Correlations Between Multilingual Language Model Geometry and Crosslingual Transfer Performance

Cheril Shah[1], **Yashashree Chandak**[1*], **Atharv Mahesh Mane**[2*],
Benjamin K. Bergen[3], **Tyler A. Chang**[3]

[1]Pune Institute of Computer Technology
[2]Birla Institute of Technology and Science, Pilani, Goa
[3]University of California San Diego
{shahcheril311, chandakyashashree304, atharvmeenakshi}@gmail.com
{bkbergen, tachang}@ucsd.edu

Abstract

A common approach to interpreting multilingual language models is to evaluate their internal representations. For example, studies have found that languages occupy distinct subspaces in the models' representation spaces, and geometric distances between languages often reflect linguistic properties such as language families and typological features. In our work, we investigate whether geometric distances between language representations correlate with zero-shot crosslingual transfer performance for POS-tagging and NER in three multilingual language models. We consider four distance metrics, including new metrics that identify a basis for a multilingual representation space that sorts axes based on their language-separability. We find that each distance metric either only moderately correlates or does not correlate with crosslingual transfer performance, and metrics do not generalize well across models, layers, and tasks. Although pairwise language separability is a reasonable predictor of crosslingual transfer, representational geometry overall is an inconsistent predictor for the crosslingual performance of multilingual language models.

Keywords: multilinguality, explainability, evaluation methodologies

1. Introduction

Pre-trained multilingual language models represent multiple languages in a single vector space, a feature which is hypothesized to enable their impressive crosslingual transfer capabilities (Conneau et al., 2020). Still, languages occupy distinct subspaces in the common model embedding space (Chang et al., 2022), and geometric distances between languages correlate with phylogenetic distances (Rama et al., 2020) and typological similarities (Choenni and Shutova, 2022). Similarities and differences between language geometries impact downstream model performance for parallel sentence retrieval (Libovický et al., 2020; Pires et al., 2019) and raw language modeling performance (Chang et al., 2022).

It is then natural to hypothesize that geometric distances between languages might also correlate with crosslingual transfer capabilities (e.g. fine-tuning on language A and evaluating on language B), which vary substantially across language pairs in multilingual language models (Pires et al., 2019; Wu and Dredze, 2019, 2020). Crosslingual transfer capabilities correlate with features such as syntactic, geographic, and genetic similarities between languages (Karthikeyan et al., 2020; Philippy et al., 2023), along with shared morphological systems (Gerz et al., 2018) and writing systems (Fujinuma et al., 2022). However, the effects of geometric distances on fine-tuned crosslingual transfer between languages have not been investigated.[1] If connections between representational geometry and crosslingual transfer are established, then we may better predict model performance on zero-shot crosslingual transfer and potentially improve model training for better performance.

Thus, we consider four geometric metrics to quantify the distances between languages in three multilingual language models, and we study their correlations with crosslingual transfer performance for part-of-speech (POS) tagging and named entity recognition (NER). Although pairwise language separability is a reasonable indicator for transfer performance, geometric measures in general do not consistently correlate with transfer performance across all models, layers, or tasks. These results suggest that geometric features are extremely noisy signals for multilingual model performance.[2]

2. Related Work

Previous work has used geometric measures between languages in the representation space to ex-

*Equal second-authorship.

[1] Philippy et al. (2023) quantify the evolution of languages' representation spaces during fine-tuning, without focusing on language distances before fine-tuning.

[2] Code is available at: https://github.com/Cheril311/Crosslingual_geometry

LREC-COLING 2024, pages 4059–4066

plain multilingual model behavior and to design better cross-lingual transfer mechanisms. Nakashole (2018) use the structure of word embedding spaces for different languages to create neighborhood-sensitive mappings for word translation. Similarly, Alaux et al. (2019) align the word embedding spaces for multiple languages into a single vector-space. Chang et al. (2022) use subspace distances and LDA-based analyses to study how information in different languages is encoded along orthogonal language-sensitive and language-neutral axes. Finally, Shah et al. (2023) use PCA to demonstrate the separability of representation spaces for different language families, measuring distances between languages in semantic space. Our work evaluates whether these geometric measures are correlated with downstream crosslingual transfer performance.

