Third Workshop on NLP for Similar Languages, Varieties and Dialects 2016 (VarDial 3)

Osaka, Japan
12 December 2016

ISBN: 978-1-5108-3328-9

Printed from e-media with permission by:

Curran Associates, Inc.
57 Morehouse Lane
Red Hook, NY 12571

CURRAN ASSOCIATES INC.
proceedings
.com

Some format issues inherent in the e-media version may also appear in this print version.

Copyright© (2016) by the Association for Computational Linguistics
All rights reserved.

Printed by Curran Associates, Inc. (2016)

For permission requests, please contact the Association for Computational Linguistics
at the address below.

Association for Computational Linguistics
209 N. Eighth Street
Stroudsburg, Pennsylvania 18360

Phone: 1-570-476-8006
Fax: 1-570-476-0860

acl@aclweb.org

Additional copies of this publication are available from:

Curran Associates, Inc.
57 Morehouse Lane
Red Hook, NY 12571 USA
Phone: 845-758-0400
Fax: 845-758-2633
Email: curran@proceedings.com
Web: www.proceedings.com

Table of Contents

Discriminating between Similar Languages and Arabic Dialect Identification: A Report on the Third DSL Shared Task
Shervin Malmasi, Marcos Zampieri, Nikola Ljubešić, Preslav Nakov, Ahmed Ali and Jörg Tiedemann 1

Discriminating Similar Languages with Linear SVMs and Neural Networks
Çağrı Çöltekin and Taraka Rama 15

LSTM Autoencoders for Dialect Analysis
Taraka Rama and Çağrı Çöltekin 25

The GW/LT3 VarDial 2016 Shared Task System for Dialects and Similar Languages Detection
Ayah Zirikly, Bart Desmet and Mona Diab 33

Processing Dialectal Arabic: Exploiting Variability and Similarity to Overcome Challenges and Discover Opportunities
Mona Diab 42

Language Related Issues for Machine Translation between Closely Related South Slavic Languages
Maja Popović, Mihael Arcan and Filip Klubička 43

Romanized Berber and Romanized Arabic Automatic Language Identification Using Machine Learning
Wafia Adouane, Nasredine Semmar and Richard Johansson 53

How Many Languages Can a Language Model Model?
Robert Östling 62

Automatic Detection of Arabicized Berber and Arabic Varieties
Wafia Adouane, Nasredine Semmar, Richard Johansson and Victoria Bobicev 63

Automatic Verification and Augmentation of Multilingual Lexicons
Maryam Aminian, Mohamed Al-Badrashiny and Mona Diab 73

Faster Decoding for Subword Level Phrase-based SMT between Related Languages
Anoop Kunchukuttan and Pushpak Bhattacharyya 82

Subdialectal Differences in Sorani Kurdish
Shervin Malmasi 89

Enlarging Scarce In-domain English-Croatian Corpus for SMT of MOOCs Using Serbian
Maja Popović, Kostadin Cholakov, Valia Kordoni and Nikola Ljubešić 97

Arabic Dialect Identification in Speech Transcripts
Shervin Malmasi and Marcos Zampieri 106

DSL Shared Task 2016: Perfect Is The Enemy of Good Language Discrimination Through Expectation–Maximization and Chunk-based Language Model
Ondřej Herman, Vit Suchomel, Vít Baisa and Pavel Rychlý 114

Byte-based Language Identification with Deep Convolutional Networks
Johannes Bjerva 119

Classifying ASR Transcriptions According to Arabic Dialect
Abualsoud Hanani, Aziz Qaroush and Stephen Taylor 126

UnibucKernel: An Approach for Arabic Dialect Identification Based on Multiple String Kernels
Radu Tudor Ionescu and Marius Popescu 135

A Character-level Convolutional Neural Network for Distinguishing Similar Languages and Dialects
Yonatan Belinkov and James Glass 145

HeLI, a Word-Based Backoff Method for Language Identification
Tommi Jauhiainen, Krister Lindén and Heidi Jauhiainen 153

ASIREM Participation at the Discriminating Similar Languages Shared Task 2016
Wafia Adouane, Nasredine Semmar and Richard Johansson 163

Comparing Two Basic Methods for Discriminating Between Similar Languages and Varieties
Pablo Gamallo, Iñaki Alegria, José Ramom Pichel and Manex Agirrezabal 170

Advances in Ngram-based Discrimination of Similar Languages
Cyril Goutte and Serge Léger 178

Discrimination between Similar Languages, Varieties and Dialects using CNN- and LSTM-based Deep Neural Networks
Chinnappa Guggilla 185

Language and Dialect Discrimination Using Compression-Inspired Language Models
Paul McNamee 195

Arabic Language WEKA-Based Dialect Classifier for Arabic Automatic Speech Recognition Transcripts
Areej Alshutayri, Eric Atwell, Abdulrahman Alosaimy, James Dickins, Michael Ingleby and Janet Watson 204

An Unsupervised Morphological Criterion for Discriminating Similar Languages
Adrien Barbaresi 212

QCRI @ DSL 2016: Spoken Arabic Dialect Identification Using Textual Features
Mohamed Eldesouki, Fahim Dalvi, Hassan Sajjad and Kareem Darwish 221

Tuning Bayes Baseline for Dialect Detection
Hector-Hugo Franco-Penya and Liliana Mamani Sanchez 227

Vanilla Classifiers for Distinguishing between Similar Languages
Sergiu Nisioi, Alina Maria Ciobanu and Liviu P. Dinu 235

N-gram and Neural Language Models for Discriminating Similar Languages
Andre Cianflone and Leila Kosseim 243