Book Reviews

Introduction to Arabic Natural Language Processing

Nizar Y. Habash
(Columbia University)

Morgan & Claypool (Synthesis Lectures on Human Language Technologies, edited by Graeme Hirst, volume 10), 2010, xvii+167 pp; paperbound, ISBN 978-1-59829-795-9, $40.00; ebook, ISBN 978-1-59829-796-6, $30.00 or by subscription

Reviewed by
Imed Zitouni
IBM T. J. Watson Research Center

The Arabic language, which is the mother tongue of more than 300 million people, presents significant challenges to many natural language processing (NLP) applications (Farghaly and Shaalan 2009). Arabic is a highly inflected and derived language. The scale of Arabic computational linguistic research work is now orders of magnitude beyond what was available a decade ago. Hence, it becomes important to find a book that introduces necessary background information for working with Arabic in research or development on NLP and computational linguistics. Habash's book *Introduction to Arabic Natural Language Processing* is an important step toward that goal. It introduces Arabic linguistic phenomena, shows how the Arabic language can be handled by computers, and presents resources and tools available for Arabic.

The book focuses on introducing the language, as opposed to explaining how to build or conduct research in a specific NLP research area. It is not meant to be an in-depth description of how to build specific NLP applications. However, it highlights necessary issues to be aware of in NLP research and application building. As an example, readers will find a solid overview of core tasks including morphological analysis, generation, tokenization, part-of-speech (POS) tagging, and even parsing. But it does not contain detailed instructions on how to build any of these technologies. The author's approach is to highlight Arabic-specific issues (e.g., how Arabic morphology interacts with different approaches to POS tagging) and to leave language-independent technologies such as core POS tagging or parsing to other resources. Readers will find a nice note on machine translation from and to Arabic (Chapter 8) that addresses the multilingual aspects of working with Arabic.

The book is meant for students, researchers, and engineers who are interested in working in Arabic computational linguistics and NLP. No prior knowledge of the Arabic language is required. The book mainly discusses Modern Standard Arabic (MSA), but the author doesn't hesitate to refer to Arabic dialect occasionally, especially in the early chapters. Readers not familiar with Arabic get a good introduction to the language and how it can be handled by computers. Readers with prior knowledge of Arabic are reminded of the basics needed for computational research, some of which may not be intuitive to a person familiar with the language only linguistically as opposed to computationally.

The book starts with an introductory chapter called "What is Arabic?". This chapter briefly explains the difference between Arabic the language and Arabic dialects. It then gives a better understanding of what the reader will learn from the following seven chapters. In the subsequent chapters, the author carefully discusses resources available

in the field for each specific topic. This book also has five appendixes that are of great interest to readers.

Arabic Script. This chapter discusses the Arabic scripts as used in MSA. It is not only a linguistic description of the Arabic script but also a discussion of computer encoding and text input and display for Arabic. Another aspect discussed in this chapter that is important in any NLP application is orthographic transliteration and orthographic normalization. Several transliteration approaches (including the well-known Buckwalter transliteration method) are also presented here. These approaches are simply methods to produce a one-to-one mapping between Arabic and Latin characters, sometimes needed for non-Arabic readers. One can work directly with the Arabic script. One application directly related to Arabic scripts is handwriting recognition, which the author gives a short note about in the last section, with pointers to important starting references.

Arabic Phonology and Orthography. This chapter nicely introduces MSA phonology and shows how the Arabic spelling rules can be used to map between Arabic phonology and script. This knowledge is important for applications such as proper name transliteration, spelling correction, automatic speech recognition, and speech synthesis. These applications are also briefly discussed with references in the last section of this chapter.

Arabic Morphology. This chapter has the lion's share of this book in terms of size and content. This is expected because Arabic morphology is challenging and it is central when working on any Arabic NLP application. This chapter has the advantage of helping new scientists and engineers remove some confusion about the large amount of terminology and disambiguate terms about Arabic morphology frequently used in the community. This chapter is not meant to be a complete reference but rather a detailed introduction to understanding challenges in Arabic morphology.

