Direction-Aware Skip-Gram: Improving Word Embedding

Changho Yoon
Department of Computer Science Engineering
UNIST
ho9938@unist.ac.kr

Minsik Lee
Department of Computer Science Engineering
UNIST
lms0522@unist.ac.kr

Abstract

In this project, we venture into the realm of word embedding techniques with a specific goal - to alleviate the directionality limitations of the traditional Skip-Gram model while enhancing its efficiency and accuracy. To achieve this, we propose the Direction-Aware Skip-Gram model, an innovative technique that incorporates the forward and backward positions of surrounding words relative to the central word during the embedding process. Our findings show that the Direction-Aware Skip-Gram model outperforms the traditional model in learning word similarity according to test results, even with a smaller embedding size. Furthermore, the behavior of these models in response to changes in the embedding size provides a new understanding of their dynamics. This study presents a significant advancement in word embedding technologies with the potential to greatly contribute to the field of natural language processing.

1 Introduction

In Efficient Estimation of Word Representations in Vector Space(Mikolov et al., 2013)[1], the authors introduce the word embedding technique called Word2Vec, which efficiently captures semantic similarities between words and represents them in vector space. The main goal of this paper is to propose a method that uses less time and memory while effectively learning high-quality word vectors that capture relationships between words.

The Word2Vec model is widely utilized in various natural language processing tasks, such as machine translation, sentiment analysis, and named entity recognition, due to its effectiveness in measuring semantic similarity between words. As demonstrated by the experimental results in the paper, Word2Vec outperforms existing word embedding methods in terms of speed and accuracy, resulting in exceptional performance. Moreover, Word2Vec enables the use of semantic relationships between words based on vector similarities, exemplified by analogies like 'king $-$ man $+$ woman $\approx$ queen'. This characteristic has further applications in transfer learning for a diverse range of language tasks.

However, there is a need to explore new approaches that can overcome the limitations related to the directionality inherent in the traditional Skip-Gram model. Therefore, we aim to investigate an embedding method that considers the forward and backward positions of surrounding words relative to the central word, going beyond the existing Skip-Gram model. By doing so, we seek to address the issues of directionality that may arise in the traditional Skip-Gram and strive to develop an improved word embedding technique. Through this research, we hope to present novel word embedding technologies that can enhance both efficiency and accuracy.

Preprint. Under review.

2 Related Work

Since the inception of Word2Vec, numerous advancements and extensions have been introduced to enhance its applicability and efficiency in capturing complex word relationships. These improvements have greatly benefited the field of natural language processing. Some notable extensions include Distributed Representations of Words and Phrases and their Compositionality (Mikolov et al., 2013)[2], which refines the original Word2Vec model and introduces Negative Sampling; FastText (Bojanowski et al., 2017)[3], which captures morphological information using character n-grams and accelerates training time; and ELMo (Peters et al., 2018)[4], a novel approach that generates deep contextualized word representations using bidirectional language models, leading to substantial performance gains in various NLP tasks.

There were some further researches to overcome the limitations related to the directionality inherent in the traditional Skip-Gram model. Structured Skip-Gram (SSG) Model (Ling et al., 2015)[5] tried to achieve it by training different matrices for each position of the context words. However, it causes the explosion of the number of parameters and operations and brings serious increase of training time. Directional Skip-Gram (DSG) Model (Song et al., 2018)[6] added new term concerning the direction of a context word to original Skip-Gram's probability distribution. But there is an ambiguity about reflection of the directionality depending on the hyperparameters as it trains inclusion term and direction term separately.

3 Approach

3.1 Original Skip-gram Model

The original Skip-gram model has been proposed in [1]. It tries to maximize classification of a word based on another word in the same sentence.

It consists of input, projection, and output layers. It uses each current word as an input to a log-linear classifier with continuous projection layer, and predict words within a certain range before and after the current word.

