Automatic Detection of Arabicized Berber and Arabic Varieties

Wafia Adouane[1], Nasredine Semmar[2], Richard Johansson[3], Victoria Bobicev[4]
Department of FLoV, University of Gothenburg, Sweden[1]
CEA Saclay – Nano-INNOV, Institut CARNOT CEA LIST, France[2]
Department of CSE, University of Gothenburg, Sweden[3]
Technical University of Moldova[4]
wafia.gu@gmail.com, nasredine.semmar@cea.fr
richard.johansson@gu.se, vika@rol.md

Abstract

Automatic Language Identification (ALI) is the detection of the natural language of an input text by a machine. It is the first necessary step to do any language-dependent natural language processing task. Various methods have been successfully applied to a wide range of languages, and the state-of-the-art automatic language identifiers are mainly based on character n-gram models trained on huge corpora. However, there are many languages which are not yet automatically processed, for instance minority and informal languages. Many of these languages are only spoken and do not exist in a written format. Social media platforms and new technologies have facilitated the emergence of written format for these spoken languages based on pronunciation. The latter are not well represented on the Web, commonly referred to as under-resourced languages, and the current available ALI tools fail to properly recognize them. In this paper, we revisit the problem of ALI with the focus on Arabicized Berber and dialectal Arabic short texts. We introduce new resources and evaluate the existing methods. The results show that machine learning models combined with lexicons are well suited for detecting Arabicized Berber and different Arabic varieties and distinguishing between them, giving a macro-average F-score of 92.94%.

1 Introduction

Automatic Language Identification (ALI) is a well-studied field in computational linguistics, since early 1960's, where various methods achieved successful results for many languages. ALI is commonly framed as a categorization.[1] problem. However, the rapid growth and wide dissemination of social media platforms and new technologies have contributed to the emergence of written forms of some varieties which are either minority or colloquial languages. These languages were not written before social media and mobile phone messaging services, and they are typically under-resourced. The state-of-the-art available ALI tools fail to recognize them and represent them by a unique category; standard language. For instance, whatever is written in Arabic script, and is clearly not Persian, Pashto or Urdu, is considered as Arabic, Modern Standard Arabic (MSA) precisely, even though there are many Arabic varieties which are considerably different from each other.

There are also other less known languages written in Arabic script but which are completely different from all Arabic varieties. In North Africa, for instance, Berber or Tamazight[2], which is widely used, is also written in Arabic script mainly in Algeria, Libya and Morocco. Arabicized Berber (BER) or Berber written in Arabic script is an under-resourced language and unknown to all available ALI tools which misclassify it as Arabic (MSA).[3] Arabicized Berber does not use special characters and it coexists with Maghrebi Arabic where the dialectal contact has made it hard for non-Maghrebi people to distinguish

[1] Assigning a predefined category to a given text based on the presence or absence of some features.

[2] An Afro-Asiatic language widely spoken in North Africa and different from Arabic. It has 13 varieties and each has formal and informal forms. It has its unique script called Tifinagh but for convenience Latin and Arabic scripts are also used. Using Arabic script to transliterate Berber has existed since the beginning of the Islamic Era (L. Souag, 2004).

[3] Among the freely available language identification tools, we tried Google Translator, Open Xerox language and Translated labs at http://labs.translated.net.

Proceedings of the Third Workshop on NLP for Similar Languages, Varieties and Dialects,
pages 63–72, Osaka, Japan, December 12 2016.