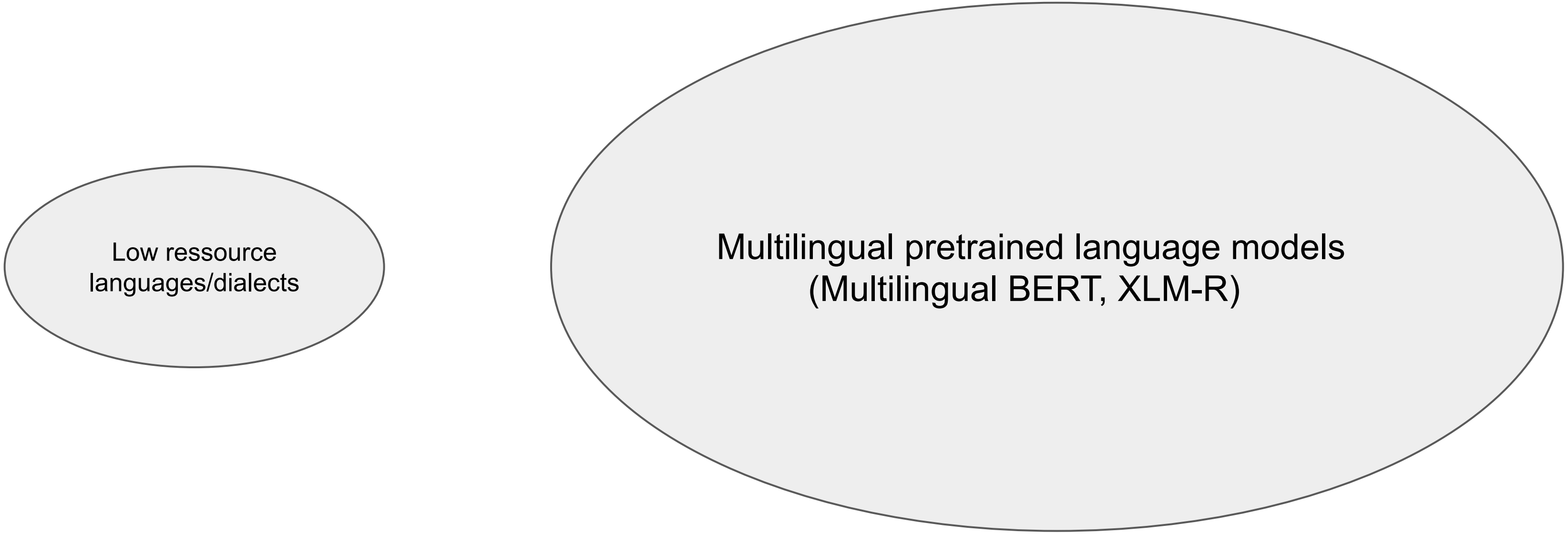
Transfer Learning on an unseen North-African Arabic Dialect

Benjamin Muller
PhD Student ALMANACH INRIA, Paris
joint work with Djamé Seddah and Benoit Sagot

December 2019, Bar-Ilan University

Research Question

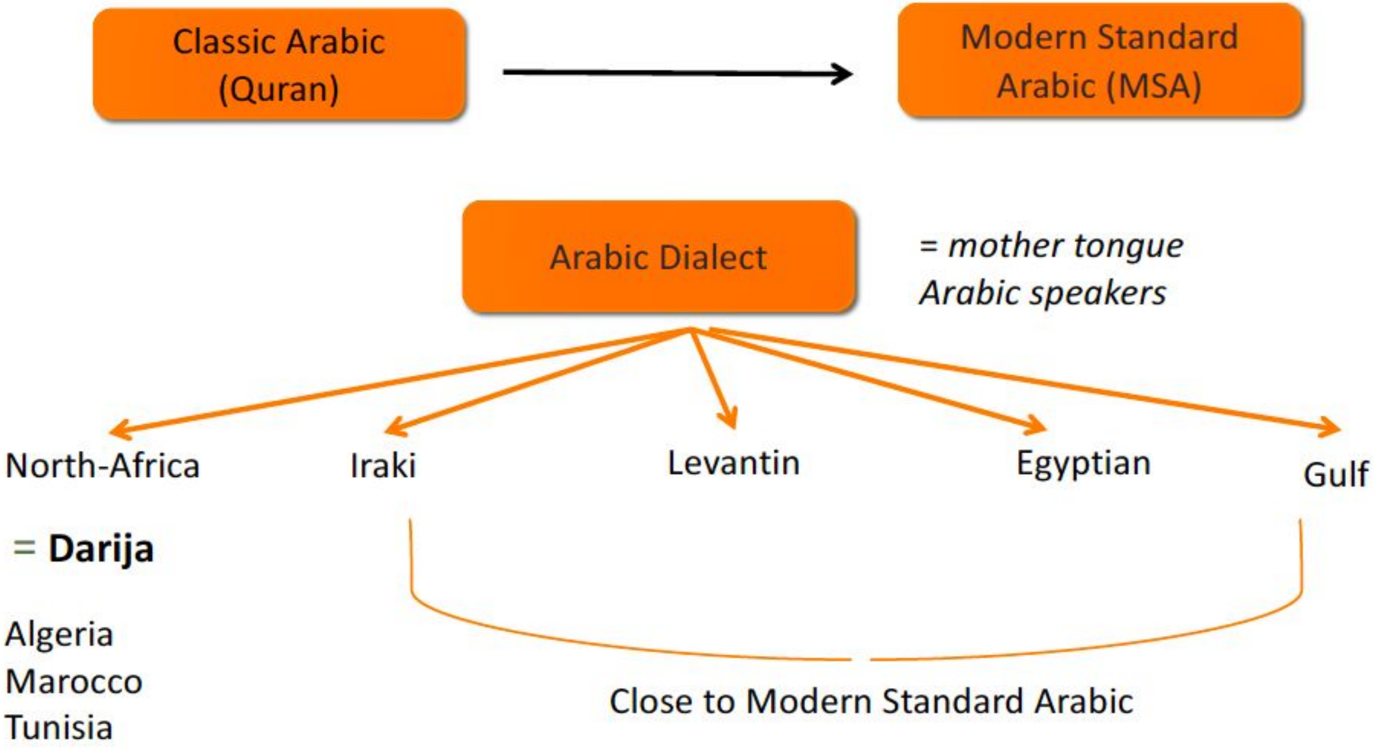


Can we use **multilingual large scale pretrained language models** to improve **low resource NLP ?**

