Interpretable Word Embedding Contextualization

Kyoung-Rok Jang
School of Computing
KAIST
Daejeon, South Korea
kyoungrok.jang@kaist.ac.kr

Sung-Hyon Myaeng
School of Computing
KAIST
Daejeon, South Korea
myaeng@kaist.ac.kr

Sang-Bum Kim
Naver Corp.
Seongnam-si, South Korea
sangbum.kim@navercorp.com

Abstract

In this paper, we propose a method of calibrating a word embedding, so that the semantic it conveys becomes more relevant to the context. Our method is novel because the output shows clearly which senses that were originally presented in a target word embedding become stronger or weaker. This is possible by utilizing the technique of using sparse coding to recover senses that comprises a word embedding.

1 Introduction

In this paper we propose a method of generating contextualized word embeddings. What we mean by 'contextualized' is that standard embeddings such as Skip-gram and GloVe are modified to reflect their contexts. For instance, *apple* appeared in fruit-implying context should become more similar to *fruit* than it was in the prior state.

We need contextualized embeddings because not all information contained in an embedding is helpful for modelling accurately the meaning of a word in context (e.g. company-related senses of *apple* in fruit-implying context). Since word embeddings are trained on unconstrained variation of contexts, using word embeddings as-is is like taking the risk of feeding our subsequent models (e.g. classifiers) with noises that are not relevant to the given context.

We formulate our task as calibrating senses contained in embeddings so that contextually relevant senses (e.g. fruit-ness) becomes stronger and the others (e.g. company-ness) become weaker. To achieve this we utilize the technique of recovering standard word embeddings as linear composition of different senses, proposed by (Murphy et al., 2012; Faruqui et al., 2015; Arora et al., 2016). After applying the technique word embeddings are transformed into high dimensional (e.g. 2,500) and sparse (only small portion of dimensions are nonzero) embeddings. This makes our method *interpretable* since extracted senses can give us "a succinct description of which other words co-occur with a specific word sense". More detailed explanation is presented in Section 3.

Using the technique we first decompose word embeddings of a target word (to be contextualized) and context words into linear composition of senses, then identify strong senses extracted from context and regard them as contextually relevant. Finally we use the contextually relevant senses for calibrating the senses contained in a target word.

2 Task and Model

2.1 Task

We show that our method is effective by applying it to Word Sense Discrimination (WSD) task. For brevity we present only one instance of sense discrimination: discriminating *apple* as a *fruit* or a *company* depending on given context. The result is described in Section 3

2.2 Embedding Decomposition

To contextualize a target word's embedding, we should first decompose the participating word embeddings (i.e. target and context words) into linear composition of different senses. As a result we obtain high dimensional and sparse embeddings, in which few 'activated' dimensions represent significant senses reside in a word embedding.

For our preliminary experiment we use Non-negative Sparse Embedding (NNSE) proposed by (Murphy et al., 2012). We use NNSE partly because pre-trained word embeddings are publicly available at it's official website[1].

[1] https://www.cs.cmu.edu/ bmurphy/NNSE/

Proceedings of the 2018 EMNLP Workshop BlackboxNLP: Analyzing and Interpreting Neural Networks for NLP, pages 341–343
Brussels, Belgium, November 1, 2018.

2.3 Contextualization

Figure 1 shows the baseline method of performing word sense discrimination only using embeddings of a target word and context words (Kober et al., 2017). Basically it takes a sum of a target word and context word embeddings, and then decide whether the target word in the context has the same sense by calculating cosine similarity. In our work, we modify the composition (i.e. contextualization) step.

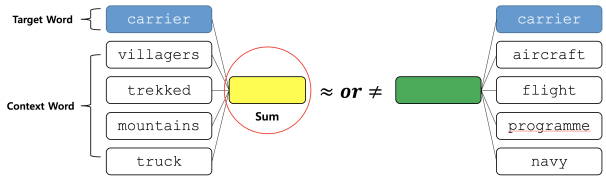

Figure 1: The baseline method. The red circle indicates the step where we make our modification.

