Australasian Language Technology Association Workshop 2016

Proceedings of the Workshop

Editor:
Trevor Cohn

5–7 December 2016
Monash University
Caulfield, Australia

Australasian Language Technology Association Workshop 2016

(ALTA 2016)

http://alta2016.alta.asn.au

Online Proceedings:

http://alta2016.alta.asn.au/proceedings

Gold Sponsor:

Silver Sponsors:

Bronze Sponsors:

Volume 14, 2016 ISSN: 1834-7037

ALTA 2016 Workshop Committees

Workshop Co-Chairs

- Gholamreza Haffari (Monash University)
- Andrew Mackinlay (IBM Research)

Workshop Programme Chair

- Trevor Cohn (University of Melbourne)

Programme Committee

- Oliver Adams, University of Melbourne
- Timothy Baldwin, University of Melbourne
- Julian Brooke, University of Melbourne
- Alicia Burga, Universitat Pompeu Fabra - UPF
- Mark Dras, Macquarie University
- Long Duong, University of Melbourne
- Dominique Estival, Western Sydney University
- Ben Hachey, University of Sydney / Hugo.ai
- Graeme Hirst, University of Toronto
- Vu Hoang, University of Melbourne
- Nitin Indurkhya, University of New South Wales
- Sarvnaz Karimi, CSIRO
- Alistair Knott, University of Otago
- François Lareau, Université de Montréal
- Shervin Malmasi, Macquarie University
- Nitika Mathur, University of Melbourne
- Meladel Mistica, Intel
- Diego Mollá, Macquarie University
- Anthony Nguyen, CSIRO
- Joel Nothman, University of Sydney
- Scott Nowson, Accenture
- Bahadorreza Ofoghi, University of Melbourne
- Nagesh Panyam-Chandrasekarasastry, University of Melbourne
- Cécile Paris, CSIRO
- Lizhen Qu, Data61
- Will Radford, Hugo.ai
- Andrea Schalley, Griffith University
- Rolf Schwitter, Macquarie University
- Ehsan Shareghi, Monash University
- Hanna Suominen, Australian National University
- Karin Verspoor, University of Melbourne
- Ming Zhou, Microsoft Research Asia
- Ingrid Zukerman, Monash University

Preface

This volume contains the papers accepted for presentation at the Australasian Language Technology Association Workshop (ALTA) 2016, held at Monash University in Caulfield, Australia on 5–6 December 2016.

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 20 peer-reviewed papers, including 13 long papers and 7 short papers. We received a total of 28 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 2016 includes a presentations track, following on from 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. In total 11 of 12 submissions where selected for presentation. 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 2016 continues the tradition of including a shared task, this year addressing cross-KB coreference. Participation is summarised in an overview paper by organisers Andrew Chisholm, Ben Hachey, and Diego Mollá. 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 the put into maintaining the high standards of our reviewing process; the chairs Gholamreza Haffari and Andrew Mackinlay for coordinating all the logistics that go into running the workshop, from arranging the space, catering, budgets, sponsorship and more; the shared task organisers Andrew Chisholm, Ben Hachey, and Diego Mollá; our keynote speakers Mark Steedman, Hercules Dalianis and Steven Bird for agreeing to share their perspectives on the state of the field; and the tutorial presenters Wray Buntine, Simon Gog and Matthias Petri for their efforts towards the two tutorial sessions. We would like to acknowledge the constant support and advice of the ALTA Executive Committee.

Finally, we gratefully recognise our sponsors: Capital Markets CRC, Google, CSIRO/Data61, Voicebox and Monash University. 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.

Trevor Cohn
ALTA Programme Chair

ALTA 2016 Programme

* Denotes shared session with ADCS.

