BERTologiCoMix: How does Code-Mixing interact with Multilingual BERT?

Sebastin Santy, Anirudh Srinivasan, Monojit Choudhury
Microsoft Research, India

Microsoft

Code-Mixing + BERTology = BERTologiCoMix

Code Mixing and Code-Switching

Life ko face kiijiye with himmat and faith in yourself
"Face life with courage and faith in self"
She lives en una casa blanca
"She lives in a white house"

BERTology

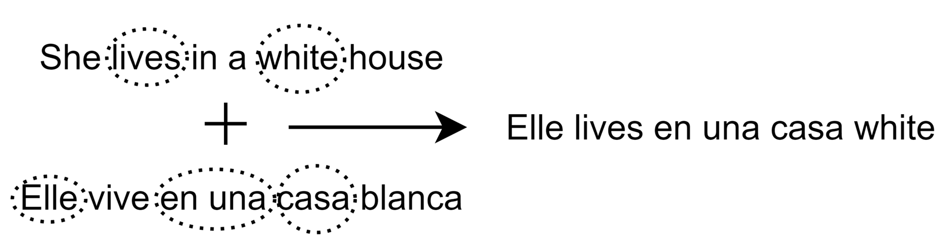
Series of studies probing BERT and its representations (Rogers et. al., 2020)

Questions we ask:

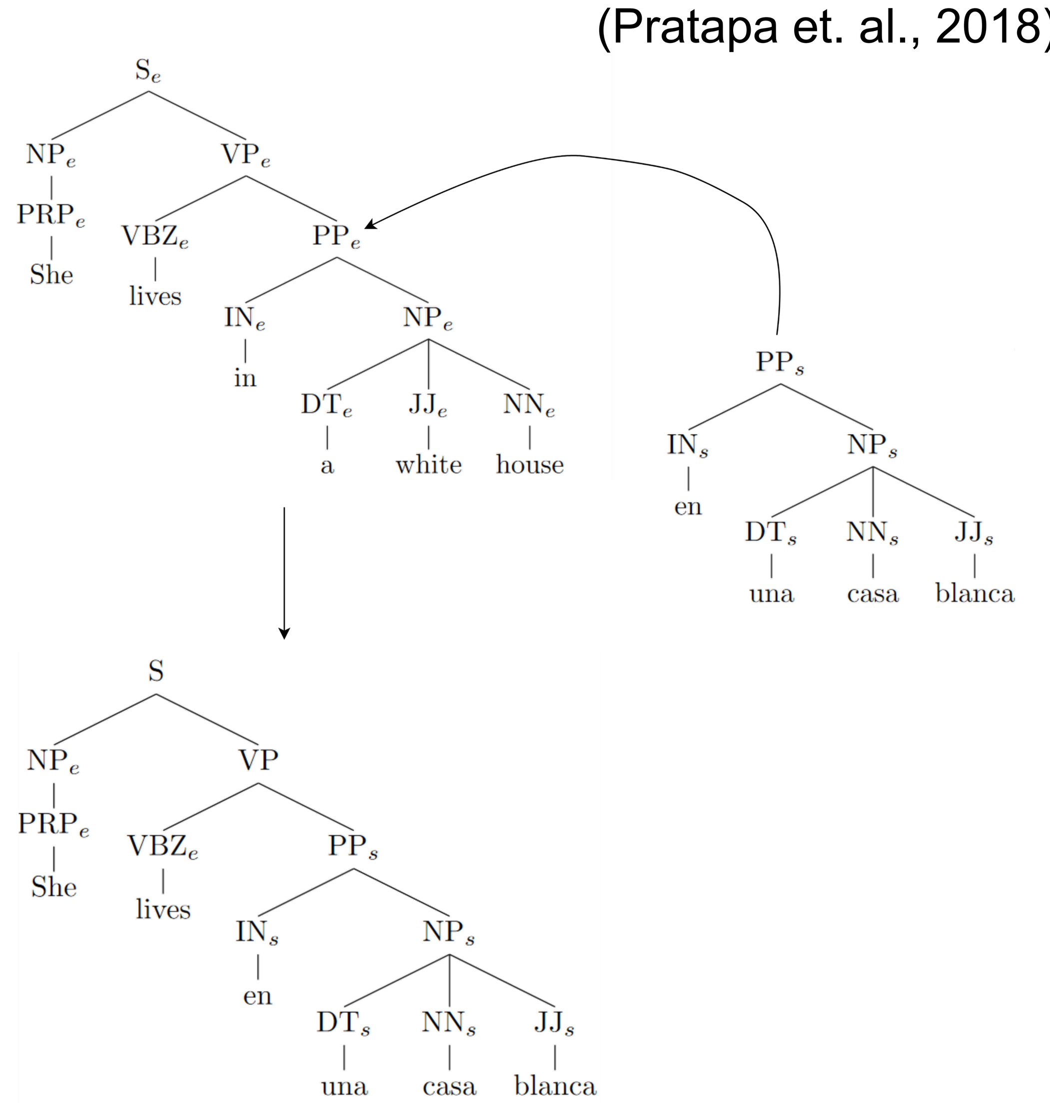
- What type of CM is ideal for mBERT finetuning?
- What changes happen to mBERT while finetuning?

