The ASMAT project - Arabic Social Media Analysis Tools

Fatiha Sadat
University of Quebec in Montreal
201 President Kennedy
Montreal, QC, H2X 3Y7, Canada
sadat.fatiha@uqam.ca

List of partners
Atefeh Farzindar, NLP Technologies Inc. http://www.nlptechnologies.ca 52, Le Royer Street W., Montréal, Québec, Canada, H2Y 1W7 farzindar@nlptechnologies.ca

Summary

The main objective of the ASMAT project – *Arabic Social Media Analysis Tools*, is to make available a comprehensive set of language resources and tools covering Arabic dialects in social media context.

Current Arabic NLP tools are capable of analysing large part of standard Arabic, but fail short of handling the dialects and the social media domain. To this end, the project aims to create tools for Arabic language and its varieties following certain tasks: (1) language and dialect identification; (2) dialect to standard (MSA) mapping and vice versa; (3) automatic machine translation from any Arabic dialect to English and French. More specifically, the ASMAT project deals with the *Maghrebi* (*North African*) Arabic dialects for machine translation with very scarce resources.

Parts of the ASMAT project, such as dialect identification for all varieties of Arabic language and a systematic rule-based mapping of the Tunisian dialect to MSA were achieved on December 2013, with an industrial collaboration with NLP Technologies, under NSERC Engage[1] grant. Our latest evaluations showed that Naive Bayes classifiers based on character bi-gram model and trained on data extracted from forums and blogs on 18 Arabic dialects could identify the 18 different Arabic dialects with a considerable overall accuracy of 98% on social media texts. A successful identification of which sentence in written in which dialect could guide the system in using the specific pre-processing tools for the respective dialectal portions.

We have already achieved a rule-based system that converts any text of social media in Tunisian dialect to MSA. We are working on the construction of more linguistic resources for the Tunisian dialect and MSA that will help build a hybrid statistical and rule-based MT system integrated in the ASMAT project. Finally, the translation from the Tunisian dialect to French and/or English will be completed through MSA as a pivot language.

Future works of the ASMAT project are concerned by all varieties of Arabic dialects for machine translation, starting from the Maghrebi.

The ASMAT project will be funded from late 2014 by additional research grants for a longer period.

[1] http://www.nserc-crsng.gc.ca/Professors-Professeurs/RPP-PP/Engage-engagement_eng.asp