Not-so-true to the Origins: Language Clusters in Modern Times

Anonymous ACL submission

Abstract

The shared lexicon can reveal genealogical relationships between languages in a linguistic area. However, widespread cross-linguistic borrowings have increasingly blurred traditional phylogenetic distinctions based on lexical similarities, leading to a distorted perception of language clusters based on prior diachronic knowledge. To better understand language clusters at a synchronic level, including the influence of borrowings, this study investigates the relatedness of 9 Indic languages by leveraging the lexical knowledge of pre-trained language models: mBERT, XLM-RoBERTa, IndicNLP, and MuRIL. We extract the embeddings of cognate reflexes from the CogNet dataset for the selected languages. By performing hierarchical agglomerative clustering on the embedding-based cosine similarity scores of language pairs, we identify language clusters that reflect contemporary language groupings, carefully considering the impact of borrowings. This study also aims to assess how well word embedding-based lexical similarity aligns with string similarity-based genealogical clustering and the actual phylogenetic groupings. The results demonstrate that cognates play a crucial role in extracting phylogenetic signals, even when using pre-trained language models.

1 Introduction

Most often, the historical relatedness of languages comes from inferences of linguistic phylogenies in historical-comparative linguistics using word sets that share a common origin regardless of their meaning, and barring borrowed words (Trask, 2000; Campbell, 2020; List et al., 2022). Such sets of cognates are composed of tuples of cognate *reflexes* pertaining to basic vocabulary items such as body parts, colors, numbers, etc. This is a *de facto* choice given the resistance of such words to be affected by the process of borrowing. And, as a result, we have language groupings that reflect linguistic histories but not necessarily the present.

There are cases of the basic vocabulary items also being borrowed, for example, the complementizer 'that' in Hindi 'ki' has been borrowed from Classical Persian 'ki'. Such words have been so well integrated into the linguistic system of the language that they cannot be teased apart from the native vocabulary items, therefore it is worthwhile to include borrowings to infer *phylogenetic* signals based on lexical data.

For handling such cases computationally, there have been attempts to simplify the definition of cognates, following Kondrak et al. (2003) which states that two words in different languages are cognates if they have the same meaning and present a similarity in orthography, resulting from a *proposed* underlying etymological relationship due to common ancestry or borrowing. CogNet v2 (Batsuren et al., 2019) is curated on the basis of this definition and therefore is a suitable choice for our experiments.

For our experiments, we use *contextual* embeddings of cognate reflexes by studying the embedding representations of pre-trained language models (PLMs): mBERT, XLM-RoBERTa, IndicNLP, and MuRIL. The last two models were specifically adapted to process Indic languages. These contextualized embeddings can recover more lexical relation knowledge than static embeddings, and therefore have a substantial amount of lexical knowledge (Vulić et al., 2020), which makes them ideal for our study. The research questions that we pose are:

RQ1. Can word embeddings from PLMs be used to capture phylogenetic information?

RQ2. Is considering an embedding-based representation of cognate reflexes useful in extracting signals of language-relatedness?

We conducted experiments for 9 Indic languages representing the Indo-Aryan (Bengali, Gujarati,

Concept ID	Gloss	Languages							
		Bengali	Gujarati	Hindi	Marathi	Kannada	Malayalam	Tamil	Telugu
n03046257	clock	ghori	ghadiyal	ghari	ghadyal	gadiyara	ghatikaram	kattikaram	gadiyaram
n00729285	lesson	path	-	path	-	patha	patham	patam	pathyamu
n07873464	pilaf	polao	pulav	pulav	pulav	palav	biriyani	pulav	pulav
n03580615	internet	intaronet	internet	internet	internet	antarjala	inrarnerr	intarnet	intarnet

Table 1: Cognate reflexes for some of the selected languages in CogNet. The concepts are labeled using IDs and transliterations for the reflexes are provided.

Hindi, Punjabi, and Marathi) and Dravidian (Kannada, Malayalam, Tamil, and Telugu) language families. We highlight that borrowings influence language clusters, which we demonstrate by dendrograms generated through agglomerative hierarchical clustering of language distances based on cosine similarities between language pairs.

2 Related Work

There has been a significant amount of research done in inferring phylogenetic signals through phoneme sequence comparison in historical linguistics such as Kondrak (2000), List (2014, 2016), and List et al. (2017).

Rosa and Žabokrtský (2015) introduce a language similarity measure based on distributions of coarse POS tags in the source and target POS-tagged corpora for delexicalized parsing.

Bella et al. (2021) compute pairwise similarity for 338 languages from CogNet (Batsuren et al., 2019). They also stress the handicap of traditional comparative-historical linguistic methods of using basic vocabulary items as they do not provide information about the present state of lexicons.

3 Data and Resources

We use CogNet[1] version 2.0, a large-scale multilingual cognate database. It contains 8.1 million sense-tagged word pairs in 338 languages and 35 writing systems.

Table 1 provides a brief view of the data. The concept n03046257 "clock" can be found in the cited languages, whereas there are also concepts such as n00729285 "lesson". However, they are present in languages like Gujarati and Marathi but are not captured in CogNet. Words like "pilaf" of Dravidian origin according to Wiktionary can have different equivalents such as *biriyani* (which is of Persian origin) in Malayalam. This again indicates the relaxed definition of cognates used in CogNet. Furthermore, we also find examples where one language uses the calque of English words such as Kannada *antarjala* for "internet".

We extract contextualized word embeddings for cognate reflexes for every ID of concepts[2] from the PLMs. The PLMs trained on huge data have already learned a lot about language structure and semantics, making the features they produce highly informative. We extract word embeddings from **mBERT** (Devlin et al., 2018), **XLM-RoBERTa** (Conneau et al., 2019), **IndicBERT** (Kakwani et al., 2020), and **MuRIL** (Khanuja et al., 2021). IndicBERT is a multilingual ALBERT (Lan et al., 2020) model pre-trained exclusively in 12 major Indic languages, and MuRIL is a pre-trained BERT model in 17 Indic languages and their transliterated counterparts. We limit our analysis to 9 Indic languages, as all the selected models have been pre-trained in these languages.

