

Fifth Workshop on NLP for Similar Languages, Varieties and Dialects (VarDial 2018)

Santa Fe, New Mexico, USA
20 August 2018

ISBN: 978-1-5108-6916-5

Printed from e-media with permission by:

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



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

Copyright© (2018) by the Association for Computational Linguistics
All rights reserved.

Printed by Curran Associates, Inc. (2018)

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

| | |
|---|-----|
| <i>Language Identification and Morphosyntactic Tagging: The Second VarDial Evaluation Campaign</i> Marcos Zampieri, Shervin Malmasi, Preslav Nakov, Ahmed Ali, Suwon Shon, James Glass, Yves Scherrer, Tanja Samardžić, Nikola Ljubešić, Jörg Tiedemann, Chris van der Lee, Stefan Grondelaers, Nelleke Oostdijk, Dirk Speelman, Antal van den Bosch, Ritesh Kumar, Bornini Lahiri and Mayank Jain | 1 |
| <i>Encoder-Decoder Methods for Text Normalization</i> Massimo Lusetti, Tatyana Ruzsics, Anne Göhring, Tanja Samardžić and Elisabeth Stark | 18 |
| <i>A High Coverage Method for Automatic False Friends Detection for Spanish and Portuguese</i> Santiago Castro, Jairo Bonanata and Aiala Rosá | 29 |
| <i>Sub-label dependencies for Neural Morphological Tagging – The Joint Submission of University of Colorado and University of Helsinki for VarDial 2018</i> Miikka Silfverberg and Senka Drobac | 37 |
| <i>Part of Speech Tagging in Luyia: A Bantu Macrolanguage</i> Kenneth Steimel | 46 |
| <i>Tübingen-Oslo Team at the VarDial 2018 Evaluation Campaign: An Analysis of N-gram Features in Language Variety Identification</i> Çağrı Çöltekin, Taraka Rama and Verena Blaschke | 55 |
| <i>Iterative Language Model Adaptation for Indo-Aryan Language Identification</i> Tommi Jauhiainen, Heidi Jauhiainen and Krister Lindén | 66 |
| <i>Language and the Shifting Sands of Domain, Space and Time (Invited Talk)</i> Timothy Baldwin | 76 |
| <i>UnibucKernel Reloaded: First Place in Arabic Dialect Identification for the Second Year in a Row</i> Andrei Butnaru and Radu Tudor Ionescu | 77 |
| <i>Varying image description tasks: spoken versus written descriptions</i> Emiel van Miltenburg, Ruud Koolen and Emiel Kraemer | 88 |
| <i>Transfer Learning for British Sign Language Modelling</i> Boris Mocialov, Helen Hastie and Graham Turner | 101 |
| <i>Paraphrastic Variance between European and Brazilian Portuguese</i> Anabela Barreiro and Cristina Mota | 111 |
| <i>Character Level Convolutional Neural Network for Arabic Dialect Identification</i> Mohamed Ali | 122 |
| <i>Neural Network Architectures for Arabic Dialect Identification</i> Elise Michon, Minh Quang Pham, Josep Crego and Jean Senellart | 128 |
| <i>HeLI-based Experiments in Discriminating Between Dutch and Flemish Subtitles</i> Tommi Jauhiainen, Heidi Jauhiainen and Krister Lindén | 137 |

| | |
|--|-----|
| <i>Measuring language distance among historical varieties using perplexity. Application to European Portuguese.</i> | |
| José Ramom Pichel Campos, Pablo Gamallo and Iñaki Alegria | 145 |
| <i>Comparing CRF and LSTM performance on the task of morphosyntactic tagging of non-standard varieties of South Slavic languages</i> | |
| Nikola Ljubešić | 156 |
| <i>Computationally efficient discrimination between language varieties with large feature vectors and regularized classifiers</i> | |
| Adrien Barbaresi | 164 |
| <i>Character Level Convolutional Neural Network for German Dialect Identification</i> | |
| Mohamed Ali | 172 |
| <i>Discriminating between Indo-Aryan Languages Using SVM Ensembles</i> | |
| Alina Maria Ciobanu, Marcos Zampieri, Shervin Malmasi, Santanu Pal and Liviu P. Dinu..... | 178 |
| <i>IIT (BHU) System for Indo-Aryan Language Identification (ILI) at VarDial 2018</i> | |
| Divyanshu Gupta, Gourav Dhakad, Jayprakash Gupta and Anil Kumar Singh..... | 185 |
| <i>Exploring Classifier Combinations for Language Variety Identification</i> | |
| Tim Kreutz and Walter Daelemans | 191 |
| <i>Identification of Differences between Dutch Language Varieties with the VarDial2018 Dutch-Flemish Subtitle Data</i> | |
| Hans van Halteren and Nelleke Oostdijk | 199 |
| <i>Birzeit Arabic Dialect Identification System for the 2018 VarDial Challenge</i> | |
| Rabee Naser and Abualsoud Hanani | 210 |
| <i>Twist Bytes - German Dialect Identification with Data Mining Optimization</i> | |
| Fernando Benites, Ralf Grubenmann, Pius von Däniken, Dirk von Grünigen, Jan Deriu and Mark Cieliebak | 218 |
| <i>STEVENDU2018's system in VarDial 2018: Discriminating between Dutch and Flemish in Subtitles</i> | |
| Steven Du and Yuan Yuan Wang | 228 |
| <i>Using Neural Transfer Learning for Morpho-syntactic Tagging of South-Slavic Languages Tweets</i> | |
| Sara Meftah, Nasredine Semmar, Fatiha Sadat and Stephan Raaijmakers | 235 |
| <i>When Simple n-gram Models Outperform Syntactic Approaches: Discriminating between Dutch and Flemish</i> | |
| Martin Kroon, Masha Medvedeva and Barbara Plank | 244 |
| <i>HeLI-based Experiments in Swiss German Dialect Identification</i> | |
| Tommi Jauhiainen, Heidi Jauhiainen and Krister Lindén | 254 |
| <i>Deep Models for Arabic Dialect Identification on Benchmarked Data</i> | |
| Mohamed Elaraby and Muhammad Abdul-Mageed | 263 |
| <i>A Neural Approach to Language Variety Translation</i> | |
| Marta R. Costa-jussà, Marcos Zampieri and Santanu Pal | 275 |
| <i>Character Level Convolutional Neural Network for Indo-Aryan Language Identification</i> | |
| Mohamed Ali | 283 |

| | |
|---|-----|
| <i>German Dialect Identification Using Classifier Ensembles</i> | |
| Alina Maria Ciobanu, Shervin Malmasi and Liviu P. Dinu | 288 |