Instructions for EACL 2021 Proceedings

Anonymous EACL submission

Abstract

Proto-word reconstruction is an effective method to help linguistics researchers study language evolution. It consists of reconstructing the ancient word by using the modern words from the daughter languages. In this paper, we aim to extend the work by (Meloni et al., 2019) and create our own system for automatically reconstructing proto-words. Using data compiled from both (Swadesh, 1971) and (Ciobanu and Dinu, 2014), ...TODO

1 Introduction

Recent approaches on the automated reconstructions of proto-words made use of phylogenetic methods, see Bouchard-Côté et al. (2013) for Austronesian and Hruschka et al. (2015) for Turkic. Another approach is to use a neural decoder-encoder model with attention, a pipeline used in machine translation, to map sequences of cognate sounds in related words to a single sound in the (supposedly unknown) parent language. Examples of this approach are Ciobanu and Dinu (2018) and Meloni et al. (2019). While the phylogenetic approaches relied on phonological features of the sounds involved to reconstruct proto-words (and to derive the most likely family tree), the latter two make use of character embeddings to represent sounds. In contrast, Dekker (2018) used a neural machine translation model in conjunction with feature encodings.

How plausible is using a machine translation pipeline for proto-word reconstruction? While a machine translation system maps a sequence of words in one language onto a sequence of words in another language (while preserving the meaning), the above approaches map a sequence of sounds or characters pertaining to a word in a cognate set to a single sound in the proto-language. Since sounds do not make up the meaning of a word in the same sense as words make up the meaning of a sentence (with every word contributing an identifiable part to it), we suspect that a more simple recurrent or even non-recurrent architecture may as well capture sound correspondences.

Another feature of the above approaches is that they seem not to benefit from more a more fine-grained transcription: if IPA transcriptions are used as input to a model, reconstruction accuracy is always lower compared to the ordinary Latin orthography. In this paper, we will investigate whether a more coarse transcription as that used in the context of the ASJP database (Brown et al., 2008) does improve reconstruction accuracy.

We also want to explore how automatic methods can be applied to smaller datasets, and to language families for which the proto-language is not well established. Therefore, we compile a custom romance dataset and compare the performance of our models on it to their performance with a dataset used in the prior studies.

2 Methods

2.1 Models

Here we discuss the different models we use to reconstruct the proto words. We use several different models to evaluate how well they reconstruct each proto word, given a cognate set. We use a simple recurrent neural network (RNN), a long short-term memory (LSTM) network, and a simple feedforward neural network.

For both the RNN and the LSTM, we use the stochastic gradient descent algorithm for optimization. For the feedforward neural network, we use the Adam algorithm for optimization.

As for the loss function, each network uses cosine similarity. This is because we are wanting to measure the similarity between each character of each derived language in comparison to the character of

the reconstructed proto word in question.
For the parameters, we use the default parameters provided by Tensorflow for our models.

2.1.1 Simple Recurrent Neural Network

To construct our simple recurrent neural network, we use the Keras Application Program Interface (API) (Chollet, 2015) from Tensorflow (Abadi et al., 2015). Following the example that Tensorflow gives, we use a sequential model consisting of an embedding layer, a simple recurrent neural network layer, and a dense layer.
As for parameters, the only parameter that can be tuned is the context dimension, how many units the RNN has. Tensorflow has a default value of 128 for the context dimension.

2.1.2 Long Short-Term Memory Neural Network

To construct our long short-term memory neural network, we again use the Keras API from Tensorflow. As we did with the RNN, we also follow the example that Tensorflow gives. We use a sequential model consisting of an LSTM layer that takes five inputs (consisting of the five characters of each of the romance languages), two dense layers: one to avoid narrowing of the signal from 128 to 10 in one step and another to represent the size of the character embedding.

2.1.3 Feedforward Neural Network

To construct our feedforward neural network, we use Keras. In contrast with our RNN and LSTM models, we manually construct this network instead of using an API. The network is constructed by creating five hidden dense layers (for each romance language), flattening the layers to create one all-encompassing layer, and then having a final dense layer as the output.

