BERT is Not an Interlingua and the Bias of Tokenization

Jasdeep Singh[1], Bryan McCann[2], Caiming Xiong[2], Richard Socher[2]
Stanford University[1], Salesforce Research[2]
jasdeep@cs.stanford.edu {bmccann,cxiong,rsocher}@salesforce.com

Abstract

Multilingual transfer learning can benefit both high- and low-resource languages, but the source of these improvements is not well understood. Cananical Correlation Analysis (CCA) of the internal representations of a pretrained, multilingual BERT model reveals that the model partitions representations for each language rather than using a common, shared, interlingual space. This effect is magnified at deeper layers, suggesting that the model does not progressively abstract semantic content while disregarding languages. Hierarchical clustering based on the CCA similarity scores between languages reveals a tree structure that mirrors the phylogenetic trees hand-designed by linguists. The subword tokenization employed by BERT provides a stronger bias towards such structure than character- and word-level tokenizations. We release a subset of the XNLI dataset translated into an additional 14 languages at https://www.github.com/salesforce/xnli_extension to assist further research into multilingual representations.

1 Introduction

Natural language processing (NLP) in multilingual settings often relies on transfer learning between high- and low-resource languages. Word embeddings trained with the Word2Vec (Mikolov et al., 2013b) or GloVe (Pennington et al., 2014) algorithms are trained with large amounts of unsupervised data and transferred to downstream tasks-specific architectures in order to improve performance. Multilingual word embeddings have been trained with varying levels of supervision. Parallel corpora can be leveraged when data is available (Gouws et al., 2015; Luong et al., 2015), monolingual embeddings can be learned separately (Klementiev et al., 2012; Zou et al., 2013; Hermann and Blunsom, 2014) and then aligned using dictionaries between languages (Mikolov et al., 2013a; Faruqui and Dyer, 2014), and cross-lingual embeddings can be learned jointly through entirely unsupervised methods (Conneau et al., 2017; Artetxe et al., 2018).

Contextualized word embeddings like CoVe, ElMo, and BERT (McCann et al., 2017; Peters et al., 2018; Devlin et al., 2018) improve a wide variety of natural language tasks (Wang et al., 2018; Rajpurkar et al., 2016; Socher et al., 2013; Conneau et al., 2018). A *multilingual* version of BERT trained on over 100 languages achieved state-of-the-art performance across a wide range of languages as well. Performance for low-resource languages has been further improved by additionally leveraging parallel data (Lample and Conneau, 2019) and leveraging machine translation systems for cross-lingual regularization (Singh et al., 2019).

Prior work in zero-shot machine translation has investigated the extent to which multilingual neural machine translation systems trained with a shared subword vocabulary Johnson et al. (2017); Kudugunta et al. (2019) learn a form of interlingua, a common representational space for semantically similar text across languages. We aim to extend this study to language models pretrained with multilingual data in order to investigate the extent to which the resulting contextualized word embeddings represent an interlingua.

Canonical correlation analysis (CCA) is a classical tool from multivariate statistics (Hotelling, 1992) that investigates the relationships between two sets of random variables. Singular value and projection weighted variants of CCA allow for analysis of representations of the same data points from different models in a way that is invariant to affine transformations (Raghu et al., 2017; Morcos et al., 2018), which makes them particularly suitable for analyzing neural networks. They have

Proceedings of the 2nd Workshop on Deep Learning Approaches for Low-Resource NLP (DeepLo), pages 47–55
Hong Kong, China, November 3, 2019.
https://doi.org/10.18653/v1/P17

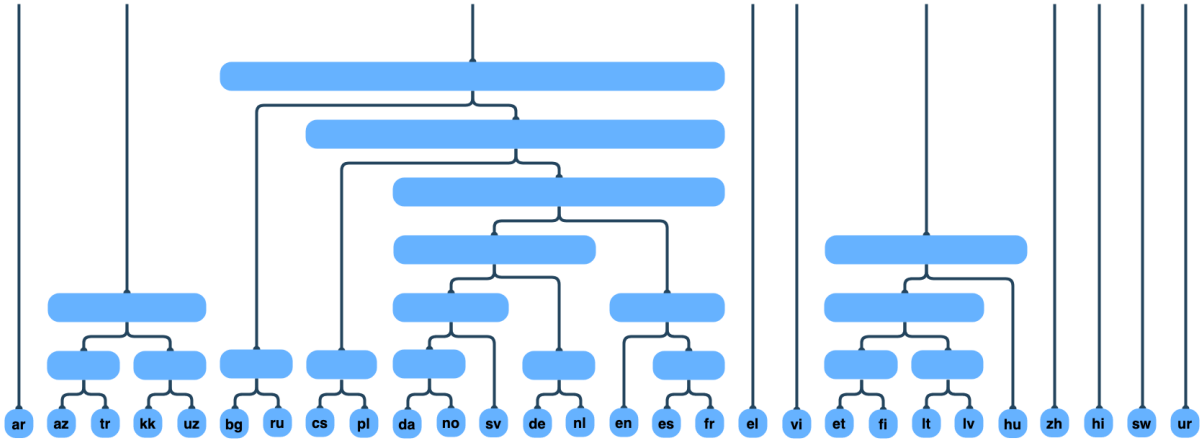


Figure 1: Agglomerative clustering of Languages based on the PWCCA similarity between their represenations, generated from layer 6 of a pretrained multilingual uncased BERT.

