Quantifying Interlinguality In Foundation Models

Anonymous ACL submission

Abstract

Large multilingual language models show remarkable zero-shot cross-lingual transfer performance on a range of tasks. Follow-up works hypothesized that these models internally project representations of different languages into a shared interlingual space. However, they produced contradictory results. In this paper, we correct the prior work claiming that "BERT is not an Interlingua" and show that with the proper choice of sentence representation different languages actually do converge to a shared space in such language models. Lastly, we apply our own advice and quantify interlinguality across six pretrained models and 378 language pairs. 1

1 Introduction

Large-scale multilingual language models (LMs) such as mBERT (Devlin et al., 2019) and XLMR (Conneau et al., 2020a) achieve remarkable results on a variety of cross-lingual transfer tasks (Hu et al., 2020b; Liang et al., 2020). Follow-up works performed representational similarity analysis comparing encoded sentences in different languages (Singh et al., 2019; Muller et al., 2021; Conneau et al., 2020b). However, they come up with two opposite conclusions.

In particular, Singh et al. (2019) concluded that "mBERT is not an Interlingua". They observed that the model separates representations for each language rather than using a common, shared, interlingual space while increasing divergence at deeper layers. Muller et al. (2021), on the contrary, found that approximately the first half of the mBERT layers align different languages together. See Figure 1 for our reproduction of this conflicting results.

This begs the question of which one should we rely on and why. Muller et al. (2021) backs up their representational analysis using probing task in the layer-wise ablation setting; (Conneau et al., 2020b) also directly supports this conclusion. Singh et al. (2019), on the other hand, provided explanation based on possible tokenization bias issue. In any case, we believe that cross-lingual representational analysis should provide unambiguous results, so in this work we investigate this issue in depth.

Specifically, we set the following research questions:

• What do results in literature diverge?
• How do we interpret conflicting results?
• How should we quantify interlinguality in multilingual language models in the future?
• Does the resulting interlingual pattern generalizes across different multilingual language models?
• Does the resulting interlingual pattern generalizes across large number of languages?

We find that using PWCCA with CLS-pooling is the only metric-pooling combination that yields divergence patter, and does so for a particular reason. We then advice usage of mean-pooled representations with CKA/SVCAA algorithm when doing

1The code for replicating our results will be publicly available

Figure 1: mBERT is not an "interlingua" (left) vs mBERT is an "interlingua" (right). Reproduced from Singh et al. (2019) and Muller et al. (2021).
cross-lingual measurements. Lastly, we apply our own advice and quantify interlinguality across six pretrained models and 378 language pairs showing robustness of the interlingual pattern.

2 Resolving Conflicting Literature

2.1 Setup and Background

2.1.1 Data and model

We do a setup similar to the one of Singh et al. (2019) and use the mBERT-based model and four parallel datasets (en-fr, en-de, en-ru, en-et; 10k examples for each pair). The parallel corpus is composed from Singh et al. (2019)’s extension of XNLI dataset (Conneau et al., 2018).

We embed the source and target sentences with mBERT and pool the CLS tokens or perform a mean-pooling over tokens from each layer for each language pair. Next, we compare two parallel sets of sentence representations using the PWCCA (Morcos et al., 2018), (Kornblith et al., 2019) or SVCCA (Raghu et al., 2017) similarity measure.

2.1.2 Representational Similarity Algorithms

In this subsection we give a reader brief intuitive explanations and refer to the Kornblith et al. (2019) for a systematic mathematical description of the correlation similarity measures.

In general, all metrics compare two parallel sets of vectors by maximizing correlations (score 1 represents perfect similarity).

CCA is the correlation based similarity analysis method. As formulated by Morcos et al., CCA “identifies the ’best’ (maximizing correlation) linear relationships (under mutual orthogonality and norm constraints) between two sets of multidimensional variates”. CCA score can be a mean of the resulting correlation coefficients.

SVCCA reduces sensitivity of CCA to particular dimensions by performing Support Vector Decomposition on parallel vectors first, and then applying CCA on the resulting components. The number of resulting CCA coefficients is the hyperparameter for the SVCCA, we use 20 in this work.

CKA is another similarity measure which works by computing pairwise dot products between two parallel sets of vectors and correlating resulting distance matrices.

PWCCA is an extension of the original CCA (Hardoon et al., 2004) that weights resulting CCA correlation coefficients based on their importance, instead of taking simple mean.

Also, we highlight that PWCCA is only invariant to the translation and isotropic scaling. CKA is also invariant to the orthogonal transforms, and SVCCA is invariant to any invertible linear transform. Another important distinction of PWCCA is that representations are not directly centered at the first step, unlike in CKA and SVCCA.

2.2 Identifying the Issue

In this section we aim to find what caused the discrepancy between results in related works. They use different data and code, but we successfully reproduced the results with our dataset and code. However, they also differ in their choice of sentence representation (CLS-pooling for Singh et al. (2019) and mean=pooling for Muller et al. (2021)) and similarity measuring algorithm (PWCCA for Singh et al. (2019) and CKA for Muller et al. (2021)).

