Australasian Language Technology Association Workshop 2015

Proceedings of the Workshop

Editors:
Ben Hachey
Kellie Webster

8–9 December 2015
Western Sydney University
Parramatta, Australia

Australasian Language Technology Association Workshop 2015
(ALTA 2015)

http://www.alta.asn.au/events/alta2015

Online Proceedings:
http://www.alta.asn.au/events/alta2015/proceedings/

Silver Sponsors:

Bronze Sponsors:

Volume 13, 2015 ISSN: 1834-7037

ALTA 2015 Workshop Committees

Workshop Co-Chairs

- Ben Hachey (The University of Sydney — Abbrevi8 Pty Ltd)
- Kellie Webster (The University of Sydney)

Workshop Local Organiser

- Dominique Estival (Western Sydney University)
- Caroline Jones (Western Sydney University)

Programme Committee

- Timothy Baldwin (University of Melbourne)
- Julian Brook (University of Toronto)
- Trevor Cohn (University of Melbourne)
- Dominique Estival (University of Western Sydney)
- Gholamreza Haffari (Monash University)
- Nitin Indurkhya (University of New South Wales)
- Sarvnaz Karimi (CSIRO)
- Shervin Malmasi (Macquarie University)
- Meladel Mistica (Intel)
- Diego Mollá (Macquarie University)
- Anthony Nguyen (Australian e-Health Research Centre)
- Joel Nothman (University of Melbourne)
- Cécile Paris (CSIRO – ICT Centre)
- Glen Pink (The University of Sydney)
- Will Radford (Xerox Research Centre Europe)
- Rolf Schwitter (Macquarie University)
- Karin Verspoor (University of Melbourne)
- Ingrid Zukerman (Monash University)

Preface

This volume contains the papers accepted for presentation at the Australasian Language Technology Association Workshop (ALTA) 2015, held at Western Sydney University in Parramatta, Australia on 8–9 December 2015.

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 16 peer-reviewed papers, including 12 long papers and 4 short papers. We received a total of 20 submissions for long and short papers. Each paper was reviewed by three members of the program committee. Great care was taken to avoid all conflicts of interest.

ALTA 2015 introduces an experimental presentations track. 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 executive committee to ensure relevance, with 4 of 5 submissions included in the programme.

ALTA 2015 continues the tradition of including a shared task, this year addressing the identification of French cognates in English text. Participation is summarised in an overview paper by organisers Laurianne Sitbon, Diego Mollá and Haoxing Wang. 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 their valuable time and effort; the local organisers Dominique Estival and Caroline Jones for handling physical logistics and coordination with the Confluence 2015 programme; our keynote speaker Mark Johnson for agreeing to share his perspective on the state of the field; and the panelists Tim Baldwin [Moderator], Grace Chung, David Hawking, Maria Milosavljevic and Doug Oard for agreeing to share their experience and insights. We would like to acknowledge the constant support and advice of the ALTA Executive Committee, and the valuable guidance of previous co-chairs.

Finally, we gratefully recognise our sponsors: Data61/CSIRO, Capital Markets CRC, Google, Hugo/Abbrevi8 and The University of Sydney. Most importantly, their generous support enabled us to offer travel subsidies to all students presenting at ALTA. Their support also funded the conference dinner and student paper awards.

Ben Hachey and Kellie Webster
ALTA Workshop Co-Chairs

ALTA 2015 Programme

* denotes sessions shared with ADCS.

