Mitigating the Linguistic Gap with Phonemic Representations for Robust Multilingual Language Understanding

Anonymous ACL submission

Abstract

Approaches to improving multilingual language understanding often require multiple languages during the training phase, rely on complicated training techniques, and—importantly—struggle with significant performance gaps between high-resource and low-resource languages. We hypothesize that the performance gaps between languages are affected by linguistic gaps between those languages and provide a novel solution for robust multilingual language modeling by employing phonemic representations (specifically, using phonemes as input tokens to LMs rather than subwords). We present quantitative evidence from three cross-lingual tasks that demonstrate the effectiveness of phonemic representation, which is further justified by a theoretical analysis of the cross-lingual performance gap[1].

1 Introduction

In an era of large language models, natural language processing has promised to bring us together. However, there are gaps between how human language technologies perform in some languages (like English) and others (most of the languages of the world). Language technologies exhibit **performance gaps** between languages. To some degree, this is due to differences in resourcedness, but we hypothesize that **linguistic gaps**—the chasms that separate languages from one another structurally and lexically—also play a role. We further hypothesize that techniques that reduce linguistic gaps will also reduce performance gaps. In this paper, we focus on one such technique: representing languages phonetically using the International Phonetic Alphabet (IPA).

Figure 1 illustrates our motivation for using IPA as a universal language representation. The key contributions of this paper are as follows:

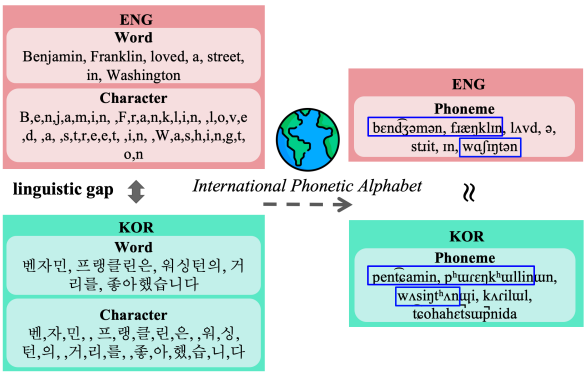

Figure 1: Example of word, character, and phoneme units for a sentence (English and Korean).

- We demonstrate the advantage of using phonemic representation (i.e., IPA) for language modeling, particularly as a robust multilingual representation.
- We empirically validate the effectiveness of phonemic representations by comparing the performance gap and linguistic gap across languages with subword or character representations.
- We further explain the empirical observations with theoretical analysis from domain generalization literature, by treating the linguistic gap as the domain gap caused by lexical and syntactic disparities.

2 Related Work

Cross-lingual Transfer Learning approaches aim to transfer knowledge from one language to enhance NLP performance in another. While adversarial training methods (Lample et al., 2018; Chen et al., 2018, 2021; Yu et al., 2023) attempt to learn a language-agnostic representation, their training can be unstable (Balcan et al., 2023). Non-adversarial joint learning approaches (Cotterell and

[1]Our code is available here: https://anonymous.4open.science/r/ipa-for-multilingual-nlu/README.md

Heigold, 2017; Gu et al., 2018; Wei et al., 2021; Zheng et al., 2021) also present promising results, but necessitate significant multilingual data. Recently, Yang et al. (2022) revealed the empirical correlation between cross-lingual transferability and representation discrepancy, yet they do not provide any theoretical statement to justify such numerical analysis.

Multilingual Large Language Models (MLLMs) and the pre-train/fine-tune paradigm have become exceptionally popular, including for multilingual LMs (Devlin et al., 2018; Conneau and Lample, 2019; Conneau et al., 2020; Chi et al., 2021; Workshop et al., 2022; Wei et al., 2023). Although MLLMs have demonstrated remarkable potential in diverse multilingual tasks, there is a significant performance gap between different languages, especially in the case of high-resource language versus low-resource language (Wu and Dredze, 2020; Zhao et al., 2023). In this work, we propose a novel paradigm for robust multilingual language modeling: IPA as a universal language representation.