2.2 Data

2.2.1 Alphabets & self-compiled dataset

We compile cognate sets based on the Swadesh lists (Swadesh, 1971) for the five romance languages used in Ciobanu and Dinu (2014) (Italian, Spanish, French Portuguese & Romanian) using two different alphabets: The International Phonetic Alphabet (IPA) and the alphabet of the Automated Similarity Judgment Program (ASJP). The data source was mostly from the etymological section of the Wiktionary [1]. The ASJP alphabet comprises all ASCII characters and is used primarily for phylogenetic inference. The advantage of using the ASJP alphabet in addition to IPA is that wordlists containing 40 or 100 words from the Swadesh list are available for many language families, and likely the first data collected for any unknown language. Also, the feature set contained in ASJP is less fine-grained than that of IPA, describing sound classes rather than individual sounds.

$$\begin{pmatrix} k & a & p & u & - & - & \\ k & a & p & o & - & - & \\ k & a & b & e & 8 & a & \\ S & E & f & - & - & - & \\ k & a & b & u & - & - & \\ & k & a & p & - & - & - \end{pmatrix}$$

Figure 1: Aligned item 38 (HEAD) from Swadesh data. {-} are placeholder tokens indicating either loss in one or innovation in another daughter language.

We also add manually aligned cognate sets to our self-compiled data in order to investigate whether introduction of top-down knowledge, i. e. hypotheses about the relationships between the sounds in the daughter languages improve the quality of reconstructions. **??** shows the manual alignment for item 38 from the Swadesh list.

Since the feature values of the individual characters are not fixed, we use those given in (Brown et al., 2008). From these alphabets, we derive multi-hot embeddings based on the phonological features of a sound represented by a character of either alphabet, and one-hot character embeddings from the surface characters. That means that each character is once encoded as a vector of [0, 1] for each phonological feature, and once as one row in a identity matrix where rows and columns represent the characters of the alphabet.

Since Latin is not the direct ancestor of the romance languages, and none of the languages our dataset actually continues (e. g.) the Latin nominative in nouns or the passive inflection in verbs, we replace the Latin lemma form with either the stem of the accusative singular in nouns or the ”active” infinitive in verbs. This deviates from all prior approaches, but since it is well established that Classical Latin is not the direct ancestor of the romance language family (compare Dworkin (2016)), and most lemmas in the individual languages do not continue (e. g.) the Latin nominative or infinitive, we decide to use a form that more closely matches the actual ancestor of the romance words.

[1] https://www.wiktionary.org

	I (Swadesh list)	**II** (Ciobanu 2014)
Languages	Italian, Spanish, French, Portuguese, Romanian	
Ancestor	Latin	
Size	100	3218
Train set	80	2574
Test set	20	644
Aligned	yes	no
Alphabets	IPA, ASJP, Latin	

Table 1: Datasets

Therefore we replace Latin *pe:s* with a normalized accusative singular **pede*, and a passive infinitive as *mori:* with its active counterpart **mori:re*, which is the ancestor of Spanish *morir*, French *mourir*, etc. We think that this approach is valid since no romance language actually preserves the nominative form, which therefore should be impossible to reconstruct if only the romance data is known.

2.2.2 Dataset from Ciobanu and Dinu (2014)

We also test our model on a larger dataset. We choose the romance data provided by Ciobanu and Dinu (2014) because it was used by Meloni et al. (2019). One drawback of this dataset is that it consists mainly of direct loans from Latin, which in addition to the conservative spelling conventions makes it somewhat artificial compared to data that is collected from oral sources where no transcription is available. We want to circumvent this drawback by using our smaller dataset containing cognate sets that actually mirror the phonological developments leading to the modern languages. This is different from phylogenetic approaches where loans serve the goal of the reconstruction, i. e. reconstructing a family tree that mirrors language contact. Since we want to reconstruct proto-words and we opted to have only some aberrant items that represent either loans from Latin or lexical replacement.

As a baseline we also compile a Latin alphabet version of our dataset to test our models under the same conditions as in Meloni et al. (2019).

3 Results

We report edit distance percentiles, mean edit distance and the mean edit distance normalized for word length for each test condition. An edit distance of 0 means that the reconstructed word differed in 0 characters from the ground truth, i. e. a perfect reconstruction. We observe that generally speaking the feedforward model/model **B** yields slightly better results than the recurrent model/model **A**. However the effect is stronger in the data from Ciobanu and Dinu (2014), with the lowest average edit distance being 1.39 for the Latin with model **A** opposed to 1.6 with model **B**. Our self-compiled data set **I** proved harder for either model, with edit distances being on average one character higher than for dataset **II**. Differences in model performance for the two alphabets were small, with performance on the IPA alphabet being slightly worse than performance on the ASJP alphabet, in line with the results of Meloni et al. (2019). This effect is strongest for dataset **II**. Orthographic (one-hot) encodings did not influence model performance significantly, though the runs with the Latin alphabet on model **A** on dataset **II** produced the best overall results. Contrary to our intuition, there is no clear tendency as to whether manual alignments increase reconstruction accuracy. However, in the ASJP condition mean edit distances are smallest for the aligned data.

