Transformed Protoform Reconstruction

Young Min Kim[*] and **Kalvin Chang**[*] and **Chenxuan Cui** and **David Mortensen**
Language Technologies Institute, Carnegie Mellon University
{youngmik, kalvinc, cxcui, dmortens}@cs.cmu.edu

Abstract

Protoform reconstruction is the task of inferring how morphemes or words sounded in ancestral languages of a set of daughter languages. Meloni et al. (2021) achieved the state-of-the-art on Latin protoform reconstruction with an RNN-based encoder-decoder with attention model. We update their model with the state-of-the-art seq2seq model—the Transformer. Our model outperforms their model on a suite of different metrics on two different datasets: Meloni et al.'s Romance data of 8,000+ cognates (spanning 5 languages) and a Chinese dataset (Hóu, 2004) of 800+ cognates (spanning 39 varieties). We also probe our model for potential phylogenetic signal contained in the model. Our code is publicly available [1].

1 Introduction

Languages change over time and sometimes diverge into multiple daughter languages. The common ancestor of a set of genetically related languages is their proto-language. While there are proto-languages such as Latin that are attested, they are the exception[2]. Reconstructed words and morphemes in proto-languages are called protoforms. The task of reconstructing unattested proto-languages is called protoform reconstruction.

Historical linguists reconstruct proto-languages by identifying systematic sound changes that can be inferred from correspondences between attested daughter languages (see Table 1). They compare the sounds between a set of cognates, or words with a common ancestor, to develop hypotheses about the types and chronologies of sound changes.

	'tooth'	'two'	'ten'	
English	**t**ooth	**t**wo	**t**en	**t**
Dutch	**t**and	**t**wee	**t**ien	**t**
German	**Z**ahn	**z**wei	**z**ehn	**z**
PWG	***t**anþ	***t**wai-	***t**ehun	*__t__

Table 1: Sound correspondences in West Germanic Languages and Proto-West-Germanic (PWG).

This task is inherently data-constrained, especially for under-documented languages. Such data scarcity makes it a particularly difficult task for contemporary neural network architectures such as the Transformer (Vaswani et al., 2017), which are data hungry.

The contributions of this paper are as follows:

- Application of the Transformer architecture to the protoform reconstruction task, achieving state of the art performance, contrary to expectation.

- Expansion of prior digital versions of Hóu (2004)'s Chinese dataset to include a total of 804 cognate sets across 39 modern varieties and Middle Chinese.

2 Related Work

Applying machine learning to protoform reconstruction is not new. Bouchard-Côté et al. (2013) learn an unsupervised protoform reconstruction model for the large Oceanic language family using Monte Carlo Expectation Maximization (Dempster et al., 1977; Bouchard-Côté et al., 2008), supervising the model with a gold phylogeny and using a probabilistic, generative model of sound change. He et al. (2022) modernize an earlier version of Bouchard-Côté et al. (2013)'s model with RNNs for a 4 language subset of Romance, but they rely on a bigram language model of Latin, making their model technically not unsupervised.

---

[*] Equal contribution
[1] https://github.com/cmu-llab/acl-2023
[2] In fact, the proto-language from which Romance languages like Spanish and Italian are descended is not identical to Classical Latin but is, rather, a closely related and sparsely attested language sometimes called Proto-Romance or Vulgar Latin.

List et al. (2022) apply an SVM classifier to supervised reconstruction by treating sound correspondences as training examples. Note that there were no word boundaries in the input matrix; that is, all sound correspondences across the training set are flattened into one matrix. Furthermore, each language has an independent phonemic inventory. To learn contextual information, the authors experiment with adding features encoding the position of phonemes, among others.

Ciobanu and Dinu (2018) learn a conditional random field (Lafferty et al., 2001) using n-gram features for supervised reconstruction and ensemble 5 daughter-to-protoform models. They use a dataset of 3,218 complete cognate sets spanning Latin (the proto-language) and 5 Romance languages: Romanian, French, Italian, Spanish, Portuguese.

Meloni et al. (2021) employ a GRU-based seq2seq approach (Cho et al., 2014) to Latin protoform reconstruction and achieve state-of-the-art character edit distances. They extend Dinu and Ciobanu (2014)'s Romance data using data from Wiktionary—for a total of 8,799 cognate sets across 5 Romance languages plus Latin—in both orthographic and phonetic (IPA) representations. In their model, all entries comprising the cognate set are concatenated together in a fixed order to form a training example. Chang et al. (2022) applied Meloni et al. (2021)'s architecture to the reconstruction of Middle Chinese on a dataset of 5000+ cognate sets spanning 8 languages they compiled from Wiktionary. [3]

Fourrier (2022) compares statistical machine translation, RNN, and Transformer architectures for protoform reconstruction, but they evaluate their results using BLEU scores (Papineni et al., 2002) instead of edit distance. They find that their Transformer model did not outperform the RNN models on protoform reconstruction. In addition, their multilingual NMT (neural machine translation) model predicts many languages instead of one target language and is trained on bilingual pairs for protoform reconstruction (e.g. Italian-Latin and Spanish-Latin), unlike comparative reconstruction. In contrast, we encode the entire cognate set consisting of multiple daughter languages (5 for the Romance dataset; 39 for Chinese) and predict the corresponding protoform.

