

Australasian Language Technology Workshop 2007

Proceedings of the Workshop

Workshop Chairs:
Nathalie Colineau
Mark Dras

10–11 December 2007
Melbourne Zoo
Melbourne, Australia

Proceedings of the Australasian Language Technology Workshop 2007

URL: <http://www.alta.asn.au/events/alta2007/>

Sponsors:



ISSN 1834-7037 (for the online proceedings)

Preface

This volume contains the papers accepted for presentation at this year's Australasian Language Technology Workshop (ALTA2007), held at the Melbourne Zoo, Melbourne, Australia, on December 10–11, 2007. This is the fifth annual installment of the workshop in its most recent incarnation, and the continuation of an annual workshop series that has existed under various guises since the early 90s.

The goals of the workshop are:

- to bring together the growing Language Technology (LT) community in Australia and New Zealand and encourage interactions;
- to encourage interactions between this community and the international LT community;
- to foster interaction between academic and industrial researchers;
- to encourage dissemination of research results;
- to provide a forum for the discussion of new and ongoing research and projects;
- to provide an opportunity for the broader artificial intelligence community to become aware of local LT research; and, finally,
- to increase visibility of LT research in Australia, New Zealand and overseas.

This year has seen quite some activity in language technology in Australia, with PACLING 2007, the 10th Conference of the Pacific Association for Computational Linguistics, being hosted in September also in Melbourne. In addition, this year ALTA is co-located with the Australian Document Computing Symposium (ADCS), including a joint session of papers and talks of interest to both communities.

This year's Australasian Language Technology Workshop includes regular talks as well as poster presentations and student posters. Of the 23 papers submitted, 16 papers were selected by the program committee for publication and appear in these proceedings. Of these, 10 are oral presentations and 6 are poster presentations. Additionally, we have included 3 student posters to encourage feedback on early results. Each full-length submission was independently peer reviewed by at least two members of the international program committee, in accordance with the DEST requirements for E1 conference publications.

We would like to thank all the authors who submitted papers; the members of the program committee for the time and effort they contributed in reviewing the papers; and our invited speakers, Sophia Ananiadou, Nick Thieberger and Justin Zobel. Our thanks also go to local organisers Nicola Stokes and Lawrence Cavedon, to members of the ALTA executive for their support in organising the workshop, and to our sponsors (NICTA, CSIRO, Inference Communications, Appen, and the University of Melbourne), who enabled us in particular to support student participation in the event.

Nathalie Colineau and Mark Dras
Programme co-chairs

Workshop Co-Chairs

Nathalie Colineau (CSIRO, Sydney)
Mark Dras (Macquarie University, Sydney)

Workshop Local Organiser

Nicola Stokes (NICTA and University of Melbourne, Melbourne)

Programme Committee

Timothy Baldwin (University of Melbourne, Melbourne)
Francis Bond (NTT Communication Science Labs, Japan)
Lawrence Cavedon (National ICT Australia and RMIT, Melbourne)
Fang Chen (National ICT Australia, Sydney)
Eric Choi (National ICT Australia, Sydney)
Nigel Collier (NII – National Institute of Informatics, Japan)
Dominique Estival (Appen Pty Ltd, Sydney)
Tanja Gaustad van Zaanen (Appen Pty Ltd, Sydney)
Michael Haugh (Griffith University, Brisbane)
Achim Hoffman (University of NSW, Sydney)
Chiori Hori (ATR – Advanced Telecommunication Research Institute International, Japan)
Ben Hutchinson (Google, Sydney)
Kentarō Inui (NAIST – Nara Institute of Science & Technology, Japan)
Kazunori Komatani (Kyoto University, Japan)
Richard Leibbrandt (Flinders University, Adelaide)
Chris Manning (Stanford University, US)
Kathleen McCoy (University of Delaware, US)
Nobuaki Minematsu (University of Tokyo, Japan)
Diego Mollá Aliod (Macquarie University, Sydney)
Hwee Tou Ng (National University of Singapore, Singapore)
Jon Patrick (University of Sydney, Sydney)
Mark Pedersen (University of Queensland, Brisbane)
David Powers (Flinders University, Adelaide)
Long Qiu (National University of Singapore, Singapore)
Hendra Setiawan (National University of Singapore, Singapore)
Tony Smith (Waikato University, Hamilton, NZ)
Seyed M. M. Tahaghoghi (RMIT University, Melbourne)
James A. Thom (RMIT University, Melbourne)
Dongqiang Yang (Flinders University, Adelaide)
Menno van Zaanen (Macquarie University, Sydney)
Ingrid Zukerman (Monash University, Melbourne)

