Semisupervised Neural Proto-Language Reconstruction

Liang Lu[1], Peirong Xie[2], David R. Mortensen[1]

[1]Carnegie Mellon University, [2]University of Southern California   lianglu@cs.cmu.edu, louisxie@usc.edu, dmortens@cs.cmu.edu

TL;DR: We introduce the novel task of semisupervised protoform reconstruction. Informed by historical linguists' comparative method, we propose the DPD architecture for this task, which outperforms baseline methods in almost all situations.

The Task

Given descendant words (reflexes in a cognate set) of the same ancestral word, reconstruct the ancestral word (protoform).

? 'neighbor' (Latin)

vwazin	*vitʃina*	*beθina*	*vizine*
<voisine>	<vicina>	<vecina>	<vizinha>
(French)	(Italian)	(Spanish)	(Portuguese)

Answer: ***wikinam*** <vicinam>

Example from (Meloni et al., 2021; Ciobanu and Dinu, 2018)

The Comparative Method

- The regularity principle: sound changes are regular
- Reflexes should be derivable deterministically from reconstructions using a single set of sound change rules
- Laborious for humans to apply in practice for large datasets

Semisupervised Reconstruction

It is likely that historical linguists only have a limited number of protoforms to work with at the early stage of a reconstruction project. In such a scenario, labeled training data is scarce for neural reconstruction models. Unlabeled cognate can be useful to semisupervised reconstruction models if used effectively. Consider the following example drawn from Tangkhulic Languages:

Gloss	'grandchild'	'bone'	'breast'	'laugh'
Kachai	ð ɐ	r ɐ	n ɐ	n i
Huishu	r u k	r u k	n u k	n u k
Ukhrul	r u	r u	n u	n u
Reference Protoform	d u	r u	n u	n ɨ
Labeled?	Yes	Yes	No	No
Model Sees...	d u	r u	(hidden)	(hidden)
Supervised Model	d u	r u	n u	n u
DPD	d u	r u	n u	n ✻

Where ✻ is something other than **u**. Notice that if we try to infer the Kachai reflexes from the reconstructions:

Supervised Model			DPD		
d u	→	ð ɐ	d u	→	ð ɐ
r u	→	r ɐ	r u	→	r ɐ
n u	→	n ɐ	n u	→	n ɐ
n u	☹ →	n i	n ✻	☺ →	n i

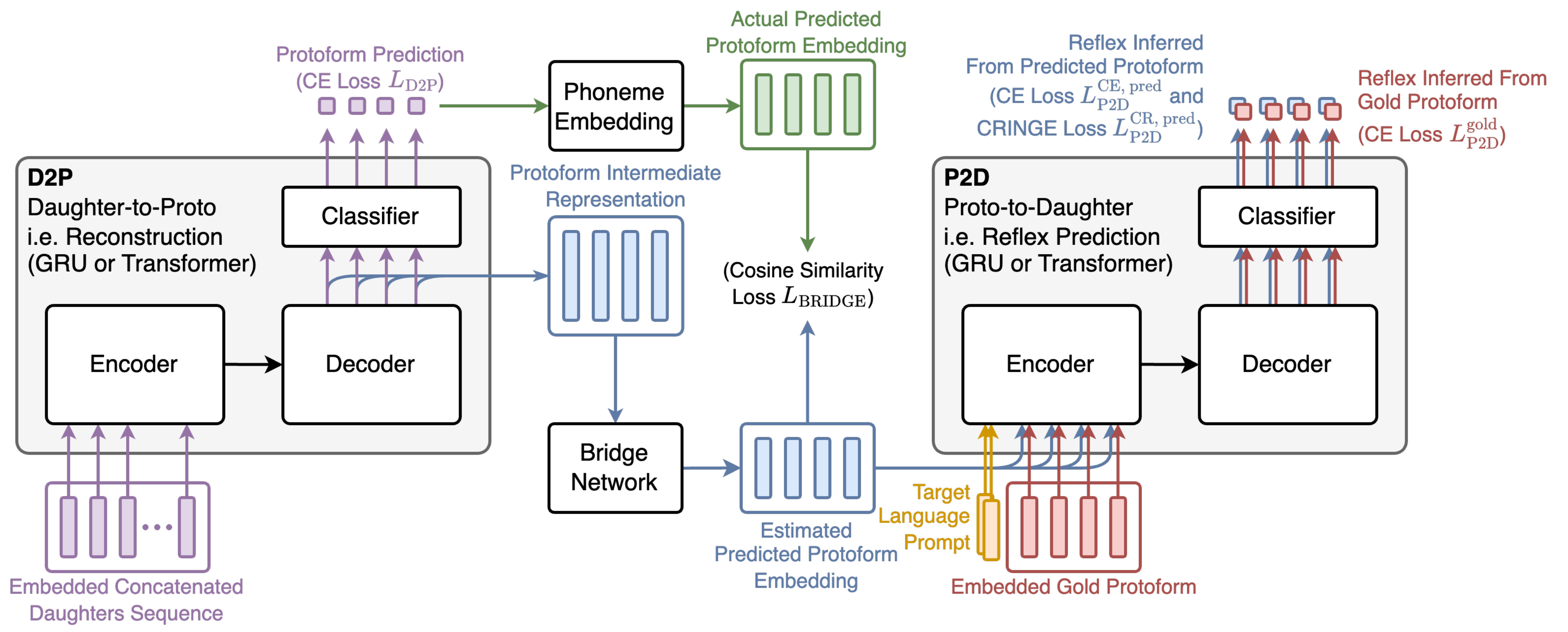
The DPD (Daughter-to-Proto-to-Daughter) Model

