When Being unseen by mBERT is just the beginning
Handling New Languages With Multilingual Language Models

Benjamin Muller, January 2021, ITU Copenhagen
Joint work with Antonis Anastasopoulos, Benoît Sagot and Djamé Seddah

Inria

Context

Large Scale Multilingual Language Models are now available for the top **100~ highest-resource languages** (mBERT, XLM-R, mT5)

Large Scale Multilingual Language Models can **outperform** **Monolingual** Language models and **reach state-of-the-art** on their **pretraining** languages (Conneau et. al 2020)

Large Scale Multilingual Language Models encodes different **pretraining languages** in **a shared sub-space** (Pires et. al 2019, Chi et. al 2020)

Still, **Large Scale Multilingual Language Models** are limited by the **curse of multilinguality** (Conneau et. al 2020)

Research Question

Low resource languages/dialects

Multilingual pretrained language models (Multilingual BERT, XLM-R, mT5)

Can **Large Scale Multilingual Language Models** improve **NLP for Low-Resource Languages ?**

Outline

1. How to handle **Unseen Languages** with Multilingual Language Models?
2. **The Three Categories** of **Unseen Languages** (Easy, Intermediate, Hard)
3. How to handle **Hard Languages**?

Framework

Given pretrained Multilingual Language Model (e.g. **mBERT**).

We want to use this model on a **target language** that **has not been seen (i.e. unseen)** during **pretraining** (e.g. Swiss German) for a given task (e.g. Parsing).

We assume that we have a sufficient amount of **raw data** and **annotated data** in the **target language.**

How to use Multilingual Models for Unseen Languages ?

- **Fine-tune the model directly on the task** with **annotated data** in the **target** Language

$$\mathrm{X}_i \rightarrow p_{\theta_0}(X|\dot{X})$$

1. **Pretraining**
on a **Multilingual** corpora

2. **Task-Specific fine-tuning**
on the **unseen Target Language**

$$\widetilde{Y_i}, \widetilde{X_i}, \theta_0 \rightarrow p_{\widetilde{\theta_1},\alpha}(\widetilde{Y}|\widetilde{X})$$

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

How to use Multilingual Models for Unseen Languages ?

- Step 1: **Adapt the model in an Unsupervised way** with its Mask-Language Model objective (mBERT+MLM)
- Step 2: **Fine-tune** in a task-specific way

$$\mathrm{X}_i \rightarrow p_{\theta_0}(X|\dot{X})$$

1. **Pretraining** on a **multilingual** corpora (e.g. mBERT)

$$\widetilde{X}_i, \theta_0 \rightarrow p_{\widetilde{\theta}_0}(\widetilde{X}|\dot{\widetilde{X}})$$

2. **Unsupervised Language Adaptation**

3. **Task-Specific fine-tuning** on the **unseen** **Target Language**

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

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

Experiment 1

17 typologically diverse unseen languages

mBERT (trained on 104 languages with Wikipedia data)

Experimenting with **NER** (WikiAnn), **POS** tagging (UD) and **Dependency Parsing** (UD)

Raw Data using Web Crawled Corpus (**OSCAR**) or Wikipedia

Baselines

- **Monolingual Language Model** trained from scratch on the target language
- Strong **non-contextual baselines**: stanza / udpipe 2.0

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

Can mBERT be useful for unseen languages ?

- Does mBERT **outperform** **non-contextual baselines** on such languages?
- Does mBERT **outperform** **non-contextual baselines after unsupervised fine-tuning**?
- Does mBERT **outperform** **monolingual language** models trained from scratch ?