At the input layer, center word is one-hot encoded. The dimension of the input vector is $|V|$, where V is set of the vocabulary.

$$x \in \mathbb{R}^{|V|}$$

The input layer is then projected to a projection layer P that has dimensionality N, using a projection matrix W. The projection layer is shared for all words (not just the projection matrix); thus, all output words are assumed to get projected into the same position.

$$v_c = Wx \in \mathbb{R}^n$$

Finally, output layer is derived by using an output word matrix W'. It passes through the softmax function and eventually becomes a probability vector. Because output vectors of each position within the context window share same output word matrix, they have same values regardless of their position.

$$z = W'v_c \in \mathbb{R}^{|V|}$$
$$\hat{y} = softmax(z) \in \mathbb{R}^{|V|}$$

In training process, it uses cross-entropy as a loss function. Because output vectors are not position-specific, it cannot implement position-specific training.

$$\mathcal{L}(\hat{y}, y) = -\sum_{j=0, j\neq m}^{2m} \sum_{k=1}^{|V|} y_k^{(c-m+j)} \log \hat{y}_k$$
$$\hat{\theta} = \arg\min_{\theta} \mathcal{L}(\hat{y}, y)$$

3.2 Direction-Aware Skip-Gram Model

Our proposed architecture is similar to the Skip-gram model. It tries to maximize classification of a word based on another word in the same sentence. But instead of calculating context words regardless of the position, it predicts words considering their relative position to the center word.

In Direction-Aware Skip-Gram Model, output layer is derived by using two different output word matrices. W_prev, W_next is output word matrix for context words followed by center word, following center word, respectively. It passes through the softmax function and eventually becomes a probability vector. Because output vectors of each position within the context window uses two different output word matrices, their values vary depending on their position.

$$z_{prev} = W_{prev} v_c \in \mathbb{R}^{|V|}$$
$$z_{next} = W_{next} v_c \in \mathbb{R}^{|V|}$$
$$\hat{y}_{prev} = softmax(z_{prev}) \in \mathbb{R}^{|V|}$$
$$\hat{y}_{next} = softmax(z_{next}) \in \mathbb{R}^{|V|}$$

In training process, it uses cross-entropy as a loss function. Because output vectors are position-specific, it can implement position-specific training.

$$\mathcal{L}(\hat{y}, y) = -\sum_{j=1}^{m}\sum_{k=1}^{|V|}(y_k^{(c-j)} \log \hat{y}_{prev,k} + y_k^{(c+j)} \log \hat{y}_{next,k})$$

$$\hat{\theta} = \arg\min_{\theta} \mathcal{L}(\hat{y}, y)$$

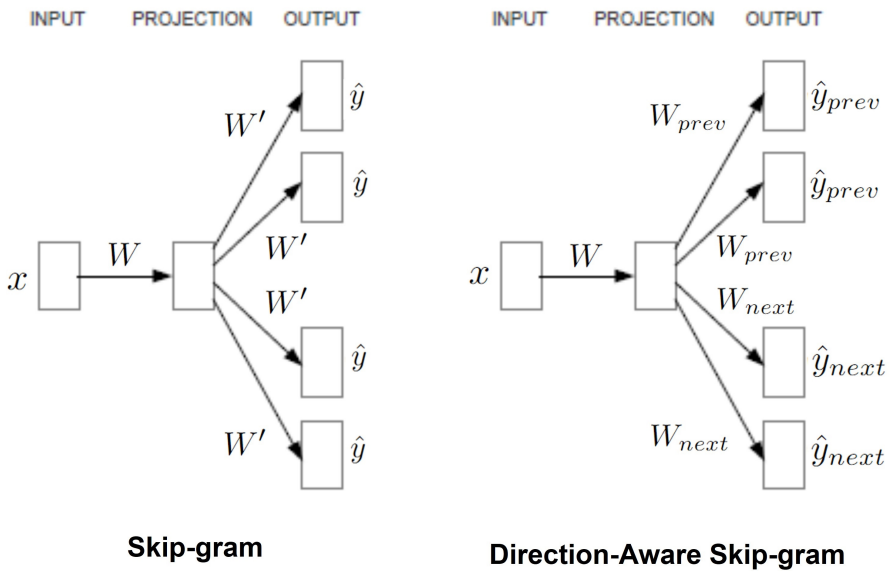


Figure 1: Original Skip-Gram and new Direction-Aware Skip-Gram architecture.

4 Experiments

4.1 Data

The embeddings were trained on the dump of English Wikipedia articles. We used 0.1-percent version of the dataset, because of the time and resource limitation of the Google Colab account. It contains approximately 2 million word tokens.

4.2 Evaluation method

We used one quantitative, two qualitative evaluation methods.

As a quantitative evaluation, we compared the word similarities the model produce to human-annotated word similarities. We used Spearman's rank correlation coefficient as the evaluation metric. We used MEN and SimLex-999 dataset as the word similarity dataset.

