Exploring Alignment in Shared Cross-Lingual Spaces[*]

Basel Mousi Nadir Durrani Fahim Dalvi
Majd Hawasly Ahmed Abdelali[†]
{bmousi,ndurrani,faimaduddin}@hbku.edu.qa
Qatar Computing Research Institute, HBKU Research Complex, Doha, Qatar

Extended Abstract

The emergence of multilingual contextualized embeddings has been a ground-breaking advancement, in the ever-evolving landscape of natural language processing. Adept at capturing the linguistic nuances across different languages, these embeddings have spurred a multitude of studies (Pires et al., 2019; Dufter and Schütze, 2020; Papadimitriou et al., 2021) seeking to understand the underlying mechanisms. How these models achieve multilinguality without explicit cross-lingual supervision during training is a particularly interesting question to answer.

Cross-lingual embeddings are designed to encode linguistic concepts that bridge equivalent semantic meaning across diverse languages. The question is: how well is this achieved in practice? When considering two arbitrary languages, *how well aligned are the embeddings of those languages?* and *how language agnostic are these multilingual embeddings in reality?* Addressing these questions necessitates a comprehensive approach.

In high-dimensional spaces, neural language models exhibit a capability to group words with shared linguistic associations, as highlighted by Mikolov et al. (2013). Expanding upon this foundational insight, recent research endeavors (Michael et al., 2020; Dalvi et al., 2022; Fu and Lapata, 2022) delved into conducting representation analysis within pre-trained models. Our objective, in this work, is to uncover encoded concepts within multilingual models and analyze their *alignment* and *overlap* across various languages within the latent space. We discover latent concepts by applying clustering to the underlying contextualized representations. The premise is that these clusters potentially signify latent concepts, encapsulating the language knowledge assimilated by the model. We build our work on top of this foundation to quantify concept *alignment* and *overlap* within multilingual latent space. We propose two metrics CALIGN and COLAP to quantify these two aspects and carry out analysis to study the following questions:

- To what extent do latent spaces across languages exhibit *alignment* and *overlap* in multilingual models?
- How does this change as the models are tuned towards any downstream NLP task?
- How do the multilingual latent spaces transform for zero-shot scenarios?

Methodology

Our approach focuses on discovering and analyzing latent concepts within multilingual neural language models to understand their alignment and overlap across languages. Figure 1 provides an overview of our methodology, which involves clustering contextualized embeddings to identify these encoded concepts (Dalvi et al., 2022). We introduce two novel metrics to analyze the alignment and overlap within multilingual models for this purpose: CALIGN (Concept Alignment) and COLAP (Concept Overlap).

CALIGN measures how well concepts in one language align with their counterparts in another, capturing the semantic coherence within the multilingual framework. This metric assesses the alignment of encoded concepts by identifying semantically equivalent tokens across different languages. Figure 2 shows sample aligned concepts.

COLAP investigates overlapping cross-lingual latent spaces within the model's representation, highlighting multilingual concepts that bring together words from multiple languages into a close latent space. This metric helps in understanding the intricate relationships between concepts across

[*]Work was published at ACL24
[†]Ahmed contributed to the project while he was at QCRI.