been used to explore learning dynamics and representational similarity in computer vision (Morcos et al., 2018) and natural language processing (Saphra and Lopez, 2018; Kudugunta et al., 2019).

We analyze multilingual BERT using projection weighted canonical correlation analysis (PWCCA) between representations from semantically similar text sequences in mulitple languages. We find that the representations from multilingual BERT can be partitioned using PWCCA similarity scores to reflect the linguistic and evolutionary relationships between languages. This suggests that BERT does not represent semantically similar data points nearer to each other in a common space as would be expected of an interlingua. Rather, representations in this space are primarily organized around features that respect the natural differences and similarities between languages. Our analysis shows that the choice of tokenization can heavily influence this space. Subword tokenization, in contrast to word and character level tokenization, provides a strong bias towards discovering these linguistic and evolutionary relationships between languages. As part of our experiments, we translated a subset of the XNLI data set into an additional 14 languages, which we publicly release to assist further research into multilingual representations.

2 Background and Related Work

2.1 Natural Language Inference and XNLI

The Multi-Genre Natural Language Inference (MultiNLI) corpus (Williams et al., 2017) uses data from ten distinct genres of English language for the the task of natural language inference (prediction of whether the relationship between two sentences represents entailment, contradiction, or neither). XNLI (Conneau et al., 2018) is an evaluation set grounded in MultiNLI for cross-lingual understanding (XLU) in 15 different languages that include low-resource languages such as Swahili and Urdu. XNLI serves as the primary testbed for bench marking multilingual understanding. We extend a subset of XNLI to an additional 14 languages for our analysis.

2.2 CCA

Deep network analyses techniques focusing on the weights of a network are unable to distinguish between several invariances such as permutation and scaling. CCA (Hotelling, 1992) and variants that use Singular Value Decomposition (SVD) (Raghu et al., 2017) or projection weighting (Morcos et al., 2018) are apt for analyzing the activations of neural networks because they provide a similarity metric that is invariant to permutations and scaling of neurons. These methods also allow for comparisons between representations for the same data points from different neural networks where there is no naive alignment from neurons of one network to another.

Given a dataset $X = \{x_1, \ldots x_n\}$, let $L_1 \in \mathbb{R}^{m_1 \times n}$ and $L_2 \in \mathbb{R}^{m_2 \times n}$ be two sets of neurons. Often these sets correspond to layers in neural networks. CCA transforms L_1 and L_2 to $a_1^\top L_1$ and $b_1^\top L_2$ respectively where the pair of canonical variables $\{a_1, b_1\}$ is found by maximizing the correlation between the transformed subspaces:

