Multi-Dialect Machine Translation (MuDMaT)

Natural Science and Engineering Research Council of Canada (NSERC)
Research Project
NSERC 356097-2008

List of partners

Fatiha Sadat, University of Quebec in Montreal, QC, Canada (coordinator)

Project duration: January 2014 — December 2017

Summary

The Multi-Dialect Machine Translation (MuDMaT) project aims to encourage research and development of Machine Translation (MT) systems for less resourced languages and their variants or dialects. More specifically, the MuDMaT project deals with three Maghrebi (North African) Arabic dialects for machine translation with very scarce resources: the Tunisian, the Algerian and the Moroccan. Many ideas of this project can be applied to any less-resourced language variant or dialect.

In this project, an Arabic dialect can play a role as a source dialect in machine translation system when translating into French with considering the Modern Standard Arabic (MSA) as pivot language. Moreover, this dialect can play a role as a target dialect when translating from French and/or MSA. A third translation module focuses on translations from a dialect into another dialect using MSA as pivot language.

At the current stage, MuDMaT targets building hybrid statistical and rule-based machine translation systems from multiple Arabic dialects into MSA and French and vice versa. Statistical machine translation based on parallel corpora has been very successful and widely used in major translation systems' engines. Our interest in this project focuses on comparable corpora, which are defined as monolingual corpora covering roughly the same subject area or author's name or dates in different languages but without being exact translations of each other. In our project, comparable corpora are built by mining the World Wide Web and more specifically the social media such as blogs. Other linguistic resources such as lexicons (and grammar) that are automatically extracted from the Web or collaboratively built (through crowdsourcing) are exploited in this multi-dialect translation system.

The project has already been running for a year and a demonstration using the Tunisian dialect in a rule-based translation system for translating texts into MSA and French was achieved. We are working towards the construction of more linguistic resources such as comparable and parallel corpora for the Tunisian dialect and MSA that will help enhance the hybrid statistical and rule-based MT system. The other North African Arabic dialects will be included in the rule-based translation system during this research project.

The availability of the multi-dialect machine translation system will be through the years 2015, 2016 and 2017.