All Languages are not equal: Swiss vs. Uyghur

Swiss German

- **Latin** script
- Closely Related to **German** (high resource language)
- Around **500 mb** of available **raw data**
- **Annotated data** for POS/Parsing

Native Speakers: **~7 million**

Uyghur

- **Arabic** script
- Relatively Close to **Turkish**, a mid-resource language (written in the **latin script**)
- Around **100MB** of available **raw data**
- Annotated for **POS/Parsing/NER**

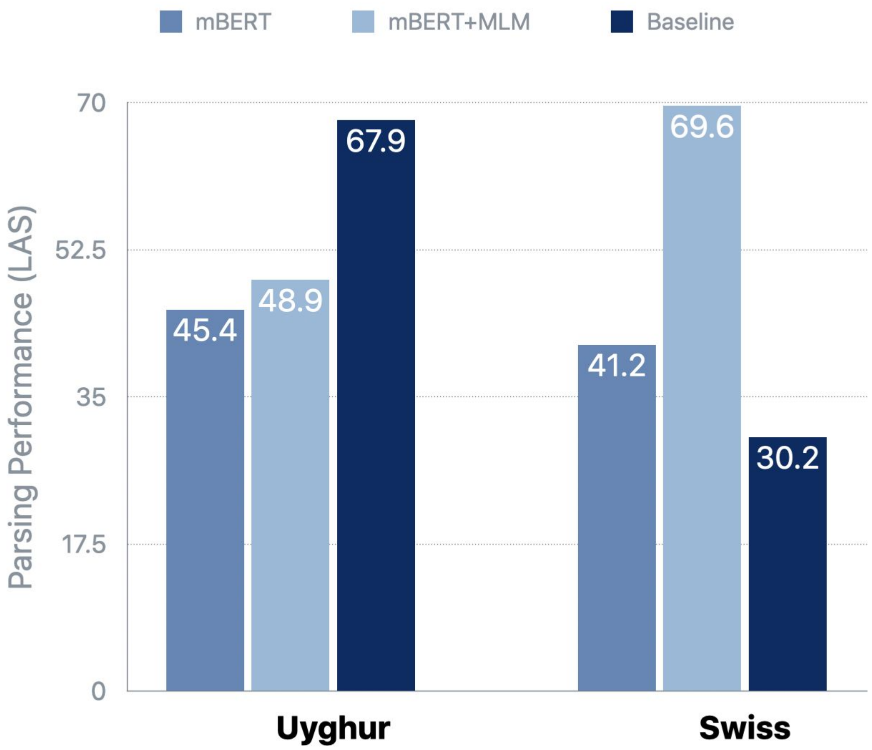
Native Speakers: **~10.4 million**

All Languages are not equal: Swiss vs. Uyghur

Multilingual BERT provides **decent performance** on **Swiss German**

Unsupervised Adaptation leads to **exceeding state-of-the-art performance on Swiss German**

mBERT completely fails on Uyghur even after Unsupervised Adaptation



The **Three Categories** of Unseen Languages

- **Easy Languages**

If mBERT **outperforms the non-contextual baseline**, we consider the language **Easy**

- **Intermediate Languages**

If mBERT does not outperform the non-contextual baselines, but outperforms it **after Unsupervised fine-tuning**, we consider the **Language Intermediate**

- **Hard Languages**

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

Easy Languages

Model	UPOS				LAS				NER			
	mBERT	mBERT+MLM	MLM	Baseline	mBERT	mBERT+MLM	MLM	Baseline	mBERT	mBERT+MLM	MLM	Baseline
Faroese	96.3	96.5	91.1	95.4	84.0	86.4	67.6	83.1	52.1	58.3	39.3	44.8
Naija	89.3	89.6	87.1	89.2	71.5	69.2	63.0	68.3	-	-	-	-
Swiss German	76.7	78.7	65.4	75.2	41.2	69.6	30.0	32.2	-	-	-	-
Mingrelian	-	-	-	-	-	-	-	-	53.6	68.4	42.0	48.2

Table 1: **Easy Languages** POS, Parsing and NER scores comparing mBERT, mBERT+MLM and monolingual MLM to strong non-contextual baselines when trained and evaluated on unseen languages. Baselines are LSTM based models from UDPipe-future (Straka, 2018) for parsing and POS tagging and Stanza (Qi et al., 2020) for NER.

