An Isotropy Analysis in the Multilingual BERT Embedding Space

Anonymous ACL submission

Abstract

Several studies have explored various advantages of multilingual pre-trained models (e.g., multilingual BERT) in capturing shared linguistic knowledge. However, their limitations have not been paid enough attention to. In this paper, we investigate the representation degeneration problem and outlier dimensions in multilingual contextual word representations (CWRs) of BERT. We show that though mBERT exhibits no outliers among its representations, its multilingual embedding space is highly anisotropic. Furthermore, our experimental results demonstrate that similarly to their monolingual counterparts, increasing the isotropy of multilingual embedding spaces can significantly improve their representation power and performance. Our analysis indicates that, although the degenerated directions vary in different languages, they encode similar linguistic knowledge, suggesting a shared linguistic space among languages.

1 Introduction

The multilingual BERT model (Devlin et al., 2019, mBERT), pre-trained on 104 languages with no supervision, has shown impressive ability in capturing linguistic knowledge across different languages (Pires et al., 2019). Many studies have explored the encoded knowledge in multilingual CWRs using probing tasks and under zero-shot setting (Wu and Dredze, 2019; K et al., 2020; Chi et al., 2020). Following the probing studies, in this paper, we investigate the multilingual embedding space of BERT, focusing on its geometry in terms of isotropy. Previous research has shown that many pre-trained models, such as GPT-2 (Radford et al., 2019), BERT, and RoBERTa (Liu et al., 2019) have degenerated embedding spaces that downgrade their semantic expressiveness (Ethayarajh, 2019; Cai et al., 2021; Rajaee and Pilehvar, 2021). Several proposals have been put forward to overcome this challenge (Gao et al., 2019; Zhang et al., 2020).

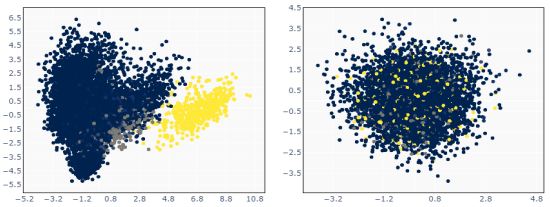


Figure 1: Degenerated (left) and isotropic (right) embedding spaces for Arabic. Frequency-based distribution can be easily detected using two top PCs in the space (lighter colors indicate higher frequency). See Appendix A for more languages.

However, to our knowledge, no study has so far been conducted on the degeneration problem in the multilingual embedding space.

Using two well-known metrics, we evaluate isotropy in the mBERT embedding space for three different languages: English, Arabic, and Spanish. We find that the representation spaces are massively anisotropic in all these languages. Extending our study to other structural properties of multilingual space, we investigate outliers, specific dimensions with consistently high values, in multilingual CWRs (Kovaleva et al., 2021). Our findings reveal that, as opposed to pre-trained BERT, the multilingual space does not involve any major outliers. This indicates that the suggestion of Luo et al. (2021) on the role of positional embeddings on the emergence of outliers may not be valid. Furthermore, we study the outliers' effects on similarity-based metrics (e.g., cosine similarity) using multilingual CWRs. We show that, unlike monolingual CWRs where a few dimensions dominate the cosine similarity metric (Timkey and van Schijndel, 2021a), all dimensions of multilingual representations have almost a uniform contribution to such metrics. Moreover, our analysis reveals that word frequency plays an important role in the distribution of the multilingual embedding space: words with similar frequencies create distinct local regions in the embedding space.

In analyzing multilingual space, we take a further step toward making the space isotropic. By applying a cluster-based isotropy enhancement method (Rajaee and Pilehvar, 2021), we demonstrate that increasing isotropy of multilingual embedding space can result in significant performance improvements on downstream tasks. Our frequency analysis and the remarkable performance improvement in the zero-shot setting denote that the feature space of mBERT has a similar structure across different languages.

2 Isotropy

In this section, we first provide reasons for the importance of isotropy in the embedding space. Then, we introduce the metrics used to quantify embedding space isotropy.

Geometrically, in anisotropic embedding space, embeddings occupy a narrow cone. This brings about an over-estimation of the cosine similarity of word embeddings (Gao et al., 2019). In other words, randomly sampled words will have high cosine similarity. As a result, anisotropic distribution reduces the effectiveness of similarity-based metrics.

2.1 Metrics

To quantify isotropy, we utilize two well-known metrics based on cosine similarity and principal components (PCs).

