An LSTM Architecture for Phonotactically-Informed Word Segmentation

Sara Ng
University of Washington
sbng@uw.edu

Abstract

Word segmentation is a first-step to many tasks in speech processing and natural language processing (NLP). In this work, I outline framework for predicting word boundaries from phonemic transcriptions of natural speech data. Using an LSTM architecture on TIMIT data using phonotactically-informed feature embeddings, the framework is able to predict word boundaries with high fidelity.

1 Introduction

The task of phoneme segmentation is as follows: given a stream of symbolic phones representing an audio signal, determine whether each phone is on a word boundary, word internal, or an isolate.

This paper proceeds as follows. Section 2 describes previous studies in both the psychological realities of word segmentation and computational implementations. Section 3 provides an overview of the data and methods used in two experiments. Section 4 provides the empirical results of these experiments and offers comparison to other methods of segmentation. Section 5 offers some intuitions behind the findings of the experiments, as well as possible further directions for this body of work.

2 Related Work

There is a wealth of previous work, both in theoretical linguistic and computational bodies, on issues in speech segmentation. Section 2.1 provides some background on how segmentation is thought to occur in humans, and Section 2.2 overviews some of the recent approaches to automatic word segmentation in the computational realm.

2.1 In Humans

The field of child acquisition provides several plausible models for how humans learn word boundaries as native (L1) speakers.

In their study of 8-month-olds, Johnson & Jusczyk (2001) show that purely statistical models poorly mirror the performance of children as they parse parents' directed speech. In fact, they posit that phonotactics, the interactions of sounds and their changes based on word position, are being learned by the novel speakers as highly-active heuristics. In addition, there are supra-segmental cues, such as stress timing and prosody, that greatly impact children's models of segmentation (Johnson & Jusczyk, 2001). Therefore, a psychologically-motivated computational model of word segmentation would ideally be able to learn information about the symbolic phonological features in a signal and their possible deviations from an underlying representation, as well as higher-level speech cues in the production. In this work, I focus on the features available at the phone level only, as this allows for the use of text transcriptions without human-created stress marking.

The unpublished manuscript by Gambell & Charles (2006) suggests that computational models of speech segmentation benefit from non-statistical approaches as well. Using child-directed speech samples from the CHILDES corpus (MacWhinney & Snow, 1985), the authors show that both statistical models and statistical models ensembled with ontologically motivated stress constraints under-perform hand-crafted systems that leverage featural information on the segments (Gambell & Charles, 2006).

2.2 In Computation

From the computational perspective, it may seem un-intuitive to propose learning word segmentations from transcribed sounds; after all, if a phonetic transcription and pronouncing dictionary are available, as is the case with TIMIT, the task of aligning phones with the dictionary entries is trivial and highly faithful. However, producing models based on transcribed phones may offer downstream advantages (see Section 5). Fleck (2008) show how phoneme n-grams and durational silence cues can be used cross-linguistically to induce word segmentation. Goldwater (2006) also show highly faithful results for this task in another statistical framework.

Related computational tasks, however, are essential to speech processing. For example, much work has been done on word segmentation in languages like written Chinese or Arabic, where there aren't white space word boundary cues as in Romanized text. In these cases, character and sub-character-based architectures, especially LSTMs and EM, have given high performance (see Chen et al. (2015), Cai & Zhao (2016), Kitagawa & Komachi (2017), and Peng & Schuurmans (2001), for a sample). In the spoken domain, the analogous task of signal segmentation is essential as a first step to automatic speech recognition (ASR). It is not unreasonable to say that either of these tasks may benefit from additional input based on discrete phonetic information like the kind that can be found in transcriptions or pronouncing dictionaries.

3 Methodology

3.1 Data

For this study, I employ a subset of the TIMIT corpus (Garofolo et al., 1993). This data set includes lab recordings of read English speech, alongside human-created annotations and meta data. The elicitations are designed to be phonologically illustrative across varieties of American English, with portions of the data highlighting dialectical variation (Zue & Seneff, 1996). The advantage of using TIMIT data is that not only is the speech and text available for each utterance, but also human-crafted phonetic transcriptions, alignments by sample, and a bespoke phonetic dictionary for the data. What's more, meta data for speaker dialect, race, and gender are associated to each sample.

For this study, I follow the suggested data division in Garofolo et al. (1993) for TIMIT data, where 60% of utterances are used in training, 20% for validation, and 20% for testing. Additionally, I remove all sentences from the test set that were spoken by any speaker in the training (the "SA" sentences in the corpus' own encoding). Table 1 summarizes the used data set.

