ALTA 2019

Proceedings of the 17th Workshop of the Australasian Language Technology Association

4–6 December, 2019

UTS

Sydney, Australia

Sponsors

ALTA is extremely grateful to the following sponsors who helped make ALTA as accessible to as many NLP researchers as possible

Platinum

SINTELIX

Gold

Australian Government
Department of Defence

Google

UTS

DATA 61

CSIRO

Bronze

IBM **Research** AI

Introduction

Welcome to the 17th edition of the Annual Workshop of the Australasian Language Technology Association (ALTA 2019) in Sydney, Australia. The purpose of ALTA is to promote language technology research and development in Australia and New Zealand. Every year ALTA hosts a workshop which is the key local forum for socialising research results in natural language processing and computational linguistics, with presentations and posters from students, industry, and academic researchers. This year ALTA 2019 is being hosted by the University of Technology Sydney and we acknowledge and pay our respects to the Gadigal people of the Eora Nation, the Boorooberongal people of the Dharug Nation, the Bidiagal people and the Gamaygal people upon whose ancestral lands the university stands.

In total we received 36 paper submissions. We accepted 8 long papers (of 14 submissions), 7 short papers (of 22 submissions) to appear as oral presentations in the programme, giving a total of 15 paper presentations (42% of submissions). Of the 36 submissions 23 were first-authored by students, with submissions from 6 of the 8 states and territories of Australia. We are extremely grateful to the Programme Committee members for their time and their detailed and helpful comments and reviews, both locally and abroad. This year we had committee members from all over the globe including Sweden, Scotland, USA, UAE, and Germany, and 16% of the committee being made up of our near neighbours in New Zealand.

Overall, there are 6 sessions of oral presentations in the programme, two of which are jointly organised with the Australasian Document Computing Symposium (ADCS 2019), starting each day with an ALTA keynote talk. The main workshop follows a tutorial on *NLP for Healthcare in the Absence of a Healthcare Dataset* guided by Sarvnaz Karimi and Aditya Joshi (CSIRO Data61). To encourage a broader participation of the local NLP community we organised a poster session, jointly with ADCS, of which 10 papers were included. These papers have undergone the same double-blind review process as the oral presentations. In addition, this year we also ran a shared task on sarcasm detection organised by Diego Molla-Aliod (University of Macquarie) and Aditya Joshi (CSIRO Data61).

The talks from our keynote speakers reflect the main themes in our workshop as well as the direction in which our field is taking us. Nicholas Evans and Ben Foley present their work on the *New wings for the Library of Babel: The transcription challenge for the world's 7000+ languages*. We face the challenge of creating resources and adapting methodologies for the low-resource domain – which most of the worlds languages fit into – as well as the many tasks that we undertake in the field. Mark Johnson presents research in *Building new kinds of Natural Language Understanding and Conversational AI with Deep Learning*, which reflects the continuing trend towards neural and deep learning methods in natural language processing as well as the need to look beyond the sentence by taking into consideration context from the discourse and document level, to develop more naturalistic language interfaces with AI agents.

ALTA 2019 is very grateful for the financial support generously offered by our sponsors, without which the running of these events to bring together the NLP community of the Australasian region would be a challenge.

We very much hope that you will have an enjoyable and inspiring time at ALTA 2019!

Meladel Mistica, Andrew MacKinlay and Massimo Piccardi

Sydney, December 2019

Organisers:

Local Chair: Massimo Piccardi
Program Chairs: Meladel Mistica
General Chair: Andrew MacKinlay

Program Committee:

Abeed Sarker, Afshin Rahimi, Alistair Knott, Andrea Schalley, Antonio Jimeno, Ben Hachey, Benjamin Boerschinger, Brian Hur, Daniel Beck, David Martinez, Diego Molla, Dominique Estival, Gabriela Ferraro, Gholamreza Haffari, Hamed Hassanzadeh, Hanna Suominen, Hiyori Yoshikawa, Jennifer Biggs, Jeremy Nicholson, Jey-Han Lau, Jojo Wong, Karin Verspoor, Kristin Stock, Laurianne Sitbon, Lawrence Cavedon, Lizhen Qu, Mahsa Mohaghegh, Mariano Phielipp, Mark Dras, Markus Luczak-Roesch, Michael Witbrock, Myunghee Kim, Nitika Mathur, Nitin Indurkhya, Parma Nand, Rolf Schwitter, Sarvnaz Karimi, Scott Nowson, Sharon Gao, Shervin Malmasi, Spandana Gella, Stephen Wan, Sumithra Velupillai, Sunghwan Mac Kim, Timothy Baldwin, Trevor Cohn, Will Radford, Wray Lindsay Buntine, Xiang Dai, Xiuzhen Zhang, Yitong Li

