Mitigating the Linguistic Gap with Phonemic Representations for Robust Cross-lingual Transfer

Haeji Jung[1], Changdae Oh[2], Jooeon Kang[3], Jimin Sohn[4], Kyungwoo Song[5], Jinkyu Kim[1], David R. Mortensen[6]

[1]Korea University, [2]University of Wisconsin-Madison, [3]Sogang University, [4]GIST, [5]Yonsei University, [6]Carnegie Mellon University

Abstract

Approaches to improving multilingual language understanding often struggle with significant performance gaps between high-resource and low-resource languages. While there are efforts to align the languages in a single latent space to mitigate such gaps, how different input-level representations influence such gaps has not been investigated, particularly with phonemic inputs. We hypothesize that the performance gaps are affected by representation discrepancies between these languages, and revisit the use of phonemic representations as a means to mitigate these discrepancies. To demonstrate the effectiveness of phonemic representations, we present experiments on three representative cross-lingual tasks on 12 languages in total. The results show that phonemic representations exhibit higher similarities between languages compared to orthographic representations, and it consistently outperforms grapheme-based baseline model on languages that are relatively low-resourced. We present quantitative evidence from three cross-lingual tasks that demonstrate the effectiveness of phonemic representations, and it is further justified by a theoretical analysis of the cross-lingual performance gap.

1 Introduction

Large language models have significantly advanced natural language processing, offering improved capabilities across numerous languages. However, substantial **performance gaps** remain, particularly between high-resource languages like English and the majority of the world's low-resource languages. While these gaps are partly driven by discrepancies in data availability and quality, recent studies suggest that **linguistic gaps**—potentially caused by structural and lexical differences—also contribute significantly to these disparities .

Cross-lingual transfer techniques, which aim to adapt to arbitrary target language, have shown promise with the advancement of pre-trained multilingual language models (Devlin et al., 2019; Conneau et al., 2020; Clark et al., 2022). However, they continue to face challenges, particularly with low-resource languages. One line of prior research has focused on mitigating these gaps through cross-lingual representation alignment (Zhang et al., 2022; Wu and Monz, 2023; Stap et al., 2023), but these efforts often overlook the impact of varying input representations on performance consistency across languages.

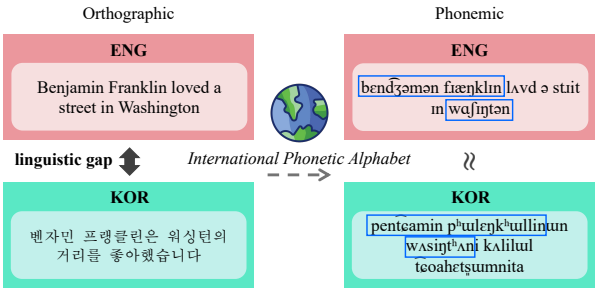

Figure 1: Example of orthographic and phonemic input representations of a sentence (English and Korean).

In this work, we explore the use of phonemic representations written in International Phonetic Alphabet (IPA) characters as a robust input representation (see Figure 1) to reduce linguistic gaps and, consequently, performance gaps across languages. We define the *linguistic gap* as the representation discrepancy between embedding vectors and the *performance gap* as the relative difference in downstream task performances between languages, to analyze the impact of phonemic representations in cross-lingual adaptation.

Our empirical analysis shows that phonemic representations consistently reduce linguistic gaps between languages compared to orthographic character-based models. This reduction in linguistic gaps directly correlates with smaller performance gaps in tasks such as cross-lingual natural

language inference (XNLI), named-entity recognition (NER), and part-of-speech (POS) tagging, demonstrating the potential of phoneme-based models to enhance cross-lingual transfer across diverse languages. We further support these findings with theoretical analysis from domain generalization literature, where we frame the performance gap as a consequence of linguistic gaps driven by lexical and syntactic differences.

Our key contributions are as follows:

- We revisit the use of phonemic representations (IPA) as a universal input strategy to reduce performance gaps across languages in multilingual language models.
- We empirically demonstrate the effectiveness of phonemic representations by comparing them with subword and character-based models, highlighting their ability to minimize both performance and linguistic gaps.
- We provide a theoretical explanation for the observed benefits of phonemic representations, drawing parallels between linguistic gaps in multilingual settings and domain gaps in domain generalization literature.

2 Related Works

2.1 Cross-lingual Transfer with Multilingual Language Model

Cross-lingual transfer learning aims to improve performance on low-resource languages (LRLs) by leveraging data from high-resource languages. Models like mBERT (Devlin et al., 2019) and XLM-R (Conneau et al., 2020), trained on hundreds of languages, have demonstrated effective cross-lingual adaptation by leveraging large multilingual pre-train datasets (Fujinuma et al., 2022; Wu and Dredze, 2020; Conneau et al., 2020). However, significant performance discrepancies remain between languages due to differences in data availability, script types, and language families (Wu and Dredze, 2020; Muller et al., 2021a; Bagheri Nezhad and Agrawal, 2024). This "performance gap" has been systematically evaluated in benchmarks such as XTREME (Hu et al., 2020), highlighting the need for methods that can ensure more consistent performance across languages.

2.2 Cross-lingual Representation Gap

One approach to reducing performance gaps focuses on narrowing the representation gap between languages. Multilingual pre-training enables models to learn shared representation space for multiple languages. (Singh et al., 2019) and (Muller et al., 2021b) both analyze the representations of pre-trained multilingual models and observe that lower layers are responsible for this cross-lingual alignment. (Yang et al., 2022) employs mixup (Zhang et al., 2018) to bring representations closer together, improving performance by reducing their distance in the latent space. Other works show a strong correlation between representation distance and machine translation performance, suggesting that improved alignment leads to better transfer results (Wu and Monz, 2023; Stap et al., 2023). While these studies provide valuable insights into the benefits of aligning cross-lingual representations, they do not explore how variations in input-level representations, such as the use of phonemic representations instead of orthographic characters, might affect this alignment. This paper investigates how phonemic representations can further reduce cross-lingual gaps.