- Designed to mimic the comparative method's workflow, which involves checking whether the reflexes are recoverable from the reconstruction
- Gradients flow from P2D into D2P, allowing the reconstruction sub-network to learn even in the absence of protoform labels

Reflexes —D2P→ **Reconstruction?** —P2D→ **Reflexes?**

What's a good reconstruction that would allow P2D to recover the reflexes?

Can the reflexes be derived from the reconstruction proposed by D2P?



Loss Calculation

$$L_{\mathrm{overall}} = \alpha_1 L_{\mathrm{D2P}} + \alpha_2 L_{\mathrm{P2D}}^{\mathrm{CE,\,pred}} + \alpha_3 L_{\mathrm{P2D}}^{\mathrm{CR,\,pred}} + \alpha_4 L_{\mathrm{P2D}}^{\mathrm{gold}} + \alpha_5 L_{\mathrm{BRIDGE}}$$

where $\alpha_{\{1\ldots5\}}$ are constants

Reconstruction Input and Output Format

Input sequence (concatenated reflexes with markers and separators)

[Cantonese]:mei̯˧[Mandarin]:mei̯˨˩˦*[Wu]:me̯˧˦*

↓ D2P

Target prediction (Middle Chinese)

mij³

Reflex Prediction Input and Output Format

Input Sequence	Input Sequence	Input Sequence
[Cantonese]mij³	[Mandarin]mij³	[Wu]mij³
↓ P2D	↓ P2D	↓ P2D
Target output	Target output	Target output
mei̯˧	mei̯˨˩˦	me̯˧˦

Experiments

Strategies

- **Weak baselines:** Supervised only (**SUPV**), Bootstrapping (**BST**), Π-model (**ΠM**)
- **Strong baseline:** Π-model with Bootstrapping (**ΠM-BST**)
- **DPD-based:** Plain DPD (**DPD**), DPD with Bootstrapping (**DPD-BST**), DPD merged with Π-model (**DPD-ΠM**), DPD-ΠM with Bootstrapping (**DPD-ΠM-BST**)

Sub-network Architectures

D2P and P2D both being **GRU** or both being Transformer (**Trans**)

Datasets

WikiHan (Chang et al., 2022) for Middle Chinese reconstruction, **Romance** (Meloni et al., 2021; Ciobanu and Dinu, 2018) for Latin reconstruction

