A Joint Matrix Factorization Analysis of Multilingual Representations

Zheng Zhao Yftah Ziser Bonnie Webber Shay B. Cohen

School of Informatics, University of Edinburgh

Overview

- This work presents an analysis tool using joint matrix factorization to investigate how multilingual pre-trained models capture linguistic information.
- An alternative to probing, our analysis tool uses subpopulation analysis [2], to compare the latent representations of multilingual and monolingual models.
- The comparison is done using **PARAFAC2** [1], which jointly factorizes representations into components that can be analyzed and recombined to uncover underlying structures and patterns in the data.
- A large-scale study on 33 languages and 17 morphosyntactic categories shows that different layers of multilingual representations encode morphosyntactic information at different levels, depending on the language and the categories.

Methodology

- Subpopulation analysis: compare the representations of control model $\mathbf{C}$ trained on data of interest and experimental model $\mathbf{E}$ trained on additional data from different sources;
- Here we have a multilingual model $\mathbf{E}$, and monolingual models $\mathbf{C}_\ell$ for $\ell \in [L]$;
- Two sets of representations (per language ℓ): $\mathbf{Z}_\ell, \mathbf{X}_\ell$ from data fed to $\mathbf{E}$ and $\mathbf{C}_\ell$;
- Obtain the covariance matrix $\mathbf{\Omega}_\ell$, defined as: $\mathbf{\Omega}_\ell = \mathbf{Z}_\ell^\top \mathbf{X}_\ell$;
- Apply PARAFAC2 on the set of joint matrices, decomposing each $\mathbf{\Omega}_\ell$ into:

$$\mathbf{\Omega}_\ell \approx \mathbf{U}_\ell \mathbf{\Sigma}_\ell \mathbf{V}^\top \quad (1)$$

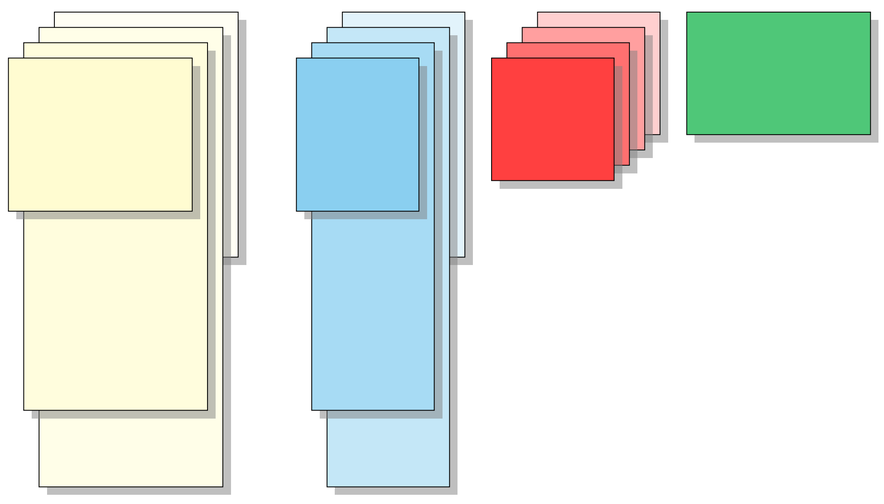


Figure 1. A diagram of the matrix factorization that PARAFAC2 performs.

We call the elements on the diagonal of $\mathbf{\Sigma}_\ell$ *signatures* of the language, which assesses if similar directions in $\mathbf{V}$ are important for the transformation to monolingual space.

Phylogenetic Tree

We are able to construct a phylogenetic tree using factorization outputs and hierarchical clustering between languages.

