

# **Australasian Language Technology Association Workshop 2017**

**Proceedings of the Workshop**



**Editors:**

**Jojo Sze-Meng Wong  
Gholamreza Haffari**

**6–8 December 2017  
Queensland University of Technology  
Brisbane, Australia**

# Australasian Language Technology Association Workshop 2017

(ALTA 2017)

<http://alta2017.alta.asn.au>

Online Proceedings:

<http://alta2017.alta.asn.au/proceedings>

## Gold Sponsors:



## Silver Sponsors:



## Bronze Sponsors:



# **ALTA 2017 Workshop Committees**

## **Workshop Chairs**

- Stephen Wan (Data61)
- Jojo Sze-Meng Wong (Monash University)

## **Workshop Programme Chairs**

- Jojo Sze-Meng Wong (Monash University)
- Gholamreza Haffari (Monash University)

## **Programme Committee**

- Oliver Adams, University of Melbourne
- Timothy Baldwin, University of Melbourne
- Benjamin Borschinger, Google
- Julian Brooke, University of Melbourne
- Lawrence Cavedon, RMIT University
- Trevor Cohn, The University of Melbourne
- Nathalie Colineau, DST Australia
- Mark Dras, Macquarie University
- Dominique Estival, Western Sydney University
- Gabriela Ferraro, CSIRO Data61
- Hamed Hassanzadeh, The Australian e-Health Research Centre
- Nitin Indurkha, University of New South Wales
- Sarvnaz Karimi, CSIRO Data61
- Mac Kim, CSIRO Data61
- Yitong Li, The University of Melbourne
- Teresa Lynn, Dublin City University
- Andrew MacKinlay, IBM Research
- Diego Mollá-Alliod, Macquarie University
- Anthony Nguyen, The Australian e-Health Research Centre
- Bahadorreza Ofoghi, The University of Melbourne
- Sylvester Orimaye, East Tennessee State University
- Cécile Paris, CSIRO Data61
- Matthias Petri, The University of Melbourne
- Lizhen Qu, CSIRO Data61
- Will Radford, Red Marker
- Abeed Sarker, University of Pennsylvania
- Rolf Schwitter, Macquarie University
- Ehsan Shareghi, Monash University
- Laurianne Sitbon, Queensland University of Technology
- Hanna Suominen, The Australian National University
- Karin Verspoor, The University of Melbourne
- Wei Wang, University of New South Wales
- Ingrid Zukerman, Monash University

## Preface

This volume contains the papers accepted for presentation at the Australasian Language Technology Association Workshop (ALTA) 2017, held at Queensland University of Technology in Brisbane, Australia on 6–8 December 2017.

The goals of the workshop are to:

- bring together the Language Technology (LT) community in the Australasian region and encourage interactions and collaboration;
- foster interaction between academic and industrial researchers, to encourage dissemination of research results;
- provide a forum for students and young researchers to present their research;
- facilitate the discussion of new and ongoing research and projects;
- increase visibility of LT research in Australasia and overseas and encourage interactions with the wider international LT community.

This year's ALTA Workshop presents 13 peer-reviewed papers, including 10 long papers and 3 short papers. We received a total of 23 submissions for long and short papers. Each paper was reviewed by three members of the program committee, using a double-blind protocol. Great care was taken to avoid all conflicts of interest.

ALTA 2017 includes a presentations track, following the workshops since 2015 when it was first introduced. This aims to encourage broader participation and facilitate local socialisation of international results, including work in progress and work submitted or published elsewhere. Presentations were lightly reviewed by the ALTA chairs to gauge overall quality of work and whether it would be of interest to the ALTA community. Offering both archival and presentation tracks allows us to grow the standard of work at ALTA, to better showcase the excellent research being done locally.

ALTA 2017 continues the tradition of including a shared task, this year on correcting OCR errors. Participation is summarised in an overview paper by organisers Diego Mollá-Alliod and Steve Cassidy. Participants were invited to submit a system description paper, which are included in this volume without review.

We would like to thank, in no particular order: all of the authors who submitted papers; the programme committee for the time and effort they put into maintaining the high standards of our reviewing process; the co-chair Stephen Wan for coordinating the logistics that go into running the workshop, from arranging the space, catering, budgets, sponsorship and more; the shared task organisers Diego Mollá and Steve Cassidy; our keynote speakers Lewis Mitchell and Robert Dale for agreeing to share their perspectives on the state of the field; and the tutorial presenter Ben Hachey for his efforts towards the three parts of the tutorial. We would like to acknowledge the constant support and advice of the ALTA Executive Committee.