Phonemic Language Representation has been applied in previous studies (Bharadwaj et al., 2016; Chaudhary et al., 2018; Dalmia et al., 2019; Hu et al., 2019; Nguyen et al., 2023), but primarily for speech recognition tasks or where they were not focused on the linguistic gap between languages. Our work revisits phonemic representation for MLLMs to mitigate the linguistic gap for the first time.

3 Experimental Setup

In this section, we briefly describe the experiment setup in terms of models, datasets, and tasks. Refer to Appendix A for further details.

3.1 Models and Data Preprocessing

We compare LMs with three different types of language representation: subword, character, and phoneme. All models[2] are pre-trained on multilingual data that covers around 100 languages from Wikipedia dump files. We employ an off-the-shelf MLLMs, multilingual BERT (mBERT; Devlin et al. (2018)). For a **subword-based** model which is trained with byte-pair encoding. For a **character-based** model, we utilize CANINE (Clark et al., 2022), which is a tokenization-free LM that directly maps each character to its codepoint by hashing. It is pre-trained on the same data and training objectives as mBERT. For a **phoneme-based** model, we adopt XPhoneBERT (Nguyen et al., 2023), which has the same model architecture as mBERT.

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. For a fair comparison between the representations, we primarily compare phoneme-based model to character-based one instead of directly comparing it to the subword-based model (i.e., mBERT), and leave further improvements of overall performance as future work.

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, we inserted white space between every character to make it compatible with XPhoneBERT's tokenizer.

3.2 Downstream Tasks

We adopted the cross-lingual generalization benchmark tasks suggested in XTREME (Hu et al., 2020). Among them, we selected two types of tasks; classification and structured prediction. For evaluation languages, we choose eight languages of three groups—high-resource (`eng`, `deu`, `fra`), low-resource (`urd`, `hin`, `swa`), and typologically distinct (`kor`, `ukr`)—to analyze the impact of phonemic representation with respect to each category.

Sentence-level Classification. We employ **XNLI** (Conneau et al., 2018) dataset, which is a representative benchmark for the natural language inference task on the cross-lingual generalization setting.

Token-level Classification. We choose **POS tagging** and **NER** as our testbed for structured prediction tasks, both requiring predicting labels for each word in sentences. We utilize the corpora from Universal Dependencies[4] for POS tagging, and WikiAnn (Pan et al., 2017) with train, dev, test splits following (Rahimi et al., 2019) for NER.

4 Results

Phoneme-based Model on Low-Resource Languages. The phoneme-based model shows

[2] pre-trained weights were obtained from https://huggingface.co/models

[3] https://github.com/dmort27/epitran

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

Method	Language		
	ENG	SWA (Δ from ENG)	URD (Δ from ENG)
Subword	80.80	62.93 (17.87)	61.57 (19.23)
Character	75.02	59.72 (15.30)	56.55 (18.46)
Phoneme	71.89	60.88 (**11.00**)	56.10 (**15.78**)

Table 1: Accuracy (%) on XNLI task. ENG, SWA, URD refer to English, Swahili, and Urdu, respectively. Phonemic representation shows relatively small performance gaps compared to other representations.

promising results compared to other models, especially for low-resource languages. As shown in Table 1, the phoneme-based model has the smallest performance gap between English and other low-resource languages – swa, urd. Furthermore, while subword-based mBERT achieves the highest scores, the performance disparity across models narrows when it comes to low-resource languages. Table 2 also suggests that the phoneme-based model exhibits superiority in addressing low-resource languages. For NER, on languages like urd, hin, and swa, the phoneme-based model significantly outperforms character-based model, highlighting the capability of the phoneme-based model's generalization over the low-resource languages.

Performance Gap Across Languages. We observe that the phoneme-based model performs the most consistently across languages. The leftmost panel in Figure 3, shows how each language representation results in performance gaps across different languages. Here, the phoneme-based model comes with the lowest performance gap. Table 2 also shows the phoneme-based model's robustness in terms of performance gap across languages. From the table, standard devision and mean difference both indicate how the results for all languages differ from each other. The phoneme-based model ends up with a smaller standard deviation in both NER and POS tasks.