2.2.1 Is this a similarity algorithm issue?

To answer the question of this subsection, we compute all similarity between CLS-pooled representations for all of three main algorithms. Figure 2 presents the results. It shows that as we change similarity measure from PWCCA, we get convergence pattern similar to one in (Muller et al., 2021).

This is good insight, but works also differ in sentence representation type, lets try to see what is hidden there in the next subsection.

2.2.2 Is this a representation type issue?

To answer the question of this subsection, we compute all similarity measures for representations, but this time obtained with averaging individual token representations (mean-pooling). Figure 3 presents the results. It shows that as we change pooling type from CLS, we get rid of divergence pattern Singh et al. (2019) had with CLS-pooling.

PWCCA pattern convergence pattern is less pronounced then with other similarities, but does not contradict (Muller et al., 2021).

At the end, the divergence pattern only occurs when we use CLS-pooling and PWCCA metric at the same time, so in the next subsections we aim to find out what is so special about this particular combination.
2.3 Debunking Divergence Pattern

So we only get divergence pattern when we use PWCCA measure over CLS-pooled representations. In the next subsection we analyse what is so specific about PWCCA and CLS in our setting.

2.3.1 Debunking PWCCA

Two crucial differences between PWCCA and SVCCA/CKA is that are that PWCCA is not invariant to orthogonal transformation and does not explicitly center representations as the first step. This suggest we look at the directions of the word vectors.

We run a simple experiment with cosine similarity and CLS-pooling, mean-pooling, and pooling the 1st token (next after CLS) to get a glance about directional nature of contextual vector space. Specifically, we compute average cosine similarity between English sentences and their target (parallel) translations. At the same time, we compute average cosine similarity between each pairs of vectors in the vector space. Results are in Figure 4

We can see that for the first five layers in CLS-pooling, cosine similarities between parallel sentences are close to one, just as cosine for random sentences. So they all point about at the same direction, and there is no straightforward distinction between translations and not translations.

Because centering is not directly performed for PWCCA and the because the algorithm is not invariant to rotations (Kornblith et al., 2019), it probably gets biased by these similarities and gives high scores.

However, this behaviour at the levels 0-4 also suggests we should verify the very validity of CLS as an underlying sentence representation.

2.3.2 Debunking CLS

Our goal is to find out how "interlingual" mBERT is across layers. "Interlinguality" is not a well defined concept. However, a its key property is that cross-lingual representations in multiple languages should have close representations. Moreover, these representations should be far closer to representations of other non-parallel sentences.

So let us see how well CLS satisfies these properties. To keep neutral regarding similarity measure, we step away from direct representational similarity analysis and setup a probing task which measures desired properties. We use the same data and for
Figure 4: Cosine similarities of sentence representations under three different pooling strategies. "Target" box in legend means that (average) cosine similarity was measured between parallel sentences ("parallel") or arbitrary pairs of sentences ("random").

Figure 5: Accuracy of closest sentence vector to each source sentence being actual translation of this sentence. Measured over three pooling strategies and averaged over four languages (as before).

Each English sentence we find the closest target sentence in opposite language (out of all 10k targets) and declare "1" if the closest sentence was actual translation of source. We declare "0" otherwise. Then we compute accuracy of this matching task for our language pairs.

We also repeat this experiment where we pool the 1st token from each sentence to get a reference point for comparison. We present (averaged over languages) results in Figure 5.

Figure shows that CLS-pooling scores almost zero accuracy at layers 0-4, which suggests it is not useful representation to rely upon when measuring interlinguality, including using CCA-like measures. Matching by first token is better then matching by CLS at these layers. In later layers CLS gets higher then first token pooling, but mean-pooling is still about 0.3 accuracy points above.

However, Figure 5 also reveals that other representational similarity metrics match the figure’s pattern, which is empirical evidence for their usefulness. PWCCA is the only exception assumably for the reasons we discussed before.

Mean-pooling or CLS-pooling: CLS-pooling is used as a sentence representation because at the last layer mBERT uses it to predict the next sentence in the corpus during pretraining.

At the last layer, next sentence prediction task directly uses CLS token. However, CLS at other hidden layers does not have any clear signal to meaningfully represent the sentence due to the token mixing procedure in the Transformer layers. Each token position carries pieces of information about itself as well pieces about other tokens, and CLS is just a regular token similarly to others at these layers. Consequently, representing sentences with CLS for hidden layers is conceptually suboptimal. Mean pooling, on the contrary, gathers distributed information about all tokens at each layer and thus free of abovementioned flaw.

In this section we showed that using CLS-pooling and PWCCA measure are suboptimal for interlinguality representational analysis, and when used together they result in the pattern opposite to the expedient one.

