Australasian Language Technology Association Workshop (ALTA 2016)

Caulfield, Australia
5 – 7 December 2016

Editor:

Trevor Cohn

ISBN: 978-1-5108-3316-6
Contents

Invited talks 1

Tutorials 4

Long papers 6

Improving Neural Translation Models with Linguistic Factors
Cong Duy Vu Hoang, Reza Haffari and Trevor Cohn 7

Evaluation of Medical Concept Annotation Systems on Clinical Records
Hamed Hassanzadeh, Anthony Nguyen and Bevan Koopman 15

The Benefits of Word Embeddings Features for Active Learning in Clinical Information Extraction
Mahnoosh Kholghi, Lance De Vine, Laurianne Sitbon, Guido Zuccon and Anthony Nguyen 25

Syndromic Surveillance using Generic Medical Entities on Twitter
Pin Huang, Andrew MacKinlay and Antonio Jimeno Yepes 35

Syndromic Surveillance through Measuring Lexical Shift in Emergency Department Chief Complaint Texts
Hafsah Aamer, Bahadorreza Ofoghi and Karin Verspoor 45

Unsupervised Pre-training With Seq2Seq Reconstruction Loss for Deep Relation Extraction Models
Zhuang Li, Lizhen Qu, Qiongkai Xu and Mark Johnson 54

ASM Kernel: Graph Kernel using Approximate Subgraph Matching for Relation Extraction
Nagesh C Panyam, Karin Verspoor, Trevor Cohn and Rao Kotagiri 65

N-ary Biographical Relation Extraction using Shortest Path Dependencies
Gitansh Khirbat, Jianzhong Qi and Rui Zhang 74

Phonotactic Modeling of Extremely Low Resource Languages
Andrei Shcherbakov, Ekaterina Vylomova and Nick Thieberger 84

The Role of Features and Context on Suicide Ideation Detection
Yufei Wang, Stephen Wan and Cecile Paris 94

Featureless Domain-Specific Term Extraction with Minimal Labelled Data
Rui Wang, Wei Liu and Chris McDonald 103

An Effect of Background Population Sample Size on the Performance of a Likelihood Ratio-based Forensic Text Comparison System: A Monte Carlo Simulation with Gaussian Mixture Model
Shunichi Ishihara 113

Short papers 122

How Challenging is Sarcasm versus Irony Classification?: A Study With a Dataset from English Literature
Aditya Joshi, Vaibhav Tripathi, Pushpak Bhattacharyya, Mark Carman, Meghna Singh, Jaya Saraswati and Rajita Shukla 123

Learning cascaded latent variable models for biomedical text classification
Ming Liu, Gholamreza Haffari and Wray Buntine 128

Temporal Modelling of Geospatial Words in Twitter
Bo Han, Antonio Jimeno Yepes, Andrew MacKinlay and Lianhua Chi 133

NER for Medical Entities in Twitter using Sequence to Sequence Neural Networks
Antonio Jimeno Yepes and Andrew MacKinlay 138

An empirical study for Vietnamese dependency parsing
Dat Quoc Nguyen, Mark Dras and Mark Johnson 143

:telephone::person::sailboat::whale::okhand: ; or â€œCall me Ishmaelâ€ How do you translate emoji?
Will Radford, Ben Hachey, Bo Han and Andy Chisholm 150

Presenting a New Dataset for the Timeline Generation Problem
Xavier Holt, Will Radford and Ben Hachey 155

ALTA Shared Task papers 160

Overview of the 2016 ALTA Shared Task: Cross-KB Coreference
Andrew Chisholm, Ben Hachey and Diego Mollá 161

Disambiguating Entities Referred by Web Endpoints using Tree Ensembles
Gitansh Khirbat, Jianzhong Qi and Rui Zhang 165

Filter and Match Approach to Pair-wise Web URI Linking
S. Shivashankar, Yitong Li and Afshin Rahimi 170

Pairwise FastText Classifier for Entity Disambiguation
Cheng Yu, Bing Chu, Rohit Ram, James Aichinger, Lizhen Qu and Hanna Suominen 175