Learning Phone Embeddings for Word Segmentation of Child-Directed Speech

Anonymous ACL submission

Abstract

This paper presents a novel model that learns and exploits embeddings of phone ngrams for word segmentation in child language acquisition. Embedding-based models are evaluated on a phonemically transcribed corpus of child-directed speech, in comparison with their symbolic counterparts using the common learning framework and features. Results show that learning embeddings significantly improves performance. A detailed analysis of the learned embeddings show that they are informative for both word segmentation and phonology in general.

1 Introduction

Segmentation is a prevalent problem in language processing. Both humans and computers process language as a combination of linguistic units, such as words. However, spoken language does not include reliable cues to word boundaries that are found in many writing systems. The hearers need to extract words from a continuous stream of sounds using their linguistic knowledge and the cues in the input signal. Although the problem is still non-trivial, competent language users utilize their knowledge of the input language, e.g., the (mental) lexicon, to a large extent to aid extraction of lexical units from the input stream.

Word segmentation in early language acquisition is especially interesting and challenging, as early language learners barely have a lexicon or any other linguistic knowledge to start with. Consequently, it has been studied extensively through psycholinguistic experiments (Cutler and Butterfield, 1992; Jusczyk et al., 1999; Jusczyk et al., 1993; Saffran et al., 1996; Jusczyk et al., 1999; Suomi et al., 1997; van Kampen et al., 2008) and computational modeling (Cairns et al., 1994; Christiansen et al., 1998; Brent and Cartwright, 1996; Brent, 1999; Venkataraman, 2001; Xanthos, 2004; Goldwater et al., 2009; Johnson and Goldwater, 2009).

The majority of the state-of-the-art computational models use symbolic representations for input units. Due to Zipf's law, most linguistic units, however, are rare and thus the input provides little evidence for their properties that are useful for solving the task at hand. In machine learning terms, the learner has to deal with the data sparseness problem due to the rare units whose parameters cannot be estimated reliably. A model using distributed representations can counteract the data sparseness problem by exploiting the similarities between the units for parameter estimation. This has motivated the introduction of *embeddings* (Bengio et al., 2003; Collobert et al., 2011), a family of low-dimensional, real-valued vector representation of features that are learned from data. Unlike purely symbolic representations, such distributed representations allow input units that appear in similar contexts to share similar vectors (embeddings). The model can, then, exploit the similarities between the embeddings during segmentation and learning.

This paper studies the learning and use of embeddings of phone[1] uni- and bi-grams for computational models of word segmentation in child language acquisition. Our work is inspired by recent success of embeddings in NLP (Devlin et al., 2014; Socher et al., 2013), especially in Chinese word segmentation (Zheng et al., 2013; Pei et al., 2014). However, this work differs from Chinese word segmentation models in two aspects. (1) The model (Section 2) learns from a phonemically transcribed corpus of child-directed speech

[1]We use the therm *phone* as a theory-neutral term for the distinct (phonetic) segments in the input.

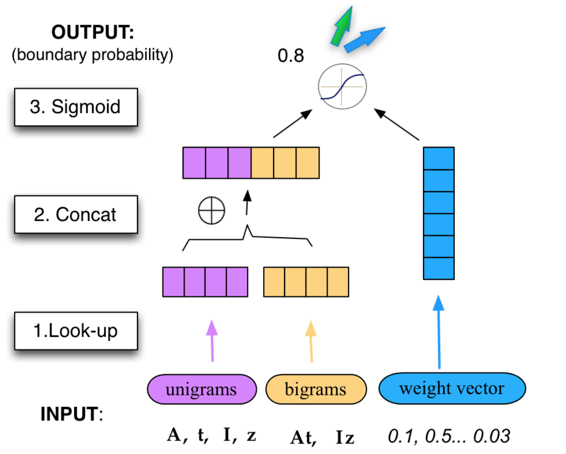
(Section 3.1) instead of large written text input. (2) The learning (Section 2.2) only relies on utterance boundaries in input as opposed to explicitly marked word boundaries. Although the number of phone types is small, higher level ngrams of phones inevitably increase the severity of data sparseness. Thus we expect embeddings to be particularly useful when larger phoneme ngrams are used as input units.

The contributions of this paper are three-fold: (1) A novel model that constructs and uses embeddings of phone ngrams for word segmentation in child language acquisition; (2) Empirical evaluations of symbolic and embedding representations for this task on the benchmark data, which suggest that learning embeddings boosts the performance; and (3) A deeper analysis of the learned embeddings through visualizations and clustering, showing that the learned embeddings capture information relevant to segmentation and phonology in general.

In the next section we define the distributed representations we use in this study, *phone-embeddings*, and a method for learning the embeddings and the segmentation parameters simultaneously from a corpus without word boundaries. Then we present a set of experiments for comparing embedding and symbolic representations (Section 3). We show our visualization and clustering analyses of the learned embeddings (Section 4) before discussing our results further in the context of previous work (Section 5) and concluding the paper.

2 Learning segmentation with phone embeddings

2.1 The architecture of the model



The position between t and I in "WAtIzIt" is being predicted.

Figure 1: Architecture of our model.

Figure 1 shows the architecture of the proposed embedding-based model. Our model takes the embeddings of phone uni- and bi-grams in the local window for each position in an utterance, and predicts whether that position is a word boundary. The embeddings for the phone ngrams are learned *jointly* with the segmentation model. The model has the following three components:

Look-up table maps phone ngrams to their corresponding embeddings. In this study, for each position j, we consider the 4 unigrams ($c_{j-1}, c_j, c_{j+1}, c_{j+2}$) and 2 bigrams ($c_{j-1}c_j$ and $c_{j+1}c_{j+2}$) that are in a window of 4 phones of positions j. The phone c_{j-1} represents the one to the left of the current position and so on.