Results – Performance when 10% of the Cognate Sets Are Labeled

WikiHan

Architecture	Strategy	ACC% ↑	TED ↓	TER ↓	FER ↓	BCFS ↑
Transformer	DPD-ΠM-BST (ours)	**40.50%** ❶❷❸❹	**1.0075** ❶❷❸❹	**0.2360** ❶❷❸❹	**0.0970** ❶❷❸❹	**0.6707** ❶❷❸❹
	DPD-BST (ours)	39.06% ❶❷❸❹	1.0367 ❶❷❸❹	0.2428 ❶❷❸❹	0.0997 ❶❷❸❹	0.6630 ❶❷❸❹
	DPD-ΠM (ours)	37.72% ❶❷❸❹	1.0791 ①❷③	0.2528 ①❷③	0.1022 ❶❷❸❹	0.6472 ❷③
	DPD (ours)	39.50% ❶❷❸❹	1.0356 ❶❷❸❹	0.2426 ❶❷❸❹	0.0993 ❶❷❸❹	0.6564 ❶❷❸❹
	ΠM-BST	34.21%	1.1489	0.2691	0.1106	0.6371
	BST (Lee, 2013)	34.78%	1.1455	0.2683	0.1109	0.6334
	ΠM (Laine and Aila, 2017)	34.30%	1.1699	0.2740	0.1122	0.6209
	SUPV	33.25%	1.1891	0.2785	0.1140	0.6138
GRU	DPD-ΠM-BST (ours)	**39.74%** ❶❷❸❹	**1.0280** ❶❷❸❹	**0.2408** ❶❷❸❹	**0.0972** ❶❷❸❹	**0.6683** ❶❷❸❹
	DPD-BST (ours)	35.89% ①❷③	1.1025 ①❷③④	0.2582 ①❷③④	0.1039 ①❷③④	0.6493 ①❷③④
	DPD-ΠM (ours)	37.90% ❶❷❸❹	1.0697 ①❷③④	0.2506 ❶❷❸❹	0.1006 ❶❷❸❹	0.6517 ❶❷❸❹
	DPD (ours)	34.51% ①③④	1.1538 ①③④	0.2703 ①③④	0.1091 ③④	0.6278 ①③④
	ΠM-BST	34.99% ①②③	1.1479 ①③	0.2689 ①③	0.1077 ①③	0.6354 ①②③
	BST (Lee, 2013)	28.18%	1.3092	0.3067	0.1208	0.5939
	ΠM (Laine and Aila, 2017)	32.59%	1.2047	0.2822	0.1137	0.6166
	SUPV	28.16%	1.3257	0.3105	0.1234	0.5835

