

A Appendices

A.1 Sample Breakdown by Label

Table 1 gives breakdown of the Assamese-Bengali and Bengali-Assamese train/test splits based on their labels. Since we distinguish cognates from loanwords but otherwise do not single out loanwords in our datasets, loanwords may exist in the other categories. Given the phonetic similarity between loanwords and their sources, where loanwords do exist in our data, they are overwhelmingly likely to be in the hard negative category.

	<i>as-bn</i>		<i>bn-as</i>	
	train	test	train	test
Cog.	306	303	306	300
HN	776	769	721	716
Syn.	329	327	317	316
Rnd.	304	301	304	299
Total	1715	1700	1648	1631

Table 1: Number of Hard-Negatives (HN), Synonyms (Syn.), Cognates (Cog.), and Random pairs (Rnd.) in Assamese-Bengali and Bengali-Assamese train/test sets.

A.2 ALBERT (Monolingual Assamese Configuration)

Table 2 gives configuration details of the monolingual Assamese Transformer model that we trained for this research.

A.3 Further Details on Effects of Phonetic Features

Of the 6 phonetic edit distances we used, Hamming Feature Distance (divided by maximum length) and Partial Hamming Distance (divided by maximum length) appear to be the most correlated with cognate status according to the weights assigned to them by the logistic regressor. This suggests that Hamming distance’s (?) focus on using the minimum number of substitutions to transform one string into another works well for similar languages like Assamese and Bengali where most individual phonemes are largely preserved between cognate words.

Interestingly, the Dolgo Prime Distance variant gets a low (usually negative) weight in almost all feature combinations. This is interesting and suggests that Dolgo Prime Distance is not useful here due to it unduly conflating multiple phonemes into

the same class. The Dolgopolsky-inspired stable phoneme classes used by PanPhon places /ʃ/ in the “coronal fricatives” class, while /x/ is in the “velar/postvelar obstruents” class. The unvoiced velar fricative /x/ is unique to Assamese and rare among Indian languages (?) and we know well that Bengali and Assamese have a regular /ʃ/-/x/ sound correspondence. So, as Dolgo Prime distance splits these up into different classes, when using this metric cognate words containing these corresponding sounds will have phonetic distance added to them when in fact they are regularly corresponding.

Parameters	Config
architecture	AlbertForMaskedLM
attention_probs_dropout_prob	0.1
bos_token_id	2
classifier_dropout_prob	0.1
embedding_size	128
eos_token_id	3
hidden_act	gelu
hidden_dropout_prob	0.1
hidden_size	768
initializer_range	0.02
inner_group_num	1
intermediate_size	3072
layer_norm_eps	1e-05
max_position_embeddings	514
num_attention_heads	12
num_hidden_groups	1
num_hidden_layers	6
position_embedding_type	"absolute"
transformers_version	"4.18.0"
vocab_size	32001

Table 2: ALBERT Model configuration trained on monolingual Assamese corpus.