Phonetic, Semantic, and Articulatory Features in Assamese-Bengali Cognate Detection

Abhijnan Nath, Rahul Ghosh, and Nikhil Krishnaswamy

VarDial 2022, October 16, 2022, Gyeongju, South Korea

Colorado State University

Outline

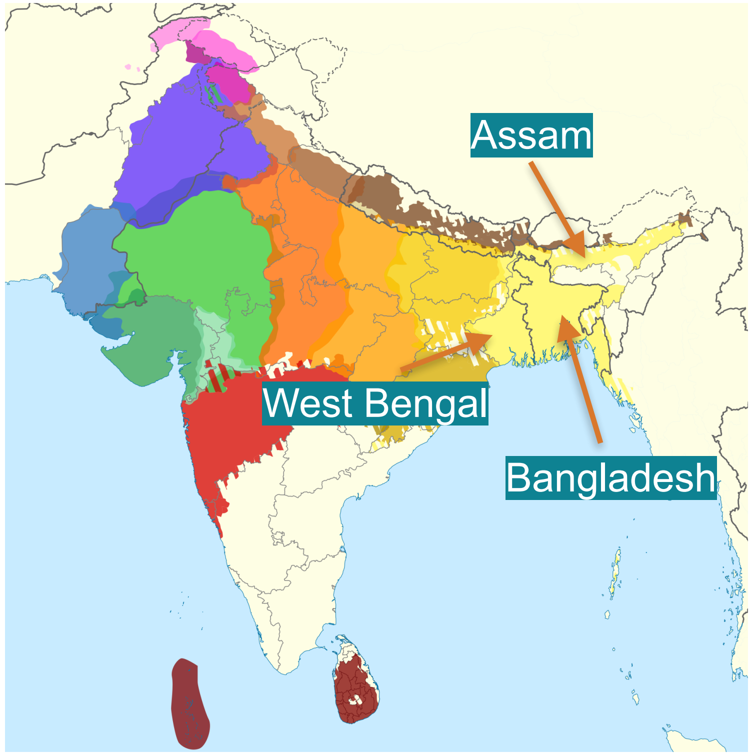
- Introduction
- Datasets
- Orthographic and Phonetic Edit Distance
- Phonetic Alignment Network
- Semantic Similarity
  - Assamese-ALBERT
  - Affine Transformation Between Embedding Spaces
- Results
- Discussion
- Conclusions and Future Work

Introduction: Cognate Detection

- Cognates: words inherited from a common ancestor
  - Sound shift, semantic change in effect
  - Cognates may not be obvious
  - e.g., extreme example: English. “two” vs. Armenian երկու (*erku*)
- Unlike loanwords, cognates are *necessarily* subject to diachronic sound change
- Cognates are crucial to historical linguistic applications, e.g., reconstructing ancestral forms
- NLP research in this area often conflates loanwords and cognates (e.g., Kondrak (2001)): **we do not adopt this definition**
- Focus languages: **Assamese, Bengali** (Eastern India, Bangladesh)
- Combine phonetic, orthographic, articulatory alignment, and semantic features
- Apply novel affine transformation technique to language models

Introduction: Assamese and Bengali

- Bengali (`bn`): 262 million speakers
  - Primarily in West Bengali (India) and Bangladesh
- Assamese (`as`): 15 million speakers
  - Primarily in Assam (India)
- **Descent**: Early Indo-Aryan >> Magadhi Prakrit >> Bengali-Assamese languages
- Similar grammatical features (classifying affixes/“measure words”), common phonetic innovation (e.g., Skt. /ə/ → /ɔ/, loss of contrastive vowel length), same script

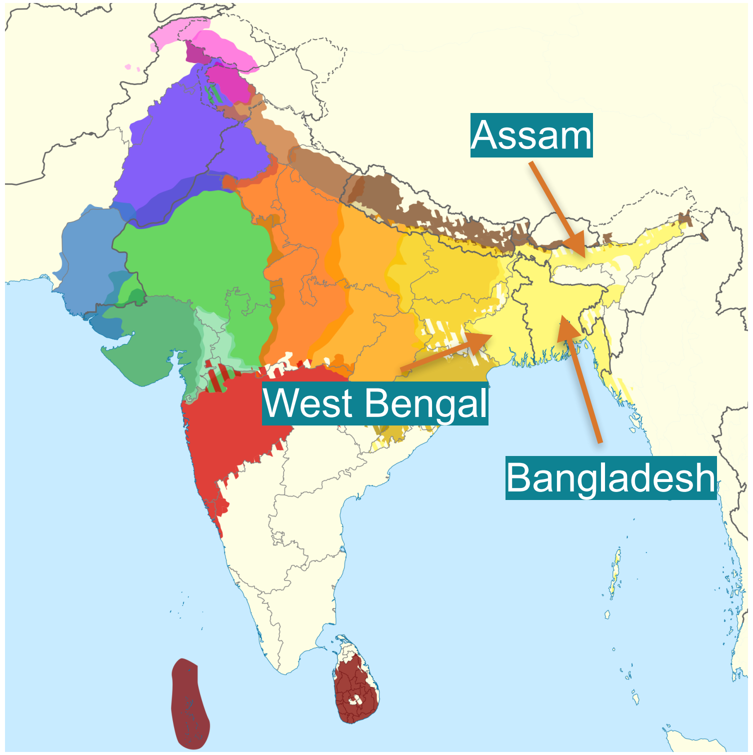


https://en.wikipedia.org/wiki/Eastern_Indo-Aryan_languages

Introduction: Assamese and Bengali

- **Some important differences in sound pattern**

Glyph	Bengali	Assamese	Glyph	Bengali	Assamese
চ	/tʃ/	/s/	ঢ	/ɖɦ/	/dɦ/
ছ	/tʃʰ/	/s/	ত	/t̪/	/t/
জ	/dʒ/	/z/	থ	/t̪ʰ/	/tʰ/
ঝ	/dʒɦ/	/z/	দ	/d̪/	/d/
ট	/ʈ/	/t/	ধ	/d̪ɦ/	/dɦ/
ঠ	/ʈʰ/	/tʰ/	স, শ, ষ	/ʃ/	/x/
ড	/ɖ/	/d/	র/ৰ	/r/	/ɹ/



