Elucidating Conceptual Properties from Word Embeddings

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

Abstract

In this paper, we introduce a method of identifying the components (i.e. dimensions) of word embeddings that strongly signifies *properties* of a word. By elucidating such properties hidden in word embeddings, we could make word embeddings more interpretable, and also could perform property-based meaning comparison. With the capability, we can answer questions like "To what degree a given word has the property *cuteness*?" or "In what perspective two words are similar?". We verify our method by examining how the strength of property-signifying components correlates with the degree of *prototypicality* of a target word.

1 Introduction

Modeling the meaning of words has long been studied and served as a basis for almost every kind of NLP tasks. Most recent word modeling techniques are based on neural networks, and the word representations produced by such techniques are called word embeddings, which are usually low-dimensional, dense vectors of continuous-valued components. Although word embeddings have been proved for their usefulness in many tasks, the question of what are represented in them is understudied.

Recent studies report empirical evidence that indicates word embeddings may reflect some *property* information of a target word (Erk, 2016; Levy et al., 2015). Learning the properties of a word would be helpful because many NLP tasks can be related to "finding words that possess similar properties", which include finding synonyms, named entity recognition (NER). Without a method for explicating what properties are contained in embeddings, however, researchers have mostly focused on improving the performance in well-known semantic benchmark tasks (e.g. SimLex-999) as a way to find better embeddings.

Performing well in such benchmark tasks is valuable but provides little help in understanding the inside of the black box. For instance, it is not possible to answer to questions like "To what degree a given word has the property *cuteness*?".

One way to solve this problem is to elucidate properties that are encoded in word embeddings and associate them with task performances. With the capability, we can not only enhance our understanding of word embeddings but also make it easier to make comparisons among heterogeneous word embedding models in more coherent ways. Our immediate goal in this paper is to show the feasibility of explicating properties contained in word embeddings.

Our research can be seen as an attempt to increase the *interpretability* of word embeddings. It is in line with an attempt to provide a human-understandable explanation for complex machine learning models, with which we can gain enough confidence to use them in decision-making processes.

There has been a line of work devoted to identifying components that are important for performing various NLP tasks such as sentiment analysis or NER (Faruqui et al., 2015; Fyshe et al., 2015; Herbelot and Vecchi, 2015; Karpathy et al., 2016; Li et al., 2016a; Li et al., 2016b; Rothe et al., 2016). Those works are analogous to ours in that they try to inspect the role of the components in word embedding. However, they just attempt to identify key features for specific tasks rather than elucidating properties. In contrast, our question is "what comprises word embeddings?" not "what components are important for performing well in a specific task?"

Proceedings of the 1st Workshop on Sense, Concept and Entity Representations and their Applications, pages 91–95,
Valencia, Spain, April 4 2017.

2 Feasibility Study

2.1 Background

Word embeddings can be seen as representing concepts of a word. As such, we attempt to design an property-related experiment around manipulation of concepts. In particular, we bring in the category theory (Murphy, 2004) where the notion of category is defined to be "grouping concepts that share similar properties". In other words, properties have a direct bearing on concepts and their categories, according to the theory.

On the other hand, researchers have argued that some concepts are more *typical* (or central) than others in a category (Rosch, 1973; Rosch, 1975). For instance, *apple* is more typical than *olive* in the fruit category. The typicality is a graded phenomenon, and may rise due to the strength of 'essential' properties that make a concept a specific category.

The key ideas from the above are 1) concepts of the same category share similar properties and 2) some concepts that have strong essential properties are considered more typical in specific category, and they guided our experiment design.

2.2 Design

The goal of this study is to show the feasibility of sifting property information from word embeddings. We assume that a concept's property information is captured and distributed over one or more components (dimensions) of embeddings during the learning process. Since the concepts that belong to the same category are likely to share similar properties, there should be some salient components that are shared among them. We call such components as SIG-PROPS (for significant properties) of a specific category.

In this feasibility study, we hypothesize that the strength of SIG-PROPS is strongly correlated with the degree of concept's typicality. This is based on the theory introduced in Section 2.1, that the typicality phenomenon rises due to the strength of essential properties a target concept possesses. So the concept that has (higher/lower) SIG-PROPS values than others should be (more typical/less typical) than other concepts.

2.3 Datasets

For our experiment dealing with typicality of concepts, we needed both (pre-trained) word embeddings and a dataset that encodes typicality scores of concepts to a set of categories. Below we describe two datasets we used in our experiment: HyperLex and Non-Negative Sparse Embedding (NNSE).

2.3.1 Dataset: Non-Negative Sparse Embedding (NNSE)

One desirable quality we wanted from the word embeddings to be used in our experiment is that there should be clear contrast between informative and non-informative components. In ordinary dense word embeddings, usually every component is filled with a non-zero value.