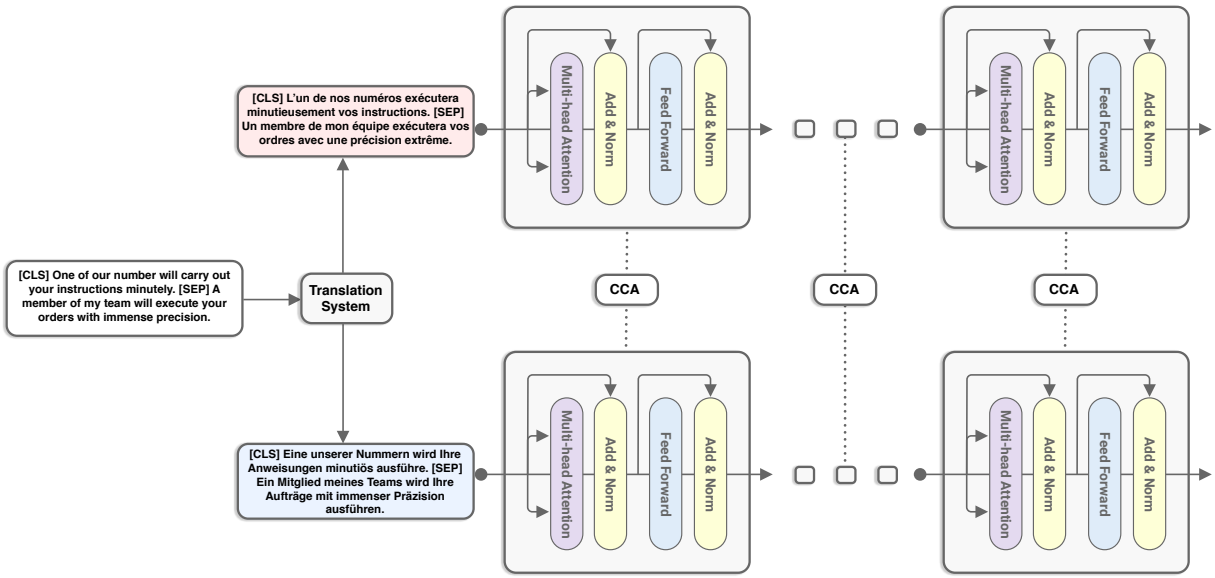


Figure 2: How CCA is used to compare the representations of different languages at different layers in BERT.

$$\rho_1 = \max_{a \in \mathbb{R}^{m_1}, b \in \mathbb{R}^{m_2}} \frac{\langle a^\top L_1, b^\top L_2 \rangle}{\|a^\top L_1\| \cdot \|b^\top L_1\|}$$

Given the set $\{a_1, b_1\}$, we can find another pair $\{a_2, b_2\}$:

$$\rho_2 = \max_{a_2 \in \mathbb{R}^{m_1}, b_2 \in \mathbb{R}^{m_2}} \frac{\langle a_2^\top L_1, b_2^\top L_2 \rangle}{\|a_2^\top L_1\| \cdot \|b_2^\top L_1\|}$$

under the constraints that $\langle a_2, a_1 \rangle = 0$ and $\langle b_2, b_1 \rangle = 0$. This continues until $\{a_{m'}, b_{m'}\}$ and $\rho_{m'}$ have been found such that $m' = \min(m_1, m_2)$.

The average of $\{\rho_1, \ldots \rho_{m'}\}$ is often used as an overall similarity measure, as in related work exploring multilingual representations in neural machine translation systems (Kudugunta et al., 2019) and language models (Saphra and Lopez, 2018). Morcos et al. (2018) show that in studying recurrent and convolutional networks, replacing a weighted average leads to a more robust measure of similarity between two sets of activations. For the rest of the paper we use this PWCCA measure to determine the similarity of two sets of activations. The measure lies between [1,0] with 1 being identical and 0 being no similarity between the representations.

CCA is typically employed to compare representations for the same inputs for different models or layers (Raghu et al., 2017; Morcos et al., 2018). We also use CCA to compare representations from the same neural network when fed two translated versions of the same input (Figure 2).

3 Experiments and Discussion

We use the uncased multilingual BERT model (Devlin et al., 2018) and the XNLI data set for most experiments. Multilingual BERT is pretrained on the Wikipedia articles form 102 languages and is 12-layers deep. The XNLI dataset consists of the MultiNLI dataset translated into 15 languages. The uncased multilingual BERT model does not contain tokenization or pretaining for Thai so we focus our analysis on the remaining 14 languages. To provide a more robust study of a broader variety of languages, we supplement the XNLI data set by further translating the first 15 thousand examples using google translate. This allows for analysis of representations for 14 additional languages including Azerbaijani, Czech, Danish, Estonian, Finish, Hungarian, Kazakh, Latvian, Lithuanian, Dutch, Norwegian, Swedish, Ukrainian, and Uzbek.

Following the standard approach to using BERT (Devlin et al., 2018), a [CLS] token is prepended to each example, which consists of a premise, a [SEP] token and a hypothesis. The [CLS] token has been pretrained to extract intersentence relationships between the sentences that follow it. When fine-tuned on XNLI, the final representations for the [CLS] token is used to predict the relationship between sentences as either entailment, contradiction or neutral. This [CLS] token

can be thought of as a summary embedding for the input as a whole. We analyze the activations of this [CLS] token for the same XNLI examples across all available languages (Figure 2). These representations have 768 neurons computed over 15 thousand datapoints.

For all fine-tuning experiments we use the hyperparameters and optimization strategies recommended by (Devlin et al., 2018) unless otherwise specified. We use a learning rate warm-up for 10% of training iterations and then linearly decay to zero. The batch size is 32 and the target learning rate after warm-up is $2e-5$.

3.1 Representations across languages are less similar in the deeper layers of BERT

Figure 3 demonstrates that for all language combinations tested, the summary representation (associated with the [CLS] token) for semantically similar inputs translated into multiple languages is most similar at the shallower layers of BERT, close to the initial embeddings. The representations steadily become more dissimilar in deeper layers until the final layer. The jump in similarity in the final layer can be explained by the common classification layer that contains only three classes. In order to finally choose an output class, the network must project towards one of the three embeddings associated with those classes (Liu et al., 2019).

