Arabic Dialect Processing

Mona Diab and Nizar Habash
Columbia University

The existence of dialects for any language constitutes a challenge for Natural Language Processing (NLP) in general since it adds another set of variation dimensions from a known standard. The problem is particularly interesting and challenging in Arabic and its different dialects, where the diversion from the standard could, in some linguistic theories, warrant a classification as a different language. This problem would not be as pronounced if standard Arabic were to be a living language, however it is not. Any realistic and practical approach to processing Arabic will have to account for dialectal usage since it is so pervasive. In this tutorial, we will attempt to highlight different dialectal phenomena and how they migrate from the standard and why they pose challenges to NLP. Our tutorial will have four different parts:
First, we will give you a background layout of issues for standard Arabic NLP. Then, we will present a high level generic view of dialects and different aspects of them that are of interest for the NLP community, addressing both text and speech issues in addition to standardization issues. We will focus in depth on two aspects of dialect processing in the third and fourth parts of the tutorial, namely, dialectal morphology and dialectal syntactic parsing. Throughout the presentation we will make references to the different resources available and draw contrastive links with standard Arabic and English. We will provide links to recent publications and available toolkits/resources for all four sections.

This tutorial is designed for computer scientists and linguistics alike. The tutorial will provide NLP system developers/researchers with necessary background information for working with the Arabic and its dialects, which have recently become a focus of an increasing number of projects in computational linguistics. Previous versions of the tutorial were given at AMTA 2006 and NAACL 2007: http://www.ccls.columbia.edu/oldweb/cadim/ArabicDialectTutorialAMTA2006.pdf

Short Bios of Speakers

Mona Diab received her PhD in 2003 in the Linguistics department and UMIACS, University of Maryland College Park. Her PhD work focused on lexical semantic issues and was titled Word Sense Disambiguation within a Multilingual Framework. Mona is currently an associate research scientist at the Center for Computational Learning Systems, Columbia University. Her research includes work on word sense disambiguation, automatic acquisition of natural language resources such as dictionaries and taxonomies, unsupervised learning methods, lexical semantics, cross language knowledge induction from both parallel and comparable corpora, Arabic NLP in general, tools for processing Arabic(s), computational modeling of Arabic dialects, Arabic syntactic and semantic parsing.

Dr. Diab served as co-chair – together with Kareem Darwish and Nizar Habash - of the Workshop on Computational Approaches to Semitic Languages (ACL 2005). She was also a senior member in the 2005 JHU summer workshop on Parsing Arabic Dialects. In 2005, she co-founded the Columbia Arabic Dialect Modeling (CADIM) group together with Nizar Habash and Owen Rambow. She has published over 30 articles in different conferences, journals and workshops. Mona has presented her work in numerous lectures and tutorials both for academic and industrial audiences.

Nizar Habash received his PhD in 2003 from the Computer Science Department, University of Maryland College Park. His Ph.D. thesis is titled Generation-Heavy Hybrid Machine Translation. He is currently an Associate research scientist at the Center for Computational Learning Systems in Columbia University. His research includes work on machine translation, natural language generation, lexical semantics, morphological analysis, generation and disambiguation, computational modeling of Arabic dialects, and Arabic dialect parsing.

Dr. Habash served as co-chair for the Workshop on Computational Approaches to Semitic Languages (ACL 2005) and also the Workshop on Machine Translation for Semitic Languages (MT Summit 2003). In 2005, he co-founded the Columbia Arabic Dialect Modeling (CADIM) group. He is the vice-president of the Semitic Language Special Interest Group in the Association of Computational Linguistics. Finally, he served as research program co-chair for AMTA 2006.

Dr. Habash has published over 30 articles in international conferences and journals and has given numerous lectures and tutorials for academic and industrial audiences.

Mona's website: http://www.cs.columbia.edu/~mdiab
Nizar's website: http://www.cs.columbia.edu/~habash
CADIM website: http://www.ccls.columbia.edu/oldweb/cadim