Cross Lingual Transfer with Multilingual Language Models

Benjamin Muller

Seminar CSLP JHU, March 2022

Inria

Acknowledgment

This presentation summarize work done in collaboration and under the supervision of:

- Benoit Sagot, INRIA Paris
- Djamé Seddah, INRIA Paris
- Antonis Anastasopoulous, GMU
- Yanai Elazar, Bar Ilan University

Motivation

Most languages are not studied by the NLP community

- Only a few dozen languages benefit from the best models
- Our SOTA models are English-centric

Hundreds of Millions of people have smartphones but no access to good search engines, ASR, translation... (Blasi et al. 2021)

Motivation

Most languages are not studied by the NLP community

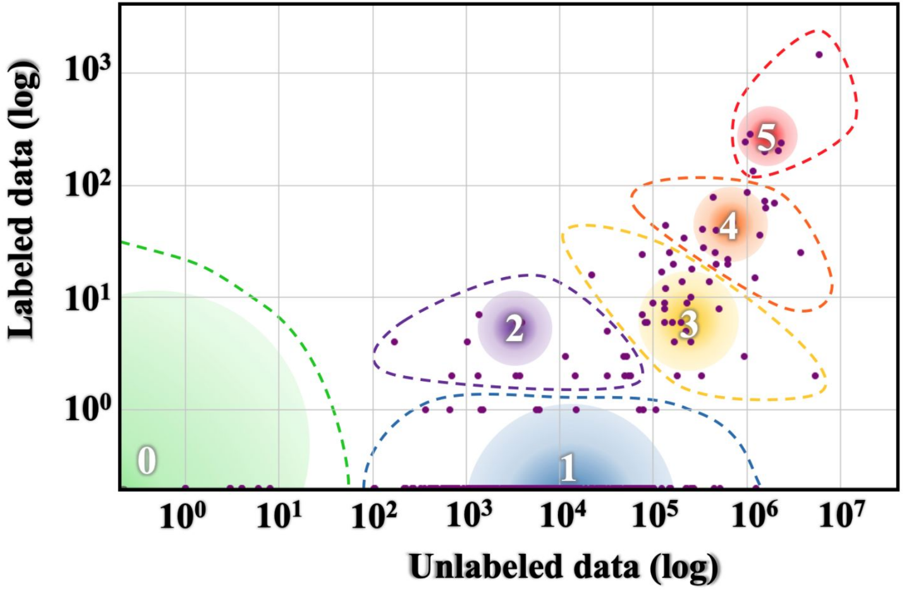
- Only a few dozen languages benefit from the best models
- Our SOTA models are English-centric

Hundreds of Millions of people have smartphones but no access to good search engines, ASR, translation... (Blasi et al. 2021)

How to build better NLP models for the largest number of low-resource languages?

Why low-resource languages?

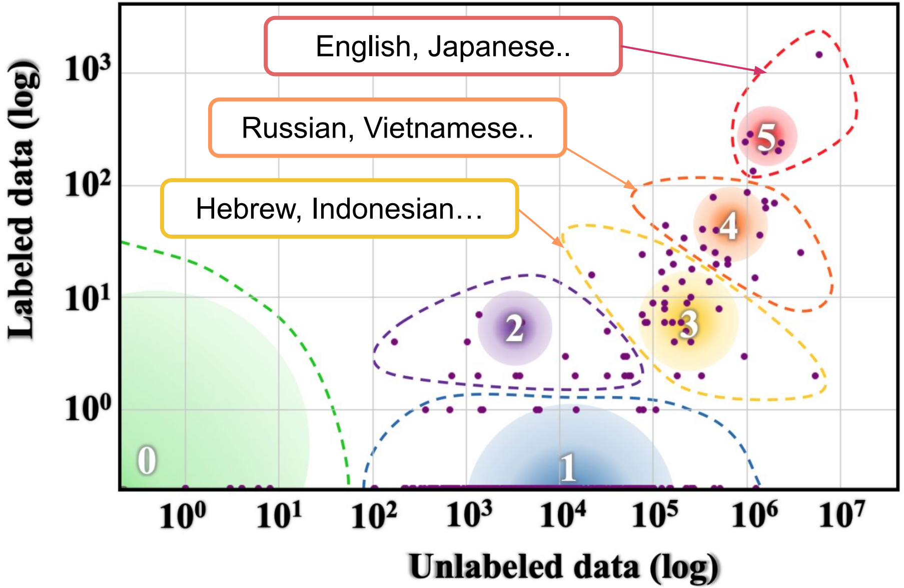
- About **7000 languages** in the world (Ethnologue 2021)



(Joshi et al. 2020)

Why low-resource languages?

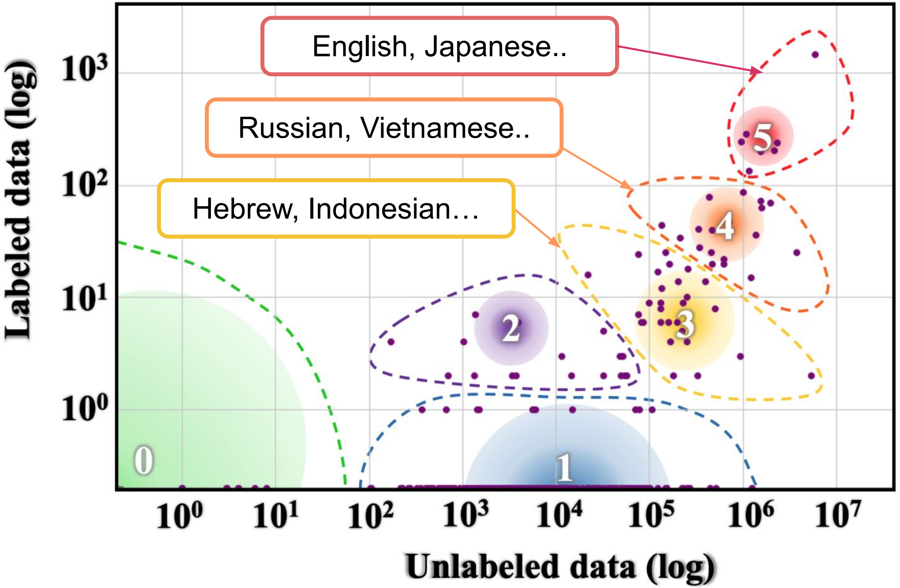
- About 7000 languages in the world (Ethnologue 2021)
- Only a few dozens (3 to 5) benefit from progress in NLP



(Joshi et al. 2020)

Why low-resource languages?

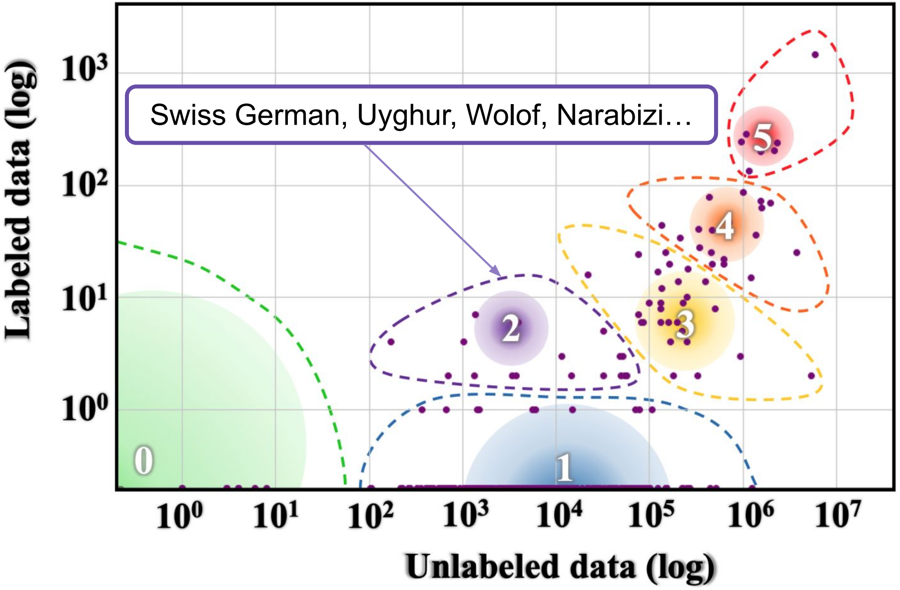
- About 7000 languages in the world (Ethnologue 2021)
- Only a few dozens (3 to 5) benefit from progress in NLP
- Thousands of languages are left-out



(Joshi et al. 2020)

Why low-resource languages?

- About 7000 languages in the world (Ethnologue 2021)
- Only a few dozens (3 to 5) benefit from progress in NLP
- Thousands of languages are left-out
- Focus on the "Hopefuls" languages (Category 2)



(Joshi et al. 2020)

E.g. North African Arabic Dialect: *Narabizi*

- Used online by millions of people
- Non-standard, code-mixing with French, very rich morphology
- Very small raw corpus available (10mb~) & very few annotated datasets
- Usually written in the Latin Script (Arabizi)

“Mrhba, Ana **3**rbi mn dzaye“:

مرحبا,أنا عربي من الجزائر

Hi, I’m arabic from Algeria

(Seddah et al. 2020)

Research Question and Framework

Context: “Large-Scale” Language Models are **great transfer learners** (Devlin et al. 2018, Pires et al. 2019)

Research Question and Framework

Context: "Large-Scale" Language Models are **great transfer learners** (Devlin et al. 2018, Pires et al. 2019)

How can we use language models efficiently for low-resource languages?

Research Question and Framework

Context: “Large-Scale” Language Models are **great transfer learners** (Devlin et al. 2018, Pires et al. 2019)

How can we use language models efficiently for low-resource languages?