Concatenation. To predict the segmentation for position j, the embeddings of the phone uni- and bi-gram features are *concatenated* into a single vector, *input embedding*, $\mathbf{i}_j \in \mathbb{R}^{NK}$, where $K = 6$ is the number of uni- and bi-gram used and $N = 50$ is the dimension of the embedding of each ngram.

Sigmoid function. The model then computes the sigmoid function (1) of the dot product of the input embedding $\mathbf{i}_j$ and the weight vector $\mathbf{w}$. The output is a score $\in [0, 1]$ that denotes the probability that the current position being a word boundary, which we call *boundary probability*.

$$f(j) = \frac{1}{1 + exp\left(-\mathbf{w} \cdot \mathbf{i}_j\right)} \qquad (1)$$

2.2 Learning with utterance edge and random sampling

Our model learns from utterances that have word boundaries removed. It, however, utilizes the *utterance boundaries* as positive instances of word boundaries. Specifically, the position before (after) the first (last) phone of a utterance is the left (right) boundary of the first (last) word in the utterance. For these positions, dummy symbols are used as the two leftmost (rightmost) phones. Moreover, one position within the utterance is randomly sampled as negative instance. Although such randomly sampled instances are not guaranteed to actually be negative ones, sampling balances the positive instances, which makes learning possible.

The training follows an online learning strategy, processing one utterance at a time and updating the parameters after processing each utterance. The trainable parameters are the weight vector and the embeddings of the uni- and bi-grams. For each position j, the boundary probability is computed with the current parameters. Then the parameters are updated by minimizing the *cross-entropy loss function* as in (2).

$$J_j = -\left[y_j \log f(j) + (1 - y_j) \log (1 - f(j))\right] \quad (2)$$

In formula (2), $f(j)$ is the boundary probability estimated in (1) and y_j is its presumed value, which is 1 and 0 for utterance boundaries and sampled intra-utterance positions, respectively. To offset over-fitting, we add an L2 regularization term ($||\mathbf{i}_j||^2 + ||\mathbf{w}||^2$) to the loss function, as follows:

$$J_j = J_j + \frac{\lambda}{2}\left(||\mathbf{i}_j||^2 + ||\mathbf{w}||^2\right) \quad (3)$$

The λ is a factor that adjusts the contribution of the regularization term. To minimize the regularized loss function, which is is still convex, we perform *stochastic gradient descent* to iternatively update the embeddings and the weight vector in turn, each time considering the other as constant. The gradients and update rules are similar to that of logistic regression model as in Tsuruoka et al. (2009), except that the input embeddings $\mathbf{i}$ are also updated besides the standard weight vector.

In particular, the gradient of input embeddings $\mathbf{i}_j$ for each particular position j is computed according to (4), where $\mathbf{w}$ is the weight vector and y_j is the assumed label. The input embeddings are then updated by (5), where α is the learning rate.

$$\frac{\partial J_j}{\partial \mathbf{i}_j} = \left(f\left(j\right) - y_j\right) \cdot \mathbf{w} + \lambda \mathbf{i}_j \quad (4)$$

$$\mathbf{i}_j = \mathbf{i}_j - \alpha \frac{\partial J_j}{\partial \mathbf{i}_j} \quad (5)$$

2.3 Segmentation via greedy search

The word segmentation of utterances is a greedy search procedure using the learned model. It irreversibly predicts segmentation for each position j ($1 \leq j \leq N =$ utterance length), one at a time, in a left-to-right manner. If the boundary probability given by the model > 0.5 (or ≤ 0.5), the current position is predicted as word boundary (or non-boundary). The segmented word sequence is built from the predicted word boundaries in the utterance.

3 Experiments and the results

The learning framework described in Section 2 can also be adopted for symbolic representations where the ngram features for each position are represented by a sparse *binary vector*. In the symbolic representation, each distinct uni- or bi-gram is represented by a distinct dimension in the input vector. In that case, the learning framework is equivalent to a *logistic regression* model, the training of which only updates the weight vector but not the feature representations. In this section, we run experiments to compare the performances of embedding- and symbolic-based models using the same learning framework with the same features. Before presenting the experiments and the results, we describe the data and evaluation metrics.

3.1 Data

In the experiments reported in this paper, we use the *de facto* standard corpus for evaluating segmentation models. The corpus was collected by Bernstein Ratner (1987) and converted to a phonemic transcription by Brent and Cartwright (1996). The original corpus is part of the CHILDES database (MacWhinney and Snow, 1985). Following the convention in the literature the corpus will be called the *BR corpus*. Since our model does not know the locations of true boundaries, we do not make training and test set distinction, following previous literature.