Types of Code-Mixing

$(l{-}\mathrm{CM})$ – lexical Code−Mixing (random replacement)



$(g{-}\mathrm{CM})$ – generated Code-Mixing (synthetic) (Pratapa et. al., 2018)



$(r-\mathrm{CM})$ – real Code-Mixing (naturally occurring)

Downstream Task Experiments

$m\langle\,\rangle$ - stock mBERT i.e., without finetuning

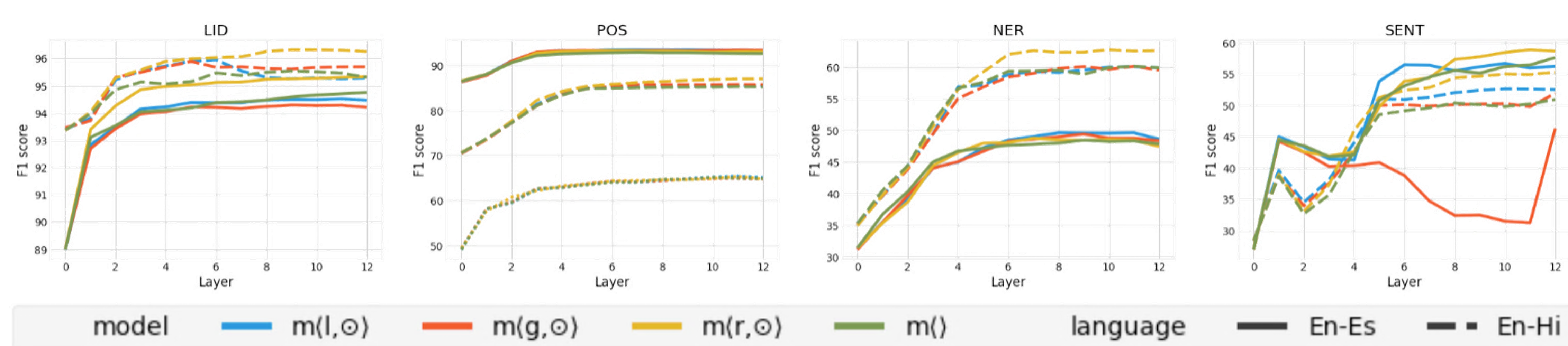
$m\langle l,\odot\rangle$ - mBERT finetuned on $(l-\mathrm{CM})$ $m\langle g,\odot\rangle$ - mBERT finetuned on $(g-\mathrm{CM})$ $m\langle r,\odot\rangle$ - mBERT finetuned on $(r-\mathrm{CM})$

GLUECoS Benchmark (Khanuja et al., 2020) consists of varied code-mixing tasks

Sentiment, NER, POS, Language ID, QA, NLI | English-Spanish ($enes$) and English-Hindi ($enhi$)