As a qualitative evaluation, we drew t-SNE graphs of 500 most frequent word embeddings. We evaluated the model by observing how clusters are formed and how are the words in the cluster semantically close to each other. We also calculated 10 closest words for some common query words in cosine similarity. We evaluated the model by observing how are the calculated words semantically close to the query word.

4.3 Experimental details

We first made a vocabulary by extracting 50000 most frequent words from the dataset. Words that are not contained in the vocabulary were treated as unknown token while training. With this vocabulary, we made list of skip-grams in sliding-window fashion. Each skip-gram has the index of the center word of the window as a key, and the indices of the context words of the window as a value.

When training, we iterated batch training until it uses the whole dataset for each epoch. For each iteration, We made input batch which is size of 128 by extracting separated skip-grams from the dataset. We intentionally used separated skip-grams for one batch for the efficient training. We used the Adam optimizer to complement our time and resource limitation.

Additionally, when training Direction-Aware Skip-Gram model, we assigned more epochs as default. It is because of the time limitation. If we had enough time and resource, we could give the enough epoch to the models and comprehend the dataset to the fixed point. However, we cannot conduct such experiment. So we decided to give more epochs to the Direction-Aware Skip-Gram model, because it has more parameters to train compared to the original Skip-Gram model. And we used this result for a analysis in Section 5.

5 Result and Analysis

5.1 Default Settings

Model	Epoch	Window-size	Embedding-size	MEN	SimLex-999
Skip-Gram	10	2	64	0.2386	0.0688
Direction-Aware Skip-Gram	15	2	64	0.2404	0.0966

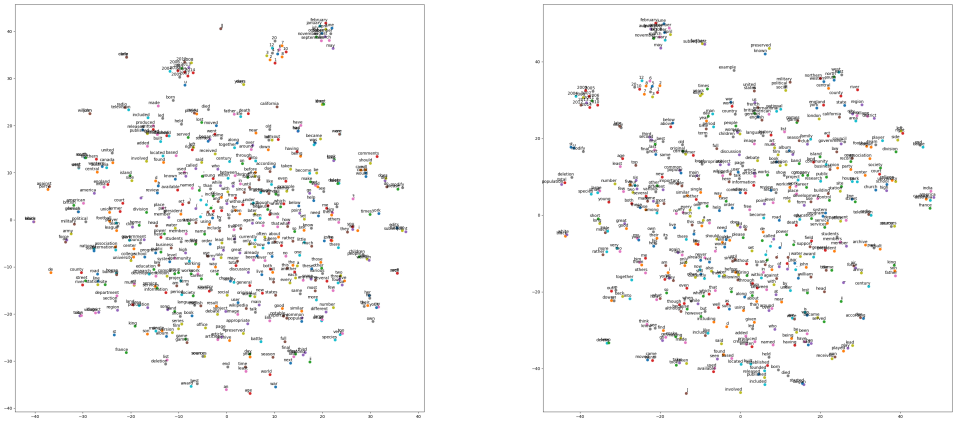
For our default configurations, we set the Skip-Gram model with epoch = 10, window-size = 2, and embedding-size = 64. In contrast, the Direction-Aware Skip-Gram was set to epoch = 15, window-size = 2, and embedding-size = 64. The difference in the number of epochs between the two models is discussed in detail in the experiments section. The MEN test evaluates the performance of 'relevance' learning, while the SimLex-999 test assesses the learning of 'similarity'.

Upon evaluating each model after training, we found similar scores in the MEN test. However, the Direction-Aware Skip-Gram model displayed marginally superior performance in the SimLex-999 test. This evidence suggests that while the default models exhibit similar 'relevance' performance, the Direction-Aware Skip-Gram model performs slightly better in terms of 'similarity'.

We can see similar results in the t-SNE graphs. It visualizes high-dimensional data into the 2-dimentional space. About the months and years, both graph show the similar clusters. But about auxiliary verbs and pronouns, Direction-Aware Skip gram shows more dense and segregated clusters comparing to the original Skip-Gram. It shows that Direction-Aware Skip-Gram model is slightly better to catch the similarity among the words.