https://en.wikipedia.org/wiki/Eastern_Indo-Aryan_languages

Datasets

- Data extracted from Wiktionary categories
  - `[Descendent]_terms_derived_from_Sanskrit`
- Exclude affixes, numerals, non-phonetic/syllabic terms
- Took union of Bengali and Assamese results, then subset where paired terms had same ancestor
  - Checking against common ancestry removes loanwords
- Convert words to IPA using Epitran (Mortensen et al., 2018)
  - Created custom Epitran G2P for Assamese

Descendant	Ancestor	# Cognates
Assamese	Sanskrit	205
Bengali	Sanskrit	335

Cognate counts per language

Datasets

- Complete dataset with non-cognate samples:
  - **Hard negatives** (phonetically similar non-cognates)
    - PanPhon (Mortensen et al., 2016) calculates 6 edit distances between every cognate and every lemma in other language
    - Closest ≤6 phonetic neighbors selected (e.g., **Asm. কথা (/kɔtʰa/) “word”, Beng. কটা (/kɔʈa/) “how many”**)
  - **Synonyms** (semantically similar non-cognates)
    - Exploit Wiktionary metadata to extract synonyms for each gathered cognate where available
    - e.g., **Asm. কুটুম (/kutum/) “family”, Beng. রিশতাদার (/riʃt̪ad̪ar/) “relatives”**
  - **Randoms** (no discernable relation)
    - Randomly paired words in the two languages
    - Exclude pairs already in cognates, hard negatives, or synonyms subsets

Datasets

- Final step: native speaker verification
- Concatenate all data splits into Assamese-Bengali, Bengali-Assamese, and bidirectional datasets
- Approx. 50/50 train/test split

	as-bn		*bn-as*	
	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

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

Orthographic Similarity

- Assamese and Bengali both use Bengali (Eastern Nagari) script
- Orthographic similarity is just Levenshtein distance between words

পঞ্চাশ টাকা

পঞ্চাশ টকা

Orthographic Similarity

- Assamese and Bengali both use Bengali (Eastern Nagari) script
- Orthographic similarity is just Levenshtein distance between words

পঞ্চাশ টাকা

পঞ্চাশ টকা

Orthographic Similarity

- Assamese and Bengali both use Bengali (Eastern Nagari) script
- Orthographic similarity is just Levenshtein distance between words

পঞ্চাশ টাকা ⇒

পঞ্চাশ টকা ⇒

Orthographic Similarity

- Assamese and Bengali both use Bengali (Eastern Nagari) script
- Orthographic similarity is just Levenshtein distance between words

পঞ্চাশ টাকা	⇒	/pɔntʃaʃ ʈaka/
পঞ্চাশ টকা	⇒	/pɔnsax tɔka/

Orthographic Similarity

- Assamese and Bengali both use Bengali (Eastern Nagari) script
- Orthographic similarity is just Levenshtein distance between words

পঞ্চাশ টাকা	⇒	/pɔntʃaʃ ʈaka/
পঞ্চাশ টকা	⇒	/pɔnsax tɔka/

Textual edit distance may be useful OR misleading!

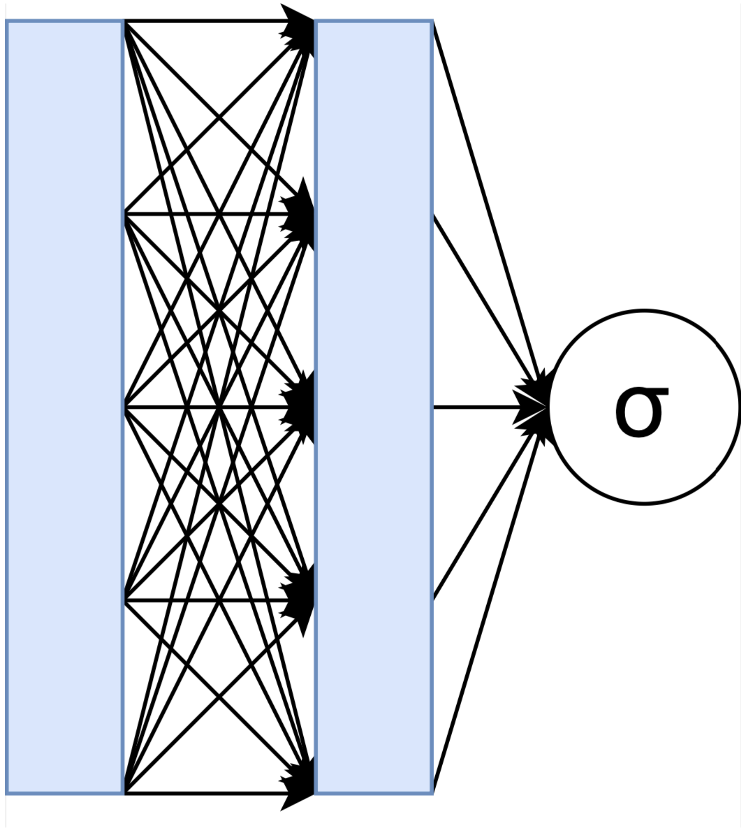
Phonetic Similarity

- 6 edit distances from PanPhon
  - Fast Levenshtein Distance
  - Dolgo Prime Distance
  - Feature Edit Distance
  - Hamming Feature Distance
  - Weighted Feature Distance
  - Partial Hamming Feature Distance
  - … all normalized by the maximum length of the words in the pair

	pɔntʃaʃ vs. pɔnsax
Fast Levenshtein	0.428571428571429
Dolgo Prime	0.285714285714286
Feature Edit	0.157738095238095
Hamming Feature	0.172619047619048
Weighted Feature	1.375
Partial Hamming Feature	0.169642857142857