	Sentences	**Speakers**	**Phones (transcribed)**	**Phones (translated)**
Train	3,696	463	134,627	121,190
Test	1,344	168	48,628	43,981

Table 1: Data Summary

One of the central questions of this research is whether phonotactics, not just phones themselves, inform segmentation. While TIMIT transcriptions are not sufficiently fine-grained to show some of the intuitively informative phonotactic features for word segmentation (e.g. lack of release on on final obstruents), the multiple encodings of the samples can be used to approximate gradient effects of fine-grained phonotactic distinctions. Further detail follows in Section 3.2.1.

In a data set like TIMIT, human-crafted phonetic transcriptions should reflect real-time phonotactic variation. For example, one should expect the duration of word-initial obstruents to be longer than their word-final counterparts (Turk & Shattuck-Hufnage, 2000). At the feature-level, word-internal vowels should be reduced or centralized (Flemming & Johnson, 2007). It is the hope of this study that such phenomenon will be acquired tacitly by a neural model.

3.2 Architecture

The TIMIT data set represents a subset of uniquely clean speech data available at its data size. The inclusion of human phonetic transcription is especially valuable in testing assumptions about the utility of phonotactics. Based on the availability of such data, I design two experiments on word segmentation: one based on human phonetic transcriptions, and one based on input from a bespoke pronouncing dictionary.

3.2.1 Pre-processing

To encourage the neural architecture to learn theoretically motivated features, I hand-craft feature embeddings for the two types of data.

In **Experiment 1**, the phone sequence is taken directly from human transcription. To determine

the tag for each segment, its sample duration was compared to the listed sample durations for the sentence at the word level.

In **Experiment 2**, the word sequence itself was passed through the TIMIT pronouncing dictionary to produce a novel phone sequence. While the given dictionary includes marking for primary and secondary stress, this was excluded from consideration and removed from the final string (see Section 5 for further discussion). In cases where the pronouncing dictionary had more than one possible pronunciation, e.g. for different parts of speech, one was chosen at random.

In both experiments, the features *voicing, manner, nasality, place, syllabicity, height, backness, rounding, rhoticity,* and *second diphthong segment* were used as these are minimally distinctive for English. Each feature was encoded using the minimum number of bits, resulting in an initial embedding of size 25. For each sample, speaker dialect and gender were concatenated to the embedding, resulting in vectors of length 29. For Experiment 1, the additional feature *duration* of the phone was added as a single dimension. In each sample, characters marking initial and final silences were excluded. Epenthetic pauses and pad characters were equally treated as zero vectors. Table 2 shows the possible distinctive features.

Feature (# bits)	**Possible Values**
voicing (2)	voiced, voiceless, N/A
manner (3)	stop, affricate, fricative, flap, oral closure, lateral approximate, approximant, N/A
nasality (2)	yes, no, N/A
place (4)	bilabial, labiodental, dental, alveolar, postalveolar, palatal, velar, labiovelar, glottal, N/A
syllabicity (2)	yes, no, N/A
height (3)	high, near-high, higher-mid, mid, lower-mid, near-low, low, N/A
backness (3)	front, near-front, central, near-back, back, N/A
rounding (2)	yes, no, N/A
rhoticity (2)	yes, no, N/A
diphthong 2nd (2)	'i', 'u', N/A
dialect (3)	1-8
gender (1)	M, F
duration (1)	# in samples

Table 2: Distinctive Features for Word Embeddings

For each sentence in both experiments, the phone sequence was encoded in this way and then right-padded to yield sequences of even length.

3.2.2 Baseline

While rough estimations can be compared between the findings of my experiments and the experimentation in Gambell & Charles (2006), for a more meaningful comparison I conducted a statistically-motivated baseline based on the same embeddings used in both experiments. For each unique phone, a prediction was chosen at random from a probability distribution on the training set. Table 3 shows the results.

		Precision	Recall	F1
Transcribed	Train	0.744	0.744	0.744
	Test	0.734	0.734	0.734
Translated	Training	0.758	0.758	0.758
	Test	0.749	0.749	0.749

Table 3: Evaluation Metrics on Baseline for 2 Experimental Conditions

Figures 1-2 show the confusion matrices for the training and test set on the embeddings from transcribed data, and Figures 3-4 show the confusions for the embeddings translated from the pronouncing dictionary.

3.2.3 LSTM

For **Experiment 1**, an LSTM, consisting of

1. a 128-neuron LSTM layer,
2. a 128-neuron dense layer with ReLU activation,
3. a 256-neuron LSTM layer,
4. a dropout layer with rate 0.5, and
5. an output SoftMax

was trained for 25 epochs in batches of 12 with early stopping condition on the validation loss (δ=10e-4). For **Experiment 2**, a similar LSTM was trained with batches of 5 and a finer minimum delta for early stopping (10e-5). All hyperparameters were tuned using the performance on the training data. Both systems were optimized on categorical cross-entropy using Adam.

