Natural Language Processing for Low-resourced Code-switched Colloquial Languages

The Case of Algerian Language

Wafia Adouane

Thesis for the degree of Doctor of Philosophy in computational linguistics, to be publicly defended, by due permission of the dean of the Faculty of Arts at the University of Gothenburg.

September 2, 2020 at 17:00
C350, Humanisten, Renströmsgatan 6, Gothenburg

Department of Philosophy, Linguistics and Theory of Science
Centre for Linguistic Theory and Studies in Probability

Title	Natural Language Processing for Low-resourced Code-switched Colloquial Languages – The Case of Algerian Language
Author	Wafia Adouane
Language	English
Keywords	Natural language processing, Deep neural networks, Low-resourced language, Colloquial language, Code-switch, Dialectal Arabic, User-generated data, Non-standardised orthography, Algerian language
ISBN	978-91-7833-958-7 (print) 978-91-7833-959-4 (pdf)

Abstract

In this thesis we explore to what extent deep neural networks (DNNs), trained end-to-end, can be used to perform natural language processing tasks for code-switched colloquial languages lacking both large automated data and processing tools, for instance tokenisers, morpho-syntactic and semantic parsers, etc. We opt for an end-to-end learning approach because this kind of data is hard to control due to its high orthographic and linguistic variability.

This variability makes it unrealistic to either find a dataset that exhaustively covers all the possible cases that could be used to devise processing tools or to build equivalent rule-based tools from the bottom up. Moreover, all our models are language-independent and do not require access to additional resources, hence we hope that they will be used with other languages or language varieties with similar settings.

We deal with the case of user-generated textual data written in Algerian language as naturally produced in social media. We experiment with five natural language processing tasks, namely Code-switch Detection, Semantic Textual Similarity, Spelling Normalisation and Correction, Sentiment Analysis, and Named Entity Recognition. For each task, we created a dataset from user-generated data reflecting the real use of the language.

Our experimental results in various setups indicate that end-to-end DNNs combined with character-level representation of the data are promising. Further experiments with advanced models, such as Transformer-based models, could lead to even better results. Completely solving the challenge of code-switched colloquial languages is beyond the scope of this experimental work. Even so, we believe that this work will extend the utility of DNNs trained end-to-end to low-resource settings. Furthermore, the results of our experiments can be used as a baseline for future research.