Articulatory Alignment

- For each word pair, gather 21 articulatory features from PanPhon
- Word pair features concatenated and padded
- Fed into feedforward neural network
  - 2 hidden layers, 512 neurons each, ReLU activation
  - 5,000 training epochs, BCE loss, Adam optimization
  - Output is pre-sigmoid DNN logit value
    - Alignment score is feature in final classification task



Semantic Similarity

- Past work on cognate detection has focused mostly on phonetic similarity
- Modern language models allow for quantitative measurement of semantic similarity
- Four large multilingual language models (MLMs): **MBERT**, **XLM-R**, **IndicBERT**, and **MuRIL**
  - **MBERT**, **XLM-R** trained on ~100 languages (MBERT does not contain Assamese, XLM-R trained on little Assamese data)
  - **IndicBERT**, **MuRIL** specialized on Indian languages
- Monolingual Assamese Model (ALBERT variant)
  - Trained on Assamese Wikidumps, OSCAR, PMIndia, and Common Crawl, ~14M Assamese tokens, BERT MLM loss function

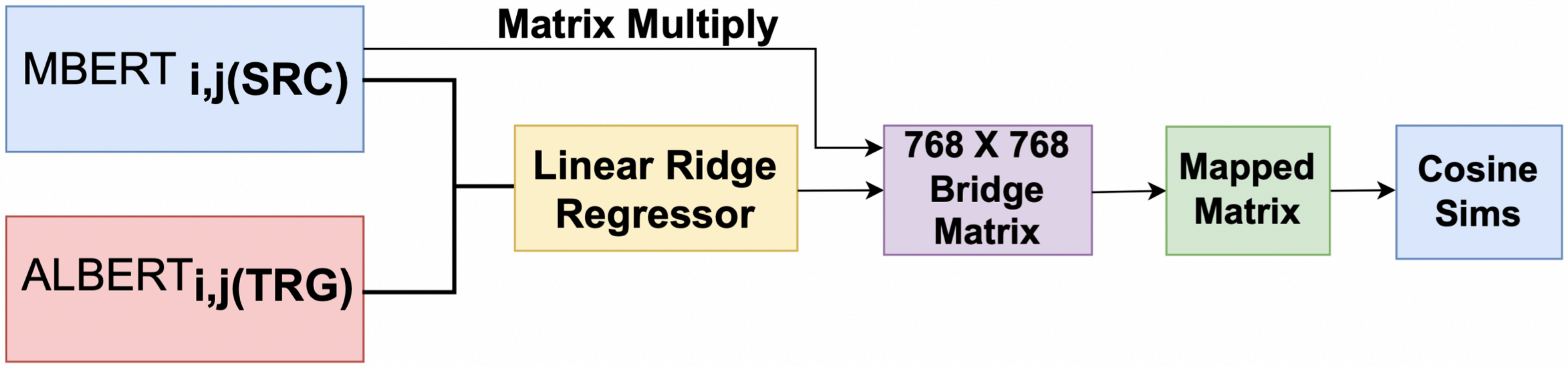
Assamese-ALBERT

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

ALBERT Model configuration trained on monolingual Assamese corpus.

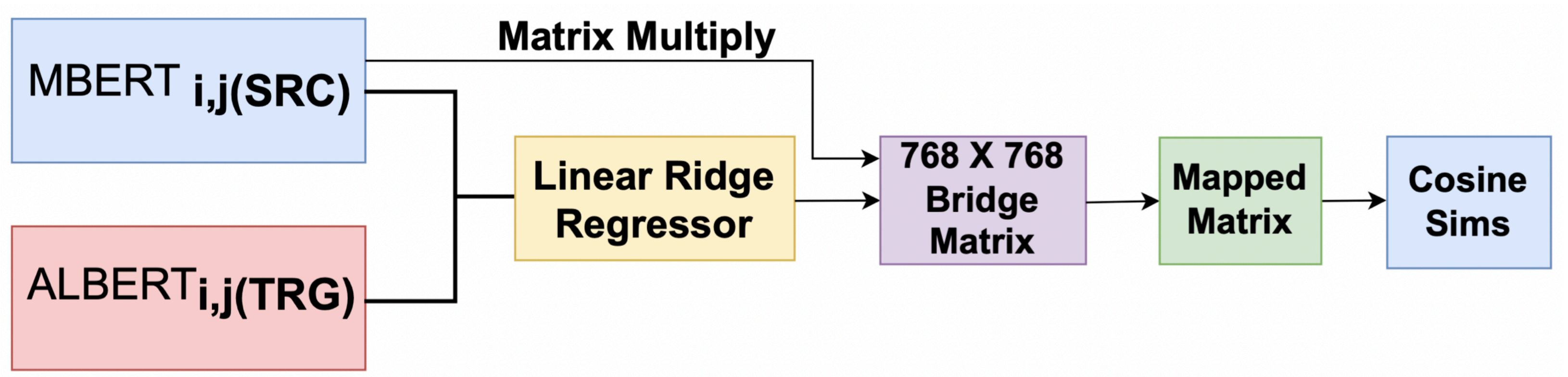
Semantic Similarity

- Getting semantic similarity between words:
  - Input “sentence” `<bos>`<word>`<eos>` into MLM, extract `<bos>` last hidden state, take cosine similarity between vectors
- **Problem**: treatment of Bengali and Assamese is not equal in MLMs
  - e.g., MBERT: no Assamese, XLM-R: weak Assamese (5M training tokens)
  - To provide additional Assamese semantics, map monolingual vectors into multilingual space
- **Problem**: vectors from different model spaces are not directly comparable