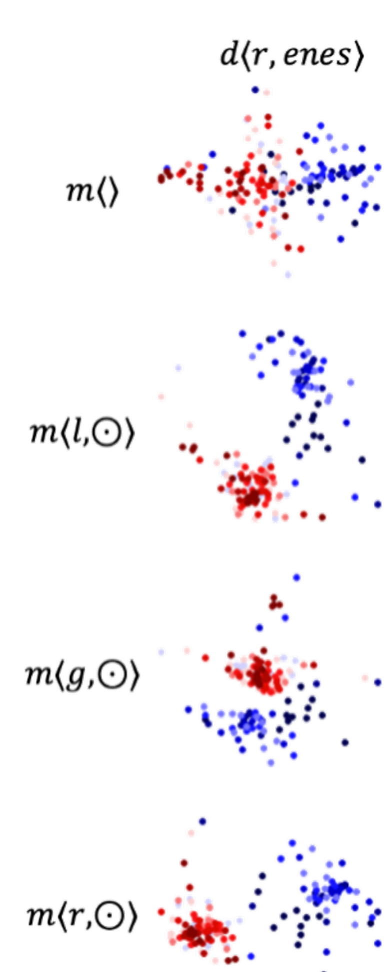
	SENT		NER		POS			LID		QA	NLI
model	*enes*	*enhi*	*enes*	*ehi*	*enes*	*enhi*	*enhi*	*enes*	*enhi*	*enhi*	*enhi*
$m_{\langle\rangle}$	$67.81_{\pm2.5}$	$\mathbf{58.42}_{\pm1.1}$	$59.50_{\pm0.9}$	$75.55_{\pm0.6}$	$93.35_{\pm0.2}$	$87.49_{\pm0.1}$	$63.40_{\pm0.5}$	$95.99_{\pm0.0}$	$\mathbf{95.80}_{\pm0.4}$	$71.95_{\pm0.8}$	$\mathbf{63.25}_{\pm1.9}$
$m_{\langle l,\odot\rangle}$	$68.07_{\pm1.5}$	$58.08_{\pm0.8}$	$59.39_{\pm1.0}$	$76.53_{\pm1.0}$	$\mathbf{93.84}_{\pm0.1}$	$88.00_{\pm0.2}$	$\mathbf{64.09}_{\pm0.2}$	$96.09_{\pm0.1}$	$95.32_{\pm0.9}$	$70.53_{\pm3.5}$	$62.94_{\pm2.7}$
$m_{\langle g,\odot\rangle}$	$68.64_{\pm1.5}$	$57.90_{\pm1.1}$	$59.88_{\pm0.7}$	$76.86_{\pm0.6}$	$93.74_{\pm0.1}$	$87.79_{\pm0.2}$	$63.79_{\pm0.2}$	$96.06_{\pm0.0}$	$95.41_{\pm0.8}$	$70.11_{\pm1.8}$	$55.19_{\pm6.5}$
$m_{\langle r,\odot\rangle}$	$\mathbf{68.51}_{\pm0.7}$	$58.25_{\pm0.8}$	$\mathbf{60.46}_{\pm0.6}$	$\mathbf{76.86}_{\pm0.5}$	$93.68_{\pm0.1}$	$\mathbf{88.00}_{\pm0.0}$	$63.38_{\pm0.0}$	$\mathbf{96.12}_{\pm0.0}$	$94.60_{\pm0.2}$	$\mathbf{73.54}_{\pm3.9}$	$60.00_{\pm5.7}$

Probing for layer-wise performance on different downstream tasks (Tenney et al., 2019)



Differential Visualization

How does stock mBERT change with continued-pretraining on $(l-\mathrm{CM})$, $(g-\mathrm{CM})$ or $(r-\mathrm{CM})$?



Intra-head distance (Clark et al.)

