An LSTM Architecture for Phonotactically-Informed Word Segmentation

Sara Ng, Dept. of Linguistics

Idea: *Phonotactics,* the way that sounds interact with one another, inform word boundaries

Goal: Given a **phone**, determine its **place** in the word

Motivation

- Experiments show that humans don't learn boundaries statistically!
- Productive phonotactics have **specific rules at word boundaries**

Data

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

TIMIT (LDC93S1) - sentences spoken in 7 dialects of American English, meant to illustrate phonological diversity (1993)

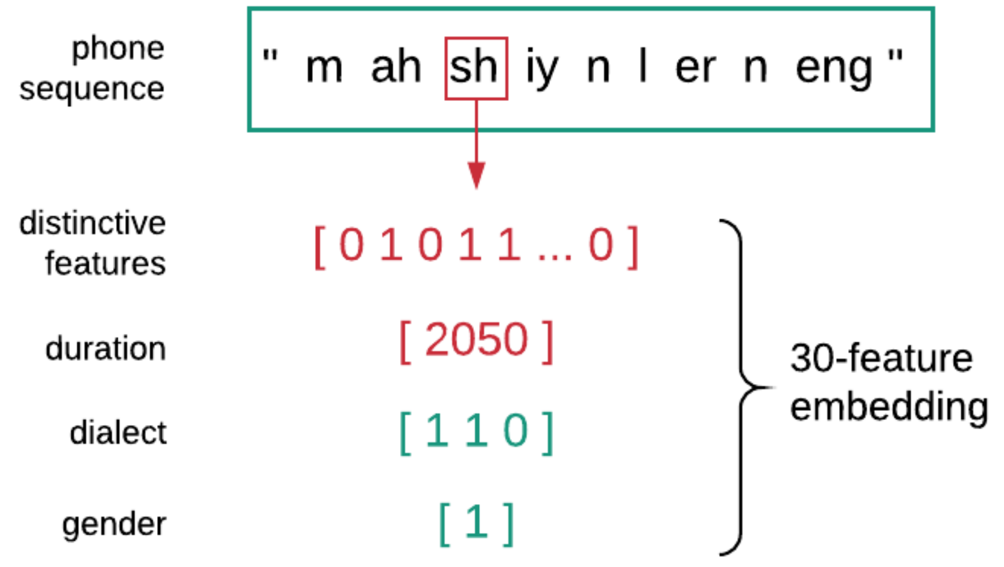
Each elicitation has:

1) human **phonetic transcription**
2) list of spoken **words**
3) information about **dialect** & **gender**
4) **durations** of phones & words

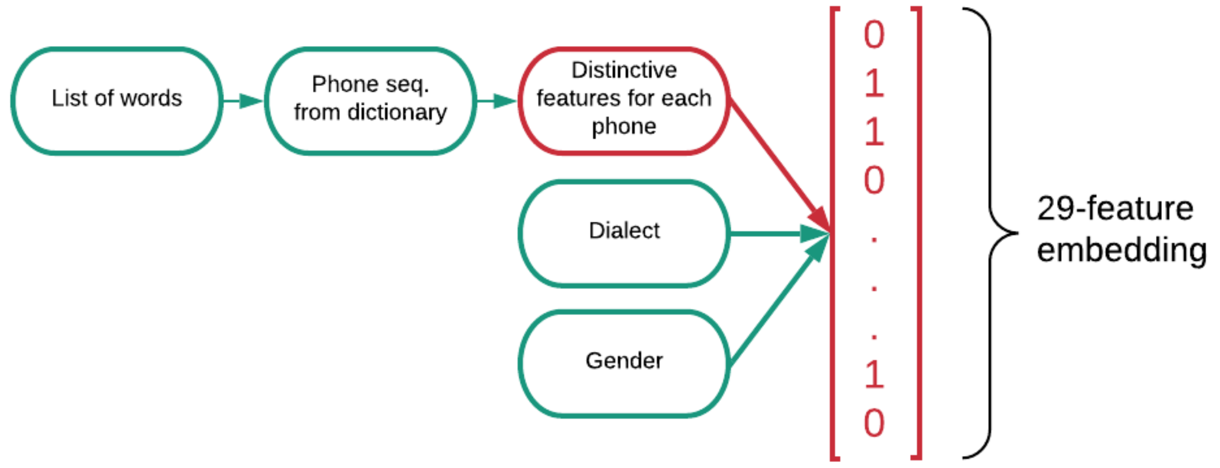
The corpus has its own **pronouncing dictionary**

Design

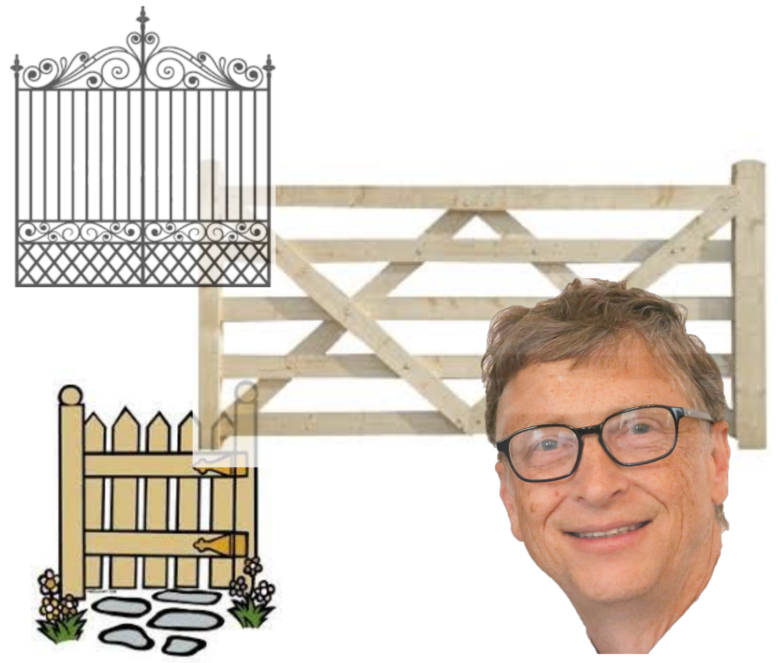
Experiment 1: Transcriptions



Experiment 2: Translations


List of words
Phone seq. from dictionary
Distinctive features for each phone
Dialect
Gender
0
1
1
0
.
.
.
1
0
29-feature embedding

Architecture



Fast Facts:

- batch size = 12/5
- epochs = 25
- early stopping on valid loss
- categorical cross entropy
- dropout = 0.5
- 2 LSTM layers
- Internal transform with ReLU
- final transform with SoftMax

Results (**translated**/translated)

	Precision	Recall	F1
Train	**0.966**/0.884	**0.964**/0.824	**0.965**/0.853
Valid	**0.943**/0.889	**0.940**/0.825	**0.941**/0.856
Test	**0.888**/0.888	**0.883**/0.824	**0.885**/0.855

Summary

- performance comparable to combined statistical/ linguistic heuristics on translations
- poor performance on transcribed data —possibly paucity of features?
- robust within transcription system
- language-specific embeddings reduce dimensionality, but can be modified to accept multilingual features

Next

- go from the signal directly
- look at a more interesting language for which dictionary is available (i.e Arabic)
- use as measure of grained-ness of transcription

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.

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.
Elizabeth K Johnson and Peter W Jusczyk. Word segmentation by 8-month-olds: When speech cues count more than statistics. 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.

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