The trend towards dissimilarity in deeper layers suggests that contextualization in BERT is not a process of semantic abstraction as would be expected of an interlingua. Though semantic features common to the multiple translations of the input might also be extracted, the similarity between representations is dominated by features that differentiate them. BERT appears to preserve and refine features that separate the inputs, which we speculate are more closely related to syntactic and grammatical features of the input.

Representations at the shallower layers, closer to the subword embeddings, exhibit the highest degree of similarity. This provides further evidence for how a subword vocabulary can effectively span a large space of languages.

3.2 Representations diverge with depth after fine-tuning

In the previous set of experiments, BERT had only been pretrained on the unupservised masked language modeling objective. In that setting the

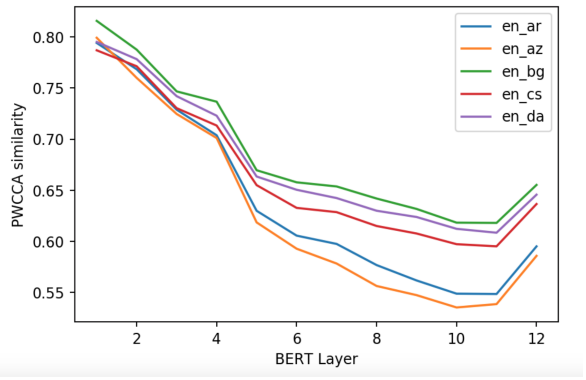


Figure 3: The similarity between representations of different languages decreases deeper into a pretrained uncased multilingual BERT model. Here we show the similarity between English and 5 other languages as a function of model depth

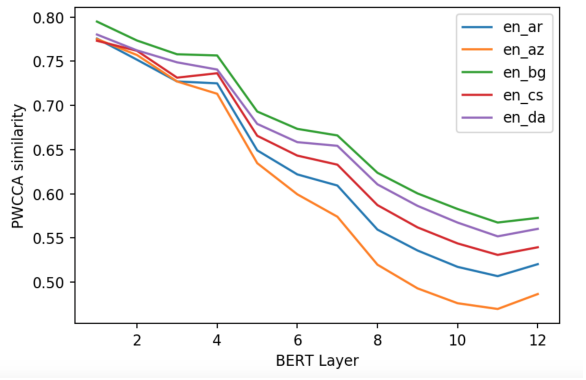


Figure 4: The similarity between representations of different languages decreases deeper into an uncased multilingual BERT model finetuned on XNLI.

[CLS] token was trained to predict whether the second sentence followed the first. This does not align well with the XNLI task in cases in which the hypothesis would not likely follow the premise in the corpora used for pretraining. To alleviate concerns that this might influence the representational similarity, we repeat the above experiments after fine-tuning BERT on several languages. Figure 4 confirms that representations for semantically similar inputs in different languages diverge in PWCCA similarity in deeper layers of BERT.

3.3 Deeper layers change more dramatically during fine-tuning

We also notice that during fine-tuning, the deepest layers of BERT change the most according to PWCCA similarity . In these experiments, we use PWCCA in the more standard setting, in which identical inputs are provided to two different models in order to get two set of neuron activations.

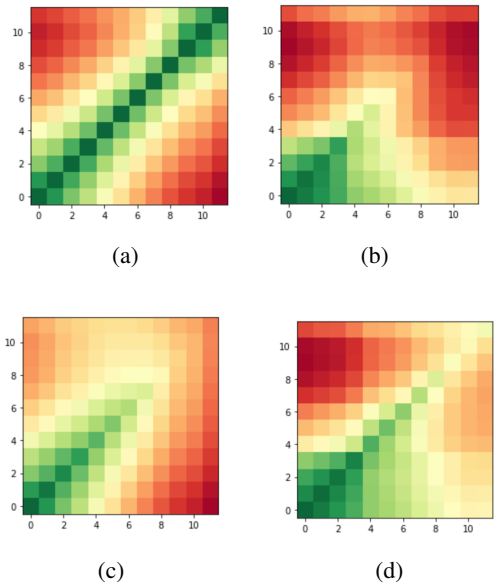


Figure 5: CCA experiments showing the finetuning behavior of BERT

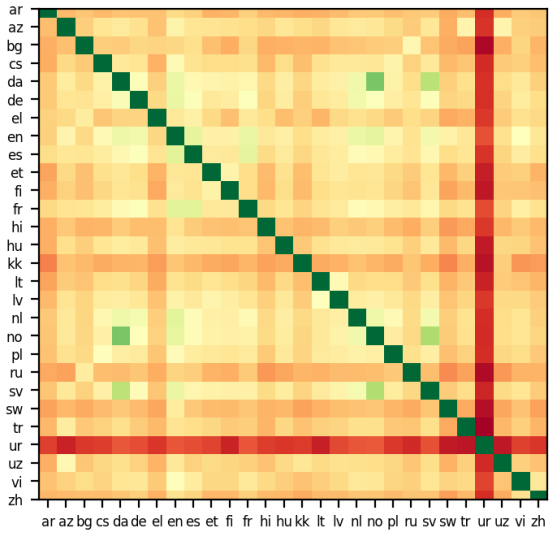


Figure 6: PWCCA generated similarity matrix between languages.

