Unsupervised Cognate Identification with Variational Autoencoders

Marlon Betz

August 13, 2016

Contents

1 Introduction

2 Motivation

3 Related Research

4 Architecture

Hence, the model should have three major components:

1. The phonemes should be embedded in a feature space, where similar phonemes should cluster in similar subspaces of the feature space.

2. The words as sequences of such phoneme embeddings should themselves be embedded in another feature space, where words with similar shape should cluster among each other.

3. The word embeddings are then clustered in such a way that words that appear together in a cluster are assigned a common label, which is then predicted cognate class.

4.1 Phoneme Vectorization

4.1.1 Hand-crafted Vectorization Models

4.1.2 Data-driven Embeddings

4.2 Word Embeddings

4.2.1 Autoencoders

4.2.2 Variational Autoencoders

4.3 Clustering

4.3.1 Affinity Propagation

5 Evaluation

5.1 Data

5.2 Results

6 Resume

7 Acknowledgements

For training the phoneme embeddings, I used the word2vec implementations provided by the gensim package [Řehůřek and Sojka2010]. The Autoencoder was implemented with Keras [Chollet2015] and Tensorflow [Abadi et al.2015]. The clustering algorithms used here were provided by scikit-learn [Pedregosa et al.2011]. All code connected to this thesis can be found on my github [1]

References

[Abadi et al.2015] Abadi, M., Agarwal, A., Barham, P., Brevdo, E., Chen, Z., Citro, C., Corrado, G. S., Davis, A., Dean, J., Devin, M., Ghemawat, S., Goodfellow, I., Harp, A., Irving, G., Isard, M., Jia, Y., Jozefowicz, R., Kaiser, L., Kudlur, M., Levenberg, J., Mané, D., Monga, R., Moore, S., Murray, D., Olah, C., Schuster, M., Shlens, J., Steiner, B., Sutskever, I., Talwar, K., Tucker, P., Vanhoucke, V., Vasudevan, V., Viégas,

[1] https://github.com/marlonbetz/BA

F., Vinyals, O., Warden, P., Wattenberg, M., Wicke, M., Yu, Y., and Zheng, X. (2015). TensorFlow: Large-scale machine learning on heterogeneous systems. Software available from tensorflow.org.

[Chollet2015] Chollet, F. (2015). Keras. https://github.com/fchollet/keras.

[Pedregosa et al.2011] Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Prettenhofer, P., Weiss, R., Dubourg, V., Vanderplas, J., Passos, A., Cournapeau, D., Brucher, M., Perrot, M., and Duchesnay, E. (2011). Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research*, 12:2825–2830.

[Řehůřek and Sojka2010] Řehůřek, R. and Sojka, P. (2010). Software Framework for Topic Modelling with Large Corpora. In *Proceedings of the LREC 2010 Workshop on New Challenges for NLP Frameworks*, pages 45–50, Valletta, Malta. ELRA. http://is.muni.cz/publication/884893/en.