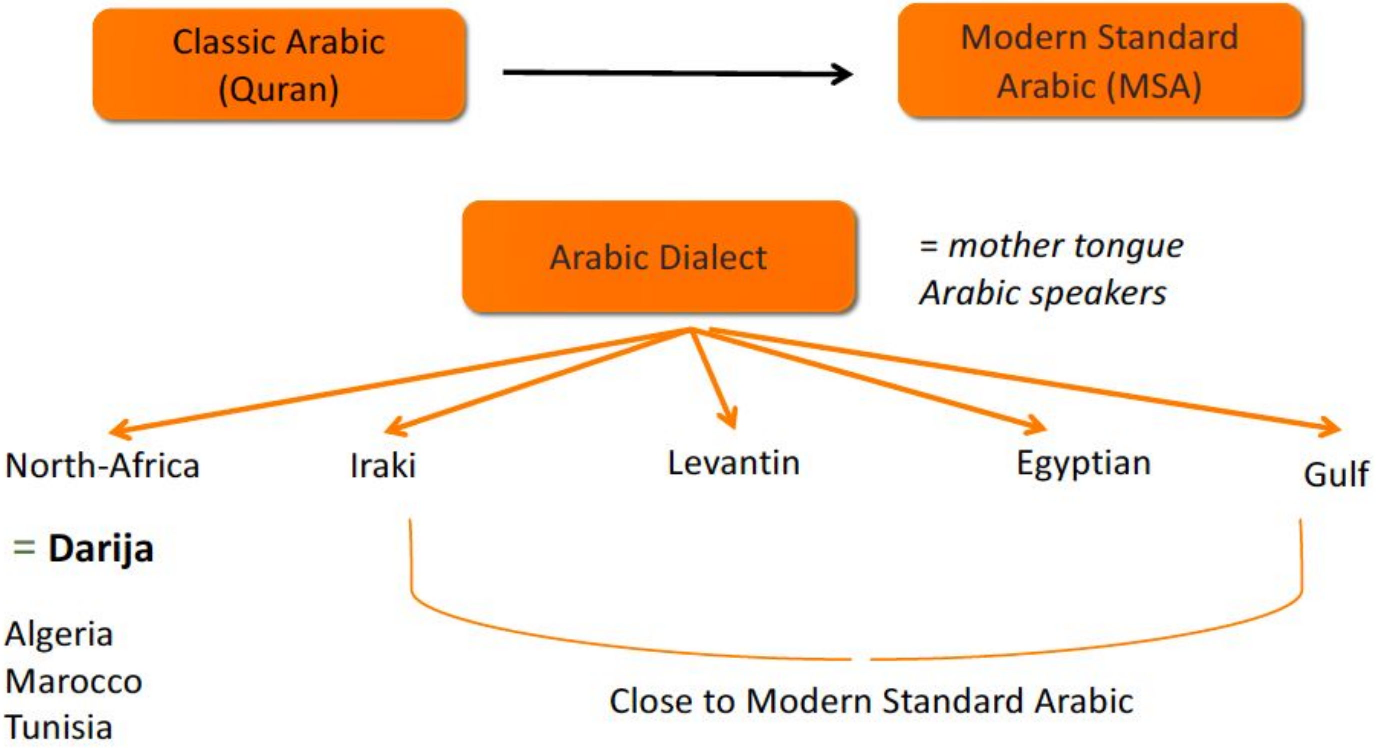
Outline

1. North African Arabic Dialect : Arabizi
2. Transfer Learning using pre-trained language model
3. Part-of-Speech Tagging on Arabizi with Multilingual BERT
    - Scenario 1 : Low resource
    - Scenario 2 : Zero-shot Cross Lingual
    - Scenario 3 : Unsupervised Adaptation

North African Arabic Romanized dialect : **Arabizi**



North African Arabic Romanized dialect : **Arabizi**



+ **Maltese** : descended from Siculo-Arabic a Maghrebi Arabic dialect

North African Arabic Romanized dialect : **Arabizi**

Definition

Spontaneous transliteration and non-normalized that emerged to ease internet communication between Arabic speakers

"Mrhba, Ana **3**rbi mn dzaye

مرحبا,أنا عربي من الجزائر *Hi, I'm arabic from Algeria*

Arabizi properties

1. Diacritics sign replaced by vowels a i e u or y (unlike MSA)
2. Use of digit numbers to cope with Arabic letters absent from latin alphabet (ascii actually)
3. Code mixed with European languages (mostly French)
4. No norms (spontaneous usage), high degree of variability between arabizi speakers

E.g : **Why : wa3lach w3alh 3alach 3lache** (Arabizi)

All : ekl kal kolach koulli kol (Arabizi)

Many : beaucoup boucoup bcp (French)

Transfer Learning

- Machine Learning core problem

I.I.d assumption :

$$X_i, Y_i \rightarrow p_\theta(Y|X)$$

- Transfer Learning core problem

Given $\widetilde{Y}, \widetilde{X} \neq Y, X$ $\quad p_\theta(\widetilde{Y}|\widetilde{X})$ $\quad \widetilde{Y}, \widetilde{X}$ other domain, language, task...

➔ What performance can we expect from p_θ and why ?

➔ How to do better ?

Transfer Learning with language models

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

Pretraining

In-Domain Task specific Fine tuning

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

Transfer Learning with language models

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

Pretraining

In-Domain Task specific Fine tuning

$$Y_i, X_i, \theta_0 \rightarrow p_{\theta_{1,\alpha}}(Y|X)$$

$$p_{\theta_{1,\alpha}}(\widetilde{Y}|\widetilde{X})$$

Cross Domain/Language evaluation

Transfer Learning with language models

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

Pretraining

In-Domain Task specific Fine tuning

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

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

Cross Domain/Language evaluation

Domain/Language Task Specific Fine tuning

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

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

Transfer Learning with language models

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

Pretraining

$$\widetilde{X_i}, \theta_0 \rightarrow p_{\widetilde{\theta_0}}(\widetilde{X}|\dot{\widetilde{X}})$$

Unsupervised adaptation

In-Domain Task specific Fine tuning

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

Domain/Language Task Specific Fine tuning

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

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

Domain/Language Specific Evaluation

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

Transfer Learning with language models

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

Pretraining

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

In-Domain Task specific Fine tuning

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

Unsupervised adaptation

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

Domain/Language Task Specific Fine tuning

By analysing those results

➔ Can we design fine-tuning strategies for out of domain low resource language ?

➔ Can we gain insights on what's (not) captured by the language model at each step of the training process ?

So Far...

- Arabizi : a highly variable and low resource romanized arabic dialect
- Transfer Learning based on Language modelling

How does a pre-trained language model perform an unseen Arabizi ?

➔ Zero-shot transfer
➔ Low ressource
➔ Unsupervised adaptation

How does the French code-mixing impact those performances ?

Arabizi Datasets

- Extraction from Common Crawl using language identifier based on fastext (linear models on bag of n-grams)
- UD-Like treebank enhanced with glose, and **word level language id (**⅓ French tokens)

Treebank + Raw Dataset (Data from Upcoming release) :

Data	Sent	Tokens
Gold-Standard Code-Mixed	9,372	203,386
Annotated	1,434	22,465
Crawled+CommonCrawl Extract	49,523	1,729,411

Table 1: Summary of Dataset (In annotated, 1,172 sentences for training, 146 for validation and 148 for test)

Multilingual Bert on Arabizi

Multilingual BERT (mBERT)

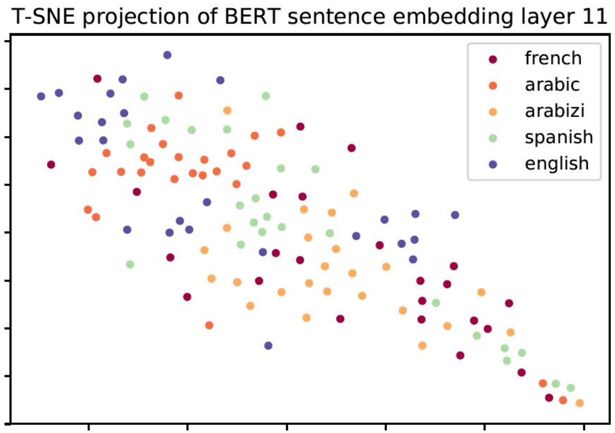
- Trained on concatenation of wikipedia corpus on 104 languages [Delvin et al. 2019]
- Include MS Arabic and French
- Do not include Arabizi or any related dialects (even distant one like Maltese)
- Word-Pieces allows to handle any sequence of alphanumeric characters