The two different models are different checkpoints of pretrained, multilingual BERT over the course of fine-tuning. Figures 5a - 5d follow a similar structure in which the PWCCA value is computed between all the layers of two models. The matrices are symmetric in expectation, but noise during optimization creates slight asymmetry. Figure 5a shows a baseline case demonstrating that a pretrained, multilingual BERT compared with itself has a PWCCA value of 1 with strong diagonal showing the identity between each layer. The off-diagonal entries show that layer-wise similarity depends on relative depth. This successive and gradual changing of representations is precisely the behavior we expect from networks with residual connections (Raghu et al., 2017).

We compare the pretrained multilingual BERT to a converged BERT fine-tuned on XNLI in Figure 5b and 5c. We use representations for the [CL] token in Figure 5b and the first token in the premise for Figure 5c. Both confirm that the function of early layers remains more similar as the network is fine-tuned. We find this trend to hold for a wide range of tasks including SST and QNLI as well.

Figure 5d shows that deep pretrained networks also converge bottom up during fine-tuning by comparing the representations a quarter of the way through fine-tuning with those of a converged model. Most of the remaining change in the representations between a quarter of the way through training and convergence happens in the later layers. Therefore the changes to middle layers we observe in Figure 5b happen during the first quarter of training.

3.4 Phylogentic Tree of Language Evolution

Figures 3 and 4 show that the representations learned for different languages diverge as we go deeper into the network, as opposed to converging if the network were learning an interlingua or a shared representation space. However, the relative similarities between languages clearly varies for different pairs and changes as a function of depth. To further investigate the internal relationships between representations learned by BERT we create a similarity matrix using PWCCA between all 28 languages for all 12 layers in BERT. For the PWCCA calculations we use the representations of the [CLS] token to generate L_1 and L_2. The resulting similarity matrix for Layer 1 is shown in Figure 6.

To visualize these relationships this matrix can be converted to a phylogentic tree using a cluster algorithm. In Figure 1 we use unweighted pair group method with arithmetic mean (UPGMA), a simple agglomerative (bottom-up) hierarchical clustering method (Sokal, 1958) to generate a phylogentic tree from the representations from Layer 6 of BERT. The generated phylogentic tree closely resembles the language tree constructed by linguists to explain the relationships and evolution of human languages. The details of the linguis-

tic evolutionary phylogentic tree of languages is still debated and a tree model faces some limitations as not all evolutionary relationships are completely hierarchical and it can not easily account for horizontal transmissions. However many of the commonly known relationships between languages are embedded in BERT's representations. We see that BERT's space of its internal representations is finely partitioned into families and subfamiles of languages.

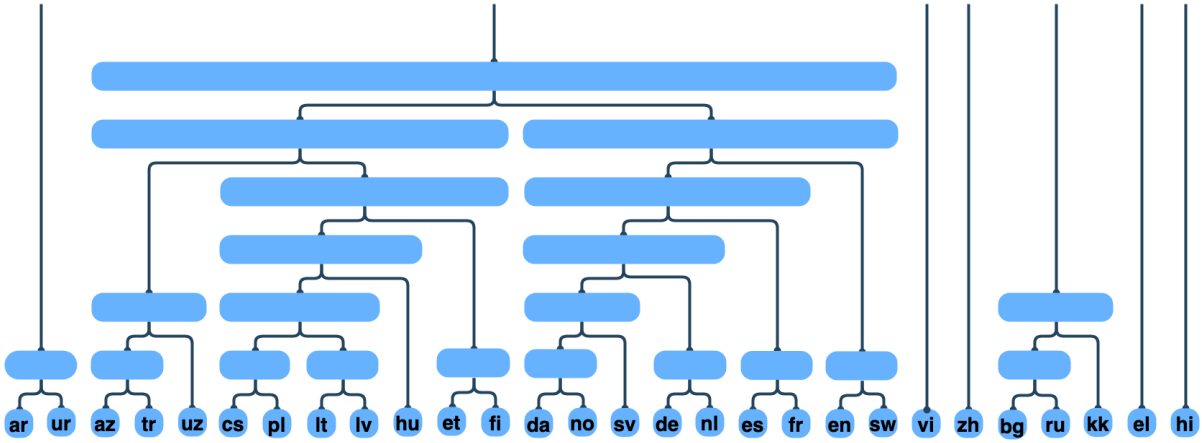
Northern Germanic languages are clustered together and the Western Germanic languages are clustered together before being combined together into the pro-germanic family. Romantic languages are clustered together before the Romantic and Germanic families are combined together. BERT's internal representation of English is in-between that of the Germanic and Romantic sub-families. This captures evolution and structure of English, which is considered a Germanic language but borrows heavily from romantic and latin languages. By varying the layer at which the representations are used to create the phylogentic tree, different structures emerge. Sometimes English is clustered with German before it is combined with the romantic languages, but mostly BERT seems to classify English as a romantic language.