The Non-Negative Sparse Embedding (NNSE) (Murphy et al., 2012) fulfills the condition in the sense that insignificant components are set to zero. The NNSE component values falling between 0 and 1 (non-negative) are generated by applying the non-negative sparse coding algorithm (Hoyer, 2002) to ordinary word embeddings (e.g. word2vec).

2.3.2 Dataset: HyperLex

HyperLex is a dataset and evaluation resource that quantifies the extent of the semantic category membership (Vulić et al., 2016). A total of 2,616 concept pairs are included in the dataset, and the strength of category membership is given by native English speakers and recorded in *graded* manner. This graded category membership can be interpreted as a 'typicality score' (1–10). Some samples are shown in Table 1.

Concept	Category	Score
basketball	activity	10
spy	agent	8
handbag	bowl	0

Table 1: A HyperLex sample. The score is the answer to the question "To what degree is concept a type of category?"

3 Experiment

3.1 Preparation

We first prepared pre-trained NNSE embeddings. The authors released pre-trained model on their website[1]. We used the model trained with dependency context ('Dependency model' on the website), because as reported in (Levy and Goldberg, 2014), models trained on dependency context tend

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

to prefer *functional* similarity (hogwarts — sunnydale) rather than *topical* similarity (hogwarts — dumbledore)[2]. The embeddings are more sparse than ordinary embeddings and have 300 components.

Next we fetched HyperLex dataset at the author's website[3]. To make the settings suitable to our experiment goal, we selected categories with the following criteria:

1. The categories and instances must be concrete nouns (e.g. food). This is because people are more coherent in producing the properties of concrete nouns (Murphy, 2004). So the embeddings of concrete nouns should contain more clear property information than other types of words.

2. The categories must contain enough number of instances (not 1 or 2). This is to gain reliable result.

3. Some categories are sub-category of another selected category while others are not related. This is to see the discriminative and overlapping effect of identified SIG-PROPS between categories. Related categories should share a set of strong SIG-PROPS, while unrelated categories shouldn't.

As a result, we selected five categories: *food*, *fruit* (sub-category of food), *animal*, *bird* (sub-category of animal), and *instrument*. We fetched the concepts that belong to the categories and then filtered out those that aren't contained in the pre-trained NNSE embeddings. The final size of each category is shown in Table 2.

Category	# of Concepts
food	54
fruit	9
animal	46
bird	16
instrument	14

Table 2: The size of selected categories

In the next section, we explain how we identified SIG-PROPS of each category.

[2]We thought "sharing similar function" is more compatible with the notion of *sharing similar properties*. The topical similarity is less indicative of having properties in common.

[3]http://people.ds.cam.ac.uk/iv250/hyperlex.html

3.2 Identification of SIG-PROPS

The goal of this step is to find SIG-PROPS that might represent each category. Simply put, SIG-PROPS are the components that have on average high value compared to other components of the concepts in the same category. We identified SIG-PROPS by 1) calculating an average value of each component across the concepts with the same category, then 2) choosing those components whose average value is above h. We empirically set h to 0.2.

Category	SIG-PROPS	
	Comp. ID	Avg.
instrument	c88	0.806
	c258	0.769
animal	**c154**	0.587
	c265	0.221
bird	**c154**	0.550
	c265	0.213
food	c207	0.298
	c233	0.269
fruit	c229	0.492
	c27	0.369
	c156	0.349
	c44	0.264
	c233	0.206

Table 3: SIG-PROPS of each category. The strings in "Comp. ID" column are the component IDs (c1–c300). "Avg." column indicates the average value of the component across all the concepts under that category.

Table 3 shows the identified SIG-PROPS. The number of SIG-PROPS is different across categories. Interestingly, there is component overlap between taxonomically similar categories ('c154' and 'c265' between *animal–bird*, 'c233' between *food–fruit*), while there is none between unrelated categories (*instrument–animal–food*).

This initial observation is encouraging for our feasibility study in that indeed SIG-PROPS can play a role of distinguishing or associating categories. We argue that the identified SIG-PROPS strongly characterize each category, showing that we can associate vector components with properties.

3.3 Correlation between SIG-PROPS and concepts typicality scores

In this section, we check how the strength of SIG-PROPS correlates with the typicality scores. Note that the range of SIG-PROPS values differ from category to category — those for instrument are especially high, which might indicate they are highly

representative.

Our assumption is that if the identified SIG-PROPS truly represent the essential quality of a category, the strength of SIG-PROPS should be proportional to concepts' typicality scores (equation 1).

$$Strength(\text{SIG-PROPS}) \propto Typicality(Concept) \quad (1)$$

We observe this phenomenon by calculating Pearson correlation between the strength of SIG-PROPS and typicality scores. For instance, suppose we calculate the correlation between 'c88' (88th component) of *instrument* concepts and their typicality score. We inspect the instrument concepts one by one, collect their 'c88' values (x) and typicality scores (y), and then measure the tendency of changes in the two variables.

