

Australasian Language
Technology Association
Workshop 2008

Proceedings of the Workshop



Workshop Chairs:
Nicola Stokes
David Powers

8-10 December 2008
CSIRO ICT Centre
Hobart, Australia

Proceedings of the Australasian Language Technology Association
Workshop 2008 (ALTA 2008)

URL: <http://www.alta.asn.au/events/alta2008/>

Sponsors:



ISSN 1834-7037 (online proceedings)

Preface

This volume contains the papers accepted for presentation at this year's Australasian Language Technology Workshop (ALTA2008), held at the CSIRO Tasmanian ICT Centre, Hobart, Australia, on December 8-10, 2008. This is the sixth annual installment of the workshop since the formation of ALTA, and the continuation of an annual workshop series that existed under various guises throughout the 1990s.

The goals of the workshop are:

- to bring together the Language Technology (LT) community in Australasia;
- to encourage interactions with the international LT community;
- to foster interaction between academic and industrial researchers;
- to encourage dissemination and peer evaluation of research results;
- to provide a forum for the discussion of new and ongoing research and projects;
- to allow the broader AI community to become aware of local LT research;
- to increase visibility of LT research in Australia, New Zealand and overseas;
- to promote, support and encourage the participation of students in LT research.

Once again, this year ALTA is co-located with the Australasian 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 which rank equally as papers. The number of papers submitted increased by 52% compared to last year, and 20 papers were selected by formal review to appear in these proceedings. This represents an increase of just 25% in the number of papers accepted, and acceptance rate is thus 57% as compared to 70% last year. Ten of the submissions came from outside Australasia, and two submissions were accepted from Indonesia.

Of the 20 accepted papers, 14 are oral presentations and 6 are poster presentations. 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. The conference is classified as category B in the Computing Research and Education Association of Australasia (CORE) classification system which will form the basis of ranking computer science conference publications for ARC.

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; Rosie Jones (Yahoo!) for joining us as Keynote Speaker, as well as ADCS for their joint support of the conference and keynotes. Our thanks also go to local organizers Dipak Bhandari and Shlomo Berkovsky (CSIRO), to webmasters Darius Pfitzner and Richard Leibbrandt (Flinders University) for their logistical support, to members of the ALTA executive for their assistance in organizing the workshop, and to our sponsors (NICTA and CSIRO) who enabled us in particular to support student participation and accommodation, as well as free registration for all early registrants.

David Powers and Nicola Stokes
Programme Co-Chairs

Workshop Co-Chairs

Nicola Stokes (University College Dublin)
David Powers (Flinders University)

Workshop Local Organizers

Dipak Bhandari (CSIRO, Tasmanian ICT Centre)
Shlomo Berkovsky (CSIRO)

Reviewers

Timothy Baldwin (University of Melbourne, Melbourne)
Alan Black (Carnegie-Mellon University)
Lawrence Cavedon (National ICT Australia, Melbourne)
Eric Choi (National ICT Australia, Sydney)
Nathalie Colineau (CSIRO)
Nigel Collier (National Institute of Informatics, Japan)
Fintan Costello (University College Dublin)
Hal Daume (University of Utah)
Mark Dras (Macquarie University, Sydney)
Ben Hachey (University of Edinburgh)
Michael Haugh (Griffith University, Nathan, Queensland)
Graeme Hirst (University of Toronto)
Achim Hoffman (University of NSW, Sydney)
Kazunori Komatani (Kyoto University)
Sarvnaz Karimi (NICTA, Victoria)
Su Nam Kim (National University of Singapore, Singapore)
Andrew Lampert (CSIRO)
Richard Leibbrandt (Flinders University, Adelaide)
David Martinez (University of Melbourne)
Diego Mollá Aliod (Macquarie University, Sydney)
Martina Naughton (University College Dublin, Ireland)
Ani Nenkova (University of Pennsylvania)
Jon Patrick (University of Sydney, Sydney)
Cecile Paris (CSIRO)
Darius Pfitzner (Flinders University, Adelaide)
Benoit Sagot (INRIA, France)
Hendra Setiawan (National University of Singapore, Singapore)
Seyed M M Tahaghoghi (RMIT University, Melbourne)
James A Thom (RMIT University, Melbourne)
Dongqiang Yang (Flinders University, Adelaide)
Steven Wan (Macquarie University, Sydney)
Zing Wei (University of Massachusetts)
Menno van Zaanen (Tilburg University, Netherlands)

Programme

Monday December 8, 2008

ALTA/ADCS Joint Session

3:45pm-5:00pm:

- Sumukh Ghodke and Steven Bird, "Querying linguistic annotations" (ADCS)
- Karl Grieser, Timothy Baldwin, Fabian Bohnert, and Liz Sonenberg, "Using collaboratively constructed document collections to simulate real world object comparisons" (ADCS)
- Katie Bell and James R Curran, "Answer Attenuation in Question Answering" (ALTA)
- Clint Burfoot, "Using multiple sources of agreement information for sentiment classification of political transcripts" (ALTA)

5:00pm-6:00pm: Joint ADCS/ALTA Keynote

- "Syntactic and Semantic Structure in Web Search Queries"
Rosie Jones, Yahoo!

