Proto-Word Reconstruction with RNNs Update 24.08

Julius Steuer Morgan Wixted

Quick Timeline

Week	Julius	Morgan
0 17.08	Collect Romance cognates from the web and align the data	Set up the neural net
1 24.08	Aligned IPA dataset based on Romance data	Train the neural net with IPA embeddings
2 31.08	ASJP/ASCII dataset from IPA	Train the neural net with ASJP/ASCII embeddings (baseline)
3 07.09	Experiment with Ciobanu's data (IPA, ASJP, ASCII)	Compare model performance to baseline
4 14.09	Start writing the paper	Evaluate model performance on Ciobanu's data
5 21.09	Write the paper	Write the paper
6 28.09	Try another dataset, tune model parameters	Try another dataset, tune model parameters

Morgan: Work from last week 17.8 to 24.8

- Setting up the neural network
  - Using the neural machine translation with attention network from TensorFlow.
  - First, got it working with the data provided by TensorFlow.
  - Then, worked to get it working with our data.
- Neural network with our data
  - Challenges:
    - Neural network provided by TensorFlow uses string data for the input and target languages.
    - Our data has string data for the input but an array for the target language.
    - So, had to have input for both a string an an array.
- Status:
  - Can successfully pass one word (at character level) through the network.

Morgan: Work for this week

- Train the Neural Net with IPA embeddings
  - Get a full cognate set to pass through the network.
  - Create a pipeline that can extract the cognate sets from a csv file containing the IPA embeddings and pass them through the neural network.

Julius: Work from last week (17.08-24.08)

- Create IPA & ASJP feature embeddings (multi-hot-vectors)
  - Started with IPA, then used LingPy to go to ASJP
  - Easier to do it this way, as IPA contains more information
- Compile romance dataset based on Swadesh list (1971)
  - Latin (proto), Italian, Spanish, French, Portuguese, Romanian
- Align cognate loci in proto-word and reconstruction

Latin	Italian	Spanish	French	Portuguese	Romanian
Ego	io	So	3e	ew	yew
-Ego	-i-o	-S-o	-3-e	-e-w	ye-w

Julius: Work for next week

- Transform IPA/ASJP characters to binary ASCII/UTF-16 encodings
  - Same Swadesh data
- Reproduce baseline task from Meloni (2018)
  - Replace bare + aligned phonological encodings to orthographical representation