How is linguistic information **organized** in a **multilingual** language model, and can we use this internal structure?

Using a linear language classifier, we **project** an LM's embeddings towards Language X or Y, and evaluate the semantics of the result. We found that our alterations push the model towards the *prior* of the intended language, rather than boosting the semantic equivalent.

Counterfactually Probing Language Identity in Multilingual Models

Anirudh Srinivasan, Venkata S Govindarajan, Kyle Mahowald

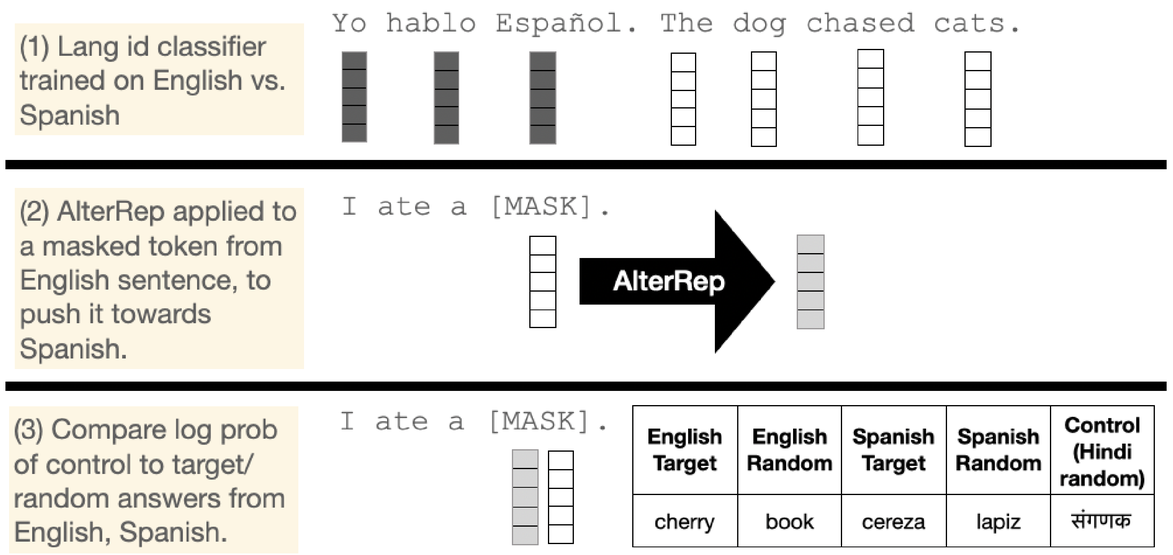

FIGURE 1: Overview of our proposed experiments.

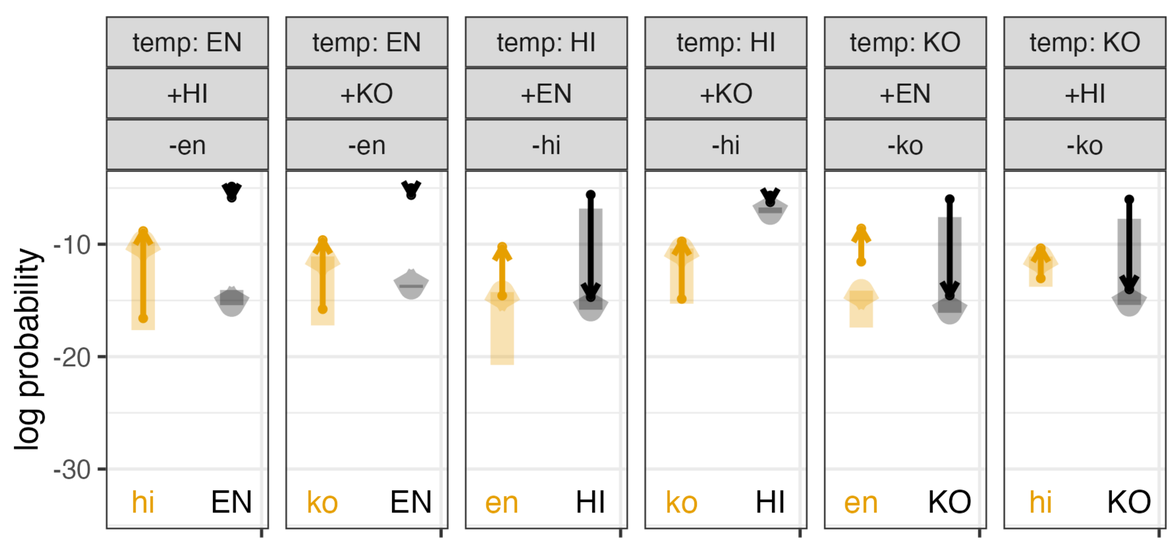

FIGURE 2: mBERT, pushing tokens in opp. direction to template.