3.2 Evaluation metrics

As a measure of success, we report F-score, the harmonic mean of *precision* and *recall*. F-score is a well-known evaluation metric originated in information retrieval (van Rijsbergen, 1979). The calculation of these measures depend on true positive (TP), false positive (FP) and false negative (FN) values for each decision. Following earlier studies, we report three varieties of F-scores. The *boundary* F-score (**BF**) considers individual boundary decisions. The *word* F-score (**WF**) quantifies the accuracy of recognizing word tokens. And the *lexicon* F-scores (**LF**) are calculated based on the gold-standard lexicon and lexicon learned by the model. For details of the metrics, see Goldwater et al. (2009). Besides these standard scores we also present over-segmentation (**EO**) and under-segmentation (**EU**) error rate (lower is better) defined as:

$$\mathrm{EO} = \frac{\mathrm{FP}}{\mathrm{FP}+\mathrm{TN}} \qquad \mathrm{EU} = \frac{\mathrm{FN}}{\mathrm{FN}+\mathrm{TP}}$$

where TN is true negatives of boundaries. Besides providing a different look at the models' behavior, it is straightforward to calculate the statistical uncertainty around them since they resemble N Bernoulli trials with a particular error rate, where N is number of boundary and word-internal positions for EU and EO respectively.

3.3 Experiments

To show the differences between the symbolic and embedding representations, we train both models on the BR corpus, and present the performance and error scores on the complete corpus. The training of all models use the linear decay scheme of learning rate with the initial value of 0.05 and the regularization factor is set to 0.001 throughout the experiments. Table 1 presents the results, including standard errors for EO and EU, for *emb*(edding)- and *sym*(bolic)-based models using unigram features (*uni*) and unigram+bigram features (*all*), respectively.

Model	EO	EU	BF	WF	LF
emb/all	**6.4±0.1**	**17.3±0.2**	**82.9**	**68.7**	**42.6**
sym/all	8.1±0.1	25.8±0.2	75.9	60.2	31.6
emb/uni	15.8±0.1	**10.6±0.3**	**77.4**	**59.1**	**40.7**
sym/uni	**13.2±0.1**	21.7±0.2	73.4	54.4	29.4

Table 1: Performance of embedding and symbolic models. Numbers in percentage.

Table 1 shows the average of the results obtained from 10 independent runs. For each run, we take the scores from the 10th iteration of the whole dataset, where the scores are stabilized. All models learn quickly and have good performance after the first iteration already. And the differences between the scores of subsequent iterations are rather small.

4 Visualization and Interpretation

The experiment results in the previous section show that learning embeddings jointly with a segmentation model, instead using symbolic representations, leads to a boost of segmentation performance. Nevertheless, it is not straightforward to interpret embeddings, as the "semantics" of each dimension is not pre-defined as in symbolic representations. In this section, we use visualization and clustering techniques to interpret the information captured by the embeddings.

Phone symbols in the BR corpus. We use the BR corpus for visualization as in the experiments. The transcription in the BR corpus use symbols that, unfortunately, can not be converted to International Phonetic Alphabet (IPA) in a context-free, deterministic way. Thus we keep them as they are and suggest readers who are unfamiliar with such symbols to refer to Appendix A.

4.1 Embeddings encode segmentation roles

Segmentation roles of phone ngrams. We first investigate the correspondence of the embeddings to the metrics that are indicative for segmentation decisions. For distinguishing word-boundary positions from word-internal positions as in segmentation models, it is helpful to know whether a particular phone unigram/bigram is more likely to occur at the beginning of a word (*word-initial*), at the end of a word (*word-final*), in the middle of a word (*word-medial*), or has a *balanced distribution* of above positions. For a phone bigram, it can also *corss word-boundary*. We call such tendencies of phone ngrams as *segmentation roles*.

We hypothesize that the embeddings that are learned by our model can capture segmentation roles: the embeddings of phone ngrams of the same segmentation role are similar to each other and are dissimilar to the phone ngrams of different segmentation roles. To test this, we use principal component analysis (PCA) to project the embeddings of phone uni- and bi-grams that are learned in our model into two-dimension space, where the resulting vectors preserve 85% and 98% of the variance in the original 50-dimension uni- and bigram embeddings, respectively. We then plot such PCA-projected 2-D vectors of the phone ngrams in Figure 2, where the geometric distances between data points reflect the (dis-)similarities between the original embeddings of phone ngrams. These data points are color coded to demonstrate the dominant segmentation role of each phone ngram. A phone ngram is categorized as *word-initial*, *word-medial*, *word-final* or *corss word-boundary* (only applicable for bigrams), if the ngram co-occur more than 50% of the time with the corresponding segmentation roles according to the gold standard segmentation. If none of the roles reaches