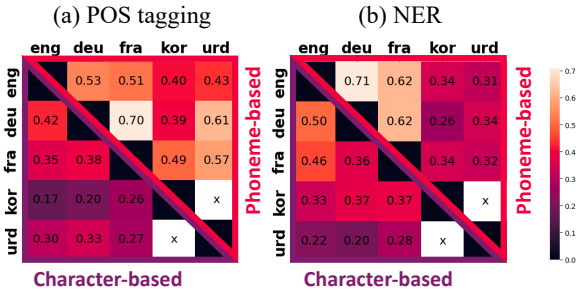

Figure 2: Linguistic gaps across languages. Upper and lower triangular elements indicate pairwise linguistic gaps derived with phoneme-based model and character-based model, respectively. Lighter color indicates larger CKA score, which means smaller discrepancy. Upper triangular elements show relatively lighter colors, implying smaller discrepancies across languages.

Linguistic Gap of Different Representations. To analyze the potential of phonemes as a robust representation for multilingual language modeling, we check the gap between different languages over different representations. As in (Yang et al., 2022), we use centered kernel alignment (CKA) to compute the representation similarity. Figure 3 shows that phonemic representation shows higher CKA and lower Sinkhorn distance compared to other representations, meaning that the phonemic representations from different languages are relatively closer to each other than those of subword or character representations.

Figure 2 also illustrates the linguistic gap between languages with their pairwise similarity. After fine-tuning each model on downstream tasks, we compute the discrepancy between different languages using held-out parallel data. It can be clearly stated that the upper triangle part (which corresponds to the phonemic representation model) of the map has large values that indicate a smaller linguistic gap.

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 B for its definition) to quantify the distance between two language distributions.

Method	Language								Performance gap		Linguistic Gap
	EN	DE	FR	KO	UK	UR	HI	SW	Std. (↓)	Mean diff. (↓)	Mean CKA (↑)
Named Entity Recognition											
Character	88.80	77.99	90.87	**80.17**	**85.61**	75.88	73.33	74.84	6.68	0.10	0.34
Phoneme	**91.36**	**81.36**	**91.93**	79.33	83.67	**76.41**	**82.19**	**78.06**	**5.79**	**0.08**	**0.43**
Part-of-Speech Tagging											
Character	**73.54**	**81.20**	**80.55**	70.1	**84.28**	93.15	-	-	8.14	0.12	0.30
Phoneme	70.50	78.68	78.10	**77.35**	83.75	**93.96**	-	-	**7.88**	0.12	**0.52**

Table 2: Performance of POS tagging and NER across different languages. Std. refers to the standard deviation of the scores across the languages, and Mean diff. indicates average pairwise difference of F1 scores. Mean CKA represents the average linguistic gap between languages.

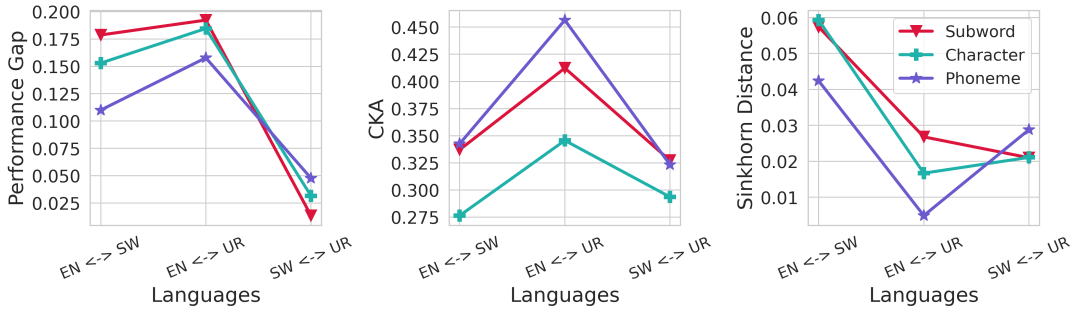

Figure 3: Qualitative analysis of performance gap (difference of accruacy) on XNLI task. (Left) the absolute difference between performance across two languages, (center) centered kernel alignment (CKA) scores to measure cross-lingual embedding similarity, and (right) Sinkhorn distance on the output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

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 B). 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 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).

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 that 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. Thus, we can not ensure the effectiveness of IPA representation when adopted within modern large language models, such as LLaMa 2 (Touvron et al., 2023). Additionally, we performed evaluation with a limited languages (up to 8), 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.

