Resources and Applications for Dialectal Arabic

The case of Levantine

Chatrine Qwaider

Department of Philosophy, Linguistics and Theory of Science

Thesis submitted 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, on May 25, 2022, at 15:00, in J439, Lilla Hörsalen, Humanisten, Renströmsgatan 6, Gothenburg.
Faculty opponent: Muhammad Abdul-Mageed, University of British Columbia.

UNIVERSITY OF GOTHENBURG

Title	Resources and Applications for Dialectal Arabic
Author	Chatrine Qwaider
Language	English with a Swedish abstract
Department	Department of Philosophy, Linguistics and Theory of Science
ISBN	978-91-8009-803-8 (print) ISBN 978-91-8009-804-5 (pdf)

This is a thesis about the computational study of Dialectal Arabic (DA). In particular, the thesis studies DA, with a special emphasis on Levantine Arabic, and develops tools and resources for the computational study of Dialectal Arabic Natural Language Processing (DANLP). It investigates the creation of fine-grained resources that can be used for a variety of computational tasks, and a number of effective models that can deal with the complexity of fine-grained dialectal data. Dialect Identification (DI), as well as Sentiment Analysis (SA) are the Natural Language Processing (NLP) tasks investigated in this thesis.

In the first part (Study 1 and Study 2), I study the DI task on both coarse-grained and fine-grained levels. For this reason, I build the first annotated Levantine (SHAMI) Dialect Corpus (SDC). Furthermore, I explore the ability of n-gram language models, Machine Learning (ML) algorithms and ensemble learning techniques to classify and detect 26 Arabic varieties. In the second part, I conduct a linguistic study to measure the lexical distance between MSA and DA, and between the dialects themselves. This is done to check whether transferring knowledge from one variety to another is possible. In the third part, studies 4,5 and 6, I explore Arabic Sentiment Analysis (SA). I investigate the idea of knowledge transfer between MSA and the dialects using SA as a case study. Furthermore, I implement various models such as the pre-trained language model BERT, Deep Learning (DL), ML and feature engineering approaches to detect the sentimental polarity of DA data. I introduce two valuable resources for this task, one focusing on Levantine sentiment (Shami-Senti), and the other for DA in general (ATSAD). I exploit different ways of annotation, e.g. human, lexicon-based and automatic distant supervision annotation. The last study is about choosing the best model for DI and SA. I exploit well-known models and approaches using various kinds of DA resources.

The thesis contributes to the field of DANLP in a number of ways. The introduced valuable resources can be seen as a stepping stone for a deeper investigation and understanding of issues in DANLP. They are also reliable and can be used by researchers to address different NLP tasks. The cross-dialectal linguistic studies will open up prospects for researchers to fine-tune models and transfer knowledge among Arabic varieties. A big part of the contribution lies in designing DI and SA models. I implement several ML models that use feature engineering approaches and N-gram language models to identify the dialect or detect the sentiment. For DI, I design and implement an ensemble learning model that is able to handle fine-grained dialects. Additionally, I exploit the usage of DL models on different SA dialectal datasets and achieve competitive results. For both tasks, I exploit the recent pre-trained language models and perform a comparison to choose the best model. I also implement a semi-supervised approach for automatic labelling and annotating data with the help of self-training techniques to improve the performance of the dataset. These models will help researchers dive deeper into DANLP and create practical and industrial systems.

Keywords: Dialectal Arabic Natural Language Processing, Computational Linguistics, Dialect Identification, Sentiment Analysis, Machine Learning, Deep Learning, Language modelling