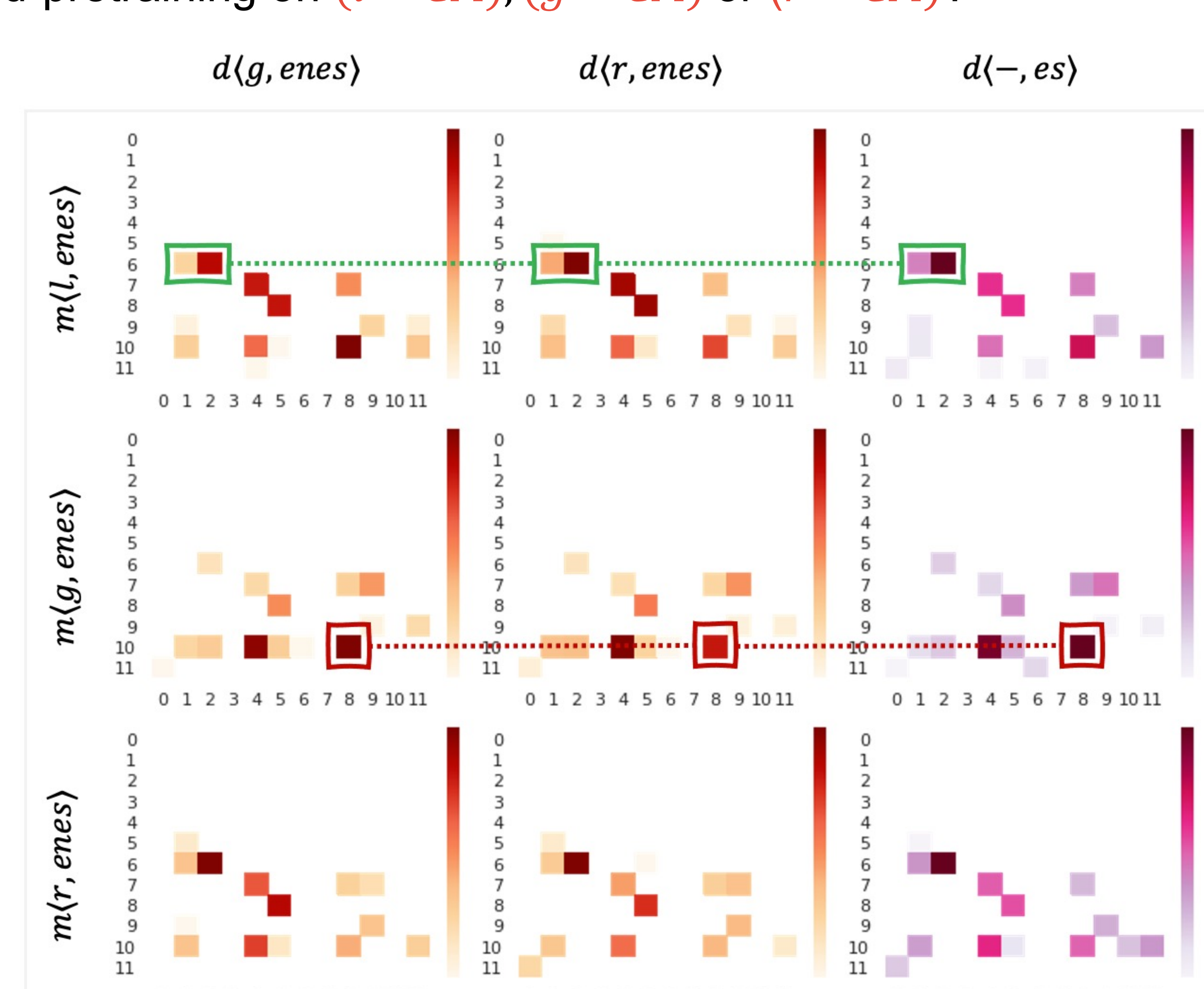
$$\Delta_m = JS(\mathbf{H}_{i,j}^{m\langle\odot,\odot\rangle}(\text{token}), \mathbf{H}_{i,j}^{m\langle\rangle}(\text{token}))$$

As more CM is introduced, we can see that that there are formation of tighter clusters between layers of mBERT.

Inter-head distance between fine-tuning with different data

$$\sum_{\text{token}\in\text{sentence}} JS(\mathbf{H}_{i(\text{token})}, \mathbf{H}_{j(\text{token})})$$

