Mach Translat
DOI 10.1007/s10590-011-9087-8

BOOK REVIEW

Nizar Y. Habash, Introduction to Arabic natural language processing (Synthesis lectures on human language technologies)

Morgan & Claypool, 2010, xiv + 166 pp

Khaled Shaalan

Received: 13 January 2011 / Accepted: 28 January 2011

Arabic is rich in morphology and syntax. It is normally written with optional diacritics and without the notion of capitalization. These characteristics make dealing with Arabic a challenge for both learners and researchers. My experience in the Arabic natural language processing (ANLP) research area allows me to say that the key to fostering a research or developing an application in the ANLP field lies in getting insights into the standard layer-based structure of linguistic phenomena (phonology, morphology, syntax and semantics) as well as in recognizing the interaction between them. Until now, there was no such an available introductory resource that can fulfill these requirements. For example, a simple *Google* search of the term "Arabic Natural Language Processing" results in research groups, research papers, tutorials and presentations, companies, and scholars. Consequently, tangible efforts had to be spent by any beginner of the ANLP field, whether a scientist, linguist, developer, or student, in going through different material which might be either irrelevant or too advanced. Therefore, the purpose and the significance of this book are clear from where it stands. Also, it is adequately classified by the publisher as belonging to the "human language technologies" series.

This book gives a sufficient solid introductory background on ANLP. This makes it the first of its kind. The author has a broad background in computational linguistics, in general, and ANLP, in particular. He also has a marvelous research track record and has been very active in serving the research communities. The book is clear about its intended audience. It is the most suitable for anyone who would like to get a fundamental background about ANLP for research, study, or development purposes. It is very well written. This book amazingly takes you gradually from the ground

K. Shaalan (✉)
Faculty of Engineering and Information Technology, The British University in Dubai,
UAE, P.O. Box 345015, Dubai, UAE
e-mail: khaled.shaalan@buid.ac.ae

Published online: 23 February 2011

up, beginning with the basics of the Arabic script and preprocessing tasks up to the syntactic and semantic representations and processing of the Arabic sentence. It does not only describe the concepts of Arabic linguistic knowledge but also explains some technical details necessary to achieve ANLP tasks. Arabic has a complex and rich morphology, which results in a very large vocabulary to cover. Morphology is central in working on Arabic NLP because of its important interactions with both orthography and syntax. An Arabic inflected verb can form a complete sentence, e.g. the verb "سمعتك" (heard-I-you) contains a complete syntactic structure in just a one-word sentence. Therefore, two comprehensive chapters are dedicated to dealing with all the important aspects of this subject. Across the book, the presentation, discussion, and well-motivated examples show very interesting content. The book presents both linguistic-knowledge-based and empirical based state-of-the-art ANLP approaches and relevant resources. Throughout the book, the balance in addressing these approaches is driven by the achievements of research in each approach. In my opinion, the level of detail is sufficient and would absolve the reader of the need to study any further foundation books on Arabic processing. Nevertheless, it enables the reader to smoothly proceed to advanced topics.

The book is organized in eight chapters. It begins with two introductory chapters about the historic and social levels of the Arabic language, and about the peculiarities of its script. The next five chapters contain sufficient cohesion material for the ANLP linguistic phenomena and their correlated tasks. The last chapter contains a supplementary material on the subject of Arabic machine translation. Appendices include very useful complementary material about resources, organizations, and terminology standards. The book cites very remarkable and motivating references in relevant sections. It is clear that the book targets beginners, who are expected to frequently search the book to find distant pages that talk about terms of their interest. As a result, I found that having an index of terms would have been helpful for these readers. In a few places, there are references to colored text (e.g. *A* in green in Fig. 2.6) which I found a little bit confusing given that my version of the book is printed in black and white with grayscale.

