ACL-IJCNLP 2015

ACL 2015
Workshop on Noisy User-generated Text

Proceedings of the Workshop

July 31, 2015
Beijing, China

Sponsors

©2015 The Association for Computational Linguistics

Order print-on-demand copies from:

Curran Associates
57 Morehouse Lane
Red Hook, New York 12571
USA
Tel: +1-845-758-0400
Fax: +1-845-758-2633
curran@proceedings.com

ISBN 978-1-941643-69-3

Introduction

The WNUT 2015 workshop focuses on a core set of natural language processing tasks on top of noisy user-generated text, such as that found on social media, web forums and online reviews. Recent years have seen a significant increase of interest in these areas. The internet has democratized content creation leading to an explosion of informal user-generated text, publicly available in electronic format, motivating the need for NLP on noisy text to enable new data analytics applications. The workshop is an opportunity to bring together researchers interested in noisy text with different backgrounds and encourage crossover.

The workshop this year features two shared tasks, (a) Text Normalization and (b) Twitter Named Entity Recognition, to facilitate comparison of different approaches and help advance the state of the art. Because this is a fast-moving area, there is a lack of standardized datasets, and papers published in the same year may not compare against each other. By organizing these shared tasks we hope to help develop standardized evaluations and promote research on NLP in noisy text.

The program this year includes 8 papers in the main track, 8 system description papers in the Twitter Named Entity Recognition track, and 9 system description papers in the Text Normalization track. All the papers are presented as short talks and as well as posters. There are also 4 invited speakers, Tim Baldwin, Brendan O'Connor, Anders Søgaard and Joel Tetreault, with each of their talks covering a different aspect of NLP for user-generated text.

We would like to thank the Program Committee members who reviewed the papers this year. We would also like to thank the workshop participants. Last, a word of thanks also goes to our two sponsors: Microsoft Research and IBM Research.

Wei Xu, Bo Han and Alan Ritter
Co-Organizers

Organizers:

Wei Xu (University of Pennsylvania)
Bo Han (IBM Research)
Alan Ritter (The Ohio State University)

Program Committee:

David Bamman (Carnegie Mellon University)
Kalina Bontcheva (University of Sheffield)
Claire Cardie (Cornell University)
Colin Cherry (National Research Council Canada)
Grzegorz Chrupała (Tilburg University)
Leon Derczynski (University of Sheffield)
Jacob Eisenstein (Georgia Institute of Technology)
Jennifer Foster (Dublin City University)
Eric Fosler-Lussier (The Ohio State University)
Kevin Gimpel (Toyota Technological Institute at Chicago)
Weiwei Guo (Columbia University)
Dirk Hovy (University of Copenhagen)
Jing Jiang (Singapore Management University)
Emre Kiciman (Microsoft Research)
Wang Ling (Carnegie Mellon University)
Xiaohua Liu (University of Montreal)
Preslav Nakov (Qatar Computing Research Institute)
Miles Osborne (Bloomberg)
Kristen Parton (Facebook)
Ellie Pavlick (University of Pennsylvania)
Daniel Preoţiuc-Pietro (University of Pennsylvania)
Roi Reichart (Technion-IIT)
Alla Rozovskaya (Columbia University)
Nathan Schneider (University of Edinburgh)
Djamé Seddah (University Paris-Sorbonne)
Richard Sproat (Google)
Maosong Sun (Tsinghua University)
Oren Tsur (Harvard University)
Benjamin Van Durme (Johns Hopkins University)
Svitlana Volkova (Johns Hopkins University)
Lu Wang (Cornell University)
Jun-Ming Xu (University of Wisconsin-Madison)
Xiaojin Zhu (University of Wisconsin-Madison)

Invited Speakers:

Tim Baldwin (The University of Melbourne)
Brendan O'Connor (University of Massachusetts Amherst)
Anders Søgaard (University of Copenhagen)
Joel Tetreault (Yahoo! Research)

Table of Contents

Minority Language Twitter: Part-of-Speech Tagging and Analysis of Irish Tweets
Teresa Lynn, Kevin Scannell and Eimear Maguire 1

Challenges of studying and processing dialects in social media
Anna Jørgensen, Dirk Hovy and Anders Søgaard 9

Toward Tweets Normalization Using Maximum Entropy
Mohammad Arshi Saloot, Norisma Idris, Liyana Shuib, Ram Gopal Raj and AiTi Aw 19

Five Shades of Noise: Analyzing Machine Translation Errors in User-Generated Text
Marlies van der Wees, Arianna Bisazza and Christof Monz 28

A Normalizer for UGC in Brazilian Portuguese
Magali Sanches Duran, Maria das Graças Volpe Nunes and Lucas Avanço 38

