Evaluating Word Embeddings for Language Acquisition

Raquel G. Alhama[1,2] **Caroline Rowland**[1,3] **Evan Kidd**[1,3,4,5]
[1]Language Development Department, Max Planck Institute for Psycholinguistics
[2]Department of Cognitive Science & Artificial Intelligence, Tilburg University
[3]Donders Institute for Brain, Cognition & Behaviour, Radboud University
[4]The Australian National University
[5]ARC Centre of Excellence for the Dynamics of Language
rgalhama@tilburguniversity.edu,
{caroline.rowland,evan.kidd}@mpi.nl

Abstract

Continuous vector word representations (or word embeddings) have shown success in capturing semantic relations between words, as evidenced by evaluation against behavioral data of adult performance on semantic tasks (Pereira et al., 2016). Adult semantic knowledge is the endpoint of a language acquisition process; thus, a relevant question is whether these models can also capture *emerging* word representations of young language learners. However, the data for children's semantic knowledge across development is scarce. In this paper, we propose to bridge this gap by using Age of Acquisition norms to evaluate word embeddings learnt from child-directed input. We present two methods that evaluate word embeddings in terms of (a) the semantic neighbourhood density of learnt words, and (b) convergence to adult word associations. We apply our methods to bag-of-words models, and find that (1) children acquire words with fewer semantic neighbours earlier, and (2) young learners only attend to very local context. These findings provide converging evidence for validity of our methods in understanding the prerequisite features for a distributional model of word learning.

1 Introduction

Word embeddings have a long tradition in Computational Linguistics. There exist a range of methods to derive word embeddings based on the distributional paradigm, such that words with similar embeddings are semantically related. These embeddings are often evaluated either extrinsically, on how well they boost performance on a certain task, or intrinsically, by comparing representations against behavioral data from tests of semantic similarity, synonymity, analogy or word association (Pereira et al., 2016).

Adult semantic knowledge is the culmination of a language acquisition process; therefore, a relevant question is whether these models can also capture *emerging* word representations of language learners. A capacity for distributional analysis is a basic assumption of all theories of language acquisition: children are capable of performing distributional analyses over their input from a young age (Saffran et al., 1996), motivating the use of word embeddings for modelling language acquisition. However, the evaluation of emergent word representations is far from straightforward, as there is no availability of the kind of semantic judgements that we have for adults.

This paper presents two methods for evaluating word embeddings for language acquisition. We apply our methods to two bag-of-words models, and evaluate them on the acquisition of nouns in English-speaking children[1].

2 Models

Bag-of-words models offer a good starting point to evaluate word representations in the context of language acquisition, given their minimal assumptions on knowledge of word order: once the context of a word is determined, the order in which words appear in this context is ignored by these type of models. We explore a range of hyperparameter configurations of two models: a 'context-counting' model involving a PPMI matrix compressed with Singular Value Decomposition (SVD), and the Skipgram with Negative Sampling (SGNS)

[1]We share the code for these methods at https://github.com/rgalhama/wordrep_cmcl2020

Proceedings of the Workshop on Cognitive Modeling and Computational Linguistics, pages 38–42
Online Event, November 19, 2020.
https://doi.org/10.18653/v1/P17

version of *word2vec* (Mikolov et al., 2013). Note that, although these models have been found to implicitly optimize the same shifted-PPMI matrix (Levy and Goldberg, 2014), they are unlikely to obtain the same results without careful parameter alignment. Our goal by selecting these two approaches is to increase the variability of model performance within the bag-of-words paradigm.

The hyperparameters we explore include: window size [1,2,3,4,5,7,10], minimum frequency threshold [10,50,100], dynamic window (for SGNS), negative sampling in SGNS [0,15] (and its equivalents as shifted-PPMI), eigenvalue in SVD [0,0.5,1]. We restrict our analyses to vectors of size 100. We use the Hyperwords package from Levy et al. (2015).

3 Data

We trained the models on transcriptions of child-directed speech, i.e. samples of naturalistic productions in the linguistic environment of a child. We extracted the child-directed speech data from the CHILDES database (MacWhinney, 2000), for all the varieties of English, for ages ranging from 0 to 60 months. We used the `childesr` library to extract the child-directed utterances (Sanchez et al., 2019) [2]. Word tokens were coded at the lemma level. The resulting dataset contains a total number of 3,135,822 sentences, 34,961 word types, and 12,975,520 word tokens.