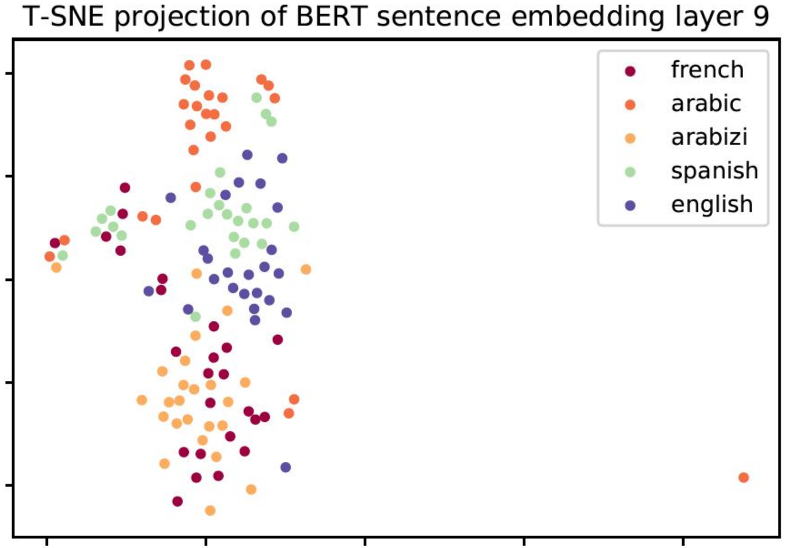
Recent Work

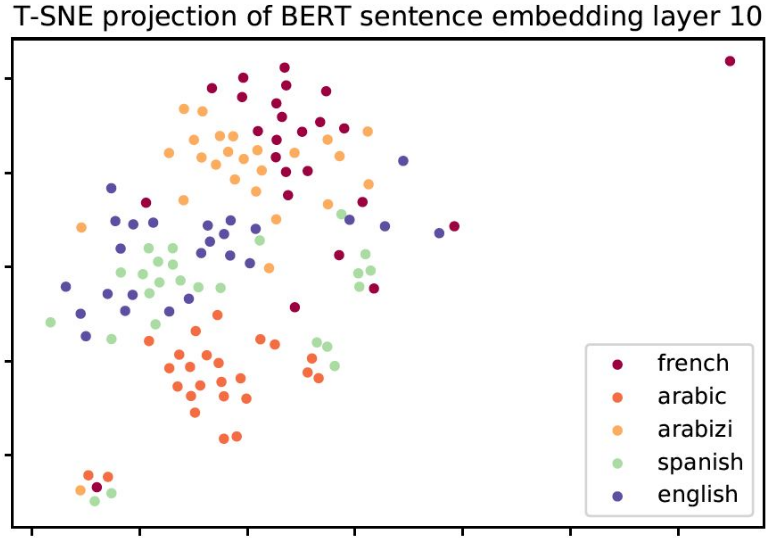
- *How multilingual is multilingual BERT ?* [T.Pires et al. 2019]
- *Unsupervised Domain Adaptation of Contextualized Embeddings for Sequence Labeling ?* [Han Jacob Eisenstein 2019]

mBERT on Arabizi : 1st clue

- Layer embedding of sentences sampled from Arabizi French, Arabic, English, Spanish







mBERT on Arabizi : 2nd clue

Accuracy Masked Language Models

Data/Model	mBERT
Arabizi	26.4
Wiki French	43.05

Table 1: Impact of unsupervised fine tuning on Mask Language Modelling top-1 accuracy. Arabizi measured on Test set of Arabizi treebank. Wikipedia measured on 5000 sentences sampled from french Wikipedia

- For unseen dialect it is decent performance
- Now how does it perform on a specific task ?

Experiments : 3 Transfer scenarios

Focusing on **Part-of-Speech Tagging**

- **Scenario 1 : Low Ressource**

mBERT ⇒ Unsupervised Adaptation $BERT_{AZ}$ ⇒ POS Fine-tuned on Arabizi ⇒ Arabizi Test

mBERT ⇒ POS Fine-tuned on Arabizi ⇒ Arabizi Test

- **Scenario 2 : Zero-shot Cross-Lingual Transfer**

mBERT ⇒ POS Fine-tuned on Source Language ⇒ Arabizi Test

- **Scenario 3 : Unsupervised Adaptation**

mBERT ⇒ Unsup. Adaptation $BERT_{AZ}$ ⇒ POS Fine-tuned on Source Language ⇒ Arabizi Test

Baselines

- **Majority class + PUNCT** prediction
- **StanfordNLP** (State-of-the-art) in POS tagging with French Fastext word embeddings [P Qi et al 2018]
- **Udpipe 1.0** (non neuronal baseline)

➔ How does mBERT perform compare to other models ?

- Random Transformer
- Random Transformer trained on 50k sentences as a Masked-Language Model

➔ What’s the value of pretraining on 104 languages in processing Arabizi ?

Fine-tuning details

- Part-of-Speech Fine tuning
  - Control annotated dataset size ⇒ Training on 1500~ annotated sentences in all scenarios
  - Fine-tuning with Adam for up to 10 epochs

- Unsupervised Fine-tuning (following [Han Jacob Eisenstein 2019])
  - Fine-tuning using dynamic Masked Language Modelling Objective (50k sentence)
  - Learning rate Linear Warm Up 10% of training and Linear Decay

Experiments : Low Resource Scenario

Scenario	Model	Training data	Accuracy UPOS	Δ to mBERT
Low Resource	mBERT	Arabizi	81.15	-
Low Resource	mBERT_{AZ}	Arabizi	83.36	+2.21
Low Resource	MLM 50k	Arabizi	73.33	-7.82
Low Resource	Random Transformer	Arabizi	63.42	-17.73
Low Resource	StanfordNLP	Arabizi	**84.20**	+3.05
Low Resource	Udpipe	Arabizi	73.35	-7.80

Table 3: Low Resource : training on Arabizi train set (1172 sentences) evaluated on test set (148)

- ➔ 1500 sentences on Arabizi Treebank leads to useable POS tags
- ➔ mBERT is 2.21 above "SOTA" model and 7 points above baseline
- ➔ +17.7 compare to random transformer
- ➔ Unsupervised Adaptation leads to significant +2.21 gain

Experiments : Zero Shot Cross Lingual Transfer

Scenario	Model	Training data	Accuracy UPOS
Zero shot CL transfer	mBERT	French	**39.11**
Zero shot CL transfer	mBERT	MS Arabic	16.55
”	mBERT	Maltese	36.11
”	mBERT	Vietnamese	16.92
Cross Lingual transfer	StanfordNLP	French	31.90
Cross Lingual transfer	Random Transformer	French	30.29
Bottom line	Majority class + Punct	Arabizi	20.49

Table 4: Zero shot cross lingual transfer

- Unsurprisingly French is the one that transfer the best to Arabizi
- No transfer happening across scripts [Pires et al. 2017 confirmed]
- Pretraining on 104 languages account for +7.7 points in FR→Arabizi transfer

Experiments : Unsupervised Adaptation

Scenario	Model	Training data	Accuracy UPOS
Zero shot CL transfer	mBERT	French	39.11
Adaptation (50k raw sentences)	mBERT_{AZ}	French	**50.99**
Adaptation (50k raw sentences)	MLM	French	36.23
Cross Lingual transfer	StanfordNLP	French	31.90
Bottom line	Majority class + Punct	Arabizi	20.49

Table 5: Unsupervised Adaption to Arabizi

➔ Unsupervised Adaptation leads to +10 points (confirming [Han Jacob Eisenstein 2019] on Arabizi)

So Far again...