2.3 Phonemic Representations for Multilingual Language Modeling

Phonemes, typically represented by International Phonetic Alphabet (IPA) characters, are the perceptual sounds of a language. Phonemic representations offer a language-agnostic input that can enhance multilingual modeling, especially for LRLs. By using phonological features that are less dependent on specific orthographic systems, these representations offer a language-agnostic alternative that can help bridge performance gaps across languages. Previous studies have shown that using the IPA characters as input can enhance performance in cross-lingual tasks such as named entity recognition (Chaudhary et al., 2018; Bharadwaj et al., 2016; Leong and Whitenack, 2022) and machine translation (Chaudhary et al., 2018; Sun et al., 2022), particularly for low-resource languages. Similarly, Sohn et al. (2024) report that phoneme-based models outperform other baselines on target languages unseen during pre-training. While these works demonstrate the potential of phonemic representations in language modeling, few have explored the specific embeddings and representations of phonemes. Although some studies have developed pre-defined phoneme embeddings (e.g., PanPhon (Mortensen et al., 2016), Phoible (Moran and McCloy, 2019)) and learned embeddings from masked language modeling (Li et al.,

2023; Jia et al., 2021; Sundararaman et al., 2021; Zhang et al., 2022), there is limited understanding of how these embeddings function in cross-lingual contexts.

We utilize XPhoneBERT (Nguyen et al., 2023), a model pre-trained with phonemes across approximately 100 languages, to investigate how using phonemic representations as input can mitigate cross-lingual performance discrepancies. Our empirical and theoretical analyses provide new insights into the benefits of phonemic representations for multilingual language modeling, particularly in terms of narrowing the cross-lingual linguistic gap and performance gap.

3 Experimental Setup

In this section, we describe the experiment setup in terms of models, datasets, and downstream tasks, including the selected target languages and details for preprocessing. Additionally, we provide details on evaluation strategies, particularly on quantifying the performance and linguistic gap.

3.1 Models

We employ three masked language models that are pre-trained on multilingual corpus that covers around 100 languages from Wikipedia dump files[1]: mBERT (Devlin et al., 2019), CANINE (Clark et al., 2022), and XPhoneBERT (Nguyen et al., 2023). Each model is trained on different types of language representation.

Multilingual BERT (mBERT) is a **subword-based** model that utilizes WordPiece algorithm for tokenization. During pre-training, mBERT learns to perform masked language modeling (MLM) and next sentence prediction (NSP).

CANINE is a multilingual **character-based** model that is trained on the same corpus with the same training objective as mBERT. CANINE is a tokenization-free language model that directly maps each unicode character to its codepoint by hashing. This prevents unknown tokens, enabling the model to handle a large amount of distinct characters.

Lastly, XPhoneBERT is a **phoneme-based** model trained to do MLM. XPhoneBERT follows the pre-training scheme of XLM-R (Conneau et al., 2020), so NSP is not employed in its pre-training. This model takes as input the sequence of IPA characters, where the input data are created from original text by G2P conversion followed by phoneme segmentation.

While character-level models are known to better generalize to low-resource languages (Clark et al., 2022), their general performance falls behind subword-based models. To specifically compare input representations–phonemes versus orthographic scripts–we minimize the impact of different tokenization units by focusing on phoneme-based models versus character-based models, rather than directly comparing with subword-based models like mBERT. Nevertheless, we include mBERT results for the XNLI task to highlight its significant performance drop on low-resource languages. For other tasks, we report results from phoneme and character-based models to ensure a fair comparison, and leave further improvements of character-level models in overall performance as future work.

3.2 Downstream Tasks

We adopted the cross-lingual generalization benchmark tasks suggested in XTREME (Hu et al., 2020).

Token-level Classification. We choose **POS tagging** and **NER** as our testbed for structured prediction tasks. Both tasks require labeling each token from the model. These types of tasks were previously analyzed as being relatively independent from the data size of each language used for pre-training the language model (Hu et al., 2020). We find this particularly suitable in our scenario where two models with different pre-training strategy are compared. For datasets, we utilize the corpora from Universal Dependencies[2] for POS tagging, and WikiAnn (Pan et al., 2017) with train, dev, test splits following Rahimi et al. (2019) for NER.

Sentence-level Classification. XTREME supports two sentence-level classification tasks. This type of task requires semantic understanding of given sentences to make a prediction. We employ **XNLI** (Conneau et al., 2018) dataset, which is a representative benchmark for the natural language inference task on cross-lingual generalization setting. This task requires the model to classify the relation of two given sentences into three different classes.

[1] pre-trained weights are obtained from https://huggingface.co/models

[2] https://universaldependencies.org/ , v2.13, 148 languages, released Nov 15, 2023.

3.3 Performance Gap

We analyze performance gaps of each model for all downstream tasks. As we are interested in how different models with different input types performs consistent across languages rather than their absolute overall performance, we take the relative percentage difference (RPD) (Miller, 2011) to derive the performance gap. Here, we define RPD as

$$\text{RPD}(L_i, L_j) = \frac{|\text{S}(L_i) - \text{S}(L_j)|}{\frac{1}{2}(\text{S}(L_i) + \text{S}(L_j))} \times 100, \quad (1)$$

where $\text{S}(L_i)$ represents the performance for the language L_i. This is used to analyze the performance gap, which specifically computes the relative performance gaps across languages.

3.4 Linguistic Gap

To compute representation discrepancy across languages, we use FLORES+ (Costa-jussà et al., 2022) corpus which contains parallel sentences of more than 200 languages. We employ devtest set of each language subset, which contains 1,012 sentences.