For Layers 6 through 12 the trees generated are almost or exactly like Figure 1. At these layers, Azerbaijani, Turkish, Kazakh, and Uzbek are grouped into the same family although these languages span multiple scripts and have had their official scripts changed multiple times allowing for the possible introduction of confounding differences. Interestingly, at these same layers in Figure 3, languages seem to diverge the most from each other. This seems to suggest that instead of finding a shared latent space for all of the languages, as would be necessary for an interlingua, BERT is actually carefully partitioning its space in a fashion that linguistic and evolutionary relationships are preserved between languages. Trees generated from Layers 1 through 5 seem to make more mistakes than those of later layers. We see that these trees end up failing to group Azerbaijani, Turkish, Kazakh, and Uzbek into the same family, often leaving Kazakh out which is written in Cyrillic script. We can hypothesis that BERT's internal representations are more reliant on the identity of shared subwords earlier in the network as opposed to later in the network. As a matter of fact, if we use agglomerative clustering to construct a tree from a matrix of subword overlap counts (Figure 7a), we find that it almost exactly matches the tree constructed from BERT's earlier layers.

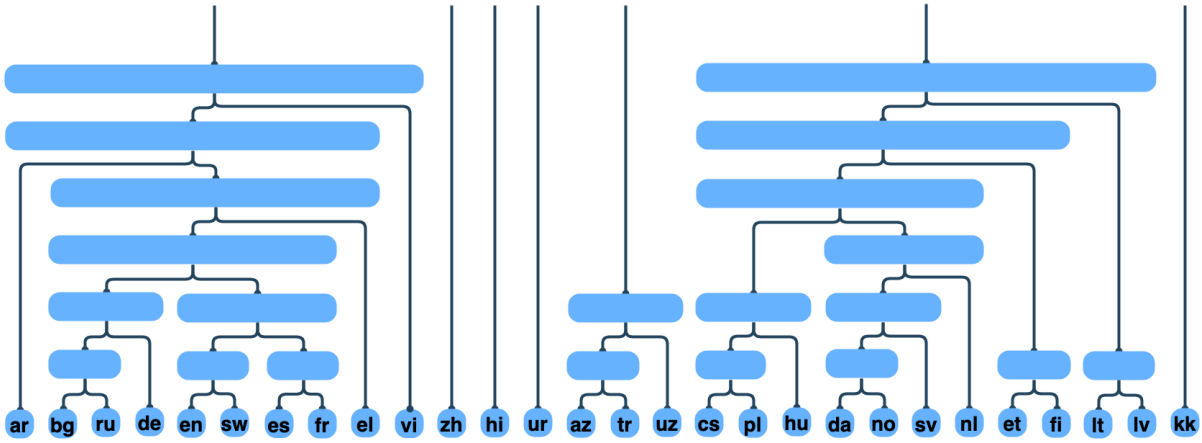
It seems that BERT's shared multilingual subword vocabulary (Mikolov et al., 2012; Sennrich et al., 2015) provides it with a strong bias towards what linguistic relationships exist between human languages. Instead of then fusing the representations of different languages into one shared representation during training, BERT actually successively refines this partitioned space to better reflect the linguistic relationships between languages at higher layers (Figure 1).