To evaluate the models, we used data collected with the MacArthur-Bates Communicative Development Inventory forms (CDI). These are forms, given to parents of young children, that contain checklists of common early acquired words. Parents complete the forms according to whether their child *understands* or *produces* each of those words. These forms are collected at different ages, and thus can be used to estimate the Age of Acquisition (AoA) of words. We used all the variants of English 'Words & Sentences' CDIs from the Wordbank database (Frank et al., 2017), with the exception of those involving twins (as significant differences have been observed in the language development of twins and singletons, Tomasello et al., 1986). We estimated the AoA of a word by considering that a word is acquired at the age at which at least 50% of the children in the sample produced a given word.

[2] http://childes-db.stanford.edu/about.html

4 Method 1: Neighbourhood Density

Our first evaluation method is inspired by prior work on human word learning, presented in Hills et al. (2010). In their work, the authors modeled the emerging network of semantic associations that children build during language acquisition. Their model consists of a simple word co-occurrence matrix, where all the counts greater than zero are flattened into a count of one, resulting in a binary matrix. The authors view the resulting matrix as a network of associations, where words are connected only if they have co-occurred. The number of connections of each word is then used as an index, which the authors call Contextual Diversity (CD). This index has been repeatedly shown to predict language acquisition phenomena, such as the age of acquisition of words in different syntactic categories (Hills et al., 2010; Stella et al., 2017) and individual differences between typically developing children and late talkers (Beckage et al., 2011).

We propose a variant evaluation method that takes *token* co-occurrences into account. Because of the binarization of the co-occurrence matrix, the CD index is an indicator of *type* co-occurrences, and is therefore agnostic to co-occurrence frequency. The models we work with, on the contrary, are sensitive to co-occurrence frequencies, providing a more fine-grained characterization of the semantic space.

Our method works as follows. First, we derived the semantic networks based on the cosine distance between representations. This required us to set a minimum cosine similarity threshold θ to determine if two words are connected, which we treat as a hyperparameter (with values [.6, .7, .8, .9]). Second, given this network, we counted the number of neighbours of each word as the number of other words connected to it. We refer to this index as neighbourhood density (ND). Third, we computed the Pearson's r correlation between this index and the AoA norms.

Figure 1 shows the distribution of the computed metric. Note that these correlations cannot be expected to be of the same order as those found when evaluating against adult ratings, since age of acquisition is predicted by a variety of factors, of which distributional information is only one, and it is subject to greater individual differences than adult semantic knowledge. Therefore, moderate but significant correlations are generally consid-

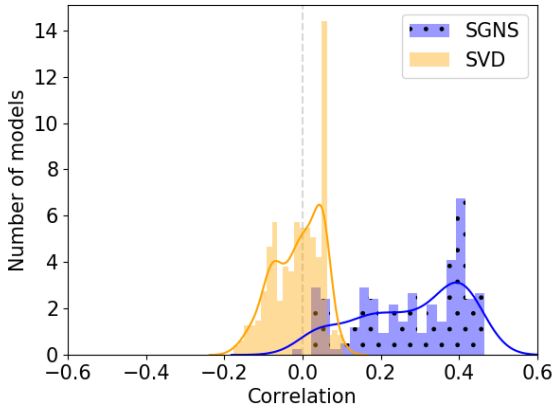

Figure 1: Histogram of Pearson's r correlations between ND and AoA, for SGNS and SVD models.

ered meaningful. As a reference, the CD index, has a correlation of $r = 0.32$ in our dataset [3].

As can be seen, the SGNS model is more likely to provide a semantic space that correlates with AoA, and some configurations yield an effect size comparable (even larger) than the CD metric. This indicates that the SGNS model builds word representations in a way that reflects the relative difficulty of each word, and thus offers a good starting point for understanding how children use distributional context for vocabulary acquisition. The fact that the correlation is positive prompts the prediction that, when co-occurrence frequency is incorporated in the model, words inhabiting less dense neighbourhoods are acquired earlier. This finding suggests that semantic neighbours may act as competitors in the process of word learning.