After training each model on each downstream task, we utilize each model to obtain similarity in their representations. We adopt mean-pooling to obtain sentence representations and Centered Kernel Alignment (CKA) (Kornblith et al., 2019) to measure the similarity, which Del and Fishel (2021) has recommended for robust analysis on cross-lingual similarity. CKA is defined as,

$$\text{CKA}(\mathbf{X}, \mathbf{Y}) = \frac{\|\mathbf{X}^T\mathbf{Y}\|_2^2}{\|\mathbf{X}^T\mathbf{X}\|_2\|\mathbf{Y}^T\mathbf{Y}\|_2}, \quad (2)$$

where features $\mathbf{X}$ and $\mathbf{Y}$ are from different languages. They are extracted from the input embedding layers as we are interested in how different input types (i.e., orthographic vs. phonemic) affect cross-lingual alignment, and Muller et al. (2021b) finds that cross-lingual alignment happens in the lower layers of the model. We use this similarity scores computed with CKA to refer to *linguistic gaps*, where smaller CKA score means larger linguistic gap.

3.5 Implementation Details

Models were trained for 30 epochs on a single NVIDIA A5000 GPU for POS tagging, 30 epochs a single NVIDIA A40 GPU for NER, and 20 epochs on NVIDIA A6000 for XNLI. For all experiments, batch size was set to 128 and AdamW (Loshchilov and Hutter, 2018) optimization was used. Additionally, cosine learning rate scheduler was adopted with its initial learning rate set by grid search. Learning rates used for each model on each language are in the supplementary material.

3.6 Data Preparation

Languages. To evaluate token-level tasks, we selected 10 languages with diverse typoloigical background—English(eng), French(fra), Russian(rus), Italian(ita), Hungarian(hun), Ukrainian(ukr), Korean(kor), Turkish(tur), Finnish(fin), and Hindi(hin). First four languages are high-resource languages, where English, French, and Italian are written in Latin scripts and Russian in Cyrillic. The other languages are pre-trained on each model with moderate or small amount of data, and are written in diverse scripts, such as Hangul, Cyrillic and Devanagari. For further analysis using sentence-level tasks, we chose two low-resource languages—Swahili(swa), and Urdu(urd)—to compare with a representative high-resource language, English(eng).

Preprocessing. In order to prepare inputs for a phoneme-based model, we employed G2P (Grapheme-to-Phoneme) conversion to obtain an IPA version of the input. This conversion was done with Epitran[3] (Mortensen et al., 2018) , an external tool for G2P conversion. After converting to IPA, phoneme segmentation with a python package, segments[4], to identify each phoneme. Lastly, to make it compatible with XPhoneBERT's tokenizer, white space was inserted between every phoneme.

4 Results and Analysis

Here, we present our observations and analyses of the results. We first discuss the behavior of phoneme-based model towards low-resource languages and writing systems, which contributes to robust cross-lingual performance. Next, we delve into the performance and linguistic gaps of phoneme-based models through empirical and theoretical analyses.

4.1 Phoneme-based Model on Low-Resource Languages and Writing Systems

We observe that phoneme-based model shows promising performance in low-resource languages

[3]https://github.com/dmort27/epitran
[4]https://pypi.org/project/segments/

Method	Language										Performance gap		Linguistic Gap
	eng	fra	rus	ita	hun	ukr	kor	tur	fin	hin	Std. (↓)	Mean RPD (↓)	Mean CKA (↑)
Named Entity Recognition													
Character	87.13	91.27	91.80	92.26	93.14	93.88	84.11	92.92	90.45	87.68	0.0316	4.02	0.4584
Phoneme	83.61	89.42	89.60	90.56	91.89	92.76	87.19	92.35	89.23	88.23	**0.0259**	**3.52**	**0.7195**
Part-of-Speech Tagging													
Character	96.62	95.54	87.91	96.06	74.57	85.79	86.71	90.49	91.78	96.81	0.0692	8.77	0.4593
Phoneme	95.94	96.35	86.69	96.37	85.87	91.32	85.82	91.11	93.76	96.94	**0.0455**	**5.80**	**0.7204**

Table 1: Performance of POS tagging and NER across different languages. Std. refers to the standard deviation of the scores across the languages, and Mean RPD indicates average relative difference of F1 scores between different languages. Mean CKA represents the average linguistic gap between languages.

and writing systems (scripts). Results from Table 1 show that phoneme-based model outperforms the character-based model on NER task, in languages written in scripts other than major scripts[5]—Korean and Hindi. This can be attributed to the fact that named entities, such as geopolitical or personal names, are often pronounced similarly across languages. When different writing systems and scripts are used, models may struggle to align such entities. However, representing them in IPA characters that reflect their pronunciations helps the model to better align these entities, resulting in better cross-lingual transfer. This results align with findings from Muller et al. (2021a); Sohn et al. (2024), which focus on unseen languages, whereas we observe this phenomenon with diverse 'seen' languages.

Results also demonstrates the potential of phoneme-based model in addressing low-resource languages. As shown in Table 2, the phoneme-based model achieves a smaller gap when transferred to low-resource languages such as Swahili and Urdu, compared to other baselines. This finding is further analyzed in Section 4.2

4.2 Performance Gap Across Languages

We observe that the phoneme-based model consistently exhibits the smallest performance gap across diverse languages, highlighting its robustness in cross-lingual tasks. In Table 1, we present the standard deviation (Std.) and average percentage difference (Mean diff.) for all models, which reflect the variability in performance across different languages. The phoneme-based model exhibits both a lower standard deviation and a smaller average percentage difference in the NER and POS tasks, demonstrating its relatively stable performance across different languages.

[5]Latin and Cyrillic are scripts that are used the most during the pre-training phase.

Method	Language				
	eng	swa		urd	
	Acc.	Acc.	Δ from eng (Rel./Abs.)	Acc.	Δ from eng (Rel./Abs.)
Subword	80.80	62.93	24.87 / 17.87	61.57	27.01 / 19.23
Character	75.02	59.72	22.71 / 15.30	56.55	28.08 / 18.47
Phoneme	71.89	60.88	**16.59 / 11.01**	56.10	**24.67 / 15.79**

Table 2: Accuracy (%) and relative/absolute performance gaps on XNLI task. eng, swa, and urd refer to English, Swahili, and Urdu, respectively, and relative difference is computed with RPD. Phonemic representation shows relatively small performance gaps compared to other representations.

