Learning word representations in a developmentally realistic order

Chad DeChant
Computer Science Department
Columbia University
chad.dechant@columbia.edu

Daniel Bauer
Computer Science Department
Columbia University

Abstract

We suggest that learning word representations in an order similar to that which children learn may affect properties of those representations. We propose the creation of synthetic datasets to study this question and introduce a method and tool to reorder existing datasets so that new words appear over time in an order that matches existing data on natural word order acquisition in humans. We show that word representations learned in a more natural order differ in some respects from those learned in the usual all at once, shuffled fashion, including showing greater stability. We also see that how much word representations change during training depends on when they are introduced.

1 Introduction

Children learn the words of a language slowly and over an extended period of time. At their fastest, from ages 10 - 12 years, they may learn an average of twelve new words a day (Anglin et al., 1993). The words they learn at any given time depend on their age, their environment, and what words they already know (Bloom, 2002). This is in stark contrast to the way that natural language processing techniques produce word representations, which typically make use of very large corpora presented without regard to order. We suggest that it is worth investigating whether learning word representations in a more natural order would affect the usefulness or properties of such representations.

The age at which a word is acquired by a person can have significant and lasting effects on its use later in life. Words learned at younger ages are recognized more quickly than those learned at older ages (Juhasz, 2005) and have been found to have more semantic associations in a free association task (Brysbaert et al., 2000). Steyvers and Tenenbaum (2005) theorized that words learned early have a special place in shaping the network of semantic associations that gradually develops over time with the addition of new words.

A significant obstacle to any attempt to develop word representations respecting a more natural ordering is the difficulty of finding enough quality and diverse data to train on. Though there are some relevant corpora, including datasets of dialogue interactions involving children (MacWhinney, 2000) (Sullivan et al., 2020) and children's literature (Hill et al., 2015a), larger and more diverse datasets are needed. Given the difficulty and cost of collecting such data, we create a tool to modify existing datasets to mimic the order of word learning in children. All that is needed is a text corpus and an ordering of words. Because estimates of typical age of acquisition of words exist for dozens of languages, this tool can be used widely.

We illustrate its use by reordering a dataset of English language books. We see that the representations learned on the reordered data perform similarly on standard benchmarks to representations learned on the same text presented without reordering. However, we also show that the semantic spaces induced by our method as evident in the relationships of word representations to each other differ from the unordered approach. Notably, we find that representations learned in order are more stable.

2 Technique

The creation of an ordered dataset requires an order and a dataset. Kuperman et al. (2012) provide estimates for age of acquisition for approximately thirty thousand English words. They define age of acquisition to be the earliest age a word is understood, not necessarily produced. To these we add a supplementary list of inflections of those words (e.g. plural, past tense, etc.) which we rank immediately following the forms of the associated words

found in the original age of acquisition estimate list (e.g. 'dogs' follows 'dog').

We use text from the Children's Book Test (Hill et al., 2015a) dataset (stripped of duplications) and a cleaned subset of the Project Gutenberg books corpus (Lahiri, 2014) as the basis for our dataset. For each word in this combined dataset we assign an ordering from the augmented ordering list mentioned above. Each sentence is then assigned the highest estimated age of acquisition of all the words contained in it. Sentences with words not found in our augmented order list of words are discarded except when those unknown words are classified as named entities, in which case the presence of the named entities has no effect on a sentence's ordering. We ignore named entities when training word representations. Sentences are then ordered from youngest to oldest estimated ages and broken into tranches for in order training.

Training skip gram word vectors (Mikolov et al., 2013) begins with a tranche of sentences containing only the 500 earliest learned words. After one pass through these sentences, the word vectors are saved for later comparison and then used in continuing to train on a new tranche of sentences containing the next 500 possible words (not all words or inflections of words are found in our dataset). Note that these new sentences may also contain any of the previously added words. Within a tranche sentences are reordered randomly so that each of the five runs which goes into our averaged reported results has a different sentence order within a particular tranche as well as different initialization values for each word vector.

For comparison we also train skip gram models more traditionally, using the same data but fed to the model in a randomly shuffled order. To show relative training dynamics of the two approaches we show in the charts that follow how word vectors at various stages of training perform when each approach has seen the same number of training examples.