Can probing techniques give insights into how information is organized and used in multilingual pretrained language models? To answer this question, we use **AlterRep**, a recent counterfactual probing technique, to test whether we can do a form of 'translation', effectively *pushing embeddings* towards a particular language.

By first training a linear classifier on a multilingual model's token embeddings, we can generate 'counterfactual' embeddings by projecting onto the null space of the classifier, and subsequently pushing in the direction of language X or Y. By evaluating the predictions of masked tokens in context *after* this intervention, we can infer if multilingual models encode tokens with **language-neutral** and **language-specific** components.

We experiment with multilingual BERT and XLM-Roberta in monolingual and code-mixed language settings.

Original sentence	I ate a *cherry*
Masked input to model	I ate a [MASK]
Mask replaced with target language word	I ate a *cereza*
Mask replaced with random target language word	I ate a *lapiz*
Mask replaced with third language word	I ate a *kirsikka*

TABLE 1: Example of how we replace a masked word.

When we push in the *opposite direction of the template*, the template language probabilities plummet both for the target and random words (see Figures 2 & 3), and the pushed to language probabilities increase significantly. Pushing in the *same direction as the template* essentially pushes it towards the language prior. Importantly, the probability of words in random control languages do not increase under either intervention.

We conclude that while multilingual models have linearly extractable (and manipulatable) language-specific and language-neutral components, we found no evidence that this can be used for translation.

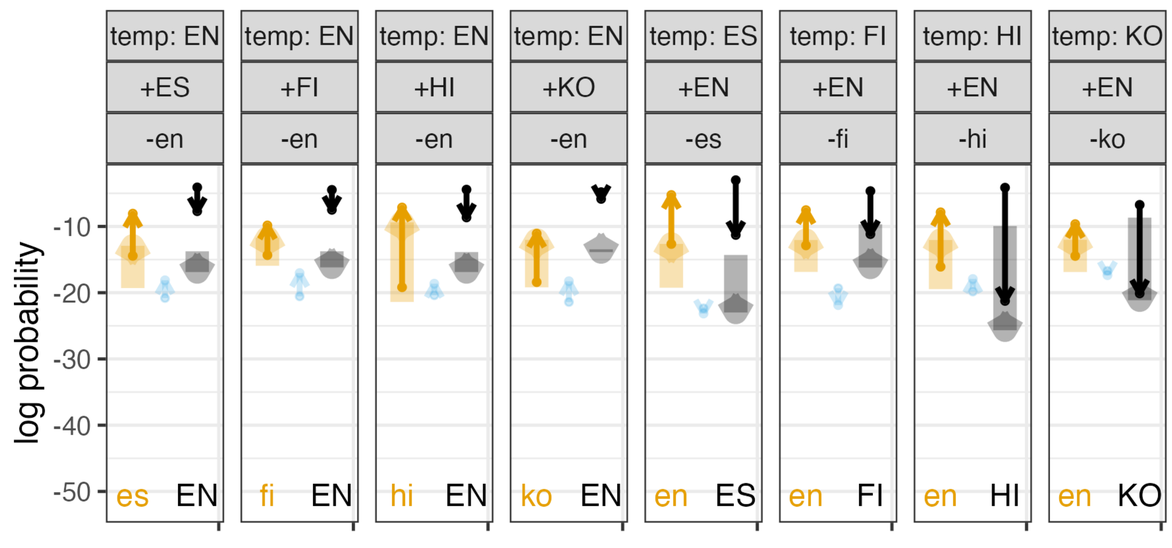

FIGURE 3: mBERT, pushing tokens in opp. direction to template in code-mixed setting.

Code available online at github.com/venkatasg/multilingual-counterfactual-probing

TEXAS
The University of Texas at Austin