Arabic Computational Morphology

Text, Speech and Language Technology

VOLUME 38

Arabic Computational Morphology

Knowledge-based and Empirical Methods

Edited by

Abdelhadi Soudi
Ecole Nationale de I'Industrie Minérale, Rabat, Morocco

Antal van den Bosch
Tilburg University, The Netherlands

Günter Neumann
Deutsches Forschungszentrum für Künstliche Intelligenz, Saarbrücken, Germany

Springer

A C.I.P. Catalogue record for this book is available from the Library of Congress.

ISBN 978-1-4020-6045-8 (HB)
ISBN 978-1-4020-6046-5 (e-book)

Published by Springer,
P.O. Box 17, 3300 AA Dordrecht, The Netherlands.

www.springer.com

Printed on acid-free paper

Contents

Preface

One of the advantages of having worked in a field for twenty years is that you have an opportunity to watch research areas grow from infancy into maturity. The present book represents a marriage of two such fields: computational morphology and Arabic computational linguistics.

In the mid 1980s, Koskenniemi had just published his landmark (1983) thesis on Two-Level Morphology. Prior to Koskenniemi there had of course been work on computational morphology dating back all the way to the 1960s, but the field had never been a major focus of research. Koskenniemi changed that by taking the computational framework of finite-state transducers, proposed by Ron Kaplan and Martin Kay, and making it actually work in a real system. Koskenniemi provided a practical implementation of a well-defined computational model, and this in turn led to an explosion of work in finite-state and other approaches to morphology over the ensuing twenty years. Data-driven methods, which gained popularity in the late 1980s took a few years to make an impact on computational morphology, but in the last decade there has been a significant amount of work, particularly in the area of self-organizing methods for morphological induction.

In the mid 1980s, with notable exceptions like early work by Beesley, there was next to nothing being done on Arabic. There simply were not the resources, nor were there very many people who both had the linguistic training and knowledge of Arabic, as well as training in natural language processing. All of this has changed. Now there are quite a few resources for Arabic including the roughly 400 million words of Modern Standard Arabic newswire text in the Arabic Gigaword corpus, the Penn Arabic Treebank, the Prague Dependency treebank, Tim Buckwalter's publicly available morphological analyzer, as well as a growing set of resources for Colloquial Arabic, including the Egyptian, Levantine, Iraqi and Gulf dialects. As evidenced by the contributors to this volume, there are now a large number of computational linguists with a knowledge of Arabic. And perhaps most importantly, there is a widespread interest in the community as a whole in Arabic language processing.

Like all good marriages, the union of computational morphology with Arabic language processing is one fraught with complexity; for Arabic seems almost to have been specially engineered to maximize the difficulties for automatic processing. The famous Semitic "root-and-pattern" morphology defies a straightforward implementation in terms of morpheme concatenation, and this has spawned a wide variety of different computational solutions, many of which are represented in various chapters in this volume. Students of writing systems have speculated that this

root-and-pattern morphology was ultimately responsible for the second interesting and difficult property of Arabic (and several other Semitic languages), namely that the writing system is impoverished in that a fair amount of phonological information is simply missing in the script. In its normal everyday use, the script systematically fails to represent not only most vowel information (is درس *DRS /darasa/, /durisa/*, ...?), but also information on consonant gemination (is كتب *KTB /kataba/* or */kattaba/*?), as well as both vowel and nunation information in the nominal case system (is ولد *WLD /waladu/, /waladun/, /waladin/*, ...?). If this weren't enough, as Tim Buckwalter shows, the advent of Unicode has failed to standardize Arabic encoding, so that in dealing with real texts, one has to be prepared to do a fair amount of low level normalization; to some extent the differences reflect regional variants (such as the use of */alif maqSūra/* for */ya/* in Egyptian texts), but in other cases they reflect the fact that for all of its attempts at rigid design, Unicode still allows for a fair amount of "wiggle room": the same issue comes up in the encoding of South Asian languages using Brahmi-derived scripts.

The chapters in this volume attest both to the wide variety, and to the sophistication of the work being done on the computational analysis of Arabic morphology, both in terms of approaches to morphological analysis, as well as in applications of such work to other areas such as machine translation and information retrieval. To be sure, part of the reason for the increased interest in Arabic language processing is due to greater funding opportunities for work on Arabic, and this in turn has been fueled by various important political events of the past few years. But it would be shortsighted to view this as the sole justification for an increased interest in Arabic. Forms of Arabic are spoken by roughly 250 million people in an area spanning North Africa to the Persian Gulf. It is the official language of over 20 countries. It is a significant minority language of a number of sub-saharan African countries. And there is a large expatriate population spread throughout the world. Arabic is thus one of the world's major languages. History, especially that of the last hundred years, has not been kind to the Arabic-speaking peoples, and they have not had an economic clout proportional to their population. This is bound to change sooner or later, and there will be an increasing need for tools that allow one to use Arabic in the digital world as easily as one can now use English.

The work represented in this book is an important milestone along the path towards that goal. I commend the editors — Abdelhadi Soudi, Antal van den Bosch, and Günter Neumann — and all of the contributing authors on its publication.

I wish to thank Elabbas Benmamoun for helpful feedback on an earlier draft of this preface.

Richard Sproat
University of Illinois at Urbana-Champaign