- mBERT **reaches good performance** **out-of-the box** on the Easy Languages
  Easy Languages seem **closely related** to a language that is **in the pretraining** corpora (e.g. Faroese to Icelandic)

Intermediate Languages

Model	UPOS				LAS				NER			
	mBERT	mBERT+MLM	MLM	Baseline	mBERT	mBERT+MLM	MLM	Baseline	mBERT	mBERT+MLM	MLM	Baseline
Maltese	92.0	**96.4**	92.05	96.0	74.4	**82.1**	66.5	79.7	61.2	**66.7**	62.5	63.1
Narabizi	81.6	**84.2**	71.3	**84.2**	56.5	**57.8**	41.8	52.8	-	-	-	-
Bambara	90.2	**92.6**	78.1	92.3	71.8	75.4	46.4	**76.2**	-	-	-	
Wolof	92.8	**95.2**	88.4	94.1	73.3	**77.9**	62.8	77.0	-	-	-	-
Erzya	89.3	**91.2**	84.4	91.1	61.2	**66.6**	47.8	65.1	-	-	-	-
Livvi	83.0	**85.5**	81.1	84.1	36.3	**42.3**	35.2	40.1	-	-	-	-
Mari	-	-	-	-	-	-	-	-	55.2	**57.6**	44.0	56.1

Table 2: **Intermediate Languages** POS, Parsing and NER scores comparing mBERT, mBERT+MLM and monolingual MLM to strong non-contextual baselines when trained and evaluated on unseen languages.

mBERT **highly benefits from Unsupervised Adaptation** leading to efficiently process those languages