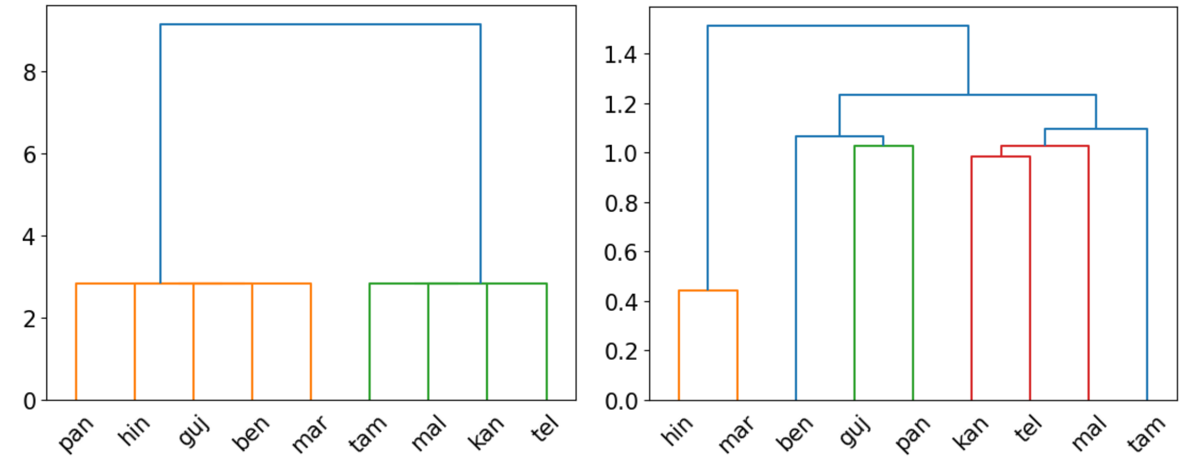
4 Experiments

The experimental setup[3] mainly comprises the extraction of contextualized embeddings. The motivation is that these dense vector representations make it easier to perform cross-lingual lexical similarity calculations without language-specific normalizations like transliteration, etc. For extracting the embeddings from the PLMs, we tokenize the words using the respective tokenizer of the PLMs and extract the *contextual* representations, i.e. we process each word in isolation, which may not capture its intended meaning in real language use. However, this is not a concern for our experiments, since cognate sets are initially curated based on semantic concepts, and we assume that the PLMs would still provide useful embeddings. We take the output of the last hidden state and get the embedding of each word by averaging the sub-word embeddings.

[1] http://cognet.ukc.disi.unitn.it

[2] Sometimes for a given concept there is a tuple of synonyms available in the data, but for our experiments, we randomly select only one of them.

[3] Python libraries like spaCy and scikit-learn were used for our experiments, along with ChatGPT.



(a) GOLD

(b) PAIR

Figure 1: The language clusters are derived from mBERT. The Glottolog-based GOLD clustering identifies two primary groups: Indo-Aryan and Dravidian. The PAIR clustering highlights modifications in the language groupings.

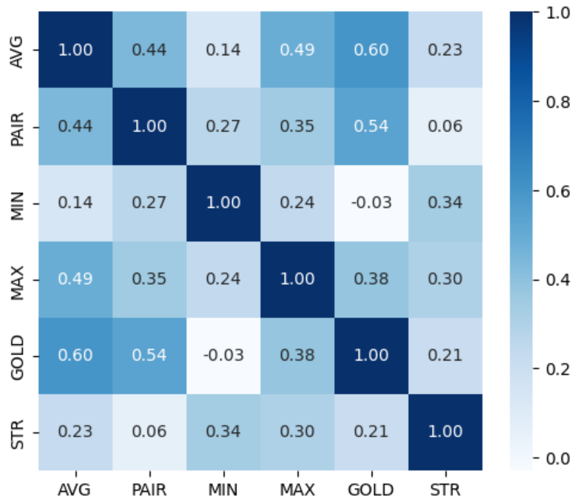


Figure 2: The Spearman's rank correlation for all the experimental setups for mBERT.

Our experimental design is composed of 4 different embedding-based methods and 2 control setups:

AVG: Here, the embeddings for all the words in a language present in CogNet were extracted and averaged, and then the cosine similarity across the thus obtained *language* embeddings were calculated for the language pairs. In this method, CogNet serves as a wordlist for the individual languages.

PAIR: Here, we considered the embeddings of cognate reflex pairs, computing their cosine similarity and averaging the values to obtain a single similarity score for each language pair. As shown in Table 1, some concepts lack cognate reflexes. Therefore, for each language pair, only concepts with available cognate reflexes were included in the calculation. This is also true for the following embedding-based methods.

MIN: For each word in a given language, the most distant pair was identified based on the minimum cosine similarity across all words in the other language, without explicitly considering the corresponding cognate reflex. The resulting scores were then averaged to mitigate data sparsity.

MAX: For each word in a given language, the closest pair was identified based on the maximum cosine similarity across all words in the other language, without explicitly considering the corresponding cognate reflex. The final score was then obtained by averaging across all such tuples for language pairs.

GOLD: For comparing the thus obtained clusterings or dendrograms based on embeddings, we induce a *phylogenetic* tree for the selected languages based on Glottolog[4]. The distances between the languages are assigned based on their membership within the same language families and subfamilies. For example, Tamil (`tam`) and Hindi (`hin`) are assigned a language similarity score of 0.5 because they belong to different families, but Tamil and Malayalam (`mal`) are assigned a score of 1 as they are both Dravidian languages. The choice of the scores is arbitrary.

STR: We also want to compare our embedding-based clusters with string similarity-based ones;

[4] https://glottolog.org/

hence we use the lexical similarity scores[5] provided by Bella et al. (2021). They calculate the cognate-based similarity between languages using Levenshtein distance on words, with Latin transliterations for different scripts. A smoothing factor prevents excessive penalization of dissimilar cognates. The similarity score is normalized by the harmonic mean of lexicon sizes to account for differences in vocabulary size and completeness.

4.1 Results: Constructing Phylogenetic Trees from Embeddings

The negative logarithms for the cosine similarities obtained from all 4 embedding-based methods were taken to represent the language distances and the dendrograms (See Appendix A) were constructed using Ward's minimum variance[6] method. Figure 1 illustrates the visualization for mBERT-based language clusters. It shows that mBERT to a fair extent captures phylogenetic information. For example, all Dravidian languages Kannada (`kan`), Telugu (`tel`), Tamil, and Malayalam are part of the same cluster. The Indo-Aryan languages form two exclusive groups, one with Hindi and Marathi (`mar`) and the other with Bengali (`ben`), Gujarati (`guj`), and Punjabi (`pan`).

4.2 Results: Correlations with Glottolog

We calculated the Spearman rank correlation for the language distances obtained from all experimental setups (See Appendix A). We find positive correlations for embedding-based methods for mBERT with GOLD, except for MIN (Figure 2), suggesting that mBERT is sensitive to cognates or, in other words, it captures well the phonological and semantic information of cognate reflexes, also indicated by higher correlations in the case of AVG and PAIR with GOLD. The negative correlation of MIN with the control methods indicates that semantically closer translation equivalents are not ideal candidates to align with phylogenetic clusters, whereas cognates are more suitable.