3 Datasets

We train and test our model on Romance and Sinitic (Chinese) language datasets. For Romance languages, we use Meloni et al. (2021)'s dataset which consists of 8,799 cognate sets of Romanian, French, Italian, Spanish, Portuguese words and the corresponding Latin form (approximately, a protoform). There are two versions of this dataset: phonetic and orthographic. The phonetic dataset (Rom-phon) represents words with IPA symbols whereas the orthographic dataset (Rom-orth) represents words in the orthographic form of each language. We preserved all diacritics, except for vowel length. This dataset is an extension of Dinu and Ciobanu (2014)'s original dataset of 3,218 cognate sets, which is not publicly available. Refer to Table 2 for more information.

3.1 Expanding digital versions of Hóu (2004)

For Sinitic languages, we created a dataset of Middle Chinese and its modern daughter languages. Middle Chinese is an unattested language, and we thus have to rely on Baxter and Sagart (2014)'s reconstructions of forms corresponding to 4,967 Chinese characters. We scraped Wiktionary to obtain Hóu (2004)'s phonetic representations of their modern reflexes. [4] The resulting dataset contains 804 cognate sets of 39 modern Sinitic languages and the corresponding reconstructed Middle Chinese word. List (2021)'s version previously had 894 cognate sets across 15 varieties.

4 Model

We propose a Transformer-based encoder-decoder architecture (Vaswani et al., 2017) because such models have produced state-of-the-art results on many sequence processing tasks. Transformers are by reputation data hungry, though, which poses a challenge to our problem setting, where the number of available training examples is often very small.

---

[3] The original dataset contains 21,000 cognate sets, but only 5000+ had at least 3 daughter entries and were used as input to the model.

[4] https://en.wiktionary.org/wiki/Module: zh/data/dial-pron/documentation originally had 1,023 characters, but only 804 had reconstructions from Baxter and Sagart (2014).
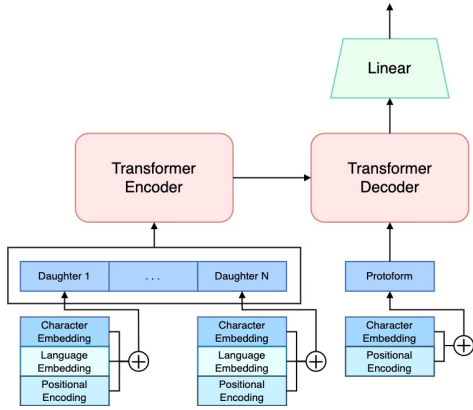
Figure 1: Diagram of our encoder-decoder architecture. Additive positional encoding and language embedding are applied to each daughter sequence before all daughter sequences are concatenated into a single sequence.

We modify the standard encoder-decoder architecture to accommodate the structure of our datasets, where multiple daughter sequences correspond to a single protoform sequence. Like Meloni et al. (2021), the daughter sequences are concatenated into a single sequence before being fed into the encoder. Because we only care about the relative position between tokens within each daughter sequence but not across daughter sequences, positional encoding is applied to each individual daughter sequence before concatenation. Along with positional encoding, an additive language embedding is applied to the token embeddings to differentiate between input tokens of different daughter languages.

5 Experiments

5.1 Baselines

We compare our Transformer model to a variety of baselines. For Meloni et al. (2021), we use Chang et al. (2022)'s PyTorch re-implementation and reran a Bayesian hyperparameter search using WandB (Biewald, 2020) to ensure a more fair comparison (since our model is tuned with WandB as well). We also include the random daughter (randomly designate a daughter form as the protoform and assume no sound change) and the majority constituent baselines (predict the most common phoneme in each syllable constituent) from Chang et al. (2022). For the SVM and CoRPaR classifiers (List et al., 2022), we experiment with different contextual features, such as Pos (position), Str (prosodic structure), and Ini (whether or not the phoneme appears word-initially or word-finally).

We publish results on Meloni et al. (2021)'s full set of 8,799 cognates but cannot redistribute this set due to Dinu and Ciobanu (2014)'s restrictions. For reproducibility, we include results on Meloni et al. (2021)'s public subset of 5,419 cognates in the Appendix (Table 7), both of which include vowel length. Observe that these results are worse than those obtained on the full set, suggesting that the RNN and Transformer are dependent on a wealth of training data.

5.2 Preprocessing

In all our datasets, we merge diacritics to their base segments to form a multi-character token. For instance, the sequence [t, ʰ] is concatenated to [tʰ]. This ensures that phonemes are treated as one token. For Chinese, tone contours (a sequence of tones) are treated as one token. When multiple pronunciation variants are listed for a single Chinese character, we arbitrarily pick the first one.

6 Results and Discussion

6.1 Evaluation criteria

We evaluate the predicted protoforms using edit distance (Levenshtein et al., 1966), normalized edit distance (edit distance normalized by the length of the target) and accuracy (the percentage of protoforms that are reconstructed without any mistakes). Like Chang et al. (2022), we also use feature error rate calculated using articulatory feature vectors from PanPhon (Mortensen et al., 2016) because it reflects the phonetic similarity between the prediction and the gold protoform. For datasets with phonetic transcriptions (Romance-phonetic and Chinese), we use phoneme edit distance and normalized phoneme edit distance. As List (2019) suggests, we use B-Cubed F Scores (Amigó et al., 2009) to capture the structural similarity between the gold and predicted protoforms (0: structurally dissimilar, 1: similar). With the exception of character and phoneme edit distance, the metrics enable fair comparison across different language families, which will differ in the average word length.