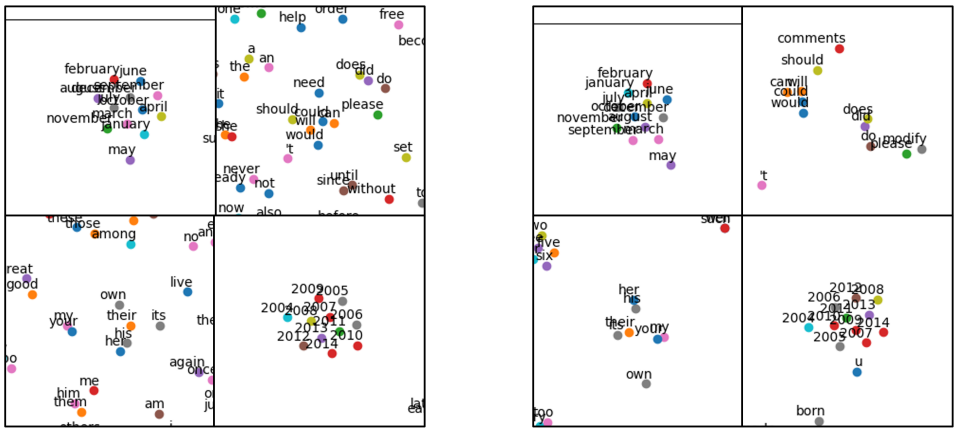
Closest word result shows similar results among original Skip-Gram and Direction-Aware Skip-Gram model. They found closest words of verbs and adverbs relatively easily, while they couldn't find the nouns which have specific meaning. It may because verbs and adverbs have variety of synonyms which have very close meaning with a query word, while nouns don't have many of them.

Overall, we can conclude that Direction-Aware Skip-Gram shows slightly better performance over the original Skip-Gram model. We analyze that the main reason of this difference comes from the awareness toward the directionality. However, number of epochs could be additional factor, because we gave different epochs to the models, by the reason mentioned in section 4.



(a) Original Skip-Gram (b) Direction-Aware Skip-Gram

Figure 2: t-SNE graphs of Original Skip-Gram and Direction-Aware Skip-Gram



(a) Original Skip-Gram (b) Direction-Aware Skip-Gram

Figure 3: major clusters found in the t-SNE graphs

5.2 Varying Epoch

In our next experiment, we kept all factors constant and varied the number of epochs. This approach revealed that both the Skip-Gram and Direction-Aware Skip-Gram models show enhanced performance as the number of epochs increase. The Direction-Aware Skip-Gram model, in particular, demonstrates superior performance with fewer epochs and continues to slightly outperform with more epochs.

Increasing number of epochs has especially strong effects in our case, because our models are trained in relatively small epochs, so they have more parameters to learn. We analyze that large difference in

Query Word	Closest Words
can	could will cannot would must might files couldn excludes
good	bad scholarly mahdi reasonable nice interesting recurring biscay notable
number	handful fraction eight percentage lot amount roller scale height
january	july december april october august february november june march
word	meaning means capitalism congregation pseudonym possessed fantom lyrics
first	second last final ramiz ninth fourth inaugural next third
because	though aware if instead accused follower when whom fillings
special	multi-channel permanent sliding incidental fare police substantial noddy
computer	technology management software midtown nintendo normed bmc cmai global
result	debate puritans speedy goalposts follicles consequence sockpuppet portrait

Table 1: closest words extracted by Original Skip-Gram Model

Query Word	Closest Words
can	could cannot will might must would shouldn should didn
good	bad reasonable constructive strategic decent beep stub happy helpful
number	percentage unabashed couple few variety valid large posthumous amount
january	march october december november april september july june august
word	phrase lizard ip learner egg spelling snake words
first	second last third fifth earliest latest fourth next ninth
because	struve if although naylor when portions saen vasiukov bump
special	temporary volunteer technical financial magical covert magistrate clandestine
computer	gaming document rna virtual bosch monitoring analytical digital ballistics
result	supportive osmosis part consequence mausoleum turn first-hand buoyancy cyc

Table 2: closest words extracted by Direction-Aware Skip-Gram Model

Model	Epoch	Window-size	Embedding-size	MEN	SimLex-999
Skip-Gram	5	2	64	0.0852	0.0439
	10			0.2386	0.0688
	20			0.2749	0.1351
Direction-Aware Skip-Gram	7	2	64	0.1638	0.0595
	15			0.2404	0.0966
	30			0.2983	0.1410

