Semantics- and Syntax-Related Subvectors in the Skip-Gram Embeddings (Student Abstract)

Maxat Tezekbayev, Zhenisbek Assylbekov, Rustem Takhanov
Department of Mathematics, School of Sciences and Humanities,
Nazarbayev University, Nur-Sultan, Kazakhstan

Abstract

We show that the skip-gram embedding of any word can be decomposed into two subvectors which roughly correspond to semantic and syntactic roles of the word.

Introduction

Assuming that words have already been converted into indices, let $\{1, \ldots, n\}$ be a finite vocabulary of words. Following the setups of the widely used WORD2VEC (Mikolov et al. 2013) model, we consider *two* vectors per each word i:

- $\mathbf{w}_i$ is an embedding of the word i when i is a center word,
- $\mathbf{c}_i$ is an embedding of the word i when i is a context word.

We follow the assumptions of Assylbekov and Takhanov (2019) on the nature of word vectors, context vectors, and text generation, i.e.

1. A priori word vectors $\mathbf{w}_1, \ldots, \mathbf{w}_n \in \mathbb{R}^d$ are i.i.d. draws from isotropic multivariate Gaussian distribution: $\mathbf{w}_i \overset{\text{iid}}{\sim} \mathcal{N}\left(\mathbf{0}, \frac{1}{d}\mathbf{I}\right)$, where $\mathbf{I}$ is the $d \times d$ identity matrix.
2. Context vectors $\mathbf{c}_1, \ldots, \mathbf{c}_n$ are related to word vectors according to $\mathbf{c}_i = \mathbf{Q}\mathbf{w}_i$, $i = 1, \ldots, n$, for some orthogonal matrix $\mathbf{Q} \in \mathbb{R}^{d \times d}$.
3. Given a word j, the probability of any word i being in its context is given by

$$p(i \mid j) \propto p_i \cdot e^{\mathbf{w}_j^\top \mathbf{c}_i} \quad (1)$$

where $p_i = p(i)$ is the unigram probability for the word i.

Hypothesis. Under the assumptions 1–3 above, Assylbekov and Takhanov (2019) showed that each word's vector $\mathbf{w}_i$ splits into two approximately equally-sized subvectors $\mathbf{x}_i$ and $\mathbf{y}_i$, and the model (1) for generating a word i in the context of a word j can be rewritten as

$$p(i \mid j) \approx p_i \cdot e^{\mathbf{x}_j^\top \mathbf{x}_i - \mathbf{y}_j^\top \mathbf{y}_i}.$$

Interestingly, embeddings of the first type ($\mathbf{x}_i$ and $\mathbf{x}_j$) are responsible for pulling the word i into the context of the word j, while embeddings of the second type ($\mathbf{y}_i$ and $\mathbf{y}_j$) are responsible for pushing the word i away from the context of the word j. We hypothesize that the $\mathbf{x}$-embeddings are more related to semantics, whereas the $\mathbf{y}$-embeddings are more related to syntax. In what follows we provide a motivating example for this hypothesis and then empirically validate it through controlled experiments.

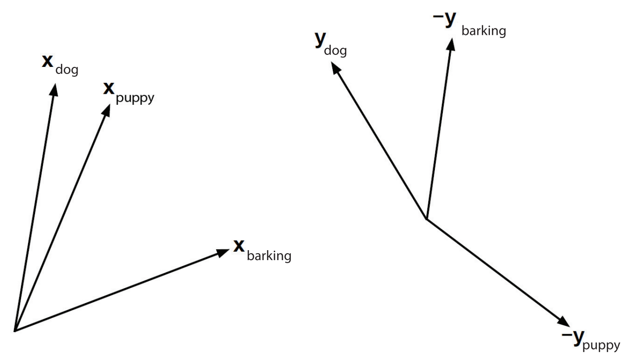

Figure 1: $\mathbf{x}$- and $\mathbf{y}$-embeddings

Motivating Example

Consider a phrase

the dog barking at strangers

The word 'barking' appears in the context of the word 'dog' but the word vector $\mathbf{w}_{\text{barking}}$ is not the closest to the word vector $\mathbf{w}_{\text{dog}}$ (see Table 2). Instead, these vectors are split

$$\mathbf{w}_{\text{dog}}^\top = [\mathbf{x}_{\text{dog}}^\top; \mathbf{y}_{\text{dog}}^\top]$$
$$\mathbf{w}_{\text{barking}}^\top = [\mathbf{x}_{\text{barking}}^\top; \mathbf{y}_{\text{barking}}^\top]$$

in such way that the quantity $\mathbf{x}_{\text{dog}}^\top \mathbf{x}_{\text{barking}} - \mathbf{y}_{\text{dog}}^\top \mathbf{y}_{\text{barking}}$ is large enough. We can interpret this as follows: the word 'barking' is semantically close enough to the word 'dog' but is not the closest one: e.g. $\mathbf{w}_{\text{puppy}}$ is much closer to $\mathbf{w}_{\text{dog}}$ than $\mathbf{w}_{\text{barking}}$; on the other hand the word 'barking' syntactically fits better being next to the word 'dog' than 'puppy', i.e. $-\mathbf{y}_{\text{dog}}^\top \mathbf{y}_{\text{puppy}} < -\mathbf{y}_{\text{dog}}^\top \mathbf{y}_{\text{barking}}$. This combination of semantic