- By reusing pretrained multilingual language models (mBERT, XLM-R, mT5...)
- By adapting them
- By training new models from scratch

Research Question and Framework

Context: “Large-Scale” Language Models are **great transfer learners** (Devlin et al. 2018, Pires et al. 2019)

How can we use language models efficiently for low-resource languages?

- By reusing pretrained multilingual language models (mBERT, XLM-R, mT5...)
- By adapting them
- By training new models from scratch

Tasks: POS tagging, Dependency Parsing , NER

Language Modeling Framework

Standard Setting

1. Pretraining

$$p_{\theta_0}(X)$$

Language Modeling Framework

Standard Setting

1. Pretraining

$$p_{\theta_0}(X)$$

2. Fine-Tuning

$$p_{\theta_1}(Y|X, \theta_0)$$

Language Modeling Framework

Standard Setting

1. Pretraining $p_{\theta_0}(X)$

2. Fine-Tuning $p_{\theta_1}(Y|X, \theta_0)$

3. Evaluation $p_{\theta_1}(Y|X)$

Language Modeling Framework

	Standard Setting
1. Pretraining	$p_{\theta_0}(X)$
2. Fine-Tuning	$p_{\theta_1}(Y\|X, \theta_0)$
3. Evaluation	$p_{\theta_1}(Y\|X)$

➔ Improves SOTA on high-resource languages

Requirements:

1. A lot of computing power
2. A lot of data (~GB, (Martin et al. 2019)

Language Modeling Framework

Standard Setting

1. Pretraining $p_{\theta_0}(X)$

2. Fine-Tuning $p_{\theta_1}(Y|X, \theta_0)$

3. Evaluation $p_{\theta_1}(Y|X)$

➔ Improves SOTA on high-resource languages

Requirements:

1. A lot of computing power
2. A lot of data (~GB, (Martin et al. 2019))

Research Question

➔ Can we be **more efficient?**

Language Modeling Framework

Standard Setting

Zero-Shot Cross-Lingual Transfer

1. Pretraining

$$p_{\theta_0}(X)$$

2. Fine-Tuning

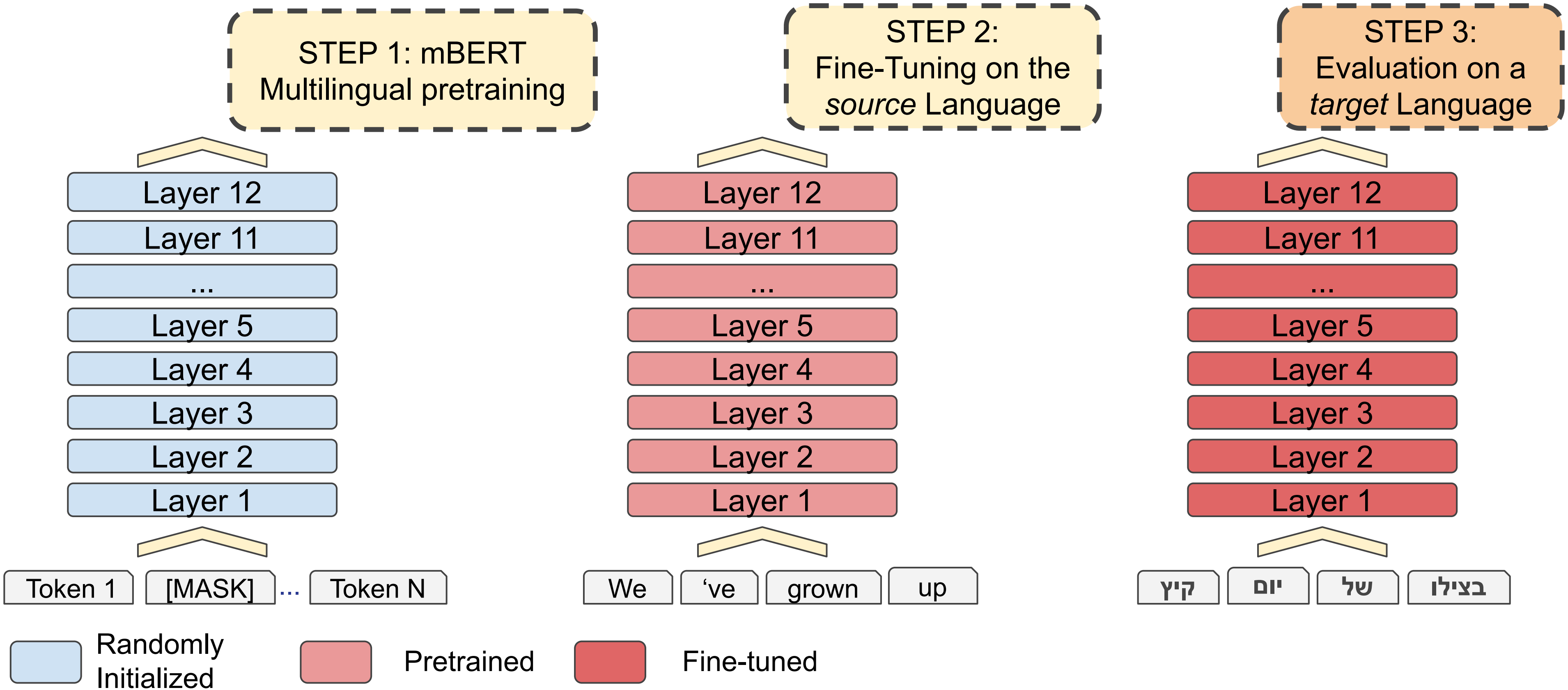
$$p_{\theta_1}(Y|X, \theta_0)$$

3. Evaluation

$$p_{\theta_1}(Y|X)$$

$$p_{\theta_1}(\widetilde{Y}|\widetilde{X})$$

Zero-Shot CL Transfer with mBERT



Zero-Shot Cross-Lingual Transfer with mBERT

mBERT fine-tuned for Parsing **on English**

- Reaches **non-trivial performance** on all target languages
- **This transfer is** surprising because the model was **trained on no annotated data in the target** and **no parallel data**

SOURCE - TARGET	MBERT
same-language performance	
EN - ENGLISH	90.0
cross-lingual performance	
EN - FRENCH	74.0
EN - GERMAN	70.4
EN - RUSSIAN	62.5
⋮	⋮
EN - X (MEAN)	53.2

Table: Dependency Parsing Performance (LAS score) of mBERT fine-tuned on English

Zero-Shot Cross-Lingual Transfer with mBERT

mBERT fine-tuned for Parsing **on English**

- Reaches **non-trivial performance** on all target languages
- **This transfer is surprising** because the model was trained on no annotated data in the target and no parallel data

➔ How does mBERT perform cross-lingual transfer?

SOURCE - TARGET	MBERT
same-language performance	
EN - ENGLISH	90.0
cross-lingual performance	
EN - FRENCH	74.0
EN - GERMAN	70.4
EN - RUSSIAN	62.5
⋮	⋮
EN - X (MEAN)	53.2

Table: Dependency Parsing Performance (LAS score) of mBERT fine-tuned on English

Understanding the Zero-Shot Cross-Lingual (CL) Transfer abilities of mBERT

Understanding the behaviour of Deep-Learning models is inherently difficult (cf. Bertology (Rogers et al. 2020))

Understanding the Zero-Shot Cross-Lingual (CL) Transfer abilities of mBERT

Understanding the behaviour of Deep-Learning models is inherently difficult (cf. Bertology (Rogers et al. 2020))

For Zero-Shot Cross-Lingual (CL) Transfer:

- (Chi et.al 2020) found "universal grammar relations in mBERT" with probing
- (Artexte 2019, Conneau et. al 2020) found emerging cross-lingual structure in monolingual language models
- (Dufter et. al 2020) found that **shared** special tokens (e.g.[MASK]), position vectors and masking are key elements of multilinguality

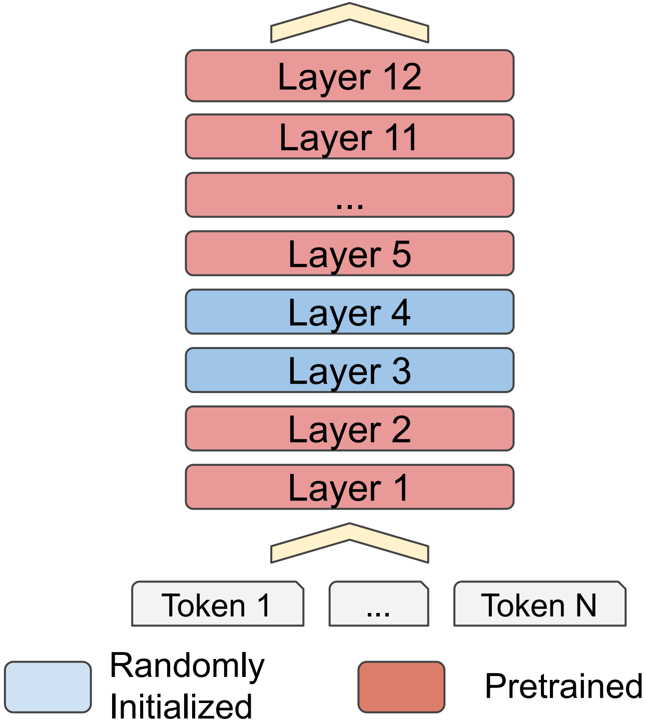
Understanding Zero-Shot Cross-Lingual Transfer abilities of mBERT

