T3: Arabic Dialect Processing Tutorial

Mona Diab, Nizar Habash

ABSTRACT

The existence of dialects for any language constitutes a challenge for 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 views, warrant a classification as different languages. This problem would not be as pronounced if Modern Standard Arabic (MSA) were the native language of some sub group of Arabic speakers, 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, how they migrate from the standard and why they pose challenges to NLP. This area of research (dialects in general and Arabic dialects in particular) is gaining a lot of interest. For example, the DARPA-funded BOLT program starting this year will only consider dialectal varieties for its effort on Arabic. Furthermore, there was a workshop on dialect processing as part of EMNLP 2011.

This tutorial has four different parts: First, we contextualize the question of Arabic dialects from a sociolinguistic and political perspective. Second, we present a discussion of issues in relevant to Arabic NLP; this includes generic issues common to MSA and dialects, and MSA specific issues. In the third part, we detail dialectal linguistic issues and contrast them to MSA issues. In the last part, we review the state-of-the-art in Arabic dialect processing covering several enabling technologies and applications, e.g., dialect identification, speech recognition, morphological processing (analysis, disambiguation, tokenization, POS tagging), parsing, and machine translation. Throughout the presentation we will make references to the different resources available and draw contrastive links with standard Arabic and English. Moreover, we will discuss annotation standards as exemplified in the Treebank. We will provide links to recent publications and available toolkits/resources for all four sections.

This tutorial is designed for computer scientists and linguists alike. No knowledge of Arabic is required (though, we recommend taking a look at Nizar Habash's Arabic NLP tutorialhttp://www1.ccls.columbia.edu/~cadim/presentations.html which will be reviewed as part of the tutorial.)

OUTLINE

1. **Introduction**

   Introduction to the question of Arabic dialects from sociolinguistic and political perspectives (20 min)

2. **General (Standard/Dialectal) Arabic linguistic issues and their relevance to NLP**

   Orthography, Phonology, Morphology, Syntax (60 min)

3. **Coffee Break**

   (20 min)

4. **Generic dialect issues from an NLP perspective**

   Orthography, Phonology, Morphology, Syntax (40 min)

5. **State-of-the-art in a sample of applications for Arabic dialects**

   Speech recognition, Morphological processing, Parsing, Machine Translation (40 min)

BIOS

Mona Diab
850 Interchurch Center MC 7717
475 Riverside Drive
New York, NY 10115
Office 212-870-1290, Fax 212-870-1285
mdiab--AT--ccls.columbia.edu
http://www1.ccls.columbia.edu/~mdiab/

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 a 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.

Nizar Habash
850 Interchurch Center MC 7717
475 Riverside Drive
New York, NY 10115
Office 212-870-1289, Fax 212-870-1285
habash@ccls.columbia.edu
http://www.nizarhabash.com

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 a 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. Nizar recently published a book entitled "Introduction to Arabic Natural Language Processing".