Table 2 provides additional evidence by showing that the phoneme-based model achieves a smaller gap in performance between English and other low-resource languages—Swahili(swa) and Urdu(urd)—compared to other models. We report both relative and absolute differences in performance, with the relative difference calculated as described in Section 3.3.

While subword-based mBERT achieves the highest scores, the performance gaps between models narrow when applied to low-resource languages, with outperforming the phoneme-based model by 8.91% in English and by 2.05% and 5.47% in Swahili and Urdu, respectively. This reflects subword LM's significant performance drops on low-resource languages, while highlighting the phoneme-based LM's robustness in cross-lingual transfer to such languages. The leftmost panel of Figure 3 also illustrates the performance gaps of each model, where the phoneme-based model predominantly displays lower gaps compared to others.

These metrics collectively suggest that phone-

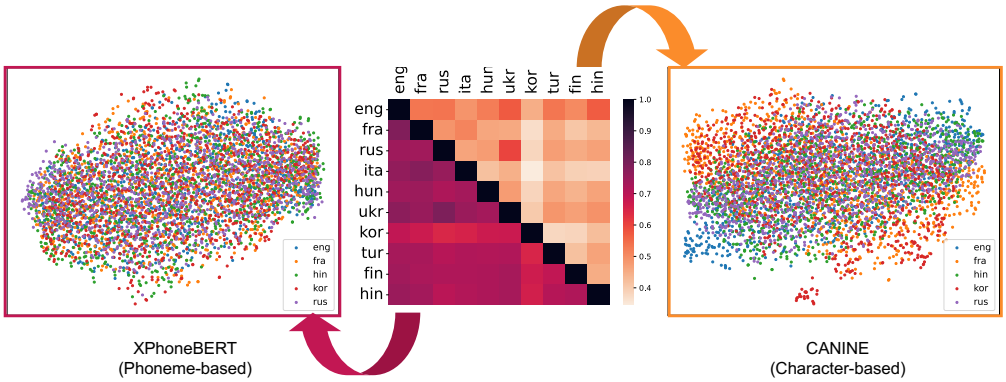

Figure 2: Linguistic gaps across languages in each model. (Center) Upper and lower triangular elements of the heatmap indicate pairwise linguistic gaps derived with character-based model and phoneme-based model, respectively. Darker color indicates larger CKA score, which means smaller discrepancy. Lower triangular elements show relatively darker colors, implying smaller discrepancies across languages of phoneme-based model. (Left, right) T-SNE plots for each model are shown with only five languages, for better visibility.

mic representations offer a more consistent performance in multilingual settings, reducing the disparities typically observed when models are applied to languages with varying resource availability.

4.3 Linguistic Gap of Different Representations

To investigate the potential of phonemes as a robust representation for multilingual language modeling, we analyze the linguistic gap between languages using different input representations. Following Yang et al. (2022); Muller et al. (2021b), we use linear CKA to quantify representation similarity across languages. Figure 2 shows the pairwise similarities between languages, with the lower triangle of the heatmap, which corresponds to phonemic representations, demonstrating higher similarity values. This indicates a smaller linguistic gap compared to models that use orthographic inputs, contributing to a smaller performance gap. Moreover, the t-SNE plots placed in both sides show how the distributions of the representations from different languages resemble each other. Phoneme-based model exhibits more similar distribution across languages.

Figure 3 further supports these observations by showing the linguistic gap after fine-tuning on the XNLI task. The plot in the center illustrates that phonemic representations have higher CKA scores than other baseline models, indicating closer alignment between language representations. As XNLI directly learns to build a sentence representation during fine-tuning, we extract the representation from the last hidden layer unlike in other token-level tasks. Additionally, by using Sinkhorn distance to compare the logit space, we observe that the phoneme-based model shows lower distances, reflecting more consistent predictions across languages.

These results highlight the potential of phonemic representations to address the performance gaps that challenge multilingual language models, particularly in bridging the gap between high-resource and low-resource languages by more similar representations.

4.4 Connecting Performance Gap and Linguistic Gap

Correlation Analysis. Meanwhile, one may speculate the low-performance gap of the phoneme-based model can be driven by the low English performance rather than reducing the linguistic gap. To clarify this, we simulate 15 repeated runs (with different random seeds) of phonemic representation using 10% of the XNLI train dataset over English, Swahili, and Urdu. After computing the best performance per each language, Sinkhorn distance (S-Dist), and CKA between English and the other two languages, we conducted correlation analyses by performing hypothesis tests with Spearman's rank correlation coefficient and Kendall's Tau.

As can be seen from Table 3, rather than the English performance, S-Dist and CKA have stronger correlations, indicating that the linguistic gap has stronger correlations that are statistically significant (with a significant level less than 0.01).

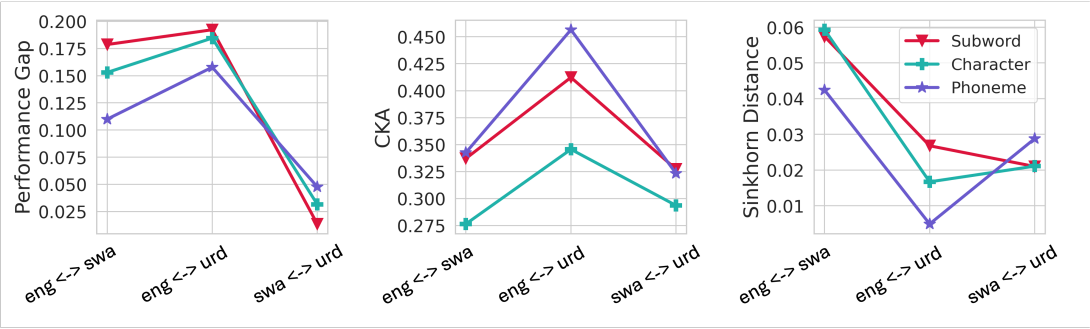

Figure 3: Qualitative analysis of performance gap (difference of accruacy) on XNLI task. (Left) the absolute difference between performance across two languages, (center) centered kernel alignment (CKA) scores to measure cross-lingual embedding similarity, and (right) Sinkhorn distance on the output probability space. Phonemic representation shows relatively small performance gaps w.r.t. eng ↔ swa and eng ↔ urd, and these gaps are correlated with similarity and discrepancy on the embedding space (CKA) and logit space (Sinkhorn distance).