Finally, we gratefully recognise our sponsors: Capital Markets CRC, Sintelix, Google, CSIRO/Data61 and Queensland University of Technology. Importantly, their generous support enabled us to offer travel subsidies to all students presenting at ALTA, and helped to subsidise conference catering costs and student paper awards.

Jojo Sze-Meng Wong  
Gholamreza Haffari

ALTA Programme Chairs

## ALTA 2017 Programme

### Wednesday, 6 December 2017

---

\*Tutorial Session 1 (Monash Caulfield, B214)

13:00–17:00      Tutorial: Ben Hachey  
*Active Learning ... and Beyond!*

13:00–14:15      *Part 1: From Zero to Hero*

14:15–14:30      Break

14:30–15:45      *Part 2: Live Shared Task*

15:45–16:00      Break

16:00–17:00      *Part 3: Wild Blue Yonder*

---

## Thursday, 7 December 2017

---

|   |   |
|---|---|
| Opening & Keynote (Room P421)                                 |   |
| 9:00–9:15   | Opening   |
| 9:15–10:15  | Keynote 1 (from ADCS): Dan Russell<br><i>What do you really need to know? Learning and knowing in the age of the Internet</i>   |
| <hr/>   |   |
| 10:15–10:45   | Morning tea   |
| <hr/>   |   |
| Session 1: Machine Learning and Applications (Room P521)      |   |
| 10:45–11:05   | Paper: Leonardo Dos Santos Pinheiro and Mark Dras<br><i>Stock Market Prediction with Deep Learning: A Character-based Neural Language Model for Event-based Trading</i>         |
| 11:05–11:25   | Paper: Fei Liu, Trevor Cohn and Timothy Baldwin<br><i>Improving End-to-End Memory Networks with Unified Weight Tying</i>  |
| 11:25–11:45   | Paper Shivashankar Subramanian, Trevor Cohn, Timothy Baldwin and Julian Brooke<br><i>Joint Sentence-Document Model for Manifesto Text Analysis</i>                              |
| 11:45–12:05   | Paper: Ming Liu, Gholamreza Haffari, Wray Buntine and Michelle Ananda-Rajah<br><i>Leveraging Linguistic Resources for Improving Neural Text Classification</i>                  |
| 12:05–12:15   | Paper: Hamideh Hajiabadi, Diego Molla-Aliod and Reza Monsefi<br><i>On Extending Neural Networks with Loss Ensembles for Text Classification</i>                                 |
| <hr/>   |   |
| 12:15–13:15   | Lunch   |
| <hr/>   |   |
| Session 2 & Keynote (Room P421)                               |   |
| 13:15–14:15   | Keynote 2: Lewis Mitchell<br><i>What do you really need to know? Learning and knowing in the age of the Internet</i>  |
| 14:15–14:30   | Paper: Shiwei Zhang, Xiuzhen Zhang and Jeffrey Chan (ADCS short paper)<br><i>A Word-Character Convolutional Neural Network for Language-Agnostic Twitter Sentiment Analysis</i> |
| 14:30–14:45   | Paper: Lance De Vine, Shlomo Geva and Peter Bruza (ADCS short paper)<br><i>Efficient Analogy Completion with Word Embedding Clusters</i>  |
| 14:45–15:05   | Paper: Aili Shen, Jianzhong Qi and Timothy Baldwin<br><i>A Hybrid Model for Quality Assessment of Wikipedia Articles</i>  |
| 15:05–15:15   | Paper: Diego Molla-Aliod<br><i>Towards the Use of Deep Reinforcement Learning with Global Policy For Query-based Extractive Summarisation</i>                                   |
| <hr/>   |   |
| Session 3: Translation and Low Resource Languages (Room P521) |   |
| 15:45–16:05   | Presentation: Inigo Jauregi Unanue, Lierni Garmendia Arratibel, Ehsan Zare Borzeshi and Massimo Piccardi<br><i>English-Basque Statistical and Neural Machine Translation</i>    |
| 16:05–16:25   | Presentation: Euna Kim<br><i>Study on the Role of Machine Translation in Social Network Services in terms of User Centered Orientation: A Case Study of Instagram</i>           |
| 16:25–16:45   | Presentation: Yunsil Jo<br><i>Study on Documentary Translation for Dubbing</i>  |
| 16:45–17:05   | Paper: Oliver Adams, Trevor Cohn, Graham Neubig and Alexis Michaud<br><i>Phonemic Transcription of Low-Resource Tonal Languages</i>   |
| 17:05–17:25   | Presentation: Hanieh Poostchi, Ehsan Zare Borzeshi and Massimo Piccardi<br><i>BiLSTM-CRF for Persian Named-Entity Recognition</i>   |
| <hr/>   |   |
| 17:25   | End of Day 1  |
| <hr/>   |   |
| 19:00   | Dinner  |
| <hr/>   |   |