Tuesday, 8 December 2015

Session 1 (Parramatta City Campus, Level 6, Room 9-10)*	
09:15–09:30	Opening remarks
09:30–10:30	Keynote: Doug Oard *Information Abolition*
10:30–11:00	Morning tea
Session 2 (Parramatta City Campus, Level 6, Room 9-10)	
11:00–11:20	Presentation: Trevor Cohn *Unlimited order Language Modeling with Compressed Suffix Trees*
11:20–11:40	Long paper: Caroline Mckinnon, Ibtehal Baazeem and Daniel Angus *How few is too few? Determining the minimum acceptable number of LSA dimensions to visualise text cohesion with Lex*
11:40–12:00	Long paper: Ping Tan, Karin Verspoor and Tim Miller *Structural Alignment as the Basis to Improve Significant Change Detection in Versioned Sentences*
12:00–12:20	Presentation: Kellie Webster and James Curran *Challenges in Resolving Nominal Reference*
12:20–1:20	Lunch
Session 3 (Parramatta City Campus, Level 6, Room 9-10)*	
1:30–2:00	ADCS paper: Viet Phung and Lance De Vine *A study on the use of word embeddings and PageRank for Vietnamese text summarization*
2:00–2:20	Long paper: Mahmood Yousefi Azar, Kairit Sirts, Len Hamey and Diego Mollá *Query-Based Single Document Summarization Using an Ensemble Noisy Auto-Encoder*
2:20–2:40	Long paper: Lan Du, Anish Kumar, Massimiliano Ciaramita and Mark Johnson *Using Entity Information from a Knowledge Base to Improve Relation Extraction*
2:40–2:50	Flash presentation: Hanna Suominen *Preview of CLEF eHealth 2016*
2:50–3:00	Short break
3:00–4:00	Panel: Tim Baldwin [Moderator], Grace Chung, David Hawking, Maria Milosavljevic and Doug Oard *NLP & IR in the Wild*
4:00–4:30	Afternoon tea
Session 4 (Parramatta City Campus, Level 6, Room 9-10)	
4:30–4:50	Long paper: Julio Cesar Salinas Alvarado, Karin Verspoor and Timothy Baldwin *Domain Adaption of Named Entity Recognition to Support Credit Risk Assessment*
4:50–5:10	Presentation: Ben Hachey, Anaïs Cadilhac and Andrew Chisholm *Entity Linking and Summarisation in a News-driven Personal Assistant App*
5:10–5:30	Long paper: Shungwan Kim and Steve Cassidy *Finding Names in Trove: Named Entity Recognition for Australian Historical Newspapers*
6:00	Conference dinner @ Collector Hotel

Wednesday, 9 December 2015

Session 5 (Parramatta South Campus, Rydalmere, Building EA, Room 2.14)	
9:30–9:50	Long paper: Jennifer Biggs *Comparison of Visual and Logical Character Segmentation in Tesseract OCR Language Data for Indic Writing Scripts*
9:50–10:10	Long paper: Daniel Frost and Shunichi Ishihara *Likelihood Ratio-based Forensic Voice Comparison on L2 speakers: A Case of Hong Kong native male production of English vowels*
10:10–10:30	Long paper: Kairit Sirts and Mark Johnson *Do POS Tags Help to Learn Better Morphological Segmentations?*
10:30–11:00	Morning tea
Session 6 (Parramatta South Campus, Rydalmere, Building EA, Room 2.14)	
11:00–11:20	Business Meeting
11:20–11:30	Awards
11:30–11:45	Shared task: Laurianne Sitbon, Diego Mollá and Haoxing Wang *Overview of the 2015 ALTA Shared Task: Identifying French Cognates in English Text*
11:45–12:00	Shared task: Qiongkai Xu, Albert Chen and Chang Li *Detecting English-French Cognates Using Orthographic Edit Distance*
12:00–12:10	Short paper: Fiona Martin and Mark Johnson *More Efficient Topic Modelling Through a Noun Only Approach*
12:10–12:20	Short paper: Dat Quoc Nguyen, Kairit Sirts and Mark Johnson *Improving Topic Coherence with Latent Feature Word Representations in MAP Estimation for Topic Modeling*
12:20–12:30	Short paper: Joel Nothman, Atif Ahmad, Christoph Breidbach, David Malet and Tim Baldwin *Understanding engagement with insurgents through retweet rhetoric*
12:30–12:40	Short paper: Sam Shang Chun Wei and Ben Hachey *A comparison and analysis of models for event trigger detection*
Session 7 (Parramatta South Campus, Rydalmere, Building EE, Foyer)	
12:40–1:30	Lunch
1:30–2:30	Poster session
Session 8 (Parramatta South Campus, Rydalmere, Building EA, Room G.18)*	
2:30–3:30	Keynote: Mark Johnson *Computational Linguistics: The previous and the next five decades*
3:30–4:00	Afternoon tea
Session 8 (Parramatta South Campus, Rydalmere, Building EA, Room 2.14)	
4:00–4:20	Long paper: Hamed Hassanzadeh, Diego Mollá, Tudor Groza, Anthony Nguyen and Jane Hunter *Similarity Metrics for Clustering PubMed Abstracts for Evidence Based Medicine*
4:20–4:40	Long paper: Lance De Vine, Mahnoosh Kholghi, Guido Zuccon, Laurianne Sitbon and Anthony Nguyen *Analysis of Word Embeddings and Sequence Features for Clinical Information Extraction*
4:40–5:00	Long paper: Shervin Malmasi and Hamed Hassanzadeh *Clinical Information Extraction Using Word Representations*
5:00–5:20	Presentation: Andrew MacKinlay, Antonio Jimeno Yepes and Bo Han *A System for Public Health Surveillance using Social Media*
5:20–5:30	ALTA closing

Contents