References

Maria-Florina Balcan, Rattana Pukdee, Pradeep Ravikumar, and Hongyang Zhang. 2023. Nash equilibria and pitfalls of adversarial training in adversarial robustness games. In *International Conference on Artificial Intelligence and Statistics*, pages 9607–9636. PMLR.

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 R Mortensen, Chris Dyer, and Jaime G 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.

Aditi Chaudhary, Chunting Zhou, Lori Levin, Graham Neubig, David R Mortensen, and Jaime G Carbonell. 2018. Adapting word embeddings to new languages with morphological and phonological subword representations. *arXiv preprint arXiv:1808.09500*.

Weile Chen, Huiqiang Jiang, Qianhui Wu, Börje Karlsson, and Yi Guan. 2021. AdvPicker: Effectively Leveraging Unlabeled Data via Adversarial Discriminator for Cross-Lingual NER. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 743–753, Online. Association for Computational Linguistics.

Xilun Chen, Yu Sun, Ben Athiwaratkun, Claire Cardie, and Kilian Weinberger. 2018. Adversarial deep averaging networks for cross-lingual sentiment classification. *Transactions of the Association for Computational Linguistics*, 6:557–570.

Zewen Chi, Li Dong, Furu Wei, Nan Yang, Saksham Singhal, Wenhui Wang, Xia Song, Xian-Ling Mao, Heyan Huang, and Ming Zhou. 2021. InfoXLM: An information-theoretic framework for cross-lingual language model pre-training. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 3576–3588, Online. 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, Édouard 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.

Alexis Conneau and Guillaume Lample. 2019. Cross-lingual language model pretraining. *Advances in neural information processing systems*, 32.

Alexis Conneau, Ruty Rinott, Guillaume Lample, Adina Williams, Samuel R. 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*. Association for Computational Linguistics.

Ryan Cotterell and Georg Heigold. 2017. Cross-lingual character-level neural morphological tagging. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pages 748–759, Copenhagen, Denmark. Association for Computational Linguistics.

Siddharth Dalmia, Xinjian Li, Alan W Black, and Florian Metze. 2019. Phoneme level language models

for sequence based low resource asr. In *ICASSP 2019-2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 6091–6095. IEEE.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2018. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*.

Jiatao Gu, Hany Hassan, Jacob Devlin, and Victor O.K. Li. 2018. Universal neural machine translation for extremely low resource languages. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, pages 344–354, New Orleans, Louisiana. 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.

Ke Hu, Antoine Bruguier, Tara N Sainath, Rohit Prabhavalkar, and Golan Pundak. 2019. Phoneme-based contextualization for cross-lingual speech recognition in end-to-end models. *arXiv preprint arXiv:1906.09292*.

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.

Guillaume Lample, Alexis Conneau, Marc'Aurelio Ranzato, Ludovic Denoyer, and Hervé Jégou. 2018. Word translation without parallel data. In *International Conference on Learning Representations*.

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

Paulius Micikevicius, Sharan Narang, Jonah Alben, Gregory Diamos, Erich Elsen, David Garcia, Boris Ginsburg, Michael Houston, Oleksii Kuchaiev, Ganesh Venkatesh, et al. 2018. Mixed precision training. In *International Conference on Learning Representations*.

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).

Linh The Nguyen, Thinh Pham, and Dat Quoc Nguyen. 2023. Xphonebert: A pre-trained multilingual model for phoneme representations for text-to-speech. *arXiv preprint arXiv:2305.19709*.

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.

Hugo Touvron, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Timothée Lacroix, Baptiste Rozière, Naman Goyal, Eric Hambro, Faisal Azhar, et al. 2023. Llama: Open and efficient foundation language models. *arXiv preprint arXiv:2302.13971*.

Xiangpeng Wei, Haoran Wei, Huan Lin, Tianhao Li, Pei Zhang, Xingzhang Ren, Mei Li, Yu Wan, Zhiwei Cao, Binbin Xie, et al. 2023. Polylm: An open source polyglot large language model. *arXiv preprint arXiv:2307.06018*.