Details of each chapter follow. Chapter 1 gives a brief distinction of Arabic language variants and clarifies that the scope focuses on Modern Standard Arabic (MSA). MSA is the official written language of the Arab world. It is the primary language of the media and education. Although MSA is syntactically, morphologically and phonologically based on Classical Arabic, it has become more modern by dropping some aspects such as diacritics. It is made clear from the beginning that the scope of the book is neither dialectal Arabic nor Quranic (Classical) Arabic.

Chapter 2 deals with the Arab script that is used to write MSA text. The rules of handwriting, typesetting, and digitizing Arabic script are perfectly explained. The discussion sheds light on specific aspects of Arabic script that are essential for computational applications, including: (1) how to identify word segments, (2) how to combine (or edit) letters taking into consideration the fact that Arab letters have initial, medial, final and isolated variants, (3) how to choose or standardize the encoding format for representing, visualizing, and matching Arabic script, and (4) how to predict the phonemic or phonetic form of the MSA text for speech synthesis purposes. A significant addition by Habash and his colleagues is the contribution made to standardize Arabic

transliteration, the so-called Habash-Soudi-Buckwalter transliteration (HSB), which solves conflicts that might arise with some platforms. As the chapter starts with a note on Arabic variants, I think, it would have been worth to include the efforts in the task of mapping of such variants to MSA.

Chapter 3 is devoted to phonology and issues related to speech and spelling of Arabic. Although the book does not mean to present speech processing tasks, it is useful to get a background on the representation of the sound system of Arabic as well as on the process of handling it; especially, if we are going to interface with a speech processing input or output. I found that the part on Arabic orthography is very useful because it clarifies the mapping of letters to sound. Four natural language processing (NLP) tasks are discussed: name transliteration, spelling correction, speech recognition, and speech synthesis. It is worth noting that in some Arab countries, expatriates must use both their English and Arabic transliterated name in official documents. The lack of name transliteration standards and the nature of the language make the transliteration process a challenge for both Arabs and non-Arab expatriates due to transliteration variants which impact named entity recognition, machine translation and other NLP tasks. I agree with the author that the problem is catastrophic when it takes a political or security dimension worldwide, such that some people might be unjustly treated. The presentation of the chapter is analytical and points out a wide range of solutions. On the other hand, I found that the description of the spelling correction task is interesting; however, I must point out that the level of detail does not correlate with the importance of the topic. The section devoted to speech processing tasks is rather brief, as I believe it should be, although it contains an excellent abstracted description.

Chapters 4 and 5 on morphology, by far the most interesting (unsurprisingly, considering Prof. Habash's profile), and are adequately presented and motivated. As said above, Arabic morphology is a core linguistic phenomenon. An Arabic word is very rich in form and meaning. As such, it deserves a careful and detailed study. It is difficult to fully address the topic in one chapter which, in my opinion, justifies the division into two chapters. Chapter 4 presents in a really brilliant form a comprehensive analysis and discussions of various aspects of Arabic morphology. The presentation of approaches to Arabic morphology is well explained and accurate. Habash proposes two approaches:

- Form-based morphology, i.e. based on the individual word components, the relation between components and how they relate to the overall word form.
- Functional morphology, i.e. based on the function of the word component and its syntactic and semantic roles.

In this presentation Habash places concatenative morphology and templatic morphology in the first approach and cliticization, inflection, and derivation in the second. The discussion delves deep into the traditional and modern (as in natural language processing and computational linguistics literature) accounts of Arabic morphology trying to bridge the gaps in views. The presentation is an amazing, complete, and convenient approach to Arabic morphology. The rest of the chapter provides clear and useful insights into Arabic word morphology in a systematic way. The description of the inflectional and derivational morphology is fascinating, such that it gives the reader the necessary Arabic morphology background about nominal and verbal

forms, and their associated features. Nevertheless, complex and idiosyncratic morphological expressions are explained. The presentation has a quantitative flavor based on observations from Arabic treebanks and the web. Accordingly, indications made based on commonality and frequency.

Chapter 5 deals with different computational morphology tasks and subtasks. The chapter emphasizes those issues related to the practical implementation of morphological tools, including the resources needed for them. The presentation discusses what Habash considers as common computational morphology tasks: morphological analysis, morphological generation, tokenization, and part-of-speech tagging. This covers also enabling technologies that are required for full-fledged applications. The description of tasks is related to both approaches given in the previous chapter and the interrelation between these tasks. The topic is introduced from various dimensions: contextual vs non-contextual, shallow vs deep, coarse vs fine-grained, and analytical vs generative. This gives insights on different ways of representing and processing units of the Arabic word. I found the comparative analysis of the available tools very informative and helpful when taking decisions about the suitability of these tools to the task in question**.** A reasonable addition that would be welcome here and might be interesting to the reader is the analysis and correction of ill-formed Arabic word forms which, in my opinion, are very important to applications such as intelligent tutoring systems for analyzing and providing feedback on errors made by language learners.

Chapter 6 is devoted to syntax. The first part explains the structure of phrases and sentences in Arabic. The style of presentation is fascinating. It starts with general terms (i.e. phrases, constituents, etc.) and progressively proceeds to specific Arabic terms (e.g., *idafa* and *tamyiz*). The examples are very useful to introduce different aspects about the structure of the sentence and its constituents. For simplicity, the description does not follow any specific grammar formalism or approaches to syntax. Parsing of the Arabic sentence is a very complicated task. Efforts in rule-based parsing are still experimental. With new trends toward corpus-based/statistical sentence processing, syntactic processing has progressed substantially and more concrete results have been achieved. Therefore, it is not surprising that the presentation of the second part and mostly the last part are restricted to variants of Arabic treebanks as significant language resources for building and evaluating parsers. The Penn Arabic Treebank (PATB), the Prague Arabic Dependency Treebank (PADT), and the Columbia Arabic Treebank (CATiB) are briefly discussed. The comparison between the three treebanks is very useful to distinguish important differences between them in terms of syntactic representation and linguistics content. The ANLP tasks involving syntax are not explicitly addressed. Further readings and citations point out to a long list of references for interested readers.

Chapter 7 deals with semantics. The short presentation is due to the lack of research interest and achievements in this area. The chapter introduces general terminology about semantics. It is followed by useful information about ongoing research projects on the development of Arabic semantic resources. As Arabic semantics processing is still underdeveloped, this chapter departs considerably in that it does not include rich material as compared with previous chapters.

Chapter 8 is devoted to Arabic machine translation as a representative multilingual application. It is nice to have this application singled out due to its importance.

The chapter introduces general terminology about two major approaches of machine translation: rule-based and statistical-based machine translation. To demonstrate the importance of machine translation in a multilingual setting, a comparison of linguistic characteristic has been made between Arabic and three different languages (Chinese, English, and Spanish). This is motivated by the availability of bilingual/parallel corpora. Finally, a note is presented about the state of the field of Arabic machine translation. The content of this chapter, in my opinion, familiarizes the reader with multilingual processing involving Arabic.

The book is concise, as required by the publisher, but the contents are impressive. It takes care of very specific details such as the *hamza* letter mark (which can appear above or below specific letter forms; when it appears at stem-initial positions it tends to be perceived as a diacritic) and relates it to aspects of various linguistic phenomena. Moreover, the material in text boxes that frequently appear as alerts or FAQ's are very interesting teaching material that directly draws the attention of the reader's mind. In particular, I admired the highlight such as *FAQ: How true is "Arabic has no vowels"?* I found that the book follows a reasonable approach such that the reader can easily pursue the logical sequence not only across one chapter but also across the entire book. This means that the material is brilliantly organized in such away it covers the necessary breadth and depth of its intended audience. In my opinion, for anyone who wants to understand Arabic natural language processing, this book is indispensable.