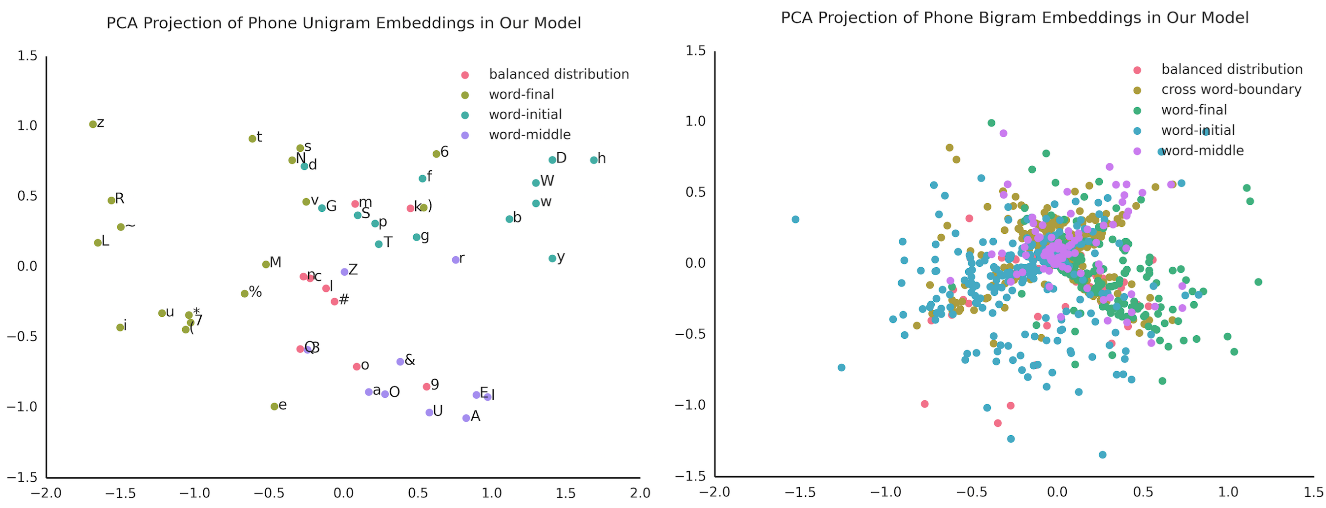


Figure 2: PCA Projections of the phone uni-gram (left) and bi-gram (right) embeddings learned in our model.

the majority, the ngram is categorized as *balanced distribution*. Note that segmentation roles are assigned using the true word boundaries, while the embeddings are learned only from utterance boundaries.

Figure 2 (left) shows that phone unigrams of the same category tend to cluster in the same neighborhood, while unigrams of distinct categories tend to locate apart from each other. This is consistent with our hypothesis on embeddings being capable of capturing segmentation roles. Figure 2 (right) shows that the distribution of phone bigrams is noisier, as many bigrams of different categories congest in the center. This suggests that bigram embeddings are less well estimated than unigrams ones, probably due to the larger number and lower relative frequencies of bigrams. Nevertheless, the *word-initial* v.s. *word-final* contrast in bigrams is still sharp, as a result of our training procedure that makes heavy use of the initial and final positions of utterances, which are also word boundaries. In summary, the information that are encoded in our phone ngram embeddings is highly indicative of correct segmentations.

4.2 Embeddings capture phonology

Hierarchical clustering of phones. Different from the previous subsection that correlates the learned embeddings with segmentation-specific metrics, we can alternatively explore the embeddings more freely to see what structures emerge from data. To this end, we apply *hierarchical agglomerative clustering* (Johnson, 1967) to the embeddings of phone unigrams to build up clusters in a bottom-up manner. Initially, each unigram embedding itself consists a cluster. Then at each step, the two most similar clusters are merged. The procedure iterates until every embedding is in the same cluster. The similarity between clusters are computed by the single linkage method, which outputs the highest score of all the pair-wise cosine similarities between the embeddings in the two clusters. Since the clustering procedure is based on pair-wise cosine similarities between embeddings, we first compute such similarity scores, composing the *similarity matrix*.

The *dendrogram* (Jones et al., 2001) that represents the clustering results is shown in Figure 3, together with the *heatmap* that represents the similarity matrix. The dendrogram draws a U-shaped link to indicate how a pair of children clusters form their parent cluster, where the dissimilarity between the two children clusters are shown by the height of the top of the U-link. The intensity of the color of each grid in the heatmap denotes the similarity between the two corresponding phone embeddings. Moreover, each lowest node, i.e. leaf, of the dendrogram is vertically aligned with the column of the heatmap that corresponds to the same phone, which is labeled using the BR-corpus symbols. Thus the dark blocks along the antidiagonal also indicate the salient clusters in which phone embeddings are similar to one another.

Phonological structure. The heatmap reveals several salient blocks, such as the one on the