5 Discussion

Concerning our RQ1, we find that the contextual embeddings from PLMs can be used to infer the phylogenetic signals and especially to observe the impact of borrowings. The use of cognates in traditional methods builds upon the similarity of phonemes in cognate reflexes, and PLMs do seem to encode that information on par with the semantic information. All of our results are based on cognate reflexes, which serve as the prime component for extracting such signals. The AVG and PAIR having higher correlations with GOLD than STR, show that cognate reflexes do enable the mapping of the phylogenetic relations, answering our RQ2.

6 Conclusion

This study clusters 9 Indic languages based on lexical similarities using contextual embeddings of cognate reflexes. We experiment with 4 PLMs and apply agglomerative hierarchical clustering based on language distances derived from different methods using cosine similarity. The resulting dendrograms reveal modern-day language groupings. Our findings suggest that contextual embeddings can effectively capture phylogenetic signals, with cognates playing a crucial role in this process.

7 Limitations

Currently, the choice of languages is confined to the language pre-trained in all the PLMs that we use. We rely on the fact that these PLMs are efficient enough to produce the embeddings without giving the cognates any sentential context and that the embedding representations extracted from the last output layer are optimal.

References

Khuyagbaatar Batsuren, Gabor Bella, and Fausto Giunchiglia. 2019. CogNet: A large-scale cognate database. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 3136–3145, Florence, Italy. Association for Computational Linguistics.

Gábor Bella, Khuyagbaatar Batsuren, and Fausto Giunchiglia. 2021. A database and visualization of the similarity of contemporary lexicons. In *Text, Speech, and Dialogue*, pages 95–104, Cham. Springer International Publishing.

Lyle Campbell. 2020. *Historical Linguistics: An Introduction*. Edinburgh University Press.

Alexis Conneau, Kartikay Khandelwal, Naman Goyal, Vishrav Chaudhary, Guillaume Wenzek, Francisco Guzmán, Edouard Grave, Myle Ott, Luke Zettlemoyer, and Veselin Stoyanov. 2019. Unsupervised cross-lingual representation learning at scale. *CoRR*, abs/1911.02116.

[5] http://ukc.disi.unitn.it/index.php/lexsim/

[6] The Ward's linkage method aims to create clusters that are compact and well-separated by minimizing the spread of data points within clusters.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2018. BERT: pre-training of deep bidirectional transformers for language understanding. *CoRR*, abs/1810.04805.

Divyanshu Kakwani, Anoop Kunchukuttan, Satish Golla, Gokul N.C., Avik Bhattacharyya, Mitesh M. Khapra, and Pratyush Kumar. 2020. IndicNLPSuite: Monolingual Corpora, Evaluation Benchmarks and Pre-trained Multilingual Language Models for Indian Languages. In *Findings of EMNLP*.

Simran Khanuja, Diksha Bansal, Sarvesh Mehtani, Savya Khosla, Atreyee Dey, Balaji Gopalan, Dilip Kumar Margam, Pooja Aggarwal, Rajiv Teja Nagipogu, Shachi Dave, Shruti Gupta, Subhash Chandra Bose Gali, Vish Subramanian, and Partha Talukdar. 2021. Muril: Multilingual representations for indian languages. *Preprint*, arXiv:2103.10730.

Grzegorz Kondrak. 2000. A new algorithm for the alignment of phonetic sequences. In *1st Meeting of the North American Chapter of the Association for Computational Linguistics*.

Grzegorz Kondrak, Daniel Marcu, and Kevin Knight. 2003. Cognates can improve statistical translation models. In *Companion Volume of the Proceedings of HLT-NAACL 2003 - Short Papers*, pages 46–48.

Zhenzhong Lan, Mingda Chen, Sebastian Goodman, Kevin Gimpel, Piyush Sharma, and Radu Soricut. 2020. ALBERT: A Lite BERT for Self-supervised Learning of Language Representations. *Preprint*, arXiv:1909.11942.

Johann-Mattis List. 2014. *Sequence comparison in historical linguistics*, volume 1. Walter de Gruyter GmbH & Co KG.

Johann-Mattis List. 2016. Beyond cognacy: Historical relations between words and their implication for phylogenetic reconstruction. *Journal of Language Evolution*, 1(2):119–136.

Johann-Mattis List, Simon J Greenhill, and Russell D Gray. 2017. The potential of automatic word comparison for historical linguistics. plos one 12 (1). e0170046.

Johann-Mattis List, Ekaterina Vylomova, Robert Forkel, Nathan Hill, and Ryan Cotterell. 2022. The SIGTYP 2022 shared task on the prediction of cognate reflexes. In *Proceedings of the 4th Workshop on Research in Computational Linguistic Typology and Multilingual NLP*, pages 52–62, Seattle, Washington. Association for Computational Linguistics.

Rudolf Rosa and Zdeněk Žabokrtský. 2015. KLcpos3 - a language similarity measure for delexicalized parser transfer. In *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing (Volume 2: Short Papers)*, pages 243–249, Beijing, China. Association for Computational Linguistics.

R.L. Trask. 2000. *The Dictionary of Historical and Comparative Linguistics*. Fitzroy Dearborn.