Results Summary

- Low Ressource
  - POS tags useable on Arabizi
  - mBERT leads to competitive POS tagger and unsupervised fine-tuning help
- Cross lingual
  - Transfer happen in same script languages but doesn't across scripts
  - French leads to the best transfer
  - Unsupervised Adaptation impacts a lot cross lingual transfer

➔ How much does code-mixing matters in those performances ?

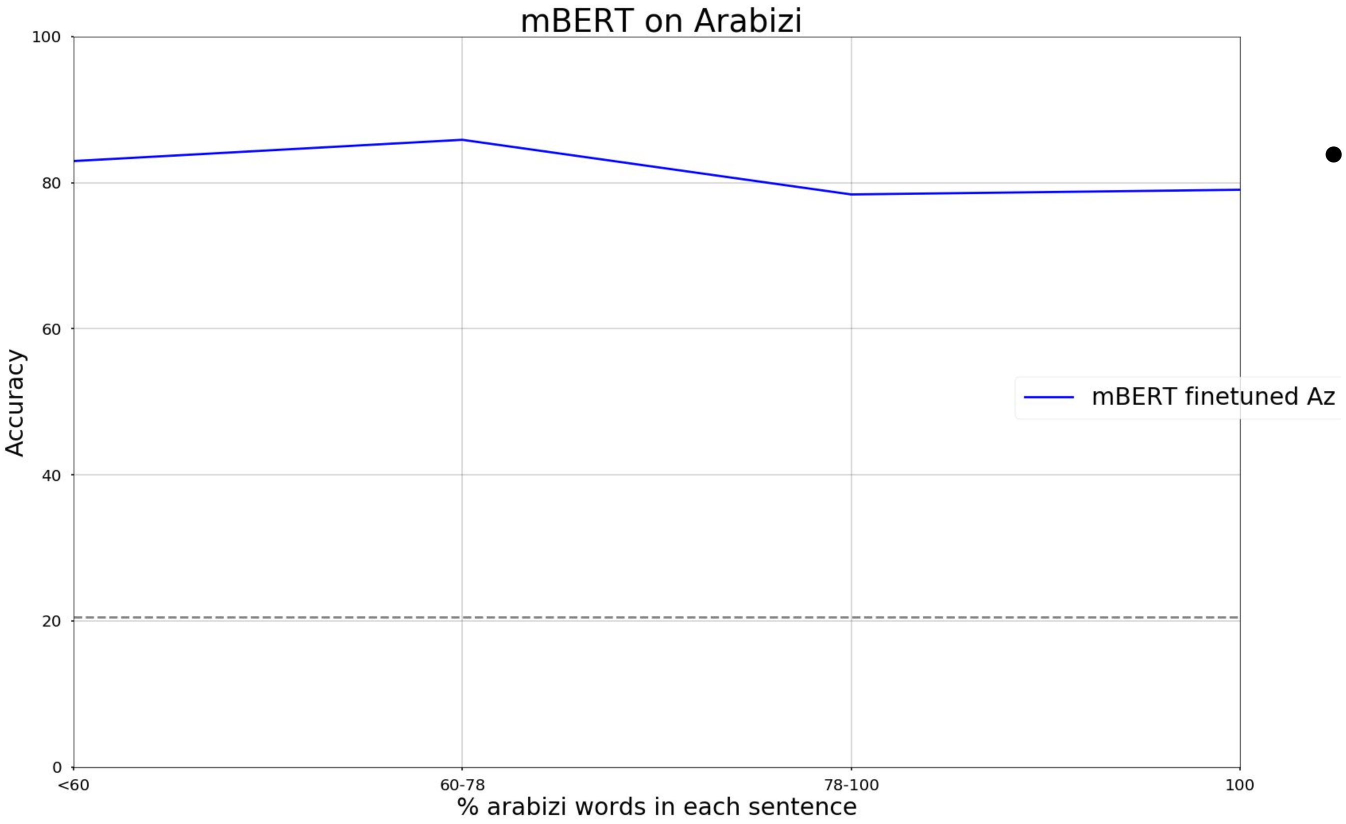
Code-mixed buckets

Proportion Arabizi % of word in sentence	60-	60-78	78-100	=100
train set number sents	322	286	283	276
test set number sents	39	38	34	36

Table 7: Code-mixed Buckets

- Split each set in buckets of controlled proportion of code-mixing (around 25%) (using the word language identifier)
- Except Low resource scenario evaluated on the Train for having more data