3. Method

We compute four metrics that capture different types of distances between languages in three multilingual language models' representation spaces. We correlate these geometric metrics with crosslingual transfer performance from English to 25 languages for POS-tagging and NER.

3.1. Models and Datasets

We extract representations and evaluate downstream crosslingual transfer performance for three multilingual language models: mBERT (Devlin et al., 2019), XLM-R (Conneau et al., 2020), and mDeBERTa-V3 (He et al., 2023). Each model has 12 Transformer layers; to extract representations, we input sentences from the OSCAR corpus (Abadji et al., 2022) and the Universal Dependencies dataset (Nivre et al., 2020), and we consider the token representations after Transformer layers 3, 8, and 11. We use roughly 13K token representations per language. We consider 26 languages that appear in the pre-training data for all the three models and that have both POS-tagging and NER data available. We use the Universal Dependencies dataset (Nivre et al., 2020) for POS-tagging and the WikiANN dataset (Pan et al., 2017) for Named Entity Recognition (NER).

3.2. Language Centroid Distances ($\mathcal{D}$)

A common way to quantify distances between languages in a multilingual representation space is to compute distances between representation centroids (means; Libovický et al., 2020; Choenni and Shutova, 2022). As in previous work, we define the centroid $\boldsymbol{c}_L \in \mathbb{R}^d$ for a language L as the arithmetic mean of the token representations in L:

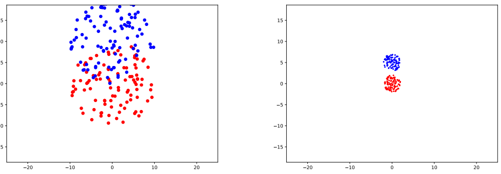

Figure 1: Subspaces with high centroid distances despite low separability (left), versus lower centroid distances with high separability (right).

$$\boldsymbol{c}_L = \frac{1}{n_L}\sum_{i=1}^{n_L} \boldsymbol{x}_i \tag{1}$$

Here, $\boldsymbol{x}_i$ is the vector representation for an individual token in context, and n_L is the total number of token representations in L. The centroid distance $\mathcal{D}$ between two languages is the Euclidean distance between their centroids.[3] We compute these distances separately for layers 3, 8, and 11.

3.3. Subspace Distances ($\mathcal{S}$)

$\mathcal{D}$ values assume that a language centroid effectively represents the entire subspace spanned by a language's token representations. However, this is often not the case for high-dimensional data (Assent, 2012). Thus, following Chang et al. (2022), we compute the distance between representation covariance matrices $\boldsymbol{K}_{L_1}, \boldsymbol{K}_{L_2} \in \mathbb{R}^{d \times d}$ for languages L_1 and L_2:

$$\mathcal{S}(\boldsymbol{K}_{L_1}, \boldsymbol{K}_{L_2}) = \sqrt{\sum_i \log^2(\lambda_i)} \tag{2}$$

Here, λ_i are the d positive real eigenvalues of $\boldsymbol{K}_{L_1}^{-1}\boldsymbol{K}_{L_2}$ (Bonnabel and Sepulchre, 2009). This metric captures dissimilarities in the shapes of the two language subspaces, after mean-centering.

3.4. Computing a Basis in Order of Language-Separability

While centroid and subspace distances ($\mathcal{D}$ and $\mathcal{S}$) allow us to quantify distances between language subspaces, they do not necessarily provide information about separations between the subspaces. Subspaces with low centroid and subspace distances between them can still be highly separated. For example, while the subspaces in Figure 1 (right) are intuitively more separated than the subspaces in Figure 1 (left), the centroid distance $\mathcal{D}$ in the left plot is higher. A metric for separability must consider differences between language centroids while accounting for the subspace covariances.

[3] We find that cosine distances between language centroids produce similar results to Euclidean distances.