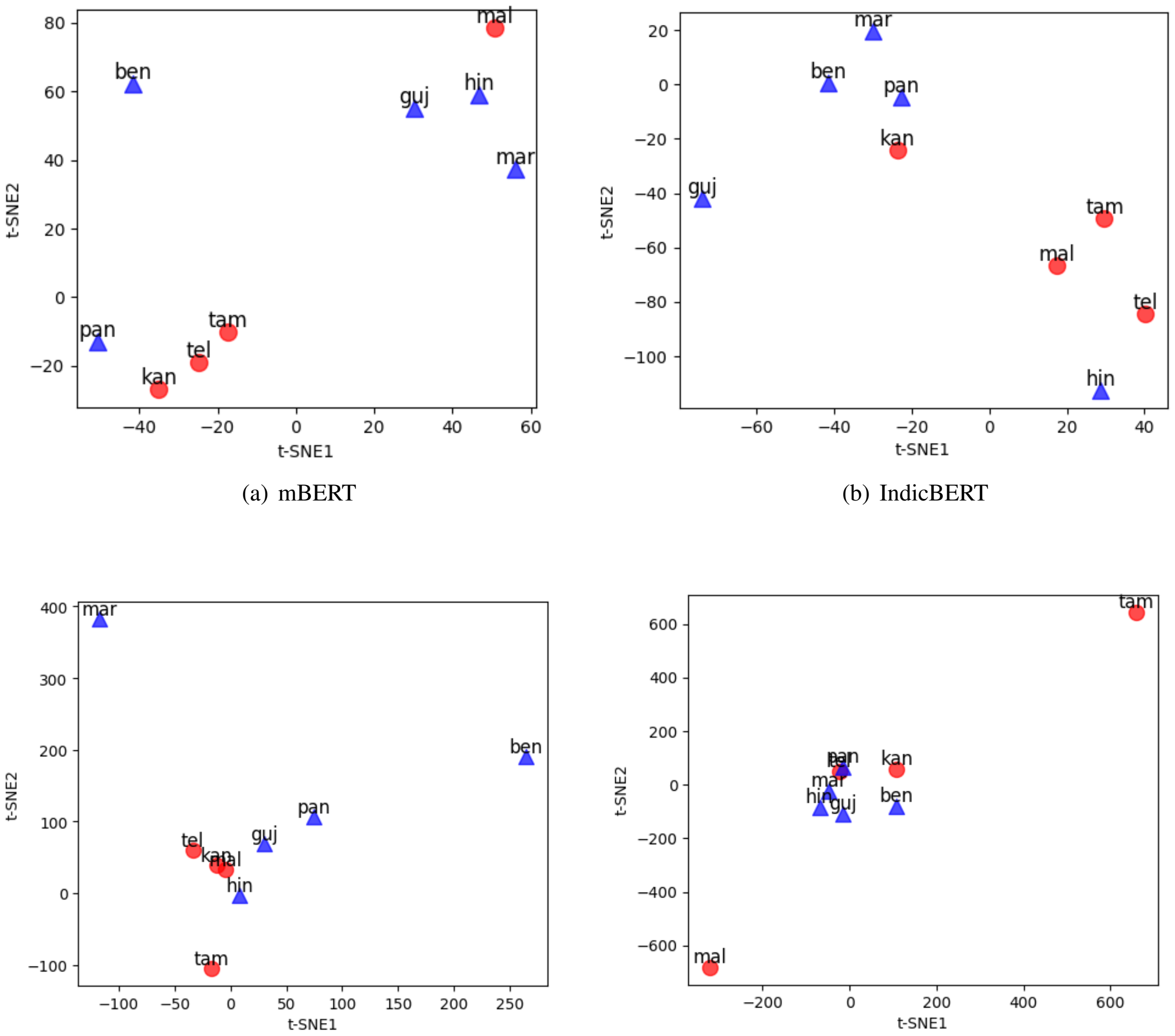
Ivan Vulić, Edoardo Maria Ponti, Robert Litschko, Goran Glavaš, and Anna Korhonen. 2020. Probing pretrained language models for lexical semantics. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 7222–7240, Online. Association for Computational Linguistics.

A Appendix

Since the CogNet is based on cognate reflexes for various semantic concepts, we examine how PLMs represent some of these reflexes (Figure 3). Additionally, we visualize the shared embedding space of the PLMs (Figures 4, 5, 6, and 7).

To interpret these results, we consider the training data used for pre-training each model. mBERT was trained on large Wikipedia corpora; XLM-RoBERTa on 2.5TB of filtered CommonCrawl data; IndicBERT on a combination of news-domain crawls and Wikipedia; and MuRIL on CommonCrawl, Wikipedia, and additional machine translation datasets for Indic languages. The results (Figure 8) show that only the mBERT-based setups exhibit higher correlations with GOLD compared to other models. This likely arises because Wikipedia articles contain relatively fewer borrowed words in Indic languages compared to CommonCrawl or news-domain data, where borrowed English vocabulary and neologisms are more prevalent.

The dendrograms (Figures 9, 10, 11, and 12) visually capture the synchronic clustering of languages, revealing a *shift* from the phylogenetic classification found in GOLD or Glottolog. This shift can be attributed to the substantial influx of foreign vocabulary into Indian languages, particularly from English. At the same time, these visualizations also reflect how PLMs structure their multilingual representation space.



(a) mBERT

(b) IndicBERT

(c) MuRIL

(d) XLM-RoBERTa

Figure 3: The word-embeddings for the cognate reflexes encoding the meaning ‘name’ in all the 9 languages. ● are the Dravidian languages and ▲ are the Indo-Aryan languages. The language codes used belong to ISO 639-3. The t-SNE plot were created using scikit-learn with perplexity of 1 and 250 iterations.

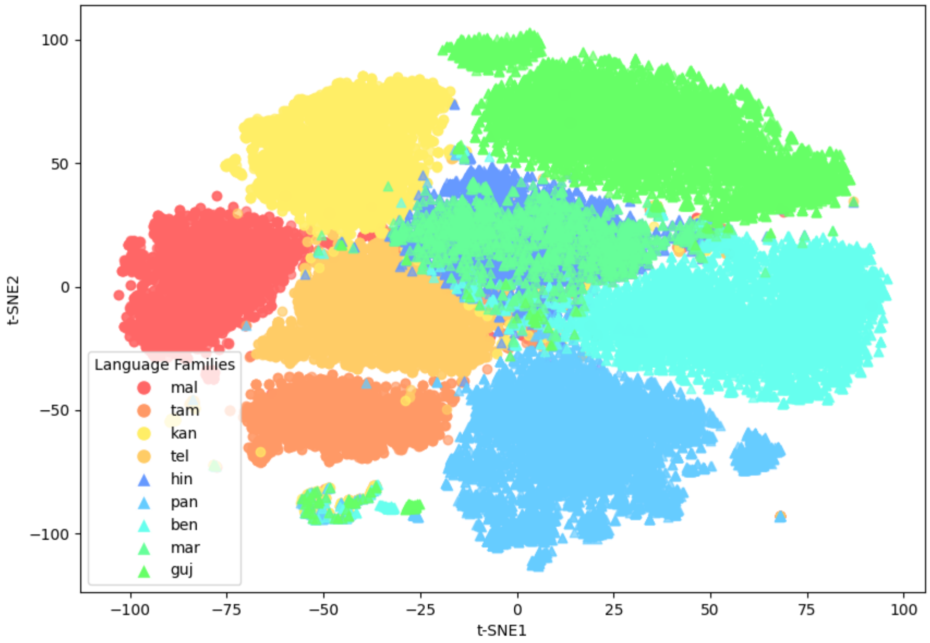


Figure 4: t-SNE representation of cognate embeddings for mBERT in the shared embeddings space. The plot was constructed with perplexity of 30, and 1000 iterations. The left section of the plot is dominated by the Dravidian languages represented with ● and the right section by Indo-Aryan languages, represented by ▲. The language codes used belong to ISO 639-3.

