Modular Sentence Encoders: Separating Language Specialization from Cross-Lingual Alignment

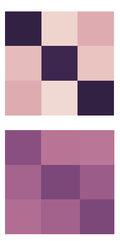
Yongxin Huang[1], Kexin Wang[1], Goran Glavaš[2], Iryna Gurevych[1]

[1]UKP Lab, Technical University of Darmstadt; [2]Center for AI and Data Science, University of Würzburg

Motivation

Curse of multilinguality in multilingual sentence encoders like LaBSE and mE5: multilingual training on shared parameters leads to **negative interference** between languages.

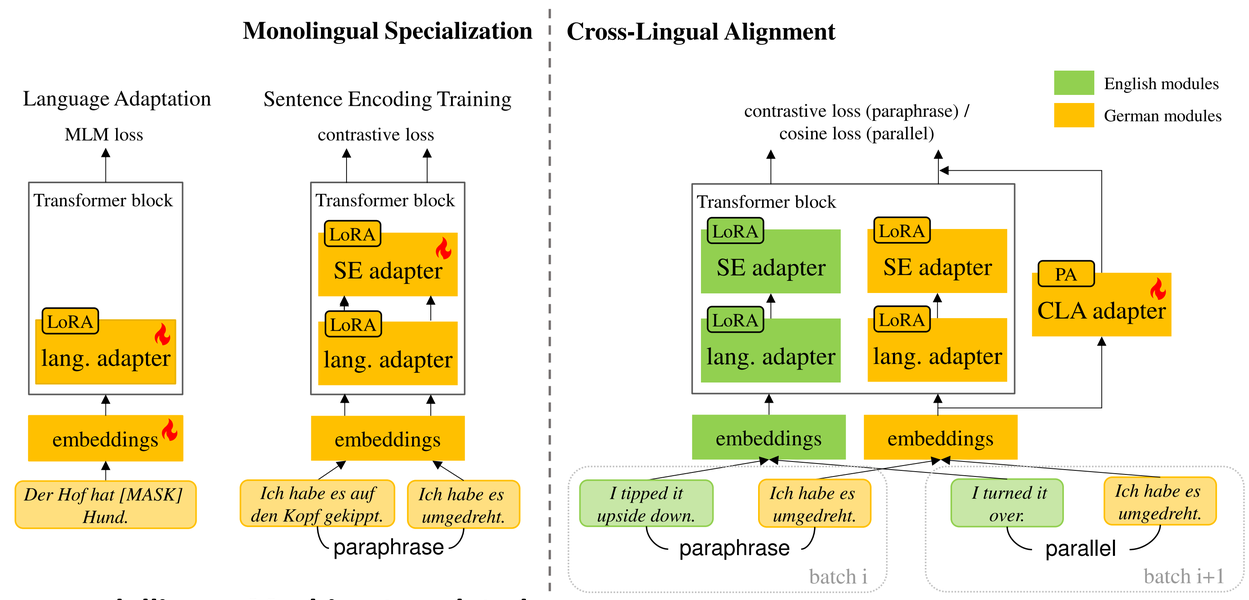
Aligning **cross-lingual** representations distorts **monolingual** semantic spaces: Cross-lingual training improves cross-lingual performance at the cost of within-language performance.



Different cross-lingual tasks place **conflicting demands** on the representation space. **Cross-lingual transfer** requires similar monolingual space structures; sentence pairs that are positive in **STS** (similar meaning) are negative in **bitext mining** (non-translations).

Method

We apply modular training to a pre-trained multilingual sentence encoder.



Language modelling with language-specific tokenizer and embedding layer.

Machine-translated data for sentence embedding training in each language.

Learning fine-grained semantic similarity with cross-lingual **paraphrases**.

Parallel data benefits cross-lingual transfer.