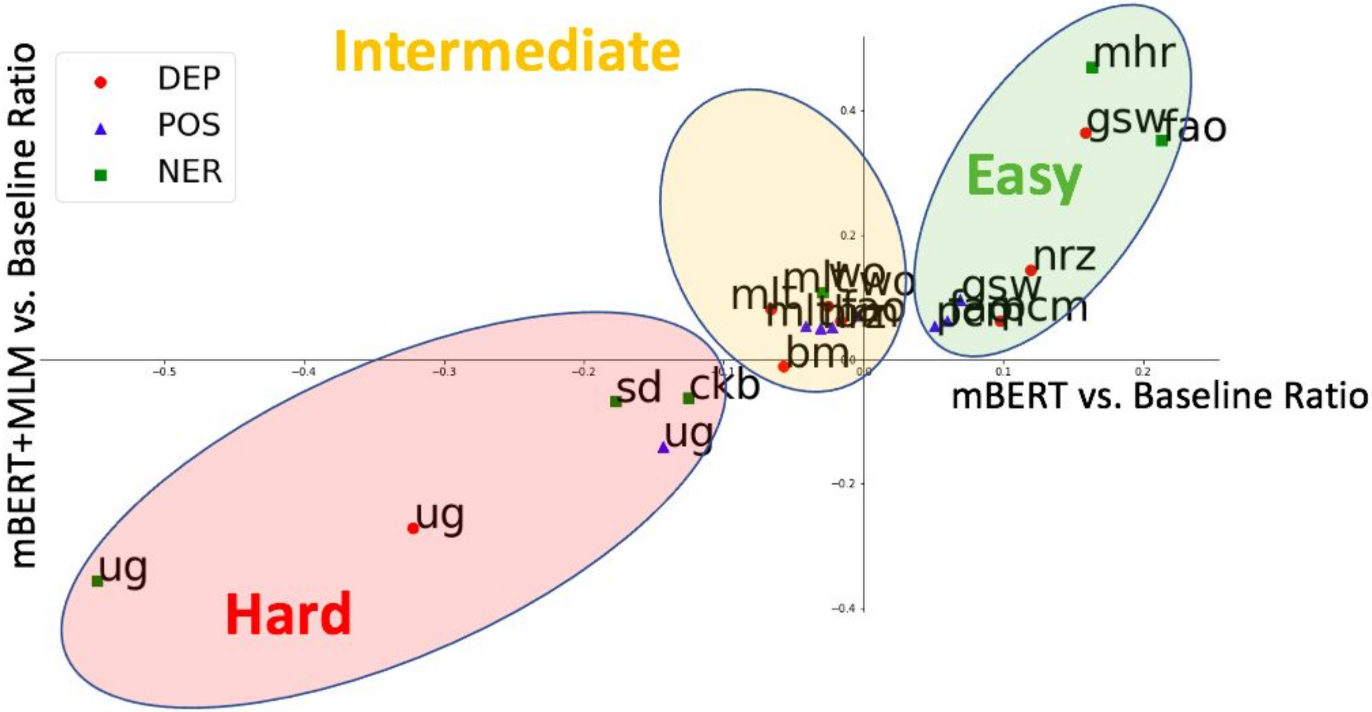
Hard Languages

Model	UPOS				LAS				NER			
	mBERT	mBERT+MLM	MLM	Baseline	mBERT	mBERT+MLM	MLM	Baseline	mBERT	mBERT+MLM	MLM	Baseline
Uyghur	77.0	88.4	87.4	**90.0**	45.5	48.9	57.3	**67.9**	24.3	34.6	41.4	**53.8**
Sindhi	-	-	-	-	-	-	-	-	42.3	47.9	45.2	**51.4**
Sorani Kurdish	-	-	-	-	-	-	-	-	70.4	75.6	80.6	**80.5**

Table 3: **Hard Languages** POS, Parsing and NER scores comparing mBERT, mBERT+MLM and monolingual MLM to strong non-contextual baselines when trained and evaluated on unseen languages.

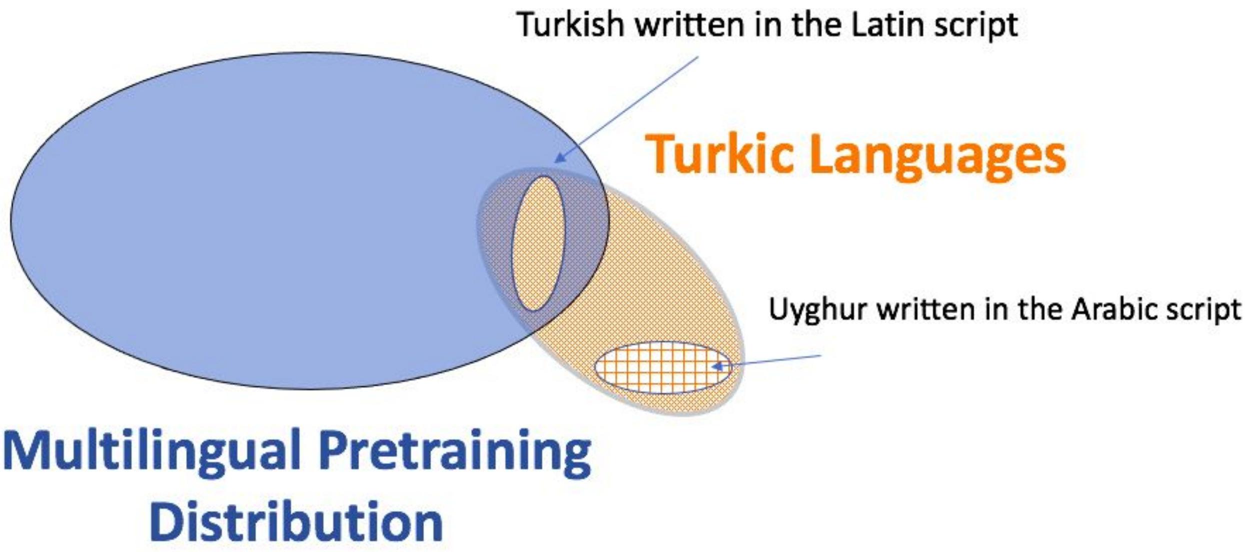
On Hard Languages, mBERT fails completely
mBERT even **outperformed by monolingual language** model trained on very **small corpora**

The **Three Categories** of Unseen Languages



Why are **Hard** Languages **Hard** ?

