Learning Phone Embeddings for Word Segmentation of Child-Directed Speech

Jianqiang Ma   Çağrı Çöltekin   Erhard Hinrichs

EBERHARD KARLS
UNIVERSITÄT
TÜBINGEN

Department of Linguistics and SFB 833

ACL CogACLL Workshop, Aug 11 2016

Outline

Introduction: word segmentation of child-directed speech

`kitty thats right kitty`

- Segmenting continuous utterances is one of the first tasks for a child acquiring language

Introduction: word segmentation of child-directed speech

`kittythatsrightkitty`

- Segmenting continuous utterances is one of the first tasks for a child acquiring language
- Spoken speech does not contain reliable cues for word boundaries

Introduction: word segmentation of child-directed speech

`ljuuzuibutsjhiuljuuz`

- Segmenting continuous utterances is one of the first tasks for a child acquiring language
- Spoken speech does not contain reliable cues for word boundaries
- Unlike you (adults), children do not have a complete lexicon either

Introduction: word segmentation of child-directed speech

`ljuuzuibutsjhiuljuuz`

- Segmenting continuous utterances is one of the first tasks for a child acquiring language
- Spoken speech does not contain reliable cues for word boundaries
- Unlike you (adults), children do not have a complete lexicon either
- There are some cues in the input that help children segment the utterances: *statistical regularities*, *stress*, *utterance boundaries*, . . .

Introduction: word segmentation of child-directed speech

`ljuuzuibutsjhiuljuuz`

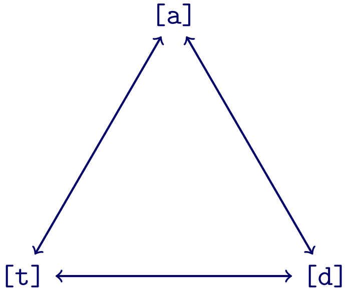
- Segmenting continuous utterances is one of the first tasks for a child acquiring language
- Spoken speech does not contain reliable cues for word boundaries
- Unlike you (adults), children do not have a complete lexicon either
- There are some cues in the input that help children segment the utterances: *statistical regularities*, *stress*, *utterance boundaries*, ...
- Computational models are particularly useful for investigating usefulness of these cues, and types of *input representations*

Introduction: motivation for using phone embeddings

- The way input is represented affects learning

Introduction: motivation for using phone embeddings

- The way input is represented affects learning
- Most segmentation models in the literature represent input as a sequence of symbols
  - → all phones are equally different from (or similar to) each other



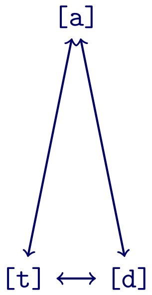
[a]: 0, 1, 0

[d]: 1, 0, 0

[t]: 0, 0, 1

Introduction: motivation for using phone embeddings

- The way input is represented affects learning
- Most segmentation models in the literature represent input as a sequence of symbols
  - → all phones are equally different from (or similar to) each other
- (Continuous) vector representations allow representing and exploiting the similarities between phones



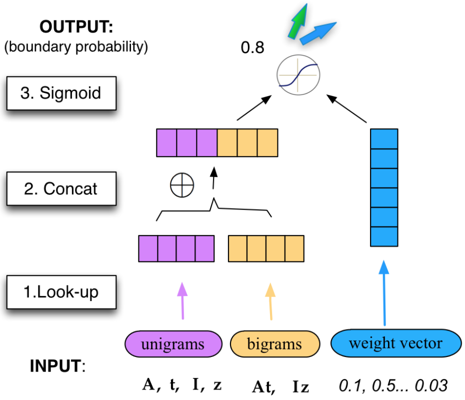
[a]: 0.00, 3.46

[d]: -0.70, 0.00

[t]: 0.70, 0.00

Outline

Model Architecture



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

- **Look-up table** maps phone ngrams to their embeddings.
- Phone embeddings are **concatenated** into a single *input embedding*
- The *boundary probability*: the **sigmoid function** of the dot product of the input embedd'n & the weight vector.

Our model jointly learns the embeddings and the weight vector.

