Massively Multilingual Transformers

– Deep Transformer nets pretrained on large **multilingual corpora** via (masked) **language modeling** objectives
  - Multilingual BERT, XLM-R, mT5

– Automatically induces shared (subword) vocabulary across all languages

– **Unsupervised** from the perspective of explicit cross-lingual signal
  - Deemed very effective for zero-shot CL transfer

„Suprising cross-lingual effectiveness of BERT"

„mBERT surprisingly good at zero-shot CL model transfer"

Massively Multilingual Transformers

– **Assumption**: after multilingual MLM pretraining mBERT can encode text from any of the languages seen in pretraining

– Automatically lends itself to **zero-shot language transfer** for downstream NLP tasks

– **mBERT** has its own *tokenizer* that can tokenize input texts from all languages seen in pre-training
  - Caveat: words from larger languages mostly have their own tokens
    - Words from smaller languages broken down into subwords which can be found across languages
    - Worst case scenario: input broken into characters

Cross-Lingual Transfer with MMTs

Zero-shot language transfer for downstream NLP tasks with mBERT:

1. Couple the mBERT Transformer with the task-specific classifier („head”)
2. Train the mBERT+classifier model jointly on source language data
   - Classifier parameters trained from scratch
   - mBERT’s Transformer parameters fine-tuned
3. Predict by feeding the target language text (tokenized with mBERT’s tokenizer) into the fine-tuned mBERT+classifier model

So...has mBERT solved zero-shot CL transfer?

– **No!** Settings in which they were evaluated were simply too favorable

„How multilingual is Multilingual BERT?" [Pires et al., ACL 19]

- **Tasks**: NER, POS; **Target languages**: DE, NL, ES

„Cross-lingual Ability of mBert: Empirical Study" [Karthikeyan et al., ICLR 20]

- **Tasks**: NER, NLI; **Target languages**: ES, HI, RU

– In most studies, the selected target languages were:

- (1) from the **same language family**,
- (2) with **large corpora in pretraining**

Zero-shot transfer performance drops

Lauscher, A., Ravishankar, V., Vulić, I., & Glavaš, G. (2020). *From Zero to Hero: On the Limitations of Zero-Shot Cross-Lingual Transfer with Multilingual Transformers*. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, 4483-4499).

Task	Model	EN	ZH Δ	TR Δ	RU Δ	AR Δ	HI Δ	EU Δ	FI Δ	HE Δ	IT Δ	JA Δ	KO Δ	SV Δ	VI Δ	TH Δ	ES Δ	EL Δ	DE Δ	FR Δ	BG Δ	SW Δ	UR Δ
DEP	B	91.2	-43.9	-46.0	-28.1	-56.4	-36.1	-50.2	-30.7	-36.1	-17.1	**-60.1**	-56.1	-14.3	-	-	-	-	-	-	-	-	-
	X	92.0	**-85.4**	-44.2	-29.7	-54.6	-39	-49.5	-26.7	-39	-23.5	-80.5	-56.0	-16.3	-	-	-	-	-	-	-	-	-
POS	B	95.8	-38.0	-35.9	-16.0	-40.1	-33.4	-34.6	-21.9	-33.4	-19.8	**-46.1**	-42.0	-9.6	-	-	-	-	-	-	-	-	-
	X	96.3	-69.2	-27.7	-14.3	-37.1	-27.3	-31.9	-17.9	-27.3	-19.0	**-77.0**	-37.3	-10.7	-	-	-	-	-	-	-	-	-
NER	B	92.4	-23.3	-11.6	-10.7	**-31.7**	-11.1	-12.8	-3.8	-11.1	-2.6	-25.7	-13.8	-6.7	-	-	-	-	-	-	-	-	-
	X	91.6	**-34.8**	-6.2	-13.7	-24.6	-16.5	-8.0	-0.9	-16.5	-2.4	-30.1	-15.6	-2.2	-	-	-	-	-	-	-	-	-
XNLI	B	82.8	-13.6	-20.6	-13.5	-17.3	-21.3	-	-	-	-	-	-	-	-11.9	-28.1	-8.1	-14.1	-10.5	-7.8	-13.3	**-33.0**	-23.4
	X	84.3	-11.0	-11.3	-9.0	-13.0	-14.2	-	-	-	-	-	-	-	-9.7	-12.3	-5.8	-8.9	-7.8	-6.1	-6.6	**-20.2**	-17.3
XQuAD	B	71.1	-22.9	-34.2	-19.2	-24.7	-28.6	-	-	-	-	-	-	-	-22.1	**-43.2**	-16.6	-28.2	-14.8	-	-	-	-
	X	72.5	**-26.2**	-18.7	-15.4	-24.1	-22.8	-	-	-	-	-	-	-	-19.7	-14.8	-14.5	-15.7	-16.2	-	-	-	-

- B = mBERT (Base), X = XLM-R (Base)
- Drops **huge** for:
  1. Distant target languages and
  2. Target languages with small pretraining corpora

Language-Specific Representation Subspaces

- In representation spaces produced by MMTs, one can still relatively easy discern language-specific subspaces