Experiments : Low Resource

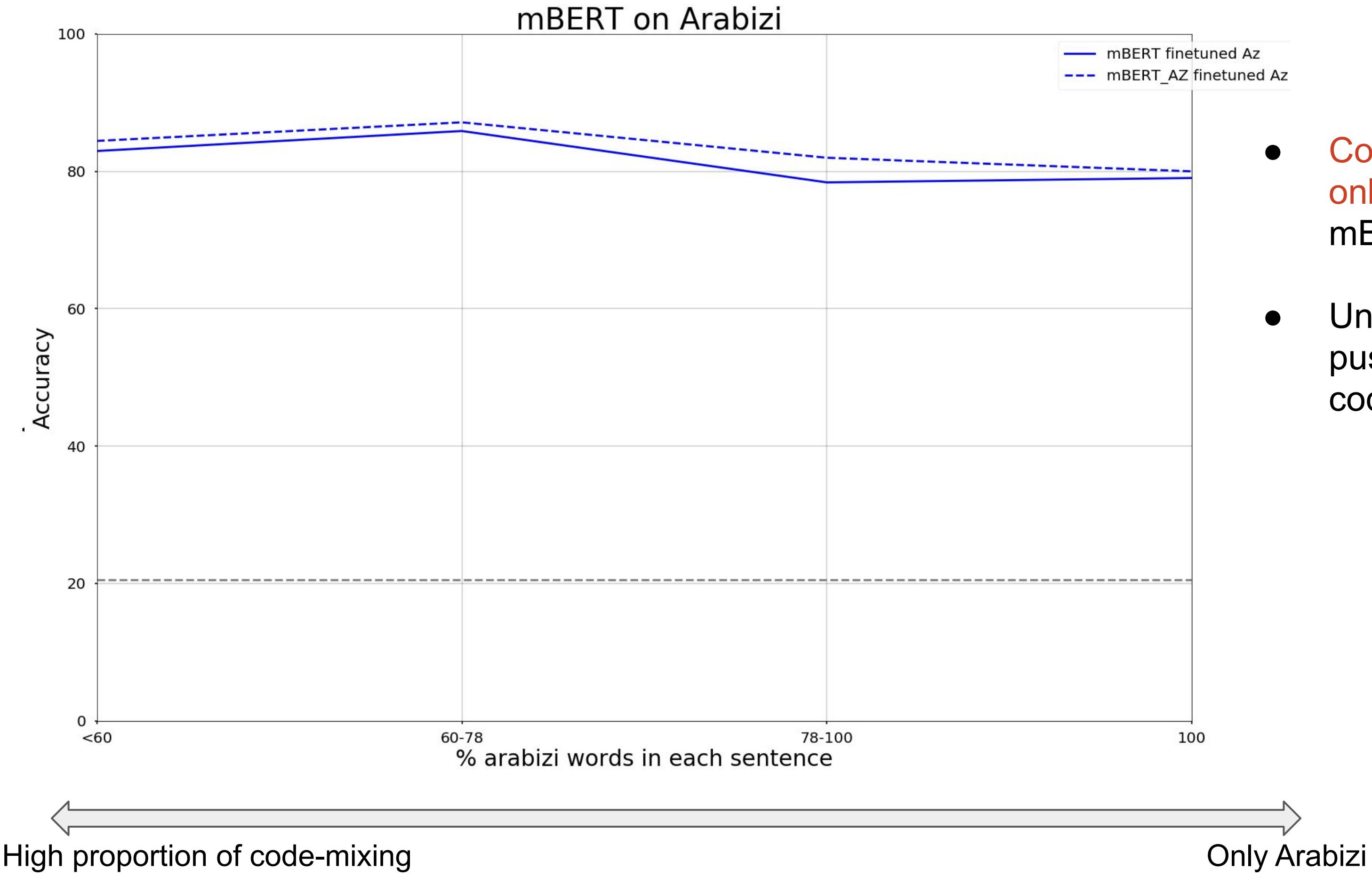


High proportion of code-mixing

Only Arabizi

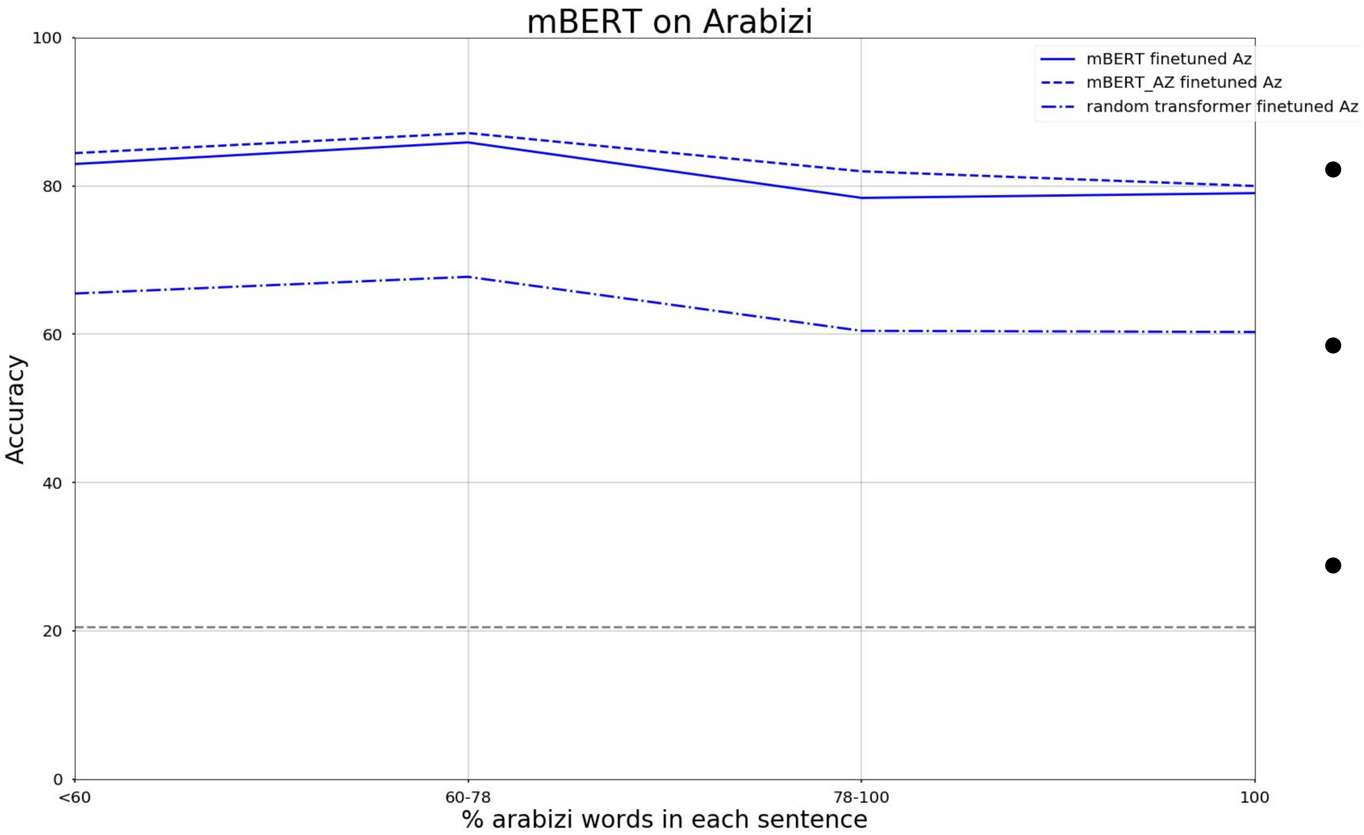
- Code-Mixing is not the only factor that explains mBERT success

Experiments : Low Resource



- Code-Mixing is not the only factor that explains mBERT success
- Unsupervised adaptation push performance across code-mixing proportions