Learning

- Learning with utterance boundaries and random sampling
  - Use *utterance boundaries* as positive instances of word boundaries
  - Randomly sample one position within the utterance as negative instance
- On-line learning
  - cross entropy loss function
  - L2 regularization
  - stochastic gradient descent

Experiments

- why? compare embeddings and symbolic representations (sparse binary vector)
- how? use the same learning framework that accommodates both representations
  - symbolic counterpart of the model: a logistic regression
- what? For both models, run on the same dataset with the same hyper parameters; report the average results of 10 runs

Experiments: dataset and evaluation metrics

- dataset: *BR corpus*: the de facto standard corpus for segmentation
  - collected and converted to transcription by Bernstein Ratner (1987) and Brent & Cartwright (1996), respectively
  - part of the CHILDES database
- evaluation metrics
  - Precision, recall and F-scores
    - **boundary F-score**
    - **word F-score**
    - **lexicon F-score**
  - over-segmentation (**EO**) and under-segmentation (**EU**) error rate

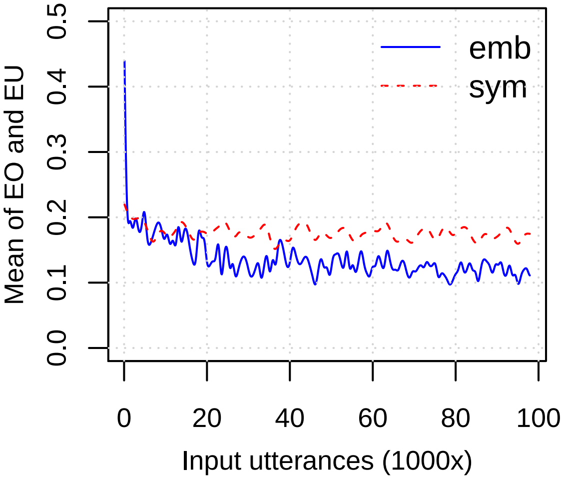
Experiments: results

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

Numbers in percentage. BF, WF, LF are F-scores (higher is better), EO and EU are error rates (lower is better).

- using embeddings boosts performance
- using all (uni- & bi-gram) features works better than only unigrams
- results are on-par with previous methods with similar cues/settings

Experiments: learning curve



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

Outline

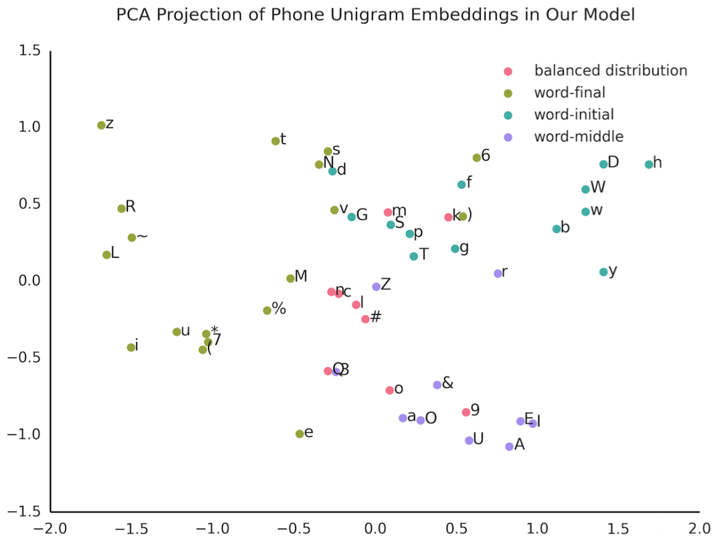
Segmentation roles

- hypothesis: the learned embeddings correspond to metrics that are indicative for segmentation decisions
- project phone embeddings to data points in 2-D space
- color-code phone points w.r.t. segmentation roles

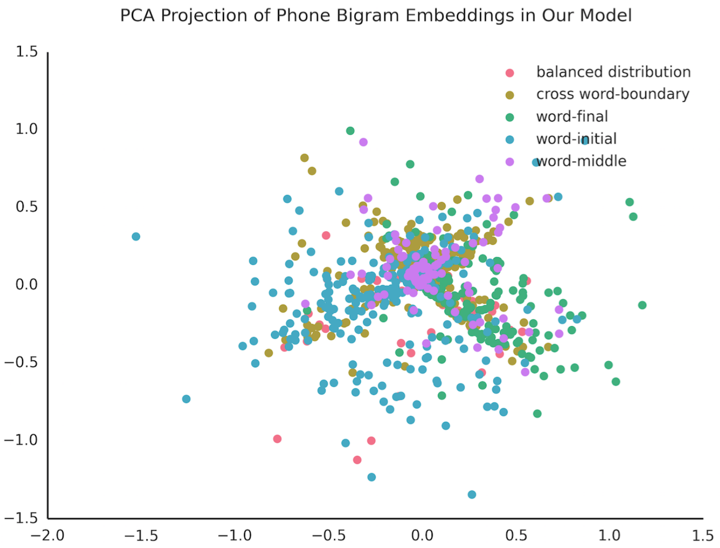
segmentation role: whether a phone ngram is more likely ($> 50\%$) to occur at a specific type of locations