Cosine Similarity. Ethayarajh (2019) used cosine similarity between random embeddings as an approximation of isotropy in the space. As mentioned before, random embeddings with an isotropic distribution have near-zero cosine similarities. The metric can be formulated as follows:

$$I_{Cos}(\mathcal{W}) = \frac{1}{N} \sum_{i=1, x_i \neq y_i}^{N} Cos(x_i, y_i) \quad (1)$$

where $x_i \in X, y_i \in Y$, X and Y are the sets of randomly sampled embeddings, and $\mathcal{W}$ is the embedding matrix. N is the number of sampled pairs that is set to 1000 in our experiments. Lower $I_{Cos}(\mathcal{W})$ values indicate higher isotropy.

Principal Components. Mu and Viswanath (2018) have proposed a metric based on principal components (PCs), which is approximated as follows:

$$I_{PC}(\mathcal{W}) \approx \frac{min_{u \in U} F(u)}{max_{u \in U} F(u)} \quad (2)$$

		mBERT		
	BERT	**Arabic**	**English**	**Spanish**
$I_{Cos}(\mathcal{W})$	0.38	0.35	0.34	0.36
$I_{PC}(\mathcal{W})$	2.61E-06	8.99E-5	2.64E-06	3.39E-05

Table 1: The isotropy of BERT and mBERT on multilingual STS, reporting based on $I_{Cos}(\mathcal{W})$ and $I_{PC}(\mathcal{W})$.

$$F(u) = \sum_{i=1}^{M} exp(u^T w_i) \quad (3)$$

where w_i is the i^{th} word embedding, M is the number of all representations in the space, U is the set of eigenvectors of the embedding matrix, and $F(u)$ is the partition function described in Equation 3. Arora et al. (2016) proved that $F(u)$ could be approximated using a constant for isotropic embedding spaces. Therefore, $I_{PC}(\mathcal{W})$ would be close to one in an isotropic embedding space

3 Analysis

For all our experiments, we opted for the multilingual BERT model (mBERT) which has a 12-layer transformer-based architecture similar to English BERT-base and has been trained on 104 languages. As our evaluation benchmark, we experimented with the multilingual and cross-lingual Semantic Textual Similarity (Cer et al., 2017, STS) that involves instances from Arabic, English, and Spanish (appendix B).

In the first place, we assess the isotropy defined as a desirable property in multilingual space and investigate outliers introduced as an influential factor on isotropy. We also expand our study to rouge dimensions disrupting similarity-based metrics used in measuring isotropy. Lastly, we analyze word frequency bias, another destructive feature, in multilingual embedding space.

3.1 Probing isotropy

As the first step, we quantify the isotropy of the mBERT and BERT embedding spaces using the two metrics. For mBERT, we separately assess the isotropy of each language in the embedding space. The results in Table 1 reveal that the anisotropy issue exists for the multilingual BERT embedding space as well as the original BERT model. Aligned with the numerical results, the illustration of multilingual CWRs in the left column of Figure 1 gives us a clear perspective of the degenerated distribution in space.

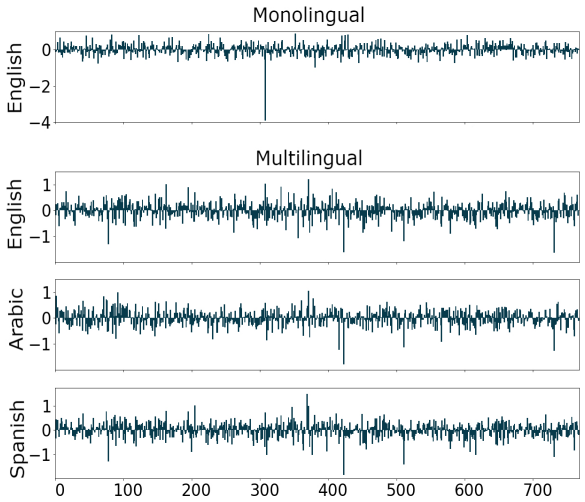


Figure 2: The average representation in English BERT (top) and mBERT (bottom). While an outlier has emerged in the former, we do not see any major outliers in the multilingual space.

3.2 Outlier Dimensions

Kovaleva et al. (2021) have found that pre-trained LMs exhibit consistent outliers, peculiar dimensions with large values, in their contextual representations across all layers. Through several experiments, they have demonstrated that disabling these outliers can notably impair the performance of pre-trained and fine-tuned LMs. These rogue dimensions can easily make the models vulnerable to adversarial attacks. Luo et al. (2021) showed that removing positional embeddings disappears the outliers, concluding that the positional information is responsible for the emergence of outliers.