Computational Morphology Tasks. This chapter aims at introducing a set of common morphologically oriented tasks needed for several NLP applications. Examples of these tasks include morphological analysis, generation, tokenization, and POS tagging. It is important to note that this chapter also presents available tools that one can use for building NLP applications. For example, readers will find an introduction to tools such as BAMA, MADA (Roth et al. 2008), and AMIRA (Diab 2009, among others) that are widely used for processing Arabic morphology.

Arabic Syntax. This chapter gives a general survey of Arabic syntax and its specific constructions such as Idafa, Tamyiz, and so on. This chapter also discusses and compares three Arabic treebanking projects that are widely used in the community: Arabic Penn Treebank (Maamouri et al. 2004), Prague Dependency Treebank (Böhmová et al. 2000) and Columbia Treebank (Habash, Faraj, and Roth 2009). The author then summarizes a few research efforts on Arabic syntactic parsing.

A Note on Arabic Semantics. Because of the smaller amount of research in this area, this chapter is meant only to give a few remarks about Arabic semantics. It discusses the set of resources developed for Arabic computational semantic modeling but leaves out discussions of various theories and representations of semantics.

A Note on Arabic Machine Translation. Compared to previous chapters, this one addresses the multilingual aspects of working with Arabic. After briefly explaining the field of machine translation (MT), the author discusses Arabic linguistic features with MT in mind. The chapter also discusses available resources and presents recent advances in MT from and to Arabic.

Appendices. The book contains five interesting appendices. They are of great value to anyone who wants to conduct research or development in Arabic computational linguistics and NLP. As an example, Appendix A will help the reader become familiar with the available repositories and networking resources for Arabic NLP. Appendices C and D are particularly important because they discuss available resources: lexicons and tools, respectively.

To summarize, *Introduction to Arabic Natural Language Processing* stands as a clear, up-to-date, and concisely written introductory reference. The book does an excellent job of introducing the Arabic language to people who have an interest in working in the field of Arabic NLP. It is also a good source of information for accessing more detailed work on Arabic NLP applications through its bibliography.

References

Böhmová, Alena, Jan Hajič, Eva Hajičová, and Barbora Hladká. 2003. The PDT: A 3-level annotation scenario. In Anne Abeillé, editor, *Treebanks: Building and Using Parsed Corpora*. Kluwer Academic Publishers, Dordrecht, pages 103–127.

Diab, Mona. 2009. Second generation AMIRA tools for Arabic processing: Fast and robust tokenization, POS tagging, and base phrase chunking. In *2nd International Conference on Arabic Language Resources and Tools*, Cairo.

Farghaly, Ali and Khaled Shaalan. 2009. Arabic natural language processing: Challenges and solutions. *ACM Transactions on Asian Language Information Processing*, 8(4) article 14.

Habash, Nizar, Reem Faraj, and Ryan Roth. 2009. Syntactic annotation in the Columbia Arabic Treebank. In *2nd International Conference on Arabic Language Resources and Tools*, Cairo.

Maamouri, Mohamed, Ann Bies, Tim Buckwalter, and Wigdan Mekki. 2004. The Penn Arabic Treebank: Building a large-scale annotated Arabic corpus. In *NEMLAR International Conference on Arabic Language Resources and Tools*, Cairo.

Roth, Ryan, Owen Rambow, Nizar Habash, Mona Diab, and Cynthia Rudin. 2008. Arabic morphological tagging, diacritization, and lemmatization using lexeme models and feature ranking. In *Proceedings of the 46th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies, Short papers*, Columbus, OH, pages 117–120.

This book review was edited by Pierre Isabelle.

Imed Zitouni is a researcher and member of the multilingual natural language processing group at IBM T. J. Watson Research Center. He received his M.Sc. and Ph.D. from the University of Nancy 1 in 1996 and 2000, respectively. His research interests include language modeling, information extraction, machine translation, machine learning, and spoken dialogue systems; e-mail: izitouni@us.ibm.com.