Invited Speakers:

Mark Johnson, Oracle and Macquarie University
Nicholas Evans, ANU; with Ben Foley, University of Queensland

Invited Talks

Mark Johnson: Building new kinds of Natural Language Understanding and Conversational AI with Deep Learning

Deep learning provides new fundamental tools, such as contextualised word embeddings and seq2seq models, that let us build new kinds of Natural Language Understanding apps faster, better and cheaper than ever before. The advanced pattern-matching capabilities of deep learning enable a new approach to app development where the system's behaviour is learnt from training data, dramatically reducing the need for manual scripting. This talk describes how we are using this technology in the Oracle Digital Assistant, focusing especially on Conversational AI. The talk ends with a discussion of how research advances in areas such as explainability, few-shot learning, data augmentation and transfer learning can help this technology achieve its full potential.

Nicholas (Nick) Evans and Ben Foley: New wings for the Library of Babel: The transcription challenge for the world's 7000+ languages

There is increasing awareness that we stand on the brink of massive knowledge loss as perhaps half of the world's languages risk not being learnt by the next generation, and of the attendant urgency of recording them in some form. Yet our conceptions for just how much we should record of each language, if we are to do justice to the intellectual richness of the oral traditions they represent, remain tragically unambitious. How much of the knowledge of English or Chinese-speaking cultures would be captured in ten hours of text, a typical amount to be recorded in a language documentation project? Compare this to the 60 million words or so we have in corpora of Classical Greek or Sanskrit, equivalent to about 6,000 hours of recordings. Is it inconceivable for modern day speech communities, seeking a deep abiding record of their language, to record and transcribe that much data? After all, ten members of a speech community, each recording three hours per day, could gather this much in a year.

The real challenge, as linguists and language community members have come to realise, is the transcription bottleneck, the fact that writing down a transcription of one hour of recording typically takes from 40 to 100 hours (and in the early phases of work almost always at the upper end). The result of this bottleneck is that even if we record something like the above amount, current language documentation methods of a few people working together over three years cannot transcribe more than around 15 hours of primary material. This does not touch the levels needed to give a rich corpus for one language, nor does it reach the one hundred hours normally cited as a necessary minimum for a deep-learning training corpus.

In this talk we describe the TAP initiative – Transcription Acceleration Project – which is a joint enterprise of language documentation fieldworkers, community language users, computational linguists, software engineers and machine learning researchers, supported by the ARC-funded Centre of Excellence for the Dynamics of Language (CoEDL). This project aims to break the impasse posed by the transcription bottleneck while maintaining the language community members' social and cultural roles. TAP's semi-automated speech recognition workflow is designed as a user-in-the-loop architecture which involves critical stakeholders in the process of creating cultural and linguistic artefacts. The tools within TAP aim to improve the transcription experience, and support new ways of working to improve the state of language documentation globally. For Australia and its neighbours, we will be able to secure a much greater proportion of the region's rich but often ignored linguistic cultural heritage – around a quarter of the world's languages – for the generations to come.

PROGRAMME

4th December (Wednesday) Tutorial, Day 1

12:30 - 1:00	Registration
13:00 - 16:30	NLP for Healthcare in the Absence of a Healthcare Dataset Sarvnaz Karimi & Aditya Joshi (CSIRO Data61)

5th December (Thursday) Day 2

8:15 - 9:00	Registration
9:00 - 9:15	Welcome to ALTA 2019
9:15 - 10:15	Keynote: Nicholas Evans (ANU) and Ben Foley (UQ) *New wings for the Library of Babel: The transcription challenge for the world's 7000+ languages* (Session Chair: Tim Baldwin)
10:15 - 10:45	MORNING TEA
10:45 - 12:15	Session 1 – Linguistic Diversity in NLP (Session Chair: Mark Dras) Long papers are 20 minutes and short papers are 12 minutes.

Towards a Robust Morphological Analyser for Kunwinjku (**ALTA Best Student Paper Award**)
William Lane and Steven Bird

From Shakespeare to Li-Bai: Adapting a Sonnet Model to Chinese Poetry (long)
Zhuohan Xie, Jey Han Lau AND Trevor Cohn

Readability of Twitter Tweets for Second Language Learners (long)
Patrick Jacob and Alexandra Uitdenbogerd

A neural joint model for Vietnamese word segmentation, POS tagging and dependency parsing (short)
Dat Quoc Nguyen

Modelling Tibetan Morphology (short)
Qianji Di, Ekaterina Vylomova and Timothy Baldwin