Correlation	Spearman's R		Kendall's T	
	coefficient	p-value	coefficient	p-value
Performance Gap <-> eng Performance	0.111	5.60E-01	0.104	4.30E-01
Performance Gap <-> **S-Dist**	**0.681**	**3.50E-05**	**0.457**	**2.00E-04**
Performance Gap <-> **CKA**	**-0.782**	**3.40E-07**	**-0.577**	**2.10E-06**

Table 3: Correlation analysis with 45 phoneme-based models. We fine-tune the phoneme-based language model XPhoneBERT on three languages, eng, swa, and urd, with 15 different random seeds and conduct two types of correlation analyses.

Theoretical Analysis. We aim to diminish the performance gap between different languages by adopting IPA as a universal language representation. Motivated by domain adaptation literature (Kifer et al., 2004; Ben-David et al., 2010), we present a theoretical justification of IPA for robust multilingual modeling by deriving a bound for cross-lingual performance gap.

Let $\mathcal{D}$ denote a domain as a distribution over text feature input $\mathcal{X}$, such as the sequence of word embeddings or one-hot vectors, and a labeling function $f : \mathcal{X} \to \{0, 1\}$. Assuming a binary classification task, our goal is to learn a hypothesis $h : \mathcal{X} \to \{0, 1\}$ that is expected to minimize a risk $\varepsilon_D(h, f) := \mathbb{E}_{x\sim\mathcal{D}}[\mathbb{I}(f(x) \neq h(x))]$ and has a small risk-deviation over two domains $\mathcal{D}_A$ and $\mathcal{D}_B$. Then, to formalize the cross-lingual performance gap, we first need a discrepancy measure between two languages. By following Ben-David et al. (2010), we adopt $\mathcal{H}$-divergence (See Appendix C for its definition) to quantify the distance between two language distributions.

Now, based on Lemma 1 and 3 of Ben-David et al. (2010), we make reasoning on performance gap over different language domains.

Theorem 4.1. *Let $h : \mathcal{X} \to [0, 1]$ be a real-valued function in a hypothesis class $\mathcal{H}$ with a pseudo dimension $\mathcal{P}dim(\mathcal{H}) = d$. If $\hat{\mathcal{D}}_A$ and $\hat{\mathcal{D}}_B$ are the empirical distribution constructed by n-size i.i.d. samples, drawn from $\mathcal{D}_A$ and $\mathcal{D}_B$ respectively, then for any $\delta \in (0, 1)$, and for all h, the bound below hold with probability at least $1 - \delta$.*

$$|\varepsilon_{\mathcal{D}_A}(h, f) - \varepsilon_{\mathcal{D}_B}(h, f)| \leq \frac{1}{2} d_{\mathcal{H}\Delta\mathcal{H}}(\hat{\mathcal{D}}_A, \hat{\mathcal{D}}_B) + 2\sqrt{\frac{d\log(2n) + \log(2/\delta)}{n}}$$

where $\mathcal{H}\Delta\mathcal{H} := \{h(x) \oplus h'(x) | h, h' \in \mathcal{H}\}$ given $\oplus$ as a xor operation (proof is in Appendix C). We see that performance gap between two lanauges is bounded from above with a distribution divergence plus an irreducible term defined by problem setup. That is, if we reduce the divergence between language distributions, the expected performance gap can also be reduced accordingly.

To investigate whether this is indeed a case or not, we provided embedding space similarity and logit-space Sinkhorn distance (Cuturi, 2013) between different languages in Figure 3. We argue

that phonemic representation's relatively mild performance gap is achieved by reducing linguistic gaps which is confirmed in the embedding space (high CKA) and final output space (low Sinkhorn distance).

5 Conclusion

Towards robust multilingual language modeling, we argue that mitigating the linguistic gap between different languages is crucial. Moreover, we advocate the use of IPA phonetic symbols as a universal language representation partially bridges such linguistic gaps without any complicated cross-lingual training phase. Empirical validation on three representative NLP tasks demonstrates the superiority of phonemic representation compared to subword and character-based language representation in terms of the cross-lingual performance gap and linguistic gap. Theoretical analysis of the cross-lingual performance gap explains such promising results of phonemic representation.

6 Limitations

While we have shown that phonemic representation induces a small cross-lingual linguistic gap, therefore a small performance gap, the absolute performance of this phonemic representation is still lacking compared to subword-level models. We spur the necessity of putting research attention to developing phoneme-based LMs. Moreover, there is no such large phonemic language model beyond the BERT-base-size architecture, so we confine the scope of our empirical validation to BERT-base-size LMs. This also means the experiments rely on existing pre-trained models, limiting control over their pre-training settings. Since the models were trained on different language sets and pre-training objectives (as noted in 3.1), it is important to verify these findings in a controlled environment. Additionally, we performed evaluation with a limited languages (up to 12), so it is unclear whether IPA language representations are effective for other numerous languages (especially low-resource ones) or not.

7 Ethics Statement

We believe there are no potential of any critical issues that harm the code of ethics provided by ACL. The social impacts of the technology—reducing performance gaps for low resource languages—will be, on the balance, positive. The data was, to the extent we can determine, collected in accordance with legal and institutional protocals.

Acknowledgements

This work was supported by Institute of Information & communications Technology Planning & Evaluation (IITP) grant funded by the Korea government(MSIT) (RS-2022-00143911, AI Excellence Global Innovative Leader Education Program)

References

Sina Bagheri Nezhad and Ameeta Agrawal. 2024. What drives performance in multilingual language models? In *Proceedings of the Eleventh Workshop on NLP for Similar Languages, Varieties, and Dialects (VarDial 2024)*, pages 16–27, Mexico City, Mexico. Association for Computational Linguistics.

Shai Ben-David, John Blitzer, Koby Crammer, Alex Kulesza, Fernando Pereira, and Jennifer Vaughan. 2010. A theory of learning from different domains. *Machine Learning*, 79:151–175.