6.2 Results

Table 3 shows that our model consistently has the best performance on all datasets with regards to most metrics. The results were averaged across 5 runs. Out of all datasets, our model performs best on the Rom-orth dataset, where we achieve a 7.0%

Language Family	Source	# varieties	Cognate sets	Proto-language
Rom-phon	Dinu and Ciobanu (2014), Meloni et al. (2021)	5	8,799	Latin
Rom-orth	Dinu and Ciobanu (2014), Meloni et al. (2021)	5	8,799	Latin
Sinitic (Chinese)	Hóu (2004)	39	804	Middle Chinese

Table 2: Statistics on both datasets used in our experiments. # varieties refers to the number of daughter varieties.
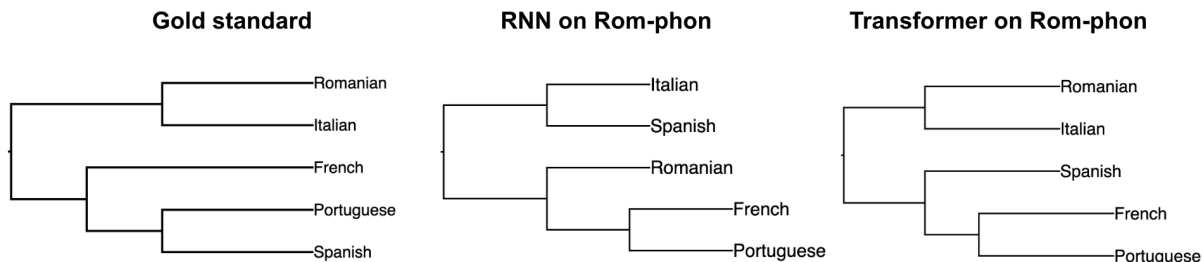

Figure 2: A gold phylogeny of Romance (left) compared with those derived by probing the RNN model (middle) and the Transformer model (right) on Rom-phon.

decrease in phoneme edit distance and a 1.43p.p improvement in accuracy relative to the RNN baseline. We observe the most dramatic performance difference with the RNN baseline on the Sinitic dataset: a 10.48% decrease in phoneme edit distance and a 5.47p.p increase in accuracy. For reproducibility, results on the publicly available portion of the Rom-phon and Rom-orth datasets are provided in Table 7 in the Appendix.

6.3 Analysis

We observe that the BCFS is relatively high for the Romance non-neural baselines compared to those of the Chinese ones. This suggests that the sound changes in the Romance datasets are more regular than that of Chinese, which corroborates List et al. (2014)'s results that more than half of the Chinese characters in their dataset could not be explained by a tree model.

We examine the errors made by the Transformer model on the Rom-phon datasest. Substitutions constitute around 61% of the errors made by the Transformer; deletions, 21%, and insertions, 18%. The highest number of substitution errors occur between [i, ɪ], [e, ɛ], [o, ɔ] and [u, ʊ]—vowel pairs that contrast only in tenseness. This is consistent with the analysis of Meloni et al. (2021), where substitutions between tense-lax vowel pairs take up the largest portion of errors.

We observe that other common substitution errors also happen between phonemes that share major phonetic features. This demonstrates that al-

though no explicit phonetic information is fed directly into the model, the model makes mistakes motivated by phonetic similarity, like Meloni et al. (2021).

We do not observe notable differences in the error statistics between the Transformer and the RNN.

6.4 Language relatedness

Inspired by Fourrier (2022), we probe our model for diachronic information on how genetically related each Romance language is to each other. We create a distance matrix between every pair of languages in a dataset by taking the cosine similarity between a pair's language embeddings. We then use sklearn (Pedregosa et al., 2011)'s implementation of the Ward variance minimization algorithm (Ward Jr, 1963) to perform hierarchical clustering on the distance matrix. We take a consensus of the dendrograms from 5 different runs using the consense program from PHYLIP (Felsenstein, 2013).

As we see in Figure 2, the Transformer captures more of the phylogenetic relationships among the languages correctly for the Rom-phon dataset. Indeed, the Generalized Quartet Distance (GQD) (Sand et al., 2013; Pompei et al., 2011; Rama et al., 2018) between the gold and predicted tree, calculated using quartetDist from the tqDist library (Sand et al., 2014), is 0.4 for the Transformer but 0.8 for the RNN. See Figure 5 in the Appendix for the results of the orthographic dataset.