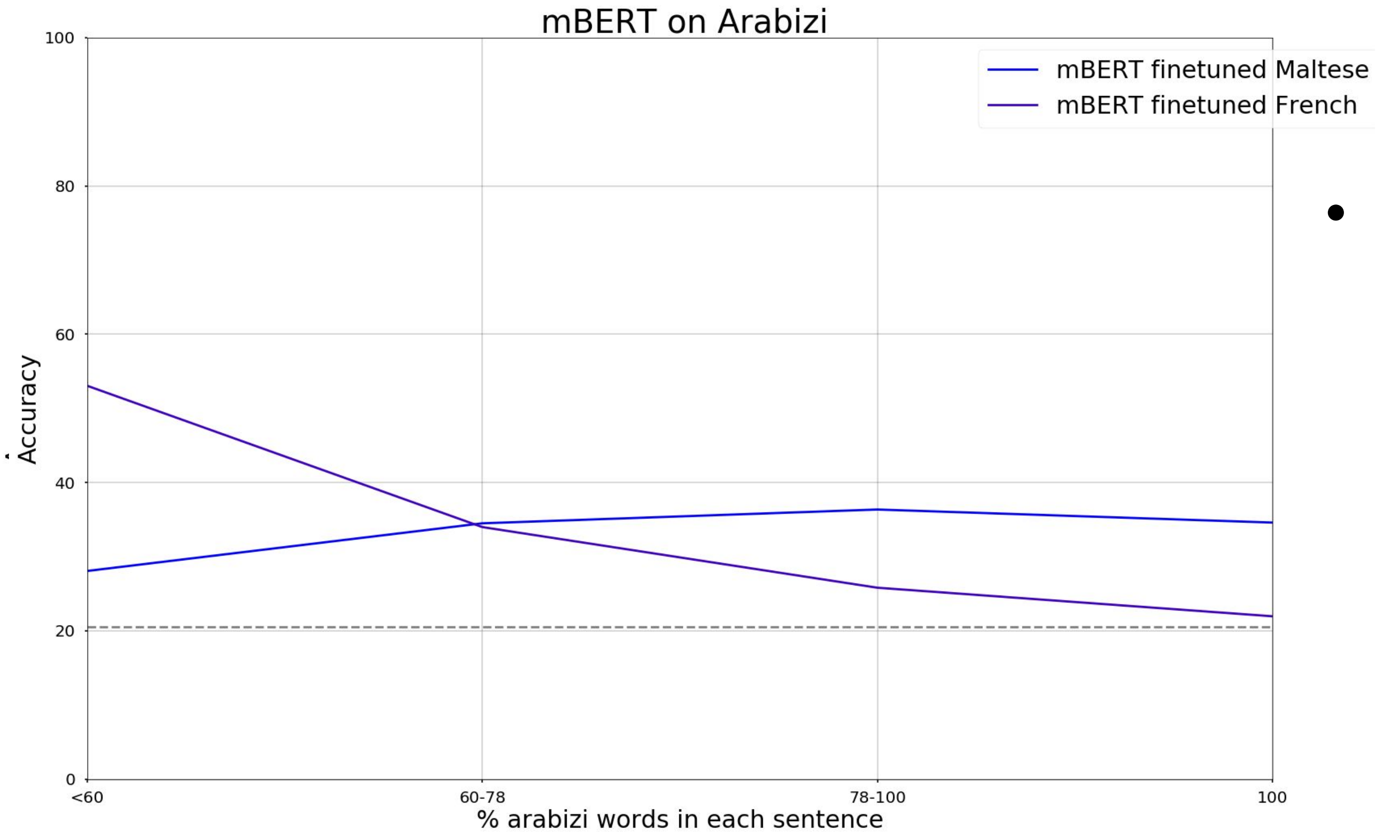
Experiments : Low Resource



High proportion of code-mixing — Only Arabizi

- Code-Mixing is not the only factor that explains mBERT success
- Unsupervised adaptation pushes performance across code-mixing proportions
- 1200 sentences enough to learn as well arabizi tokens as French tokens

Experiments : Cross-Lingual Transfer (Fr vs Maltese)

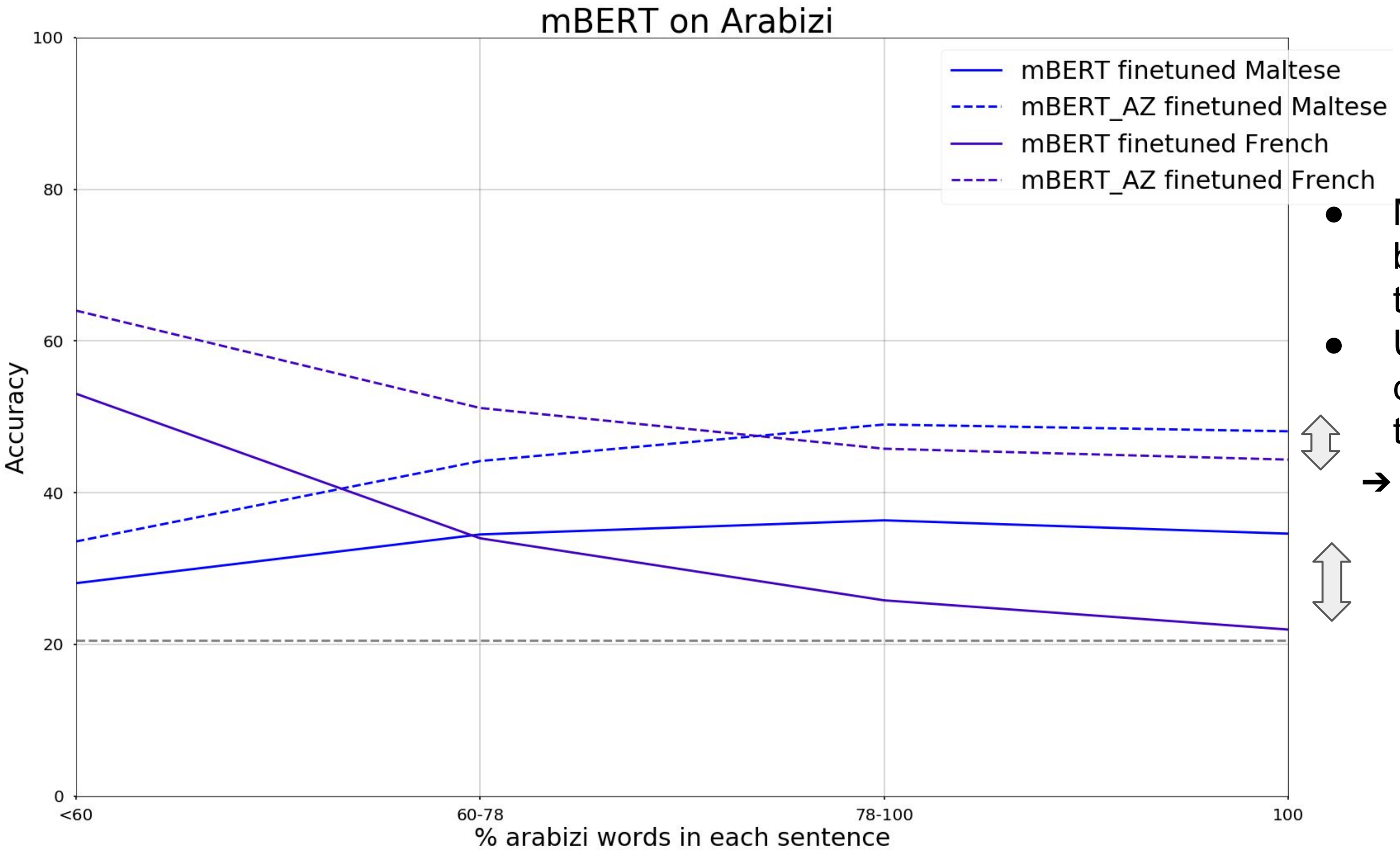


- Maltese transfer much better on Arabizi Tokens than French

High proportion of code-mixing

Only Arabizi

Experiments : Cross-Lingual Transfer (Fr vs Maltese)