Table of Contents

Invited Presentations

<i>Text Mining Techniques for Building a Biolexicon</i> Sophia Ananiadou	1
<i>Does Language Technology Offer Anything to Small Languages?</i> Nick Thieberger	2
<i>Measures of Measurements: Robust Evaluation of Search Systems</i> Justin Zobel	3

Full Papers

<i>Entailment due to Syntactically Encoded Semantic Relationships</i> Elena Akhmatova and Mark Dras	4
<i>Measuring Correlation Between Linguists' Judgments and Latent Dirichlet Allocation Topics</i> Ari Chanen and Jon Patrick	13
<i>TAT: An Author Profiling Tool with Application to Arabic Emails</i> Dominique Estival, Tanja Gaustad, Son Bao Pham, Will Radford and Ben Hutchinson	21
<i>Exploring Approaches to Discriminating among Near-Synonyms</i> Mary Gardiner and Mark Dras	31
<i>Practical Queries of a Massive n-gram Database</i> Tobias Hawker, Mary Gardiner and Andrew Bennetts	40
<i>Extending Sense Collocations in Interpreting Noun Compounds</i> Su Nam Kim, Meladel Mistica and Timothy Baldwin	49
<i>Named Entity Recognition in Question Answering of Speech Data</i> Diego Mollá, Menno van Zaanen and Steve Cassidy	57
<i>Experiments in Mutual Exclusion Bootstrapping</i> Tara Murphy and James R. Curran	66
<i>Charting Democracy Across Parsers</i> Scott Nowson and Robert Dale	75
<i>Two-Step Comprehensive Open Domain Text Annotation with Frame Semantics</i> Bahadorreza Ofoghi, John Yearwood and Liping Ma	83
<i>Question Prediction Language Model</i> Luiz Augusto Pizzato and Diego Mollá	92
<i>Exploring Abbreviation Expansion for Genomic Information Retrieval</i> Nicola Stokes, Yi Li, Lawrence Cavedon and Justin Zobel	100
<i>Parsing Internal Noun Phrase Structure with Collins' Models</i> David Vadas and James R. Curran	109
<i>An Empirical Investigation into Grammatically Constrained Contexts in Predicting Distributional Similarity</i> Dongqiang Yang and David M. W. Powers	117
<i>Dictionary Alignment for Context-sensitive Word Glossing</i> Willy Yap and Timothy Baldwin	125

<i>Statistical Machine Translation of Australian Aboriginal Languages: Morphological Analysis with Languages of Differing Morphological Richness</i>	
Simon Zwarts and Mark Dras	134

Student Papers

<i>Exploring Extensions to Machine-learning based Gene Normalisation</i>	
Benjamin Goudey, Nicola Stokes and David Martinez	143
<i>Distributional Similarity of Multi-Word Expressions</i>	
Laura Ingram and James R. Curran	146
<i>Extending CCGbank with Quotes and Multi-modal CCG</i>	
Daniel Tse and James R. Curran	149