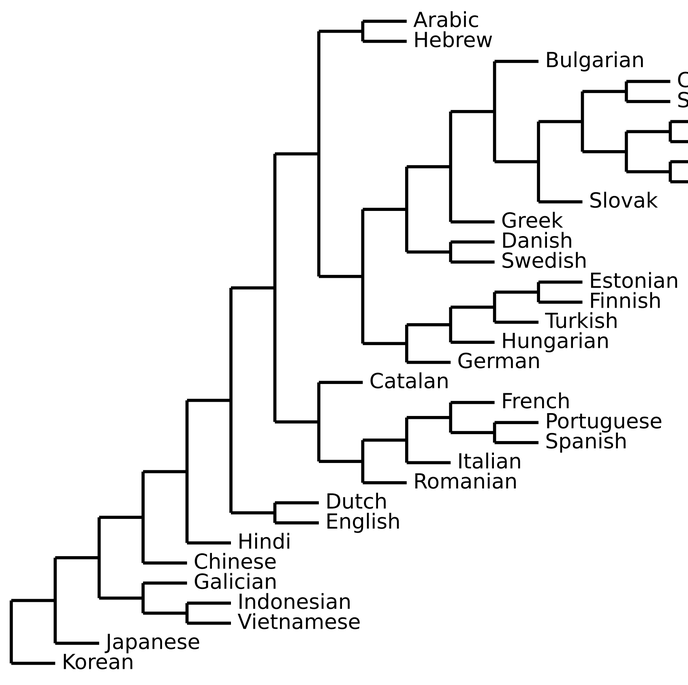


Figure 2. Phylogenetic trees of languages based on the distance between signatures vectors.

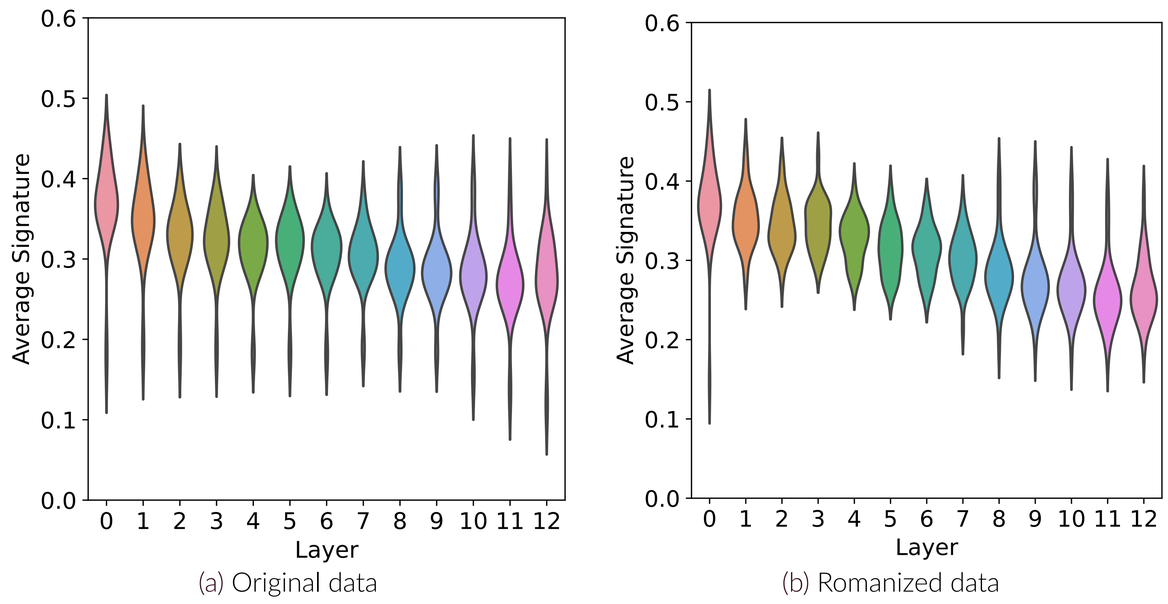
Experimental Setup

- **DATA:** CoNLL 2017 Wikipedia dump for training our models; treebanks from Universal Dependencies 2.1 for morphosyntactic features.
- **TASK:** masked language modelling (MLM).
- **MODEL:** *XLM-R* architecture for the multilingual E model; RoBERTa for the monolingual C_ℓ models. All models are trained from scratch.

Main Findings

1. The average signatures decreases across layers for all languages.

2. Multilingual models learn distinct representations for logographic writing systems (e.g., Chinese and Japanese), but romanization helps make the representations more similar.



(a) Original data (b) Romanized data

Figure 3. Comparison of average signature violin plots for all layers and languages between original data and data with Chinese and Japanese romanized.