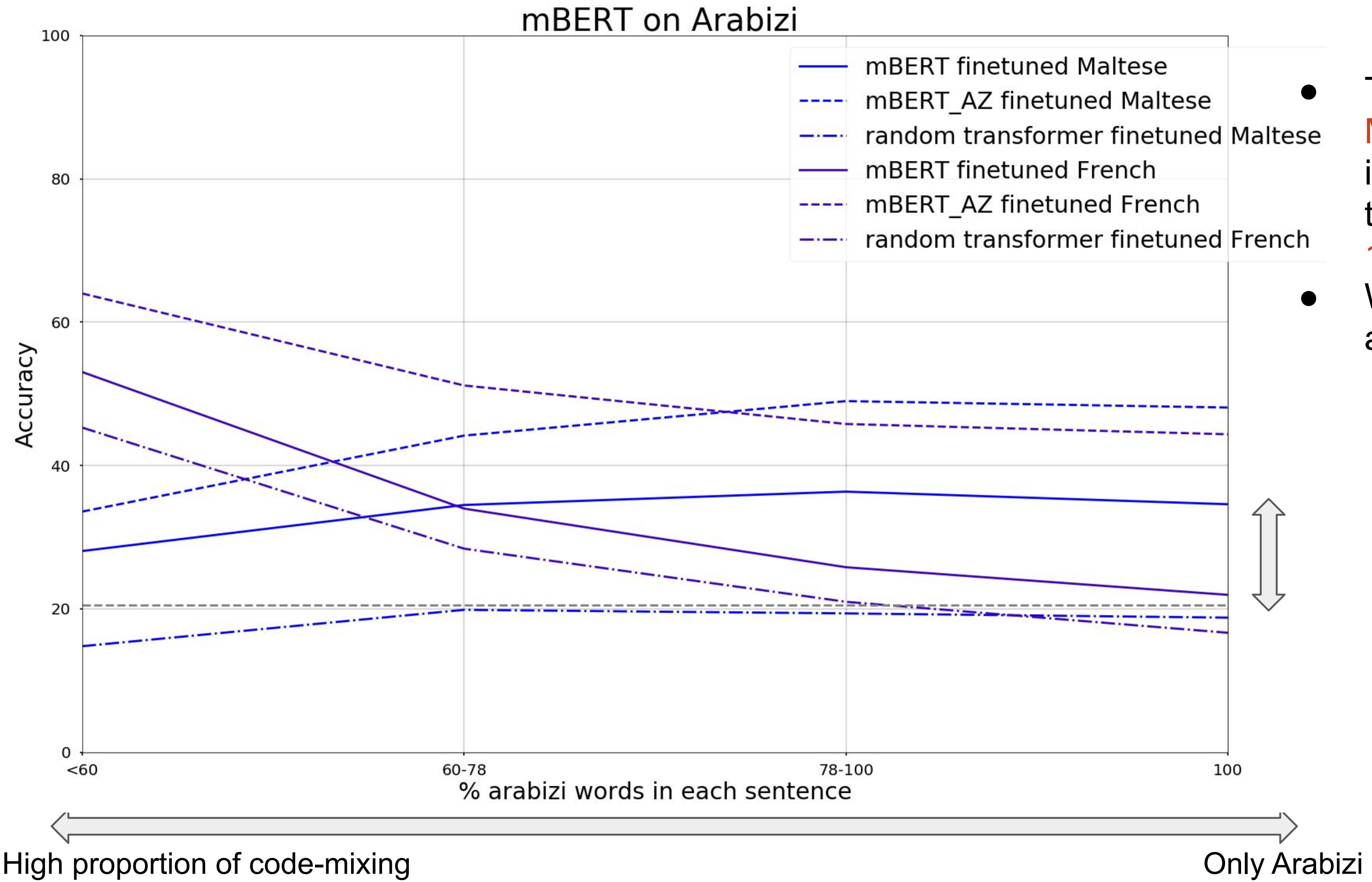
High proportion of code-mixing

Only Arabizi

- Maltese transfer much better on Arabizi Tokens than French
- Unsupervised Adaptation closes the gap on Arabizi tokens

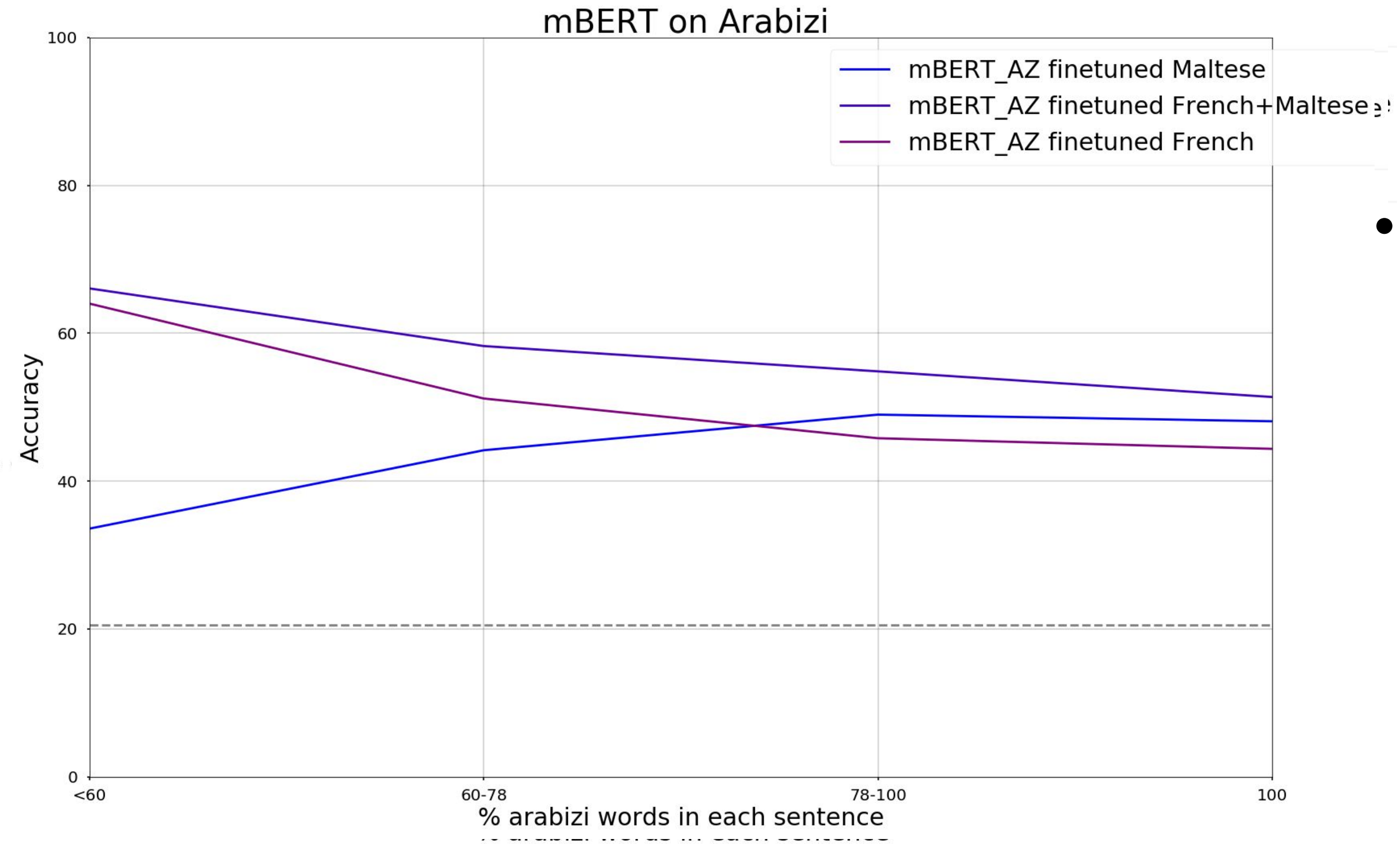
➔ Gap much smaller on Arabizi tokens after unsupervised adaptation

Experiments : Cross-Lingual Transfer (Fr vs Maltese)



- Transfer between Maltese and Arabizi is possible thanks to the pretraining on 104 languages
- While both Maltese and Arabizi are OOD!

