Australasian Language Technology Association Workshop 2018

Proceedings of the Workshop

Editors:
Sunghwan Mac Kim
Xiuzhen (Jenny) Zhang

10–12 December 2018
The University of Otago
Dunedin, New Zealand

Australasian Language Technology Association Workshop 2018

(ALTA 2018)

http://alta2018.alta.asn.au

Online Proceedings:

http://alta2018.alta.asn.au/proceedings

Gold Sponsors:

Silver Sponsor:

Bronze Sponsors:

Volume 16, 2018 ISSN: 1834-7037

ALTA 2018 Workshop Committees

Workshop Co-Chairs

- Xiuzhen (Jenny) Zhang (RMIT University)
- Sunghwan Mac Kim (CSIRO Data61)

Publication Co-chairs

- Sunghwan Mac Kim (CSIRO Data61)
- Xiuzhen (Jenny) Zhang (RMIT University)

Programme Committee

- Abeed Sarker, University of Pennsylvania
- Alistair Knott, University of Otago
- Andrea Schalley, Karlstads Universitet
- Ben Hachey, The University of Sydney and Digital Health CRC
- Benjamin Borschinger, Google
- Bo Han, Accenture
- Daniel Angus, University of Queensland
- Diego Mollá, Macquarie University
- Dominique Estival, Western Sydney University
- Gabriela Ferraro, CSIRO Data61
- Gholamreza Haffari, Monash University
- Hamed Hassanzadeh, Australian e-Health Research Centre CSIRO
- Hanna Suominen, Australian National University and Data61/CSIRO
- Jennifer Biggs, Defence Science Technology
- Jey-Han Lau, IBM Research
- Jojo Wong, Monash University
- Karin Verspoor, The University of Melbourne
- Kristin Stock, Massey University
- Laurianne Sitbon, Queensland University of Technology
- Lawrence Cavedon, RMIT University
- Lizhen Qu, CSIRO Data61
- Long Duong, Voicebox Technology Australia
- Maria Kim, Defence Science Technology
- Massimo Piccardi, University of Technology Sydney
- Ming Zhou, Microsoft Research Asia
- Nitin Indurkhya, University of New South Wales
- Rolf Schwitter, Macquarie University
- Sarvnaz Karimi, CSIRO Data61
- Scott Nowson, Accenture
- Shervin Malmasi, Macquarie University and Harvard Medical School
- Shunichi Ishihara, Australian National University

- Stephen Wan, CSIRO Data61
- Teresa Lynn, Dublin City University
- Timothy Baldwin, The University of Melbourne
- Wei Gao, Victoria University of Wellington
- Wei Liu, University of Western Australia
- Will Radford, Canva
- Wray Buntine, Monash University

Preface

This volume contains the papers accepted for presentation at the Australasian Language Technology Association Workshop (ALTA) 2018, held at The University of Otago in Dunedin, New Zealand on 10-12 December 2018.

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 10 peer-reviewed papers, including 6 long papers and 4 short papers. We received a total of 17 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 2018 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 2018 continues the tradition of including a shared task, this year on classifying patent applications. Participation is summarised in an overview paper by organisers Diego Mollá and Dilesha Seneviratne. 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 shared task organisers Diego Mollá and Dilesha Seneviratne; our keynote speakers Alistair Knott and Kristin Stock for agreeing to share their perspectives on the state of the field; and the tutorial presenter Phil Cohen for his efforts towards the tutorial of collaborative dialogue. We would like to acknowledge the constant support and advice of the ALTA Executive Committee such as budgets, sponsorship and more.

Finally, we gratefully recognise our sponsors: CSIRO/Data61, Soul Machines, Google, IBM, Seek and ARC Centre of Excellence for the Dynamics of Language. 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.

Xiuzhen (Jenny) Zhang
Sunghwan Mac Kim
ALTA Programme Chairs

ALTA 2018 Programme

Monday, 10 December 2018

Tutorial Session (Room 1.19)	
14:00–17:00	Tutorial: Phil Cohen *Towards Collaborative Dialogue*
17:00	End of Tutorial