1. What layers of the model contribute to zero-shot cross-lingual transfer?

2. What internal mechanisms support it?

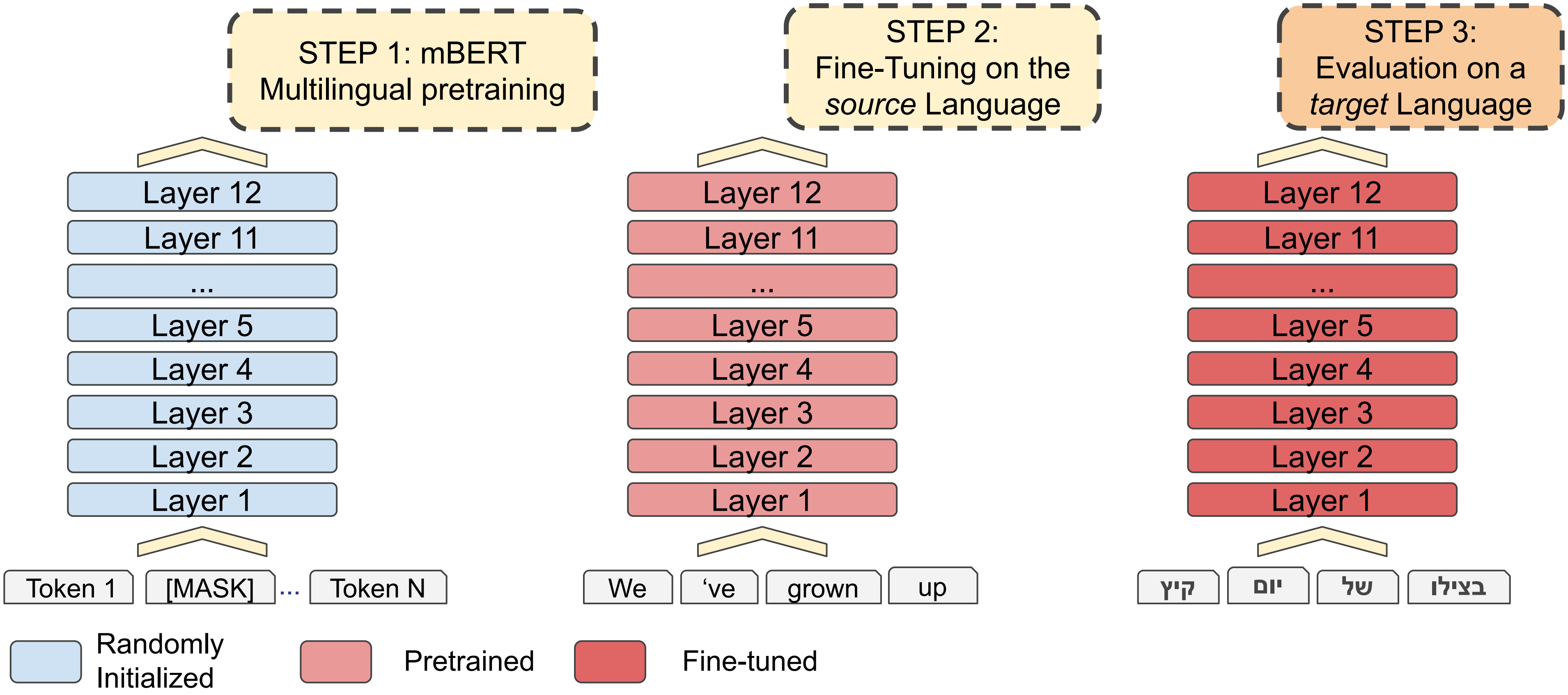
What layers contribute to CL transfer?

We introduce RANDOM-INIT as an ablation technique

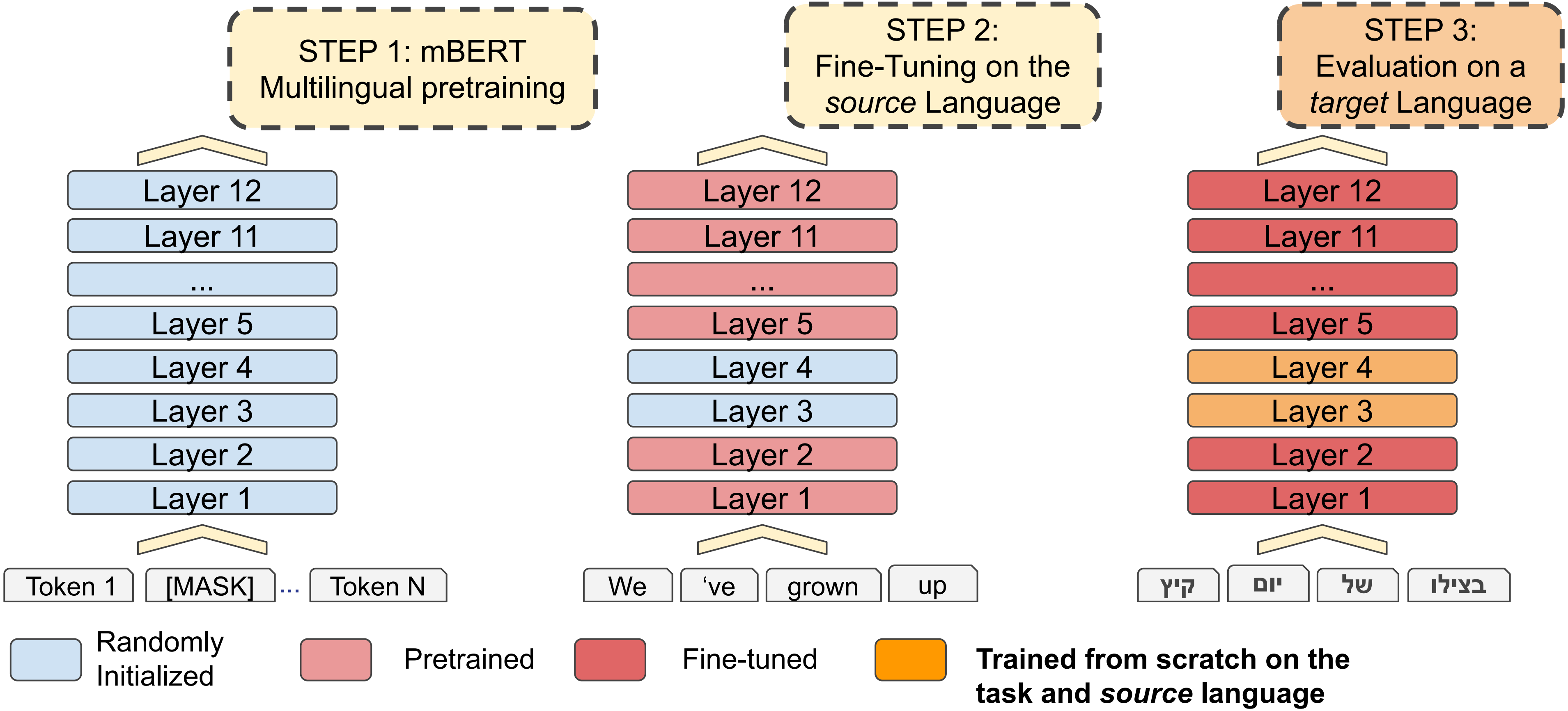


RANDOM-INIT consists of re-initializing selectively pretrained parameters before fine-tuning (e.g. layer 3 and 4)

Zero-Shot CL Transfer with mBERT



RANDOM-INIT to locate layers



What layers contribute to CL transfer?

We apply **RANDOM-INIT** to pairs of consecutive layers...

IF the performance drops in the cross-lingual setting

AND does not drop in the same-language setting....

→ these layers are critical for cross-lingual transfer

What layers contribute to CL transfer?

We apply RANDOM-INIT to **pairs of consecutive layers...**

★ The same-language performance drop is null or small across the entire model

		RANDOM-INIT *of layers*					
SRC-TRG	REF	Δ1-2	Δ3-4	Δ5-6	Δ7-8	Δ9-10	Δ11-12
				Parsing			
EN - EN	88.98	-0.96	-0.66	-0.93	-0.55	0.04	-0.09
RU - RU	85.15	-0.82	-1.38	-1.51	-0.86	-0.29	0.18
AR - AR	59.54	-0.78	-2.14	-1.20	-0.67	-0.27	0.08
EN - X	53.23	-15.77	-6.51	-3.39	-1.47	0.29	1.00
RU - X	55.41	-7.69	-3.71	-3.13	-1.70	0.92	0.94
AR - X	27.97	-4.91	-3.17	-1.48	-1.68	-0.36	-0.14

Table: Performance drop of mBERT fine-tuned for Dependency Parsing (LAS score) after applying RANDOM-INIT to pairs of layers compared to mBERT fine-tuned in a standard way (REF)

What layers contribute to CL transfer?

We apply **RANDOM-INIT** to **pairs of consecutive layers...**

- ★ The same-language performance drop is null or small across the entire model

For cross-lingual performance

- ★ Large drop in performance when RANDOM-INIT is applied to lower layers

SRC-TRG	REF	Δ1-2	Δ3-4	Δ5-6	Δ7-8	Δ9-10	Δ11-12
		RANDOM-INIT *of layers*					
				Parsing			
EN - EN	88.98	-0.96	-0.66	-0.93	-0.55	0.04	-0.09
RU - RU	85.15	-0.82	-1.38	-1.51	-0.86	-0.29	0.18
AR - AR	59.54	-0.78	-2.14	-1.20	-0.67	-0.27	0.08
EN - X	53.23	-15.77	-6.51	-3.39	-1.47	0.29	1.00
RU - X	55.41	-7.69	-3.71	-3.13	1.70	0.92	0.94
AR - X	27.97	-4.91	-3.17	-1.48	1.68	-0.36	-0.14

Table: Performance drop of mBERT fine-tuned for Dependency Parsing (LAS score) after applying RANDOM-INIT to pairs of layers compared to mBERT fine-tuned in a standard way (REF)

What layers contribute to CL transfer?