We use gensim's (Rehurek and Sojka, 2011) implementation of the skip gram model with a context window size of five, a negative to positive sample ratio of 20, only training on words that appear at least 20 times in the dataset. We use spacy (Honnibal et al., 2020) for splitting the dataset into sentences, tokenizing, and named entity recognition. The learning rate is set at 0.025 and is not changed during training. We conduct experiments using word vectors of lengths 50 and 300. Each sentence in the dataset is seen only once.

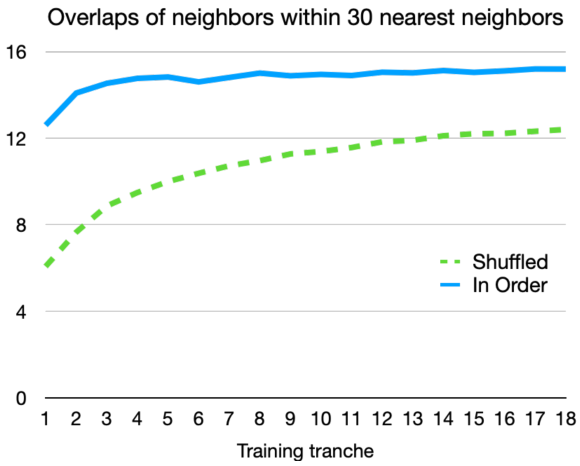

Figure 1: Average overlap of word vectors' thirty nearest neighbors between five runs of the skip gram model, both in order (solid line) and in random order (dashed line). Representations trained in a natural order are more stable at all phases of training, particularly toward the beginning.

We will release the ordering tool on github, where we will also provide instructions and maintain a list of available ordering datasets.[1]

3 Results and discussion

We compare the word vectors created in the traditional, shuffled fashion with those created by presenting the training data in an order approximating the order in which children learn the words.

First we consider the stability of models trained traditionally as opposed to in a more realistic order. Wendlandt et al. (2018) observed that word vector representations, particularly using a skip gram model, are unstable in the sense that different training runs will result in word representations whose nearest neighbors in vector space are quite different at the end of each run. Figure 1 shows the stability of vectors trained in order versus in random order for word vectors of length 50. Vectors from the in order training are consistently more stable throughout training than vectors from the randomly ordered training, ending with an average of 23% more overlaps in a word's 30 nearest neighbors in the final training period. When using vectors of length 300 the final overlap for vectors trained in order is also higher than those trained in random order, but less so, finishing just 5% higher in last training period.

[1] Link to github to be provided.

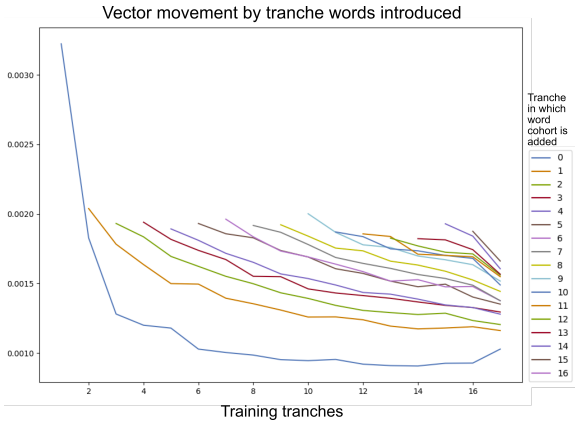

Figure 2: Movement in cosine distance of cohorts of word vectors introduced at each tranche is shown over the course of training. Cosine distance movements are normalized for word frequency in each tranche and averaged over five runs.

Second, we consider how much word vectors change over time after being introduced in a training tranche. Figure 2 shows how much each word vector cohort (i.e. the set of word vectors introduced in a particular tranche) changes in subsequent training steps following their introductory step. We plot these changes in terms of mean cosine distance of a vector from its location at the previous stage, averaged across all vectors introduced in the same cohort. Because how much a word vector is changed in a given training tranche will be affected by how often it appears, we normalize the distances shown by dividing by a word's frequency in the tranche. We find two consistent patterns. After the first word vector cohort, which has large early moves, each subsequent word vector cohort has very similar magnitude moves in the periods after their introduction. In any given training tranche, those moves are also consistently lower for words introduced in earlier cohorts; generally, the earlier a word is introduced, the less it moves in later tranches relative to how often it is seen in those later tranches.