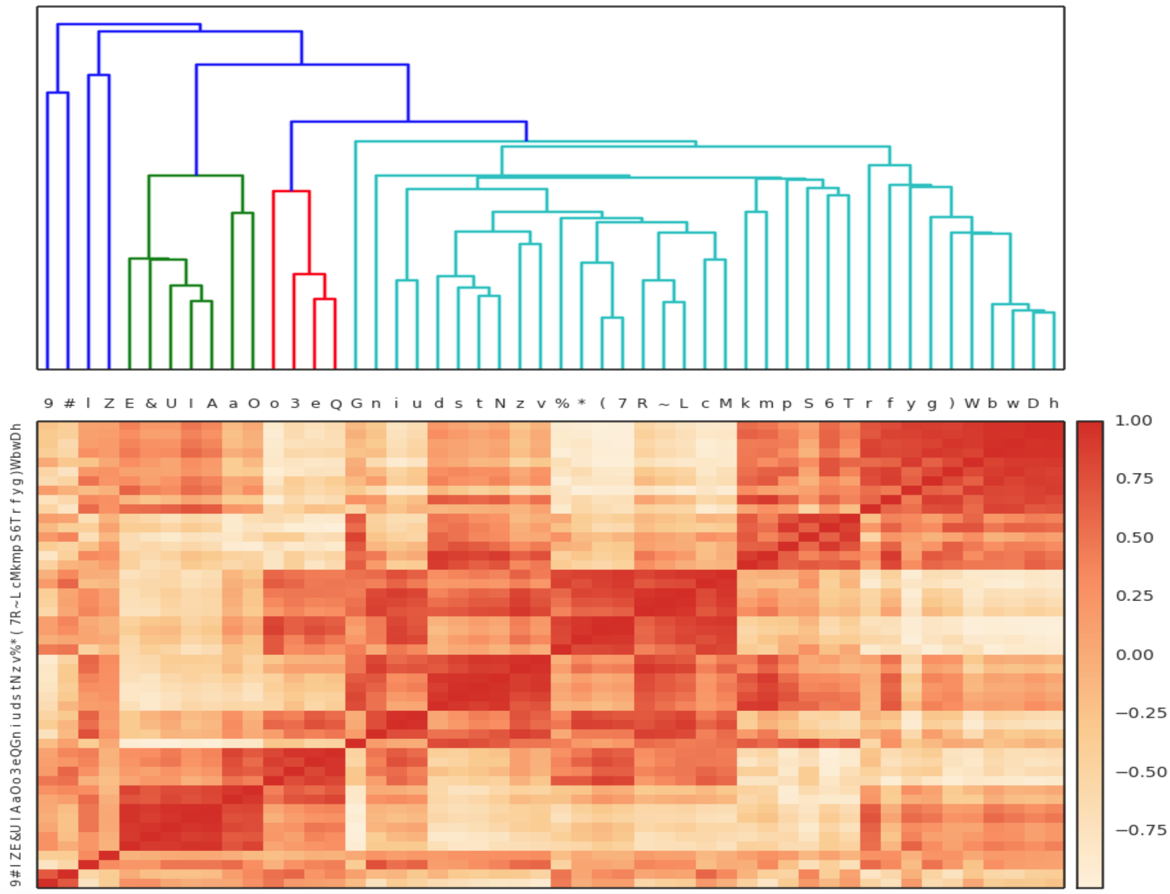


Figure 3: Hierarchical clustering and similarity matrix of phone embeddings learned by our model.

top-right corner and the one near the bottom-left corner. The former is part of a group of clusters spreading the whole right 2/3 of the dendrogram/heatmap, which mostly consists English consonants. In contrast, the latter contains short, unround vowels in English, *E*, *&*, *I* and *A*, as in b*e*t, th*a*t, b*i*t and b*u*t, respectively. It also contains the long-short vows pair *a* and *O* as in h*o*t and l*aw*. Immediately to the right of them are the cluster of compound vowels, *o*, *3*, *e*, *Q*. In general, most clusters are either consonant- or vowel-dominant, while groups of the similar vowels form sub-clusters under the big vowel cluster. Although far from perfect, the results suggest that the learned phone embeddings capture phonological features of English. On one hand, the emergence of such phonological structure is not surprising, as phonology is part of what defines a word, although our word segmentation model does not explicitly target it. On the other hand, such results are relevant as they suggest that the phonological regularities are salient and learnable from speech data even if lexical knowledge is absent.

4.3 Comparison with word2vec embeddings

We see that our phone embeddings can capture segmentation-informative and phonology-related patterns. A question remains: is this the consequence of joint learning of the embeddings with the segmentation model, or something also achievable by general-purpose embeddings? We test this by comparing our phone embeddings with the embeddings that are trained by a standard embedding construction tool, word2vec (Mikolov et al., 2013). We first preprocess the raw BR corpus to construct the phone uni- and bi-gram corpora, respectively. Then we run word2vec with *skip-gram* method for 20 iterations on the two corpora to train the embeddings for phone uni- and bi-grams, respectively. The training relies on using each ngram to predict other ngrams in the same local window. We use a window size of 4 phones in the training to be comparable with our models.

We first plot the heatmap of the unigram embeddings of the word2vec model and that of our model in Fig 4, where the embeddings of distinct phone categories in our model exhibit distinct pat-

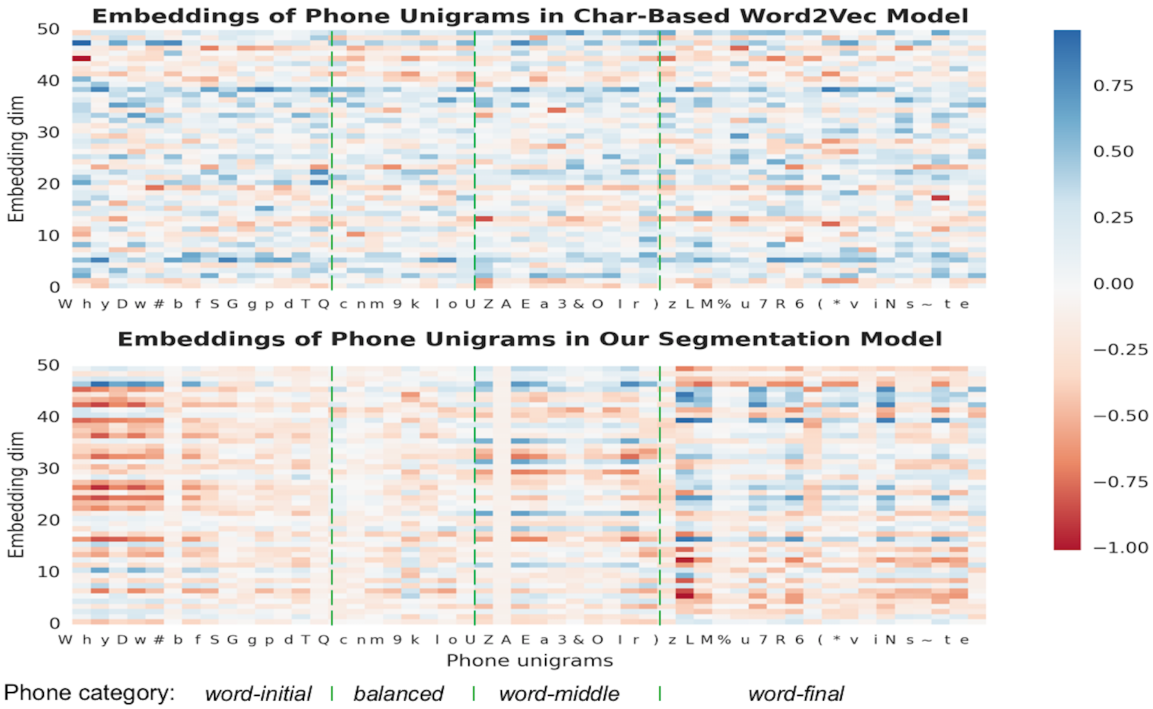


Figure 4: Heatmap of phone embeddings in word2vec (top) and our model (bottom).