Monday, 5 December 2016

*Tutorial Session 1 (Monash Caulfield, B214)	
09:00–10:15	Tutorial 1: Wray Buntine *Simpler Non-parametric Bayesian Models*
10:15–10:45	Morning tea
10:45–12:15	Tutorial 1 (continued)
12:15–1:30	Lunch
*Tutorial Session 2 (Monash Caulfield, B214)	
14:00–15:15	Tutorial 2: Simon Gog and Matthias Petri *Succinct Data Structures for Text and Information Retrieval*
15:15–15:45	Afternoon tea
15:45–16:45	Tutorial 2 (continued)

Tuesday, 6 December 2016

Session 1: Opening & Invited talk (Monash Caulfield Campus, B214)	
9:00–9:15	Opening
9:15–10:15	Invited talk: Mark Steedman *On Distributional Semantics*
10:15–10:45	Morning tea
Session 2: Translation (Monash Caulfield Campus, B214)	
10:45–11:10	Presentation: Kyo Kageura, Martin Thomas, Anthony Hartley, Masao Utiyama, Atsushi Fujita, Kikuko Tanabe and Chiho Toyoshima *Supporting Collaborative Translator Training: Online Platform, Scaffolding and NLP*
11:10–11:25	Presentation: Nitika Mathur, Trevor Cohn and Timothy Baldwin *Improving Human Evaluation of Machine Translation*
11:25–11:40	Paper: Cong Duy Vu Hoang, Reza Haffari and Trevor Cohn *Improving Neural Translation Models with Linguistic Factors*
11:40–11:55	Presentation: Daniel Beck, Lucia Specia and Trevor Cohn *Exploring Prediction Uncertainty in Machine Translation Quality Estimation*
11:55–12:00	CLEF eHealth 2017 Shared tasks
12:00–13:15	Lunch
Session 3a: Invited talk (Monash Caulfield Campus, B214)	
13:15–13:55	*Invited talk: Hercules Dalianis
13:55–14:00	Break
Session 3b: Health (Monash Caulfield Campus, B214)	
14:00–14:15	Presentation: Raghavendra Chalapathy, Ehsan Zare Borzeshi and Massimo Piccardi *An Investigation of Recurrent Neural Architectures for Drug Name Recognition*
14:15–14:30	Paper: Hamed Hassanzadeh, Anthony Nguyen and Bevan Koopman *Evaluation of Medical Concept Annotation Systems on Clinical Records*
14:30–14:45	Paper: Mahnoosh Kholghi, Lance De Vine, Laurianne Sitbon, Guido Zuccon and Anthony Nguyen *The Benefits of Word Embeddings Features for Active Learning in Clinical Information Extraction*
14:45–15:00	Presentation: Rebecka Weegar and Hercules Dalianis *Mining Norwegian pathology reports: A research proposal*
15:00–15:15	Paper: Pin Huang, Andrew MacKinlay and Antonio Jimeno *Syndromic Surveillance using Generic Medical Entities on Twitter*
15:15–15:30	Paper: Yufei Wang, Stephen Wan and Cecile Paris *The Role of Features and Context on Suicide Ideation Detection*
15:30–16:00	Afternoon tea
Session 4: Relation & Information extraction (Monash Caulfield Campus, B214)	
16:00–16:15	Presentation: Dat Quoc Nguyen and Mark Johnson *Modeling topics and knowledge bases with embeddings*
16:15–16:30	Paper: Zhuang Li, Lizhen Qu, Qiongkai Xu and Mark Johnson *Unsupervised Pre-training With Seq2Seq Reconstruction Loss for Deep Relation Extraction Models*
16:30–16:45	Presentation: Hanieh Poostchi, Ehsan Zare Borzeshi and Massimo Piccardi *PersoNER: Persian Named-Entity Recognition*
16:45–17:00	Paper: Nagesh C. Panyam, Karin Verspoor, Trevor Cohn and Rao Kotagiri *ASM Kernel: Graph Kernel using Approximate Subgraph Matching for Relation Extraction*
17:00–17:15	Paper: Gitansh Khirbat, Jianzhong Qi and Rui Zhang *N-ary Biographical Relation Extraction using Shortest Path Dependencies*

Wednesday, 7 December 2016

Session 5: Invited talk & Shared task (Monash Caulfield Campus, B214)

9:00–9:45	Invited talk: Steven Bird *Getting started with an Australian language*
9:45–10:15	Shared Task
10:15–10:45	Morning tea

Session 6: Short-papers & posters (Monash Caulfield Campus, B214)