6:00pm-7:00pm: Canapes and Drinks (Incorporating ADCS Best Paper Awards)

7:00pm-10:00pm Conference Dinner (not covered by registration)

Tuesday December 9, 2008

Session 1

9:30-10:30am: (3 papers)

- Andrew Lampert, Robert Dale and Cecile Paris, "Requests and Commitments in Email are More Complex Than You Think: Eight Reasons to be Cautious"
- Olivia March and Timothy Baldwin, "Automatic Event Reference Identification for Text Summarisation"

10:30am-11:00am: Coffee break

Session 2

11:00am-12:00pm: (2 papers)

- Susan Howlett and James Curran, "Automatic Acquisition of Training Data for Statistical Parsers"
- Tara McIntosh and James R Curran, "Weighted Mutual Exclusion Bootstrapping for Domain Independent Lexicon and Template Acquisition"

Poster Sessions (posters should be mounted by the end of morning tea time)

12:00pm-12:30pm: Speed papers I (3 speed papers = 30 mins)

- Femphy Pisceldo, Rahmad Mahendra, Ruli Manurung and I Wayan Arka, "A Two-Level Morphological Analyser for the Indonesian Language"

- Michael Fridkin and David Dowe, "Lexical Access via Phoneme to Grapheme conversion"
- Zhihan Li, Yue Xu and Shlomo Geva, "Text Mining Based Query Expansion for Chinese IR"

12:30pm-2:00pm: Lunch (posters may be previewed)

2:00pm-2:30pm: Speed papers II (3 speed papers =30 mins)

- Jonathan K Kummerfeld and James R Curran, "Classification of Verb Particle Constructions with the Google Web1T Corpus"
- Simon Zwarts and Mark Dras, "Morphosyntactic Target Language Matching in Statistical Machine Translation"
- Daniel Tse and James Curran, "Punctuation normalisation for cleaner treebanks and parsers"

2:30pm-3:30pm: - Posters

3:30pm-4:00pm: Coffee Break (poster session can continue)

Session 3

4:00pm-5:00pm: (2 papers)

- Jette Viethen and Robert Dale, "Generating Relational References: What Makes a Difference?"
- Eliza Margaretha and Ruli Manurung "Comparing the value of Latent Semantic Analysis on two English-to-Indonesian lexical mapping tasks: What is it good for?"

Wednesday December 10, 2008 (5 papers)

Session 4

9:30am-11:00: (3 papers)

- Dominick Ng, David J Kedziora, Terry T W Miu and James R Curran, "Investigating Features for Classifying Noun Relations"
- Joel Nothman, James Curran and Tara Murphy, "Transforming Wikipedia into Named Entity Training Data"
- Jeremy Nicholson and Timothy Baldwin, "Learning Count Classifier Preferences of Malay Nouns"

11:00-11:30am: Coffee break

Session 5

11:30am-1:00pm: (3 papers)

- Cecile Paris, Nathalie Colineau, Andrew Lampert and Joan Giralt Duran, "Fit it in but say it well!"
- Tobias Kuhn and Rolf Schwitter, "Writing Support for Controlled Natural Languages"

- Ari Chanen and Jon Patrick, "All-Topology, Semi-Abstract Syntactic Features for Text Categorization"

ALTA Best Paper & Best Presentation Awards

1:00pm-2:00pm: Lunch (incorporating ALTA Best Paper Awards)

2:00pm-2:45pm: ALTA Annual General Meeting

Table of Contents

Syntactic and Semantic Structure in Web Search Queries (invited talk) Rosie Jones	1
Answer Attenuation in Question Answering Katie Bell, James R Curran	2
Using Multiple Sources of Agreement Information for Sentiment Classification of Political Transcripts Clint Burfoot	11
All-Topology, Semi-Abstract Syntactic Features for Text Categorization Ari Chanen, Jon Patrick	19
Lexical Access via Phoneme to Grapheme Conversion (poster presentation) Michael Fridkin, David Dowe, Simon Musgrave	28
Automatic Acquisition of Training Data for Statistical Parsers Susan Howlett, James R Curran	37
Writing Support for Controlled Natural Languages Tobias Kuhn, Rolf Schwitter	46
Classification of Verb Particle Constructions with the Google Web1T Corpus (poster presentation) Jonathan K Kummerfeld, James R Curran	55
Requests and Commitments in Email are More Complex Than You Think: Eight Reasons to be Cautious Andrew Lampert, Robert Dale, Cecile Paris	64
Text Mining Based Query Expansion for Chinese IR (poster presentation) Zhihan Li, Yue Xu, Shlomo Geva	73
Automatic Event Reference Identification Olivia March, Timothy Baldwin	79
Comparing the Value of Latent Semantic Analysis on two English-to-Indonesian lexical mapping tasks Eliza Margaretha, Ruli Manurung	88
Weighted Mutual Exclusion Bootstrapping for Domain Independent Lexicon and Template Acquisition	97

Tara McIntosh, James R Curran	
Investigating Features for Classifying Noun Relations Dominick Ng, David J Kedziora, Terry T W Miu, James R Curran	106
Learning Count Classifier Preferences of Malay Nouns Jeremy Nicholson, Timothy Baldwin	115
Transforming Wikipedia into Named Entity Training Data Joel Nothman, James R Curran, Tara Murphy	124
Fit it in but say it well! Cecile Paris, Nathalie Colineau, Andrew Lampert, Joan Giralt Duran	133
A Two-Level Morphological Analyser for the Indonesian Language (poster presentation) Femphy Pisceldo, Rahmad Mahendra, Ruli Manurung, I Wayan Arka	142
Punctuation Normalisation for Cleaner Treebanks and Parsers (poster presentation) Daniel Tse, James Curran	151
Generating Relational References: What Makes a Difference? Jette Viethen, Robert Dale	160
Morphosyntactic Target Language Matching in Statistical Machine Translation (poster presentation) Simon Zwarts, Mark Dras	169