Dataset	Model	PED ↓	NPED ↓	Acc % ↑	FER ↓	BCFS ↑
Sinitic	Random daughter (Chang et al., 2022)	3.7702	0.8405	0%	0.2893	0.2748
	Majority constituent (Chang et al., 2022)	3.5031	0.7806	0%	0.2013	0.3695
	CorPaR (List et al., 2022)	3.2795	0.7278	0%	0.3972	0.3332
	SVM + PosStr (List et al., 2022)	1.6894	0.3692	15.52%	0.1669	0.5418
	RNN (Meloni et al., 2021)	1.0671	0.2421	35.65%	0.0899	0.6781
	Transformer (present work)	**0.9553**	**0.2150**	**41.12%**	**0.0842**	**0.7033**
Rom-phon	Random daughter (Chang et al., 2022)	6.1534	0.6914	0.06%	0.6264	0.4016
	CorPaR + PosIni (List et al., 2022)	1.6847	0.1978	22.18%	0.0728	0.7403
	SVM + PosStrIni (List et al., 2022)	1.5787	0.1861	24.69%	0.0713	0.7610
	RNN (Meloni et al., 2021)	0.9655	0.1224	52.31%	0.0384	0.8296
	Transformer (present work)	**0.8926**	**0.1137**	**53.75%**	**0.0373**	**0.8435**
Rom-orth	Random daughter (Chang et al., 2022)	4.2567	0.4854	2.97%	-	0.5147
	CorPaR + Ini (List et al., 2022)	0.9531	0.1160	47.23%	-	0.8400
	SVM + PosStr (List et al., 2022)	0.8988	0.1105	50.43%	-	0.8501
	RNN (Meloni et al., 2021)	0.5941	0.0770	69.80%	-	0.8916
	Transformer (present work)	**0.5525**	**0.0720**	**71.23%**	-	**0.9002**

Table 3: Evaluation of models and baselines using various metrics, averaged across 5 runs (same hyperparameters, different seeds). Because Rom-orth is not in IPA, character edit distance is used instead of PED, and we cannot accurately calculate FER. See Section 6.1 for an explanation of each evaluation metric. See Table 4 for the standard deviation values.

Since the Romance dataset only includes 5 daughter languages, our results are insufficient to corroborate or contradict Cathcart and Wandl (2020)'s findings: the more accurate the protoforms, the less accurate the phylogeny will be. It is not clear if the model's language embeddings are learning information that reflects shared innovations (sound changes that if shared among a set of daughter languages, would be acceptable justification for grouping them)—the only acceptable criterion for phylogenetic inference in historical linguistics (Campbell, 2013)—or if the model is learning superficial phonetic similarity.

7 Conclusion

By showing that Transformers can outperform previous architectures in protoform reconstruction despite the inherent data scarcity of the task, our work motivates future research in this area to take full advantage of the recent advancements in the Transformer space.

Accurate supervised reconstruction can help pre-

dict protoforms for cognate sets where linguists have not reconstructed one yet. Future work could reconstruct proto-languages whose linguist reconstructions are not available, by transferring knowledge learned from languages with already reconstructed protoforms. Furthermore, future work can leverage the abundance of work in unsupervised NMT to adapt our Transformer model for the unsupervised setting, a more realistic scenario for the historical linguist.

Limitations

One limitation of our work is that the RNN (Meloni et al., 2021) actually outperforms our Transformer on the Chinese dataset in Chang et al. (2022). In addition, as with other neural approaches, our model requires significant amounts of data, which is often not available to historical linguists researching less well-studied language families based on field reports. Romance and Chinese have relatively many cognate sets because the protoforms

are documented[5], but a low resource setup with 200 cognate sets would not fare well on our data-hungrier Transformer model. Furthermore, concatenating the entire cognate set may not work on language families with hundreds of languages such as Oceanic because the input sequence would be too long compared to the output protoform sequence.

Finally, we obtain our Chinese gold protoforms from Baxter and Sagart (2014)'s Middle Chinese reconstruction, which was actually a transcription of the *Qieyun*, a rhyme dictionary. Norman and Coblin (1995) disagree with relying on such a philological source and prefer comparative reconstructions that begin from daughter data. However, there is no available comparative reconstruction of Middle Chinese with protoforms corresponding to thousands of characters to use as a gold standard. Be that as it may, it seems clear that Middle Chinese as recorded in the *Qieyun* is not identical to the most recent ancestor of the Chinese languages. Its preface concedes that it is a compromise between Tang Dynasty dialects. The situation with Romance is, in some ways, comparable. Classical Latin—the variety on which we train— is not the direct ancestor of modern Romance languages. Instead, they are descended from Vulgar Latin or Proto-Romance, which is not well-attested and is primarily through graffiti and other informal inscriptions. Proto-Romance reconstructions are also not exhaustive. As a result, it is difficult to find a dataset like Meloni et al. (2021) with thousands of such ancestor forms. We are also limited to the faithfulness of espeak-ng's Latin G2P, from which Meloni et al. (2021) obtain their phonetic Romance dataset.

For most language families, protoforms are not attested. In fact, as the term is often used, protoform refers to a form that is inferred only through linguists' comparative method. We adopt the other usage for simplicity. In practice, our approach would require reconstructions made by a linguist to serve as training labels for cognate sets.

Acknowledgements

References

Enrique Amigó, Julio Gonzalo, Javier Artiles, and Felisa Verdejo. 2009. A comparison of extrinsic clustering evaluation metrics based on formal constraints. *Information retrieval*, 12(4):461–486.

William H Baxter and Laurent Sagart. 2014. *Old Chinese: A new reconstruction*. Oxford University Press.

Lukas Biewald. 2020. Experiment tracking with weights and biases. Software available from wandb.com.

Alexandre Bouchard-Côté, Dan Klein, and Michael Jordan. 2008. Efficient inference in phylogenetic indel trees. In *Advances in Neural Information Processing Systems*, volume 21. Curran Associates, Inc.

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.

Lyle Campbell. 2013. *Historical Linguistics: an Introduction*. Edinburgh University Press.