The result is shown in Table 4–8 where the column 'Rank' shows the rank of the component's correlation score (compared to other components). For instance, in Table 4 'c88' has the highest correlation with the concept's typicality score.

SIG-PROPS	Correlation	Corr. rank
c88	0.926	1st
c258	0.918	2nd

Table 4: Corr(SIG-PROPS, typicality): *instrument*

SIG-PROPS	Correlation	Corr. rank
c154	0.549	1st
c265	0.265	2nd

Table 5: Corr(SIG-PROPS, typicality): *animal*

SIG-PROPS	Correlation	Corr. rank
c154	0.783	1st
c265	0.563	2nd

Table 6: Corr(SIG-PROPS, typicality): *bird*

SIG-PROPS	Correlation	Corr. rank
c233	0.255	1st
c120	0.224	2nd
c207	0.216	4th
c192	0.030	104th

Table 7: Corr(SIG-PROPS, typicality): *food*

SIG-PROPS	Correlation	Corr. rank
c229	0.743	1st
c233	0.540	4th
c27	0.516	5th
c44	0.474	7th
c156	-0.663	85th

Table 8: Corr(SIG-PROPS, typicality): *fruit*

As the results show, there is clear tendency SIG-PROPS having high correlation with the typicality scores. Most of the SIG-PROPS showed meaningful correlation (> 0.5) with the typicality score or placed at the top in the component–typicality correlation ranking. The result strongly indicates that even when we apply the simple method of identifying SIG-PROPS and regarding them as properties, they serve as strong indicators for the concept's typicality.

4 Conclusion and Future Work

Although limited in scale, our work showed the feasibility of discovering properties from word embeddings. Not only SIG-PROPS can be used to increase the interpretability of word embeddings, but also enable us more elaborate, property-based meaning comparison.

Our next step would be checking the applicability to general NLP tasks (e.g. NER, synonym identification). Also, applying our method to word embeddings that have more granular components (e.g. 2,500) might be helpful for identifying SIG-PROPS in more granular level.

Acknowledgments

This work was supported by Institute for Information & communications Technology Promotion(IITP) grant funded by the Korea government(MSIP) (No. R0126-16-1002, Development of agro-livestock cloud and application service for balanced production, transparent distribution and safe consumption based on GS1)

References

Katrin Erk. 2016. What do you know about an alligator when you know the company it keeps? *Semantics and Pragmatics Article*, 9(17):1–63.

Manaal Faruqui, Yulia Tsvetkov, Dani Yogatama, Chris Dyer, and Noah A. Smith. 2015. Sparse overcomplete word vector representations. In *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International*

Joint Conference on Natural Language Processing (Volume 1: Long Papers), pages 1491–1500, Beijing, China, July.

Alona Fyshe, Leila Wehbe, Partha P. Talukdar, Brian Murphy, and Tom M. Mitchell. 2015. A compositional and interpretable semantic space. In *Proceedings of the 2015 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 32–41, Denver, Colorado, May–June.

Aurélie Herbelot and Eva Maria Vecchi. 2015. Building a shared world: Mapping distributional to model-theoretic semantic spaces. In *Proceedings of EMNLP*, number September, pages 22–32, Lisbon, Portugal.

P. O. Hoyer. 2002. Non-negative sparse coding. In *Neural Networks for Signal Processing - Proceedings of the IEEE Workshop*, volume 2002-Janua, pages 557–565.

Andrej Karpathy, Justin Johnson, and Li Fei-Fei. 2016. Visualizing and Understanding Recurrent Networks. *International Conference on Learning Representations (ICLR)*, pages 1–13.

Omer Levy and Yoav Goldberg. 2014. Dependency-based word embeddings. In *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 302–308, Baltimore, Maryland, June.

Omer Levy, Steffen Remus, Chris Biemann, and Ido Dagan. 2015. Do supervised distributional methods really learn lexical inference relations? In *Proceedings of the 2015 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 970–976, Denver, Colorado, May–June.

Jiwei Li, Xinlei Chen, Eduard Hovy, and Dan Jurafsky. 2016a. Visualizing and Understanding Neural Models in NLP. In *Naacl*, pages 1–10.

Jiwei Li, Will Monroe, and Dan Jurafsky. 2016b. Understanding Neural Networks through Representation Erasure. In *Arxiv*.

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

Gregory Murphy. 2004. *The big book of concepts*. MIT press.

Eleanor H. Rosch. 1973. Natural categories. *Cognitive Psychology*, 4(3):328–350.

Eleanor H. Rosch. 1975. Cognitive representations of semantic categories. *Journal of experimental psychology: General*, 104(3):192.

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, San Diego, California, June.

Ivan Vulić, Daniela Gerz, Douwe Kiela, Felix Hill, and Anna Korhonen. 2016. HyperLex: A Large-Scale Evaluation of Graded Lexical Entailment. *Arxiv*.