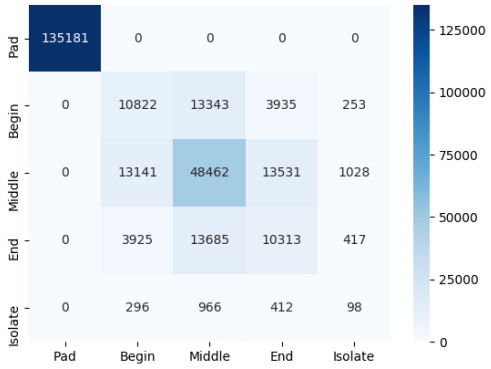

Figure 1: Baseline Transcription Training

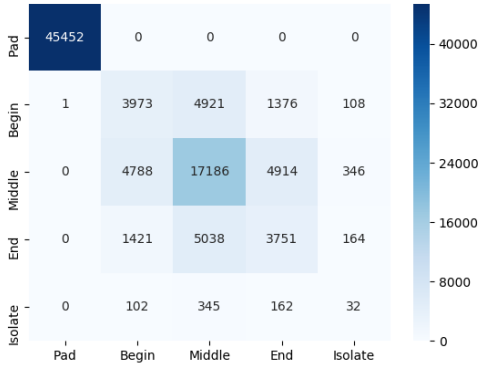

Figure 2: Baseline Transcription Testing

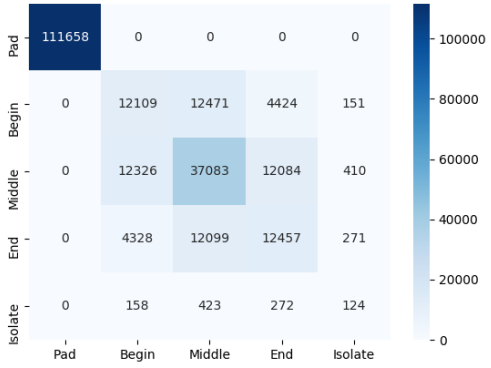

Figure 3: Baseline Translation Training

3.3 Evaluation Metric

Following the tradition of similar work on segmentation (and the admonition found in (Gambell & Charles, 2006)), I evaluate all models on precision, recall, and F1.

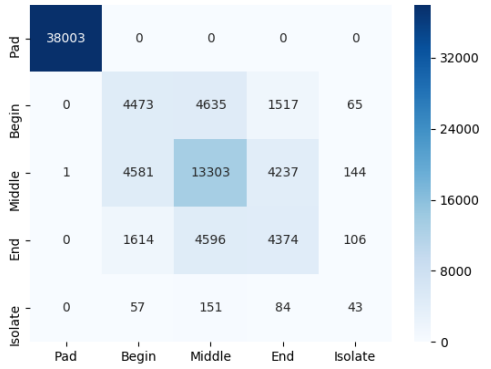

Figure 4: Baseline Translation Testing

The relevant source code is available at `github.com/SaraBlalockNg/EE511`.

4 Experimental Results

Both experiments were run on a CPU using the Keras library for python. The run time for Experiment 1 was 9 minutes 35 seconds, and the run time for Experiment 2 was 17 minutes 5 seconds. The evaluation measures are given in Tables 4 and 5, respectively.

	Precision	**Recall**	**F1**
Train	0.884	0.824	0.853
Valid	0.889	0.825	0.856
Test	0.888	0.824	0.855

Table 4: Evaluation Metrics for Experiment 1

	Precision	**Recall**	**F1**
Train	0.966	0.964	0.965
Valid	0.943	0.940	0.941
Test	0.888	0.883	0.885

Table 5: Evaluation Metrics for Experiment 2

Confusion matrices for each experiment are given in Figures 5 and 6.

5 Analysis

In general the performance of the two systems according to the standard metrics is comparable to other segmentation architectures (Fleck , 2008; Gambell & Charles, 2006), though both these

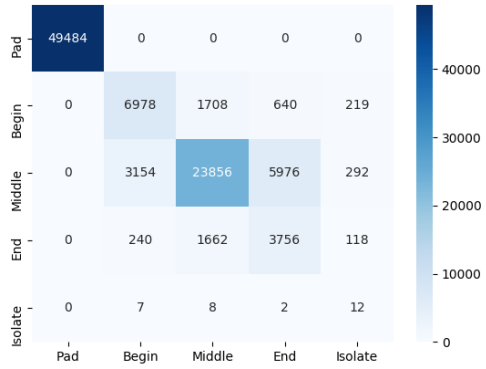

Figure 5: Confusion on Transcribed Test Set

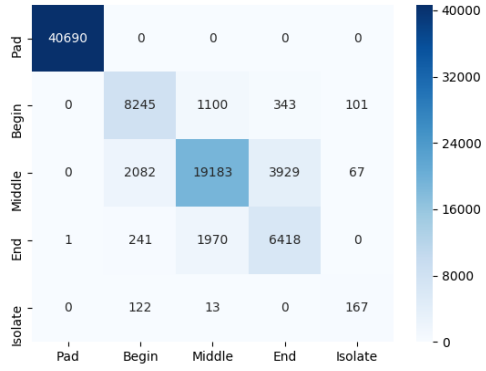

Figure 6: Confusion on Translated Test Set

studies had different data and a slightly different classification task.