Our worst performing model was the RNN model. It has the highest amount of words with a Levenstein distance of five, meaning there were many reconstructed words that did not match the original proto word by any character. This is quite surprising because (Meloni et al., 2019) used an RNN model as well, but with attention. Compared to (Meloni et al., 2019), we have quite opposite results. For example, with (Meloni et al., 2019)'s model on the IPA alphabet, the number of reconstructed words that have an edit distance up to two characters is 85.9%. In our model, there is no reconstructed word with an edit distance of less than three characters.

4 Discussion

The choice of alphabet seems to have the largest impact on reconstruction accuracy. Similar to Mel-

	A	**B**
Architecture	Recurrent	Deep feedforward
Hidden layer	LSTM layer	2 dense layers
Size	128	2 × 256

Table 2: Models

Model			**Edit Distance**					**Mean**	**Norm**
			≤ 0	≤ 1	≤ 2	≤ 3	≤ 4		
Features	**IPA**	**A**	23%	49%	67%	83%	90%	1.85	0.21
		B	29%	54%	69%	83%	90%	1.72	0.2
	ASJP	**A**	25%	50%	66%	81%	90%	1.85	0.22
		B	30%	53%	69%	83%	91%	1.71	0.2
Orthographic	**IPA**	**A**	26%	49%	70%	83%	91%	1.79	0.21
		B	30%	54%	71%	83%	92%	1.68	0.2
	ASJP	**A**	29%	52%	69%	85%	91%	1.71	0.2
		B	28%	53%	70%	84%	91%	1.71	0.2
Latin		**A**	27%	56%	74%	87%	93%	1.6	0.19
		B	34%	63%	79%	89%	92%	1.39	0.16

Table 3: Performance of both models on the test set of dataset **II**

	IPA		**ASJP**		**Latin**
Edit Distance	Vanilla	Orthographic	Vanilla	Orthographic	
≤ 0	29%	30%	30%	28%	34%
≤ 1	54%	54%	53%	53%	63%
≤ 2	69%	71%	69%	70%	79%
≤ 3	83%	83%	83%	84%	89%
≤ 4	90%	92%	91%	91%	92%
Mean	1.72	1.68	1.71	1.71	1.39
Mean, norm	0.2	0.2	0.2	0.2	0.16

Table 4: Performance on the test set for data from Ciobanu and Dinu (2014) & model **B**, trained for 10 epochs

	IPA			**ASJP**			**Latin**
Edit Distance	Vanilla	Aligned	Orthographic	Vanilla	Aligned	Orthographic	
≤ 0	12%	12%	29%	21%	16%	31%	20%
≤ 1	35%	28%	41%	55%	38%	52%	37%
≤ 2	59%	46%	65%	73%	57%	72%	66%
≤ 3	76%	73%	78%	87%	76%	85%	84%
≤ 4	87%	81%	92%	92%	90%	93%	87%
Mean	2.31	2.6	2.05	1.72	2.22	1.67	2.07
Mean, norm	0.37	0.44	0.32	0.3	0.42	0.29	0.36

Table 5: Performance on the test set for Swadesh list data & model **A**, trained for 10 epochs

	IPA			**ASJP**			**Latin**
Edit Distance	Vanilla	Aligned	Orthographic	Vanilla	Aligned	Orthographic	
≤ 0	15%	9%	14%	28%	21%	30%	18%
≤ 1	41%	22%	33%	54%	43%	51%	39%
≤ 2	61%	48%	62%	76%	59%	77%	62%
≤ 3	78%	69%	79%	89%	81%	90%	86%
≤ 4	84%	84%	91%	94%	92%	94%	94%
Mean	2.21	2.68	2.21	1.59	2.04	1.58	2.01
Mean, norm	0.35	0.45	0.35	0.27	0.39	0.28	0.34

Table 6: Performance on the test set for Swadesh list data & model **B**, trained for 10 epochs

oni et al. (2019) both models performed worse on the IPA transcriptions across both datasets. One reason for this might be that there are more features encoded in the IPA than in the ASJP. We assume that the overall better performance on the Latin orthographic data is because of the etymological information contained in, e. g., the French orthography. The overall higher reconstruction accuracy on dataset **II** may be caused by its larger size (3128 items versus 100 in dataset **I**) and it containing a large portion of direct loans from Latin. To our surprise both models fared almost equally well with either dataset, with a small advantage for the feed-forward model (**B**) observable with dataset **II**. This result should be more reliable because of the larger size of dataset **II**. This indicates that the higher accuracy scores in both Ciobanu and Dinu (2018) and Meloni et al. (2019) are indeed due to the decoder-encoder architecture with attention, which makes the machine translation pipeline a plausible choice for the task of proto-word reconstruction. However, the poor performance of both our models on the Swadesh list data in dataset **II** demonstrates that this approach relies on a high number of cognate sets being available to the researcher, which cannot be expected to be the case for language families for which no literary records and therefore no direct information about the proto language is present, e. g. the Papuan clades.

