Unsupervised Learning for Handling Code-Mixed Data: A Case Study on POS Tagging of North-African Arabizi Dialect

Abhishek Srivastava, Benjamin Muller, Djamé Seddah

Unsupervised Learning for Handling Code-Mixed Data: A Case Study on POS Tagging of North-African Arabizi Dialect

Abhishek Srivastava , Benjamin Muller*, Djamé Seddah

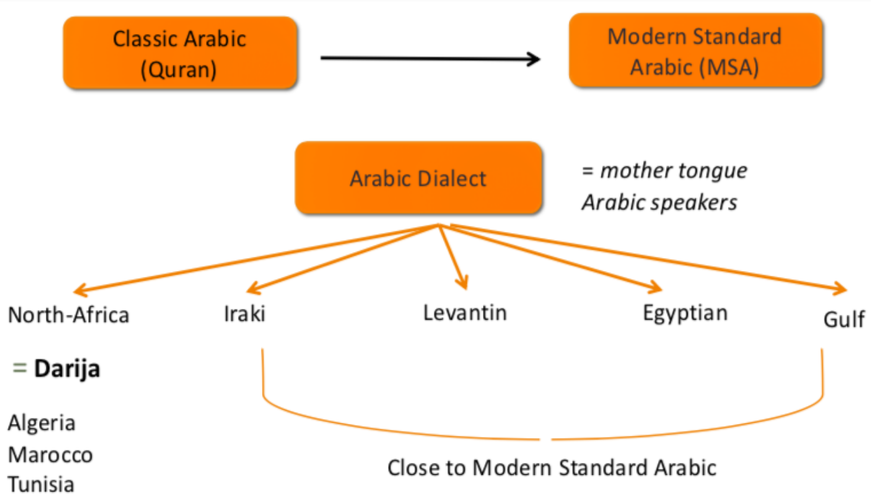
- PROJET FINANCÉ PAR L'ANR - PROJECT FUNDED BY THE ANR

Language model pretrained representation are now ubiquitous in Natural Language Processing. In this work, we present some first results in adapting those models to Out-of-Domain textual data.
Using **Part-of-Speech tagging** as our case study, we analyze the ability of **BERT** to model a complex **North-African** Dialect (**Arabizi**)**.**

Research questions

- Is BERT able to model Out-of-Domain languages such as Arabizi ?
- Can we adapt BERT in an unsupervised way to Arabizi ?

What is Arabizi ?



Definitions

- ***Dialectal Arabic*** is a variation of Classic Arabic that varies from one region to another that is spoken orally only. ***Darija*** is the one spoken in **Maghreb** (Algeria, Tunisia, Morocco).
- **Arabizi** is the name given to the transliterated language of dialectal Arabic in **Latin** script mostly found online.

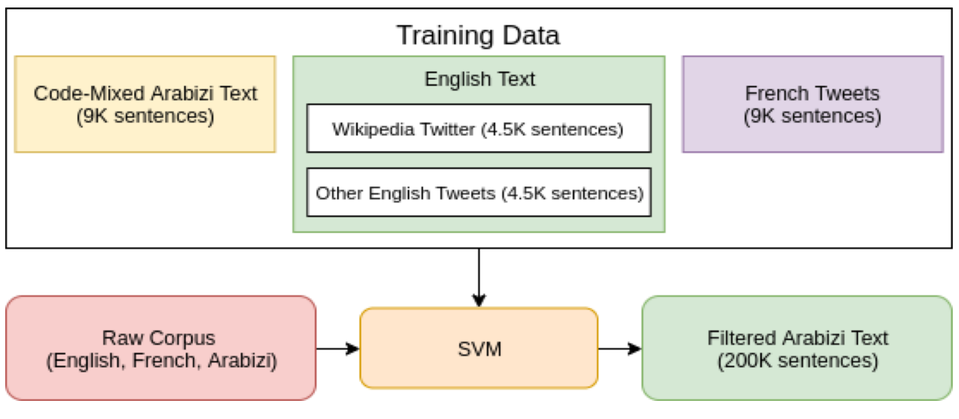
Key Property : High Variability

- **No** spelling, morphological or syntactic **fixed norms**
- Strong influence from **foreign languages**
- **Code-Switching** French / Darija

vive mca w nchalah had l'3am championi	Arabizi
long live MCA and I hope that this year we will be champions	English

Collecting and filtering raw Arabizi Data

We bootstrap a data set for Arabizi starting from 9000 sentences collected by Cotterell et al. (2014). Using keywords scraping, we collect **1 million UGC sentences** comprising French, English and Arabizi.
We filter 200k Arabizi sentences out of the raw corpus (94% F1 score) using our language identifier (cf. Figure below).