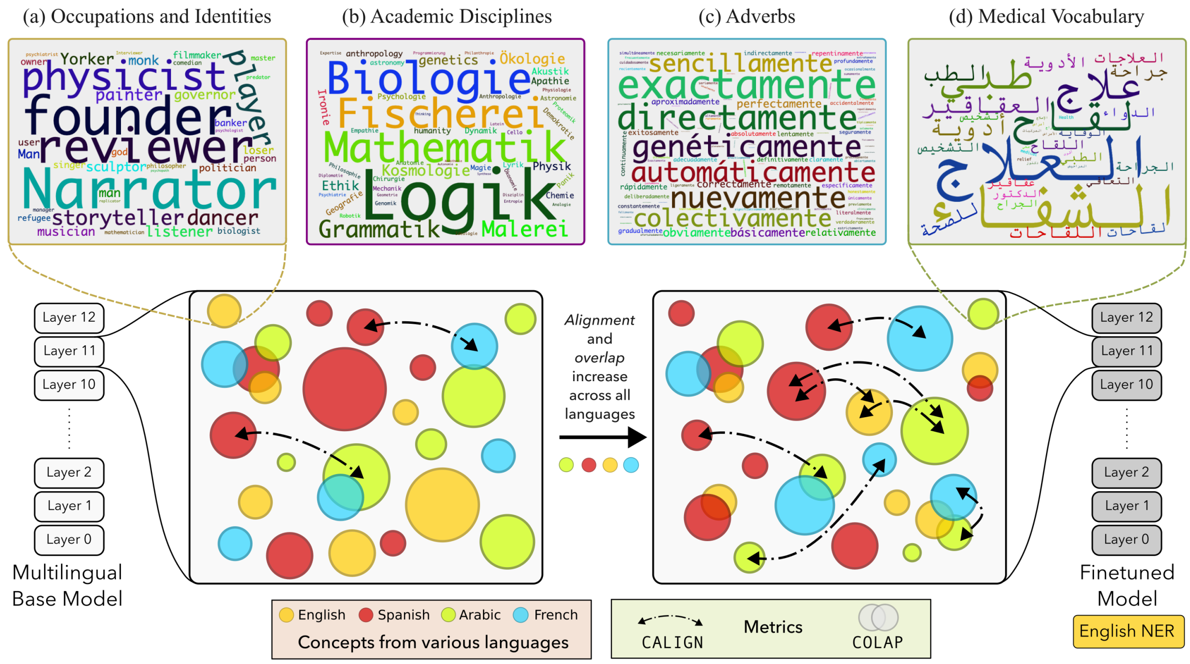


Figure 1: Overview of CALIGN and COLAP metrics in latent spaces of multilingual models, and how the space re-calibrates after fine-tuning. The top row shows concepts learned in mT5 across different languages: (a) English (b) German, (c) Spanish, (d) Arabic.

languages. Figure 3 shows sample overlapping concepts.

Setup and Results

We conducted a study employing three multilingual transformer models: mT5 (Xue et al., 2021), mBERT (Devlin et al., 2019), and XLM-RoBERTa (Conneau et al., 2020). These models were fine-tuned for three downstream NLP tasks: machine translation, named-entity recognition and sentiment analysis, spanning sequence generation, labeling and classification respectively. Our analysis revealed intriguing insights, including:

- Deeper layers in multilingual models preserve semantic concepts, contrasting with language-dependent lexical learning in lower layers, resulting in a higher alignment.
- Fine-tuning calibrates the latent space towards higher alignment and the task-specific calibration of the latent space facilitates zero-shot capabilities.
- Divergent patterns emerge in the encoder and decoder latent spaces in seq2seq models. The final layers in the decoder tend to primarily retain language specific concepts.

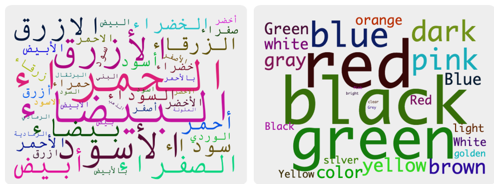
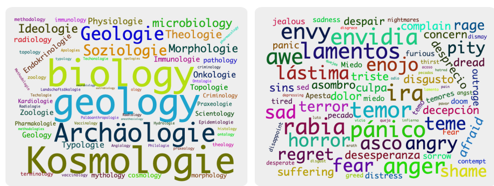


Figure 2: Aligned concepts: Colors in Arabic and English



(a) Shared olog infix (b) Emotions/Mind States

Figure 3: Overlapping concepts

- Closely related languages demonstrate higher overlap in latent space.
- The complexity of optimization function affects the extent of overlap in latent spaces
- While many model concepts exhibit multilingual traits, later layers post fine-tuning tend to retain primarily language-specific characteristics.

References

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. Association for Computational Linguistics.

Fahim Dalvi, Abdul Rafae Khan, Firoj Alam, Nadir Durrani, Jia Xu, and Hassan Sajjad. 2022. Discovering latent concepts learned in BERT. In *International Conference on Learning Representations*.

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*, NAACL-HLT '19, pages 4171–4186, Minneapolis, Minnesota, USA. Association for Computational Linguistics.

Philipp Dufter and Hinrich Schütze. 2020. Identifying elements essential for BERT's multilinguality. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 4423–4437, Online. Association for Computational Linguistics.

Yao Fu and Mirella Lapata. 2022. Latent topology induction for understanding contextualized representations.

Julian Michael, Jan A. Botha, and Ian Tenney. 2020. Asking without telling: Exploring latent ontologies in contextual representations. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing*, EMNLP '20, pages 6792–6812, Online. Association for Computational Linguistics.

Tomas Mikolov, Kai Chen, Greg Corrado, and Jeffrey Dean. 2013. Efficient estimation of word representations in vector space. In *Proceedings of the ICLR Workshop*, Scottsdale, AZ, USA.

Isabel Papadimitriou, Ethan A. Chi, Richard Futrell, and Kyle Mahowald. 2021. Deep subjecthood: Higher-order grammatical features in multilingual BERT. In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume*, pages 2522–2532, Online. 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.

Linting Xue, Noah Constant, Adam Roberts, Mihir Kale, Rami Al-Rfou, Aditya Siddhant, Aditya Barua, and Colin Raffel. 2021. mT5: A massively multilingual pre-trained text-to-text transformer. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 483–498, Online. Association for Computational Linguistics.