Chundra Cathcart and Florian Wandl. 2020. In search of isoglosses: continuous and discrete language embeddings in Slavic historical phonology. In *Proceedings of the 17th SIGMORPHON Workshop on Computational Research in Phonetics, Phonology, and Morphology*, pages 233–244, Online. Association for Computational Linguistics.

Kalvin Chang, Chenxuan Cui, Youngmin Kim, and David R. Mortensen. 2022. WikiHan: A new comparative dataset for Chinese languages. In *Proceedings of the 29th International Conference on Computational Linguistics (COLING 2022)*.

Kyunghyun Cho, Bart van Merriënboer, Dzmitry Bahdanau, and Yoshua Bengio. 2014. On the properties of neural machine translation: Encoder–decoder approaches. In *Proceedings of SSST-8, Eighth Workshop on Syntax, Semantics and Structure in Statistical Translation*, pages 103–111, Doha, Qatar. 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.

Arthur P Dempster, Nan M Laird, and Donald B Rubin. 1977. Maximum likelihood from incomplete data via the em algorithm. *Journal of the Royal Statistical Society: Series B (Methodological)*, 39(1):1–22.

Liviu Dinu and Alina Maria Ciobanu. 2014. Building a dataset of multilingual cognates for the Romanian lexicon. In *Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC'14)*, pages 1038–1043, Reykjavik, Iceland. European Language Resources Association (ELRA).

Joseph Felsenstein. 2013. Phylip (phylogeny inference package), version 3.695. Department of Genome Sciences, University of Washington, Seattle.

Clémentine Fourrier. 2022. *Neural Approaches to Historical Word Reconstruction*. Ph.D. thesis, Université PSL (Paris Sciences & Lettres).

Andre He, Nicholas Tomlin, and Dan Klein. 2022. Neural unsupervised reconstruction of protolanguage word forms. *arXiv preprint arXiv:2211.08684*.

侯精一 Jīngyī Hóu, editor. 2004. *Xiàndài Hànyǔ fāngyán yīnkù* 现代汉语方言音库 *[Phonological database of Chinese dialects]*. Shànghǎi Jiàoyù 上海教育, Shànghǎi 上海.

John D. Lafferty, Andrew McCallum, and Fernando C. N. Pereira. 2001. Conditional random fields: Probabilistic models for segmenting and labeling sequence data. In *Proceedings of the Eighteenth International Conference on Machine Learning*, ICML '01, page 282–289, San Francisco, CA, USA. Morgan Kaufmann Publishers Inc.

Vladimir I Levenshtein et al. 1966. Binary codes capable of correcting deletions, insertions, and reversals. *Soviet physics doklady*, 10(8):707–710.

Johann-Mattis List. 2019. Beyond edit distances: Comparing linguistic reconstruction systems. *Theoretical Linguistics*, 45(3-4):247–258.

Johann-Mattis List. 2021. *CLDF dataset derived from Hóu's "Phonological Database of Chinese Dialects" from 2004*. Zenodo.

Johann-Mattis List, Robert Forkel, and Nathan Hill. 2022. A new framework for fast automated phonological reconstruction using trimmed alignments and sound correspondence patterns. In *Proceedings of the 3rd Workshop on Computational Approaches to Historical Language Change*, pages 89–96, Dublin, Ireland. Association for Computational Linguistics.

Johann-Mattis List, Nelson-Sathi Shijulal, William Martin, and Hans Geisler. 2014. Using phylogenetic networks to model chinese dialect history. *Language Dynamics and Change*, 4(2):222–252.

Carlo Meloni, Shauli Ravfogel, and Yoav Goldberg. 2021. Ab antiquo: Neural proto-language reconstruction. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 4460–4473, Online. Association for Computational Linguistics.

David R. Mortensen, Patrick Littell, Akash Bharadwaj, Kartik Goyal, Chris Dyer, and Lori S. Levin. 2016. Panphon: A resource for mapping IPA segments to articulatory feature vectors. In *Proceedings of COLING 2016, the 26th International Conference on Computational Linguistics: Technical Papers*, pages 3475–3484.

Jerry L. Norman and W. South Coblin. 1995. A new approach to Chinese historical linguistics. *Journal of the American Oriental Society*, 115(4):576–584.

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. Bleu: a method for automatic evaluation of machine translation. In *Proceedings of the 40th annual meeting of the Association for Computational Linguistics*, pages 311–318.

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

Simone Pompei, Vittorio Loreto, and Francesca Tria. 2011. On the accuracy of language trees. *PloS one*, 6(6):e20109.

Taraka Rama, Johann-Mattis List, Johannes Wahle, and Gerhard Jäger. 2018. Are automatic methods for cognate detection good enough for phylogenetic reconstruction in historical linguistics? *arXiv preprint arXiv:1804.05416*.

Andreas Sand, Morten K Holt, Jens Johansen, Rolf Fagerberg, Gerth Stølting Brodal, Christian NS Pedersen, and Thomas Mailund. 2013. Algorithms for computing the triplet and quartet distances for binary general trees. *Biology*, 2(4):1189–1209.

Andreas Sand, Morten Kragelund Holt, Jens Johansen, Rolf Fagerberg, Gerth Stølting Brodal, Thomas Mailund, and Christian N. S. Pedersen. 2014. tqdist: A library for computing the quartet and triplet distances between binary or general trees. *BMC Bioinformatics*, yy(xx):ii–jj.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. *Advances in neural information processing systems*, 30.