## Friday, 8 December 2017

---

|   |  |
|---|--|
| 9:15-10:15  | Keynote 3 (fromADCS): Victor Kovalev, Redbubble (Room P421)<br><i>Solving hard problems at massive scale – applied data science research approach at Redbubble</i>               |
| Session 4: Computational Linguistics and Information Extraction (Room P521) |  |
| 10:45–10:55   | Paper: Dat Quoc Nguyen, Thanh Vu, Dai Quoc Nguyen, Mark Dras and Mark Johnson<br><i>From Word Segmentation to POS Tagging for Vietnamese</i>                                     |
| 10:55–11:15   | Paper: Shunichi Ishihara<br><i>A Comparative Study of Two Statistical Modelling Approaches for Estimating Multivariate Likelihood Ratios in Forensic Voice Comparison</i>        |
| 11:15–11:35   | Paper: Katharine Cheng, Timothy Baldwin and Karin Verspoor<br><i>Automatic Negation and Speculation Detection in Veterinary Clinical Text</i>                                    |
| 11:35–11:55   | Paper: Xiang Dai, Sarvnaz Karimi and Cecile Paris<br><i>Medication and Adverse Event Extraction from Noisy Text</i>  |
| 11:55–12:15   | Paper: Maria Myunghee Kim<br><i>Incremental Knowledge Acquisition Approach for Information Extraction on both Semi-structured and Unstructured Text from the Open Domain Web</i> |
| 12:15   | Lunch  |
| 13:15–14:15   | Keynote 4: Robert Dale, Language Technology Group Pty Ltd (Room P421)<br><i>Commercialised NLP: The state of the art</i>   |
| 14:15–15:00   | Poster Session (ALTA & ADCS)   |
| 15:00–15:30   | Afternoon Tea  |
| Shared Task Session (Room P521)   |  |
| 15:30–15:40   | Diego Molla-Aliod and Steve Cassidy<br><i>Overview of the 2017 ALTA Shared Task: Correcting OCR Errors</i>   |
| 15:40-15:50   | Gitansh Khirbat<br><i>OCR Post-Processing Text Correction using Simulated Annealing (OPTeCA)</i>   |
| 15:50–16:00   | Yufei Wang<br><i>SuperOCR for ALTA 2017 Shared Task</i>  |
| Final Session (Room P521)   |  |
| 16:00–16:15   | Best Paper and Poster Presentation Awards  |
| 16:15–16:45   | Business Meeting   |
| 16:45-17:00   | Closing  |
| 17:00   | End of Day 2   |

---

## Contents

|  |           |
|--|-----------|
| <b>Invited talks</b>   | <b>1</b>  |
| <b>Tutorials</b>   | <b>3</b>  |
| <b>Long papers</b>   | <b>5</b>  |
| <i>Stock Market Prediction with Deep Learning: A Character-based Neural Language Model for Event-based Trading</i><br><b>Leonardo Dos Santos Pinheiro and Mark Dras</b>          | <b>6</b>  |
| <i>Improving End-to-End Memory Networks with Unified Weight Tying</i><br><b>Fei Liu, Trevor Cohn and Timothy Baldwin</b>   | <b>16</b> |
| <i>Joint Sentence-Document Model for Manifesto Text Analysis</i><br><b>Shivashankar Subramanian, Trevor Cohn, Timothy Baldwin and Julian Brooke</b>                              | <b>25</b> |
| <i>Leveraging linguistic resources for improving neural text classification</i><br><b>Ming Liu, Gholamreza Haffari, Wray Buntine and Michelle Ananda-Rajah</b>                   | <b>34</b> |
| <i>A Hybrid Model for Quality Assessment of Wikipedia Articles</i><br><b>Aili Shen, Jianzhong Qi and Timothy Baldwin</b>   | <b>43</b> |
| <i>Phonemic Transcription of Low-Resource Tonal Languages</i><br><b>Oliver Adams, Trevor Cohn, Graham Neubig and Alexis Michaud</b>  | <b>53</b> |
| <i>A Comparative Study of Two Statistical Modelling Approaches for Estimating Multivariate Likelihood Ratios in Forensic Voice Comparison</i><br><b>Shunichi Ishihara</b>        | <b>61</b> |
| <i>Automatic Negation and Speculation Detection in Veterinary Clinical Text</i><br><b>Katharine Cheng, Timothy Baldwin and Karin Verspoor</b>                                    | <b>70</b> |
| <i>Medication and Adverse Event Extraction from Noisy Text</i><br><b>Xiang Dai, Sarvnaz Karimi and Cecile Paris</b>  | <b>79</b> |
| <i>Incremental Knowledge Acquisition Approach for Information Extraction on both Semi-structured and Unstructured Text from the Open Domain Web</i><br><b>Maria Myunghee Kim</b> | <b>88</b> |
| <b>Short papers</b>  | <b>97</b> |
| <i>On Extending Neural Networks with Loss Ensembles for Text Classification</i><br><b>Hamideh Hajiabadi, Diego Mollá-Alliod and Reza Monsefi</b>                                 | <b>98</b> |

