2012 AMTA 20Years

The Tenth Biennial Conference of the
Association for Machine Translation in the Americas

Computational Approaches to Arabic Script-based Languages

Ali Farghaly

Farhard Oroumchian
University of Wollongong in Dubai,
United Arab Emirates

SAN DIEGO, CA
OCTOBER 28- NOVEMBER 1, 2012

The Fourth Workshop on Computational Approaches to Arabic Script-based Languages Proceedings

AMTA 2012 - San Diego, CA , USA

November 1, 2012

ALI FARGHALY

FARHAD OROUMCHIAN (Eds.)

Preface

This is the fourth of a series of workshops designed to bring together researchers working in all languages that use the Arabic script. The absence of short vowels and other diacritic marks from the Arabic script greatly compounds the ambiguity problem which challenges NLP applications. The historical interaction between the Arabic language and culture on the one hand and the other languages and cultures that adopted the Arabic script created a lasting commonality among all Arabic script-based languages. For example a named entity recognition system in any Arabic script-based language has to deal with the problem of the lack of capitalization, the absence of short vowels, and the lack of the strict format of names that is usually observed in Western names. For example, concepts such as last names, given names, maiden names, other name are not often adhered to in names of people in countries that use the Arabic scripts. This workshop dedicates a whole session to the discussion of name matching and named entity recognition.

Since this workshop is hosted by AMTA 2012, it is not surprising to see more than a third of the accepted papers deal directly with issues Arabic and Farsi machine translation. These papers deal with challenging problems in machine translation such as the translation of idiomatic and multi word expressions, the problem of translating discourse connectives and the issues encountered in the design of open domain machine translation systems for Farsi.

We look forward to our fifth workshop which we hope we have more papers on languages other than Arabic and more work that compares challenges and solutions in one task across different Arabic script-based languages.

November 2012 Ali Farghaly

ORGANIZATION

ORGANIZING COMMITTEE

ALI FARGHALY

STANFORD UNIVERSITY AND MONTEREY PENINSULA COLLEGE, USA

FARHAD OROUMCHIAN

UNIVERSITY OF WOLLONGONG AT DUBAI, UNITED ARAB EMIRATES

INVITED SPEAKER

HASSAN SAWAF

CHIEF SCIENTIST, SAIC, USA

TITLE OF THE TALK:

More Than 20 Years of Machine Translation of Arabic-Script Languages: Overview of the History of Diverse Challenges in Research and Deployment

PROGRAM COMMITTEE

TIM BUCKWALTER	UNIVERSITY OF MARYLAND, USA
VIOLETTA CAVALLI-SFORZA	AL AKHAWAYN UNIVERSITY, MOROCCOA
SHERRI CONDON	MITRE, USA
MONA DIAB	COLUMBIA UNIVERSITY, USA
SARMAD HUSSAIN	CRULP, PAKISTAN
FARHAD OROUMCHIAN	UNIVERSITY OF WOLLONGONG IN DUBAI, UAE
KHALED SHAALAN	THE BRITISH UNIVERSITY IN DUBAI,UAE
AHMED RAFEA	THE AMERICAN UNIVERSITY IN CAIRO, EGYPT
IMED ZITOUNI	IBM, USA

WORKSHOP PROGRAM

OPENING SESSION

9:00 – 9:30 ***Ali Farghaly, Organizer***

Commonalities in Arabic Script-based Languages: An Example from Name Matching

SESSION 1

9:30 – 10:30 ***Hassan Sawaf, Invited Speaker***

More than 20 years of Machine Translation of Arabic-Script Languages:

Overview of the History of Diverse Challenges in Research and Deployment"

10:30 – 11:00 BREAK

SESSION 2 MACHINE TRANSLATION

11:00 – 11:30 **Translating English Discourse Connectives into Arabic: a Corpus-based Analysis and an Evaluation Metric**

Najeh Hajlaoui and Andrei Popescu-Belis
Idiap Research Institute

11:30 – 12:00 **Idiomatic MWEs and Machine Translation. A Retrieval and Representation Model: the AraMWE Project**

Giuliano Lancioni and Marco Boella
Roma Tre University, Italy, 2Rome University "La Sapienza", Italy

12:00 - 12:30 **Developing an Open-domain English-Farsi Translation System Using AFEC: Amirkabir Bilingual Farsi-English Corpus**

Fattaneh Jabbari, Somayeh Bakhshaei, Seyed Mohammad Mohammadzadeh Ziabary and Shahram Khadivi

Amirkabir University of Technology, Terhan, Iran

12:30 – 2:00 LUNCH

SESSION 3 ENTITY RECOGNITION

2:00 – 2:30 **ARNE - A tool for Named Entity Recognition from Arabic text**

Carolin Shihadeh and Günter Neumann

DFKI, Saarbr¨ucken, Germany

2:30 – 3:00 **Approaches to Arabic Name Transliteration and Matching in a Software Knowledge Base**

Brant Kay and Brian Rineer
SAS Institute Inc., Cary, North Carolina, USA

3:00 – 3:30 **Using Arabic transliteration to improve word alignment from French-Arabic parallel corpora**

Houda Saadane, Nasredine Semmar, Ouafa Benterki and Christian Fluhr
LIDILEM - Université Stendhal Grenoble 3, Cedex, France
Institut Supérieur Arabe de Traduction, Bir Mourad Raïs, Algérie

3:30 – 4:00 BREAK

SESSION 4 SENTIMENTS AND MORPHOLOGICAL TAGGING

4:00 – 4:30 **Preprocessing Egyptian Dialect Tweets for Sentiment Mining**

Amira Shoukry and Ahmed Rafea
The American University in Cairo, Cairo, Eygpt

4:30– 5:00 **Rescoring N-Best Hypotheses for Arabic Speech Recognition: A Syntax-Mining Approach**

Dia Eddin AbuZeina, Moustafa Elshafei, Husni Al-Muhtaseb and Wasfi Al-Khatib
Palestine Polytechnic University, Hebron, Palestine
King Fahd University of Petroleum and Minerals, Saudi Arabia

5:00 - 5:30 **Morphological Segmentation and Part of Speech Tagging for Religious Arabic**

Emad Mohamed
Carnegie Mellon University Qatar

Table of Contents

Alternate Paper: