Towards incremental learning of word embeddings using context informativeness

Alexandre Kabbach
Dept. of Linguistics
University of Geneva
Center for Mind/Brain Sciences
University of Trento

Kristina Gulordava
Dept. of Translation
and Language Sciences
Universitat Pompeu Fabra

Aurélie Herbelot
Center for Mind/Brain Sciences,
Dept. of Information Engineering
and Computer Science
University of Trento

{firstname.lastname}@{unige.ch;upf.edu;unitn.it}

Abstract

In this paper, we investigate the task of learning word embeddings from very sparse data in an incremental, cognitively-plausible way. We focus on the notion of *informativeness*, that is, the idea that some content is more valuable to the learning process than other. We further highlight the challenges of online learning and argue that previous systems fall short of implementing incrementality. Concretely, we incorporate informativeness in a previously proposed model of nonce learning, using it for context selection and learning rate modulation. We test our system on the task of learning new words from definitions, as well as on the task of learning new words from potentially uninformative contexts. We demonstrate that informativeness is crucial to obtaining state-of-the-art performance in a truly incremental setup.

1 Introduction

Distributional semantics models such as word embeddings (Bengio et al., 2003; Collobert et al., 2011; Huang et al., 2012; Mikolov et al., 2013b) notoriously require exposure to a large amount of contextual data in order to generate *high quality* vector representations of words. This poses practical challenges when the available training data is scarce, or when distributional models are intended to mimic humans' word learning abilities by constructing reasonable word representations from limited observations (Lazaridou et al., 2017). In recent work, various approaches have been proposed to tackle these problems, ranging from task-specific auto-encoders generating word embeddings from dictionary definitions only (Bosc and Vincent, 2017, 2018), to Bayesian models used for acquiring definitional properties of words via one-shot learning (Wang et al., 2017), or recursive neural network models making use of morphological structure (Luong et al., 2013).

Arguing that the ideal model should rely on an all-purpose architecture able to learn from *any* amount of data, Herbelot and Baroni (2017) proposed a model called Nonce2Vec (N2V), designed as a modification of Word2Vec (W2V; Mikolov et al., 2013b), refactored to allow incremental learning. The model was tested on two datasets: a) the newly introduced *definitional* dataset, where the task is to learn a nonce word from its Wikipedia definition; and b) the *chimera* dataset of Lazaridou et al. (2017), where the task is to reproduce human similarity judgements related to a novel word observed in 2-6 randomly extracted sentences. The N2V model performed much better than W2V on both datasets but failed to outperform a basic additive model on the chimera dataset, leading the authors to hypothesise that their system would need to perform content selection to deal with the potentially uninformative chimera sentences.

There are two motivations to the present work. The first is to provide a formal definition of the notion of *informativeness* applied to both sentential context (as a whole) and context words (taken individually). To do so, we rely on the intuition that an *informative* context is a context that is more *specific* to a given target, and that this notion of *context specificity* can be quantified by computing the entropy of the probability distribution generated by a language model over a set of vocabulary words, given the context.

The secondary motivation of this work lays in considerations over incrementality. We show that N2V itself did not fully implement its ideal of 'online' concept learning. We also point out that architectures that have outperformed N2V since its inception actually move even further from this ideal. In contrast, we attempt to make our architecture as close as possible to a realistic belief update system, and we demonstrate that informativeness

Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics: Student Research Workshop, pages 162–168
Florence, Italy, July 28 - August 2, 2019. ©2019 Association for Computational Linguistics

is an essential part of retaining acceptable performance in such a challenging setting.

2 Related work

The original Nonce2Vec (N2V) model is designed to simulate new word acquisition by an adult speaker who already masters a substantial vocabulary. The system uses some 'background' lexical knowledge in the shape of a distributional space acquired over a large text corpus. A novel word is then learnt by using information present in its context sentence. To achieve its goal, N2V proposes some modifications to the original Word2Vec architecture to make it suitable for novel word learning. Three main changes are suggested: a) the learning rate should be greatly heightened to allow for learning via backpropagation using only the limited amount of data; b) random subsampling should be reduced to a minimum so that all available data is used; and c) embeddings in the background should be 'frozen' so that the high learning rate is prevented from 'unlearning' old lexical knowledge in favour of the scarce, and potentially uninformative, new context information.

Recent work outperforms N2V on the chimera (Khodak et al., 2018; Schick and Schütze, 2019a) and the definitional (Schick and Schütze, 2018) datasets (see §1 for a description of the datasets). However, they both deviate from the original motivation of N2V which is to learn word representations *incrementally* from *any* amount of data. Instead, those state-of-the-art models rely—at least in part—on a learned linear regression matrix which is fixed for a given corpus and thus does not lend itself well to incremental learning.

Our work follows the philosophy of N2V but pushes the notion of incrementality further by identifying aspects of the original system that in fact do not play well with true online learning. For a start, the original model is not fully incremental as it adopts a *one-shot* evaluation setup where each test instance is considered individually and where the background model is reloaded from scratch at each test iteration. This does not test how the system would react to learning *multiple* nonces one after the other (as humans do in the course of their lives). Related to this, whilst 'freezing' background vectors makes sense as a safety net when using very high learning rates, it similarly goes against the notion of incrementality. In any realistic setup, indeed, we would like newly learnt words to inform our background lexical knowledge and become part of that background themselves, being refined over time and contributing to acquiring the next nonce. Following this philosophy, the system we present in this paper does away with freezing previously learnt embeddings and does not reload the background model at each test iteration.

We should further note that our work on informativeness echoes recent research on the use of *attention* mechanisms. Vaswani et al. (2017), followed by Devlin et al. (2018) and Schick and Schütze (2019b), have shown that such mechanisms can provide very powerful tools to build sentence and contextualised word embeddings which are amenable to transfer learning tasks. However, we note that from our point of view, these systems suffer from the same problem as the previously mentioned architectures: the underlying model consists of a large set of parameters which can be used to learn a task-specific regression. It is not designed to be updated with each new encountered experience.

3 Model

Let us consider *context* to be defined as a window of $\pm n$ words around a given *target*. We define two specific functions: *context informativeness* (CI) which characterises how informative an entire context is with respect to its corresponding target; and *context word informativeness* (CWI) which characterises how informative a *particular context item* is with respect to the target. For instance, if target *chases* is seen in context $c = \{$*the, cat, the, mouse*$\}$, the context informativeness is the informativeness of c, and the context word informativeness can be computed for each element in c, with the expectation, in this case, that *the* might be less informative than *cat* or *mouse*. The CWI measure is dependent on CI, as we proceed to show.

3.1 Context informativeness

Let us consider a sequence c of n context items $c = c_1 \dots c_n$. We define the *context informativeness* of a context sequence c as:

$$CI(c) = 1 + \frac{1}{\ln(|\mathcal{V}|)} \sum_{w \in \mathcal{V}} p(w|c) \ln p(w|c) \quad (1)$$

CI is a slight modification of the Shannon entropy $H = -\sum_{w \in \mathcal{V}} p(w|c) \ln p(w|c)$, normalised over the cardinality of the vocabulary $|\mathcal{V}|$ to output values in $[0, 1]$. In this work, we use a CBOW model (Mikolov et al., 2013a) to obtain the probability distribution $p(w|c)$. We use CBOW because it is the simplest word-generation model which takes the relation between context words into account, i.e., in contrast to skipgram.

A context will be considered maximally informative ($CI = 1$) if only a single vocabulary item is predicted to occur in context c with a non-null probability. Conversely, a context will be considered minimally informative ($CI = 0$) if all vocabulary items are predicted to occur in context c with equal probability. CI should therefore quantify how *specific* a given context is regarding a given target.

3.2 Context word informativeness

Let us consider $c_{\neq i} = c_1 \ldots c_{i-1}, c_{i+1} \ldots c_n$ to be the sequence of context items taken from c to which the i^{th} item has been removed. We define the *context word informativeness* of a context item c_i in a context sequence c as:

$$CWI(c_i) = CI(c) - CI(c_{\neq i}) \tag{2}$$

CWI outputs values in $[-1, 1]$: a context word c_i will be considered maximally informative ($CWI = 1$) if removing it from a maximally informative context leads to a minimally informative one. Conversely, a context word c_i will be considered minimally informative ($CWI = -1$) if removing it from a minimally informative context leads to a maximally informative one.

3.3 CWI-augmented Nonce2Vec

As explained in §2, N2V introduces several high-level changes to the W2V architecture to achieve learning from very sparse data. In practice, this translates into the following design choices: a) nonces are initialised by summing context word embeddings (after subsampling); and b) nonces are trained with an adapted skipgram function incorporating decaying window size, sampling and learning rates at each iteration, while all other vectors remain frozen. The learning rate is computed via $\alpha = \alpha_0 e^{-\lambda t}$ with a high α_0.

The modifications we propose are as follows: i) we incorporate informativeness into the initialisation phase by summing over the set of context words with positive CWI only; ii) we train on the entire context without subsampling and window decay; and iii) we remove freezing and compute the learning rate as a function of CWI for each context item c_i via:

$$\alpha(c_i) = \alpha_{max} \frac{e^{\tanh(\beta * CWI(c_i)) + 1} - 1}{e^2 - 1} \tag{3}$$

The purpose of equation 3 is to modulate the learning rate depending on the context word informativeness for a context–target pair: α should be maximal ($\alpha = \alpha_{max}$, where α_{max} is a hyperparameter) when context is maximally informative ($CWI = 1$) and minimal ($\alpha = 0$) when context is minimally informative ($CWI = -1$). The function $x \mapsto \frac{e^{x+1}-1}{e^2-1}$ is therefore designed as a logistic “S-shape” function with domain $[-1, 1] \rightarrow [0, 1]$. In practice, CWI values are highly dependant on the language model used and may end up all being close to 0 (± 0.01 with our CBOW model for instance). The $tanh$ function and the β parameter are therefore added to compensate for this effect that would otherwise produce identical learning rates for all target-context pairs, regardless of CWI values.

4 Experimental setup and evaluation

To test the robustness of the results of Herbelot and Baroni (2017), we retrain a skipgram background model with the same hyperparameters but from the more recent Wikipedia snapshot of January 2019, and obtain a similar correlation ratio on the MEN similarity dataset (Bruni et al., 2014): $\rho = 0.74$ vs $\rho = 0.75$ for Herbelot and Baroni (2017). Probability distributions used for computing CI and CWI are generated with a CBOW model trained with gensim (Řehůřek and Sojka, 2010) on the same Wikipedia snapshot as our skipgram background model, and with the same hyperparameters. For the CWI-based learning rate computation, we set $\alpha_{max} = 1$, chosen according to α_0 in the original N2V for fair comparison; and $\beta = 1000$, chosen given min and max CWI values output by CBOW to produce $tanh(\beta * x)$ values distributed across $[-1, 1]$ and apply a learning rate $\alpha_{max} = 1$ to maximally informative context words.

We report results on the *definitional* and the *chimera* datasets (see §1). The definitional dataset contains first sentences from Wikipedia for 1000 words: e.g. *Insulin is a peptide hormone produced by beta cells of the pancreatic islets*, where the task is to learn the nonce *insulin*. Evaluation is performed on 300 test instances in terms of Median Rank (MR) and Mean Reciprocal Rank (MRR). That is, for each instance, the Reciprocal Rank of the *gold* vector (the one that would be obtained by training standard W2V over the entire corpus) is computed over the sorted list of neighbours of the *predicted* representation.

The chimera dataset simulates a nonce situation where speaker encounters words for the first time in naturally-occurring (and not necessarily informative) sentences. Each *nonce* instance in the data is associated with 2 (L2), 4 (L4) or 6 (L6) sentences showing the nonce in context, and a set of six word *probes* human-annotated for similarity to the nonce. For instance, the nonce *VALTUOR* is shown in *Canned sardines and VALTUOR between two slices of wholemeal bread and thinly spread Flora Original [...]*, and its similarity assessed with respect to *rhubarb, onion, pear, strawberry, limousine* and *cushion*. Evaluation is performed on 110 test instances by computing the Spearman correlation between the similarities output by the system for each nonce-probe pair and the similarities from the human subjects.

We evaluate both datasets using a *one-shot* setup, as per the original N2V paper: each nonce word in considered individually and the background model is reloaded at each test iteration. We further propose an *incremental* evaluation setup where the background model is loaded only once at the beginning of testing, keeping its word vectors modifiable during subsequent learning, and where each newly learned nonce representation is added to the background model. As performance in the incremental setup proved to be dependent on the order of the test items, we report average and standard deviation scores computed from 10 test runs where the test set is shuffled each time.

5 Results

5.1 Improving additive models

Herbelot and Baroni (2017) show that a simple additive model provides an extremely strong baseline for nonce learning. So we first measure the contribution of our notion of informativeness to the context filtering module of a sum model. Comparison takes place across four settings: a) *no filter*, where all words are retained; b) *random*, which applies standard subsampling with a sample rate of 10,000, following the original N2V approach; c) *self*, where all items found in training with a frequency above a given threshold are discarded;[1] and d) *CWI*, which only retains context items with a positive CWI value.

Our results on the definitional dataset, displayed in Table 1, show a consistent hierarchy of filters with the SUM CWI model outperforming all other SUM models, in both one-shot and incremental evaluation setups. Results on the chimera dataset, displayed in Table 2, are not as clear-cut, although they do exhibit a similar trend on both L4 and L6 test sets, with the notable result of achieving state-of-the-art performances with our SUM CWI model on the L4 and L6 test sets in incremental setup, and near state-of-the-art performance on the L6 test set in one-shot setup. This confirms once again that additive models can provide very robust baselines.

Qualitatively, the contribution of each filter on the definitional dataset can be exemplified on the following sentence, with nonce word *Honeywell*: *"Honeywell International Inc is an American multinational conglomerate company that produces [...] aerospace systems for a wide variety of customers from private consumers to major corporations and governments."*. The *no-filter* additive model outputs a rank of 383 (the gold vector for *honeywell* is found to be the 383th closest neighbour of the predicted vector). The *random* model randomly removes most (but not all) high frequency words before summing, outputting a rank of 192 (filtered-out words include also content words like *international* or *company*). The *self*-information model reduces the size of the context words set even further by removing all high-frequency words left over by the random process (rank 170). Finally, the *CWI* model outputs the best rank at 85, removing all function words while keeping some useful high-frequency words such as *international* or *company*.

5.2 Improving neural models in one-shot settings

Our results for neural models are also displayed in Table 1 and Table 2: *as-is* refers to the original N2V system; *CWI init* is N2V as-is with CWI-based context filtering instead of subsampling; and *CWI alpha* is a model with a CWI-based

[1] We take the log of the $sample\_int$ values computed by gensim for each word during training, keeping only items with log values above 22, which gave us the best performances overall.

	one-shot		incremental	
Model	MR	MRR	MR	MRR
SOTA	**49**	**.1754**	–	–
SUM no-filter	5,969	.0087	6,461± 225	.0014±.0002
SUM random	3,047	.0221	3,113± 179	.0071±.0012
SUM self	1,769	.0242	2,095±125	.0121±.0008
SUM CWI	935	.0374	**961±24**	**.0322±.0011**
N2V as-is	955	.0477	81,705±14,076	.0096±.0038
N2V CWI init	540	.0493	70,992±17,312	.0079±.0025
N2V CWI alpha	763	.0404	**983±175**	**.0341±.0021**

Table 1: Performance of various additive (SUM) and neural (N2V) models on the definitional dataset, measured in terms of Median Rank (MR) and Mean Reciprocal Rank (MRR). SOTA in *one-shot* evaluation setup is reported by the *Form-Context* model of Schick and Schütze (2018).

sum initialisation (as in SUM CWI), and a CWI-based learning rate computed on unfiltered context words, as detailed in §3.3.

When informativeness is incorporated to N2V in the original one-shot evaluation setup, we also observe near-systematic improvements. On the definitional dataset in Table 1, CWI init improves over the standard N2V as-is model (MR 540 vs 955; MRR .0493 vs .0477) or over the SUM CWI baseline (MR 540 vs 935; MRR .0493 vs .0374). In comparison to CWI init, our CWI alpha model provides robust performances across evaluation setups and datasets, often reaching similar if not better results than our best baseline model (SUM CWI) showing that a neural model fully based on informativeness is a more robust alternative than its counterparts. See for example Table 1 on the definitional dataset where the N2V CWI alpha model performs better than the SUM CWI model in one-shot setup (MR 763 vs. 935; MRR .0404 vs .0374) or Table 2 on the chimera dataset where it also performs better than the SUM CWI model on both the L2 (ρ .3129 vs .3074) and the L4 (ρ .3928 vs .3739) test sets and achieves state-of-the-art performance on the L4 test set.

