ACL-05

Computational Approaches to Semitic Languages

Workshop Proceedings

29 June 2005
University of Michigan
Ann Arbor, Michigan, USA

Production and Manufacturing by
Omnipress Inc.
Post Office Box 7214
Madison, WI 53707-7214

Order copies of this and other ACL proceedings from:

Association for Computational Linguistics (ACL)
75 Paterson Street, Suite 9
New Brunswick, NJ 08901
USA
Tel: +1-732-342-9100
Fax: +1-732-342-9339
acl@aclweb.org

Preface

Semitic Languages

The Semitic family includes many languages and dialects spoken by a large number of native speakers (around 300 Million). However, Semitic languages are still understudied. The most prominent members of this family are Arabic and its dialects, Hebrew, Amharic, Aramaic, Maltese and Syriac. Beyond their shared ancestry which is apparent through pervasive cognate sharing, a common characteristic of these languages is the rich and productive pattern-based morphology and similar syntactic constructions.

Previous Efforts

An increasing body of computational linguistics work is starting to appear for both Arabic and Hebrew. Arabic alone, as the largest member of the Semitic family, has been receiving a lot of attention lately in terms of dedicated workshops and conferences. These include, but are not limited to, the workshop on Arabic Language Resources and Evaluation (LREC 2002), a special session on Arabic processing in Traitement Automatique du Langage Naturel (TALN 2004), the Workshop on Computational Approaches to Arabic Script-based Languages (COLING 2004), and the NEMLAR Arabic Language Resources and Tools Conference in Cairo, Egypt (2004). This phenomenon has been coupled with a relative surge in resources for Arabic due to concerted efforts by the LDC and ELDA/ELRA. However, there is an apparent lag in the development of resources and tools for other Semitic languages. Often, work on individual Semitic languages, unfortunately, still tends to be done with limited awareness of ongoing research in other Semitic languages. Within the last four years, only three workshops addressed Semitic languages: an ACL 2002 Workshop on Computational Approaches to Semitic Languages and an MT Summit IX Workshop on Machine Translation for Semitic Languages in 2003, and the EAMT 2004, held in Malta, had a special session on Semitic languages.

Current Workshop

Welcome to the ACL 2005 Workshop on Computational Approaches to Semitic Languages. This workshop is a sequel to the ACL 2002 workshop and shares its goals of: (i) heightening awareness amongst Semitic-language researchers of shared breakthroughs and challenges, (ii) highlighting issues common to all Semitic languages as much as possible, (iii) encouraging the potential for developing coordinated approaches; and (iv) in addition, leveraging resource and tool creation for less prominent members of the Semitic language family.

We received 21 submissions, we accepted 12. The accepted papers cover several languages: Modern Standard Arabic, Dialectal Arabic, Hebrew, and Amharic. They cover a span of topics in computational linguistics, from morphological analysis and disambiguation and diacritization to information retrieval and document classification using both symbolic and statistical approaches.

We hope you enjoy reading this volume as much as we did.

The workshop organizers,

Kareem Darwish, Mona Diab, Nizar Habash

Organizers:

Kareem Darwish, German University in Cairo
Mona Diab, Columbia University Center for Computational Learning Systems
Nizar Habash, Columbia University Center for Computational Learning Systems

Program Committee:

Ibrahim A. Alkharashi (King Abdulaziz City for Science and Technology, Saudi Arabia)
Tim Buckwalter (Linguistic Data Consortium, USA)
Violetta Cavalli-Sforza (Carnegie Mellon University, USA)
Yaacov Choueka (Bar-Ilan University, Israel)
Joseph Dichy (Lyon University, France)
Martha Evens (Illinois Institute of Technology, USA)
Ali Farghaly (SYSTRAN Software, Inc.)
Alexander Fraser (USC/ISI)
Andrew Freeman (Mitre)
Alon Itai, (Technion, Israel)
George Kiraz (Beth Mardutho: The Syriac Institute, USA)
Katrin Kirchhoff (University of Washington, USA)
Alon Lavie (Carnegie Mellon University, USA)
Mohamed Maamouri (Linguistic Data Consortium, USA)
Uzzi Ornan (Technion, Israel)
Anne De Roeck (Open University, UK)
Michael Rosner (University of Malta, Malta)
Salim Roukos (IBM, USA)
Khalil Sima’an (University of Amsterdam, Netherlands)
Abdelhadi Soudi (ENIM, Rabat, Morocco)
Shuly Wintner (University of Haifa, Israel)
Remi Zajac (SYSTRAN Software, USA)

Invited Speaker:

Salim Roukos, IBM T. J. Watson Research Center

Table of Contents

Conference Program

Wednesday, June 29, 2005

9:00–9:15 Opening Remarks by Organizers

Session 1: Morphology

9:15–9:40 *Memory-based morphological analysis generation and part-of-speech tagging of Arabic*
Erwin Marsi, Antal van den Bosch and Abdelhadi Soudi

9:40–10:05 *A finite-state morphological grammar of Hebrew*
Shlomo Yona and Shuly Wintner

10:05–10:30 *Morphological Analysis and Generation for Arabic Dialects*
Nizar Habash, Owen Rambow and George Kiraz

10:30–11:00 Coffee Break

11:00–11:40 Invited Talk by Salim Roukos

Session 2: Applications I

11:40–12:05 *Examining the Effect of Improved Context Sensitive Morphology on Arabic Information Retrieval*
Kareem Darwish, Hany Hassan and Ossama Emam

12:05–12:30 *Modifying a Natural Language Processing System for European Languages to Treat Arabic in Information Processing and Information Retrieval Applications*
Gregory Grefenstette, Nasredine Semmar and Faiza Elkateb-Gara

12:30-2:15 Lunch Break

Wednesday, June 29, 2005 (continued)

Session 3: Part of Speech Tagging

2:15–2:40 *Choosing an Optimal Architecture for Segmentation and POS-Tagging of Modern Hebrew*
Roy Bar-Haim, Khalil Sima'an and Yoad Winter

2:40–3:05 *Part of Speech tagging for Amharic using Conditional Random Fields*
Sisay Fissaha Adafre

3:05–3:30 *POS Tagging of Dialectal Arabic: A Minimally Supervised Approach*
Kevin Duh and Katrin Kirchhoff

3:30–4:00 Coffee Break

Session 4: Applications II

4:00–4:25 *The Impact of Morphological Stemming on Arabic Mention Detection and Coreference Resolution*
Imed Zitouni, Jeffrey Sorensen, Xiaoqiang Luo and Radu Florian

4:25–4:50 *Classifying Amharic News Text Using Self-Organizing Maps*
Samuel Eyassu and Björn Gambäck

4:50–5:15 *Arabic Diacritization Using Weighted Finite-State Transducers*
Rani Nelken and Stuart M. Shieber

5:15–5:40 *An Integrated Approach for Arabic-English Named Entity Translation*
Hany Hassan and Jeffrey Sorensen

5:40–6:00 Parting Words & Discussion