5 Conclusion

In this paper we reproduced the result of prior studies that more fine-grained encodings presents a challenge for recontruction models. Our results also indicate that a machine-translation-like decoder-encoder architecture may be necessary to achieve high accruracy scores for datasets comparable to Ciobanu and Dinu (2014). On the other hand we assume that the poor model performance on our self-compiled dataset constitutes an inherent problem to the automated reconstruction of proto-words: if large cognate datasets are required to achieve an acceptable level of reconstruction accuracy, relying on small datasets to automatically reconstruct proto-words may be impossible.

6 Future Work

As stated in our conclusion, we want to know if large cognate datasets are required to achieve an acceptable level of reconstruction accuracy. In future work, we would like to apply the architecture used by Meloni et al. (2019) on our dataset to verify our hypothesis. If our hypothesis is correct, then relying on small datasets to automatically reconstruct proto-words may be impossible.

References

Martín Abadi, Ashish Agarwal, Paul Barham, Eugene Brevdo, Zhifeng Chen, Craig Citro, Greg S. Corrado, Andy Davis, Jeffrey Dean, Matthieu Devin, Sanjay Ghemawat, Ian Goodfellow, Andrew Harp, Geoffrey Irving, Michael Isard, Yangqing Jia, Rafal Jozefowicz, Lukasz Kaiser, Manjunath Kudlur, Josh Levenberg, Dandelion Mané, Rajat Monga, Sherry Moore, Derek Murray, Chris Olah, Mike Schuster, Jonathon Shlens, Benoit Steiner, Ilya Sutskever, Kunal Talwar, Paul Tucker, Vincent Vanhoucke, Vijay Vasudevan, Fernanda Viégas, Oriol Vinyals, Pete Warden, Martin Wattenberg, Martin Wicke, Yuan Yu, and Xiaoqiang Zheng. 2015. TensorFlow: Large-scale machine learning on heterogeneous systems. Software available from tensorflow.org.

Alexandre Bouchard-Côté, David Hall, Thomas L. Griffiths, and Dan Klein. 2013. Automated reconstruction of ancient languages using probabilistic models of sound change. *Proceedings of the National Academy of Sciences*, 110(11):4224–4229.

Cecil H. Brown, Eric W. Holman, Soeren Wichmann, and Viveka Velupillai. 2008. Automated classification of the world's languages: a description of the method and preliminary results. *Language Typology and Universals*, 61(4):285–308.

François Chollet. 2015. Keras. `https://github.com/fchollet/keras`.

Alina Maria Ciobanu and Liviu P. Dinu. 2014. Automatic detection of cognates using orthographic alignment. In *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 99–105, Baltimore, Maryland. Association for Computational Linguistics.

Alina Maria Ciobanu and Liviu P. Dinu. 2018. Ab initio: Automatic Latin proto-word reconstruction. In *Proceedings of the 27th International Conference on Computational Linguistics*, pages 1604–1614, Santa Fe, New Mexico, USA. Association for Computational Linguistics.

Peter Dekker. 2018. MSc thesis: Reconstructing language ancestry by performing word prediction with neural networks.

Steven N. Dworkin. 2016. Do romanists need to reconstruct proto-romance? *Zeitschrift für romanische Philologie*, 132(1):1–19.

Daniel J. Hruschka, Simon Branford, Eric D. Smith, Jon Wilkins, Andrew Meade, Mark Pagel, and Tanmoy Bhattacharya. 2015. Detecting Regular Sound Changes in Linguistics as Events of Concerted Evolution. *Current Biology*, 25(1):1–9.

Carlo Meloni, Shauli Ravfogel, and Yoav Goldberg. 2019. Ab antiquo: Proto-language reconstruction with rnns. *arXiv preprint arXiv:1908.02477*.

Morris Swadesh. 1971. *The origin and diversification of language*. Aldine, Atherton, Chicago.