Third, we look at the relationships between words which appear close together in vector space; how do nearby words relate to each other and how do those relationships change over time during training? We compare vectors trained in order to others trained with randomly ordered data. Figure 3 shows to what degree three relationships which can be derived from Wordnet (Miller, 1995) appear in a word's 30 nearest neighbors; the words whose neighbors are examined are those present in the simlex-999 task. The three relationships examined are: synonymy (whether two words' lemmas share a sense); hyponymy (whether a word is special case of another); and hypernomy (the inverse relationship of hyponymy). It is apparent that the words' relationships to their nearest neighbors when tracked with these three metrics do differ depending on whether the models were trained in order or in random order, though to differing degrees. Word vectors trained in a shuffled fashion have more synonyms, hypernyms, and hyponyms among their 30 nearest neighbors than word vectors trained in order. The same difference held when examining the nearest ten or 100 nearest neighbors. Whether these differences are to the advantage of one method or the other is unclear. We also examined meronyms and holonyms (the two directions of the part-whole relation) and found little to no difference between the two approaches. Characteristics of word vectors of length 300 are shown in the figure; those of length 50 showed somewhat less difference between the two training orders.

Finally, we look at how word vectors from the two approaches to training perform according to two common semantic evaluation metrics, simlex-999 (Hill et al., 2015b) and wordsim353 (Finkelstein et al., 2001). At each time step we exclude words which have not yet been seen by a model from both of these tasks. Figure 4 shows that the performance of the word vectors on these two tasks is not affected by whether the model learns the words in order or in random order. Results for vectors of length 300 are shown; results for vectors of length 50 similarly showed effectively no difference between the two training orders.

We performed the above experiments using Continuous Bag of Words (CBOW) vectors (Mikolov et al., 2013) as well and found similar results.

4 Related work

Our suggested approach is a type of curriculum learning, which has been studied in a variety of forms but not, to our knowledge, to suggest a human-like ordering of the curriculum when learning distributed word representations. More commonly a curriculum is designed around frequency statistics of words in a corpus. (Bengio et al., 2009) Curricula have also been learned. (Tsvetkov et al., 2016)

Prior work has examined what representations language models learn at various stages of train-

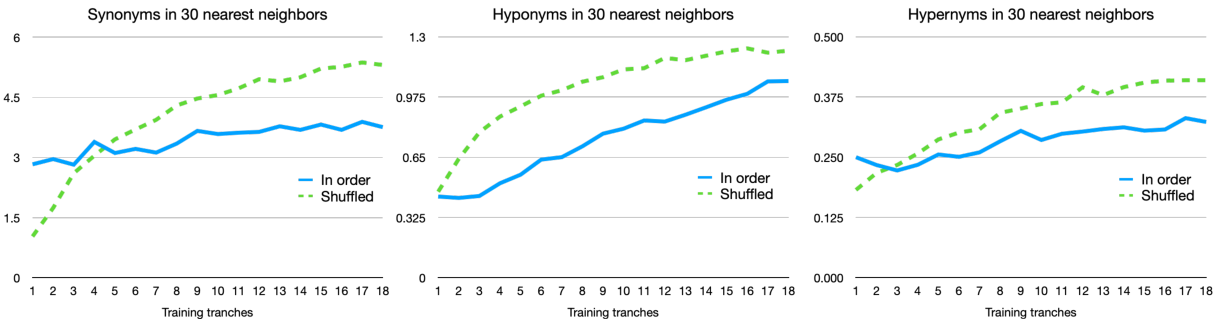

Figure 3: Wordnet relations present among the thirty nearest neighbors of words in simlex-999 as word vectors are trained in order (solid blue lines) and in a randomly shuffled order (dashed green lines). Mean values from five runs shown.