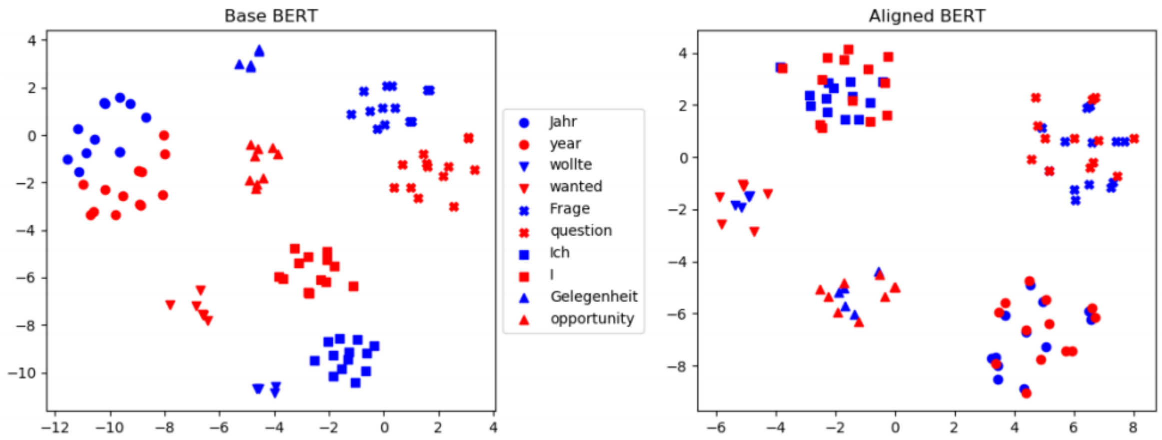

Image from [Cao et al., ‘20]

Better alignment between language subspaces...

- ...can be achieved with **bilingual supervision** (word translations of parallel data) [Wu & Conneau, ACL 20; Cao et al., ICLR 20; Hu et al., 2020]
- As with CLWEs: some bilingual/multilingual supervision ➔ better bilingual/multilingual representation space

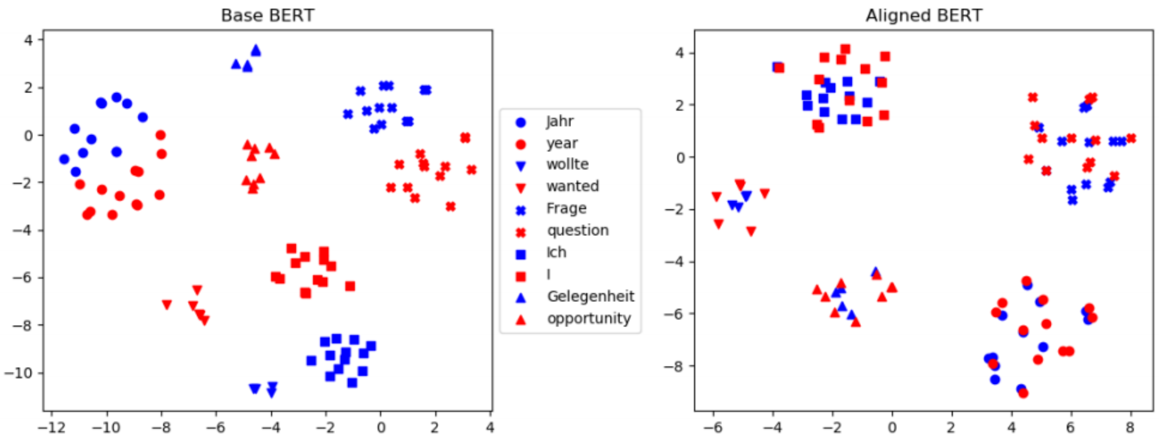

Image from [Cao et al., '20]

Choosing a Language Sample for CL Transfer Experiments

- Multilingual evaluation benchmarks should assess the expected performance of a model **across languages**
  - Sample of languages should be representative – but of what exactly?
- Findings can critically depend on the selection of languages
  - Most studies sample languages with the **largest digital footprint**
  - Such languages tend to belong to the same families (e.g., Indo-European)
  - Expected transfer performance is overestimated!

Variety sampling of languages

Ponti, E. M., Glavaš, G., Majewska, O., Liu, Q., Vulić, I., & Korhonen, A. (2020). *XCOPA: A Multilingual Dataset for Causal Commonsense Reasoning*. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing,* pp. 2362-2376.

Idea: selection according to the distribution of linguistic properties

- **Variety sampling** favors the inclusion of outlier languages

1. **Typological diversity**: entropy of distribution of linguistic properties
2. **Family index**: number of different families / sample size
3. **Geography index**: entropy of lang. distr. over 6 geographic macro-areas

	Range	XCOPA	TyDiQA	XNLI	XQUAD	MLQA	PAWS-X
Typology	[0, 1]	0.41	0.41	0.39	0.36	0.32	0.31
Family	[0, 1]	1	0.9	0.5	0.6	0.66	0.66
Geography	[0, ln 6]	1.67	0.92	0.37	0	0	0

Learning outcomes

– Now you...

1. Understand what multilingual NLP is and why we need it
2. Know the mechanisms for inducing multilingual representations spaces
    - Cross-lingual word embeddings (CLWEs)
    - Massively multilingual transformers (MMTs)
3. Understand how to use multilingual representations spaces for CL transfer