terns, whereas such distinctions are unclear in the word2vec embeddings. Then we conduct the same PCA and hierarchical clustering analyses for the word2vec embeddings, as we did for our learned embeddings. The results are shown in Figure 5 and 6, respectively. We see that word2vec embeddings capture neither segmentation-specific features nor phonological structures as our learned embeddings do, which suggests that the joint learning of the embeddings and the segmentation model is essential for the success.

5 Related Work and Discussion

Performance. The focus of this paper is investigating the usefulness of embeddings, rather than achieving best segmentation performance. Since multiple cues are useful for both segmentation by children (Mattys et al., 2005; Shukla et al., 2007) and computational models (Christiansen et al., 1998; Christiansen et al., 2005; Çöltekin and Nerbonne, 2014), our single-cue model is not expected to outperform multiple-cue ones. Nevertheless, it is instructive to compare the performance of our model with other models evaluated in similar settings and use utterance boundaries as the main cue. Using only unigrams, Daland and Pierrehumbert (2011) reports BF, WF and LF scores of 62.7%, 42.5%, and 10.1%, respectively.

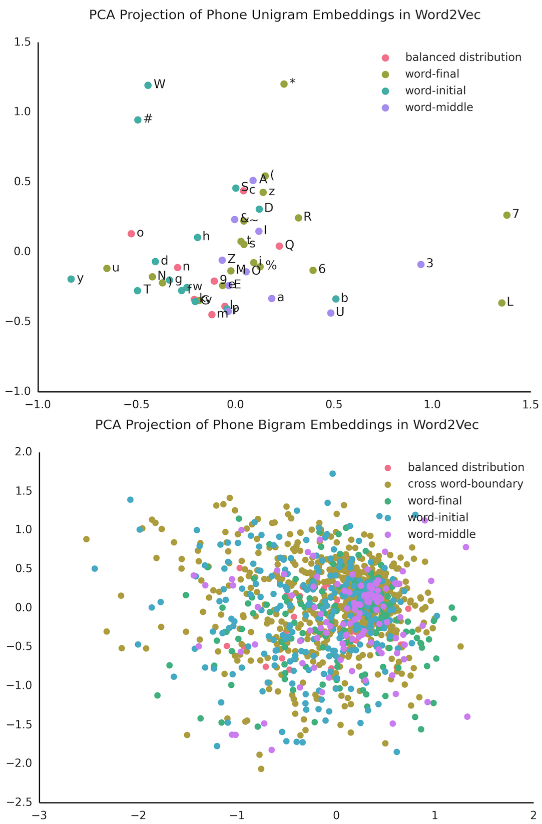


Figure 5: PCA Projections of the embeddings of phone unigrams (top) and bigrams (bottom) in word2vec models.

With a rather elaborate model, Fleck (2008) reports 82.9%, 70.7% and 36.6% (BF, WF, LF). And

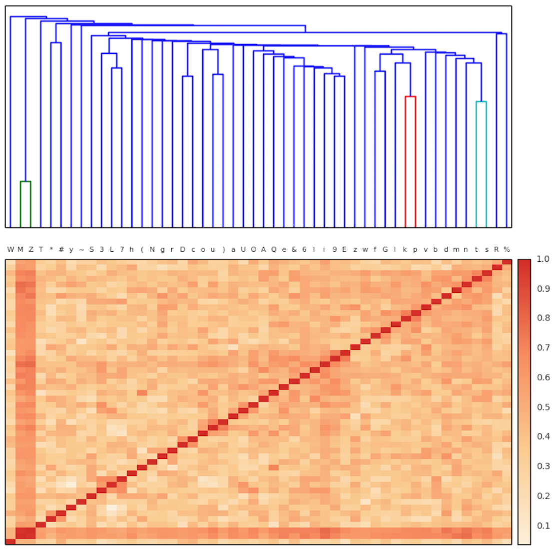


Figure 6: Hierarchical clustering and similarity matrix of phone embeddings in word2vec.

Çöltekin (2011) reports 83.8%, 71.1% and 44.9% (BF, WF, LF), with a model that combines one to three-grams.

The scores in Table 1 suggest that our model fares well in comparison to the state-of-the-art models that exploit similar learning strategies and information sources. The results also show that uni- and bi-grams are effective for segmentation. In addition, we also tried trigrams, which did not improve the results for symbolic or embedding models. This may be due to that the trigrams are too sparse, especially when our training samples only one inter-utterance position per utterance.

Embeddings boost segmentation. Table 1 demonstrates that learning embeddings instead of using symbolic representations boosts segmentation performance. This is true in both settings where the model adopts unigrams and unigram+bigrams as features, respectively. With embeddings, models apply the information obtained from frequent input units to the decisions involving infrequent units with similar representations. Hence, although embeddings are beneficial in both settings, it is not surprising that the improvement is higher for the unigrams+bigrams setting, where the data sparseness is more severe.