We first retrieve NNSE embeddings of a target word and context words. We then generate $emb_{context}$ by summing all the context embeddings to identify contextually relevant dimensions (i.e. senses) (Figure 2).

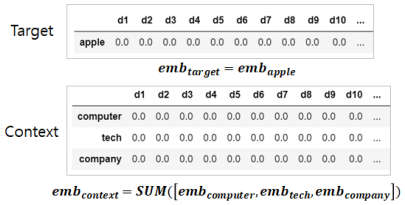

Figure 2: Calculation of target and context embedding. Since NNSE is sparse, the visual part of embeddings in this figure has all zero values.

We hypothesize that the dimensions of $emb_{context}$ that have zero value are irrelevant to the context. This is because in NNSE, a specific dimension is activated only when it represents a significant sense contained in word embeddings. So if a dimension is still deactivated after the sum of all context words, it means that the sense the dimension represents must have low importance.

So we turn off such atoms as well in the target word embedding by applying element-wise multiplication between a target word and a context embedding (Equation 1). This weakens senses that are irrelevant to context.

$$emb_{contextualized} = emb_{target} * emb_{context} \quad (1)$$

Finally, we normalize our contextualized embedding then use it in our task (Equation 2). This strengthens concepts that are relevant to context.

$$\frac{emb_{contextualized}}{\|emb_{contextualized}\|_2} \quad (2)$$

3 Preliminary Experiment and Result

In our experiment we attempt to discriminate *company* and *fruit* senses of *apple* by contextualizing with a relevant context.

Figure 3 shows the calibrated senses of *apple*. In the figure, 'd2104' means it is 2104th dimension of the embedding, and the list of words in the figure is an interpretation of the dimension (i.e. sense), which can be derived by sorting the whole vocabulary by the strength of the specific dimension in reverse order. The score is the strength of the dimension. Note that the identified dimensions are all extracted from *apple* embedding, while the values are calibrated.

The figure shows that our contextualization method is able to strengthen and weaken the senses of *apple* by reflecting the given context.

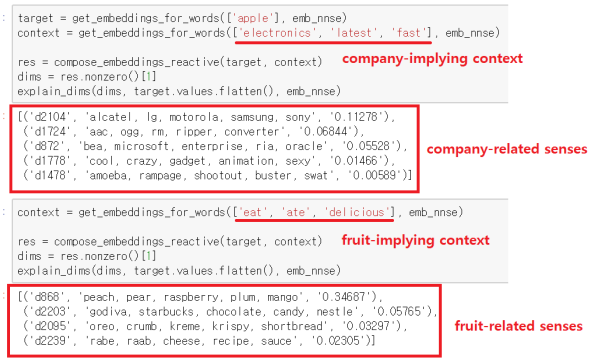

Figure 3: The calibrated senses of *apple*.

4 Discussion and Future Work

We showed that our method could be both interpretable and effective in performing a word sense discrimination task. Our method can be utilized not only to discriminate senses but to decide types of named entities or any other tasks that require inferring the context specific meaning of words. As a future work, we will try to elaborate our method and prove the efficacy of our method by testing on well-known tasks.

References

Sanjeev Arora, Yuanzhi Li, Yingyu Liang, Tengyu Ma, and Andrej Risteski. 2016. Linear Algebraic Structure of Word Senses, with Applications to Polysemy.

Manaal Faruqui, Yulia Tsvetkov, Dani Yogatama, Chris Dyer, and Noah Smith. 2015. Sparse Overcomplete Word Vector Representations.

Thomas Kober, Julie Weeds, John Wilkie, Jeremy Reffin, and David Weir. 2017. One Representation per Word - Does it make Sense for Composition? In *Proceedings of the 1st Workshop on Sense, Concept and Entity Representations and their Applications*, pages 79–90.

Brian Murphy, Partha Pratim Talukdar, and Tom Mitchell. 2012. Learning Effective and Interpretable Semantic Models using Non-Negative Sparse Embedding. *Proceedings of COLING 2012: Technical Papers*, (December 2012):1933–1950.