Shai Ben-David, John Blitzer, Koby Crammer, and Fernando Pereira. 2006. Analysis of representations for domain adaptation. In *Advances in Neural Information Processing Systems*, volume 19. MIT Press.

Akash Bharadwaj, David Mortensen, Chris Dyer, and Jaime Carbonell. 2016. Phonologically aware neural model for named entity recognition in low resource transfer settings. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pages 1462–1472, Austin, Texas. Association for Computational Linguistics.

Aditi Chaudhary, Chunting Zhou, Lori Levin, Graham Neubig, David R. Mortensen, and Jaime Carbonell. 2018. Adapting word embeddings to new languages with morphological and phonological subword representations. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 3285–3295, Brussels, Belgium. Association for Computational Linguistics.

Jonathan H. Clark, Dan Garrette, Iulia Turc, and John Wieting. 2022. Canine: Pre-training an efficient tokenization-free encoder for language representation. *Transactions of the Association for Computational Linguistics*, 10:73–91.

Alexis Conneau, Kartikay Khandelwal, Naman Goyal, Vishrav Chaudhary, Guillaume Wenzek, Francisco Guzmán, Edouard Grave, Myle Ott, Luke Zettlemoyer, and Veselin Stoyanov. 2020. Unsupervised cross-lingual representation learning at scale. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 8440–8451, Online. Association for Computational Linguistics.

Alexis Conneau, Ruty Rinott, Guillaume Lample, Adina Williams, Samuel Bowman, Holger Schwenk, and Veselin Stoyanov. 2018. XNLI: Evaluating cross-lingual sentence representations. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 2475–2485, Brussels, Belgium. Association for Computational Linguistics.

Marta R Costa-jussà, James Cross, Onur Çelebi, Maha Elbayad, Kenneth Heafield, Kevin Heffernan, Elahe Kalbassi, Janice Lam, Daniel Licht, Jean Maillard, et al. 2022. No language left behind: Scaling human-centered machine translation. *arXiv preprint arXiv:2207.04672*.

Marco Cuturi. 2013. Sinkhorn distances: Lightspeed computation of optimal transport. In *Neural Information Processing Systems*.

Maksym Del and Mark Fishel. 2021. Similarity of sentence representations in multilingual lms: Resolving conflicting literature and a case study of baltic languages. *Balt. J. Mod. Comput.*, 10.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4171–4186, Minneapolis, Minnesota. Association for Computational Linguistics.

Yoshinari Fujinuma, Jordan Boyd-Graber, and Katharina Kann. 2022. Match the script, adapt if multilingual: Analyzing the effect of multilingual pretraining on cross-lingual transferability. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1500–1512, Dublin, Ireland. Association for Computational Linguistics.

Junjie Hu, Sebastian Ruder, Aditya Siddhant, Graham Neubig, Orhan Firat, and Melvin Johnson. 2020. Xtreme: A massively multilingual multi-task benchmark for evaluating cross-lingual generalisation. In *International Conference on Machine Learning*, pages 4411–4421. PMLR.

Ye Jia, Heiga Zen, Jonathan Shen, Yu Zhang, and Yonghui Wu. 2021. Png bert: Augmented bert on phonemes and graphemes for neural tts. In *Interspeech*.

Daniel Kifer, Shai Ben-David, and Johannes Gehrke. 2004. Detecting change in data streams. In *Proceedings of the Thirtieth International Conference on Very Large Data Bases - Volume 30*, VLDB '04, page 180–191. VLDB Endowment.

Simon Kornblith, Mohammad Norouzi, Honglak Lee, and Geoffrey Hinton. 2019. Similarity of neural network representations revisited. In *Proceedings of the 36th International Conference on Machine Learning*, volume 97 of *Proceedings of Machine Learning Research*, pages 3519–3529. PMLR.

Colin Leong and Daniel Whitenack. 2022. Phone-ing it in: Towards flexible multi-modal language model training by phonetic representations of data. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 5306–5315, Dublin, Ireland. Association for Computational Linguistics.

Yinghao Aaron Li, Cong Han, Xilin Jiang, and Nima Mesgarani. 2023. Phoneme-level bert for enhanced prosody of text-to-speech with grapheme predictions. In *ICASSP 2023 - 2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 1–5.

Ilya Loshchilov and Frank Hutter. 2018. Decoupled weight decay regularization. In *International Conference on Learning Representations*.

Ronald E. Miller. 2011. *Optimization: Foundations and Applications*. John Wiley & Sons.

Steven Moran and Daniel McCloy, editors. 2019. *PHOIBLE 2.0*. Max Planck Institute for the Science of Human History, Jena.

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)*, Miyazaki, Japan. European Language Resources Association (ELRA).

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, Osaka, Japan. The COLING 2016 Organizing Committee.

Benjamin Muller, Antonios Anastasopoulos, Benoît Sagot, and Djamé Seddah. 2021a. When being unseen from mBERT is just the beginning: Handling new languages with multilingual language models. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 448–462, Online. Association for Computational Linguistics.

Benjamin Muller, Yanai Elazar, Benoît Sagot, and Djamé Seddah. 2021b. First align, then predict: Understanding the cross-lingual ability of multilingual BERT. In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume*, pages 2214–2231, Online. Association for Computational Linguistics.

Linh The Nguyen, Thinh Pham, and Dat Quoc Nguyen. 2023. XPhoneBERT: A Pre-trained Multilingual

Model for Phoneme Representations for Text-to-Speech. In *Proceedings of the 24th Annual Conference of the International Speech Communication Association (INTERSPEECH)*.

Xiaoman Pan, Boliang Zhang, Jonathan May, Joel Nothman, Kevin Knight, and Heng Ji. 2017. Cross-lingual name tagging and linking for 282 languages. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1946–1958, Vancouver, Canada. Association for Computational Linguistics.

Afshin Rahimi, Yuan Li, and Trevor Cohn. 2019. Massively multilingual transfer for NER. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 151–164, Florence, Italy. Association for Computational Linguistics.