Affine Transformation Between Embedding Spaces

- **Intuition**: if two models preserve similar information, then solving for a transformation $f(x; W)$ that minimizes distance between equivalent samples from each model should align the two embedding spaces
- Previous research from vision community (e.g., McNeely-White et al., 2020) has demonstrated interchangeability up to matrix $M_{A\to B} \in \mathbb{R}^{d_A} \times \mathbb{R}^{d_B}$ if inputs and outputs correspond to the same label
- **Here we explore the application of this finding to language models**



Affine Transformation Between Embedding Spaces

- **Process**: cognate words should represent semantically similar information
- Create sentences that capture those semantics in Assamese/Bengali
- Sentences should be simple, appropriate to the part of speech, and leave word sense unambiguous

Language	**Sentence**	**IPA**
Bengali	এটি একটি টাং	eʈi ekʈi ʈaŋ
Assamese	এইটো এটা ঠেং	eitʊ eta $\mathrm{t^h}$ɛŋ
English	This is a foot/leg	

- Insert special tokens `<m>`/`</m>` around target word mention
- Get embedding of `<m>` token in each model (e.g., Assamese-ALBERT and Bengali MBERT)
- Compute affine mappings using 338 contextual word embedding pairs (*sentence maps*) and 3415 (`as-bn`)/3279 (`bn-as`) word-only embedding pairs (*word-level maps)*

Results

- Feature key:

Abbr.	Features
ped	Phonetic Edit distances (PED)
dl	DNN logits (alignment score)
ed	PED with textual Levenstein dist.
b	All native MLMs (BERT variants)
m	All mappings w/o native MLMs
ab-am	All MLMs w/ word-level maps
ab-sm	All MLMs with sentence maps
sm	Sentence maps

Semantic features

Abbreviations for feature combinations

Results

- Two classification models: 3-layer neural net (NN) and logistic regressor (LR)
- NN better performing, LR more interpretable
- Evaluations:
  - Train on bidirectional data, evaluate on bidirectional and `bn-as` and `as-bn` data
  - Train and evaluate on `bn-as` and `as-bn` data only (pair-specific models denoted with *)

	all	*bn-as*	*as-bn*	*bn-as**	*as-bn**
P(+)	95	97	94	90	90
R(+)	93	94	92	88	87
F1(+)	94	95	93	89	88

NN classifier results (as %) for `ed-dl-ab-am` (full feature set)

Results

	all	*bn-as*	*as-bn*	*bn-as**	*as-bn**
P(+)	95	97	94	90	90
R(+)	93	94	92	88	87
F1(+)	94	95	93	89	88

NN classifier results (as %) for `ed-dl-ab-am` (full feature set)

- Slightly higher performance using Bengali baseline
- Bengali forms often preserve consonant clusters where Assamese forms do not

Bengali	**Assamese**
সাঁঝ (/ʃãdʑʱ/)	সন্ধিয়া (/xɔndʱija/)
শিক্ষা (/ʃikkʰa/)	শিকোৱা (/xikʊwa/)
মিষ্টি (/miʃʈi/)	মিঠা (/mitʰa/)

Sample false negatives

Influence of Features: Alignment

Feat.	***all***	***bn-as***	***as-bn***	***bn-as****	***as-bn****
`ed`	76	76	76	76	76
`ed-dl`	**93**	**93**	**92**	**86**	**88**
`ped`	43	43	43	42	51

F1(+) as % with and without alignment score (`dl`) and Levenshtein distance features

- **Alignment score features** add most performance boost
- Logistic regressor gives alignment features weight of ~3.2, strong correlation with cognate status
- Alignment network able to assess regular sound correspondences (e.g., /ʃ/ → /x/) better than edit distance

Influence of Features: Phonetic and Orthographic

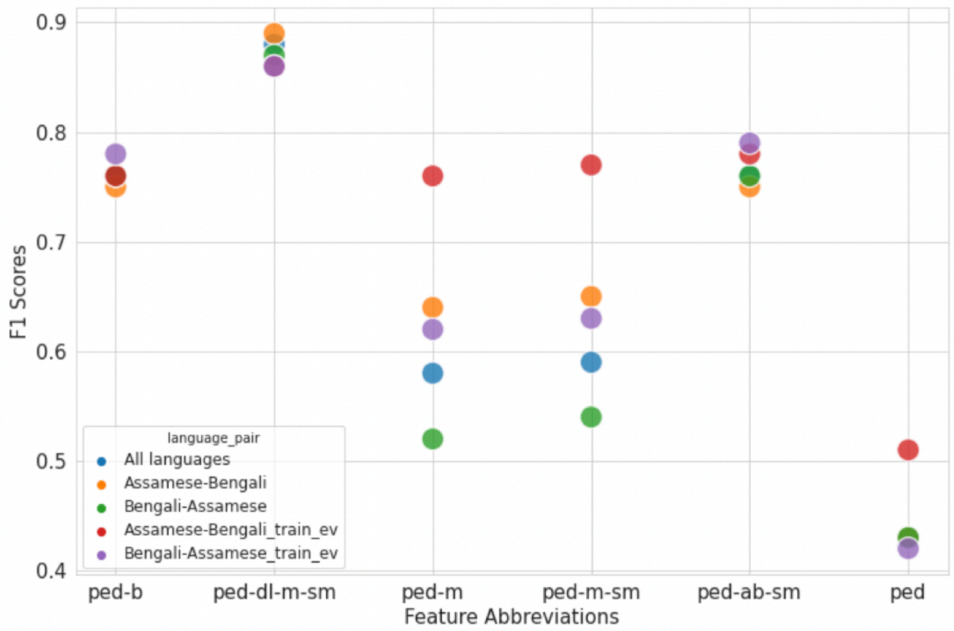
Feat.	***all***	***bn-as***	***as-bn***	***bn-as****	***as-bn****
`ed`	76	76	76	76	76
`ed-dl`	**93**	**93**	**92**	**86**	**88**
`ped`	43	43	43	42	51