USFD: Twitter NER with Drift Compensation and Linked Data
Leon Derczynski, Isabelle Augenstein and Kalina Bontcheva 48

NRC: Infused Phrase Vectors for Named Entity Recognition in Twitter
Colin Cherry, Hongyu Guo and Chengbi Dai 54

IITP: Multiobjective Differential Evolution based Twitter Named Entity Recognition
Md Shad Akhtar, Utpal Kumar Sikdar and Asif Ekbal 61

Data Adaptation for Named Entity Recognition on Tweets with Features-Rich CRF
Tian Tian, Marco Dinarelli and Isabelle Tellier 68

Hallym: Named Entity Recognition on Twitter with Word Representation
Eun-Suk Yang and Yu-Seop Kim 72

IHS_RD: Lexical Normalization for English Tweets
Dmitry Supranovich and Viachaslau Patsepnia 78

Bekli:A Simple Approach to Twitter Text Normalization.
Russell Beckley 82

NCSU-SAS-Ning: Candidate Generation and Feature Engineering for Supervised Lexical Normalization
Ning Jin 87

DCU-ADAPT: Learning Edit Operations for Microblog Normalisation with the Generalised Perceptron
Joachim Wagner and Jennifer Foster 93

LYSGROUP: Adapting a Spanish microtext normalization system to English.
Yerai Doval Mosquera, Jesús Vilares and Carlos Gómez-Rodríguez 99

IITP: Hybrid Approach for Text Normalization in Twitter
Md Shad Akhtar, Utpal Kumar Sikdar and Asif Ekbal 106

NCSU_SAS_WOOKHEE: A Deep Contextual Long-Short Term Memory Model for Text Normalization
Wookhee Min and Bradford Mott 111

USZEGED: Correction Type-sensitive Normalization of English Tweets Using Efficiently Indexed n-gram Statistics
Gábor Berend and Ervin Tasnádi 120

Shared Tasks of the 2015 Workshop on Noisy User-generated Text: Twitter Lexical Normalization and Named Entity Recognition
Timothy Baldwin, Marie-Catherine de Marneffe, Bo Han, Young-Bum Kim, Alan Ritter and Wei Xu 126

Enhancing Named Entity Recognition in Twitter Messages Using Entity Linking
Ikuya Yamada, Hideaki Takeda and Yoshiyasu Takefuji 136

Improving Twitter Named Entity Recognition using Word Representations
Zhiqiang Toh, Bin Chen and Jian Su 141

Multimedia Lab @ ACL WNUT NER Shared Task: Named Entity Recognition for Twitter Microposts using Distributed Word Representations
Fréderic Godin, Baptist Vandersmissen, Wesley De Neve and Rik Van de Walle 146

NCSU_SAS_SAM: Deep Encoding and Reconstruction for Normalization of Noisy Text
Samuel Leeman-Munk, James Lester and James Cox 154

Conference Program

Friday, July 31, 2015

9:00–10:30 Invited Talks

9:00–9:45 *Text Mining of Social Media: Going beyond the Text and Only the Text*
Tim Baldwin

9:45–10:30 *Where is Language?*
Anders Søgaard

10:30–11:00 ***Coffee Break***

11:00–12:30 Long Papers and Abstracts

11:00–11:15 *Learning finite state word representations for unsupervised Twitter adaptation of POS taggers*
Julie Wulff and Anders Søgaard

11:15–11:30 *Towards POS Tagging for Arabic Tweets*
Fahad Albogamy and Allan Ramasy

11:30–11:45 *Minority Language Twitter: Part-of-Speech Tagging and Analysis of Irish Tweets*
Teresa Lynn, Kevin Scannell and Eimear Maguire

11:45–11:00 *Challenges of studying and processing dialects in social media*
Anna Jørgensen, Dirk Hovy and Anders Søgaard

12:00–12:15 *Toward Tweets Normalization Using Maximum Entropy*
Mohammad Arshi Saloot, Norisma Idris, Liyana Shuib, Ram Gopal Raj and AiTi Aw

12:15–12:30 *Five Shades of Noise: Analyzing Machine Translation Errors in User-Generated Text*
Marlies van der Wees, Arianna Bisazza and Christof Monz

Friday, July 31, 2015 (continued)

12:30–14:00 Poster Session and Lunch

Learning finite state word representations for unsupervised Twitter adaptation of POS taggers
Julie Wulff and Anders Søgaard

Towards POS Tagging for Arabic Tweets
Fahad Albogamy and Allan Ramasy

Minority Language Twitter: Part-of-Speech Tagging and Analysis of Irish Tweets
Teresa Lynn, Kevin Scannell and Eimear Maguire