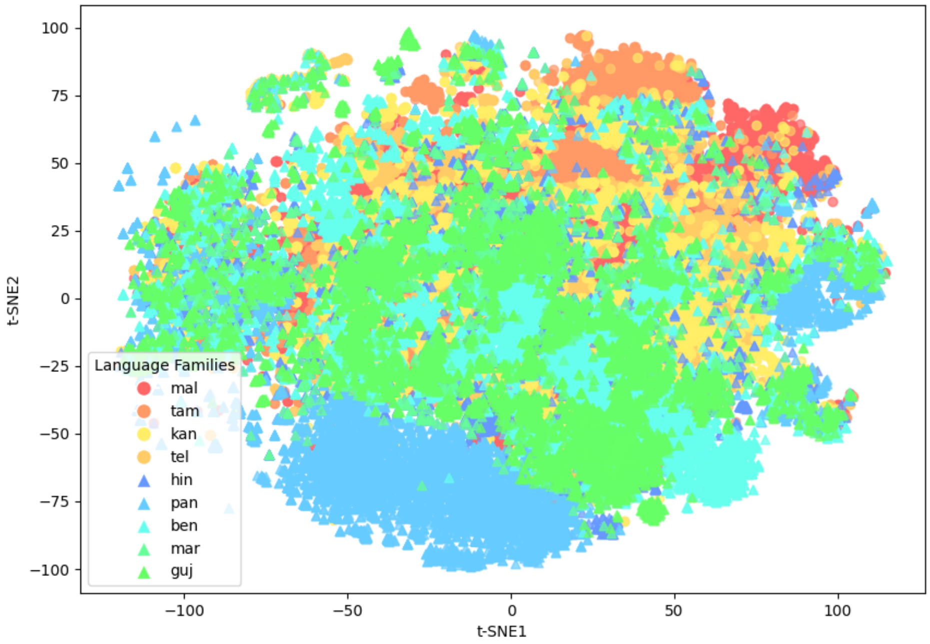


Figure 5: t-SNE representation of cognate embeddings for IndicBERT in the shared embeddings space. The plot was constructed with perplexity of 30, and 1000 iterations. The left section of the plot is dominated by the Dravidian languages represented by ● and the right section by Indo-Aryan languages, represented by ▲. The language codes used belong to ISO 639-3.

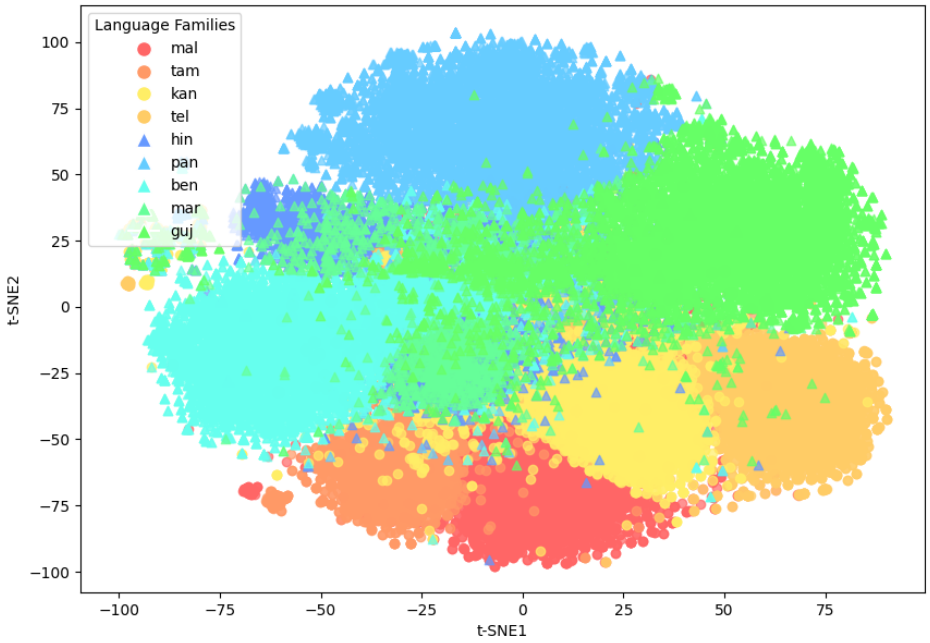


Figure 6: t-SNE representation of cognate embeddings for MuRIL in the shared embeddings space. The plot was constructed with perplexity of 30, and 1000 iterations. The left section of the plot is dominated by the Dravidian languages represented by ● and the right section by Indo-Aryan languages, represented by ▲. The language codes used belong to ISO 639-3.