F1(+) as % with and without alignment score (`dl`) and Levenshtein distance features

- **Textual Levenshtein distance** also helps performance compared to phonetic edit distance alone
- Logistic regressor gives orthographic features weight of ~-2.7, strong inverse correlation with cognate status
- Differences in pronunciation matter less when script is available (cf. English “science” /saɪən(t)s/ vs. French *science /sjɑ̃s/)*

Influence of Features: Semantic

- Adding any semantic info substantially improves on phonetic edit distance (`ped`) alone
- `ped-b` (`ped` + MLM cosine similarities) achieves performance on par with `ed` (all edit dists incl. orthographic)
- LR weights:
  - XLM-R: ~1.0
  - MBERT: ~0.4
  - MuRIL: ~0.3
  - IndicBERT: ~0.06



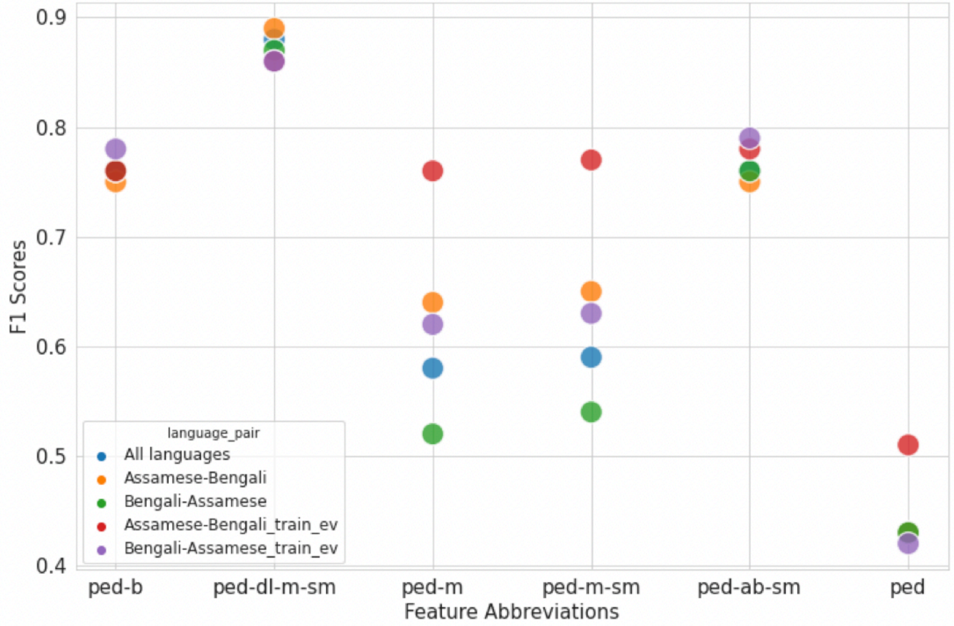
F1(+) with different semantic feature sets compared to phonetic edit distance baseline

Influence of Features: Semantic

- Adding semantic similarity is as good as adding textual Levenshtein distance, but specific retrieved cognates are different

	all		*bn-as**		*as-bn**	
	`ed`	`ped-b`	`ed`	`ped-b`	`ed`	`ped-b`
HN	18	12	12	11	6	4
Syn.	18	5	8	1	5	6
Rnd.	4	1	2	0	1	0

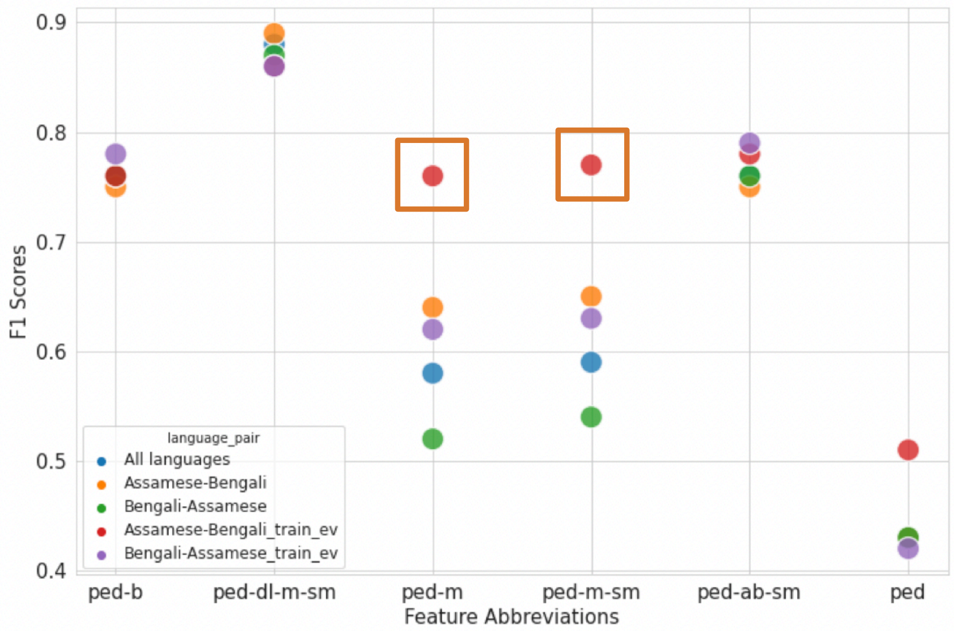
False positives using `ed` vs. `ped-b` feature sets broken down by negative example type



F1(+) with different semantic feature sets compared to phonetic edit distance baseline