To approach the issue of language separability, we compute a new orthonormal basis for a multilingual representation space, where axes are sorted by language-separability. First, we use linear discriminant analysis (LDA) on our token representations for all languages (using source language as a label) to obtain axes that maximally separate languages (Liang et al., 2021). The first LDA axis $\boldsymbol{v}_0$ is our initial most language-separable axis, and it initializes our basis as $\boldsymbol{V} = \boldsymbol{v}_0 \in \mathbb{R}^{d \times 1}$.

Then, we repeat the following. We project all token representations $\boldsymbol{X} \in \mathbb{R}^{d \times n}$ onto the existing language-separable axes $\boldsymbol{V}$, and we subtract the projections from the original representations.

$$\boldsymbol{X} - \boldsymbol{V}\boldsymbol{V}^T\boldsymbol{X} \tag{3}$$

This sets the existing language-separable axes to a fixed value across all representations, essentially removing the information encoded by those axes. We run LDA again on the adjusted representations. Then, the first LDA axis $\boldsymbol{v}_i$ is the most language-separable axis after excluding the already-identified axes. We orthonormalize $\boldsymbol{v}_i$ relative to the existing axes $\boldsymbol{V}$, and we concatenate to update the basis.

$$\boldsymbol{V} \leftarrow [\boldsymbol{V}, \boldsymbol{v}_i] \tag{4}$$

We repeat until the basis $\boldsymbol{V}$ spans the entire representation space (i.e. $\boldsymbol{V} \in \mathbb{R}^{d \times d}$). Based on how we define the axes, earlier axes are the most language-separable (representations are maximally separated by language along these axes) and later axes are the most language neutral.

3.4.1. Separability Across Axes and Models

We quantify the language-separability of an axis $\boldsymbol{v}$ in the new basis $\boldsymbol{V}$ by projecting all token representations onto $\boldsymbol{v}$ (i.e. each representation is projected to a single scalar value) and then calculating their one-way ANOVA F-statistic (i.e. variance between languages divided by variance within languages).

$$F_{\boldsymbol{v}} = \frac{\sum_L n_L (c_L - \mu)^2}{\sum_L \sum_{i=1}^{n_L} (x_{L,i} - c_L)^2} \tag{5}$$

Here, n_L is the number of representations in language L, c_L is the centroid for language L, and μ is the centroid of all representations across all languages. Individual representations in language L are denoted $x_{L,i}$. A high F-statistic indicates higher language-separability.

To identify general trends in how languages are separated across axes in different models, we plot the F-statistics for the basis vectors identified in §3.4 for each model (i.e. axes sorted by language-separability). As shown in Figure 2, mBERT is the most language-separable model for all tested layers in the initial axes (high F-statistics). XLM-R also has relatively high F-statistics in layer three for initial axes, but it is still almost half that of mBERT. However, in later axes (more language-neutral axes), XLM-R is more language-separable than mBERT. Based on this result, it may be that mBERT more aggressively concentrates language-specific information in fewer language-separable axes, allowing other axes to be more language-neutral; XLM-R may distribute language-specific information more evenly across axes.

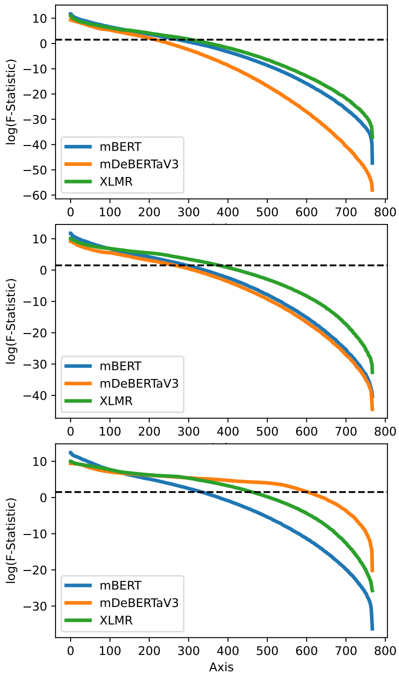

Figure 2: F-statistic curves using our new bases that sort axes by language-separability. Higher F-statistics indicate higher language-separability. The horizontal line indicates where language separability would be statistically significant for an axis at $p = 0.05$ (not adjusted for multiple comparisons). We consider all three models for layers 3 (top), 8 (middle), and 11 (bottom). All three models have the same representation dimensionality ($d = 768$).