Hypothesis: mBERT process *unseen* languages by mapping them to pretrained related languages. **We hypothesize** that this 'mapping' is possible only if **the pretraining script is consistent with the script of the target language**



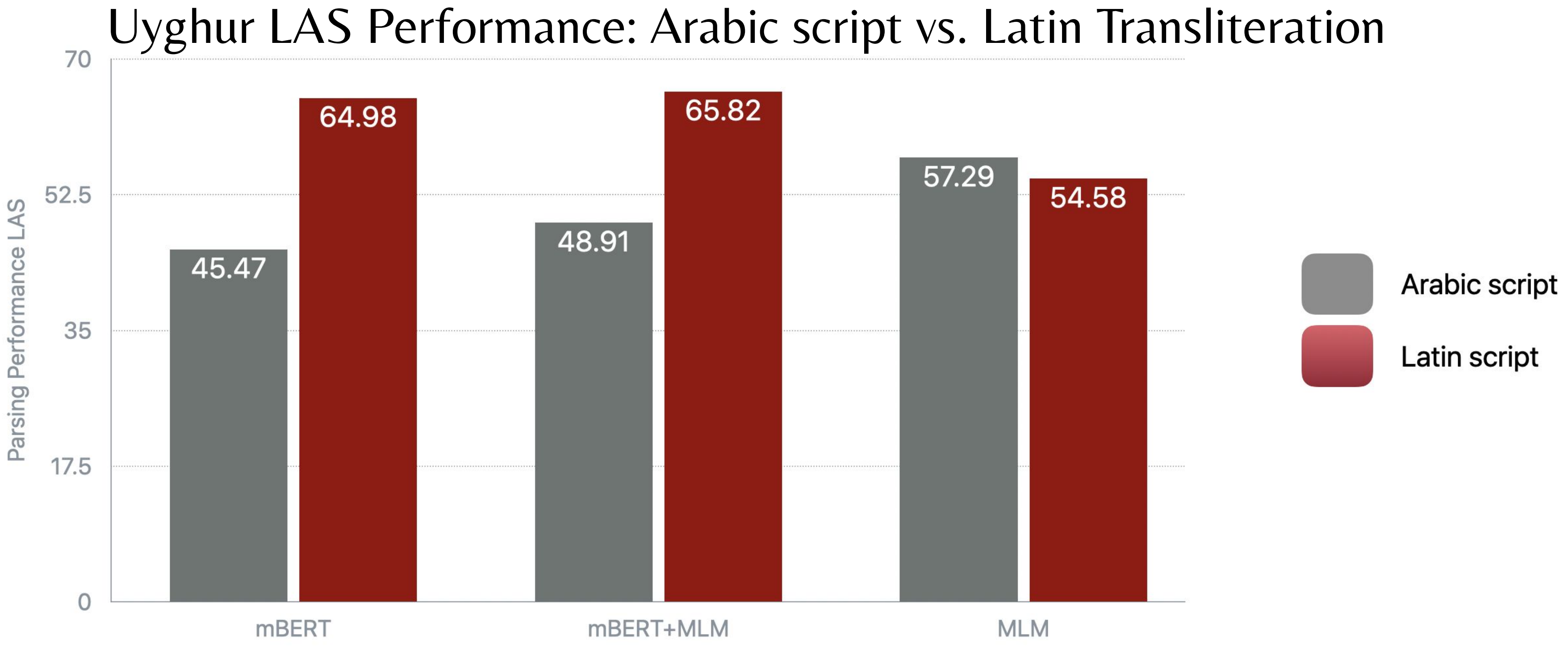
Experiment 2

1. **Transliterate** the target language
2. Run **task-fine-tuning and unsupervised fine-tuning** on the transliterated data
3. Evaluate using the transliterated data