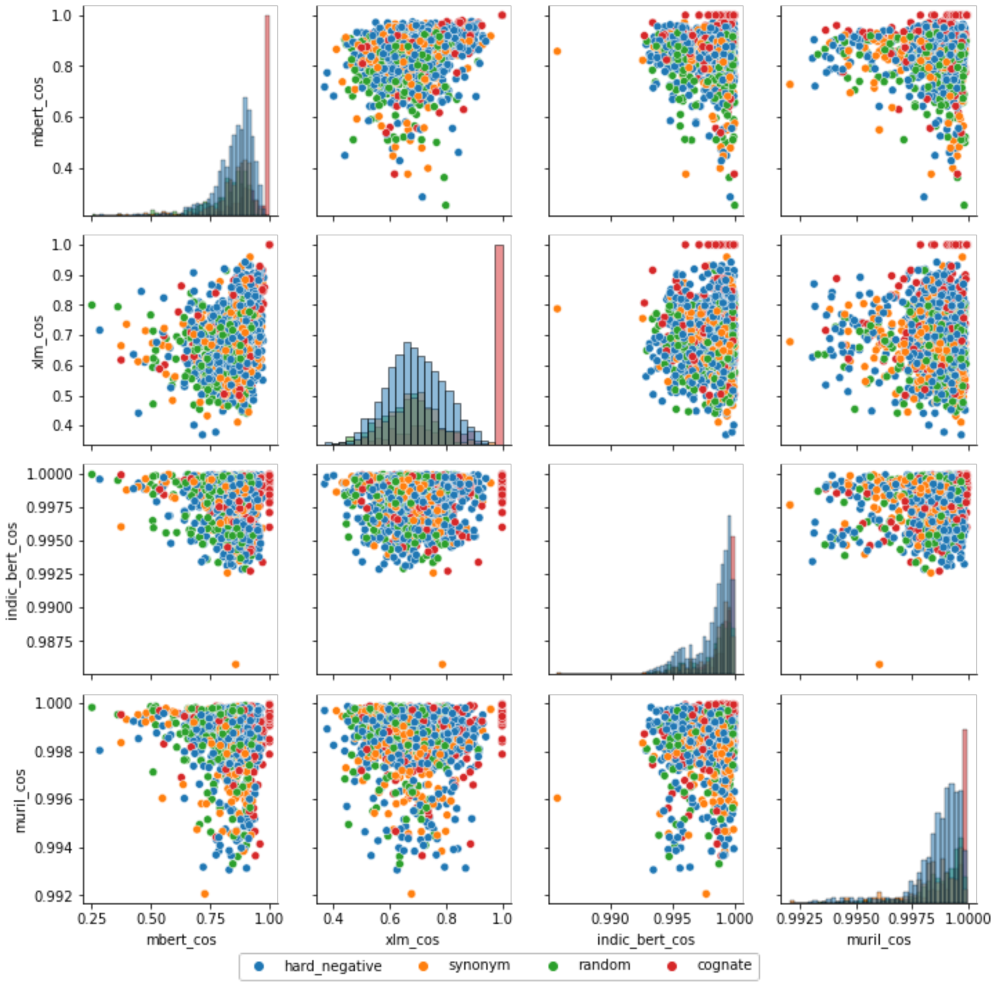
Influence of Features: Semantic

- Adding **word-level mappings** (Assamese-ALBERT → MBERT) to `ped` dramatically improves `as-bn` pair-specific model
  - F1(+) of 76%, same as using native MLM cosine similarities
  - LR weights:
    - XLM-R: ~1.0
    - MBERT: ~0.4
    - **Same as using native similarities!**
    - IndicBERT & MuRIL weights: ≈0
- Suggests that MBERT/XLM's larger training corpora create vector representations more dispersed in high-D space
- More "space" available to transform in new semantic representations
- IndicBERT/MuRIL representations clustered in tight, high-D "cone"

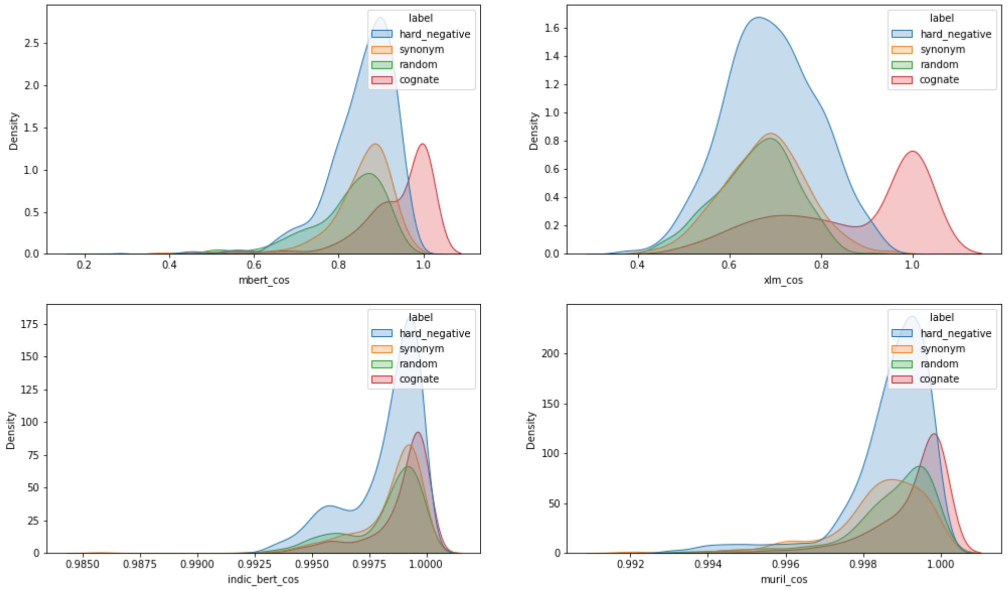


F1(+) with different semantic feature sets compared to phonetic edit distance baseline