Language-separability in mDeBERTa-V3 appears low overall in layers three and eight (low F-statistics), but high in the tail end of layer eleven; based on this result, it may be that mDeBERTa-V3 concentrates language-specific processing in later layers. These trends likely reflect differences in the models' pre-training strategies (e.g. mDeBERTa-V3 uses ELECTRA's discriminative pre-training paradigm; Clark et al., 2020) or datasets. Differences in how language-specific information is distributed in different models may explain inconsistent correlations between geometric measures and downstream model performance in different multilingual models (§4).

3.4.2. Language Isolation ($\mathcal{I}$)

We use this new basis and corresponding F-statistics to quantify the separability of individual

	POS mBERT	NER mBERT	POS mDeBERTa-V3	NER mDeBERTa-V3	POS XLM-R	NER XLM-R
Layer 3, $\mathcal{D}$	-0.65^{*}	-0.37	-0.04	-0.03	-0.40	-0.25
Layer 8, $\mathcal{D}$	-0.70^{**}	-0.55^{*}	-0.02	0.04	-0.48	-0.54
Layer 11, $\mathcal{D}$	-0.64	-0.31	-0.40	-0.32	-0.50	-0.54
Layer 3, $\mathcal{S}$	-0.53	-0.50	0.14	0.26	-0.28	-0.09
Layer 8, $\mathcal{S}$	-0.26	-0.07	0.09	0.08	-0.18	-0.04
Layer 11, $\mathcal{S}$	0.08	0.26	-0.23	-0.27	-0.16	-0.10
Layer 3, $\mathcal{I}$	-0.47	-0.50	-0.18	-0.07	-0.60	-0.47
Layer 8, $\mathcal{I}$	-0.51^{*}	-0.55	-0.38	-0.18	-0.30	-0.22
Layer 11, $\mathcal{I}$	-0.56	-0.57^{*}	-0.56	-0.51	-0.16	-0.03
Layer 3, ψ	-0.65^{*}	-0.60	-0.26	-0.17	-0.66^{*}	-0.58
Layer 8, ψ	-0.66^{*}	-0.57	-0.30	-0.03	-0.60	-0.56^{*}
Layer 11, ψ	-0.65	-0.64	-0.64	-0.37	-0.55	-0.49

Table 1: Pearson's correlation coefficient r between crosslingual transfer performance and geometric distance for each metric, model, and layer. An asterisk indicates $p < 0.05$, and two asterisks indicate $p < 0.01$ after adjusting for multiple comparisons using Bonferroni correction.

language subspaces. First, we note that the separability of an individual language will have an effect on the sum of F-statistics ($\sum_{\boldsymbol{v}} F_{\boldsymbol{v}}$) over all axes in a model. If a language L is highly isolated from other languages, then the F-statistics ($F_{\boldsymbol{v},\sim L}$) when removing L will be lower. Thus for every language L, we compute the area under the curve dividing $F_{\boldsymbol{v}}$ (F-statistic including all languages) by $F_{\boldsymbol{v},\sim L}$ (F-statistic removing L):

$$\mathcal{I}_L = \sum_{\boldsymbol{v}} \frac{F_{\boldsymbol{v}}}{F_{\boldsymbol{v},\sim L}} \tag{6}$$

A higher value of $\mathcal{I}_L$ indicates that the language L is more isolated from other languages in the multilingual representation space.