5.3 Improving incremental learning

As stated in §2, recent approaches to nonce learning have deviated from the original philosophy of N2V and in fact, N2V itself did not fully implement an incremental setting. We now show that the original N2V performance decreases significantly on both datasets in an incremental evaluation setup, without freezing of background vectors. Compare the results of the N2V as-is model in both *one-shot* and *incremental* evaluation setups on the definitional dataset in Table 1: MR 955 vs 81,705±14,076 and MRR .0477 vs .0096±.0038; and on the chimera dataset in Table 2: ρ .3412 vs .1650±.0384 on L2; ρ .3514 vs .1144±.0620 on L4 and ρ .4077 vs .1391±.0694 on L6. We find this drastic decrease in performance to be related to two distinct phenomena: 1) a *sum* effect which leads vector representations for nonces to be close to each other due to the sum initialisation creating very similar vectors in a 'special' portion of the vector space; and 2) a *snowball* effect related to the 'unfreezing' of the background space which leads background vectors to be updated by backpropagation at a very high learning rate at every test iteration, moving their original meaning towards the semantics of the new context they are encountered in. This includes vectors for very frequent words, which are encountered again in their now shifted version when a new nonce is presented to the system. This snowball effect ends up significantly altering the quality of the background model and its generated representations.

The *sum effect* is best illustrated by the decrease in performance of SUM models between *one-shot* and *incremental* setups on the definitional dataset in Table 1, as this effect has the property of specifically changing the nature and order of the nearest neighbours of the predicted nonce vectors, which is directly reflected in the MR and MRR evaluation metrics on the definitional dataset given the evaluation task. On the chimera dataset in Table 2 however, this effect does not appear to neg-

	one-shot			incremental		
Model	L2	L4	L6	L2	L4	L6
SOTA	**.3634**	.3844	**.4360**	–	–	–
SUM no-filter	.3047	.3288	.3063	.3047±.0000	.3288±.0000	.3063±.0000
SUM random	.3358	.3717	.3584	.3358±.0002	.3717±.0004	.3584±.0003
SUM self	.3455	.3638	.3651	**.3455±.0000**	.3638±.0000	.3651±.0000
SUM CWI	.3074	.3739	.4243	.3074±.0000	**.3739±.0000**	**.4243±.0000**
N2V as-is	.3412	.3514	.4077	.1650±.0384	.1144±.0620	.1391±.0694
N2V CWI init	.3002	.3482	.4218	.1451±.0265	.1522±.0396	.1225±.0544
N2V CWI alpha	.3129	**.3928**	.4181	.2970±.0262	.3000±.0268	.2678±.0408

Table 2: Performance of various additive (SUM) and neural (N2V) models on the chimera dataset, measured in terms of Spearman correlation. SOTA in *one-shot* evaluation setup on the L2 and L4 test sets are reported by the *A la carte* model of Khodak et al. (2018), while SOTA on the L6 test set is reported by the *attentive mimicking* model of Schick and Schütze (2019a).

atively impact performance given that evaluation compares correlations between gold and predicted similarity rankings of nonces with a prefixed set of probes. The *snowball* effect however is visible on both datasets in Table 1 and Table 2 when comparing performances of N2V models between *one-shot* and *incremental* setups. It proves particularly salient for neural models which do not make use of informativeness-based adaptative learning rate (all N2V models but N2V CWI alpha).