3.5 Tokenization Provides a Strong Bias Towards Knowledge of Linguistic Relationships Between Languages

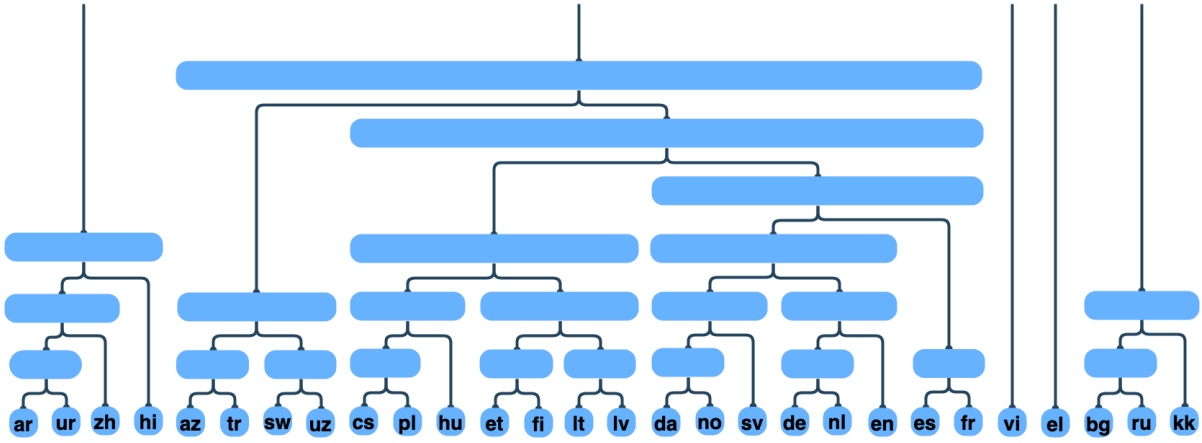
We tokenize the first 15 thousand examples from XNLI and our translated data using different tokenization methods and compute the token overlap between different languages, generating similarity matrices similar to the one shown in Figure 6. From these matrices we perform agglomerative clustering of languages to generate phylogenetic trees (Figure 7). These trees show how different tokenization schemes can embed different linguistic biases into our models. We investigate subword, word, and character level tokenization. The subword tokenization is done using BERT's learned BPE vocabulary. The word level tokenization is achieved by simply tokenizing at spaces, and the character level tokenization is done using Python's native character level string splitting. Figure 7a is generated from using subwords, and although is not as accurate as the tree generated from BERT's representation at layer 6 (Figure 1), it is still non-trivially close to a linguistically accurate depiction of human language evolution. We see that by using a shared subword vocabulary, multilingual BERT has a very strong bias to discover the linguistic relationships between languages. However, this bias is not as strong if other forms of tokenization are used. Figures 7b and 7c show the trees generated by word level and character level tokenization respectively. We see that word level tokenization splits the Romantic and Germanic languages into completely different trees, and that character level tokenization ends up combining all languages that share the Latin script regardless of their true families. Perhaps the



(a) Agglomerative clustering of languages based on subword overlap, generated from using the BERT tokenizer to tokenize the first 15 thousand examples from XNLI and our translated data.



(b) Agglomerative clustering of languages based on word overlap, generated from splitting the first 15 thousand examples from XNLI and our translated data on spaces.



(c) Agglomerative clustering of languages based on character overlap, generated from splitting the first 15 thousand examples from XNLI and our translated.

Figure 7: Different agglomerative clusterings of languages based on subword, word, and character overlap. We see that different tokenization schemes used in NLP embed different linguistic biases into models.

ability of subwords to capture these linguistic relationships between languages has contributed to their wide success in applications including machine translation and language modeling.

4 Conclusion and Future Directions

While natural language processing systems often focus on a single language, multilingual transfer learning has the potential to improve performance, especially for low-resource languages. Many previous multilingual approaches claim to develop shared representations of different languages. Recently, multilingual BERT and related models trained in an unsupervised fashion on monolingual corpora from over 100 languages achieve state of the art performance on many tasks involving low resource languages. Using Cononical Coreelation Analysis (CCA) on the internal representations of BERT, we find that it is not embedding different languages into a shared space. Rather, at deeper layers, BERT partitions the space to better reflect the linguistic and evolutionary relationships between languages. We also find that subword tokenization, in contrast to word and character level tokenization, provides a strong bias to discover linguistic and evolutionary relationships between languages.

References

Mikel Artetxe, Gorka Labaka, and Eneko Agirre. 2018. A robust self-learning method for fully unsupervised cross-lingual mappings of word embeddings. *arXiv preprint arXiv:1805.06297*.

Alexis Conneau, Guillaume Lample, Marc'Aurelio Ranzato, Ludovic Denoyer, and Hervé Jégou. 2017. Word translation without parallel data. *arXiv preprint arXiv:1710.04087*.

Alexis Conneau, Guillaume Lample, Ruty Rinott, Adina Williams, Samuel R Bowman, Holger Schwenk, and Veselin Stoyanov. 2018. Xnli: Evaluating cross-lingual sentence representations. *arXiv preprint arXiv:1809.05053*.

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*.

Manaal Faruqui and Chris Dyer. 2014. Improving vector space word representations using multilingual correlation. In *Proceedings of the 14th Conference of the European Chapter of the Association for Computational Linguistics*, pages 462–471.

Stephan Gouws, Yoshua Bengio, and Greg Corrado. 2015. Bilbowa: Fast bilingual distributed representations without word alignments.

Karl Moritz Hermann and Phil Blunsom. 2014. Multilingual models for compositional distributed semantics. *arXiv preprint arXiv:1404.4641*.

Harold Hotelling. 1992. Relations between two sets of variates. In *Breakthroughs in statistics*, pages 162–190. Springer.

Melvin Johnson, Mike Schuster, Quoc V Le, Maxim Krikun, Yonghui Wu, Zhifeng Chen, Nikhil Thorat, Fernanda Viégas, Martin Wattenberg, Greg Corrado, et al. 2017. Googles multilingual neural machine translation system: Enabling zero-shot translation. *Transactions of the Association for Computational Linguistics*, 5:339–351.