3.4.3. Pairwise Separability (ψ)

Language isolation $\mathcal{I}$ is a metric for the isolation of an individual language from all other languages. It is also interesting to consider how languages are separated from one another pairwise. Thus, we consider F-statistics ($F_{\boldsymbol{v},L_1,L_2}$) when including only representations from a pair of languages L_1 and L_2. These values are high when those two languages are highly separable. We then calculate the area under the curve:

$$\psi = \sum_{\boldsymbol{v}} \frac{F_{\boldsymbol{v},L_1,L_2}}{F_{\boldsymbol{v}}} \tag{7}$$

When two languages are highly separable relative to the overall language-separability, then ψ is high.

3.5. Downstream Task Performance

To quantify POS-tagging and NER performance, we compute F1 scores after fine-tuning a model without freezing any layers. We add only one fully connected layer for task prediction, and we use AdamW (Loshchilov and Hutter, 2019) with learning rate 5e-5 (Devlin et al., 2019). We fine-tune each model on 3 epochs of 1K English sentences, and we evaluate performance on 400 sentences in each of the 25 non-English evaluation languages: Afrikaans, Arabic, Basque, Bulgarian, Dutch, Estonian, Finnish, French, German, Greek, Hebrew, Hindi, Hungarian, Indonesian, Italian, Japanese, Mandarin, Persian, Portuguese, Russian, Spanish, Thai, Turkish, Urdu, and Vietnamese. The mean F1 score across languages and models (zero-shot transfer) is 0.69 for POS-tagging and 0.63 for NER. We correlate crosslingual transfer performance with each geometric distance metric between source and target language ($\mathcal{D}$, $\mathcal{S}$, ψ) and the geometric isolation of the target language ($\mathcal{I}$).

4. Results and Discussion

Table 1 reports correlations between different geometric distance metrics and crosslingual transfer performance for different models, layers, and tasks. In general, higher distances between languages correlate with worse transfer performance, but none of the metrics show a generalizable correlation across all models and tasks. In 43 out of 72 cases (3 layers × 2 tasks × 3 models × 4 metrics), correlations are less than $r = 0.50$.

Language centroid distances $\mathcal{D}$ correlate with POS-tagging crosslingual transfer for all layers in mBERT ($r = -0.64$ to -0.70), but the effect is much weaker for NER ($r = -0.31$ to -0.55). This aligns with results that POS information is encoded multilingually in multilingual language models across most layers (Chang et al., 2022). However, $\mathcal{D}$ is only moderately correlated with downstream performance in XLM-R ($r = -0.25$ to -0.54), and it is not correlated with downstream performance in mDeBERTa-V3 until later layers (layer 11).

Subspace distances $\mathcal{S}$ (after mean-centering) have fairly low correlations with transfer performance, with only 2 out of 18 correlations stronger than $r = -0.30$. This suggests that differing subspace means impact crosslingual transfer moreso than differing shapes (covariances).

Isolation of the target language $\mathcal{I}$ is moderately correlated with crosslingual transfer for both tasks in mBERT ($r = -0.47$ to -0.57; higher isolation correlating with worse transfer), but this effect is not observed across layers in XLM-R and mDeBERTa-V3. When considering languages pairwise, the separability ψ of source and target language is moderately correlated with downstream task performance in both mBERT and XLM-R and for both tasks ($r = -0.49$ to -0.66), but the correlations are weak for mDeBERTa-V3 except in later layers for POS-tagging. In general, correlations between geometric distances and transfer performance are higher in later layers for mDeBERTa-V3, suggesting that multilingual geometry may change more across layers in mDeBERTa-V3 than in other models. In any case, the slightly more consistent correlations for language separability ψ across models and tasks (relative to other geometric metrics) suggest that our language separability metrics encode useful geometric properties that correlate moderately with downstream crosslingual transfer.

5. Conclusion

We find inconsistent correlations between languages' geometric distances in model representation space and crosslingual transfer performance in multilingual language models. None of the evaluated geometric metrics correlate with transfer performance across all models, layers, and tasks. Of the evaluated metrics, pairwise separability of languages in late layers is a reasonable predictor for crosslingual transfer performance, but correlations are still only moderate. These results suggest that while geometric distances can provide insights into internal model mechanisms, better metrics may better correlate with downstream performance.