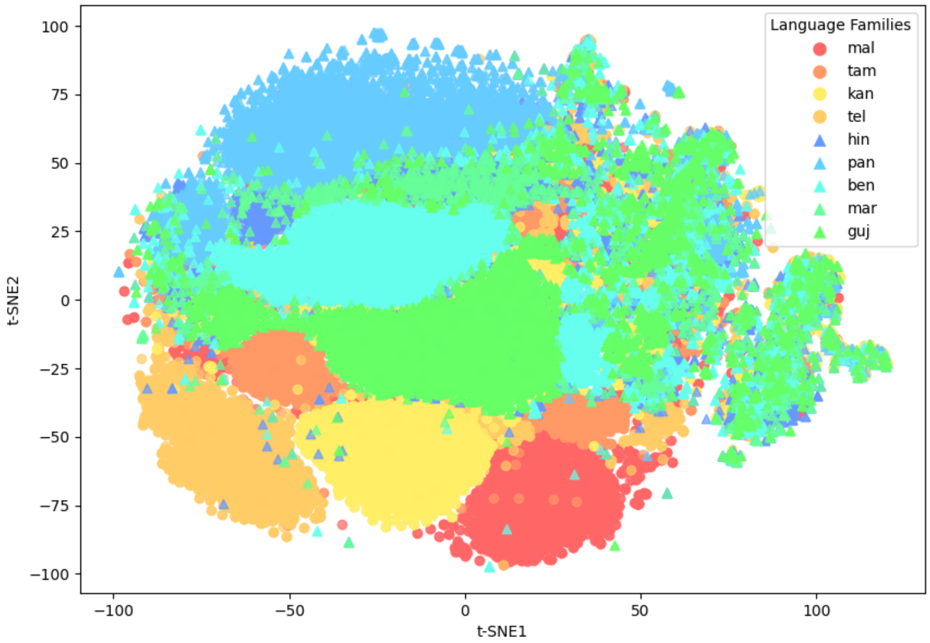
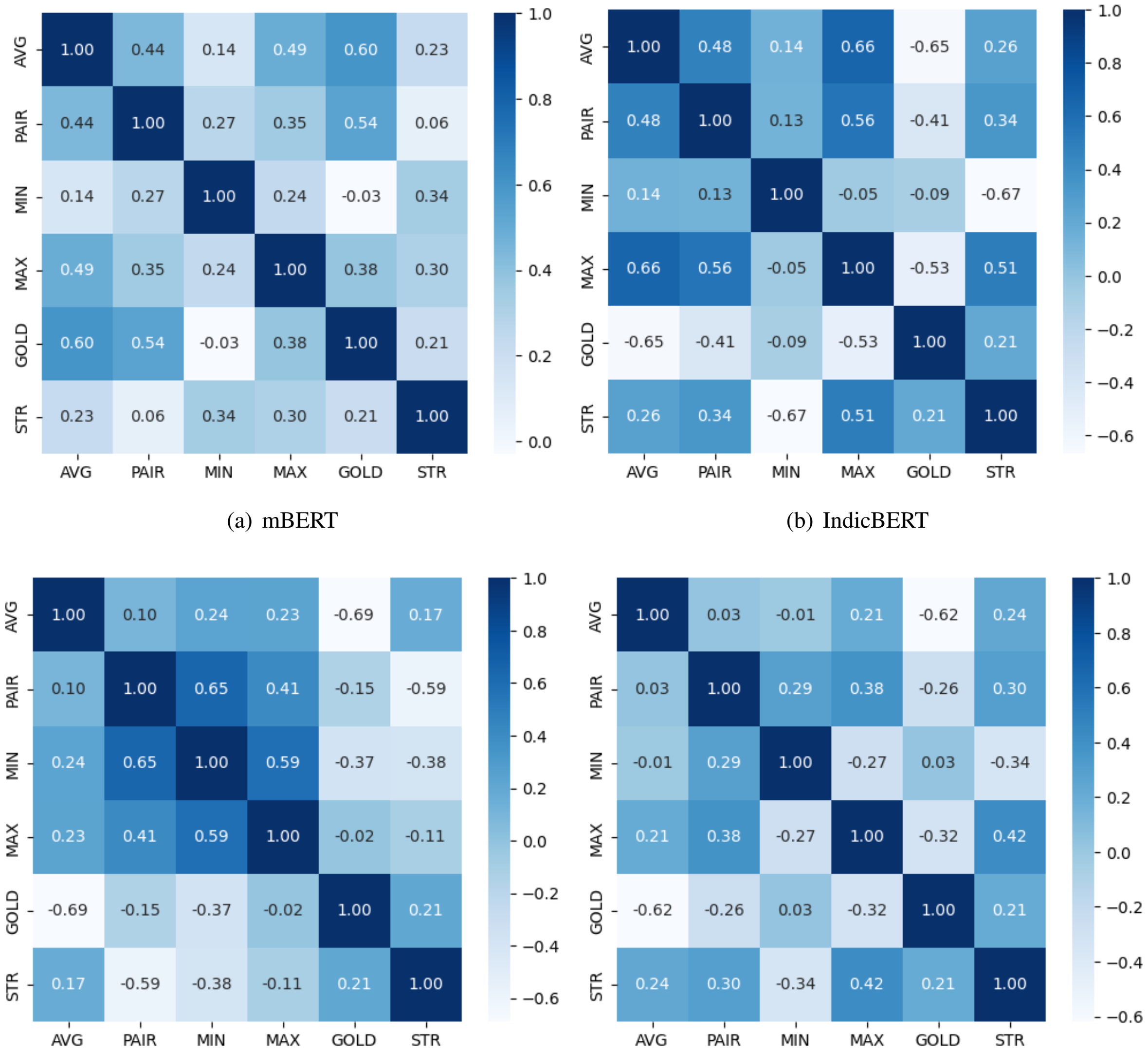


Figure 7: t-SNE representation of cognate embeddings for XLM-RoBERTa in the shared embeddings space. The plot was constructed with perplexity of 30, and 1000 iterations. The left section of the plot is dominated by the Dravidian languages represented by ● and the right section by Indo-Aryan languages, represented by ▲. The language codes used belong to ISO 639-3.