word-initial: at the beginning of a word
word-final: at the end of a word
word-medial: in the middle of a word (non-initial, non-final)
balanced distribution of above positions
cross word-boundary (only for bigrams)

Segmentation roles: unigrams



Segmentation roles: bigrams

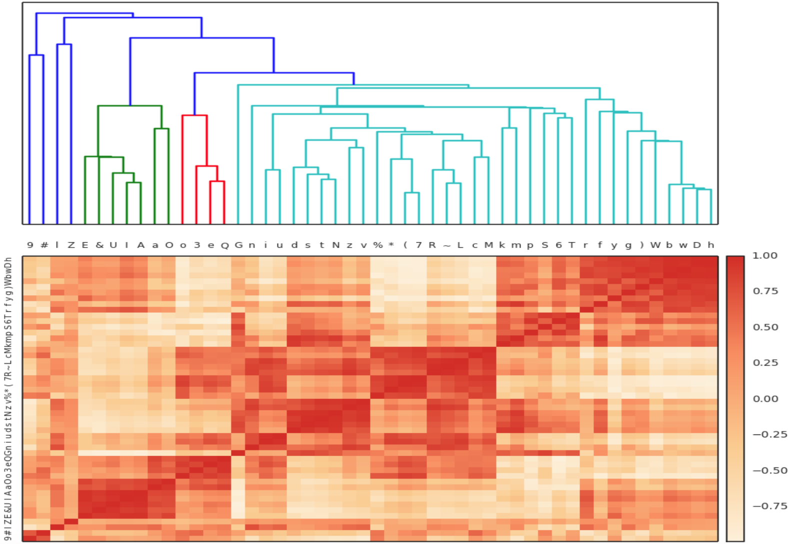


Embeddings capture phonology

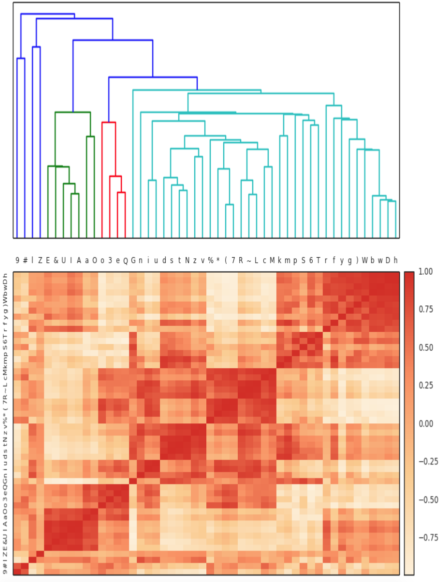
What structures emerge from the embeddings of phone unigrams?

- *similarity matrix* stores pair-wise cosine similarities between embeddings
- *hierarchical agglomerative clustering* builds clusters of phone unigrams in a bottom-up manner.
- visualize them using aligned *heatmap* and *dendrogram*

Hierarchical clustering & similarity matrix of embeddings



Embeddings capture phonology



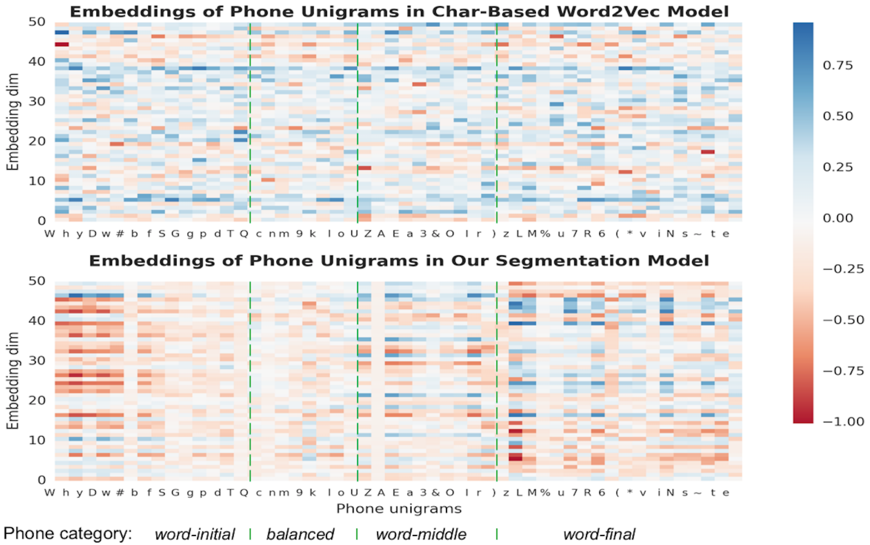
- right 2/3 are mostly consonants while the rest are mostly vowels
- similar vowels form sub-clusters under the big vowel cluster

