Tutorial Abstract

Arabic Natural Language Processing

Nizar Habash
Computational Approaches to Modeling Language (CAMeL) Lab
New York University Abu Dhabi
nizar.habash@nyu.edu

Abstract

The Arabic language continues to be the focus of an increasing number of projects in natural language processing (NLP) and computational linguistics (CL). This tutorial provides NLP/CL system developers and researchers (computer scientists and linguists alike) with the necessary background information for working with Arabic in its various forms: Classical, Modern Standard and Dialectal. We discuss various Arabic linguistic phenomena and review the state-of-the-art in Arabic processing from enabling technologies and resources, to common tasks and applications. The tutorial will explain important concepts, common wisdom, and common pitfalls in Arabic processing. Given the wide range of possible issues, we invite tutorial attendees to bring up interesting challenges and problems they are working on to discuss during the tutorial.

Type of Tutorial: Introductory.

1 Tutorial Description

The purpose of this tutorial is to provide system developers and researchers in natural language processing (NLP) and computational linguistics (CL) with the necessary background information for working with the Arabic language (Modern Standard Arabic, Classical Arabic and Arabic Dialects). The goal is to introduce Arabic linguistic phenomena that need to be addressed from orthography and phonology, to morphology, syntax and semantics, as well as to review the state-of-the-art on Arabic processing from enabling technologies and resources, to common tasks and applications. Alternative approaches will be presented and contrasted for their value in different application contexts. The tutorial will explain important concepts, common wisdom, common pitfalls, as well as basic skills for handling Arabic text, even when illiterate in the Arabic script.

2 Tutorial Outline

This tutorial introduces the different challenges and current solutions to the automatic processing of Arabic and its dialects. The tutorial has three parts (60 minutes each). The second part will be split into two portions, 30 minutes before the coffee break, and 30 minutes after.

Part 1: Arabic NLP Challenges We present the main challenges Arabic poses for NLP. Topics include Arabic script and orthography, orthographic ambiguity and noise, Arabic morphology, morphological richness, complexity and ambiguity, Arabic syntactic and semantic considerations, and Arabic dialectal variations and their challenges.

Part 2: Arabic NLP Solutions We review the state-of-the-art in Arabic NLP covering several enabling technologies and applications, e.g., transliteration schemes, morphological processing (analysis, disambiguation, tokenization, POS tagging), orthographic normalization, dialect identification, text analytics, syntactic parsing, and language modeling. Throughout the presentation we will make references to the different resources and tools available including discussing Arabic annotation standards, tools, and best practices. We will provide links to recent publications and available toolkits and resources.

Part 3: Arabic NLP New Frontiers In this section, we highlight some of the latest efforts and open problems in Arabic NLP, from work on summarization to text simplification, spelling and grammar correction, and gender rewriting. We review the various ongoing Arabic NLP shared tasks and discuss the directions the field is going into, while drawing on historical trends and patterns. This part will interactively draw on the audience and their interests in Arabic NLP.

Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing: Tutorial Abstracts, pages 9 - 10

3 Prerequisites

This is an introductory tutorial. No previous knowledge in Arabic is needed. This tutorial is designed for computer scientists and linguists alike. Acquaintance with basic formal language theory and knowledge of some programming languages will be useful.

4 Preparatory Pointers

The following are a set of optional *initial* pointers that will help the attendees maximize their learning experience.

Readings and Lectures

- A panoramic survey of natural language processing in the Arab world [Arxiv version with extended bibliography] (Darwish et al., 2021).
- Arabic Natural Language Processing: Challenges and Solutions [YouTube] (Habash, 2019).
- The Introduction to Arabic Natural Language Processing book (Habash, 2010).

Resources

- Masader+: The Arabic NLP data catalogue: [GitHub] (Alyafeai et al., 2022).
- CAMeL Tools: A suite of Arabic NLP tools [GitHub] (Obeid et al., 2020).
- Farasa: A full-stack package for Arabic Language Processing [Website] (Abdelali et al., 2016).

Sites

- SIGARAB: The ACL Special Interest Group on Arabic Natural Language Processing `http://www.sigarab.org/`, [Mailing List]
- The Arabic Natural Language Processing Workshop (WANLP) [Google Scholar]
- The Workshop on Open-Source Arabic Corpora and Processing Tools (OSACT) [Google Scholar]

Your Ideas and Questions Given the wide range of possible topics, we invite tutorial attendees to come prepared with interesting challenges and problems they are working on to discuss during the tutorial.

5 Tutorial Instructor

Nizar Habash is a Professor of Computer Science at New York University Abu Dhabi (NYUAD). He is also the director of the Computational Approaches to Modeling Language (CAMeL) Lab. Professor Habash specializes in natural language processing and computational linguistics. Before joining NYUAD in 2014, he was a research scientist at Columbia University's Center for Computational Learning Systems. He received his PhD in Computer Science from the University of Maryland College Park in 2003. He has two bachelors degrees, one in Computer Engineering and one in Linguistics and Languages. His research includes extensive work on machine translation, morphological analysis, and computational modeling of Arabic and its dialects. Professor Habash has been a principal investigator or co-investigator on over 25 research grants. And he has over 250 publications including a book entitled "Introduction to Arabic Natural Language Processing" (Habash, 2010). His website is at `www.nizarhabash.com`.

References

Ahmed Abdelali, Kareem Darwish, Nadir Durrani, and Hamdy Mubarak. 2016. Farasa: A Fast and Furious Segmenter for Arabic. In *Proceedings of the North American Chapter of the Association for Computational Linguistics*, San Diego, California.

Zaid Alyafeai, Maraim Masoud, Mustafa Ghaleb, and Maged S. Al-shaibani. 2022. Masader: Metadata sourcing for Arabic text and speech data resources. In *Proceedings of the Language Resources and Evaluation Conference*, Marseille, France.

Kareem Darwish, Nizar Habash, Mourad Abbas, Hend Al-Khalifa, Huseein T Al-Natsheh, Houda Bouamor, Karim Bouzoubaa, Violetta Cavalli-Sforza, Samhaa R El-Beltagy, Wassim El-Hajj, et al. 2021. A panoramic survey of natural language processing in the Arab world. *Communications of the ACM*, 64(4):72–81.

Nizar Habash. 2010. *Introduction to Arabic natural language processing*, volume 3. Morgan & Claypool Publishers.

Nizar Habash. 2019. Arabic natural language processing: Challenges and solutions. Grammarly AI-NLP Club #8, Kyiv, Ukraine.

Ossama Obeid, Nasser Zalmout, Salam Khalifa, Dima Taji, Mai Oudah, Bashar Alhafni, Go Inoue, Fadhl Eryani, Alexander Erdmann, and Nizar Habash. 2020. CAMeL Tools: An open source python toolkit for Arabic natural language processing. In *Proceedings of The Language Resources and Evaluation Conference*, Marseille, France.