Our notion of informativeness proves even more useful in the context of incremental nonce learning: on the definitional dataset in Table 1, our informativeness-based models, be it SUM CWI or N2V CWI alpha, achieve best (and comparable) performances (MR 961±24 vs 983±175; MRR .0322±.0011 vs .0341±.0021). Moreover, we observe that those models are able to mitigate the undesirable effects mentioned above, almost totally for the sum effect (compare the performance of the SUM CWI model in Table 1 between the incremental and one-shot setups, versus the other SUM models), and partially for the snowballing interference of the high learning rate (compare the performance of the N2V CWI alpha model in Table 1 between the incremental and one-shot setups, versus the other N2V models). Performances of our SUM CWI and N2V CWI alpha models in incremental setup approach those of the one-shot setting. On the chimera dataset in Table 2, which proves only sensitive to the snowball effect, we also observe that our N2V CWI alpha model is able to mitigate this effect, although performances of the model in incremental setup remain below those of the one-shot setup, as well as those of the additive models in incremental setup.

6 Conclusion

We have proposed an improvement of the original N2V model which incorporates a notion of informativeness in the nonce learning process. We showed that our informativeness function was very beneficial to a vector addition baseline, and could be usefully integrated into the original N2V approach, both at initialisation stage and during learning, achieving state-of-the-art results on the chimera dataset and on the definitional dataset in an incremental evaluation setup. Although our proposed notion of informativeness proved to be mostly beneficial to incremental learning, nothing prevents it from being incorporated to other non-incremental models of nonce learning, provided that those models make use of contextual information. On top of the performance improvements observed, our proposed definition of informativeness benefits from being intuitive, debuggable at each step of the learning process, and of relying on no external resource. We make our code freely available at https://github.com/minimalparts/nonce2vec.

Acknowledgements

We are grateful to our two anonymous reviewers for their helpful comments.

References

Yoshua Bengio, Réjean Ducharme, Pascal Vincent, and Christian Janvin. 2003. A neural probabilistic language model. *Journal of Machine Learning Research*, 3:1137–1155.

Tom Bosc and Pascal Vincent. 2017. Learning word embeddings from dictionary definitions only. In *Proceedings of the NIPS 2017 Workshop on Meta-Learning*.

Tom Bosc and Pascal Vincent. 2018. Auto-encoding dictionary definitions into consistent word embeddings. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 1522–1532, Brussels, Belgium. Association for Computational Linguistics.

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

Ronan Collobert, Jason Weston, Léon Bottou, Michael Karlen, Koray Kavukcuoglu, and Pavel Kuksa. 2011. Natural language processing (almost) from scratch. *Journal of Machine Learning Research*, 12:2493–2537.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2018. BERT: pre-training of deep bidirectional transformers for language understanding. *CoRR*, abs/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, Copenhagen, Denmark. Association for Computational Linguistics.

Eric Huang, Richard Socher, Christopher Manning, and Andrew Ng. 2012. Improving word representations via global context and multiple word prototypes. In *Proceedings of the 50th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 873–882, Jeju Island, Korea. 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, Melbourne, Australia. Association for Computational Linguistics.

Angeliki Lazaridou, Marco Marelli, and Marco Baroni. 2017. Multimodal word meaning induction from minimal exposure to natural text. *Cognitive Science*, 41(S4):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, Sofia, Bulgaria. Association for Computational Linguistics.

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

Tomas Mikolov, Ilya Sutskever, Kai Chen, Greg Corrado, and Jeffrey Dean. 2013b. Distributed representations of words and phrases and their compositionality. In *Proceedings of the 26th International Conference on Neural Information Processing Systems - Volume 2*, NIPS'13, pages 3111–3119, USA. Curran Associates Inc.

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.

Timo Schick and Hinrich Schütze. 2018. Learning semantic representations for novel words: Leveraging both form and context. *CoRR*, abs/1811.03866.

Timo Schick and Hinrich Schütze. 2019a. Attentive mimicking: Better word embeddings by attending to informative contexts. *CoRR*, abs/1904.01617.

Timo Schick and Hinrich Schütze. 2019b. Rare words: A major problem for contextualized embeddings and how to fix it by attentive mimicking. *CoRR*, abs/1904.06707.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. In *Advances in neural information processing systems*, pages 5998–6008.

Su Wang, Stephen Roller, and Katrin Erk. 2017. Distributional Modeling on a Diet: One-shot Word Learning from Text Only. In *Proceedings of the Eighth International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 204–213. Asian Federation of Natural Language Processing.