10:45–11:10	Short-paper lightning talks
	Aditya Joshi, Vaibhav Tripathi, Pushpak Bhattacharyya, Mark Carman, Meghna Singh, Jaya Saraswati and Rajita Shukla *How Challenging is Sarcasm versus Irony Classification?: A Study With a Dataset from English Literature* Ming Liu, Gholamreza Haffari and Wray Buntine *Learning cascaded latent variable models for biomedical text classification* Bo Han, Antonio Jimeno Yepes, Andrew MacKinlay and Lianhua Chi *Temporal Modelling of Geospatial Words in Twitter* Antonio Jimeno Yepes and Andrew MacKinlay *NER for Medical Entities in Twitter using Sequence to Sequence Neural Networks* Dat Quoc Nguyen, Mark Dras and Mark Johnson *An empirical study for Vietnamese dependency parsing* Will Radford, Ben Hachey, Bo Han and Andy Chisholm *:telephone::person::sailboat::whale::okhand: ; or âĂIJCall me IshmaelâĂİ âĂŞ How do you translate emoji?* Xavier Holt, Will Radford and Ben Hachey *Presenting a New Dataset for the Timeline Generation Problem*
11:10–12:00	Poster Session
12:00–13:15	Lunch
13:15–13:35	Business Meeting

Session 7: Applications (Monash Caulfield Campus, B214)

13:35–13:50	Paper: Hafsah Aamer, Bahadorreza Ofoghi and Karin Verspoor *Syndromic Surveillance through Measuring Lexical Shift in Emergency Department Chief Complaint Texts*
13:50–14:05	Paper: Rui Wang, Wei Liu and Chris McDonald *Featureless Domain-Specific Term Extraction with Minimal Labelled Data*
14:05–14:30	Presentation: Ehsan Shareghi *Unbounded and Scalable Smoothing for Language Modeling*
14:30–14:45	Paper: Shunichi Ishihara *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*
14:45–15:00	Presentation: Oliver Adams, Shourya Roy and Raghu Krishnapuram *Distributed Vector Representations for Unsupervised Automatic Short Answer Grading*
15:00–15:15	Paper: Andrei Shcherbakov, Ekaterina Vylomova and Nick Thieberger *Phonotactic Modeling of Extremely Low Resource Languages*
15:15–15:30	Paper: Oliver Adams, Adam Makarucha, Graham Neubig, Steven Bird and Trevor Cohn *Cross-Lingual Word Embeddings for Low-Resource Language Modeling*
15:30–16:20	Afternoon tea

Session 8: Closing (Monash Caulfield Campus, B214)

16:20–16:30	Awards for best paper and best presentation
16:30–16:45	ALTA Closing

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**

Cross-Lingual Word Embeddings for Low-Resource Language Modeling
Oliver Adams, Adam Makarucha, Graham Neubig, Steven Bird and Trevor Cohn **94**

The Role of Features and Context on Suicide Ideation Detection
Yufei Wang, Stephen Wan and Cecile Paris **105**

Featureless Domain-Specific Term Extraction with Minimal Labelled Data
Rui Wang, Wei Liu and Chris McDonald **114**

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 **124**

Short papers **133**

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 **134**

Learning cascaded latent variable models for biomedical text classification
Ming Liu, Gholamreza Haffari and Wray Buntine **139**

Temporal Modelling of Geospatial Words in Twitter
Bo Han, Antonio Jimeno Yepes, Andrew MacKinlay and Lianhua Chi **144**

NER for Medical Entities in Twitter using Sequence to Sequence Neural Networks
Antonio Jimeno Yepes and Andrew MacKinlay **149**

An empirical study for Vietnamese dependency parsing
Dat Quoc Nguyen, Mark Dras and Mark Johnson **154**

:telephone::person::sailboat::whale::okhand: ; or âĂIJCall me IshmaelâĂİ âĂŞ How do you translate emoji?
Will Radford, Ben Hachey, Bo Han and Andy Chisholm **161**

Presenting a New Dataset for the Timeline Generation Problem
Xavier Holt, Will Radford and Ben Hachey **166**

ALTA Shared Task papers **171**

Overview of the 2016 ALTA Shared Task: Cross-KB Coreference
Andrew Chisholm, Ben Hachey and Diego Mollá **172**

Disambiguating Entities Referred by Web Endpoints using Tree Ensembles
Gitansh Khirbat, Jianzhong Qi and Rui Zhang **176**

Filter and Match Approach to Pair-wise Web URI Linking
S. Shivashankar, Yitong Li and Afshin Rahimi **181**

Pairwise FastText Classifier for Entity Disambiguation
Cheng Yu, Bing Chu, Rohit Ram, James Aichinger, Lizhen Qu and Hanna Suominen **186**