Figure 7 shows the difference in the learning curves of the embedding-based and symbolic-based models, both using unigram+bigram features. The embedding model starts with a higher error rate in comparison to the symbolic one, since the vectors for each unit is randomly initialized. However, as the embeddings are updated with more input, the embedding model quickly catches up with the symbolic model and finally outperforms it, as the results in Table 1 show.

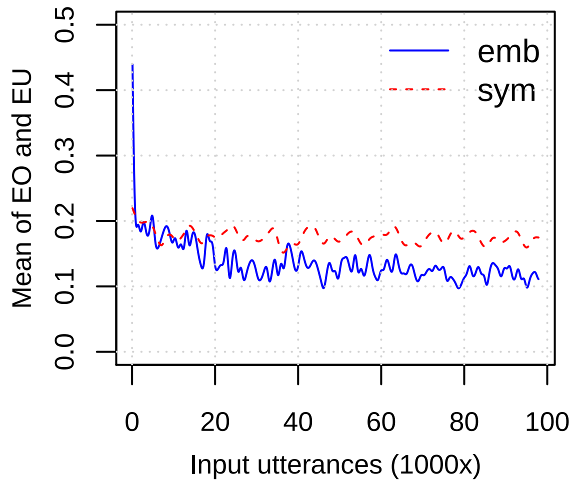


Figure 7: The mean of the error rates during the 1st iteration for the *emb*edding and *sym*bolic models.

Other distributed representations. The utterance boundary cue has been used in earlier work (Aslin et al., 1996; Stoianov and Nerbonne, 2000; Xanthos, 2004; Monaghan and Christiansen, 2010; Fleck, 2008), but not with embeddings. Distributed representations other than embeddings, however, have been common in the early connectionist models (Cairns et al., 1994; Aslin et al., 1996; Christiansen et al., 1998). Besides better performance, our model differs in that it learns the embeddings from the input, while earlier models used hand-crafted distributed representations. This allows our model to optimize representations for the task at hand.

6 Conclusion

In this paper, we have presented a model that jointly learns word segmentation and the embeddings of phone ngrams. The learning in our model is guided by the utterance boundaries. Hence, our learning method, although not unsupervised in machine learning terms, does not use any information that is unavailable to the children acquiring language. To the best of our knowledge, this is the first work of learning phone embeddings for computational models of word segmentation in child language acquisition. Compared with symbolic-based models using the same learning framework, embedding-based models significantly improve results. And the learned embeddings are indicative of not only correct segmentations, but also certain phonology structures.

References

Richard N. Aslin, Julide Z. Woodward, Nicholas P. LaMendola, and Thomas G. Bever. 1996. Models of word segmentation in fluent maternal speech to infants. In James L. Morgan and Katherine Demuth, editors, *Signal to Syntax: Bootstrapping From Speech to Grammar in Early Acquisition*, chapter 8, pages 117–134. Lawrence Erlbaum Associates.

Yoshua Bengio, Réjean Ducharme, Pascal Vincent, and Christian Janvin. 2003. A neural probabilistic language model. *The Journal of Machine Learning Research*, 3:1137–1155.

Nan Bernstein Ratner. 1987. The phonology of parent-child speech. In K. Nelson and A. van Kleeck, editors, *Children's language*, volume 6, pages 159–174. Erlbaum, Hillsdale, NJ.

Michael R. Brent and Timothy A. Cartwright. 1996. Distributional regularity and phonotactic constraints are useful for segmentation. *Cognition*, 61:93–125.

Michael R. Brent. 1999. An efficient, probabilistically sound algorithm for segmentation and word discovery. *Machine Learning*, 34(1-3):71–105.

Paul Cairns, Richard Shillcock, Nick Chater, and Joe Levy. 1994. Modelling the acquisition of lexical segmentation. In *Proceedings of the 26th Child Language Research Forum*. University of Chicago Press.

Çağrı Çöltekin and John Nerbonne. 2014. An explicit statistical model of learning lexical segmentation using multiple cues. In *Proceedings of EACL 2014 Workshop on Cognitive Aspects of Computational Language Learning*.

Morten H. Christiansen, Joseph Allen, and Mark S. Seidenberg. 1998. Learning to segment speech using multiple cues: A connectionist model. *Language and Cognitive Processes*, 13(2):221–268.

Morten H. Christiansen, Christopher M. Conway, and Suzanne Curtin. 2005. Multiple-cue integration in language acquisition: A connectionist model of speech segmentation and rule-like behavior. In J.W. Minett and W.S.-Y. Wang, editors, *Language acquisition, change and emergence: Essays in evolutionary linguistics*, chapter 5, pages 205–249. City University of Hong Kong Press, Hong Kong.

Ronan Collobert, Jason Weston, Léon Bottou, Michael Karlen, Koray Kavukcuoglu, and Pavel Kuksa. 2011. Natural language processing (almost) from scratch. *The Journal of Machine Learning Research*, 12:2493–2537.

Çağrı Çöltekin. 2011. *Catching Words in a Stream of Speech: Computational simulations of segmenting transcribed child-directed speech*. Ph.D. thesis, University of Groningen.

Anne Cutler and Sally Butterfield. 1992. Rhythmic cues to speech segmentation: Evidence from juncture misperception. *Journal of Memory and Language*, 31(2):218–236.