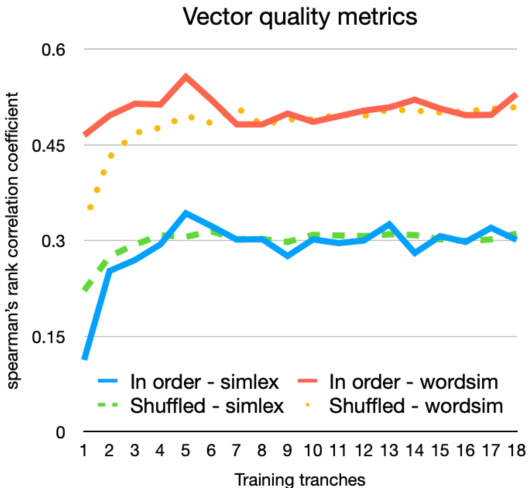

Figure 4: Performance of word vectors trained in order (solid lines) and in a randomly shuffled order (dashed lines) shown for two common assessment metrics of word vector quality, wordsim353 (top two lines) and simlex-999 (bottom two lines).

ing, though with randomly ordered training data. For example, Zhang et al. (2020) and Liu et al. (2021) study when during various amounts of training linguistic knowledge can be detected in large contextual language models.

Some work has been done to model the effects of word acquisition order, though not to develop word embeddings. Smith et al. (2001) and Sohrabi (2019) found age of acquisition effects in the behavior of a simple computational model of the mapping of a word to its phonemes, as did Chang et al. (2019), when mapping the relations between words, phonemes, and sets of semantic features.

Datasets offering age of acquisition estimates exist for many languages. For example, Frank et al. (2017) provide estimates for early acquired words in 29 languages while Xu et al. (2020) and Brysbaert et al. (2014) have estimates for large numbers of words in Mandarin and Dutch, respectively.

An increasing amount of recent work stresses the importance of embodiment in language learning and develops or suggests ways to bring that embodiment into NLP approaches to understanding language. (McClelland et al., 2019) (Hill et al., 2020) This is an acknowledgment that as language is a human activity, it is necessary to take into account more aspects of human experience when trying to understand or mimic the language faculty. We suggest that the fact that the language learning experience is extended over time is significant and should not be ignored.

5 Conclusion

We have demonstrated the use of a tool to modify existing datasets to reorder them so that the appearance of new words reflects the order that children learn words. We saw that the vectors' relationships to their neighbors differed somewhat when learned in this fashion. Most notably we found that the stability of word representations improved. It is unclear why some aspects of the word representations changed while others did not when training in order, suggesting the need for further study using synthetically ordered datasets.

Further work needs to be done to determine how this ordering may affect other word representations, including contextual representations. It is possible, even likely, that the skip gram model is not sensitive enough to fully capture the implications of training order.

Training on a much larger dataset, extension to other languages, testing this method using other metrics such as analogy tasks, and taking into account the order and pace at which children acquire syntax are all natural extensions of this work.

References

Jeremy M Anglin, George A Miller, and Pamela C Wakefield. 1993. Vocabulary development: A morphological analysis. *Monographs of the society for research in child development*, pages i–186.

Yoshua Bengio, Jérôme Louradour, Ronan Collobert, and Jason Weston. 2009. Curriculum learning. In *Proceedings of the 26th annual international conference on machine learning*, pages 41–48.

Paul Bloom. 2002. *How children learn the meanings of words*. MIT press.

Marc Brysbaert, Michaël Stevens, Simon De Deyne, Wouter Voorspoels, and Gert Storms. 2014. Norms of age of acquisition and concreteness for 30,000 dutch words. *Acta psychologica*, 150:80–84.

Marc Brysbaert, Ilse Van Wijnendaele, and Simon De Deyne. 2000. Age-of-acquisition effects in semantic processing tasks. *Acta Psychologica*, 104(2):215–226.

Ya-Ning Chang, Padraic Monaghan, and Stephen Welbourne. 2019. A computational model of reading across development: Effects of literacy onset on language processing. *Journal of Memory and Language*, 108:104025.

Lev Finkelstein, Evgeniy Gabrilovich, Yossi Matias, Ehud Rivlin, Zach Solan, Gadi Wolfman, and Eytan Ruppin. 2001. Placing search in context: The concept revisited. In *Proceedings of the 10th international conference on World Wide Web*, pages 406–414.