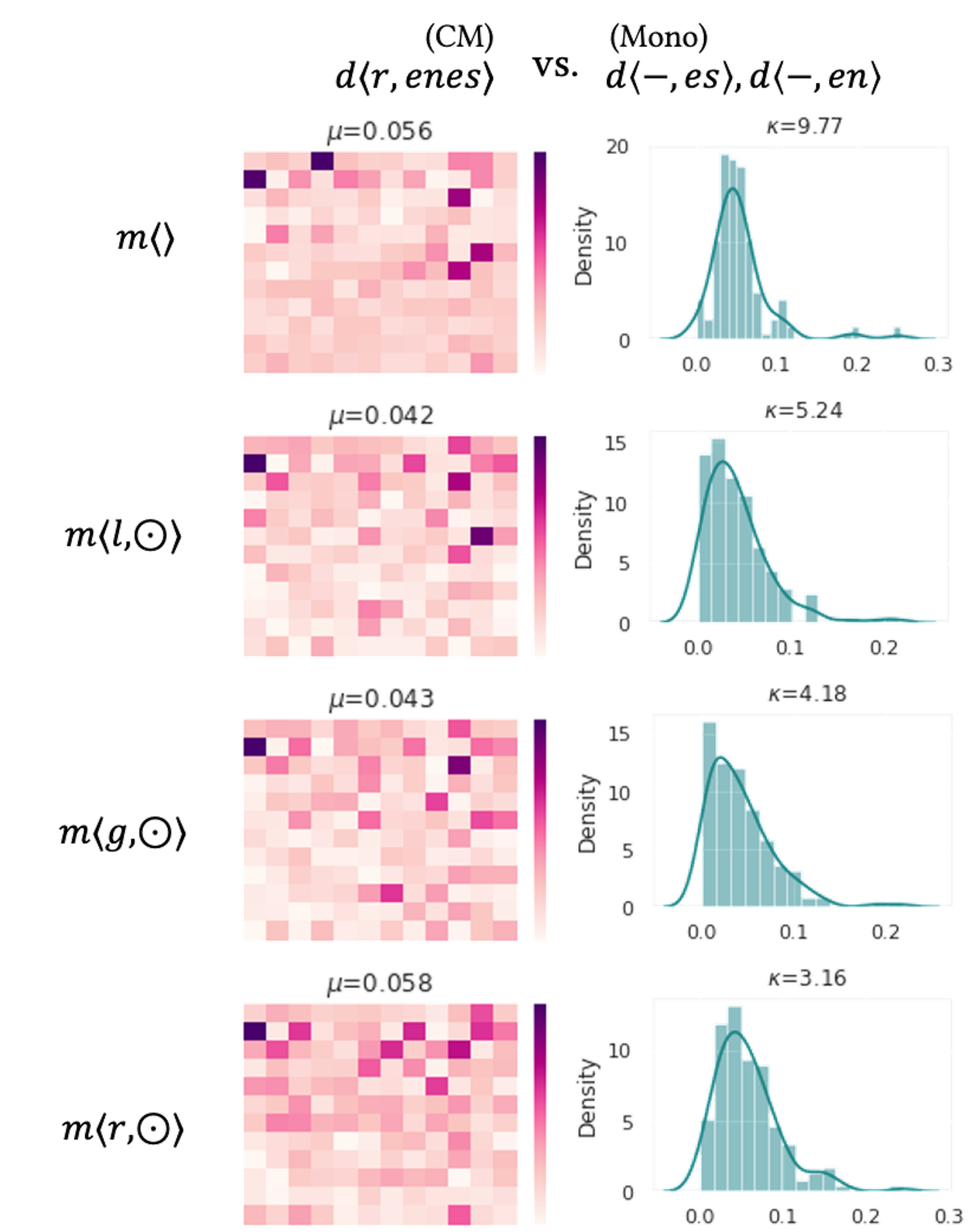
We can see that there are some common heads between different configurations.



Responsivity to Code-Mixing

Build a classifier to distinguish between Monolingual and Code-Mixed sentences using BERT attention head representations by measuring responsivity ($R_{x,y}$) (analogous to calculating information gain of features)

$$\mathcal{R}_{x,y} = H(x) - H(x|y)$$



More heads respond to CM after finetuning with $(r-\mathrm{CM})$ data as compared to either $(g-\mathrm{CM})$ and $(l-\mathrm{CM})$

References

Pratapa, et al. "Language modeling for code-mixing: The role of linguistic theory based synthetic data". *ACL* (2018).

Khanuja, et al. "GLUECoS: An Evaluation Benchmark for Code-Switched NLP." *ACL* (2020).

Tenney, et al. "BERT rediscovers the classical NLP pipeline." *ACL* (2019)

Rogers, et al. "A Primer in BERTology: What we know about how BERT works". *TACL* (2020)