12:15 - 13:15	LUNCH
13:15 - 14:15	Keynote 2: Wilson Wong (GO1) *Findability and discoverability in learning and employment*
14:15 - 15:25	Session 2 – Language Use and Applications (Shared ADCS Session, Session Chair: Alistair Moffat) ADCS papers are 25/15 mins and ALTA papers are 20/12 mins for the long/short format

Differences in language use: Insights from job and talent search (ADCS short)
Bahar Salehi, Borhan Kazimipour and Timothy Baldwin

Character profiling in low-resource language documents (ADCS short)
Tak-Sum Wong and John Lee

Towards a model for spoken conversational search (ADCS long encore presentation)
Johanne R. Trippas, Damiano Spina, Paul Thomas, Mark Sanderson, Hideo Joho and Lawrence Cavedon

A multi-constraint hinge loss for named-entity recognition (ALTA short)
Hanieh Poostchi and Massimo Piccardi

15:25 - 16:00	AFTERNOON TEA

16:00 - 17:25 Session 3 – Application and Evaluation (Session Chair: Andy MacKinlay/Massimo Piccardi)

Grounding learning of modifier dynamics: an application to color naming (short abstract presentation)
Xudong Han, Philip Shultz and Trevor Cohn
Feature-guided Neural Model Training for Supervised Document Representation Learning (short)
Aili Shen, Bahar Salehi, Jianzhong Qi and Timothy Baldwin
Red-faced ROUGE: Examining the Suitability of ROUGE for Opinion Summary Evaluation
(**ALTA 2nd Place Best Student Paper Award**)
Wenyi Tay, Aditya Joshi, Xiuzhen Zhang, Sarvnaz Karimi and Stephen Wan
Modeling Political Framing Across Policy Issues and Contexts (short)
Shima Khanehzar, Andrew Turpin and Gosia Mikolajczak
Box Embeddings for Inferring Predicate Entailment (long abstract presentation)
Ian Wood, Mark Johnson, Stephen Wan, Javad Housseini and Mark Steedman

19:00 - late DINNER

6th December (Friday) Day 3

Keynote: Mark Johnson (Macquarie University and Oracle)
9:00 - 10:00 *Building new kinds of Natural Language Understanding and Conversational AI with Deep Learning*
(Session Chair: Diego Molla-Aliod)

10:00 - 11:00 MORNING TEA AND POSTER SESSION

11:00 - 12:05 Session 4 – Parsing and Sequential Modelling (Session Chair: Trevor Cohn)

Improved Document Modelling with a Neural Discourse Parser (long)
Fajri Koto, Jey Han Lau and Timothy Baldwin
Does an LSTM forget more than a CNN? (long)
Gaurav Arora, Afshin Rahimi and Timothy Baldwin
Domain Adaptation for Low-Resource Neural Semantic Parsing (short)
Alvin Kennardi, Gabriela Ferraro and Qing Wang
A Pointer Network Architecture for Context Dependent Semantic Parsing (short)
Xuanli He, Quan Tran and Gholamreza Haffari

12:05 - 13:00 LUNCH

13:00 - 14:00 ADCS Keynote: Guido Zuccon (QUT)
Better Search, Better Health? Search engines, their evaluation and the impact on health decisions

14:00 - 15:00 Session 5 – Science and Medicine (Shared ADCS Session, Session Chair: Sarvnaz Karimi)

Detecting Chemical Reactions in Patents (**ALTA 2019 Best Paper Award**)
Hiyori Yoshikawa, Dat Quoc Nguyen, Zenan Zhai, Christian Druckenbrodt, Camilo Thorne,
Saber A. Akhondi, Timothy Baldwin and Karin Verspoor
Identifying Patients with Pain in Emergency Departments
Using Conventional Machine Learning and Deep Learning (**ALTA 2nd Place Best Paper Award**)
Thanh Vu, Anthony Nguyen, Nathan Brown and James Hughes
Learning inter-sentence, disorder-centric, biomedical relationships from medical literature (ADCS encore)
Anton van der Vegt, Guido Zuccon, Bevan Koopman

15:00 - 15:30 AFTERNOON TEA

15:30 - 16:00 ALTA GENERAL MEETING

16:00 - 16:45 Session 6 – Shared Task and Best Paper Presentations (Session Chair: Karin Verspoor)

Shared Task Introduction
Karin Verspoor
Overview of the ALTA 2019 Shared Task: Sarcasm Target Identification
Diego Molla-Aloid and Aditya Joshi
ALTA 2019 Shared Task Winner: Detecting Target of Sarcasm using Ensemble Methods
Pradeesh Parameswaran, Andrew Trotman, Veronica Liesaputra and David Eyers
ALTA 2019 Best Paper Awards
Closing Remarks

Table of Contents

Long Papers

Short Papers

Long Papers (Posters)