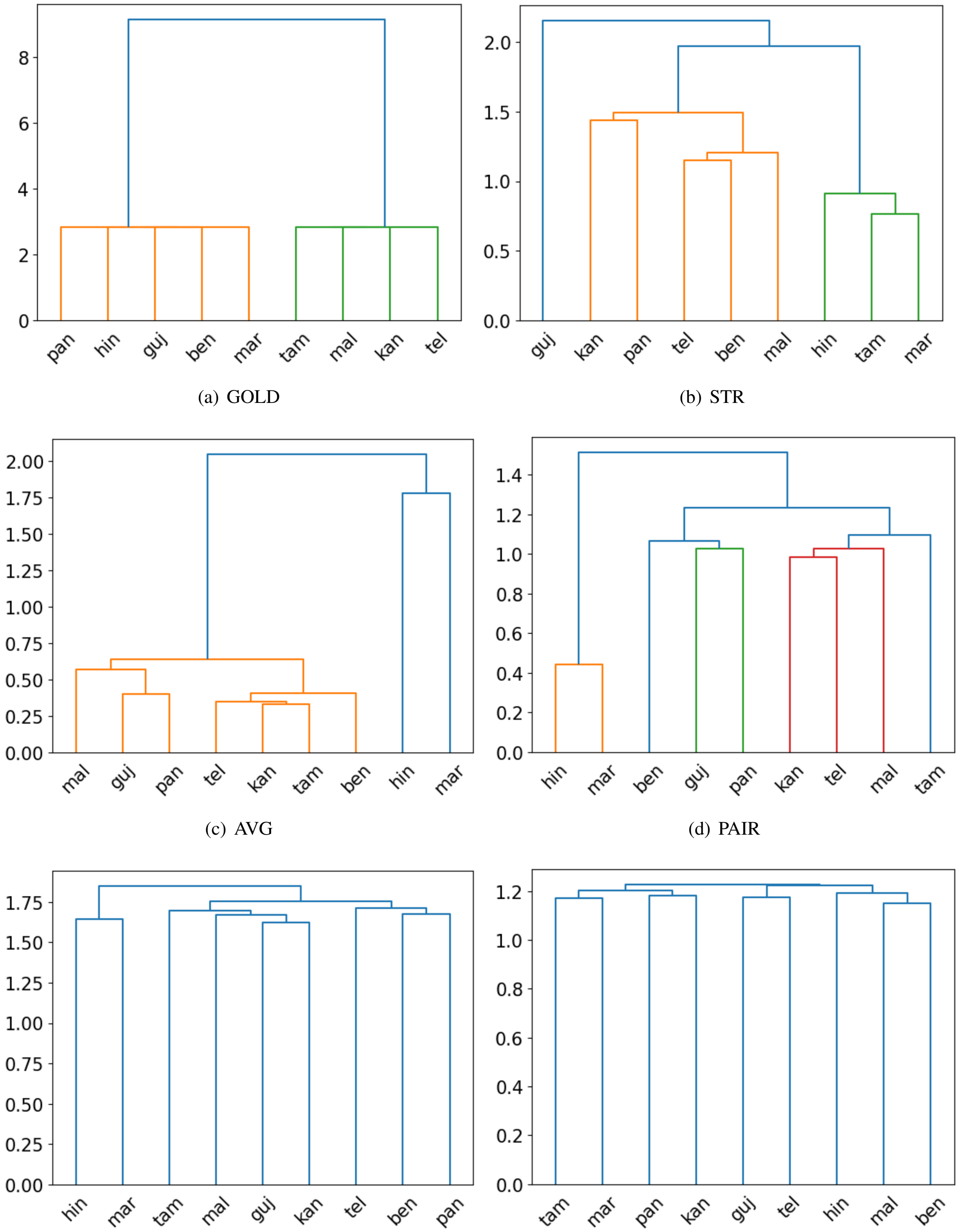
(a) mBERT

(b) IndicBERT

(c) MuRIL

(d) XLM-RoBERTa

Figure 8: The Spearman's rank correlation for all the experimental setups for all the models.



(a) GOLD

(b) STR

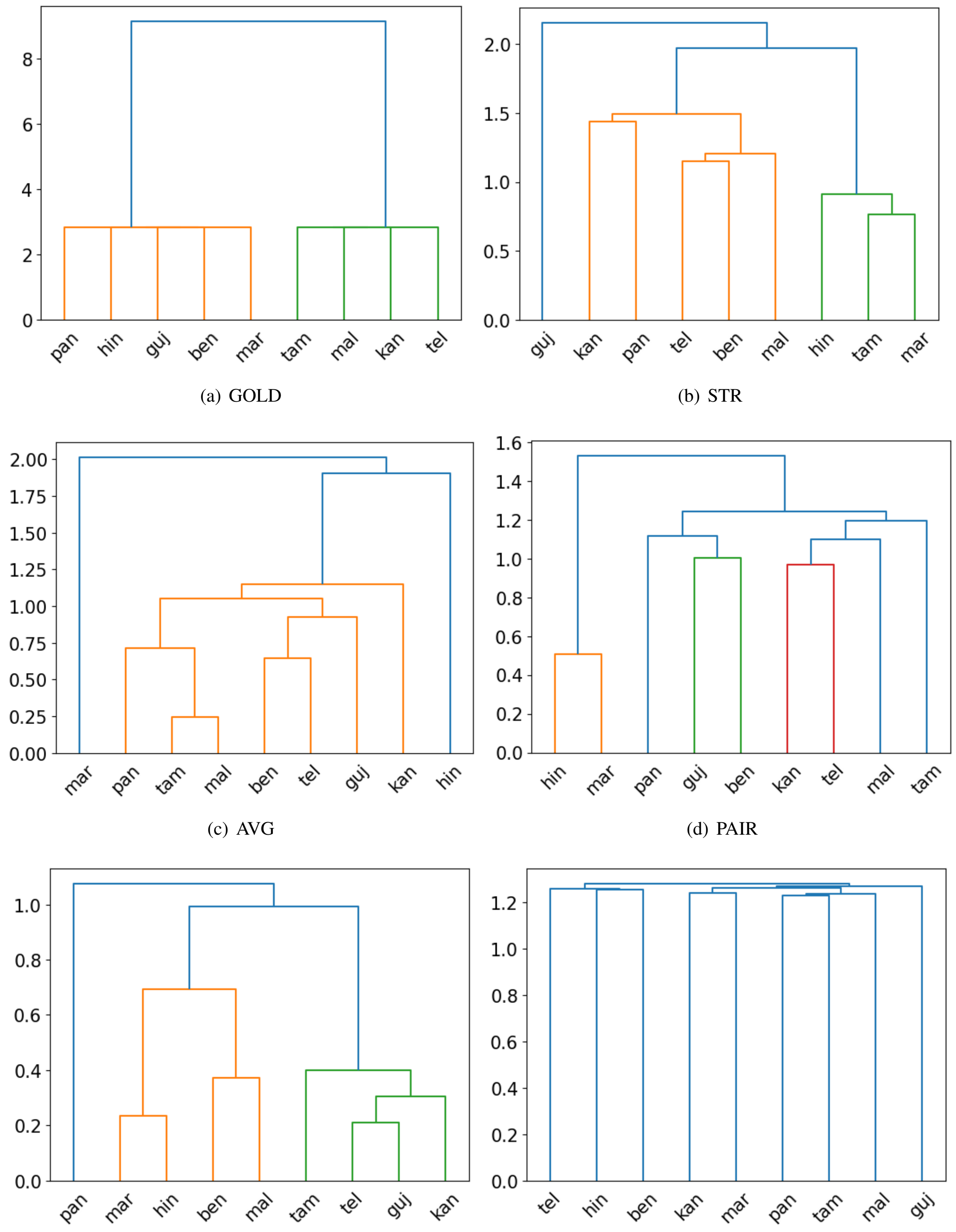
(c) AVG

(d) PAIR

(e) MIN

(f) MAX

Figure 9: The language clusters derived from mBERT.



