Cross-Linguistic Phoneme Embeddings for Computational Historical and Typological Linguistics

Marlon Betz

July 6, 2016

Contents

1 Introduction

Embeddings nowadays build the backbone of every Deep Learning NLP architecture. Due to their capacity to encode a vast amount of latent semantic and syntactic information without the need of previous manual construction, they gave rise to the contemporary Deep Learning boom, i.e. deep neural networks that can capture hidden features in data sets and by that have allowed for huge performance gains in numerous NLP fields.
While embeddings are currently being used for several linguistic units such as characters [Kim et al.2015,dos Santos and Zadrozny2014,Zhang et al.2015], words [Mikolov et al.2013, Pennington et al.2014] or entire sentences [Kiros et al.2015], proper phonemes have been largely been excluded from this trend, although there are some data that would indeed allow for it and research areas where they could be of great use, especially in the fields of Computational Typological and Historical Linguistics, which the current deep learning

boom has hardly touched yet. There, even though it is clear that some phonemes share more common features and such form natural classes, they are often treated as pure symbols that share a common distance between each other. Even phoneme representation models that do incorporate phonological features are often hand-crafted [Kondrak2000,Rama2016] or include task-specific information that inherently suffer from restricted generalization abilities when used for other tasks [Jäger2014]. Moreover, those methods usually reduce the number of possible phonemes to a minimum, getting rid of important information such as secondary or co-articulations.

In this paper, I will first a discuss a theoretical motivation for using data-driven phoneme embeddings instead of plain symbolic representations or hand-crafted feature encodings. I will then shortly take a look on related research. A big part of this paper will then review several embedding models. I will then discuss intrinsic evaluation methods for the embeddings and use those to compare the performances of the previously described models. This is then followed by a discussion of several use cases that could be interesting for those interested in data-driven approaches to Typological and Historical Linguistics. Finally, I will recapitulate the benefits and drawbacks of phoneme embeddings in a final resume.

2 Theoretical and Practical Motivation

3 Related Research

4 Embedding Models

4.1 Count-based Embedding Models

4.1.1 Latent Semantic Analysis

4.1.2 GloVe

4.1.3 SPPMI

4.1.4 SPPMI-SVD

4.2 Neural Embedding Models

4.2.1 Word2Vec

5 Evaluation

5.1 Data

5.2 Evaluation Methods

5.3 Results

6 Use Cases

6.1 Phonemic String Comparison

6.2 Modeling Sound Change

6.3 Phoneme Inventory Clustering

7 Resume

References

[dos Santos and Zadrozny2014] dos Santos, C. N. and Zadrozny, B. (2014). Learning character-level representations for part-of-speech tagging. In *ICML*, pages 1818–1826.

[Jäger2014] Jäger, G. (2014). Phylogenetic inference from word lists using weighted alignment with empirically determined weights. In *Quantifying Language Dynamics*, pages 155–204. Brill.

[Kim et al.2015] Kim, Y., Jernite, Y., Sontag, D., and Rush, A. M. (2015). Character-aware neural language models. *arXiv preprint arXiv:1508.06615*.

[Kiros et al.2015] Kiros, R., Zhu, Y., Salakhutdinov, R. R., Zemel, R., Urtasun, R., Torralba, A., and Fidler, S. (2015). Skip-thought vectors. In *Advances in neural information processing systems*, pages 3294–3302.

[Kondrak2000] Kondrak, G. (2000). A new algorithm for the alignment of phonetic sequences. In *Proceedings of the 1st North American chapter of the Association for Computational Linguistics conference*, pages 288–295. Association for Computational Linguistics.

[Mikolov et al.2013] Mikolov, T., Sutskever, I., Chen, K., Corrado, G. S., and Dean, J. (2013). Distributed representations of words and phrases and their compositionality. In *Advances in neural information processing systems*, pages 3111–3119.

[Pennington et al.2014] Pennington, J., Socher, R., and Manning, C. D. (2014). Glove: Global vectors for word representation. In *EMNLP*, volume 14, pages 1532–43.

[Rama2016] Rama, T. (2016). Siamese convolutional networks based on phonetic features for cognate identification. *arXiv preprint arXiv:1605.05172*.

[Zhang et al.2015] Zhang, X., Zhao, J., and LeCun, Y. (2015). Character-level convolutional networks for text classification. In *Advances in Neural Information Processing Systems*, pages 649–657.