Jasdeep Singh, Bryan McCann, Richard Socher, and Caiming Xiong. 2019. BERT is not an interlingua and the bias of tokenization. In *Proceedings of the 2nd Workshop on Deep Learning Approaches for Low-Resource NLP (DeepLo 2019)*, pages 47–55, Hong Kong, China. Association for Computational Linguistics.

Jimin Sohn, Haeji Jung, Alex Cheng, Jooeon Kang, Yilin Du, and David R. Mortensen. 2024. Zero-shot cross-lingual ner using phonemic representations for low-resource languages. *Preprint*, arXiv:2406.16030.

David Stap, Vlad Niculae, and Christof Monz. 2023. Viewing knowledge transfer in multilingual machine translation through a representational lens. *ArXiv*, abs/2305.11550.

Simeng Sun, Angela Fan, James Cross, Vishrav Chaudhary, Chau Tran, Philipp Koehn, and Francisco Guzmán. 2022. Alternative input signals ease transfer in multilingual machine translation. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 5291–5305, Dublin, Ireland. Association for Computational Linguistics.

Mukuntha Narayanan Sundararaman, Ayush Kumar, and Jithendra Vepa. 2021. Phoneme-bert: Joint language modelling of phoneme sequence and asr transcript. *arXiv preprint arXiv:2102.00804*.

Di Wu and Christof Monz. 2023. Beyond shared vocabulary: Increasing representational word similarities across languages for multilingual machine translation. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 9749–9764, Singapore. Association for Computational Linguistics.

Shijie Wu and Mark Dredze. 2020. Are all languages created equal in multilingual BERT? In *Proceedings of the 5th Workshop on Representation Learning for NLP*, pages 120–130, Online. Association for Computational Linguistics.

Huiyun Yang, Huadong Chen, Hao Zhou, and Lei Li. 2022. Enhancing cross-lingual transfer by manifold mixup. In *International Conference on Learning Representations*.

Guangyan Zhang, Kaitao Song, Xu Tan, Daxin Tan, Yuzi Yan, Yanqing Liu, Gang Wang, Wei Zhou, Tao Qin, Tan Lee, and Sheng Zhao. 2022. Mixed-phoneme bert: Improving bert with mixed phoneme and sup-phoneme representations for text to speech. In *Interspeech 2022*, pages 456–460.

Hongyi Zhang, Moustapha Cisse, Yann N. Dauphin, and David Lopez-Paz. 2018. mixup: Beyond empirical risk minimization. In *International Conference on Learning Representations*.

A Dataset Statistics

In Table 4, we provide the dataset statistics. For the experiments, we used train set for training and validation set for evaluation.

Dataset	Lang.	Train	Dev	Test
FLORES+	eng fra rus ita hun ukr kor tur fin hin	-	1.2k	-
XNLI	eng swa urd	393k	2.49k	5.01k
WikiAnn	eng fra rus ita hun ukr kor tur fin hin	20k 20k 20k 20k 20k 20k 20k 20k 20k 5k	10k 10k 10k 10k 10k 10k 10k 10k 10k 1k	10k 10k 10k 10k 10k 10k 10k 10k 10k 1k
UD	eng fra rus ita hun ukr kor tur fin hin	12.5k 14.5k 16k 13k 0.9k 5.5k 23k 15k 12k 13k	2k 1.5k 0.9k 0.6k 0.4k 0.7k 2k 1.6k 1.4k 1.7k	2k 0.4k 0.9k 0.5k 0.4k 0.9k 2.3k 1.6k 1.6k 1.7k

Table 4: Dataset statistics for datasets used in experiments: FLORES+, XNLI, WikiAnn, Universal Dependencies Tree Bank. For FLORES+ dataset, we used devtest set with 1,012 sentences.

B Hyperparameter sweep.

We sweep hyperparameters over grid below (in Table 5), and select the final parameters for each model based on the **best validation performance** (Accuracy for XNLI and F1-score for NER and POS Tagging).

C Details on Theoreoretical Analysis

We aim to diminish the performance gap between different languages by adopting IPA as a universal language representation. Motivated by domain adaptation literature (Kifer et al., 2004; Ben-David et al., 2010), we present a theoretical justification of IPA for robust multilingual modeling by providing a bound for cross-lingual performance gap.

Let $\mathcal{D}$ denote a domain as a distribution over text feature input $\mathcal{X}$, such as the sequence of word embeddings or one-hot vectors, and a labeling function $f : \mathcal{X} \to \{0, 1\}$. Assuming a binary classification task, our goal is to learn a hypothesis $h : \mathcal{X} \to \{0, 1\}$ that is expected to minimize a risk $\varepsilon_D(h, f) := \mathbb{E}_{x\sim\mathcal{D}}[\mathbb{I}(f(x) \neq h(x))]$ and has a small risk-deviation over two domains $\mathcal{D}_A$ and $\mathcal{D}_B$. Then, to formalize the cross-lingual performance gap, we first need a discrepancy measure between two languages. By following (Ben-David et al., 2010), we adopt $\mathcal{H}$-divergence to quantify the distance between two language distributions.

Definition C.1 ($\mathcal{H}$-divergence; Ben-David et al. (2006))**.** *Let $\mathcal{H}$ be a hypothesis class for input space $\mathcal{X}$ and a collection of subsets from $\mathcal{X}$ is denoted by $\mathcal{S}_\mathcal{H} := \{h^{-1}(1)|h \in \mathcal{H}\}$ which is the support of hypothesis $h \in \mathcal{H}$. The $\mathcal{H}$-divergence between two distributions $\mathcal{D}$ and $\mathcal{D}'$ is defined as*

$$d_\mathcal{H}(D, D') = 2 \sup_{S\in\mathcal{S}_\mathcal{H}} |\mathbb{P}_D(S) - \mathbb{P}_{D'}(S)|$$

$\mathcal{H}$-divergence is a relaxation of total variation between two distributions, and it can be estimated by finite samples from both distributions if $\mathcal{H}$ governs a finite VC dimension. Now, based on Lemma 1 and 3 of Ben-David et al. (2010), we make reasoning on performance gap over different language domains.