KDE/Pair Plots for Native Embeddings

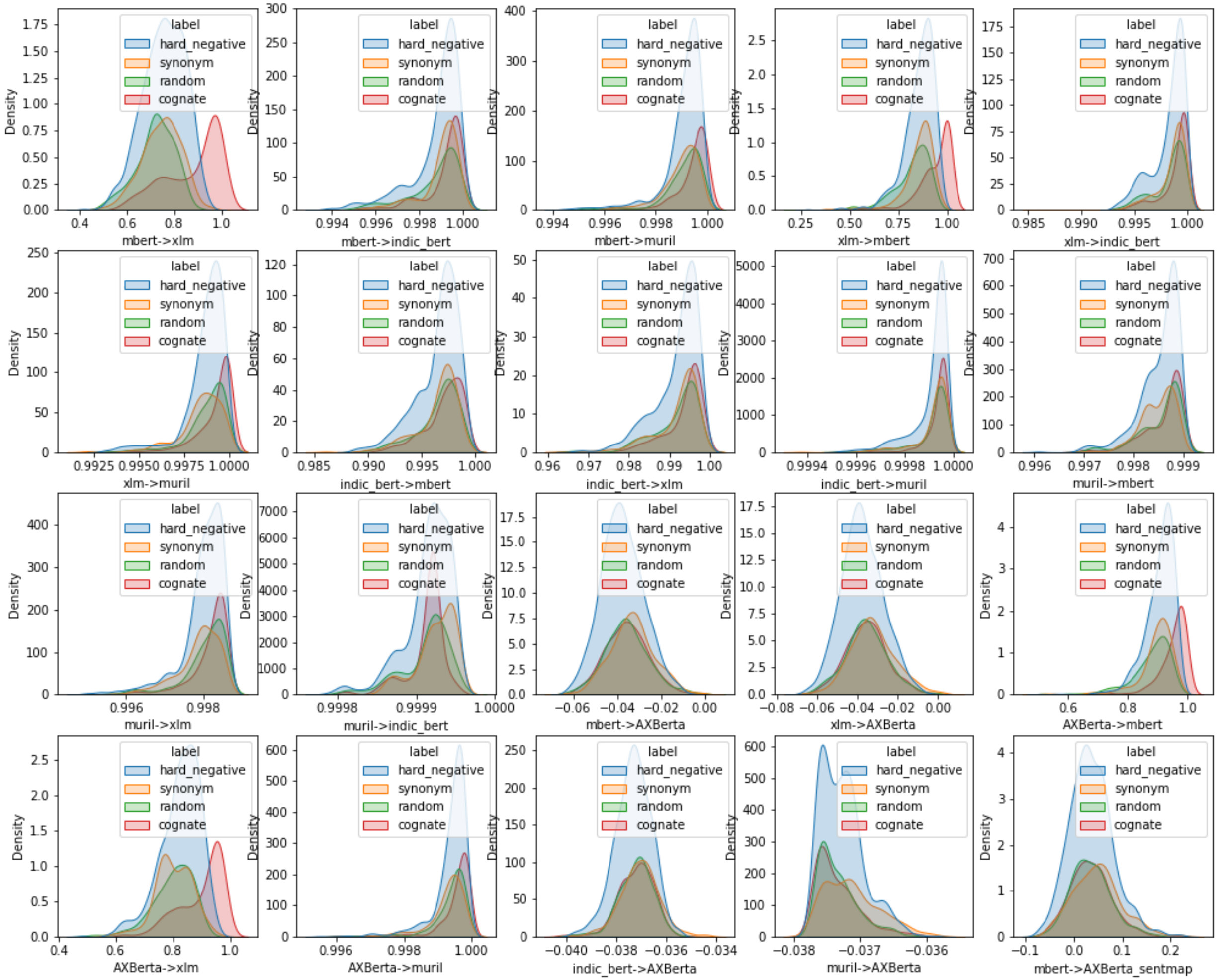


- All native embeddings encode vector semantics
- Cognate cosine similarity > Synonym cosine similarity
- Cognates distributed in distinct space
- Larger models → more dispersed space



KDE Plots for Affine-Mapped Embeddings

- Larger models → more semantic transfer into distinct space
- Smaller models → no such distinct distribution
- XLM-R → Assamese-ALBERT: little transfer, no distinct space
- Assamese-ALBERT → XLM-R: high-fidelity semantic transfer, more than MBERT
- Mapped cosine similarities similar to native cosine similarities
- XLM-R supports Assamese, MBERT does not

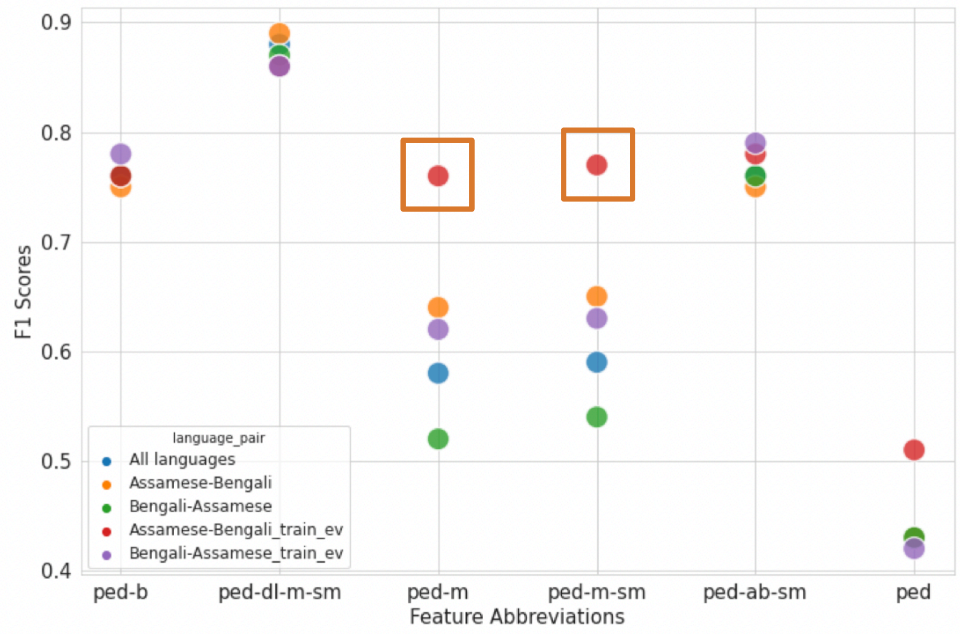


Influence of Features: Semantic

- Adding **sentence mappings** only slightly improves overall performance
  - Combination of word-level and sentence mappings most effective for pair-specific models
  - `as-bn`: largely due to reducing retrieved hard negatives
  - Affine mappings introduce semantic information to help disambiguation