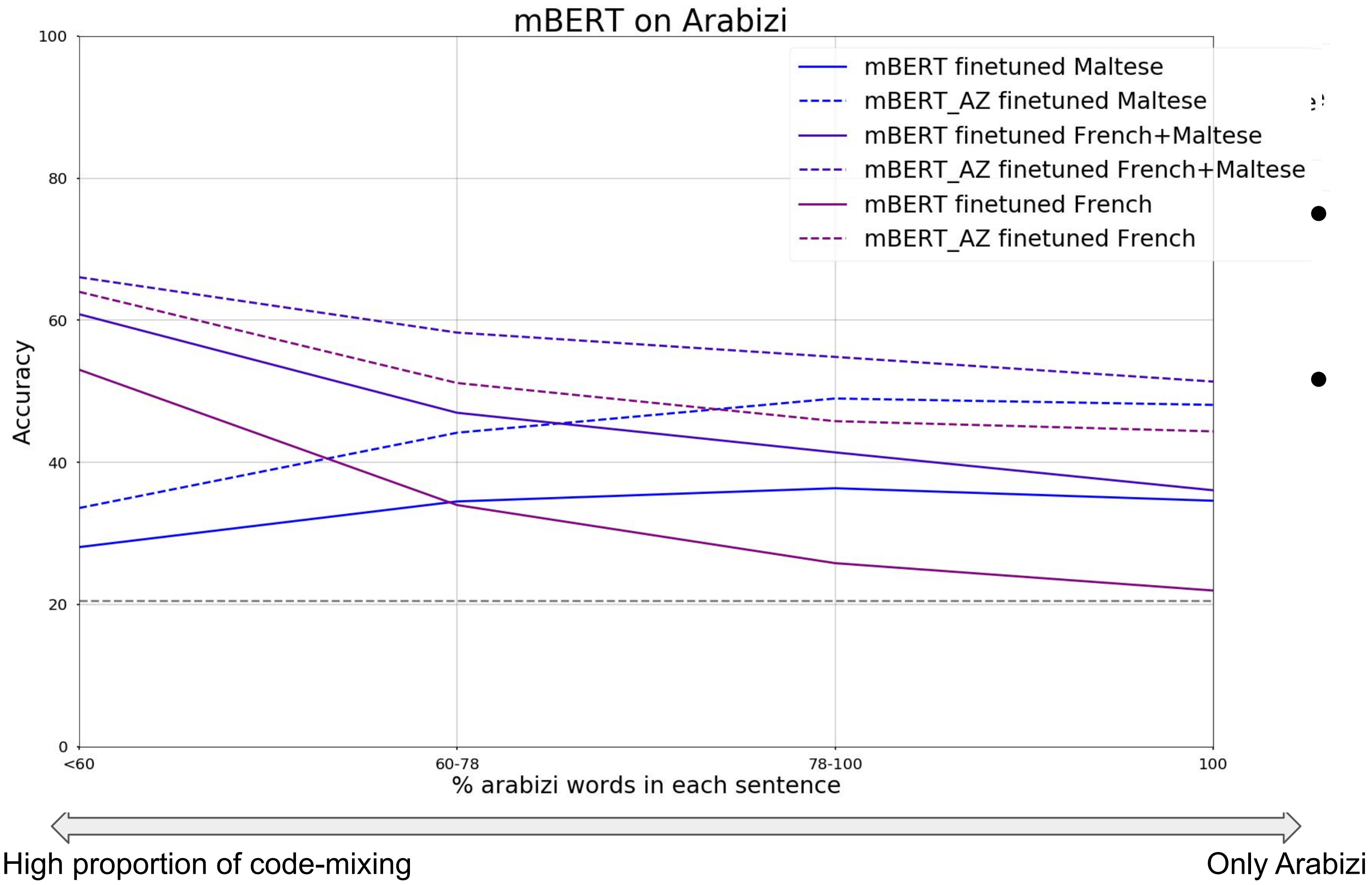
Experiments : Cross-Lingual Transfer (Fr vs Malt and concat)



- Concatenating Maltese and French leads to *best-of-both* world

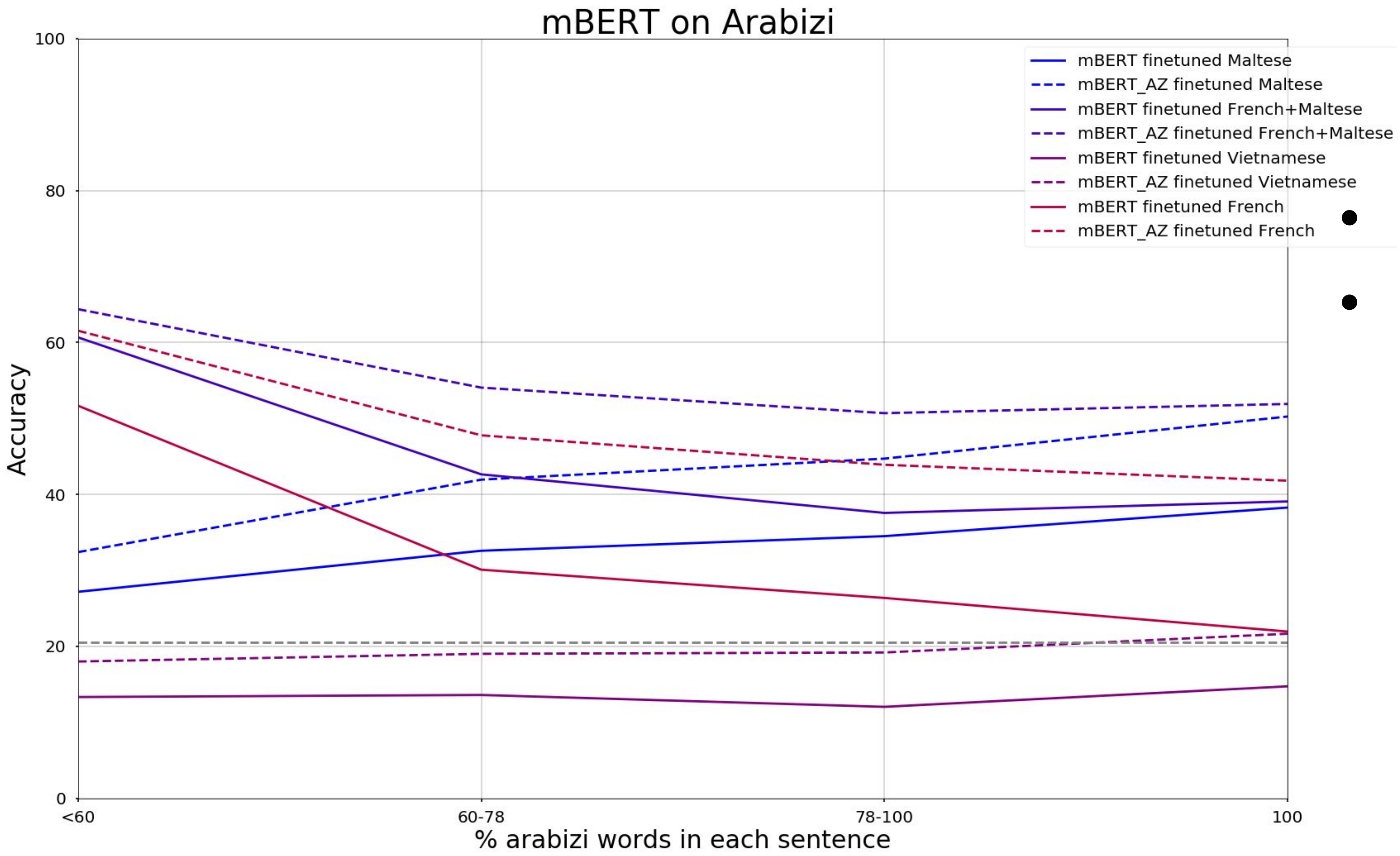
High proportion of code-mixing ⟷ Only Arabizi

Experiments : Cross-Lingual Transfer (Fr vs Malt and concat)



- Concatenating Maltese and French leads to *best-of-both* world
- Unsupervised adaptation help even more

Experiments : Cross-Lingual Transfer (with Vietnamese)



- No transfer across scripts (Vietnamese)
- Same for MSA

High proportion of code-mixing

Only Arabizi

Conclusions

- Multilingual pretrained language model are useful on unseen Dialect (low resource scenario).
  - Training in low resource scenario leads to competitive systems

- In Cross Lingual transfer
  - mBERT does not transfer across different scripts
  - Within same script : transfer is happening from multilingual representation to target unseen data
  - Unsupervised adaptation is highly impactful on the performance

➔ Structures captured at pre-training can be used and transferred to unseen languages
➔ Does Maltese to Arabizi transfer reveals Interlingua ability of multilingual BERT ?

Future Work

- Analyse representation along fine-tuning (wordpieces, embeddings,...)
- Experience with monolingual version of BERT (Roberta, CamemBERT, ...)
- Iterate on unsupervised adaptation (trick word-pieces, layer regularization...)
- Transferring across scripts ?