Theorem C.2. *Let $h : \mathcal{X} \to [0, 1]$ be a real-valued function in a hypothesis class $\mathcal{H}$ with a pseudo dimension $\mathcal{P}dim(\mathcal{H}) = d$. If $\hat{\mathcal{D}}_A$ and $\hat{\mathcal{D}}_B$ are the empirical distribution constructed by n-size i.i.d. samples, drawn from $\mathcal{D}_A$ and $\mathcal{D}_B$ respectively, then for any $\delta \in (0, 1)$, and for all h, the bound below hold with probability at least $1 - \delta$.*

$$|\varepsilon_{\mathcal{D}_A}(h, f) - \varepsilon_{\mathcal{D}_B}(h, f)| \leq \frac{1}{2} d_{\mathcal{H}\Delta\mathcal{H}}(\hat{\mathcal{D}}_A, \hat{\mathcal{D}}_B) + 2\sqrt{\frac{d\log(2n) + \log(2/\delta)}{n}}$$

where $\mathcal{H}\Delta\mathcal{H} := \{h(x) \oplus h'(x)|h, h' \in \mathcal{H}\}$ given $\oplus$ as a `xor` operation.

Task	Hyperparam	Search space	Selected parameter value		
			mBERT	CANINE	XPhoneBERT
XNLI	learning rate weight decay learning rate scheduling	[5e-6, 7e-6, 1e-5, 3e-5, 5e-5] [0.0, 1e-1, 1e-2, 1e-3] [True, False]	5e-6 0.01 True	5e-6 (en), 1e-5 (sw, ur) 0.1 (en), 0.0 (sw), 0.01 (ur) True	7e-6 (en), 3e-6 (sw, ur) 0.1 (en), 0.0 (sw), 0.01 (ur) False
NER	learning rate weight decay	[3e-5, 5e-5, 1e-4, 3e-4] 1e-2	- -	5e-5 (en, fr, it, hu, ko, tr), 1e-4 (ru, uk, fi, hi) 1e-2	3e-5 (ru, it), 5e-5 (en, fr, hu, uk, tr, fi, hi), 1e-4 (ko) 1e-2
POS	learning rate weight decay	[3e-5, 5e-5, 1e-4, 3e-4] 1e-2	- -	5e-5 (ru, uk, tr), 1e-4 (en, fr, fi, hi), 3e-4 (it, hu, ko) 1e-2	5e-5 (en), 1e-4 (fr, ru, it, hu, uk, ko, tr, fi, hi) 1e-2

Table 5: List of hyperparameter, search spaces and selected parameter values for different models applied to XNLI, NER, and POS tasks, detailing learning rate, weight decay, and learning rate scheduling for mBERT, CANINE, and XPhonemBERT, with specific configurations for optimal model performance per task.

proof of Theorem B.2. we start to prove Theorem B.2. by restating Lemma 1 of (Ben-David et al., 2010) adapted to our notation.

Lemma C.3. *Let $\mathcal{D}_A$ and $\mathcal{D}_B$ be distributions of domain A and B over $\mathcal{X}$, respectively. Let $\mathcal{H}$ be a hypothesis class of functions from $\mathcal{X}$ to $[0, 1]$ with VC dimension d. If $\hat{\mathcal{D}}_A$ and $\hat{\mathcal{D}}_B$ are the n-size empirical distributions generated by $\mathcal{D}_A$ and $\mathcal{D}_B$ respectively, then, for $0 < \delta < 1$, with probability at least $1 - \delta$,*

$$d_{\mathcal{H}}(\mathcal{D}_A, \mathcal{D}_B) \leq d_{\mathcal{H}}(\hat{\mathcal{D}}_A, \hat{\mathcal{D}}_B) + 4\sqrt{\frac{d\log(2n) + \log(2/\delta)}{n}}.$$

Then, for any hypothesis function $h, h' \in \mathcal{H}$, by the definition of $d_{\mathcal{H}\Delta\mathcal{H}}$-divergence, we have:

$$\begin{aligned} &d_{\mathcal{H}\Delta\mathcal{H}}(\mathcal{D}_A, \mathcal{D}_B) \\ &= 2 \sup_{h,h' \in \mathcal{H}} |\mathbb{P}_{x\sim\mathcal{D}_A}[h(x) \neq h'(x)] - \mathbb{P}_{x\sim\mathcal{D}_B}[h(x) \neq h'(x)]| \\ &= 2 \sup_{h,h' \in \mathcal{H}} |\varepsilon_{\mathcal{D}_A}(h, h') - \varepsilon_{\mathcal{D}_B}(h, h')| \\ &\geq 2|\varepsilon_{\mathcal{D}_A}(h, h') - \varepsilon_{\mathcal{D}_B}(h, h')| \end{aligned}$$

Now the below bound holds for any hypothesis functions $h, h' \in \mathcal{H}$ (See Lemma 3 of (Ben-David et al., 2010)).

$$|\varepsilon_{\mathcal{D}_A}(h, h') - \varepsilon_{\mathcal{D}_B}(h, h')| \leq \frac{1}{2} d_{\mathcal{H}\Delta\mathcal{H}}(\mathcal{D}_A, \mathcal{D}_B)$$

Finally, by plugging the Lemma C.3 into the above bound, we have Theorem C.2.

□

From Theorem C.2, we see that the difference between true risks across language domains is bounded by an empirical estimation of the divergence ($d_{\mathcal{H}\Delta\mathcal{H}}$) between those two domains plus an irreducible term defined by problem setup. Thus, if we reduce the divergence between language distributions, the expected performance gap can also be reduced accordingly. To investigate whether this is indeed a case or not, we provided the embedding-space similarity and logit-space Sinkhorn distance between different languages in Figure 3. We argue that phonemic representation's relatively mild performance gap is achieved by reducing linguistic gaps in the embedding space (high CKA) and final output space (low Sinkhorn distance) those are the proxy of $\mathcal{H}$-divergence.