Controlled Experiment

- Transliterate languages that are **in the pretraining corpora (e.g. Arabic)**
- **Transliterate unseen languages** to a script that does not match the pretraining corpora related languages (transliterate Mingrelian to the Latin script)

Transliterating Uyghur to the Latin Script



In the **pretraining-fine-tuning** framework, **script matters (a lot!)**

Does the script matter ?

Model	POS	LAS	NER	Model	NER
	Uyghur (Arabic→Latin)			Sorani (Arabic→Latin)	
UyghurBERT	87.4→86.2	57.3→54.6	41.4→41.7	SoraniBERT	80.6→78.9
mBERT	77.0→87.9	45.7→65.0	24.3→35.7	mBERT	70.5→77.8
mBERT+MLM	77.3→**89.8**	48.9→**66.8**	34.7→**55.2**	mBERT+MLM	75.6→**82.7**
	Buryat (Cyrillic→Latin)			Meadow Mari (Cyrillic→Latin)	
BuryatBERT	75.8→75.8	31.4→31.4	–	MariBERT	44.0→45.5
mBERT	83.9→81.6	50.3→45.8	–	mBERT	55.2→58.2
mBERT+MLM	**86.5**→84.6	**52.9**→51.9	–	mBERT+MLM	57.6→**65.9**
	Erzya (Cyrillic→Latin)			Mingrelian (Georgian→Latin)	
ErzyaBERT	84.4→84.5	47.8→47.8	–	MingrelianBERT	42.0→42.2
mBERT	89.3→88.2	61.2→58.3	–	mBERT	53.6→41.8
mBERT+MLM	**91.2**→90.5	**66.6**→65.5	–	mBERT+MLM	**68.4**→62.6

Transliterating to the Latin Script helps improve the performance for Sorani, Uyghur, and Mari

Transliteration **degrades** significantly for **Mingrelian** (Kartvelian family)

Is mBERT better in processing the Latin script ?

Model	Original Script → Latin Script POS	LAS	NER
Arabic	96.4 → 94.9	82.9 → 78.8	87.8 → 80.9
Russian	98.1 → 96.0	88.4 → 84.5	88.1 → 86.0
Japanese	97.4 → 95.7	88.5 → 86.9	61.5 → 55.6

Transliterating **Arabic, Russian and Japanese** to **the Latin script** **degrades** the performance for all tasks

This shows that **the Latin script is not inherently easier** for mBERT

Takeaways

Languages and Script are not born equal in a Multilingual Language Models

Languages closely related to High-Resource Languages written in the **same script** can **successfully** be used with Multilingual Language Models

For more **distant languages** written **in a different script**, **transliteration** is highly impactful and **unlock the power of Multilingual Models**

Open Questions

How could we make **multilingual language models** **abstract away** from the scripts they are pretrained on ?

Could transliteration help us **design better pretraining procedure** **for Multilingual Language Models** ?

Thanks!

Bibliography

How multilingual is Multilingual BERT? **[Pires et. al 2019]**

Unsupervised Cross-lingual Representation Learning at Scale **[Conneau et. al 2020]**

Finding Universal Grammatical Relations in Multilingual BERT **[Chi et. al 2020]**

On the importance of pre-training data volume for compact language models **[Micheli et. al 2020]**

Parsing with Multilingual BERT, a Small Corpus, and a Small Treebank **[Chau et. al 2020]**

First Align Then Predict, Understanding the Cross-Lingual Ability of Multilingual BERT **[Muller et. al 2020]**