Among the hyperparameters of these models, one that is particularly relevant to language acquisition is the window size, as this reveals the amount of context that children most likely attend to in the analyzed ages. To investigate this, we took the best model of our previous analyses (SGNS with window size 1, negative sampling 15, frequency threshold 10), and varied only the window size. Results are in figure 2. As can be seen, smaller window sizes have better correlation with the data, indicating that the exploited context at this age is very local. Such a result makes intuitive sense in the context of children's immature verbal memory spans, which only improve as they acquire more language.

[3]We replicated the original analyses, since we use an extended dataset (both in the case of CHILDES and the AoA norms).

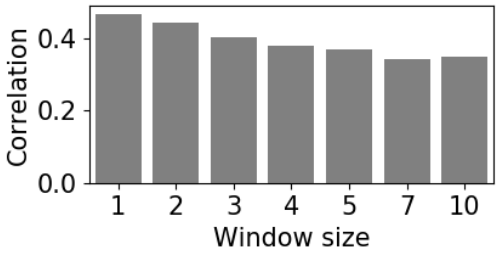

Figure 2: Pearson's r correlation depending on window size, for the best-performing SGNS model.

5 Method 2: Word Associates

Our first evaluation method above focused on the structure of the semantic spaces provided by the learnt word embeddings. Now we turn our attention to the specific lexical items and their position in the semantic space.

Children tend to under- and overextend word meaning in the first stages of acquisition, and over time they become more precise on capturing the semantics of words. A logical assumption then, is that words learnt earlier also converge earlier to adult-like semantic representations (assuming that early and late words take, on average, approximately the same amount of time to converge). We incorporated this idea in our second method by relating the AoA of words with adult free word association norms. Note that this method can be applied to other semantic tasks, but we focus on word association because it does not impose the specific type of semantic relation that words need to have (i.e. there is no distinction between similarity, analogy or others).

The dataset of free word association that we used is known as Small World of Worlds (SWOW, De Deyne et al., 2019), and it is the largest dataset of word associations in English, containing responses to over 12,000 cue words. We filtered the preprocessed version of the dataset to include only words that have been acquired before 60 months old. This results in 613 cue words, and 1839 responses (word associates) to these cues.

We then performed a similar cue-response experiment, with the best model from the previous section: for each cue, we retrieved the closest n neighbours. As in Pereira et al. (2016), we used $n = 50$, and then computed how many of these neighbours overlap with the word associates (responses) provided by human adults. However, unlike that work, our evaluation is not based directly

on the number of overlaps. Instead, we computed the Spearman rank correlation between the number of overlaps and the AoA norms, in order to quantify whether word embeddings corresponding to words learned earlier by children are also those that are converging faster to adult semantic knowledge. Figure 3 shows the result of this procedure. As can be seen, there is a statistically significant rank correlation ($\rho = -0.378$, $p < 0.001$). The negative direction confirms that words acquired earlier have a network of word associates that is more similar to those of adults, suggesting that convergence to adult semantic knowledge is at a more advanced state.

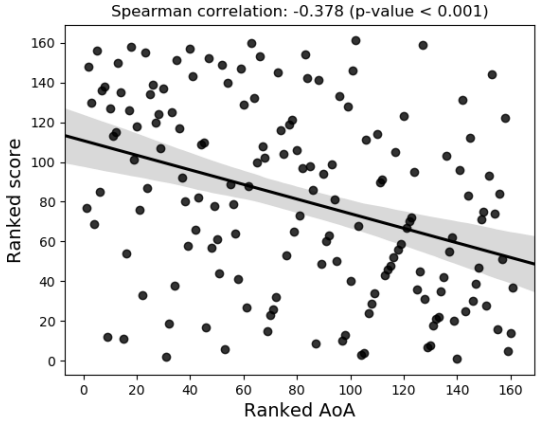

Figure 3: Ranked AoA and ranked score (number of overlaps), based on the 50 nearest neighbours in the best-performing model in the ND method.