Tuesday, 11 December 2018

Opening & Keynote (Room 1.17)	
9:00–9:15	Opening
9:15–10:15	Keynote 1 (from ADCS): Jon Degenhardt (Room 1.17) *An Industry Perspective on Search and Search Applications*
10:15–10:45	Morning tea
Session A: Text Mining & Applications (Room 1.19)	
10:45–11:05	Paper: Rolando Coto Solano, Sally Akevai Nicholas and Samantha Wray *Development of Natural Language Processing Tools for Cook Islands Māori*
11:05–11:25	Paper: Bayzid Ashik Hossain and Rolf Schwitter *Specifying Conceptual Models Using Restricted Natural Language*
11:25–11:45	Presentation: Jenny McDonald and Adon Moskal *Quantext: a text analysis tool for teachers*
11:45–11:55	Paper: Xuanli He, Quan Tran, William Havard, Laurent Besacier, Ingrid Zukerman and Gholamreza Haffari *Exploring Textual and Speech information in Dialogue Act Classification with Speaker Domain Adaptation*
11:55–13:00	Lunch
13:00–14:00	Keynote 2: Alistair Knott (Room 1.17) *Learning to talk like a baby*
14:00–14:15	Break
14:15–15:15	Poster Session 1 (ALTA & ADCS)
15:15–15:45	Afternoon tea
15:45–16:50	Poster Session 2 (ALTA & ADCS)
16:50	End of Day 1

Wednesday, 12 December 2018

9:00–10:15	Keynote 3 (from ADCS): David Bainbridge (Room 1.17) *Can You Really Do That? Exploring new ways to interact with Web content and the desktop*
10:15–10:45	Morning tea
Session B: Machine Translation & Speech (Room 1.19)	
10:45–11:05	Paper: Cong Duy Vu Hoang, Gholamreza Haffari and Trevor Cohn *Improved Neural Machine Translation using Side Information*
11:05–11:25	Presentation: Qiongkai Xu, Lizhen Qu and Jiawei Wang *Decoupling Stylistic Language Generation*
11:25–11:45	Paper: Satoru Tsuge and Shunichi Ishihara *Text-dependent Forensic Voice Comparison: Likelihood Ratio Estimation with the Hidden Markov Model (HMM) and Gaussian Mixture Model – Universal Background Model (GMM-UBM) Approaches*
11:45–11:55	Paper: Nitika Mathur, Timothy Baldwin and Trevor Cohn *Towards Efficient Machine Translation Evaluation by Modelling Annotators*
11:55–12:55	Lunch
12:55–13:55	Keynote 4: Kristin Stock (Room 1.17) *"Where am I, and what am I doing here?" Extracting geographic information from natural language text*
13:55–14:05	Break
Session C: Shared session with ADCS (Room 1.17)	
14:05–14:30	Paper: Alfan Farizki Wicaksono and Alistair Moffat (ADCS long paper) *Exploring Interaction Patterns in Job Search*
14:30–14:50	Paper: Xavier Holt and Andrew Chisholm *Extracting structured data from invoices*
14:50–15:05	Paper: Bevan Koopman, Anthony Nguyen, Danica Cossio, Mary-Jane Courage and Gary Francois (ADCS short paper) *Extracting Cancer Mortality Statistics from Free-text Death Certificates: A View from the Trenches*
15:05–15:15	Paper: Hanieh Poostchi and Massimo Piccardi *Cluster Labeling by Word Embeddings and WordNet's Hypernymy*
15:15–15:35	Afternoon tea
Session D: Word Semantics (Room 1.19)	
15:35–15:55	Paper: Lance De Vine, Shlomo Geva and Peter Bruza *Unsupervised Mining of Analogical Frames by Constraint Satisfaction*
15:55–16:05	Paper: Navnita Nandakumar, Bahar Salehi and Timothy Baldwin *A Comparative Study of Embedding Models in Predicting the Compositionality of Multiword Expressions*
Shared Task Session (Room 1.19)	
16:05–16:15	Paper: Diego Mollá-Aliod and Dilesha Seneviratne *Overview of the 2018 ALTA Shared Task: Classifying Patent Applications*
16:15–16:25	Paper: Fernando Benites, Shervin Malmasi, Marcos Zampieri *Classifying Patent Applications with Ensemble Methods*
16:25–16:35	Paper: Jason Hepburn *Universal Language Model Fine-tuning for Patent Classification*
16:35–16:45	Break
16:45–16:55	Best Paper Awards
16:55–17:20	Business Meeting & ALTA Closing
17:20	End of Day 2

Contents

Invited keynotes

Alistair Knott (University of Otago & Soul Machines)