We checked for rogue dimensions by averaging over all representations on the multilingual STS dataset. Results are shown in Figure 2. On top, the outlier dimension can be easily seen in the original BERT. However, interestingly, multilingual BERT exhibits no such outliers in its embedding space across different languages. It can be concluded that, in contrary to the suggestion of Luo et al. (2021), positional embeddings cannot be responsible for outliers, given that both multi- and mono-lingual spaces are constructed using the same training procedure involving positional encodings. We leave further investigation of outliers in contextual embedding space to future work.

3.3 Sensitivity to Rogue Dimensions

As we discussed before, cosine similarity is a widely used metric to measure the degree of isotropy in embedding space. Employing a dimension-based similarity, Timkey and van Schijndel (2021b) have shown that only a few dimensions dominate the high cosine similarity between any arbitrary representations in pre-trained LMs (e.g., BERT, RoBERTa, and XLNET). Therefore, anisotropy in such models is determined by a small fraction of dimensions (Hence, not a global property of the space). Following their approach, we compute the contribution of the i^{th} dimension in the cosine similarity of two embeddings:

$$CC_i = \frac{x_i y_i}{\|x\|\|y\|} \quad (4)$$

We compute the average cosine similarity, $I_{Cos}(\mathcal{W})$, by randomly sampling 1000 token pairs and report the average contribution of the top-three dimensions to the average cosine similarity.

	$I_{Cos}(\mathcal{W})$	First	Second	Third
BERT	0.38	0.191	0.011	0.004
English	0.34	0.032	0.030	0.021
Arabic	0.35	0.040	0.022	0.020
Spanish	0.36	0.040	0.027	0.023

Table 2: The contribution of top-three dimensions to the expected cosine similarity ($I_{Cos}(\mathcal{W})$).

Table 2 summarizes the results. Unlike the monolingual BERT, in which one dimension dominates the cosine similarity, multilingual BERT has no rogue dimensions. Hence, the anisotropic structure of the multilingual space cannot be attributed to certain dimensions.

3.4 Word frequency Bias

It has been shown that frequency plays an important role in the distribution of CWRs. Frequency-similar words make distinct local regions in the embedding space (Gao et al., 2019), with high-frequency and rare words being around the center and far from the origin, respectively (Li et al., 2020). Frequency-based distribution is a factor that hampers the expressiveness of the embedding space. So, it is essential to investigate frequency bias in the multilingual embedding space.

Figure 1 shows the distribution of word representations per word frequency.[1] As can be observed on the left, multilingual CWRs are biased toward their frequency, where words with similar frequencies

[1]We used the wordfreq library (https://pypi.org/project/wordfreq/). See Appendix C.

	Ar-Ar	Ar-En	Es-Es	Es-En	Es-En-WMT	En-En
Baseline	51.76 *(8E-5)*	10.61 *(1E-4)*	64.15 *(3E-5)*	31.26 *(5E-4)*	11.39 *(1E-4)*	60.82 *(2E-6)*
Individual	64.26 *(0.60)*	23.10 *(0.57)*	70.88 *(0.54)*	46.23 *(0.50)*	13.47 *(0.50)*	71.99 *(0.54)*
Zero-shot	52.76 *(6E-5)*	19.36 *(0.04)*	65.69 *(8E-4)*	43.82 *(0.09)*	13.68 *(8E-3)*	-

Table 3: STS performance (Spearman correlation percentage) on multi- and cross-lingual datasets using mBERT. Isotropy is reported based on $I_{PC}(\mathcal{W})$ in parentheses. Applying the cluster-based method can improve the performance on the multi- and cross-lingual datasets in both Individual and Zero-shot settings.

create clustered regions. A similar pattern can be observed for the English BERT CWRs (Rajaee and Pilehvar, 2021), with the only difference that in mBERT, low-frequency words are distributed near the origin and frequent words are far from it.

4 Isotropy Enhancement

Making the embedding space isotropic has theoretical and empirical benefits (Gao et al., 2019). In this section, we investigate the effect of isotropy enhancement for the multilingual embedding space. Several approaches have been proposed to improve isotropy in monolingual CWRs. Some requires a retraining of the model with additional objectives to address the degeneration problem (Gao et al., 2019; Li et al., 2020; Zhang et al., 2020), whereas others are applied as a light post-processing (Mu and Viswanath, 2018). We opted for the cluster-based approach of Rajaee and Pilehvar (2021) which is a recent example from the latter category. The proposed method splits the space into several clusters and discards dominant directions for each cluster. The approach also allows us to investigate the similarity of the clustered structure of the embedding space across different languages under a zero-shot setting. More details on this method can be found in Appendix D.