Findings

		Monolingual Tasks			Cross-lingual Tasks					Alignment Metrics		
		STS	STR	CLS	STS	STR	CLS	Bitext Mining		Lang. Bias		RSIM
		stsb	sick	sib	stsb	sick	sib	flores	tatoeba	stsb	sick	flores
	mE5	72.5	74.2	74	54.1	61	73.5	1.85	9.89	23.22	12.11	0.6
Training Method												
Full	mono	79.6	75.5	85.5	60.2	64.1	85.2	0.62	7.85	2.6	3.16	0.67
	mono + cross	77.4	73.1	85.4	66.7	66.9	86.5	0.26	6.33	1.05	1.14	0.74
Modular	mono	82.1	75.4	87.8	69.8	68.5	87.7	0.22	5.27	2.82	3.07	0.74
	mono + cross	81.9	76.4	88.3	73.8	70.7	88.3	0.19	5	1.33	1.73	0.8

		Monolingual Tasks					Cross-lingual Tasks						Alignment Metrics		
		STS		STR		CLS	STS		STR	CLS	Bitext Mining		Lang. Bias		RSIM
		sts17	stsb	sick	str24	sib	sts17	stsb	sick	sib	flores	tatoeba	stsb	sick	flores
	LaBSE	76.7	71.9	68	69.2	82.7	74.5	64.4	63.8	83.6	0.14	3.87	1.02	2.32	0.64
Training Method															
Full	mono	82.9	80.4	76.4	75.9	84.8	79.4	71.5	70.9	83.9	0.29	4.43	0.88	1.27	0.74
	mono + cross	80	79.2	75.1	75.4	86	76.7	72.7	71.7	86.3	0.21	4.17	0.53	0.64	0.77
Modular	mono	83.1	82.1	76.5	78.4	85.5	80.6	75.3	71.9	85	0.15	3.63	1.05	1.16	0.75
	mono + cross	82.7	82.1	76.6	78.1	85.8	80.3	76.4	72.7	85.7	0.15	3.55	0.56	0.78	0.79

	STS/STR (monolingual)	STS/STR (cross-lingual)	Classification (cross-lingual)	RSIM
	LaBSE			
paraphrase	79.7	76.5	85.0	76.0
parallel	79.1	75.9	86.2	82.0
both	79.9	76.5	85.7	79.0
	multilingual-e5-base			
paraphrase	79.1	71.9	87.6	0.8
parallel	78.0	71.2	89.0	0.8
both	79.2	72.3	88.3	0.8

Language specialization mitigates the curse of multilinguality in the monolithic model, boosting performance in both monolingual and cross-lingual tasks, even before any explicit cross-lingual training.

Cross-lingual alignment adapters further improves cross-lingual tasks and reduces language bias, without sacrificing monolingual performance.

In contrast, **cross-lingual training on full parameters** interferes with monolingual training and degrades monolingual performance.

Low **language bias** != high cross-lingual performance. Language spaces in the multilingual model (Full) can be well-aligned, but the quality of the semantic representations remains low.

Cross-lingual training strategies

- Training with only **parallel** data ➞ high isomorphism between monolingual spaces (RSIM) ➞ stronger cross-lingual transfer in classification.
- Training with only cross-lingual **paraphrase** data ➞ better at STS/STR.
- Combining both training strategies mitigates their individual shortcomings.

Evaluation on 23 languages

Semantic Textual Similarity/Relatedness: The most extensive multilingual evaluation. For the first time, evaluation on rare language pairs is enabled by combination of existing STS datasets in many languages.

- STS17 in en, ar, cs, de, es, fr, it, ko, nl, tr
- STSB in en, az, kk, ko, ky, ug, uz (low-resource)
- SICK in en, es, nl, pl
- STR24 in en, am, ha, mr, rw, te (low-resource)

Classification: SIB-200 in all the languages above

Bitext Mining:

- FLORES-200 in all the languages above
- Tatoeba in all the languages above except ky, ha, rw

Two alignment metrics

- **Language bias** in STS/STR: Does the model prefer one language over another? Measured as the performance drop when switching from *bilingual* to *multilingual* evaluation on the concatenation of all bilingual datasets.
- **Relational Similarity (RSIM):** degree of isomorphism of monolingual semantic space structures. Pearson correlation between similarities of all monolingual sentence pairs from a bilingual parallel corpus.

Links