We apply **RANDOM-INIT** to **pairs of consecutive layers...**

- ★ The same-language performance drop is null or small across the entire model

For cross-lingual performance

- ★ Large drop in performance when RANDOM-INIT is applied to lower layers
- ★ Null or Small drop when RANDOM-INIT is applied to upper layers

SRC-TRG	REF	Δ1-2	Δ3-4	Δ5-6	Δ7-8	Δ9-10	Δ11-12
			RANDOM-INIT *of layers*				
				Parsing			
EN - EN	88.98	-0.96	-0.66	-0.93	-0.55	0.04	-0.09
RU - RU	85.15	-0.82	-1.38	-1.51	-0.86	-0.29	0.18
AR - AR	59.54	-0.78	-2.14	-1.20	-0.67	-0.27	0.08
EN - X	53.23	-15.77	-6.51	-3.39	-1.47	0.29	1.00
RU - X	55.41	-7.69	-3.71	-3.13	-1.70	0.92	0.94
AR - X	27.97	-4.91	-3.17	-1.48	-1.68	-0.36	-0.14

Table: Performance drop of mBERT fine-tuned for Dependency Parsing (LAS score) after applying RANDOM-INIT to pairs of layers compared to mBERT fine-tuned in a standard way (REF)

What layers contribute to CL transfer?

Summary

- ➔ mBERT's lower layers are critical for zero-shot cross-lingual transfer
- ➔ Upper layers can be trained in a task-specific way only without harming cross-lingual transfer

SRC-TRG	REF	Δ1-2	Δ3-4	Δ5-6	Δ7-8	Δ9-10	Δ11-12
			RANDOM-INIT *of layers*				
				Parsing			
EN - EN	88.98	-0.96	-0.66	-0.93	-0.55	0.04	-0.09
RU - RU	85.15	-0.82	-1.38	-1.51	-0.86	-0.29	0.18
AR - AR	59.54	-0.78	-2.14	-1.20	-0.67	-0.27	0.08
EN - X	53.23	-15.77	-6.51	-3.39	-1.47	0.29	1.00
RU - X	55.41	-7.69	-3.71	-3.13	-1.70	0.92	0.94
AR - X	27.97	-4.91	-3.17	-1.48	-1.68	-0.36	-0.14

Table: Performance drop of mBERT fine-tuned for Dependency Parsing (LAS score) after applying RANDOM-INIT to pairs of layers compared to mBERT fine-tuned in a standard way (REF)

What internal mechanisms support CL transfer?

What happens to mBERT's hidden representations to enable this transfer?

- Measure the similarity
- between mBERT embedding of the source language and the target language
- for each layer before and after fine-tuning

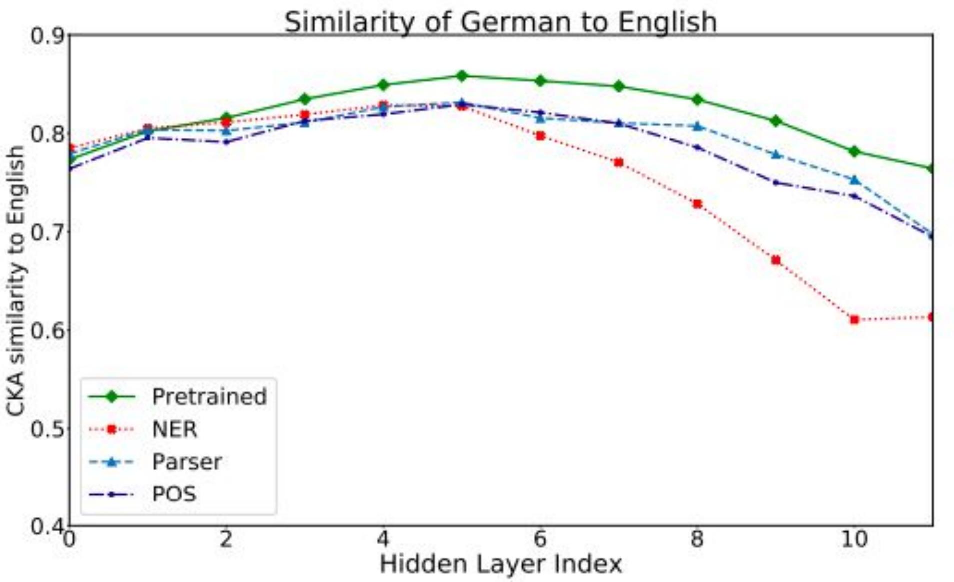


Figure: Cross-Lingual Similarity measured with the Central Kernel Alignment (CKA) between a source language (English) and a target language (German) across mBERT layers

What internal mechanisms support CL transfer?

What happens to mBERT's hidden representations to enable this transfer?

- mBERT aligns representations across languages
- This alignment occurs in the lower part of the model
- This alignment is preserved during fine-tuning

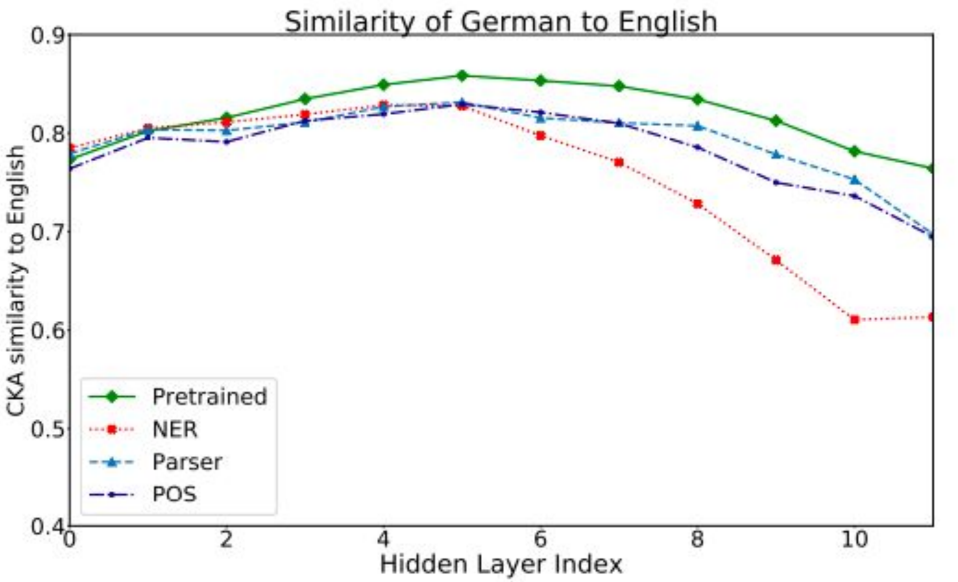


Figure: Cross-Lingual Similarity measured with the Central Kernel Alignment (CKA) between a source language (English) and a target language (German) across mBERT layers

What internal mechanisms support CL transfer?

What happens to mBERT's hidden representations to enable this transfer?

- mBERT aligns representations across languages
- This alignment occurs in the lower part of the model
- This alignment is preserved during fine-tuning
- It correlates strongly with downstream cross-lingual transfer

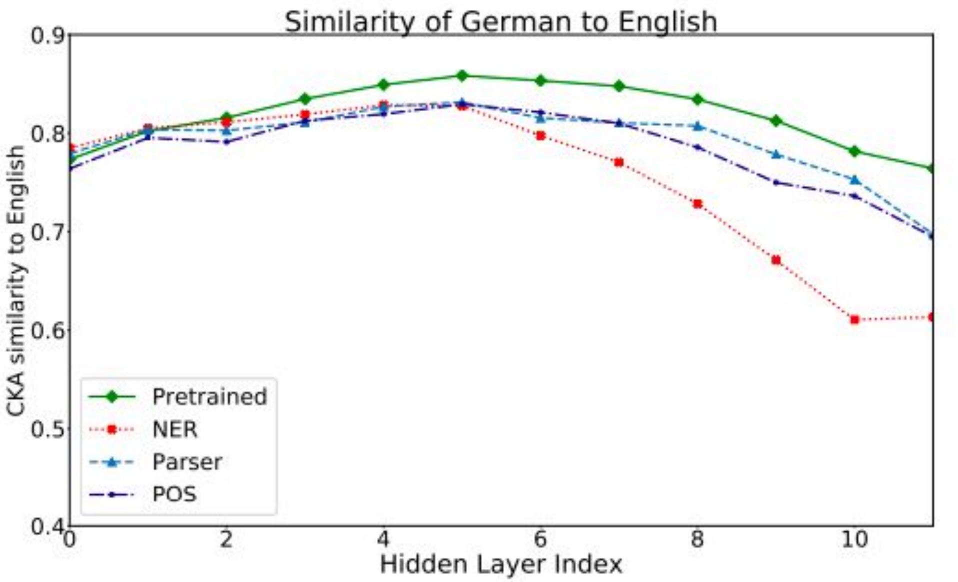


Figure: Cross-Lingual Similarity measured with the Central Kernel Alignment (CKA) between a source language (English) and a target language (German) across mBERT layers

Summary

mBERT is composed of two specific modules:

A Cross-Lingual Encoder **in the lower layers**

- **is critical for cross-lingual** transfer
- **aligns** representations across languages (preserved during fine-tuning)
- correlates strongly with downstream cross-lingual performance