|   |   |            |
|---|---|------------|
| <i>Towards the Use of Deep Reinforcement Learning with Global Policy For Query-based Extractive Summarisation</i> | <b>Diego Mollá-Alliod</b>   | <b>103</b> |
| <i>From Word Segmentation to POS Tagging for Vietnamese</i>   | <b>Dat Quoc Nguyen, Thanh Vu, Dai Quoc Nguyen, Mark Dras and Mark Johnson</b> | <b>108</b> |
| <b>ALTA Shared Task papers</b>  |   | <b>114</b> |
| <i>Overview of the 2017 ALTA Shared Task: Correcting OCR Errors</i>   | <b>Diego Mollá-Alliod and Steve Cassidy</b>                                   | <b>115</b> |
| <i>OCR Post-Processing Text Correction using Simulated Annealing (OPTeCA)</i>                                     | <b>Gitansh Khirbat</b>  | <b>119</b> |
| <i>SuperOCR for ALTA 2017 Shared Task</i>   | <b>Yufei Wang</b>   | <b>124</b> |

## **Invited talks**

# Characterising Information and Happiness in Online Social Activity

*Lewis Mitchell (Lecturer in Applied Mathematics, University of Adelaide)*

**Abstract.** Understanding the nature of influence and information propagation in social networks is of clear societal importance, as they form the basis for phenomena like "echo chambers" and "emotional contagion". However, these concepts remain surprisingly ill-defined. In studies of large online social networks, proxies for influence and information are routinely employed, leading to confusion as to whether the phenomena they underlie actually exist. In this talk I will demonstrate how online social media streams can be used as proxies for population-level health characteristics such as obesity and happiness, and introduce information-theoretic tools for constructing social networks from underlying information flows between individuals. I will present results relating individual predictability to popularity and contact volume, and introduce a paradigmatic mathematical model of information flow over social networks.

**Bio.** Lewis's research focusses on large-scale methods for extracting useful information from online social networks, and on mathematical techniques for inference and prediction using these data. He works on building tools for real-time estimation of social phenomena such as happiness from written text, and prediction of population-level events like disease outbreaks, elections, and civil unrest.

## Commercialised NLP: The State of the Art

*Robert Dale (Principal Consultant, Language Technology Group Pty Ltd)*

**Abstract.** The last few years have seen a tremendous surge in commercial interest in Artificial Intelligence, and with it, a widespread recognition that technologies based on Natural Language Processing can support valuable commercial applications. In this talk, I'll aim to give a comprehensive picture of the commercial NLP landscape, focussing on what I see as the key categories of activity: [1] virtual assistants, including chatbots; [2] text analytics and text mining technologies; [3] machine translation; [4] natural language generation; and [5] text correction technologies. In each case my goal is to sketch the history of work in the area, to identify the major players, and to give a realistic appraisal of the state of the art.

**Bio.** Robert Dale runs the Language Technology Group, an independent consultancy providing unbiased advice to corporations and businesses on the selection and deployment of NLP technologies. Until recently, he was Chief Technology Officer of Arria NLG, where he led the development of a cloud-based natural language generation tool; prior to joining Arria in 2012, he held a chair in the Department of Computing at Macquarie University in Sydney, where he was Director of that university's Centre for Language Technology. After receiving his PhD from the University of Edinburgh in 1989, he taught there for several years before moving to Sydney in 1994. He played a foundational role in building up the NLP community in Australia, and was editor in chief of the Computational Linguistics journal from 2003 to 2012. He writes a semi-regular column titled 'Industry Watch' for the Journal of Natural Language Engineering.

## **Tutorials**

## **Active Learning ... and Beyond!**

*Ben Hachey (The University of Sydney)*

This half-day session will take participants through situations they might face applying Natural Language Processing to real-world problems. We'll choose a canonical task (text classification) and focus on the main issue that faces practitioners in green fields projects — where does the data come from? Our aim is to equip participants with the theoretical background and practical skills to quickly build high-quality text classification models.