3. Unique character count and TTR for each language negatively correlate with signatures.

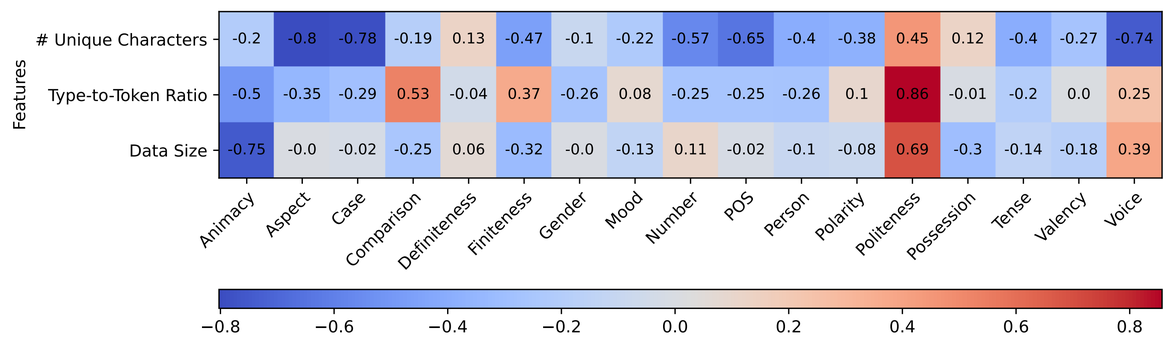


Figure 4. Pearson correlation results between the average signatures for all languages and other computed features for each morphosyntactic category.

4. No correlation between data size and average signatures in the overall dataset. However, within categories, correlation decreases across layers, suggesting language-specific properties' influence on representing morphosyntactic attributes in the multilingual model.

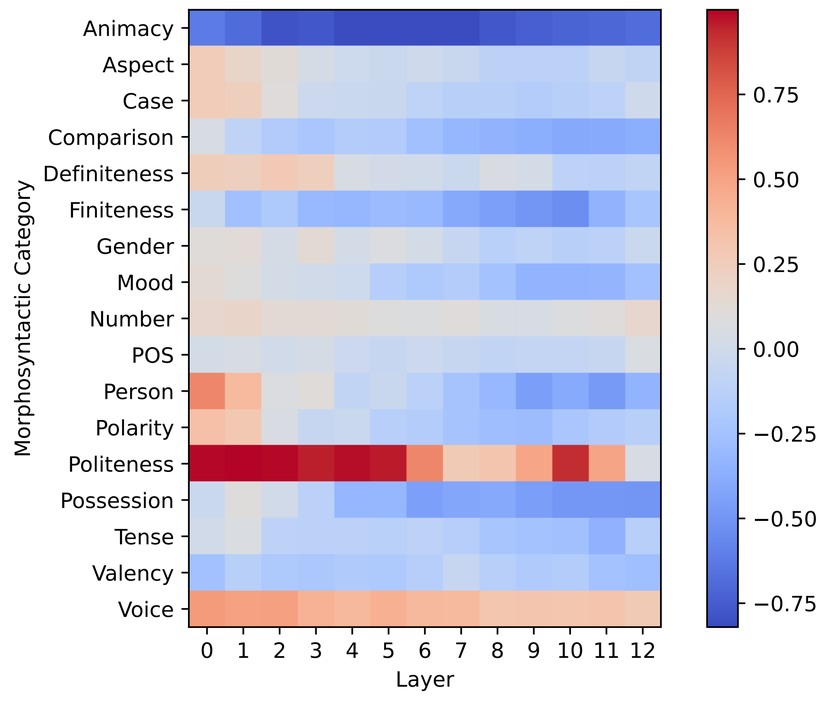


Figure 5. Pearson correlation results between the average signatures for all languages and their data size for each morphosyntactic category among all layers.

References

[1] R. A. Harshman. PARAFAC2: Mathematical and technical notes. *UCLA Working Papers in Phonetics*, 22:30–44, 1972b.

[2] Zheng Zhao, Yftah Ziser, and Shay Cohen. Understanding domain learning in language models through subpopulation analysis. In *Proceedings of the Fifth BlackboxNLP Workshop on Analyzing and Interpreting Neural Networks for NLP*, pages 192–209, Abu Dhabi, United Arab Emirates (Hybrid), December 2022. Association for Computational Linguistics.