Limitations

Although we consider three multilingual language models, our work omits several larger and more recent models such as BLOOM (Scao et al., 2022) and XGLM (Lin et al., 2022) due to compute limitations. We also only consider crosslingual transfer from English to 25 languages. Future work could focus on the properties of geometric measures in more recent multilingual models and for a more diverse set of languages. Additionally, future work might consider metrics based on model parameters themselves, or metrics based on language-specific subnetworks that causally influence outputs (e.g. Foroutan et al., 2022).

6. Bibliographical References

Julien Abadji, Pedro Ortiz Suarez, Laurent Romary, and Benoît Sagot. 2022. Towards a Cleaner Document-Oriented Multilingual Crawled Corpus. *arXiv e-prints*, page arXiv:2201.06642.

Jean Alaux, Edouard Grave, Marco Cuturi, and Armand Joulin. 2019. Unsupervised hyperalignment for multilingual word embeddings. In *International Conference on Learning Representations*.

Ira Assent. 2012. Clustering high dimensional data. *WIREs Data Mining and Knowledge Discovery*, 2(4):340–350.

Silvére Bonnabel and Rodolphe Sepulchre. 2009. Riemannian metric and geometric mean for positive semidefinite matrices of fixed rank. *SIAM Journal on Matrix Analysis and Applications*, 31:1055–1070.

Tyler Chang, Zhuowen Tu, and Benjamin Bergen. 2022. The geometry of multilingual language model representations. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 119–136, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Rochelle Choenni and Ekaterina Shutova. 2022. Investigating language relationships in multilingual sentence encoders through the lens of linguistic typology. *Computational Linguistics*, 48(3):635–672.

Kevin Clark, Minh-Thang Luong, Quoc V. Le, and Christopher D. Manning. 2020. ELECTRA: Pre-training text encoders as discriminators rather than generators. In *International Conference on Learning Representations*.

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.

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.

Negar Foroutan, Mohammadreza Banaei, Rémi Lebret, Antoine Bosselut, and Karl Aberer. 2022. Discovering language-neutral sub-networks in multilingual language models. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 7560–7575, Abu Dhabi, United Arab Emirates. 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.

Daniela Gerz, Ivan Vulić, Edoardo Maria Ponti, Roi Reichart, and Anna Korhonen. 2018. On the relation between linguistic typology and (limitations of) multilingual language modeling. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 316–327, Brussels, Belgium. Association for Computational Linguistics.

Pengcheng He, Jianfeng Gao, and Weizhu Chen. 2023. DeBERTav3: Improving deBERTa using ELECTRA-style pre-training with gradient-disentangled embedding sharing. In *The Eleventh International Conference on Learning Representations*.

Kaliyaperumal Karthikeyan, Zihan Wang, Stephen Mayhew, and Dan Roth. 2020. Cross-lingual ability of multilingual bert: An empirical study. In *International Conference on Learning Representations*.

Sheng Liang, Philipp Dufter, and Hinrich Schütze. 2021. Locating language-specific information in contextualized embeddings. *ArXiv*, abs/2109.08040.

Jindřich Libovický, Rudolf Rosa, and Alexander Fraser. 2020. On the language neutrality of pre-trained multilingual representations. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 1663–1674, Online. Association for Computational Linguistics.

Xi Victoria Lin, Todor Mihaylov, Mikel Artetxe, Tianlu Wang, Shuohui Chen, Daniel Simig, Myle Ott, Naman Goyal, Shruti Bhosale, Jingfei Du, Ramakanth Pasunuru, Sam Shleifer, Punit Singh Koura, Vishrav Chaudhary, Brian O'Horo, Jeff Wang, Luke Zettlemoyer, Zornitsa Kozareva, Mona Diab, Veselin Stoyanov, and Xian Li. 2022. Few-shot learning with multilingual generative language models. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 9019–9052, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Ilya Loshchilov and Frank Hutter. 2019. Decoupled weight decay regularization. In *International Conference on Learning Representations*.