Xiangpeng Wei, Rongxiang Weng, Yue Hu, Luxi Xing, Heng Yu, and Weihua Luo. 2021. On learning universal representations across languages. In *International Conference on Learning Representations*.

BigScience Workshop, Teven Le Scao, Angela Fan, Christopher Akiki, Ellie Pavlick, Suzana Ilić, Daniel Hesslow, Roman Castagné, Alexandra Sasha Luccioni, François Yvon, et al. 2022. Bloom: A 176b-parameter open-access multilingual language model. *arXiv preprint arXiv:2211.05100*.

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. *arXiv preprint arXiv:2205.04182*.

Pengfei Yu, Jonathan May, and Heng Ji. 2023. Bridging the gap between native text and translated text through adversarial learning: A case study on cross-lingual event extraction. In *Findings of the Association for Computational Linguistics: EACL 2023*, pages 754–769, Dubrovnik, Croatia. Association for Computational Linguistics.

Wanru Zhao, Yihong Chen, Royson Lee, Xinchi Qiu, Yan Gao, Hongxiang Fan, and Nicholas Donald Lane. 2023. Breaking physical and linguistic borders: Multilingual federated prompt tuning for low-resource languages. In *International Workshop on Federated Learning in the Age of Foundation Models in Conjunction with NeurIPS 2023*.

Bo Zheng, Li Dong, Shaohan Huang, Wenhui Wang, Zewen Chi, Saksham Singhal, Wanxiang Che, Ting

Liu, Xia Song, and Furu Wei. 2021. Consistency regularization for cross-lingual fine-tuning. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 3403–3417, Online. Association for Computational Linguistics.

A Details on Experiment Setup

A.1 Models

We compare LMs with three different types of language representation: subword, character, and phoneme. All models[5] are pre-trained on multilingual data that covers around 100 languages from Wikipedia dump files. We employ an off-the-shelf MLLMs, mBERT (Devlin et al., 2018). for a **subword-based** model which is trained on 104 languages with byte-pair encoding (vocabulary size of 110k). The model is pre-trained with the objective of masked language modeling and next sentence prediction tasks. For a **character-based** model, we utilize CANINE (Clark et al., 2022), which is a tokenization-free LM that directly maps each character to its codepoint by hashing. It is pre-trained on the same data and training objectives as mBERT. This prevents unknown tokens, enabling the model to handle a large amount of distinct characters. CANINE is pre-trained on the same data and training objectives as mBERT. For a **phoneme-based** model, we adopt XPhoneBERT (Nguyen et al., 2023), which has the same model architecture as mBERT and it is trained on 94 languages with learning objective of dynamic masked language modeling.

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. For a fair comparison between the representations, we compare phoneme-based model to character-based model instead of directly comparing it to widely used subword-based model (i.e., mBERT), and leave further improvements of overall performance as future work.

A.2 Data

G2P Conversion. In order to pass the input to phoneme-based model, we employed G2P (Grapheme-to-Phoneme) conversion on the data to obtain an IPA version of the input. This conversion was done with Epitran[6] (Mortensen et al., 2018) , an external tool for G2P conversion. After converting to IPA, we insert white space between every character to make it compatible with the XPhoneBERT tokenizer.

Languages. We categorize languages into three groups—high-resource (`eng`, `deu`, `fra`), low-resource (`urd`, `hin`, `swa`), and typologically distinct (`kor`, `ukr`)—to analyze the impact of phonemic representation with respect to each category. For high/low-resource language, we refer to Wu and Dredze (2020) and treat the languages with a wikisize under 8 as low-resource languages, and those with wikisize above 11 high-resource languages. On the other hand, typologically distinct languages are chosen with reference to English (they are orthographically and typologically different)[7].

A.3 Downstream Tasks

We adopt the cross-lingual generalization benchmark tasks suggested in XTREME (Hu et al., 2020). Among them, we selected two types of tasks; classification and structured prediction.

Sentence-level Classification. XTREME supports some sentence-level classification tasks, requiring semantic understanding of given sentences to make a prediction. We employ the **XNLI** (Conneau et al., 2018) corpus, which is well-known for cross-lingual evaluation, to train and evaluate the model on different languages.

