improved representations for words occurring fewer than 32 times, we also try the variant *AM+skip* that, in testing, replaces $v_{(w,\mathcal{C})}$ with the linear combination

$$\hat{v}_w = \beta(f_w) \cdot v_{(w,\mathcal{C})} + (1 - \beta(f_w)) \cdot v_w$$

where v_w is the skipgram embedding of w, f_w is the frequency of w and $\beta(f_w)$ scales linearly from 1 for $f_w = 0$ to 0 for $f_w = 32$.

Table 3 gives accuracy and micro F1 for several word frequency ranges. In accordance with results from previous experiments, AM performs drastically better than the baselines for up to 16 occurrences. Notably, the linear combination of skipgram and AM achieves by far the best overall results.

4.4 Chimeras

The Chimeras (CHIMERA) dataset (Lazaridou et al., 2017) consists of similarity scores for pairs of made-up words and regular words. CHIMERA provides only six contexts for each made-up word, so it is not ideal for evaluating our model. Nonetheless, we can still use it to analyze the difference of FCM (no attention) and AM (using attention). As the surface-form of the made-up words was constructed randomly and thus carries no meaning at all, we restrict ourselves to the context parts of FCM and AM (referred to as FCM-ctx and AM-ctx). We use the test set of Herbelot and Baroni (2017) and compare the given similarity scores with the cosine similarities of the corresponding word embeddings, using FCM-ctx and AM-ctx to obtain embeddings for the made-up words. Table 4 gives Spearman's ρ for our model and various baselines; baseline results are adopted from Khodak et al. (2018). We do not report results for Mimick as its representations for novel words are entirely based on their surface form. While AM performs worse than previous methods for 2–4 sentences, it drastically improves over the best result currently published for 6 sentences. Again, context attention consistently improves results: AM-ctx performs better than FCM-ctx, regardless of the number of contexts. Since A La Carte (Khodak et al., 2018), the method performing best for 2–4 contexts, is conceptually similar to FCM, it most likely would similarly benefit from context attention.

While the effect of context attention is more pronounced when there are many contexts available, we still perform a quantitative analysis of one exemplary instance of CHIMERA to better understand what AM learns; we consider the made-up word "petfel", a combination of "saxophone" and "harmonica", whose occurrences are shown in Table 5. The model attends most to sentences

model	2 sent.	4 sent.	6 sent.
skipgram	0.146	0.246	0.250
additive	**0.363**	0.370	0.360
additive – sw	0.338	0.362	0.408
Nonce2Vec	0.332	0.367	0.389
A La Carte	**0.363**	**0.384**	0.394
FCM-ctx	0.337	0.359	0.422
AM-ctx	0.342	0.376	**0.436**

Table 4: Spearman's ρ for the Chimeras task given 2, 4 and 6 context sentences for the made-up word

sentence	ρ
• i doubt if we ll ever hear a man play a **petfel** like that again	0.19
• also there were some other assorted instruments including a **petfel** and some wind chimes	0.31
• they finished with new moon city a song about a suburb of drem which featured beautifully controlled **petfel** playing from callum	0.23
• a programme of jazz and classical music showing the **petfel** as an instrument of both musical genres	0.27

Table 5: Context sentences and corresponding attention weights for the made-up word "petfel"

(2) and (4); consistently, the embeddings obtained from those sentences are very similar. Furthermore, of all four sentences, these two are the ones best suited for a simple averaging model as they contain informative, frequent words like "instrument", "chimes" and "music".

5 Conclusion

We have introduced attentive mimicking (AM) and showed that attending to informative and reliable contexts improves representations of rare and medium-frequency words for a diverse set of evaluations.

In future work, one might investigate whether attention mechanisms on the word level (cf. Ling et al., 2015) can further improve the model's performance. Furthermore, it would be interesting to investigate whether the proposed architecture is also beneficial for languages typologically different from English, e.g., morphologically rich languages.

Acknowledgments

This work was funded by the European Research Council (ERC #740516). We would like to thank the anonymous reviewers for their helpful comments.

References

Mikel Artetxe, Gorka Labaka, and Eneko Agirre. 2018. Generalizing and improving bilingual word embedding mappings with a multi-step framework of linear transformations. In *Proceedings of the Thirty-Second AAAI Conference on Artificial Intelligence*, pages 5012–5019.

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.

Ryan Cotterell, Hinrich Schütze, and Jason Eisner. 2016. Morphological smoothing and extrapolation of word embeddings. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1651–1660. Association for Computational Linguistics.

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.

Minqing Hu and Bing Liu. 2004. Mining and summarizing customer reviews. In *Proceedings of the tenth ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 168–177. ACM.

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.

Diederik Kingma and Jimmy Ba. 2015. Adam: A method for stochastic optimization. *The International Conference on Learning Representations (ICLR)*.

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

Wang Ling, Yulia Tsvetkov, Silvio Amir, Ramon Fermandez, Chris Dyer, Alan W Black, Isabel Trancoso, and Chu-Cheng Lin. 2015. Not all contexts are created equal: Better word representations with variable attention. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*, pages 1367–1372.

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.

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

Saif M Mohammad and Peter D Turney. 2013. Crowdsourcing a word–emotion association lexicon. *Computational Intelligence*, 29(3):436–465.

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.

Radim Řehůřek and Petr Sojka. 2010. Software Framework for Topic Modelling with Large Corpora. In *Proceedings of the LREC 2010 Workshop on New Challenges for NLP Frameworks*, pages 45–50, Valletta, Malta. ELRA.

Sascha Rothe, Sebastian Ebert, and Hinrich Schütze. 2016. Ultradense word embeddings by orthogonal transformation. In *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 767–777. Association for Computational Linguistics.

Alexandre Salle and Aline Villavicencio. 2018. Incorporating subword information into matrix factorization word embeddings. In *Proceedings of the Second Workshop on Subword/Character LEvel Models*, pages 66–71. Association for Computational Linguistics.

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

Cyrus Shaoul and Chris Westbury. 2010. The westbury lab wikipedia corpus.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Lukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. In *Advances in Neural Information Processing Systems 30*, pages 5998–6008. Curran Associates, Inc.

Ivan Vulić, Nikola Mrkšić, Roi Reichart, Diarmuid Ó Séaghdha, Steve Young, and Anna Korhonen. 2017. Morph-fitting: Fine-tuning word vector spaces with simple language-specific rules. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 56–68. Association for Computational Linguistics.

John Wieting, Mohit Bansal, Kevin Gimpel, and Karen Livescu. 2016. Charagram: Embedding words and sentences via character n-grams. *CoRR*, abs/1607.02789.

Yadollah Yaghoobzadeh, Katharina Kann, and Hinrich Schütze. 2018. Evaluating word embeddings in multi-label classification using fine-grained name typing. In *Proceedings of The Third Workshop on Representation Learning for NLP*, pages 101–106. Association for Computational Linguistics.