Challenges of studying and processing dialects in social media
Anna Jørgensen, Dirk Hovy and Anders Søgaard

Toward Tweets Normalization Using Maximum Entropy
Mohammad Arshi Saloot, Norisma Idris, Liyana Shuib, Ram Gopal Raj and AiTi Aw

Five Shades of Noise: Analyzing Machine Translation Errors in User-Generated Text
Marlies van der Wees, Arianna Bisazza and Christof Monz

A Normalizer for UGC in Brazilian Portuguese
Magali Sanches Duran, Maria das Graças Volpe Nunes and Lucas Avanço

USFD: Twitter NER with Drift Compensation and Linked Data
Leon Derczynski, Isabelle Augenstein and Kalina Bontcheva

Enhancing Named Entity Recognition in Twitter Messages Using Entity Linking
Ikuya Yamada, Hideaki Takeda and Yoshiyasu Takefuji

Improving Twitter Named Entity Recognition using Word Representations
Zhiqiang Toh, Bin Chen and Jian Su

NRC: Infused Phrase Vectors for Named Entity Recognition in Twitter
Colin Cherry, Hongyu Guo and Chengbi Dai

IITP: Multiobjective Differential Evolution based Twitter Named Entity Recognition
Md Shad Akhtar, Utpal Kumar Sikdar and Asif Ekbal

Friday, July 31, 2015 (continued)

Multimedia Lab @ ACL WNUT NER Shared Task: Named Entity Recognition for Twitter Microposts using Distributed Word Representations
Fréderic Godin, Baptist Vandersmissen, Wesley De Neve and Rik Van de Walle

Data Adaptation for Named Entity Recognition on Tweets with Features-Rich CRF
Tian Tian, Marco Dinarelli and Isabelle Tellier

Hallym: Named Entity Recognition on Twitter with Word Representation
Eun-Suk Yang and Yu-Seop Kim

IHS_RD: Lexical Normalization for English Tweets
Dmitry Supranovich and Viachaslau Patsepnia

Bekli:A Simple Approach to Twitter Text Normalization.
Russell Beckley

NCSU-SAS-Ning: Candidate Generation and Feature Engineering for Supervised Lexical Normalization
Ning Jin

DCU-ADAPT: Learning Edit Operations for Microblog Normalisation with the Generalised Perceptron
Joachim Wagner and Jennifer Foster

LYSGROUP: Adapting a Spanish microtext normalization system to English.
Yerai Doval Mosquera, Jesús Vilares and Carlos Gómez-Rodríguez

IITP: Hybrid Approach for Text Normalization in Twitter
Md Shad Akhtar, Utpal Kumar Sikdar and Asif Ekbal

NCSU_SAS_WOOKHEE: A Deep Contextual Long-Short Term Memory Model for Text Normalization
Wookhee Min and Bradford Mott

NCSU_SAS_SAM: Deep Encoding and Reconstruction for Normalization of Noisy Text
Samuel Leeman-Munk, James Lester, and James Cox

USZEGED: Correction Type-sensitive Normalization of English Tweets Using Efficiently Indexed n-gram Statistics
Gábor Berend and Ervin Tasnádi

Friday, July 31, 2015 (continued)

14:00–15:30 Shared Task Session

14:00–14:30 *Shared Tasks of the 2015 Workshop on Noisy User-generated Text: Twitter Lexical Normalization and Named Entity Recognition*
Timothy Baldwin, Marie-Catherine de Marneffe, Bo Han, Young-Bum Kim, Alan Ritter and Wei Xu

14:30–14:45 *Enhancing Named Entity Recognition in Twitter Messages Using Entity Linking*
Ikuya Yamada, Hideaki Takeda and Yoshiyasu Takefuji

14:45–15:00 *Improving Twitter Named Entity Recognition using Word Representations*
Zhiqiang Toh, Bin Chen and Jian Su

15:00–15:15 *Multimedia Lab @ ACL WNUT NER Shared Task: Named Entity Recognition for Twitter Microposts using Distributed Word Representations*
Fréderic Godin, Baptist Vandersmissen, Wesley De Neve and Rik Van de Walle

15:15–15:30 *NCSU_SAS_SAM: Deep Encoding and Reconstruction for Normalization of Noisy Text*
Samuel Leeman-Munk, James Lester and James Cox

15:30–16:00 ***Coffee Break***

16:00–17:30 Invited Talks

16:00–16:45 *Automated Grammatical Error Correction for Language Learners: Where are we, and where do we go from there?*
Joel Tetreault

16:45–17:30 *Are Minority Dialects "Noisy Text"?: Implications of Social and Linguistic Diversity for Social Media NLP*
Brendan O'Connor