Learning to talk like a baby

In recent years, computational linguists have embraced neural network models, and the vector-based representations of words and meanings they use. But while computational linguists have readily adopted the machinery of neural network models, they have been slower to embrace the original aim of neural network research, which was to understand how brains work. A large community of neural network researchers continues to pursue this "cognitive modelling" aim, with very interesting results. But the work of these more cognitively minded modellers has not yet percolated deeply into computational linguistics. In my talk, I will argue the cognitive modelling tradition of neural networks has much to offer computational linguistics. I will outline a research programme that situates language modelling in a broader cognitive context. The programme is distinctive in two ways. Firstly, the initial object of study is a baby, rather than an adult. Computational linguistics models typically aim to reproduce adult linguistic competence in a single training process, that presents an "empty" network with a corpus of mature language. I will argue that this training process doesn't correspond to anything in human experience, and that we should instead aim to model a more gradual developmental process, that first achieves babylike language, then childlike language, and so on. Secondly, the new programme studies the baby's language system as it interfaces with her other cognitive systems, rather than by itself. It pays particular attention to the sensory and motor systems through which a baby engages with the physical world, which are the primary means by which it activates semantic representations. I will argue that the structure of these sensorimotor systems, as expressed in neural network models, offer interesting insights about certain aspects of linguistic structure. I will conclude by demoing a model of the interface between language and the sensorimotor system, as it operates in a baby at an early stage of language learning.

Kristin Stock (Massey University)

"Where am I, and what am I doing here?" Extracting geographic information from natural language text

The extraction of place names (toponyms) from natural language text has received a lot of attention in recent years, but location is frequently described in more complex ways, often using other objects as reference points. Examples include: 'The accident occurred opposite the Orewa Post Office, near the pedestrian crossing' or 'the sample was collected on the west bank of the Waikato River, about 3km upstream from Huntly'. These expressions can be vague, imprecise, underspecified, rely on access to information about other objects in the environment, and the semantics of spatial relations like 'opposite' and 'on' are still far from clear. Furthermore, many of these kinds of expressions are context sensitive, and aspects such as scale, geometry and type of geographic feature may influence the way the expression is understood. Both machine learning and rule-based approaches have been developed to try to firstly parse expressions of this kind, and secondly to determine the geographic location that the expression refers to. Several relevant projects will be discussed, including the development of a semantic rather than syntactic approach to parsing geographic location descriptions; the creation of a manually annotated training set of geographic language; the challenges highlighted from human descriptions of location in the emergency services context; the interpretation and geocoding of descriptions of flora and fauna specimen collections; the development of models of spatial relations using social media data and the use of instance-based learning to interpret complex location descriptions.

Tutorials

Towards Collaborative Dialogue

Phil Cohen (Monash University)

This tutorial will discuss a program of research for building collaborative dialogue systems, which are a core part of virtual assistants. I will briefly discuss the strengths and limitations of current approaches to dialogue, including neural network-based and slot-filling approaches, but then concentrate on approaches that treat conversation as planned collaborative behaviour. Collaborative interaction involves recognizing someone's goals, intentions, and plans, and then performing actions to facilitate them. People have learned this basic capability at a very young age and are expected to be helpful as part of ordinary social interaction. In general, people's plans involve both speech acts (such as requests, questions, confirmations, etc.) and physical acts. When collaborative behavior is applied to speech acts, people infer the reasons behind their interlocutor's utterances and attempt to ensure their success. Such reasoning is apparent when an information agent answers the question "Do you know where the Sydney flight leaves?" with "Yes, Gate 8, and it's running 20 minutes late." It is also apparent when one asks "where is the nearest petrol station?" and the interlocutor answers "2 kilometers to your right" even though it is not the closest, but rather the closest one that is open. In this latter case, the respondent has inferred that you want to buy petrol, not just to know the location of the station. In both cases, the literal and truthful answer is not cooperative. In order to build systems that collaborate with humans or other artificial agents, a system needs components for planning, plan recognition, and for reasoning about agents' mental states (beliefs, desires, goals, intentions, obligations, etc.).

In this tutorial, I will discuss current theory and practice of such collaborative belief-desire-intention architectures, and demonstrate how they can form the basis for an advanced collaborative dialogue manager. In such an approach, systems reason about what they plan to say, and why the user said what s/he did. Because there is a plan standing behind the system's utterances, it is able to explain its reasoning. Finally, we will discuss potential methods for incorporating such a plan-based approach with machine-learned approaches.