Given the assumptions provided in the linguistic literature, it is somewhat disheartening that the LSTM trained on the more faithful phonetic transcriptions consistently under-performs the model trained on dictionary translations. There may be several reasons for this. First, it could be that the range of hyper-parameter tuning and layer selection was just better suited to the rougher phonetic translations. It may also be that suprasegmental cues that inform segmentations are more important than the phonotactics of individual phones. It is also possible that the level of granularity offered by the TIMIT annotation style is not equipped for the subtle differences within sets of productions of individual phones.

While the model in Experiment 2 performs better overall, it also seems to be suffering from at least some degree of overfitting. If I were to revise the model, I may consider stopping earlier or increasing the amount of dropout.

5.1 Further Directions

There are a few direction that I would be excited to move this work. The high-fidelity performance of the pronouncing dictionary-based model suggests that this kind of architecture can be expanded to data sets that do not have phonetic transcriptions. This would not only allow for larger data sets, but also possibly typological comparisons.

I am also interested in considering how such a system could inform signal segmentation. Do, for example, the phonotactic cues have readily-available acoustic correlates that could be learned?

Finally, there is evidence from other experiments that stress patterns are highly informative to word segmentation (Gambell & Charles, 2006). I would like to incorporate some type of learned stress marking into my system so see how it may change performance.

6 Conclusion

In this paper, I have motivated the use of phonotactic features as input to word segmentation on phonetic transcriptions. I have shown that a simple LSTM framework can predict word boundaries with fidelity comparable to both theoretically- and performance-driven frameworks. Finally, I have provided further directions which I to which I am optimistic about extended this work.

7 Acknowledgments

My sincere thanks are given to Leanne Rolston, Rik Koncel-Kedziorski, and Agatha Downey for their encouragement and stalwart “Rubber Duck” surrogacy.

References

Deng Cai and Hai Zhao. Neural word segmentation learning for Chinese. *arXiv preprint, arXiv:1606.04300*, 2016.

Xinchi Chen, Xipeng Qiu, Chenxi Zhu, Pengfei Liu, and Xuanjing Huang. Long short-term memory neural networks for Chinese word segmentation. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*, pages 1197-1206, 2015.

Margaret M Fleck. Lexicalized phonotactic word segmentation. In *Proceedings of ACL-09: HLT*, pages 130-138, 2008.

Edward Flemming and Stephanie Johnson. Rosa's roses: Reduced vowels in American English. In *Journal of the International Phonetic Association*, 37(1):83-96, Cambridge University Press, 2007.

Timothy Gambell and Charles Yang. Word Segmentation: Quick but not dirty. *Unpublished manuscript*.

John S Garofolo, Lori F Lamel, William M Fisher, Jonathan G Fiscus, and David S Pallett. DARPA TIMIT acoustic-phonetic continuous speech corpus CD-ROM. NIST speech disc 1-1.1. *NASA STI/Recon technical report n*, 93, 1993.

Sharon Goldwater. Nonparametric Bayesian Models of Lexical Acquisition, Ph.D. thesis, Brown University, 2006.

Elizabeth K Johnson and Peter W Jusczyk. Word segmentation by 8-month-olds: When speech cues count more than statistics. In *Journal of Memory and Language*, 44(4):548-567, 2001.

Yoshiaki Kitagawa and Mamoru Komachi. Long short-term memory for Japanese word segmentation. *arXiv preprint, arXiv:1709.08011*, 2017.

Brian MacWhinney and Catherine Snow. The child langauge data exchange system. In *Journal of child language, Cambridge University Press*, 12(2):271-295, 1985.

Fuchun Peng and Dale Schuurmans. A hierarchical EM approach to word segmentation. In *NLPRS*, pages 475-480, 2001.

Alice E Turk and Stefanie Shattuck-Hufnage. Word-boundary-related duration patterns in English. In *Journal of Phonetics*, 28(4):397-440, Elsevier, 2000.

Victor W Zue and Stephanie Seneff. Transcription and alignment of the TIMIT database. In *Recent Research Towards Advanced Man-Machine Interface Through Spoken Language*, 515-525, Elsevier, 1996.