Joe H Ward Jr. 1963. Hierarchical grouping to optimize an objective function. *Journal of the American statistical association*, 58(301):236–244.

A Training

We split 70%, 10%, and 20% of our dataset into train, validation, and test sets, respectively. We conduct hyperparameter searches using WandB (Biewald, 2020) and use early stopping, picking the epoch with lowest edit distance on validation data. All experiments are performed on a Ubuntu server with 4 GPUs and 20 CPUs. For both the RNN and the Transformer, Meloni et al. (2021)'s dataset takes less than 7 GPU hours to run, while Hóu (2004) takes less than 1 GPU hour. For the large Romance orthographic dataset, the RNN model has around 480,000 parameters, while the Transformer has around 800,000 parameters.

B Hyper-parameters

Refer to Table 5 and Table 6 for the best hyperparameters we found during hyperparameter search via WandB.

C Supplementary Results

In order to compare our model to earlier work, we used the Rom-phon and Rom-orth datasets from Meloni et al. (2021). However, this set includes a subset from Ciobanu and Dinu (2018) which is not freely redistributable. So that our results can be reproduced, we also computed them on the publicly available subset of Meloni et al. (2021)'s dataset, which is presented in Table 7.

Phylogenetic trees for Chinese were also extracted from the RNN and Transformer models. These are shown in Figures 3 and 4.

We also plot the dendrograms derived from the Rom-orto dataset in Figure 5.
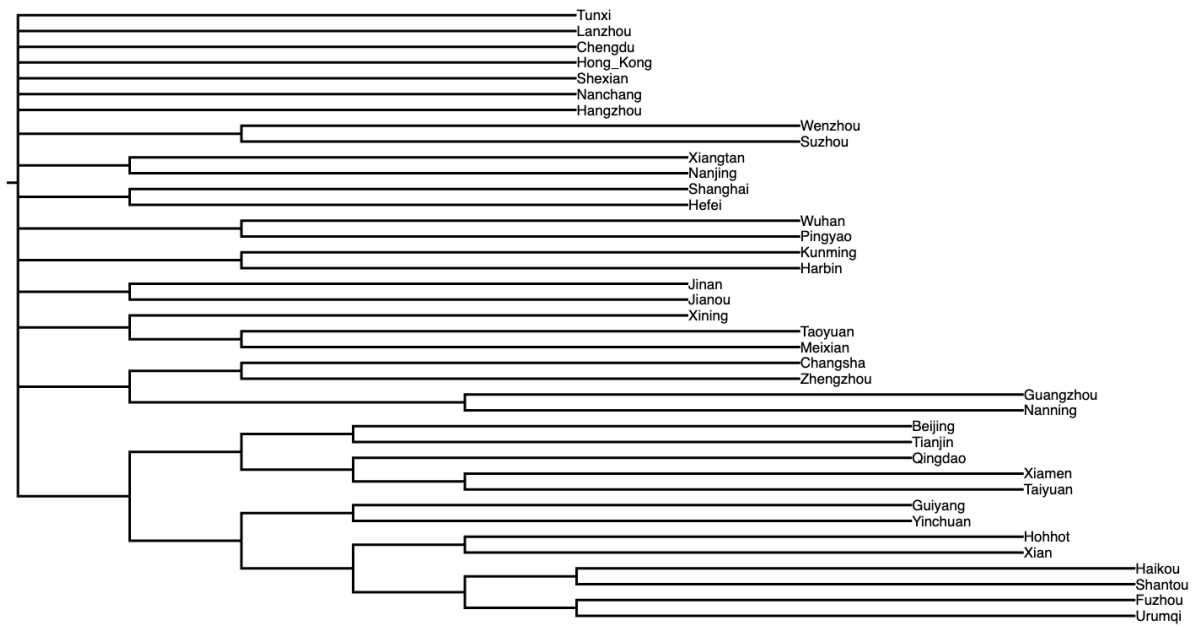
Figure 3: Consensus tree of the dendrograms from the 5 runs of the Transformer for the Chinese dataset
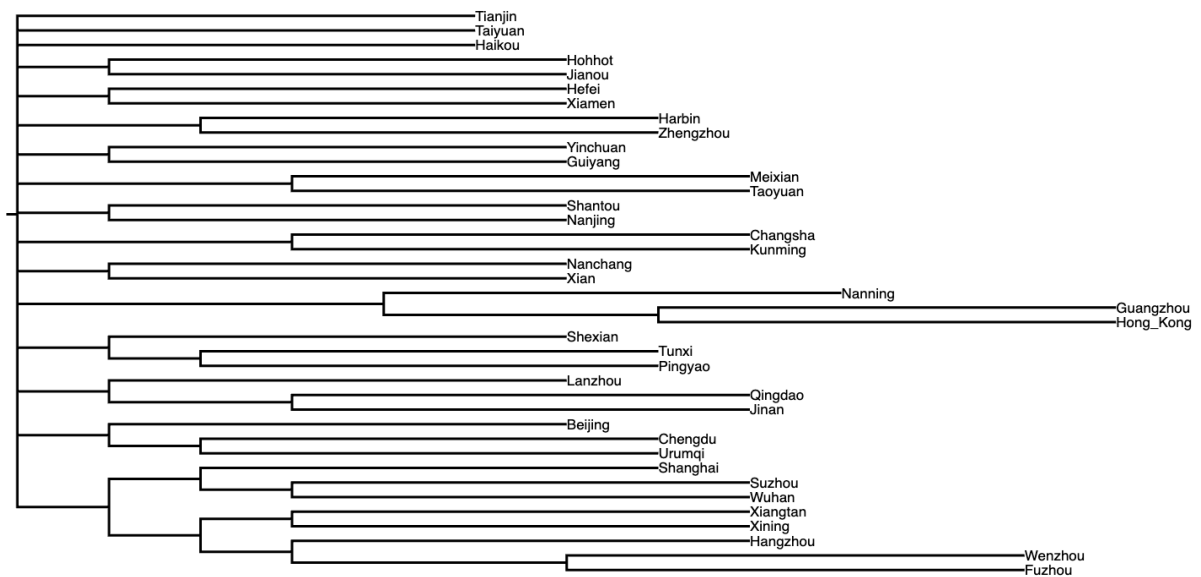