Example

- unrounded vowels *E*, *&*, *I*, *A* as in bet, that, bit and but
- long-short vowel pair *a* and *O* as in hot and law
- compound vowels, *o*, *3*, *e*, *Q* as in boat, bird, bay and bout

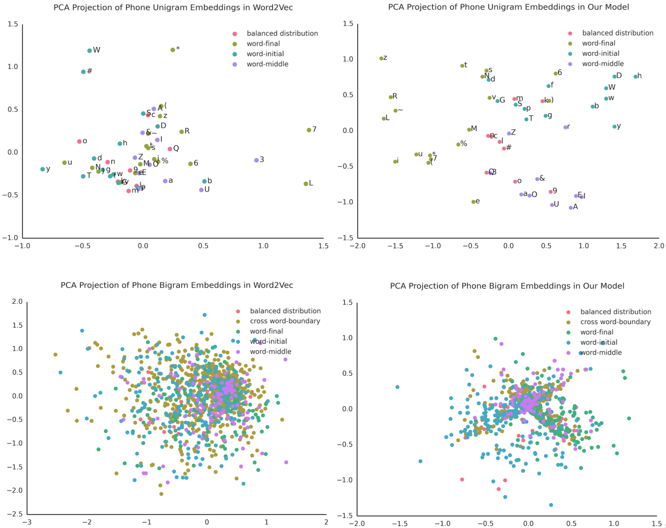
Comparison with word2vec embeddings

dimension-wise heatmaps



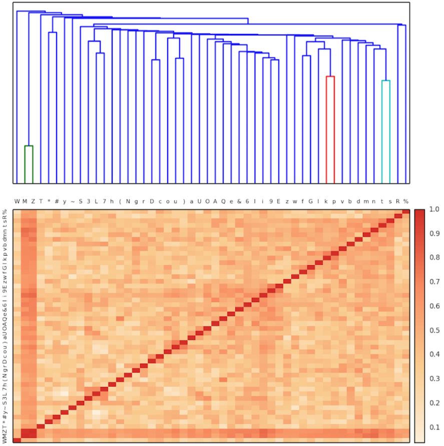
Comparison with word2vec embeddings

segmentation roles

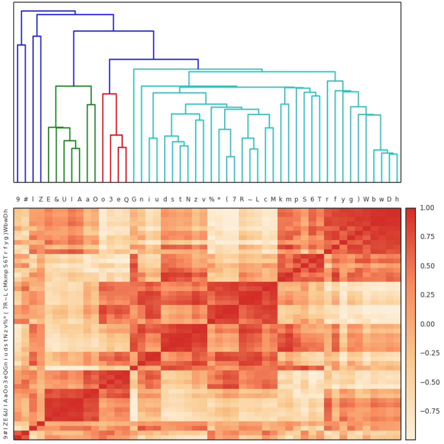


Comparison with word2vec embeddings

Phonology



embedding learned by word2vec



embeddings learned by our model

Summary

- We propose a model that jointly learns the phone embeddings and word segmentation.
- Our model relies on utterance boundaries, thus does not use any information that is unavailable to the children acquiring language.
- Using embeddings significantly improves the performance.
- The learned embeddings are informative for both word segmentation and certain phonological structures.

Summary

- We propose a model that jointly learns the phone embeddings and word segmentation.
- Our model relies on utterance boundaries, thus does not use any information that is unavailable to the children acquiring language.
- Using embeddings significantly improves the performance.
- The learned embeddings are informative for both word segmentation and certain phonological structures.

Thank you!