Token-level Classification. Structured prediction tasks from Hu et al. (2020) include **POS tagging** and **NER**. 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 (Hu et al., 2020). We find this particularly suitable in our scenario where two models with different pre-training strategy are compared. We utilize the corpora from Universal Dependencies[8] for POS tagging, and WikiAnn (Pan et al., 2017) with train, dev, test splits following (Rahimi et al., 2019) for NER. While all 8 languages are employed for NER, we do not consider `swa` for POS tagging since Universal Dependencies does not provide a Swahili treebank.

[5] pre-trained weights were obtained from https://huggingface.co/models

[6] https://github.com/dmort27/epitran

[7] Note that this categorization is not mutually exclusive and that Urdu and Hindi can also be considered as typologically distant from English.

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

Evaluation. For evaluation, we follow the common practice of each task, evaluating XNLI task with top 1 accuracy and NER, POS tagging with F1 score. All metrics used are computed via functions from scikit-learn [9] python library.

A.4 Implementation details.

We fine-tuned all models with AdamW (Loshchilov and Hutter, 2018) optimization setting batch size as 128 over 20 epochs for XNLI and 40 epochs for NER and POS Tagging tasks. Here, we adopt a cosine learning rate scheduler with a warm-up on the XNLI task. For all models, we adopt mixed precision training (Micikevicius et al., 2018) provided by PyTorch, i.e., `with autocast()` loop for saving computational cost and energy consumption. For XNLI task, we used the embedding of [CLS] token for mBERT, and used the mean of all tokens except [CLS] token for CANINE and XPhoneBERT. All experiments were done with the random seed fixed to 42.

A.5 Number of parameters of the models.

Model	Num. of params
mBERT-base	177,853,440
CANINE-c	132,082,944
XPhoneBERT	87,554,304

Table 3: The number of parameters of each model. For experiments, we use

A.6 Computational resources.

All experiments on POS tagging were done on a single NVIDIA GeForce RTX 3090, and those on XNLI and NER were done on a single NVIDIA A6000. Total GPU hours for all experiments are 29 days.

A.7 Hyperparameter sweep.

We sweep hyperparameters over grid below (in Table 4), 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).

A.8 Datasets, Statistics, and License.

In Table 5, we provide the datasets, their statistics, and license.

[9] https://scikit-learn.org/stable/index.html

B 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 B.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 B.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	[1e-5, 5e-5, 1e-4] [1e-4, 1e-3, 1e-2]	1e-4 1e-3	5e-5 1e-4	5e-5 1e-2
POS	learning rate weight decay	[1e-5, 5e-5, 1e-4, 3e-4, 1e-2] [1e-4, 1e-3, 1e-2]	1e-4 1e-4	3e-4 1e-2	1e-4 1e-2

Table 4: 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.

Dataset	Lang.	Train	Dev	Test	License
XNLI	eng swa urd	393k	2.49k	5.01k	CC BY-NC-4.0
WikiAnn	eng deu fra kor ukr urd swa hin	20k 20k 20k 20k 20k 20k 1k 5k	10k 10k 10k 10k 10k 1k 1k 1k	10k 10k 10k 10k 10k 1k 1k 1k	ODC-BY
UD	eng deu fra kor ukr urd	12.5k 13.8k 14.5k 23k 5.5k 4k	2k 0.8k 1.5k 2k 0.7k 0.6k	2k 1k 0.4k 2.3k 0.9k 0.5k	CC BY-SA 4.0 CC BY-SA 4.0 CC BY-SA 4.0 CC BY-SA 4.0 CC BY-NC-SA 4.0 CC BY-NC-SA 4.0

Table 5: Statistics and license types for datasets. The table lists the number of examples in the training, development, and testing sets for languages in the XNLI, WikiAnn, and UD datasets. It specifies the licensing conditions: CC-BY permits sharing and adaptation, CC BY-NC is for non-commercial usage, CC BY-SA mandates share-alike distributions, and CC BY-NC-SA combines non-commercial with share-alike terms. All datasets are strictly used within the bounds of these licenses.

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 B.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 B.3 into the above bound, we have Theorem B.2.

□

From Theorem B.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.