Data	Embeddings	Size	Finkelstein et al. WordSim	Bruni et al. MEN	Radinsky et al. M. Turk	Luong, Socher, and Manning Rare Words	Google	MSR
text8	$\mathbf{w} := [\mathbf{x};\mathbf{y}]$	200	.646	.650	.636	.063	.305	.319
	Only $\mathbf{x}$	100	.703	.693	.673	.149	.348	.213
	Only $\mathbf{y}$	100	.310	.102	.193	.019	.032	.128
enwik9	$\mathbf{w} := [\mathbf{x};\mathbf{y}]$	200	.664	.697	.616	.216	.518	.423
	Only $\mathbf{x}$	100	.714	.729	.652	.256	.545	.303
	Only $\mathbf{y}$	100	.320	.188	.196	.091	.096	.251

Table 1: Evaluation of word vectors and subvectors on the analogy tasks (Google and MSR) and on the similarity tasks (the rest). For word similarities evaluation metric is the Spearman's correlation with the human ratings, while for word analogies it is the percentage of correct answers. Model sizes are in number of trainable parameters.

word i	$\mathbf{w}_{\text{dog}}^\top \mathbf{c}_i$	$\mathbf{w}_{\text{dog}}^\top \mathbf{w}_i$	$\mathbf{x}_{\text{dog}}^\top \mathbf{x}_i$	$-\mathbf{y}_{\text{dog}}^\top \mathbf{y}_i$
puppy	-0.204	13.331	6.564	-6.768
barking	-0.263	10.343	5.040	-5.303

Table 2: Dot products between vectors.

proximity ($\mathbf{x}_{\text{dog}}^\top \mathbf{x}_{\text{barking}}$) and syntactic fit ($-\mathbf{y}_{\text{dog}}^\top \mathbf{y}_{\text{barking}}$) allows the word 'barking' to appear in the context of the word 'dog'.

Experiments

In this section we empirically verify our hypothesis. We train SGNS with tied weights (Assylbekov and Takhanov 2019) on two widely-used datasets, `text8` and `enwik9`,[1] which gives us word embeddings as well as their partitions:

$$\mathbf{w}_i^\top := [\mathbf{x}_i^\top; \mathbf{y}_i^\top].$$

The source code that reproduces our experiments is available at https://github.com/MaxatTezekbayev/Semantics--and-Syntax-related-Subvectors-in-the-Skip-gram-Embeddings.

$\mathbf{x}$-Subvectors Are Related to Semantics

We evaluate the whole vectors $\mathbf{w}_i$'s, as well as the subvectors $\mathbf{x}_i$'s and $\mathbf{y}_i$'s on standard semantic tasks — word similarity and word analogy. We used the HYPERWORDS tool of Levy, Goldberg, and Dagan (2015) and we refer the reader to their paper for the methodology of evaluation. The results of evaluation are provided in Table 1. As one can see, the $\mathbf{x}$-subvectors outperform the whole $\mathbf{w}$-vectors in the similarity tasks and show competitive performance in the analogy tasks. However, the $\mathbf{y}$-parts demonstrate poor performance in these tasks. This shows that the $\mathbf{x}$-subvectors carry more semantic information than the $\mathbf{y}$-subvectors.

$\mathbf{y}$-Subvectors Are Related to Syntax

We train a softmax regression by feeding in the embedding of a current word to predict the part-of-speech (POS) tag of the next word:

$$\widehat{\text{POS}}[t+1] = \text{softmax}(\mathbf{A}\mathbf{w}[t] + \mathbf{b})$$

[1] http://mattmahoney.net/dc/textdata.html. The `enwik9` data was processed with the Perl-script WIKIFIL.PL provided on the same webpage.

We evaluate the whole vectors and the subvectors on tagging the Brown corpus with the Universal POS tags. The resulting accuracies are provided in Table 3. We can see that

Embeddings	Size	Trained on `text8`	Trained on `enwik9`
$\mathbf{w} := [\mathbf{x};\mathbf{y}]$	200	.445	.453
Only $\mathbf{x}$	100	.381	.384
Only $\mathbf{y}$	100	.426	.451

Table 3: Accuracies on a simplified POS-tagging task.

the $\mathbf{y}$-subvectors are more suitable for POS-tagging than the $\mathbf{x}$-subvectors, which means than the $\mathbf{y}$-parts carry more syntactic information than the $\mathbf{x}$-parts.

Conclusion

Theoretical analysis of word embeddings gives us better understanding of their properties. Moreover, theory may provide us interesting hypotheses on the nature and structure of word embeddings, and such hypotheses can be verified empirically as is done in this paper.

Acknowledgements

This work is supported by the Nazarbayev University Collaborative Research Program 091019CRP2109, and by the Committee of Science of the Ministry of Education and Science of the Republic of Kazakhstan, IRN AP05133700.

References

Assylbekov, Z., and Takhanov, R. 2019. Context vectors are reflections of word vectors in half the dimensions. *Journal of Artificial Intelligence Research* 66:225–242.

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

Mikolov, T.; Sutskever, I.; Chen, K.; Corrado, G. S.; and Dean, J. 2013. Distributed representations of words and phrases and their compositionality. In *Proceedings of NeurIPS*, 3111–3119.