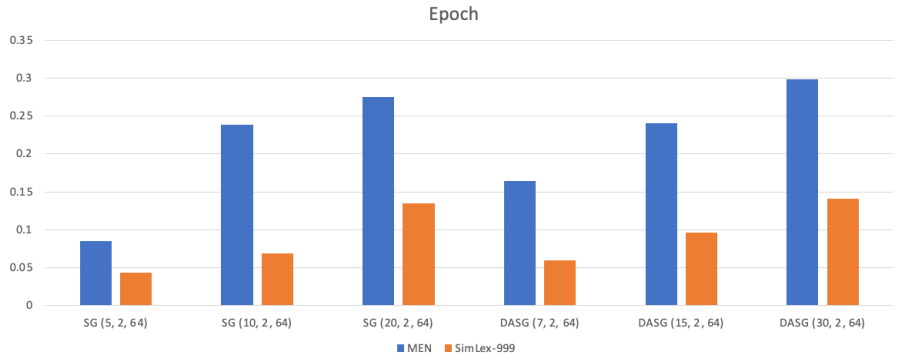


Figure 4: Original Skip-Gram and new Direction-Aware Skip-Gram with different epoch.

small epoch was effected by this factor, in company with the difference originated from the model structures themselves.

5.3 Varying Window Size

Model	Epoch	Window_size	Embedding_size	MEN	SimLex-999
Skip-Gram	10	1	64	0.1462	0.0773
		2		0.2386	0.0688
		3		0.2565	0.0569
Direction-Aware Skip-Gram	15	1	64	0.2256	0.1146
		2		0.2404	0.0966
		3		0.3054	0.1182

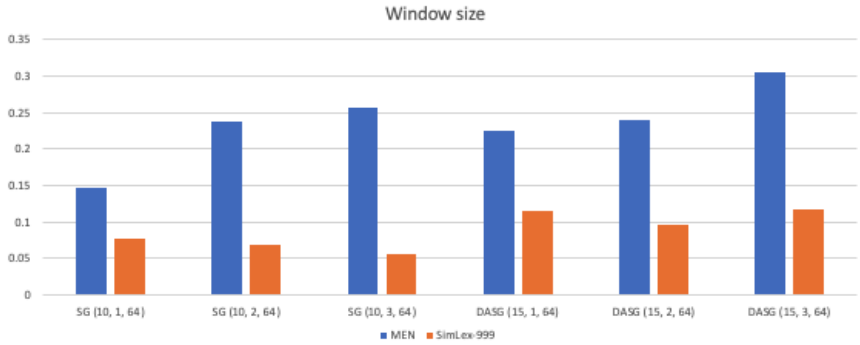


Figure 5: Original Skip-Gram and new Direction-Aware Skip-Gram with different window-size.

Next, we investigated how varying window size effects each model's characteristics while keeping all other factors constant. The MEN test results demonstrated that both the Skip-Gram and Direction-Aware Skip-Gram models' performance improved with increasing window size, similar to the effect of increasing epochs. Conversely, the SimLex-999 test showed an overall performance decreases in both models.

The result implies that increasing the window size tends to enhance 'relevance' performance while deteriorating 'similarity' performance in both models. Improved relevance measurement ability came from the expanded range of information of surrounding words. Deteriorated similarity capturing performance came from the lacking training iterations in comparison with increased amount of information. Direction-Aware Skip-Gram could relieve this effect because it assigns less information to the parameters, as it classifies context words depending on their position.

5.4 Varying Embedding Dimension

Model	Epoch	Window_size	Embedding_size	MEN	SimLex-999
Skip-Gram	10	2	32	0.1962	0.0494
			64	0.2386	0.0688
			128	0.2463	0.0792
Direction-Aware Skip-Gram	15	2	32	0.2812	0.1108
			64	0.2404	0.0966
			128	0.2503	0.0926

Finally, we examined how changes in embedding dimension affect each model, while keeping other factors fixed. Interestingly, the results showed divergent trends: as the embedding size increased, the Skip-Gram model improved in performance, while the Direction-Aware Skip-Gram model showed a decrease in performance.

Generally, increasing embedding dimension is good for performance of language model because it can capture higher order meanings. In original Skip-Gram model, it seems that the training epoch was enough to train increased parameters and achieved this goal. On the other hand, in Direction-Aware Skip-Gram model, we analyze that the number of epochs was insufficient to train increased parameter.