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.

Felix Hill, Antoine Bordes, Sumit Chopra, and Jason Weston. 2015a. The goldilocks principle: Reading children's books with explicit memory representations. *arXiv preprint arXiv:1511.02301*.

Felix Hill, Roi Reichart, and Anna Korhonen. 2015b. Simlex-999: Evaluating semantic models with (genuine) similarity estimation. *Computational Linguistics*, 41(4):665–695.

Felix Hill, Olivier Tieleman, Tamara von Glehn, Nathaniel Wong, Hamza Merzic, and Stephen Clark. 2020. Grounded language learning fast and slow. *arXiv preprint arXiv:2009.01719*.

Matthew Honnibal, Ines Montani, Sofie Van Landeghem, and Adriane Boyd. 2020. spaCy: Industrial-strength Natural Language Processing in Python.

Barbara J Juhasz. 2005. Age-of-acquisition effects in word and picture identification. *Psychological bulletin*, 131(5):684.

Victor Kuperman, Hans Stadthagen-Gonzalez, and Marc Brysbaert. 2012. Age-of-acquisition ratings for 30,000 english words. *Behavior research methods*, 44(4):978–990.

Shibamouli Lahiri. 2014. Complexity of Word Collocation Networks: A Preliminary Structural Analysis. In *Proceedings of the Student Research Workshop at the 14th Conference of the European Chapter of the Association for Computational Linguistics*, pages 96–105, Gothenburg, Sweden. Association for Computational Linguistics.

Leo Z Liu, Yizhong Wang, Jungo Kasai, Hannaneh Hajishirzi, and Noah A Smith. 2021. Probing across time: What does roberta know and when? *arXiv preprint arXiv:2104.07885*.

Brian MacWhinney. 2000. *The CHILDES project: The database*, volume 2. Psychology Press.

James L McClelland, Felix Hill, Maja Rudolph, Jason Baldridge, and Hinrich Schütze. 2019. Extending machine language models toward human-level language understanding. *arXiv preprint arXiv:1912.05877*.

Tomas Mikolov, Ilya Sutskever, Kai Chen, Greg Corrado, and Jeffrey Dean. 2013. Distributed representations of words and phrases and their compositionality. *arXiv preprint arXiv:1310.4546*.

George A Miller. 1995. Wordnet: a lexical database for english. *Communications of the ACM*, 38(11):39–41.

Radim Rehurek and Petr Sojka. 2011. Gensim–python framework for vector space modelling. *NLP Centre, Faculty of Informatics, Masaryk University, Brno, Czech Republic*, 3(2).

Mark A Smith, Garrison W Cottrell, and Karen L Anderson. 2001. The early word catches the weights. *Advances in neural information processing systems*, pages 52–58.

Ahmad Sohrabi. 2019. Age of acquisition effect: Evidence from single-word reading and neural networks. *Basic and clinical neuroscience*, 10(2):137.

Mark Steyvers and Joshua B Tenenbaum. 2005. The large-scale structure of semantic networks: Statistical analyses and a model of semantic growth. *Cognitive science*, 29(1):41–78.

Jess Sullivan, Michelle Mei, Andrew Perfors, Erica H Wojcik, and Michael C Frank. 2020. Saycam: A large, longitudinal audiovisual dataset recorded from the infant's perspective.

Yulia Tsvetkov, Manaal Faruqui, Wang Ling, Brian MacWhinney, and Chris Dyer. 2016. Learning the curriculum with bayesian optimization for task-specific word representation learning. *arXiv preprint arXiv:1605.03852*.

Laura Wendlandt, Jonathan K Kummerfeld, and Rada Mihalcea. 2018. Factors influencing the surprising instability of word embeddings. *arXiv preprint arXiv:1804.09692*.

Xu Xu, Jiayin Li, and Shulun Guo. 2020. Age of acquisition ratings for 19,716 simplified chinese words. *Behavior Research Methods*, pages 1–16.

Yian Zhang, Alex Warstadt, Haau-Sing Li, and Samuel R Bowman. 2020. When do you need billions of words of pretraining data? *arXiv preprint arXiv:2011.04946*.