(a) GOLD

(b) STR

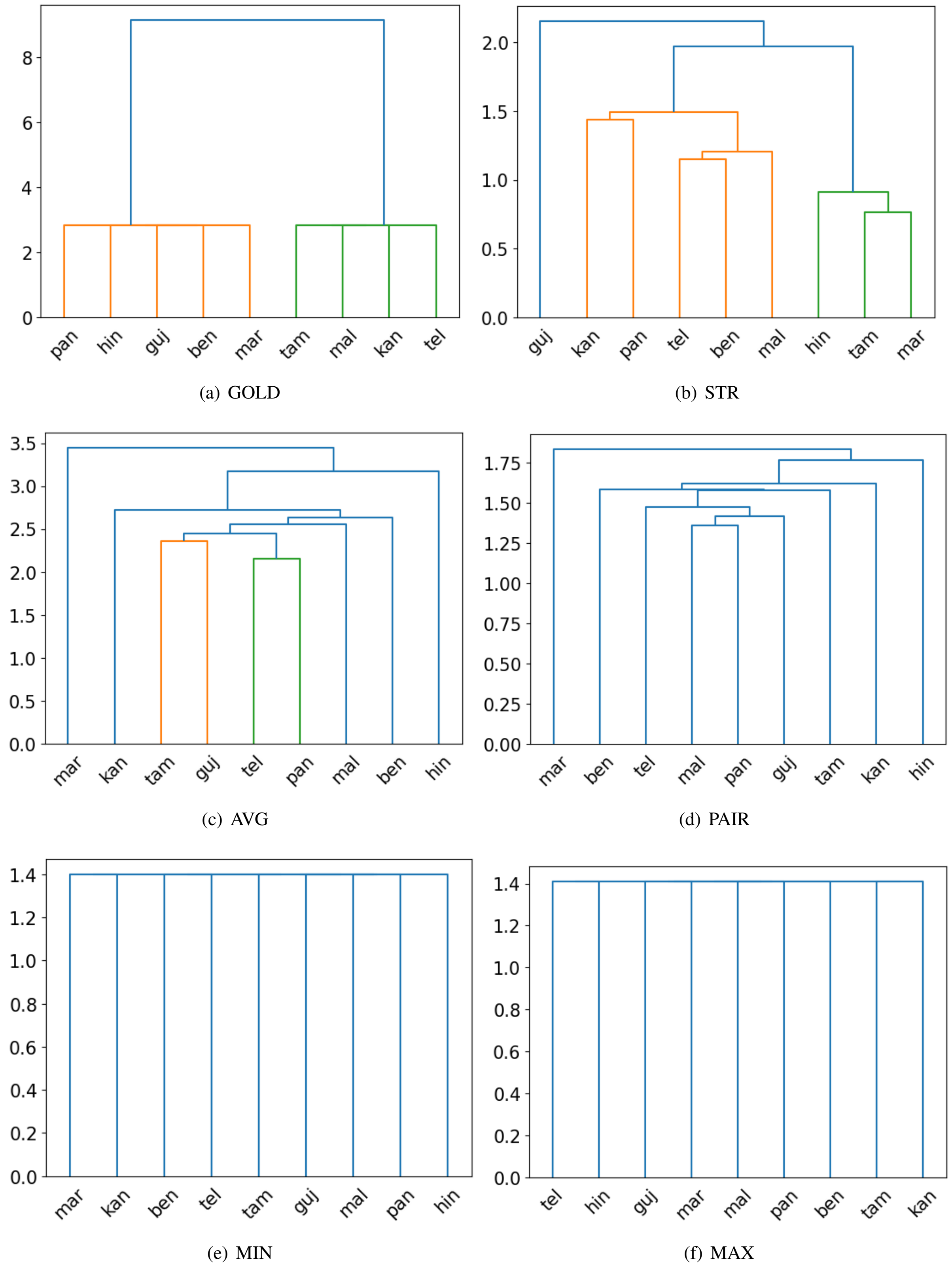
(c) AVG

(d) PAIR

(e) MIN

(f) MAX

Figure 10: The language clusters derived from IndicBERT.



(a) GOLD

(b) STR

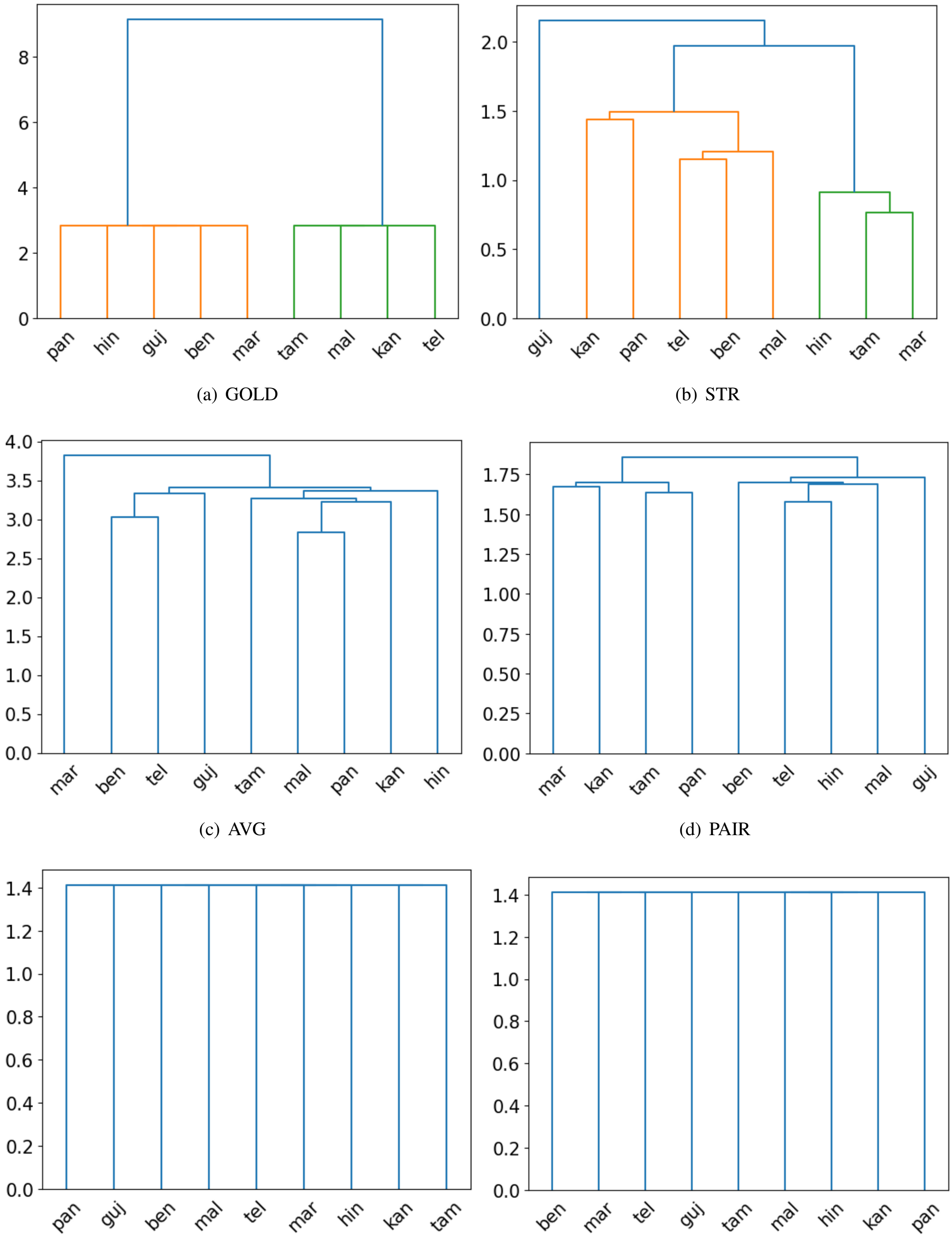
(c) AVG

(d) PAIR

(e) MIN

(f) MAX

Figure 11: The language clusters derived from MuRIL.



(a) GOLD

(b) STR

(c) AVG

(d) PAIR

(e) MIN

(f) MAX

Figure 12: The language clusters derived from XLM-RoBERTa.