4.1 Settings

We run our experiments in two different settings.

Individual. In this setting, we perform experiments individually on each language by clustering the corresponding space and applying the isotropy enhancement approach. The goal is to see whether increasing isotropy leads to performance improvement in the multilingual space and how the amount of improvement differs across cross- and multilingual tracks.

Zero-shot. In this scenario, we are interested in evaluating the shared structural properties among languages, specifically, the similarity of the encoded linguistic knowledge in the dominant directions of different languages. To this end, we obtain clusters, their means and dominant directions on the English dataset and leverage these for isotropy enhancement in other languages.

4.2 Results

The reported results in Table 3 show that increasing the isotropy in the multilingual embedding space can enhance the performance in all tracks (multi- and cross-lingual). The improvement could be attributed to the potential of the applied method in adjusting embeddings' distribution based on semantic. The visualization of the embedding space after isotropy enhancement, Figure 1 (right), clearly reveals that the frequency bias is faded after this process. Moreover, the results of the zero-shot setting suggest that the encoded information in dominant directions is similar across the languages because the improvement is compatible with the setting in which the dominant directions are obtained in each track individually.

5 Conclusion

In this paper, we provide comprehensive analyses on the geometry of multilingual embedding space through isotropy. We show that multilingual embedding spaces are highly anisotropic, which limits their semantic expressiveness. Our findings shed light on the relation between anisotropy and outliers and demonstrate that despite its anisotropic distribution, multilingual BERT has no disruptive rouge dimensions. We also investigate the other limitation of multilingual embeddings and show that they have a biased structure towards word frequency, and this distribution is similar across different languages. By applying a cluster-based method to increasing the isotropy, we significantly improve the multilingual CWRs performance and address their frequency bias.

References

Sanjeev Arora, Yuanzhi Li, Yingyu Liang, Tengyu Ma, and Andrej Risteski. 2016. A latent variable model approach to PMI-based word embeddings. *Transactions of the Association for Computational Linguistics*, 4:385–399.

Xingyu Cai, Jiaji Huang, Yuchen Bian, and Kenneth Church. 2021. Isotropy in the contextual embedding space: Clusters and manifolds. In *International Conference on Learning Representations*.

Daniel Cer, Mona Diab, Eneko Agirre, Iñigo Lopez-Gazpio, and Lucia Specia. 2017. SemEval-2017 task 1: Semantic textual similarity multilingual and crosslingual focused evaluation. In *Proceedings of the 11th International Workshop on Semantic Evaluation (SemEval-2017)*, pages 1–14, Vancouver, Canada. Association for Computational Linguistics.

Ethan A. Chi, John Hewitt, and Christopher D. Manning. 2020. Finding universal grammatical relations in multilingual BERT. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 5564–5577, 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.

Kawin Ethayarajh. 2019. How contextual are contextualized word representations? Comparing the geometry of BERT, ELMo, and GPT-2 embeddings. 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 55–65, Hong Kong, China. Association for Computational Linguistics.

Jun Gao, Di He, Xu Tan, Tao Qin, Liwei Wang, and Tieyan Liu. 2019. Representation degeneration problem in training natural language generation models. In *International Conference on Learning Representations*.

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

Olga Kovaleva, Saurabh Kulshreshtha, Anna Rogers, and Anna Rumshisky. 2021. BERT busters: Outlier dimensions that disrupt transformers. In *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*, pages 3392–3405, Online. Association for Computational Linguistics.

Bohan Li, Hao Zhou, Junxian He, Mingxuan Wang, Yiming Yang, and Lei Li. 2020. On the sentence embeddings from pre-trained language models. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 9119–9130, Online. Association for Computational Linguistics.

Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2019. Roberta: A robustly optimized bert pretraining approach. *ArXiv*, abs/1907.11692.

Ziyang Luo, Artur Kulmizev, and Xiaoxi Mao. 2021. Positional artefacts propagate through masked language model embeddings. 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 5312–5327, Online. Association for Computational Linguistics.

Jiaqi Mu and Pramod Viswanath. 2018. All-but-the-top: Simple and effective postprocessing for word representations. In *International Conference on Learning Representations*.