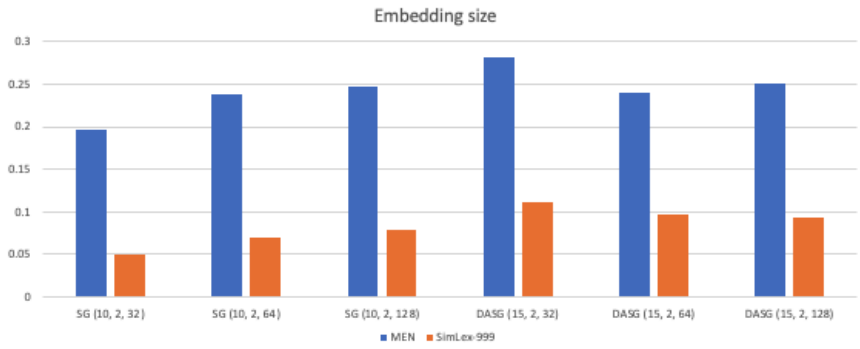


Figure 6: Original Skip-Gram and new Direction-Aware Skip-Gram with different embedding dimension.

We can clearly conclude that the difference between two models came from the number of parameters, because Direction-Aware Skip-Gram model has two separated matrices for previous and next context words.

6 Conclusion

Our experimental findings illustrate that the proposed Direction-Aware Skip-Gram model exhibits superior embedding performance overall when compared to the traditional Skip-Gram model. Both models displayed a proportionate increase in performance with more epochs, with the difference being notably prominent for fewer epochs. Regarding window size, both models demonstrated similar trends, with the Direction-Aware Skip-Gram model maintaining superior performance. The most significant distinction between the two models was observed in the embedding size. While the traditional Skip-Gram model performed better with an increase in the embedding size, the performance of the Direction-Aware Skip-Gram remained similar or even degraded in some instances.

In conclusion, our research suggests that the proposed Direction-Aware Skip-Gram model outperforms the traditional Skip-Gram in word embedding tasks. Notably, the Direction-Aware Skip-Gram model with a smaller embedding size exhibits better performance than the Skip-Gram model with an embedding size that is four times larger, implying that it can be trained more time-efficiently. Based on these results, the Direction-Aware Skip-Gram model presents a promising avenue for advancing the field of word embeddings. In particular, its superior efficiency and performance with smaller embedding sizes enable faster training and more robust representations.

The primary limitations of our work was definitely insufficient iterations to train the models. Because of it, we assigned more epochs to Direction-Aware Skip-Gram model inevitably. It may have affected results of our experiments, as we analyzed in section 5. Future research should resolve this problem by analyzing result from the experiments with more enough training time and resources. Afterward, it can explore ways to further optimize the Direction-Aware Skip-Gram model and investigate how this model can be applied to a wider range of tasks in natural language processing.

7 Contribution

Two researchers contributed to this project:

- Changho Yoon proposed the main idea of the Direction-Aware Skip gram model. He planned overall training methodology and quantitative evaluation, automatized the data extracting process. He analyzed the result from the collected data.
- Minsik Lee searched for the related work before we begin the main research. He found appropriate dataset for training and planned qualitative evaluation. He made result tables and graphs from the collected data.

References

[1] Mikolov, T., Chen, K., Corrado, G., & Dean, J. (2013). Efficient Estimation of Word Representations in Vector Space. In Proceedings of the International Conference on Learning Representations (ICLR 2013).

[2] Mikolov, T., Sutskever, I., Chen, K., Corrado, G. S., & Dean, J. (2013). Distributed Representations of Words and Phrases and their Compositionality. In Advances in Neural Information Processing Systems 26 (NIPS 2013).

[3] Bojanowski, P., Grave, E., Joulin, A., & Mikolov, T. (2017). Enriching Word Vectors with Subword Information. In Transactions of the Association for Computational Linguistics, 5, 135-146.

[4] Peters, M. E., Neumann, M., Iyyer, M., Gardner, M., Clark, C., Lee, K., & Zettlemoyer, L. (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) (pp. 2227-2237).

[5] Wang Ling, Chris Dyer, Alan W. Black, and Isabel Trancoso. 2015. Two/Too Simple Adaptations of Word2Vec for Syntax Problems. In Proceedings of the 2015 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, pages 1299–1304, Denver, Colorado. Association for Computational Linguistics.

[6] Yan Song, Shuming Shi, Jing Li, and Haisong Zhang. 2018. Directional Skip-Gram: Explicitly Distinguishing Left and Right Context for Word Embeddings. In Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 2 (Short Papers), pages 175–180, New Orleans, Louisiana. Association for Computational Linguistics.