Robert Daland and Janet B Pierrehumbert. 2011. Learning diphone-based segmentation. *Cognitive Science*, 35(1):119–155.

Jacob Devlin, Rabih Zbib, Zhongqiang Huang, Thomas Lamar, Richard Schwartz, and John Makhoul. 2014. Fast and robust neural network joint models for statistical machine translation. In *Proceedings of ACL*, pages 1370–1380.

Margaret M. Fleck. 2008. Lexicalized phonotactic word segmentation. In *Proceedings of the Annual Meeting of the Association of Computational Linguistics (ACL-08)*, pages 130–138.

Sharon Goldwater, Thomas L. Griffiths, and Mark Johnson. 2009. A Bayesian framework for word segmentation: Exploring the effects of context. *Cognition*, 112:21–54.

Mark Johnson and Sharon Goldwater. 2009. Improving nonparameteric Bayesian inference: experiments on unsupervised word segmentation with adaptor grammars. In *Proceedings of Human Language Technologies: The 2009 Annual Conference of the North American Chapter of the Association for Computational Linguistics*, pages 317–325.

Stephen C Johnson. 1967. Hierarchical clustering schemes. *Psychometrika*, 32(3):241–254.

Eric Jones, Travis Oliphant, Pearu Peterson, et al. 2001–. SciPy: Open source scientific tools for Python. [Online; accessed 2016-04-29].

Peter W. Jusczyk, Anne Cutler, and Nancy J. Redanz. 1993. Infants' preference for the predominant stress patterns of English words. *Child Development*, 64(3):675–687.

Peter W. Jusczyk, Derek M. Houston, and Mary Newsome. 1999. The beginnings of word segmentation in English-learning infants. *Cognitive Psychology*, 39:159–207.

Brian MacWhinney and Catherine Snow. 1985. The child language data exchange system. *Journal of Child Language*, 12(2):271–269.

Sven L. Mattys, Laurence White, and James F. Melhorn. 2005. Integration of multiple speech segmentation cues: A hierarchical framework. *Journal of Experimental Psychology: General*, 134(4):477–500.

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.

Padraic Monaghan and Morten H. Christiansen. 2010. Words in puddles of sound: modelling psycholinguistic effects in speech segmentation. *Journal of Child Language*, 37(Special Issue 03):545–564.

Wenzhe Pei, Tao Ge, and Chang Baobao. 2014. Max-margin tensor neural network for chinese word segmentation. In *Proceedings of ACL*, pages 239–303.

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

Mohinish Shukla, Marina Nespor, and Jacques Mehler. 2007. An interaction between prosody and statistics in the segmentation of fluent speech. *Cognitive Psychology*, 54(1):1–32.

Richard Socher, John Bauer, Christopher D Manning, and Andrew Y Ng. 2013. Parsing with compositional vector grammars. In *Proceedings of ACL*, pages 455–465.

Ivelin Stoianov and John Nerbonne. 2000. Exploring phonotactics with simple recurrent networks. In Frank van Eynde, Ineke Schuurman, and Ness Schelkens, editors, *Proceedings of Computational Linguistics in the Netherlands 1999*, pages 51–67.

Kari Suomi, James M. McQueen, and Anne Cutler. 1997. Vowel harmony and speech segmentation in finnish. *Journal of Memory and Language*, 36(3):422–444.

Yoshimasa Tsuruoka, Jun'ichi Tsujii, and Sophia Ananiadou. 2009. Stochastic gradient descent training for L1-regularized log-linear models with cumulative penalty. In *Proceedings of ACL-IJCNLP*, pages 477–485.

Anja van Kampen, Güliz Parmaksız, Ruben van de Vijver, and Barbara Höhle. 2008. Metrical and statistical cues for word segmentation: The use of vowel harmony and word stress as cues to word boundaries by 6- and 9month-old Turkish learners. In Anna Gavarro and M. Joao Freitas, editors, *Language Acquisition and Development: Proceedings of GALA 2007*, pages 313–324.

C. J. van Rijsbergen. 1979. *Information Retrieval*. Butterworth-Heinemann, 2nd edition.

Anand Venkataraman. 2001. A statistical model for word discovery in transcribed speech. *Computational Linguistics*, 27(3):351–372.

Aris Xanthos. 2004. An incremental implementation of the utterance-boundary approach to speech segmentation. In *Proceedings of Computational Linguistics in the Netherlands (CLIN) 2003*, pages 171–180.

Xiaoqing Zheng, Hanyang Chen, and Tianyu Xu. 2013. Deep learning for chinese word segmentation and pos tagging. In *Proceedings of EMNLP*, pages 647–657.

A Symbols used in BR corpus

Consonants	
Symbol	Example
D	the
G	jump
L	bottle
M	rhythm
N	sing
S	ship
T	thin
W	when
Z	azure
b	boy
c	chip
d	dog
f	fox
g	go
h	hat
k	cut
l	lamp
m	man
n	net
p	pipe
r	run
s	sit
t	toy
v	view
w	we
y	you
z	zip
~	button

(a) Consonants

Vowels	
Symbol	Example
&	that
6	about
7	bOy
9	fly
A	but
E	bet
I	bit
O	law
Q	bout
U	put
a	hot
e	bay
i	bee
o	boat
u	boot

(b) Vowels

Rhotic Vowels	
Symbol	Example
#	are
%	for
(	here
)	lure
*	hair
3	bird
R	butter

(c) Vowels with 'r' (Rhotic Vowels)

The above table of symbols is adapted from Çöltekin (2011).