3 Application

So as we identified that that PWCCA with CLS-pooling is not very suited for cross-lingual analysis and advised to use mean-pooling with SVCCA/CKA. In this section we will act on our advice and perform interlingual measurements using CKA algorithm over mean-pooled representations across six pretrained models and 378 language pairs.
3.1 Quantifying Interlinguality Across Models

3.1.1 Setup

Representational analysis in the previous work (Singh et al., 2019; Muller et al., 2021) was limited to the specific multilingual BERT model. It questions whether the pattern generalizes across other models with different training objectives, pre-training datasets, capacities, and tokenization. In this section, we repeat our analysis for five more commonly used models to address this issue.

We use six different models: uncased (uncased mBERT) and cased (uncased mBERT) versions of mBERT; next, we use XLM-MLM-100 (Lample and Conneau, 2019) which has 16 layers and was trained on single sentences. XLM-R base (Conneau et al., 2020a) is the multilingual extension of the Roberta (Liu et al., 2019) trained for longer on larger data and is “wider” than mBERT. XLM-R large uses about 3 times more parameters than XLM-R base. distil_mBERT (Sanh et al., 2019) is a twice smaller distilled version of mBERT. For all models we use the HuggingFace library. We also refer the reader to Appendix B in Conneau et al. (2020a) for a systematic comparison of models.

3.1.2 Results and Discussion

Figure 6 shows CKA similarity results for all models. We follow (Singh et al., 2019) in our choice of language pairs for this subsection.

All multilingual encoders are generally consistent in following the interlingua pattern. If we consider downstream task performance of the models (Conneau et al., 2020a; Hu et al., 2020a) we see that more performant models are generally placed higher in the graph. XLM-Rs is on top, and distill-BERT is at the bottom, while XLM-MLM and mBERT are in the middle. Interestingly, the XLM-R base is a clear leader in cross-lingual sentence similarities and outscores even the XLM-R large model in that regard, even though the large version is more cross-lingually performant. The explanation might be related to the fact that XLM-R large has ‘useless’ but cross-lingually distant dimensions

3.2 Quantifying Interlinguality Across Languages

3.2.1 Setup

This section takes a more in-depth look at interlingua in multilingual LMs and explores its structure.

We use all 378 language directions from the extended XNLI dataset Singh et al. (2019). By computing CKA across all language pairs at all layers, we determine that the 7th layer of XLM-R is the most “interlingual”. We choose to focus on XLM-R in addition to the mBERT for its performance in Section 3.1. All figures of XLM-R confirm the general conclusions made for the mBERT.

Figure 8 also highlights Thai as the most distinct language. This language was not included in the mBERT uncased pretraining. mBERT has more “Excluded” languages, which is logical since it is weaker cross-lingual model, comparing to the XLM-R.

Figure 6: Generalization over models across languages under CKA metric. Models follow similar general trend across languages which corresponds to the interlingual pattern we discuss in this work.

that the metric captured. However, we leave exploring the relation between representational similarity across models and their cross-lingual benchmark performance to future work.

In summary, here we showed that the general interlingual pattern generalizes across models.
3.2.2 Results and Discussion

Figure 7 shows the boxplot covering CKA distances between all pairs of languages for XLM-R and Figure 8 for mBERT.

Dots at the bottom suggest that there are clear outlier language pairs. The most interlingual layer for XLM-R seems to be 7th, so let us open the box at this layer with the following Figures 9, 10.

We can now clearly see that the outliers are due to Urdu and Hindi (and possibly Swahili and Thai). However, even within the shared interlingual space languages also exhibit certain relationships with each other. Thus we perform an agglomerative clustering on CKA distances to investigate this phenomenon and present the result in Figures 11, 12:
The linguistic tree in Figures 11, 12 clearly shows that languages in the right branch (that we consider to be the shared interlingual space) structure in a meaningful way. Slavic languages are together, Swahili and Thai are isolated, while Scandinavian languages are again nearby. Urdu and Hindi expectedly occupy their separate branch as outliers.

Finally, we note that CKA metric we used is invariant to translations, scaling, and orthogonal transforms of representational spaces. Despite this invariance, the tree in Figures 11, 12 that was built based on CKA distances between languages still shows the familiar linguistic hierarchy. This pushes our understanding of language representations beyond them simply residing on centroid distances without transforming the interlingual space, as Libovický et al. (2020) showed. Instead, it is clear that even on a more complex level, after pairwise “aligning” sentence representations in different languages the linguistic hierarchy pattern still emerges.

4 Conclusion

In this paper we identified, analysed, and resolved conflicting literature and derived that mean-pooling with SVCCA/CKA similarity measure is the most suitable choice for the interlingual representational similarity analysis. Next, we showed that the pattern is not specific to mBERT and is present in other multilingual language models. Finally, we analysed 378 language pairs and found that not all of the languages share the interlingual space equally.

References


Junjie Hu, Sebastian Ruder, Aditya Siddhant, Graham Neubig, Orhan Firat, and Melvin Johnson.