Workshop Programme

DAY 1 — 10 DECEMBER, 2007

- 09:15 – 09:30 Welcome and Opening
- 09:30 – 10:00 *An Empirical Investigation into Grammatically Constrained Contexts in Predicting Distributional Similarity*
Dongqiang Yang and David Powers
- 10:00 – 10:30 *TAT: An Author Profiling Tool with Application to Arabic Emails*
Dominique Estival, Tanja Gaustad van Zaanen, Son Bao Pham, Will Radford and Ben Hutchinson
- 10:30 – 11:00 Coffee Break
- 11:00 – 11:30 *Extending Sense Collocations in Interpreting Noun Compounds*
Su Nam Kim, Meladel Mistica and Timothy Baldwin
- 11:30 – 12:00 *Dictionary Alignment for Context-sensitive Word Glossing*
Willy Yap and Timothy Baldwin
- 12:00 – 12:15 Lightning preview of posters
- 12:15 – 13:45 Lunch
- 13:45 – 15:15 POSTER SESSION
- Exploring Approaches to Discriminating among Near-Synonyms*
Mary Gardiner and Mark Dras
- Practical Queries of a Massive n-gram Database*
Tobias Hawker, Mary Gardiner and Andrew Bennetts
- Named Entity Recognition in Question Answering of Speech Data*
Diego Mollá, Menno van Zaanen and Steve Cassidy
- Experiments in Mutual Exclusion Bootstrapping*
Tara Murphy and James Curran
- Two-Step Comprehensive Open Domain Text Annotation with Frame Semantics*
Bahadorreza Ofoghi, John Yearwood and Liping Ma
- Question Prediction Language Model*
Luiz Augusto Pizzato and Diego Mollá
- Exploring Extensions to Machine-Learning Based Gene Normalisation*
Benjamin Goudey, Nicola Stokes and David Martinez
- Distributional Similarity of Multi-Word Expressions*
Laura Ingram and James Curran
- Extending CCGbank with Quotes and Multi-modal CCG*
Daniel Tse and James Curran

- 15:15 – 15:30 Walk through zoo to joint session with ADCS
- 15:30 – 16:00 Coffee Break
- 16:00 – 16:45 JOINT SESSION WITH ADCS
Exploring Abbreviation Expansion for Genomic Information Retrieval
 Nicola Stokes, Yi Li, Lawrence Cavedon and Justin Zobel
A Bottom-Up Term Extraction Approach for Web-Based Translation in Chinese-English IR Systems
 Chengye Lu, Yue Xu and Shlomo Geva (ADCS authors)
Automatic Thread Classification for Linux User Forum Information Access
 Timothy Baldwin, David Martinez and Richard B. Penman (ADCS authors)
- 16:45 – 17:45 *Measures of Measurements: Robust Evaluation of Search Systems*
 Invited Speaker – Justin Zobel
- 18:00 – 19:00 Canapes and Drinks

DAY 2 — 11 DECEMBER, 2007

- 09:15 – 10:15 *Text Mining Techniques for Building a Biolexicon*
 Invited Speaker – Sophia Ananiadou
- 10:15 – 10:45 *Measuring Correlation Between Linguist’s Judgments and Latent Dirichlet Allocation Topics*
 Ari Chanen and Jon Patrick
- 10:45 – 11:15 Coffee Break
- 11:15 – 11:45 *Charting Democracy Across Parsers*
 Scott Nowson and Robert Dale
- 11:45 – 12:15 *Parsing Internal Noun Phrase Structure with Collins’ Models*
 David Vadas and James R. Curran
- 12:15 – 13:30 Lunch
- 13:30 – 14:15 ALTA Annual General Meeting
- 14:15 – 14:45 *Entailment due to Syntactically Encoded Semantic Relationships*
 Elena Akhmatova and Mark Dras
- 14:45 – 15:15 *Statistical Machine Translation of Australian Aboriginal Languages: Morphological Analysis with Languages of Differing Morphological Richness*
 Simon Zwarts and Mark Dras
- 15:15 – 15:45 Coffee Break
- 15:45 – 16:45 *Does Language Technology Offer Anything to Small Languages?*
 Invited Speaker – Nick Thieberger
- 16:45 – 17:00 Awards and Closing Remarks