Ndapa Nakashole. 2018. NORMA: Neighborhood sensitive maps for multilingual word embeddings. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 512–522, Brussels, Belgium. Association for Computational Linguistics.

Joakim Nivre, Marie-Catherine de Marneffe, Filip Ginter, Jan Hajič, Christopher D. Manning, Sampo Pyysalo, Sebastian Schuster, Francis Tyers, and Daniel Zeman. 2020. Universal Dependencies v2: An evergrowing multilingual treebank collection. In *Proceedings of the Twelfth Language Resources and Evaluation Conference*, pages 4034–4043, Marseille, France. European Language Resources Association.

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.

Fred Philippy, Siwen Guo, and Shohreh Haddadan. 2023. Identifying the correlation between language distance and cross-lingual transfer in a multilingual representation space. In *Proceedings of the 5th Workshop on Research in Computational Linguistic Typology and Multilingual NLP*, pages 22–29, Dubrovnik, Croatia. Association for Computational Linguistics.

Telmo Pires, Eva Schlinger, and Dan Garrette. 2019. How multilingual is multilingual BERT? In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 4996–5001, Florence, Italy. Association for Computational Linguistics.

Taraka Rama, Lisa Beinborn, and Steffen Eger. 2020. Probing multilingual BERT for genetic and typological signals. In *Proceedings of the 28th*

International Conference on Computational Linguistics, pages 1214–1228, Barcelona, Spain (Online). International Committee on Computational Linguistics.

Teven Le Scao, Angela Fan, Christopher Akiki, Elizabeth-Jane Pavlick, Suzana Ili'c, Daniel Hesslow, Roman Castagn'e, Alexandra Sasha Luccioni, Franccois Yvon, Matthias Gallé, Jonathan Tow, Alexander M. Rush, Stella Rose Biderman, Albert Webson, Pawan Sasanka Ammanamanchi, Thomas Wang, Benoît Sagot, Niklas Muennighoff, Albert Villanova del Moral, Olatunji Ruwase, et al. 2022. Bloom: A 176b-parameter open-access multilingual language model. *arXiv*.

Cheril Shah, Yashashree Chandak, and Manan Suri. 2023. The geometry of multilingual language models: An equality lens. In *The First Tiny Papers Track at ICLR 2023, Tiny Papers @ ICLR 2023, Kigali, Rwanda, May 5, 2023*.

Shijie Wu and Mark Dredze. 2019. Beto, bentz, becas: The surprising cross-lingual effectiveness of BERT. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 833–844, Hong Kong, China. 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.

A. Languages Included

Languages included for crosslingual transfer performance are listed in Table 2. We consider transfer from English to each non-English target language.

Language	**Language Family**
Hindi	Indo-European (Indo-Iranian branch)
German	Indo-European (Germanic branch)
Indonesian	Austronesian (Malayo-Polynesian branch)
Russian	Indo-European (Slavic branch)
Urdu	Indo-European (Indo-Iranian branch)
French	Indo-European (Romance branch)
English	Indo-European (Germanic branch)
Basque	Isolate
Greek	Indo-European (Hellenic branch)
Hebrew	Afro-Asiatic (Semitic branch)
Italian	Indo-European (Romance branch)
Mandarin	Sino-Tibetan (Sinitic branch)
Persian	Indo-European (Indo-Iranian branch)
Afrikaans	Indo-European (Germanic branch)
Hungarian	Uralic (Finno-Ugric branch)
Spanish	Indo-European (Romance branch)
Estonian	Uralic (Finno-Ugric branch)
Dutch	Indo-European (Germanic branch)
Turkish	Turkic
Finnish	Uralic (Finno-Ugric branch)
Portuguese	Indo-European (Romance branch)
Thai	Kra-Dai (Tai-Kadai branch)
Arabic	Afro-Asiatic (Semitic branch)
Bulgarian	Indo-European (Slavic branch)
Vietnamese	Austroasiatic (Vietic branch)
Japanese	Japonic

Table 2: Included languages and their language families.