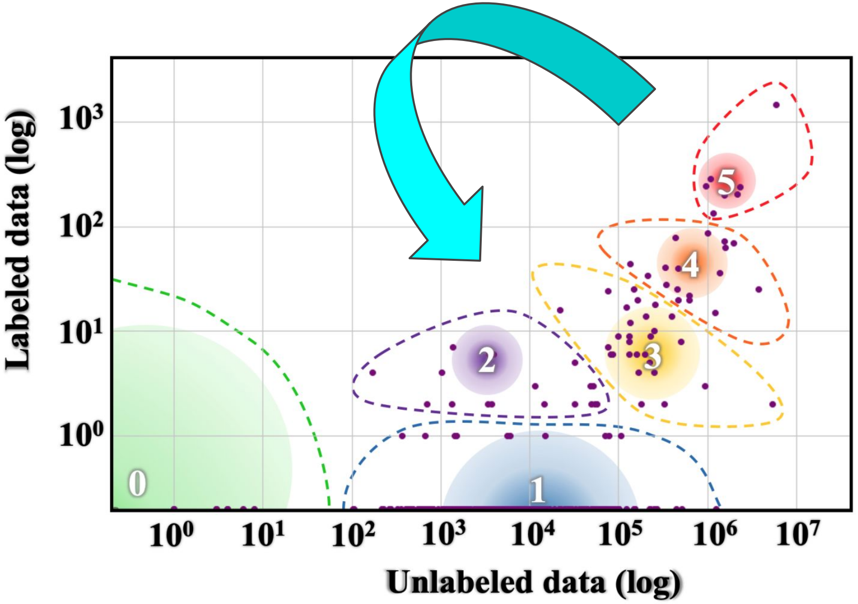
A Task-Specific Predictor **in the upper layers**

- Can be trained from scratch on the source language

Cross-Lingual Fine-Tuning Setting Focusing on Unseen Languages

- Available Language Models (mBERT, XLM-R, mT5) cover about **120 languages**
- Unseen Languages are **languages not seen in the pretraining corpora** of those models
- We focus on Category 2 Languages (small amount of data available)

➔ How can *unseen* languages benefit from CL transfer?



Joshi et. al (2020)

Language Modeling Framework

Supervised Setting

Cross-Lingual Fine-tuning

1. Pretraining

$$p_{\theta_0}(X)$$

2. Fine-Tuning

$$p_{\theta_1}(Y|X,\theta_0)$$

$$p_{\widetilde{\theta}_1}(\widetilde{Y}|\widetilde{X},\theta_0)$$

3. Evaluation

$$p_{\theta_1}(Y|X)$$

$$p_{\widetilde{\theta}_1}(\widetilde{Y}|\widetilde{X})$$

Language Modeling Framework

Supervised Setting

1. Pretraining

$$p_{\theta_0}(X)$$

2. MLM Finetuning

3. Fine-Tuning

$$p_{\theta_1}(Y|X,\theta_0)$$

4. Evaluation

$$p_{\theta_1}(Y|X)$$

Cross-Lingual Fine-tuning with MLM adaptation

$$p_{\widetilde{\theta}_0}(\widetilde{X}|\theta_0)$$

$$p_{\widetilde{\theta}_1}(\widetilde{Y}|\widetilde{X},\widetilde{\theta_0})$$

$$p_{\widetilde{\theta}_1}(\widetilde{Y}|\widetilde{X})$$

What can we do for unseen languages?

- ➔ We build a typology of unseen languages: Easy, Intermediate, Hard
- ➔ Focusing on the Hard languages, we show that the script is a critical element in cross-lingual transfer failure

What can we do for unseen languages?

➔ We build a typology of unseen languages: Easy, Intermediate, Hard

➔ Focusing on the Hard languages, we show that the script is a critical element in cross-lingual transfer failure

Related Work

- (Pfeiffer et al. 2020, 2021) used MLM and task-specific adapters for parameter-efficient CL transfer (MAD-X) or extending script coverage
- (Wang et al. 2021) showed that Ensembling Adapters trained on languages related to the target language improves zero-shot transfer

What can we do for unseen languages?

➔ We build a typology of unseen languages: Easy, Intermediate, Hard

➔ Focusing on the Hard languages, we show that the script is a critical element in cross-lingual transfer failure

Related Work

- (Pfeiffer et al. 2020, 2021) used MLM and task-specific adapters for parameter-efficient CL transfer (MAD-X) or extending script coverage
- (Wang et al. 2021) showed that Ensembling Adapters trained on languages related to the target language improves zero-shot transfer
- (Aepli and Senrich 2022) BPE-drop-out and character-level noise improves transfer between related languages

Unseen Languages

17 typologically diverse unseen languages

Language (iso)	Script	Family	#sents
Faroese (fao)	Latin	North Germanic	297K
Mingrelian (xmf)	Georg.	Kartvelian	29K
Naija (pcm)	Latin	English Pidgin	237K
Swiss German (gsw)	Latin	West Germanic	250K
Bambara (bm)	Latin	Niger-Congo	1K
Wolof (wo)	Latin	Niger-Congo	10K
Narabizi (nrz)	Latin	Semitic*	87K
Maltese (mlt)	Latin	Semitic	50K
Buryat (bxu)	Cyrillic	Mongolic	7K
Mari (mhr)	Cyrillic	Uralic	58K
Erzya (myv)	Cyrillic	Uralic	20K
Livvi (olo)	Latin	Uralic	9.4K
Uyghur (ug)	Arabic	Turkic	105K
Sindhi (sd)	Arabic	Indo-Aryan	375K
Sorani (ckb)	Arabic	Indo-Iranian	380K

We compare mBERT (w. and w/o MLM fine-tuning) with Monolingual Language Model **(MLM)** and **strong BiLSTM Baselines**

The **Three Categories** of Unseen Languages

- **Easy Languages**

If mBERT outperforms the strong BiLSTM baseline, we consider the language **Easy**

The Three Categories of Unseen Languages

- Easy Languages

If mBERT outperforms the strong BiLSTM baseline, we consider the language Easy

- Intermediate Languages

If mBERT does not outperform the strong BiLSTM baselines, but outperforms it after MLM fine-tuning, we consider the Language Intermediate

The Three Categories of Unseen Languages

- Easy Languages

If mBERT outperforms the strong BiLSTM baseline, we consider the language Easy

- Intermediate Languages

If mBERT does not outperform the strong BiLSTM baselines, but outperforms it after MLM fine-tuning, we consider the Language Intermediate

- Hard Languages

If mBERT fails in both settings we consider the language Hard.

Swiss German vs. Uyghur vs. Wolof

Swiss German

- **Latin** script
- Closely Related to German
- Around **500 mb** of available **raw data (OSCAR)**
- **Data** for POS/Parsing
- Native Speakers: ~7 million

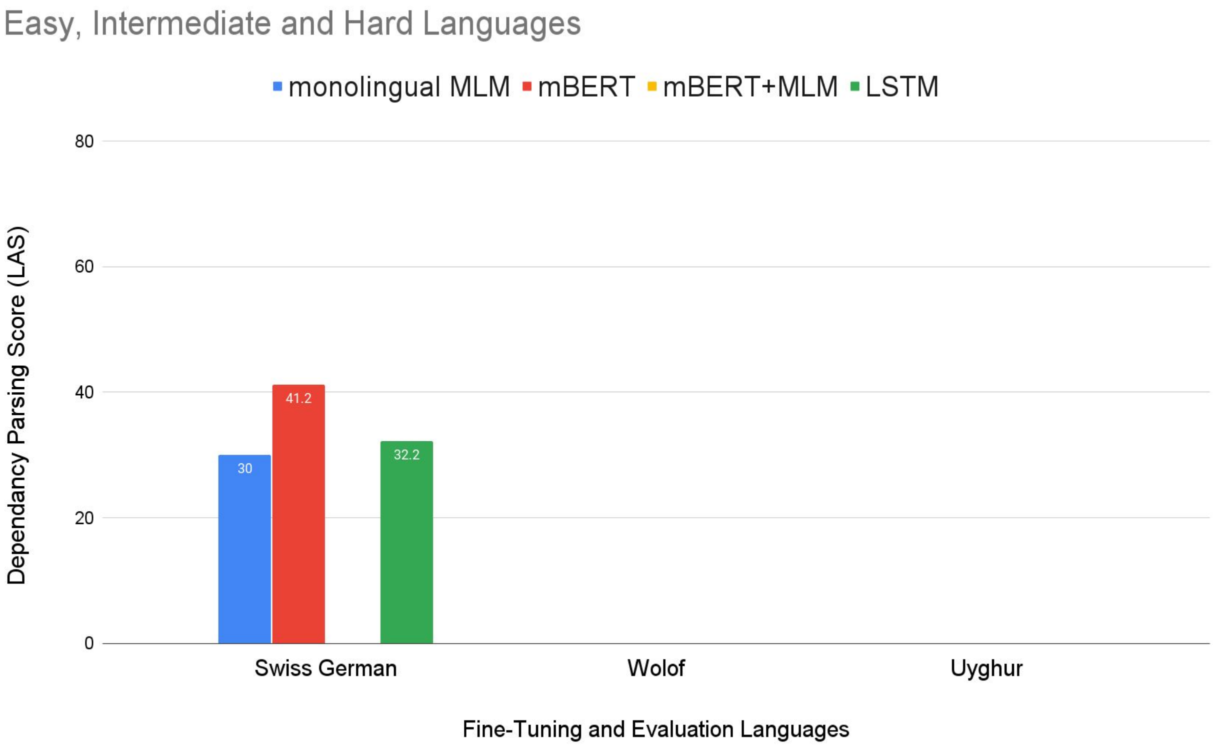
Wolof

- **Latin** script
- Related to Yoruba, Swahili
- Around **2.5 mb** of available **raw data (Wikipedia)**
- **Data** for POS/Parsing
- Native Speakers: ~5 million