One limitation of this procedure is that it requires a choice on the number of neighbours to be retrieved. In order to see how much the metric is affected by this parameter, we report the rank correlations of the previous model for several values of n. As can be seen in Figure 4, this number stabilizes after $n = 25$. The figure also shows whether this metric favours a model that did not perform well in our previous evaluation metric (SVD with window size 4, shift 15, frequency threshold 10). The graph shows that this model is consistently worse on our second evaluation method as well.

6 Conclusion

We proposed two methods to evaluate word embeddings for language acquisition. The main feature of these methods is the use of AoA norms for assessing whether the semantic organization of the word embeddings support the developmental trajectory of word learning.

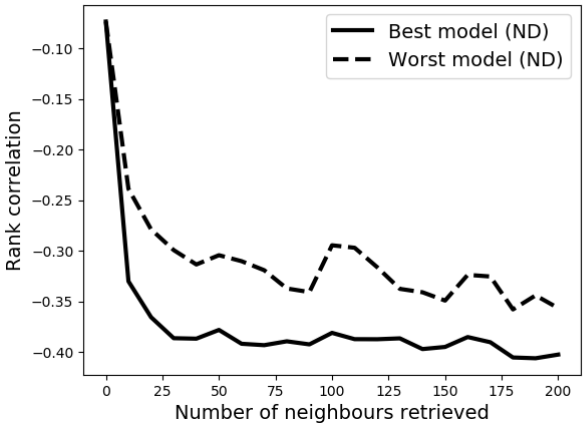

Figure 4: Rank correlations depending on the number of retrieved nearest neighbours, for the best and worst models in the previous evaluation method (ND).

The use of these metrics already prompted the discovery that (1) words with fewer neighbours are easier to acquire, suggesting competition of neighbouring words, and (2) at young age, infants only attend to very local context. The application of these methods to distributional models that incorporate additional assumptions (e.g. knowledge of word order) holds promise for further understanding of the role of distributional information in word learning.

References

Nicole Beckage, Linda Smith, and Thomas Hills. 2011. Small worlds and semantic network growth in typical and late talkers. *PloS One*, 6(5).

Simon De Deyne, Danielle J Navarro, Amy Perfors, Marc Brysbaert, and Gert Storms. 2019. The "small world of words" english word association norms for over 12,000 cue words. *Behavior research methods*, 51(3):987–1006.

Michael C Frank, Mika Braginsky, Daniel Yurovsky, and Virginia A Marchman. 2017. Wordbank: An open repository for developmental vocabulary data. *Journal of child language*, 44(3):677–694.

Thomas T Hills, Josita Maouene, Brian Riordan, and Linda B Smith. 2010. The associative structure of language: Contextual diversity in early word learning. *Journal of memory and language*, 63(3):259–273.

Omer Levy and Yoav Goldberg. 2014. Neural word embedding as implicit matrix factorization. In *Advances in neural information processing systems*, pages 2177–2185.

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

Brian MacWhinney. 2000. *The CHILDES Project: Transcription format and programs*. Lawrence Erlbaum Associates.

Tomas Mikolov, Ilya Sutskever, Kai Chen, Greg S Corrado, and Jeff Dean. 2013. Distributed representations of words and phrases and their compositionality. In *Advances in Neural Information processing systems*, pages 3111–3119.

Francisco Pereira, Samuel Gershman, Samuel Ritter, and Matthew Botvinick. 2016. A comparative evaluation of off-the-shelf distributed semantic representations for modelling behavioural data. *Cognitive neuropsychology*, 33(3-4):175–190.

Jenny R Saffran, Richard N Aslin, and Elissa L Newport. 1996. Statistical learning by 8-month-old infants. *Science*, 274(5294):1926–1928.

Alessandro Sanchez, Stephan C Meylan, Mika Braginsky, Kyle E MacDonald, Daniel Yurovsky, and Michael C Frank. 2019. childes-db: A flexible and reproducible interface to the child language data exchange system. *Behavior research methods*, 51(4):1928–1941.

Massimo Stella, Nicole M Beckage, and Markus Brede. 2017. Multiplex lexical networks reveal patterns in early word acquisition in children. *Scientific reports*, 7:46730.

Michael Tomasello, Sara Mannle, and Ann C Kruger. 1986. Linguistic environment of 1-to 2-year-old twins. *Developmental Psychology*, 22(2):169.