	*bn-as**			*as-bn**		
	`ped`	`pm`	`psm`	`ped`	`pm`	`psm`
HN	31	48	45	47	10	15
Syn.	0	4	4	6	8	6
Rnd.	0	7	2	0	2	0

Pair-specific model false positives using `ped`, `ped-m` (`pm`), and `ped-m-sm` (`psm`) feature sets broken down by negative example type



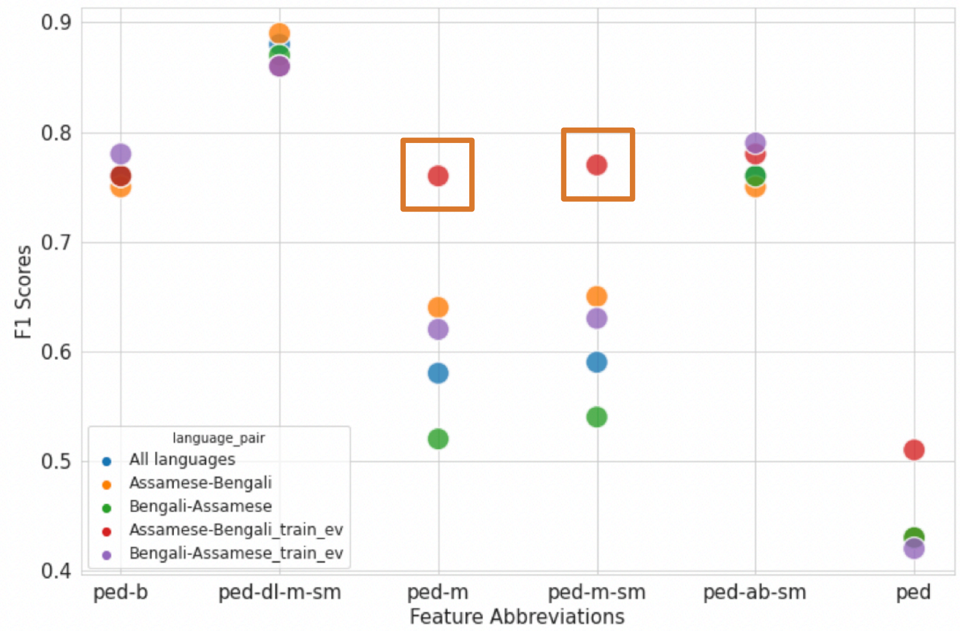
F1(+) with different semantic feature sets compared to phonetic edit distance baseline

Influence of Features: Semantic

- False negatives reduced for both pair-specific models
- Suggests geometric transformation of embeddings is useful on multiple levels
- Bringing in information specific to Assamese

	bn-as*			***as-bn****		
	`ped`	`pm`	`psm`	`ped`	`pm`	`psm`
FN	212	140	138	182	106	100

Pair-specific model false negatives using `ped`, `ped-m` (`pm`), and `ped-m-sm` (`psm`)



F1(+) with different semantic feature sets compared to phonetic edit distance baseline

Conclusion

- We have presented a high-performing method for detecting cognates between Assamese and Bengali
- Methods applied here should apply to other languages
  - Similar techniques applied to loanword detection in main conference paper (Nath et al., 2022)
  - Articulatory alignment most informative feature
  - **Unique to this paper**: affine transformation between LM embedding spaces
- Tests on different semantic representations suggest:
  1. linearly transforming vectors between model embedding spaces carries certain semantic information with high fidelity
  2. low-resource model can be mapped to a richer model's space
- **If these hypotheses hold**, transformed embeddings from a low-resourced LM can reduce computational cost involved in training and improve downstream NLP

Future Work

- Collecting putative cognates is essential in computational historical linguistics
  - Our alignment method could be adapted
    - to find regular correspondences (e.g., by training individual attention weights over a sequence)
    - to identify shared innovations
    - to reconstruct earlier word forms to reconstruct proto-languages (Bouchard-Côté et al., 2013; Jäger, 2019)
- Applications of linear mapping technique to other tasks, e.g., coreference resolution
- Further evaluating monolingual Assamese model on tasks, e.g., question answering

Thank you!

{abhijnan.nath,nkrishna}@colostate.edu

Colorado State University

References

- 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.
- Gerhard Jäger. 2019. Computational historical linguistics. *Theoretical Linguistics*, 45(3-4):151–182.
- Grzegorz Kondrak. 2001. Identifying cognates by phonetic and semantic similarity. In Second Meeting of the North American Chapter of the Association for Computational Linguistics.
- David R Mortensen, Siddharth Dalmia, and Patrick Littell. 2018. Epitran: Precision G2P for many languages. In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*.
- David R Mortensen, Patrick Littell, Akash Bharadwaj, Kartik Goyal, Chris Dyer, and Lori 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.
- Abhijnan Nath, Sina Mahdipour Saravani, Ibrahim Khebour, Sheikh Mannan, Zihui Li, and Nikhil Krishnaswamy. (2022). A Generalized Method for Automated Multilingual Loanword Detection. In *International Conference on Computational Linguistics (COLING)*. ACL.