Alexandre Klementiev, Ivan Titov, and Binod Bhattarai. 2012. Inducing crosslingual distributed representations of words. *Proceedings of COLING 2012*, pages 1459–1474.

Sneha Reddy Kudugunta, Ankur Bapna, Isaac Caswell, Naveen Arivazhagan, and Orhan Firat. 2019. Investigating multilingual nmt representations at scale. *arXiv preprint arXiv:1909.02197*.

Guillaume Lample and Alexis Conneau. 2019. Cross-lingual language model pretraining. *arXiv preprint arXiv:1901.07291*.

Nelson F Liu, Matt Gardner, Yonatan Belinkov, Matthew Peters, and Noah A Smith. 2019. Linguistic knowledge and transferability of contextual representations. *arXiv preprint arXiv:1903.08855*.

Thang Luong, Hieu Pham, and Christopher D Manning. 2015. Bilingual word representations with monolingual quality in mind. In *Proceedings of the 1st Workshop on Vector Space Modeling for Natural Language Processing*, pages 151–159.

Bryan McCann, James Bradbury, Caiming Xiong, and Richard Socher. 2017. Learned in translation: Contextualized word vectors. In *Advances in Neural Information Processing Systems*, pages 6294–6305.

Tomas Mikolov, Quoc V Le, and Ilya Sutskever. 2013a. Exploiting similarities among languages for machine translation. *arXiv preprint arXiv:1309.4168*.

Tomas Mikolov, Ilya Sutskever, Kai Chen, Greg S Corrado, and Jeff Dean. 2013b. Distributed representations of words and phrases and their compositionality. In *Advances in neural information processing systems*, pages 3111–3119.

Tomáš Mikolov, Ilya Sutskever, Anoop Deoras, Hai-Son Le, Stefan Kombrink, and Jan Cernocky. 2012. Subword language modeling with neural networks. *preprint (http://www. fit. vutbr. cz/imikolov/rnnlm/char. pdf)*, 8.

Ari Morcos, Maithra Raghu, and Samy Bengio. 2018. Insights on representational similarity in neural networks with canonical correlation. In *Advances in Neural Information Processing Systems*, pages 5727–5736.

Jeffrey Pennington, Richard Socher, and Christopher Manning. 2014. Glove: Global vectors for word representation. In *Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP)*, pages 1532–1543.

Matthew E Peters, Mark Neumann, Mohit Iyyer, Matt Gardner, Christopher Clark, Kenton Lee, and Luke Zettlemoyer. 2018. Deep contextualized word representations. *arXiv preprint arXiv:1802.05365*.

Maithra Raghu, Justin Gilmer, Jason Yosinski, and Jascha Sohl-Dickstein. 2017. Svcca: Singular vector canonical correlation analysis for deep learning dynamics and interpretability. In *Advances in Neural Information Processing Systems*, pages 6076–6085.

Pranav Rajpurkar, Jian Zhang, Konstantin Lopyrev, and Percy Liang. 2016. Squad: 100,000+ questions for machine comprehension of text. *arXiv preprint arXiv:1606.05250*.

Naomi Saphra and Adam Lopez. 2018. Understanding learning dynamics of language models with svcca. *arXiv preprint arXiv:1811.00225*.

Rico Sennrich, Barry Haddow, and Alexandra Birch. 2015. Neural machine translation of rare words with subword units. *arXiv preprint arXiv:1508.07909*.

Jasdeep Singh, Bryan McCann, Nitish Shirish Keskar, Caiming Xiong, and Richard Socher. 2019. Xlda: Cross-lingual data augmentation for natural language inference and question answering. *arXiv preprint arXiv:1905.11471*.

Richard Socher, Alex Perelygin, Jean Wu, Jason Chuang, Christopher D Manning, Andrew Ng, and Christopher Potts. 2013. Recursive deep models for semantic compositionality over a sentiment treebank. In *Proceedings of the 2013 conference on empirical methods in natural language processing*, pages 1631–1642.

Robert R Sokal. 1958. A statistical method for evaluating systematic relationships. *Univ. Kansas, Sci. Bull.*, 38:1409–1438.

Alex Wang, Amapreet Singh, Julian Michael, Felix Hill, Omer Levy, and Samuel R Bowman. 2018. Glue: A multi-task benchmark and analysis platform for natural language understanding. *arXiv preprint arXiv:1804.07461*.

Adina Williams, Nikita Nangia, and Samuel R Bowman. 2017. A broad-coverage challenge corpus for sentence understanding through inference. *arXiv preprint arXiv:1704.05426*.

Will Y Zou, Richard Socher, Daniel Cer, and Christopher D Manning. 2013. Bilingual word embeddings for phrase-based machine translation. In *Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing*, pages 1393–1398.