Uyghur

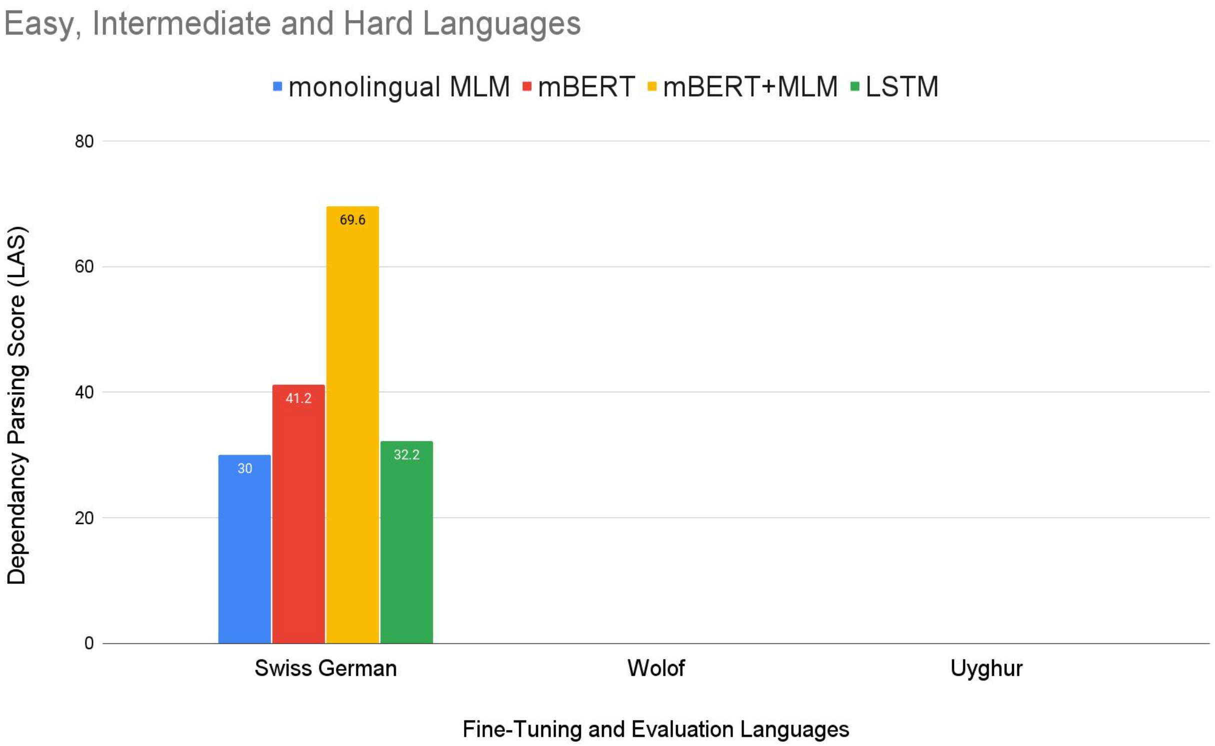
- Arabic script
- Relatively Close to **Turkish**, (written in the latin script)
- Around **100MB** of available **raw data (OSCAR)**
- Data for **POS/Parsing/NER**
- Native Speakers: ~10.4 million