*contact author benjamin.muller@inria.fr
** early access data : contact djame.seddah@inria.fr

A new Treebank

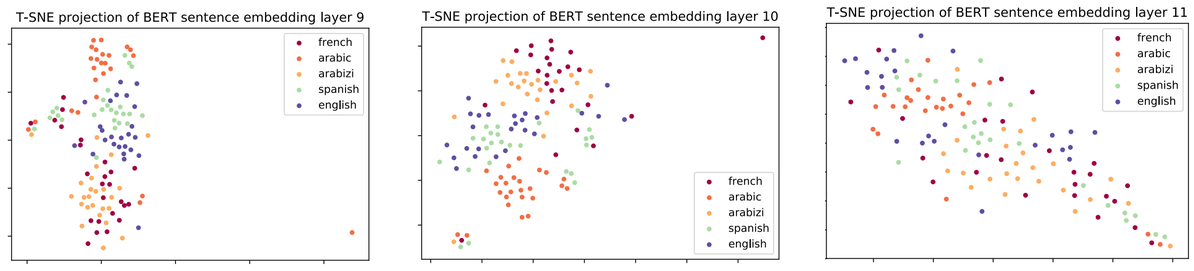
The first bottleneck in analyzing such a dialect is the lack of annotated resources. We developed a **CoNLL-U Treebank**** that includes **Part-of-Speech,** dependencies**,** and the translations of **1500 sentences** (originally posted in Facebook, Echorouk newspaper…).

Lexical Normalization

We train a clustering lexical normalizer using edit and word2vec distances. This degrades downstream performances in POS tagging.

BERT and Arabizi

We do our experiments on the released **base multilingual** version of BERT (Delvin et al. 2018) which was trained on a concatenation of Wikipedia of **104 languages**. BERT has never seen any Arabizi.

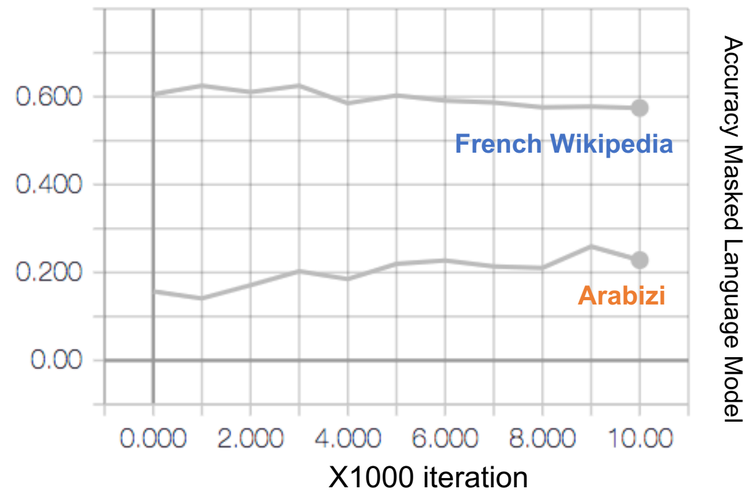


It is visible that Arabizi is related to French in BERT's embedding space

Unsupervised Fine Tuning of BERT on Arabizi

We fine-tune BERT (MLM objective) on the 200k Arabizi sentences

Figure 2 : Validation accuracy while fine tuning BERT on Arabizi data (200k sentence)



Results

Model	Accuracy
Baseline (udpipe)	73.7
Baseline + Normalization (udpipe)	72.4
BERT + POS tuning	77.3
BERT + POS tuning + Normalization (udpipe)	69.9
BERT + Unsupervised Domain fine tuning+ POS tuning	**78.3**

Final performance. Accuracy reported on the test set averaged over 5 runs

Summary

- Multilingual-BERT can be used to build a decent Part-of-Speech Tagger with a reasonable amount of annotated data
- Unsupervised adaptation improves (+1) performance in downstream POS tagging