Figure 4: Consensus tree of the dendrograms from the 5 runs of the RNN for the Chinese dataset
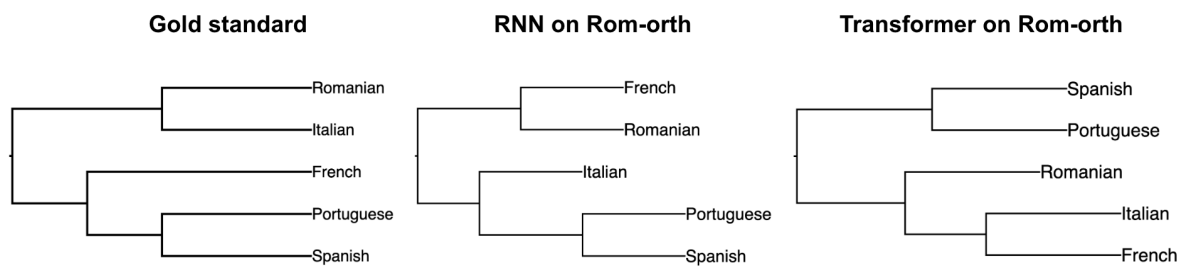
Figure 5: A gold phylogeny of Romance (left) compared with those derived by probing the RNN model (middle) and the Transformer model (right) on Rom-orto. GQD is 0.4 for both models.

Dataset	Model	PED ↓		NPED ↓		Acc % ↑		FER ↓		BCFS ↑	
Sinitic	Random daughter	3.7702		0.8405		0%		0.2893		0.2748	
	Majority constituent	3.5031		0.7806		0%		0.2013		0.3695	
	CorPaR	3.2795		0.7278		0%		0.3972		0.3332	
	SVM +PosStr	1.6894		0.3692		15.52%		0.1669		0.5418	
	RNN	1.0671 ± 0.0619		0.2421 ± 0.0140		35.65% ± 1.60%		0.0899 ± 0.0048		0.6781 ± 0.0174	
	Transformer (present work)	**0.9553** ± **0.0392**		**0.2150** ± **0.0075**		**41.12%** ± **2.3%**		**0.0842** ± **0.0070**		**0.7033** ± **0.0087**	
Rom-phon	Random daughter	6.1534		0.6914		0.06%		0.6264		0.4016	
	CorPaR +PosIni	1.6847		0.1978		22.18%		0.0728		0.7403	
	SVM +PosStrIni	1.5787		0.1861		24.69%		0.0713		0.7610	
	RNN	0.9655 ± 0.0189		0.1224 ± 0.0022		52.31% ± 0.63%		0.0384 ± 0.0011		0.8296 ± 0.0029	
	Transformer (present work)	**0.8926** ± **0.0166**		**0.1137** ± **0.0017**		**53.75%** ± **0.40%**		**0.0373** ± **0.0009**		**0.8435** ± **0.0026**	
Rom-orth	Random daughter	4.2567		0.4854		2.97%		-		0.5147	
	CorPaR +Ini	0.9531		0.1160		47.23%		-		0.8400	
	SVM +PosStr	0.8988		0.1105		50.43%		-		0.8501	
	RNN	0.5941 ± 0.0100		0.0770 ± 0.0015		69.80% ±0.22%		-		0.8916 ± 0.0019	
	Transformer (present work)	**0.5525** ± **0.0104**		**0.0720** ± **0.0017**		**71.23%** ± **0.52%**		-		**0.9002** ± **0.0017**	

Table 4: Evaluation of models and baselines using various metrics, averaged across 5 runs (same hyperparameters, different seeds), with standard deviations. Because Rom-orth is not in IPA, character edit distance is used instead of PED, and we cannot accurately calculate FER. See Section 6.1 for an explanation of each evaluation metric.