Swiss German vs. Wolof vs. Uyghur



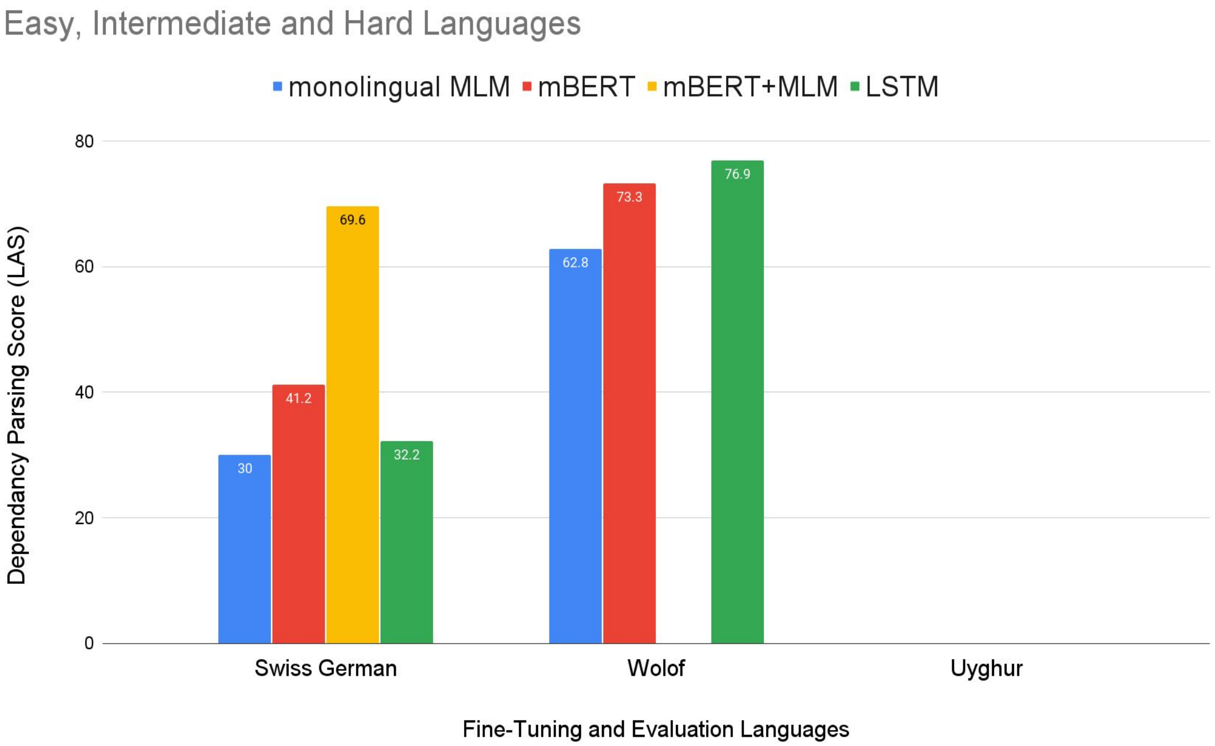
- Swiss German is Easy

Swiss German vs. Wolof vs. Uyghur



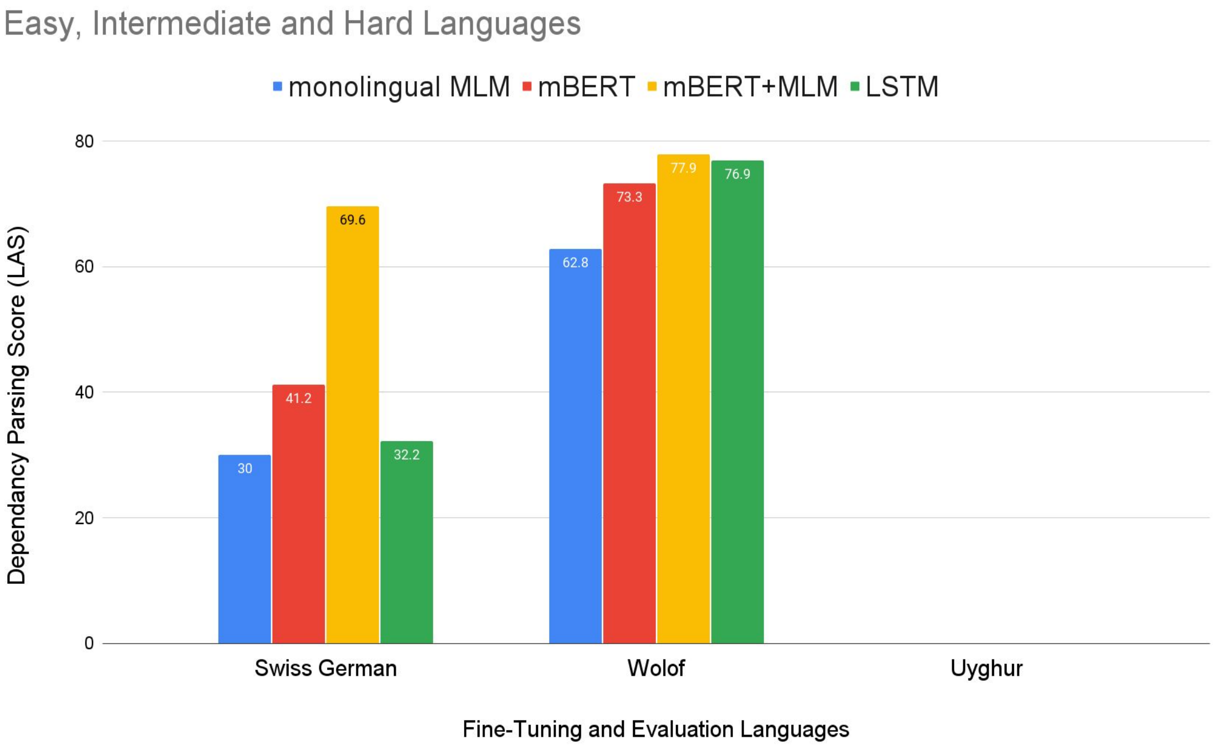
- Swiss German is Easy

Swiss German vs. Wolof vs. Uyghur



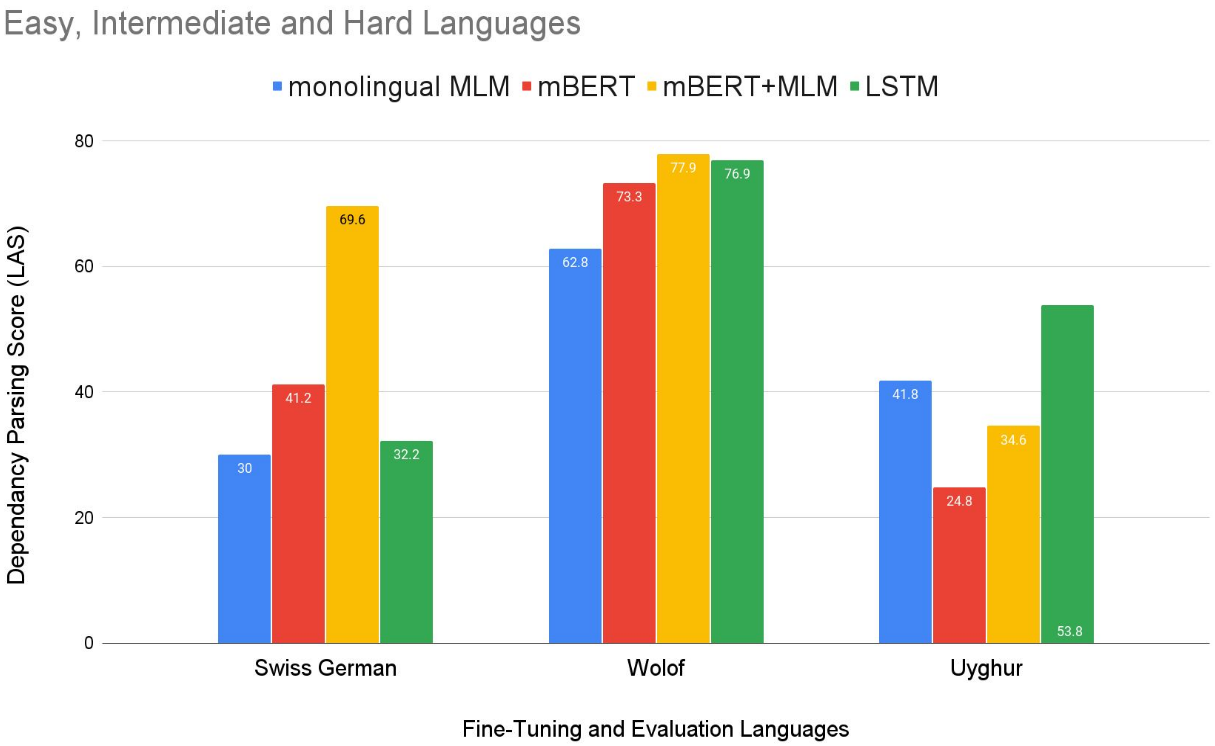
- Swiss German is Easy
- Wolof is Intermediate

Swiss German vs. Wolof vs. Uyghur



- Swiss German is Easy
- Wolof is Intermediate

Swiss German vs. Wolof vs. Uyghur



- Swiss German is Easy
- Wolof is Intermediate
- Uyghur is Hard

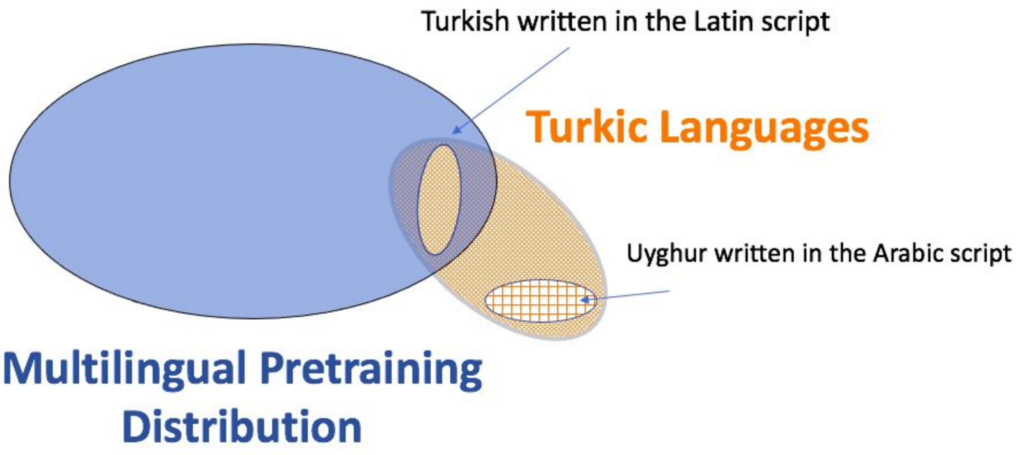
mBERT fails to compete with the baselines (3/17 are Hard)

Why are Hard Languages Hard ?

Why are Hard Languages Hard ?

Hypothesis: mBERT process *unseen* languages by mapping them **to related languages seen during the pretraining.**

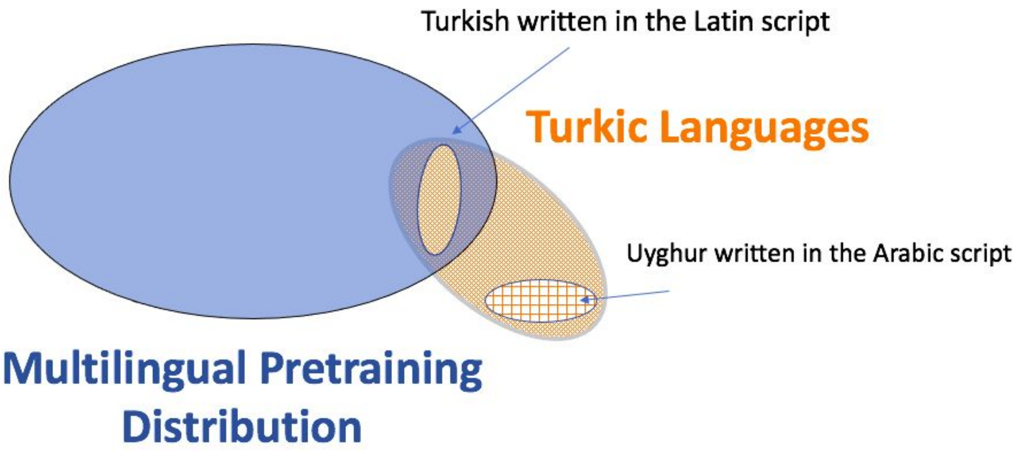
We hypothesize that this 'mapping' is possible only if **the pretraining script is the same as the script of the target language**



Why are Hard Languages Hard ?

Hypothesis: mBERT process *unseen* languages by mapping them **to related languages seen during the pretraining.**

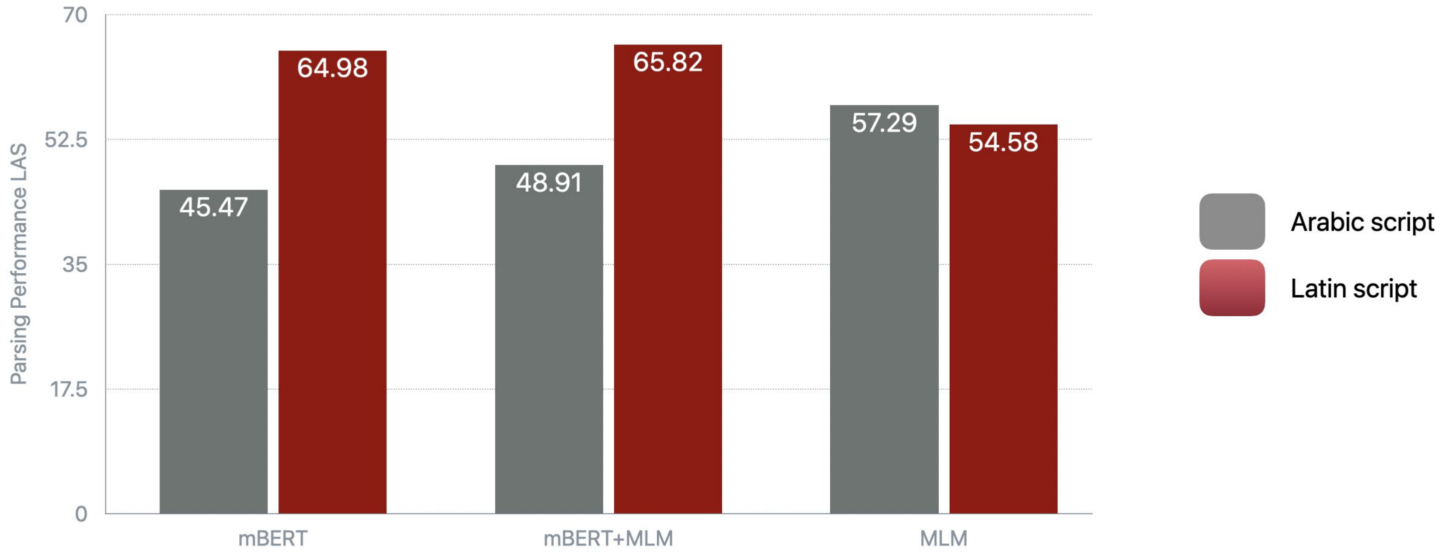
We hypothesize that this 'mapping' is possible only if **the pretraining script is the same as the script of the target language**



➔ Transliteration to control the script and run experiments on transliterated data