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.

Alec Radford, Jeff Wu, Rewon Child, David Luan, Dario Amodei, and Ilya Sutskever. 2019. Language models are unsupervised multitask learners.

Sara Rajaee and Mohammad Taher Pilehvar. 2021. A cluster-based approach for improving isotropy in contextual embedding space. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 2: Short Papers)*, pages 575–584, Online. Association for Computational Linguistics.

William Timkey and Marten van Schijndel. 2021a. All bark and no bite: Rogue dimensions in transformer language models obscure representational quality. In *EMNLP*.

William Timkey and Marten van Schijndel. 2021b. All bark and no bite: Rogue dimensions in transformer language models obscure representational quality. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 4527–4546, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

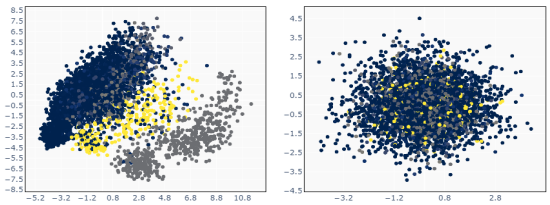
Shijie Wu and Mark Dredze. 2019. Beto, bentz, becas: The surprising cross-lingual effectiveness of BERT. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th*

International Joint Conference on Natural Language Processing (EMNLP-IJCNLP), pages 833–844, Hong Kong, China. Association for Computational Linguistics.

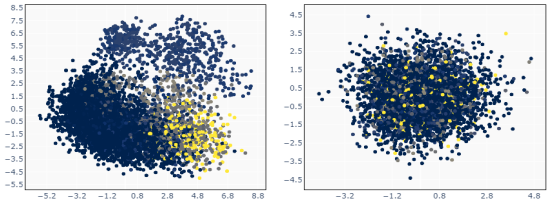
Zhong Zhang, Chongming Gao, Cong Xu, Rui Miao, Qinli Yang, and Junming Shao. 2020. Revisiting representation degeneration problem in language modeling. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 518–527, Online. Association for Computational Linguistics.

A Frequency-based Distribution

Frequency-based distribution can negatively affect the expressiveness of space. Though it is a well-known bias in pre-trained LMs (e.g., BERT and GPT-2), it is not studied in a multilingual setting. As discussed in Section 3.4, we have studied frequency bias in mBERT and demonstrated that mBERT suffers frequency-based distribution in its space like pre-trained counterparts. The illustration of this bias and the impact of the cluster-based approach on mitigating it can be found in Figure 3.



(a) English



(b) Spanish

Figure 3: Degenerated (left) and isotropic (right) embedding spaces for the two languages. Frequency-based distribution can be easily detected using two top PCs in the space (lighter colors indicate higher frequency). Eliminating top dominant directions not only makes the embedding space isotropic, but also removes frequency bias in multilingual CWRs.

B Multilingual STS Task

Multi and cross-lingual Semantic Textual Similarity (STS) is the main task in our experiments. STS is a paired sentence task in which samples have been labeled by a score in the continuous range of 0 (irrelevant) to 5 (most semantic similarity). In the multilingual tracks, in a pair, both sentences are in the same language, while sentences have different languages in the cross-lingual tracks. The reason behind choosing STS as the target task for our experiments is that Multilingual BERT has a pretty low performance on it.

In our experiments, we take the average of all tokens in a sentence as the sentence representation and consider the cosine similarity of the sentence representations in a sample as the semantic similarity score.

C Wordfreq

We have employed Wordfreq library to investigate word frequency bias in out experiments. This library obtains word frequency from the corpus containing eight different domains in 36 languages. Our target languages are in the *large* category which means their word lists cover rare words appearing at least once per 100 million words. As a result, the wordfreq could be a suitable tool for our purpose.

D Cluster-based Isotropy Enhancement

We pick the cluster-based approach (Rajaee and Pilehvar, 2021) to improve the isotropy in multilingual embedding space. In this method, the embeddings are clustered using the k-means clustering algorithm, and then dominant directions of every cluster are nulled out independently. Dominant directions have been calculated employing Principal Component Analysis (PCA). The primary key in this method is obtaining dominant principal components (PCs) of clustered areas in the embedding space separately, which makes this approach suitable for exploring the clustered structure of the multilingual CWRs.

We apply the cluster-based approach to multi and cross-lingual CWRs with two different settings, Individual and Zero-shot. The number of clusters and discarded dominant directions are chosen 7 and 12, respectively.