	Romance (phon & orth)	Sinitic
learning rate	0.00013	0.0007487
num_encoder_layers	3	2
num_decoder_layers	3	5
embedding size	128	128
n_head	8	8
dim_feedforward	128	647
dropout	0.202	0.1708861
training epochs	200	200
warmup epochs	50	32
weight decay	0	0.0000001
batch size	1	32

Table 5: Hyper-parameters used in training the Transformer

	Romance-phon	Romance-orth	Sinitic
learning rate	0.00055739	0.000964	0.000864
num_encoder_layers	1	1	1
num_decoder_layers	1	1	1
embedding size	107	51	78
hidden size	185	130	73
dim_feedforward	147	111	136
dropout	0.1808	0.323794	0.321639
training epochs	181	193	237
warmup epochs	15	15	15
batch size	8	8	4

Table 6: Hyper-parameters used in training the RNN

Dataset	Model	PED ↓	NPED ↓	Acc % ↑	FER ↓	BCFS ↑
Rom-phon	Random daughter (Chang et al., 2022)	7.1880	0.8201	0%	1.1396	0.3406
	CorPaR + Ini (List et al., 2022)	2.0885	0.2491	14.29%	0.0874	0.6799
	SVM + PosStrIni (List et al., 2022)	1.9005	0.2276	17.05%	0.0883	0.7039
	RNN (Meloni et al., 2021)	1.4581	0.1815	36.68 %	0.0592	0.7435
	Transformer (present work)	1.2516	0.1573	41.38%	0.0550	0.7790
Rom-orth	Random daughter (Chang et al., 2022)	6.3272	0.6542	0.55%	-	0.4023
	CorPaR + PosStrIni (List et al., 2022)	1.8313	0.2001	18.89%	-	0.7227
	SVM + PosStr (List et al., 2022)	1.6995	0.1867	21.66%	-	0.7454
	RNN (Meloni et al., 2021)	1.3189	0.1505	38.89%	-	0.7742
	Transformer (present work)	1.1622	0.1343	45.53%	-	0.7989

Table 7: Evaluation of models and baselines with various metrics on *Meloni et al. (2021)'s Romance* datasets, where all entries from Dinu and Ciobanu (2014) are removed, for 1 run (using the hyperparameters of the best run on the full dataset)

Latin	Romanian	French	Italian	Spanish	Portuguese
[kɔlleːktɪoːnɛm]	[kolektsie]	[kɔlɛksjɔ̃]	[kolletsione]	[kolekθjon]	[kulɨsẽ̈ʊ̃]

Table 8: One cognate set, with Latin as the protoform and all columns to its right as the daughter cognates

A   For every submission:

☑ A1. Did you describe the limitations of your work?
Section 8

☑ A2. Did you discuss any potential risks of your work?
Section 8

☑ A3. Do the abstract and introduction summarize the paper's main claims?
Section 1

☐ A4. Have you used AI writing assistants when working on this paper?
Not applicable. Left blank.

B   ☑ Did you use or create scientific artifacts?

Sections 3, 4

☑ B1. Did you cite the creators of artifacts you used?
Sections 3,4,5,6

☑ B2. Did you discuss the license or terms for use and / or distribution of any artifacts?
Sections 3, 5.1

☐ B3. Did you discuss if your use of existing artifact(s) was consistent with their intended use, provided that it was specified? For the artifacts you create, do you specify intended use and whether that is compatible with the original access conditions (in particular, derivatives of data accessed for research purposes should not be used outside of research contexts)?
Not applicable. Left blank.

☐ B4. Did you discuss the steps taken to check whether the data that was collected / used contains any information that names or uniquely identifies individual people or offensive content, and the steps taken to protect / anonymize it?
Not applicable. Left blank.

☑ B5. Did you provide documentation of the artifacts, e.g., coverage of domains, languages, and linguistic phenomena, demographic groups represented, etc.?
Section 3

☑ B6. Did you report relevant statistics like the number of examples, details of train / test / dev splits, etc. for the data that you used / created? Even for commonly-used benchmark datasets, include the number of examples in train / validation / test splits, as these provide necessary context for a reader to understand experimental results. For example, small differences in accuracy on large test sets may be significant, while on small test sets they may not be.
Table 2 and Appendix Section A

C   ☑ Did you run computational experiments?

Section 4

☑ C1. Did you report the number of parameters in the models used, the total computational budget (e.g., GPU hours), and computing infrastructure used?
Appendix A

☑ C2. Did you discuss the experimental setup, including hyperparameter search and best-found hyperparameter values?
Hyperparameter search: 5.1 Hyperparameter values: Appendix Section B

☑ C3. Did you report descriptive statistics about your results (e.g., error bars around results, summary statistics from sets of experiments), and is it transparent whether you are reporting the max, mean, etc. or just a single run?
Table 3

☑ C4. If you used existing packages (e.g., for preprocessing, for normalization, or for evaluation), did you report the implementation, model, and parameter settings used (e.g., NLTK, Spacy, ROUGE, etc.)?
5.1, 6.1, 6.3

D ☒ Did you use human annotators (e.g., crowdworkers) or research with human participants?

Left blank.

☐ D1. Did you report the full text of instructions given to participants, including e.g., screenshots, disclaimers of any risks to participants or annotators, etc.?
No response.

☐ D2. Did you report information about how you recruited (e.g., crowdsourcing platform, students) and paid participants, and discuss if such payment is adequate given the participants' demographic (e.g., country of residence)?
No response.

☐ D3. Did you discuss whether and how consent was obtained from people whose data you're using/curating? For example, if you collected data via crowdsourcing, did your instructions to crowdworkers explain how the data would be used?
No response.

☐ D4. Was the data collection protocol approved (or determined exempt) by an ethics review board?
No response.

☐ D5. Did you report the basic demographic and geographic characteristics of the annotator population that is the source of the data?
No response.