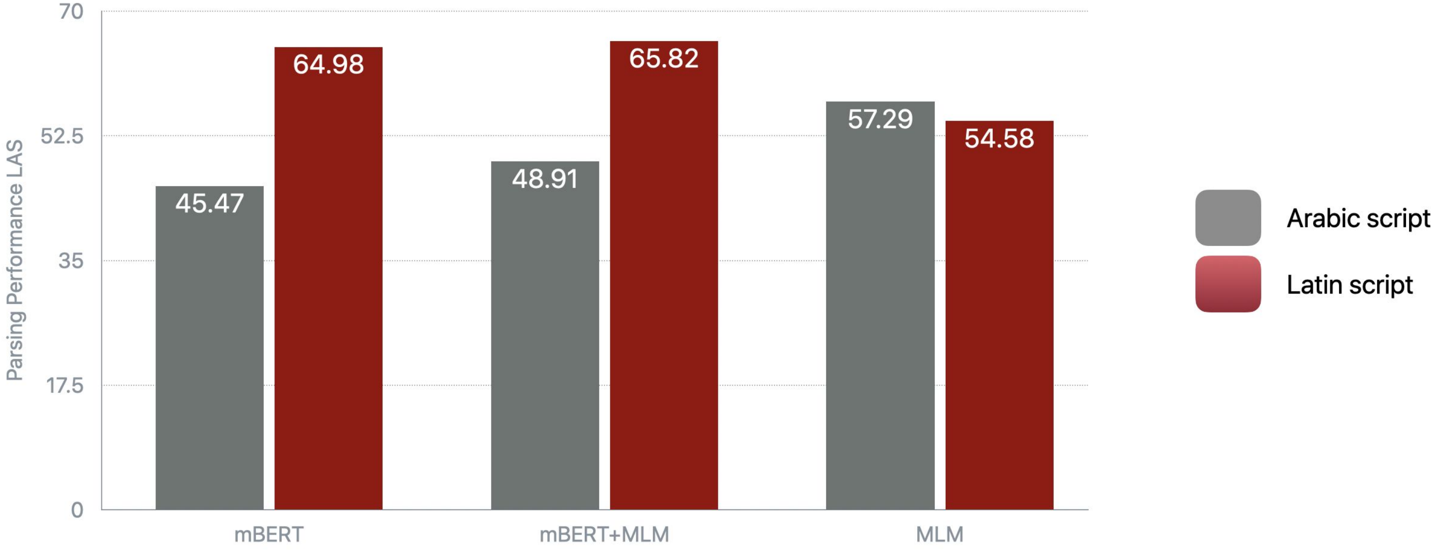
Transliterating Uyghur to the Latin Script

Uyghur LAS Performance: Arabic script vs. Latin Transliteration



Transliterating Uyghur to the Latin Script

Uyghur LAS Performance: Arabic script vs. Latin Transliteration



We validate our hypothesis on Uyghur, Sorani, Mingrelian, Mari, Buryat
As well as on seen languages like Arabic, Russian and Japanese

Takeaways

Languages and Script are not equal in Multilingual Language Models

Languages related to High-Resource Languages written in the same script can successfully be used with Multilingual LMs

For more distant languages written in a different script, transliteration is highly impactful

Conclusion

- Multilingual Language Models enables **efficient cross-lingual transfer**
- They rely on cross-lingual alignment occuring in **the lower layers**
- They are **highly impactful for low-resource languages (with MLM and task-specific fine-tuning)**
- Even for *unseen* languages **with small amount of data available**
- **When they fail,** transliterate to a better suited script

Perspectives for Low-Resource Languages

How can we make further progress?

Scaling the number of parameters

Perspectives for Low-Resource Languages

How can we make further progress?

Scaling the number of parameters

Collecting data for low resource languages

- Real-data requires **better language identification**
- **Generating synthetic data** (e.g. using dictionary)

Perspectives for Low-Resource Languages

How can we make further progress?

Scaling the number of parameters

Collecting data for low resource languages

- Real-data requires **better language identification**
- **Generating synthetic data** (e.g. using dictionary)

Better Pretraining

- **Adapters** as a **modularization framework** for cross-lingual transfer
- **Toward Multi-View models:** **i.e. beyond BPE-only models** (e.g. character and byte-level models, speech and text, image and text)

References

First Align, then Predict: Understanding the Cross-Lingual Ability of Multilingual BERT, Benjamin Muller, Yanai Elazar, Benoît Sagot, Djamé Seddah, *EACL 2021*

When Being Unseen from mBERT is just the Beginning: Handling New Languages With Multilingual Language Models, Benjamin Muller, Antonios Anastasopoulos, Benoît Sagot, Djamé Seddah, *NAACL 2021*

Building a User-Generated Content North-African Arabizi Treebank: Tackling Hell, Djamé Seddah, Farah Essaidi, Amal Fethi, Matthieu Futeral, Benjamin Muller, Pedro Javier Ortiz Suárez, Benoît Sagot, Abhishek Srivastava, *ACL 2020*

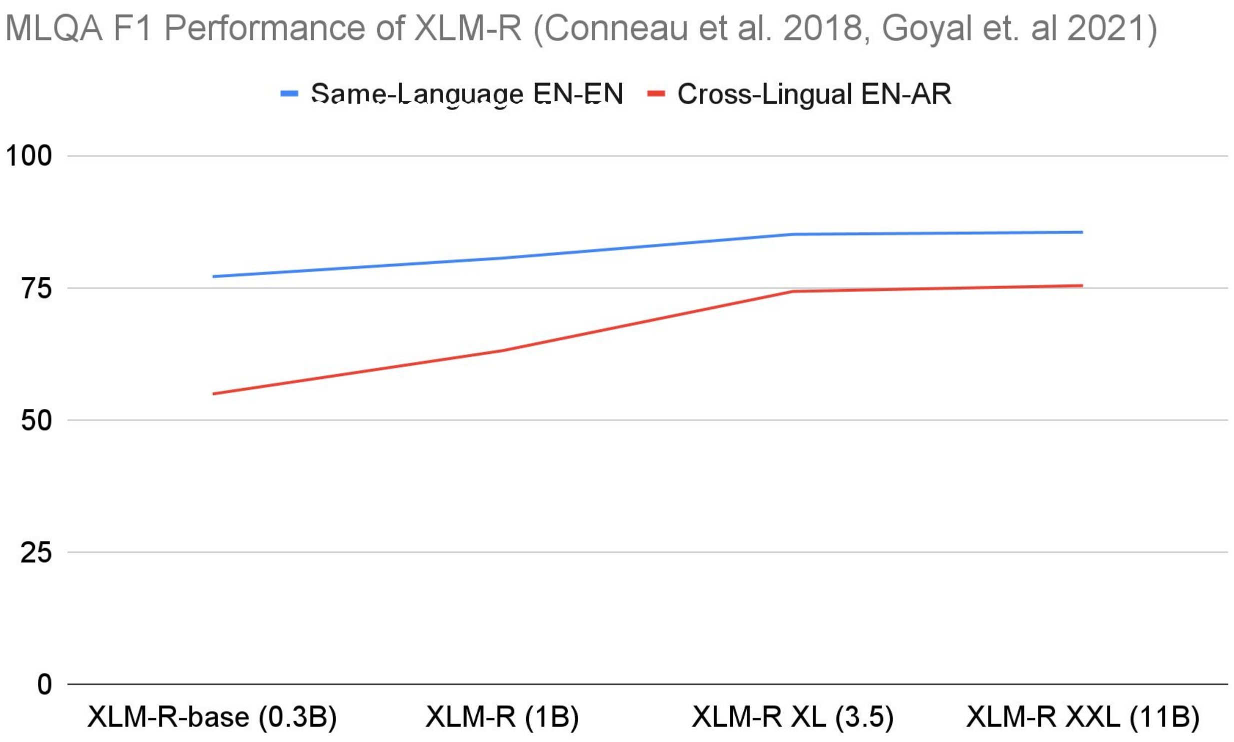
Louis Martin*, Benjamin Muller*, Pedro Javier Ortiz Suárez*, Yoann Dupont, Laurent Romary, Éric de la Clergerie, Djamé Seddah, Benoît Sagot, *ACL 2020*

Thank you!

Perspectives for Low-Resource Languages

How can we make further progress?

Scaling the number of parameters



What internal mechanisms support this transfer?

Correlating cross-lingual similarity with cross-lingual transfer

- ➔ The Cross-Lingual Similarity of mBERT hidden representations correlates strongly with cross-lingual transfer
- ➔ The higher the cross-lingual alignment inside mBERT, the better the cross-lingual transfer

Task	X-Gap vs. Cross-Lingual Similarity
Parsing	0.76
POS	0.74
NER	0.47

Table: Spearman Correlation between Cross-Lingual GAP (X-Gap) and Cross-Lingual Similarity between source and the target languages of mBERT fine-tuned on diverse tasks

What internal mechanisms support this transfer?

Correlating cross-lingual similarity with cross-lingual transfer

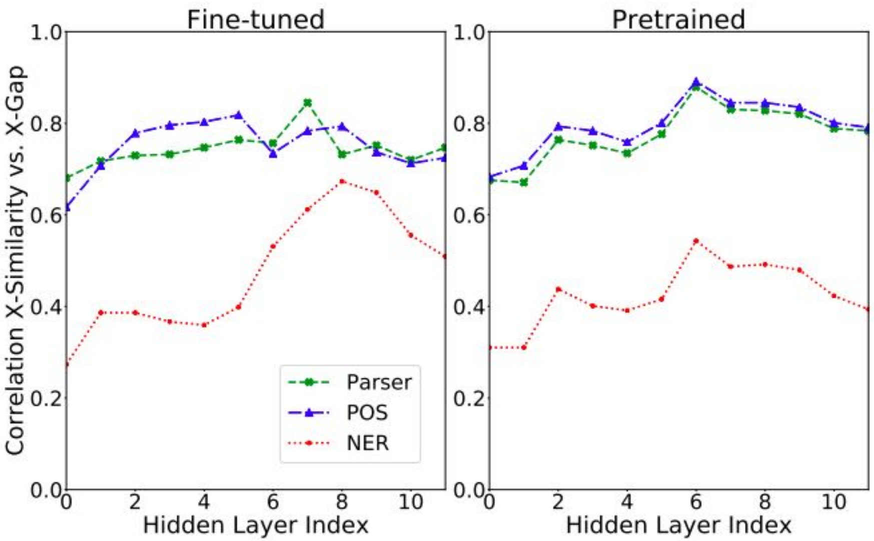


Figure: Spearman Correlation between Cross-Lingual Similarity (CKA between English and the target representations) and cross-lang gap averaged over all 17 target languages for each layer