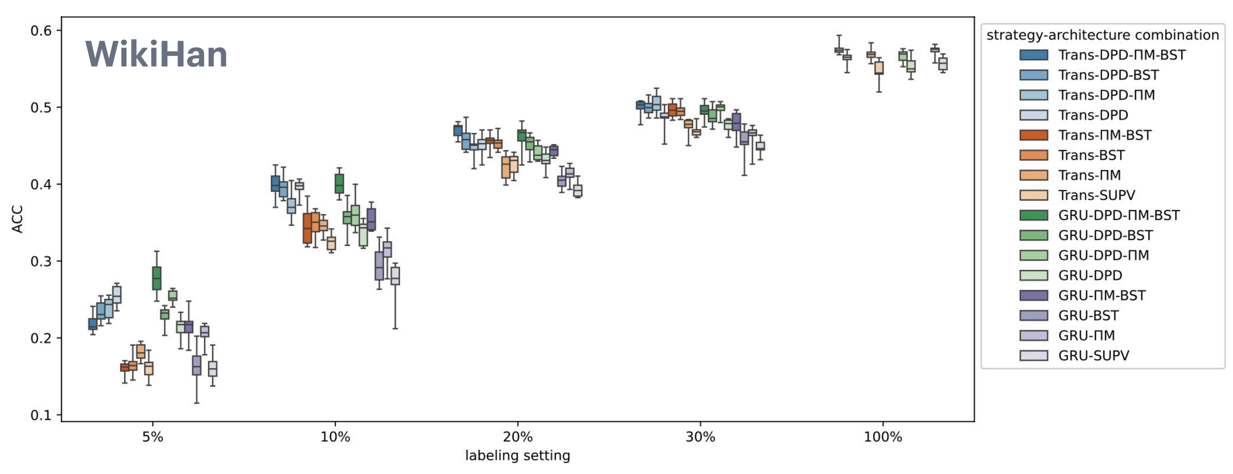
Romance

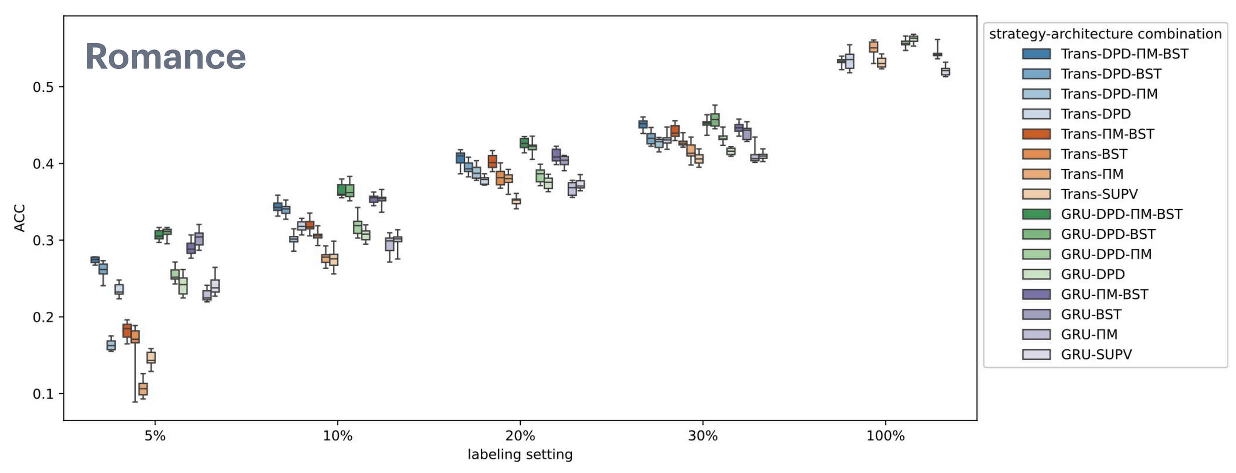
Architecture	Strategy	ACC% ↑	TED ↓	TER ↓	FER ↓	BCFS ↑
Transformer	DPD-ΠM-BST (ours)	**34.63%** ❶❷❸❹	**1.3115** ❶❷❸❹	**0.1463** ❶❷❸❹	**0.0588** ❶❷❸❹	**0.7850** ❶❷❸❹
	DPD-BST (ours)	33.51% ❶❷❸❹	1.3605 ❶❷❸❹	0.1517 ❶❷❸❹	0.0599 ❶❷❸❹	0.7763 ❶❷❸❹
	DPD-ΠM (ours)	29.24%	1.5888	0.1772	0.0732	0.7423
	DPD (ours)	31.94% ①②③④	1.5111	0.1685	0.0678 ②③	0.7529
	ΠM-BST	32.10% ①②③④	1.4005 ①②③④	0.1562 ①②③④	0.0636 ①②③④	0.7716 ①②③④
	BST (Lee, 2013)	29.95%	1.5066	0.1680	0.0704	0.7555
	ΠM (Laine and Aila, 2017)	26.97%	1.7134	0.1911	0.0796	0.7239
	SUPV	26.99%	1.7331	0.1933	0.0794	0.7218
GRU	DPD-ΠM-BST (ours)	36.78% ❶❷	1.2380 ❶❷❸❹	0.1381 ❶❷❸❹	0.0483 ❶❸❹	0.7980 ❶❷❸❹
	DPD-BST (ours)	**37.60%** ❶❷❸❹	**1.2149** ❶❷❸❹	**0.1355** ❶❷❸❹	**0.0457** ❶❷❸❹	**0.8014** ❶❷❸❹
	DPD-ΠM (ours)	31.51%	1.4892	0.1661	0.0628	0.7586
	DPD (ours)	31.12%	1.4837	0.1655	0.0608	0.7591
	ΠM-BST	35.50%	1.2970	0.1447	0.0531	0.7909 ①
	BST (Lee, 2013)	35.87%	1.2893	0.1438	0.0509	0.7908
	ΠM (Laine and Aila, 2017)	29.40%	1.5440	0.1722	0.0643	0.7517
	SUPV	30.69%	1.5018	0.1675	0.0612	0.7558

Bold: the best-performing model for each metric; ①**:** significantly better than all weak baselines (SUPV, BST, and ΠM) on dataset seed 1 with $p < 0.01$; ❶**:** significantly better than the ΠM-BST strong baseline and all weak baselines on dataset seed 1 with $p < 0.01$; ②, ③, ④, ❷, ❸, ❹**:** likewise for dataset seeds 2–4.

Results – Performance vs. Proportion of Labeled Cognate Sets

- Unsurprisingly, performance increases as the percentage of labeled data increases
- In most cases, DPD-based strategies generalize well to other labelling settings
- The performance difference between strategies is more pronounced when labeled data is scarce

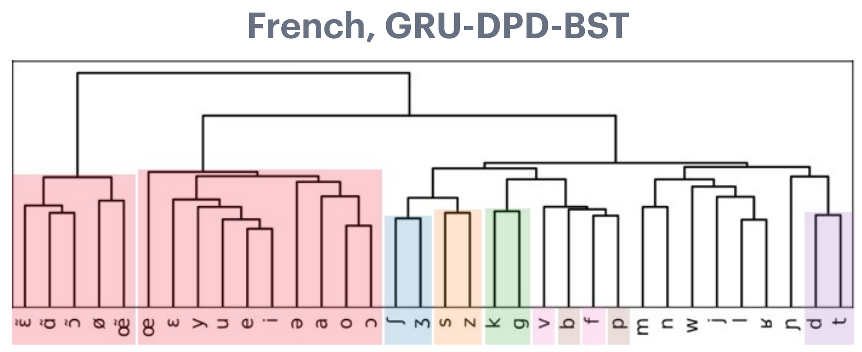




Analysis – Hierarchical Clustering of Phoneme Embedding

Phoneme embeddings learned by DPD-based strategies appear to align more with how phonologists organize phonemes.

French, GRU-DPD-BST



French, GRU-BST