Australasian Language Technology Workshop 2005

Proceedings of the Workshop

Workshop Chairs:
Timothy Baldwin
James Curran
Menno van Zaanen

10-11 December 2005
University of Sydney
Sydney, Australia

Proceedings of the Australasian Language Technology Workshop 2005

URL: `http://www.alta.asn.au/events/altw2005/`

Sponsors:

Australian Government
Department of Defence
Defence Science and Technology Organisation

NATIONAL ICT AUSTRALIA LIMITED

CSIRO

APPEN

alta

The University of Sydney

ISBN: 0-9751687-2-X

To order copies of this and other ALTA proceedings, contact `workshop@alta.asn.au`

Introduction

This volume contains the papers accepted for presentation at the Australasian Language Technology Workshop (ALTW) 2005, held at the University of Sydney, Sydney, Australia, on 10-11th of December, 2005. This is the third 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.

One innovation in this year's Australasian Language Technology Workshop was the introduction of poster presentations in addition to the regular talks. Our intention here was to optimise the presentation medium for each submission, and reviewers were accordingly instructed to independently rate the acceptability of each submitted paper first as a regular paper and second as a poster. Specifically, consideration was given to: (a) which mode of delivery was most appropriate for a given submission (e.g. papers which were felt to benefit from a more interactive presentation were preferred as posters), (b) what was the technical merit of the submission (e.g. highly technical papers which relied on detailed explanation of a series of equations were preferred as regular papers), and (c) was the submission of general interest (e.g. papers describing general results with ramifications for a range of fields were preferred as regular papers). It is important to note that research quality and technical rigour were not taken into consideration in determining whether to accept a paper as a regular paper or poster. As such, regular papers and posters are of identical academic status.

Of the 45 papers submitted, 30 papers were selected by the programme committee for publication and appear in these proceedings. Of these, 14 are regular papers and 16 are posters. 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 F1 conference publications.

We would like to thank all the authors who submitted papers, as well as the members of the program committee for the time and effort they contributed in reviewing the papers, and Dan Flickinger, Kathy McKeown and Virach Sornlertlamvanich for providing the ideal complement to the workshop with their invited talks. Our thanks go also to members of the ALTA executive, and particularly Steven Bird and Cécile Paris for encouragement and support in organising the workshop. Finally, we would like to thank the sponsors (DSTO, NICTA, CSIRO and Appen) for their generous help in supporting the workshop, and Dominique Estival as sponsorship chair.

Timothy Baldwin, James Curran, Menno van Zaanen

Organizers:

Timothy Baldwin (University of Melbourne)
James Curran (University of Sydney)
Menno van Zannen (Macquarie University)

Program Committee:

Ash Asudeh (University of Canterbury)
Eric Atwell (Leeds University)
Timothy Baldwin (University of Melbourne)
Steven Bird (University of Melbourne)
Lawrence Cavedon (NICTA Victoria)
Trevor Cohen (University of Melbourne)
James Curran (University of Sydney)
Walter Daelemans (University of Antwerp)
Robert Dale (Macquarie University)
Dominique Estival (Defence Science and Technology Organisation)
Dan Flickinger (Oslo, Saarland and Stanford Universities)
Tanja Gaustad (Appen)
Graeme Hirst (University of Toronto)
Ben Hutchinson (University of Edinburgh)
Jong-bok Kim (Kyung Hee University)
Alistair Knott (University of Otago)
Valia Kordoni (University of Saarland)
Mirella Lapata (University of Edinburgh)
Hang Li (Microsoft Research)
Diana McCarthy (University of Sussex)
Daniel Midgley (University of Western Australia)
Diego Molla (Macquarie University)
Kyonghee Paik (KLI Language and Translation)
Ajeet Parhar (Telstra Research Laboratories)
Cécile Paris (CSIRO ICT Centre)
Jon Patrick (University of Sydney)
David Powers (Flinders University)
Tony Smith (Waikato University)
Harold Somers (University of Manchester)
Nicola Stokes (NICTA Victoria)
Takaaki Tanaka (NTT Communication Science Laboratories)
Aline Villavicencio (University of Essex)
Menno van Zaanen (Macquarie University)
Simon Zwarts (Macquarie University)

Invited Speakers:

Dan Flickinger (Oslo, Saarland and Stanford Universities)
Kathy McKeown (Columbia University)
Virach Sornlertlamvanich (NICT)

Table of Contents

Dimensions of Deep Grammar Validation
Dan Flickinger 1

Text Summarization: News and Beyond
Kathy McKeown 4

From Non-segmenting Language Processing to Web Language Engineering
Virach Sornlertlamvanich 5

Disambiguating Conjunctions in Named Entities
Pawel Mazur and Robert Dale 7

Learning of Graph Rules for Question Answering
Diego Molla and Menno van Zaanen 15

A Statistical Approach towards Unknown Word Type Prediction for Deep Grammars
Yi Zhang and Valia Kordoni 24

Tagging Unknown Words with Raw Text Features
David Vadas and James R. Curran 32

POS Tagging with a More Informative Tagset
Andrew MacKinlay and Timothy Baldwin 40

Augmenting Approximate Similarity Searching with Lexical Information
James Gorman and James R. Curran 49

Word Prediction in a Running Text: A Statistical Language Modeling for the Persian Language
Masood Ghayoomi and Seyyed Mostafa Assi 57

Using Diverse Information Sources to Retrieve Samples of Low Density Languages
Andrew MacKinlay 64

Faking it: Synthetic Text-to-speech Synthesis for Under-resourced Languages – Experimental Design
Harold Somers 71

Dual-Type Automatic Speech Recogniser Designs for Spoken Dialogue Systems
Jason Littlefield and Michael Broughton 78

Efficient Knowledge Acquisition for Extracting Temporal Relations
Son Bao Pham and Achim Hoffmann 87

Formal Grammars for Linguistic Treebank Queries
Mark Dras and Steve Cassidy 96

Extracting Exact Answers using a Meta Question Answering System
Luiz Augusto Pizzato and Diego Molla 105

Multimedia Presentation of Grammatical Description: Design Issues
Simon Musgrave 113

Structuring Documents Efficiently
Robert Marshall, Steven Bird and Peter Stuckey 120

Round-trip Translation: What Is It Good For?
Harold Somers 127

Evaluating the Utility of Appraisal Hierarchies as a Method for Sentiment Classification
Jeremy Fletcher and Jon Patrick 134

Efficient Grapheme-phoneme Alignment for Japanese
Lars Yencken and Timothy Baldwin 143

Statistical Interpretation of Compound Nominalisations
Jeremy Nicholson and Timothy Baldwin 152

Paraphrase Identification by Text Canonicalization
Yitao Zhang and Jon Patrick 160

Words and Word Usage: Newspaper Text versus the Web
Vinci Liu and James R. Curran 167

Automatic Induction of a POS Tagset for Italian
Raffaella Bernardi, Andrea Bolognesi, Corrado Seidenari and Fabio Tamburini 176

A Dual-Iterative Method for Concept-Word Acquisition from Large-Scale Chinese Corpora
Guogang Tian and Cungen Cao 184

Programming With Unrestricted Natural Language
David Vadas and James R. Curran 191

Identifying FrameNet Frames for Verbs from a Real-Text Corpus
Matthew Honnibal and Tobias Hawker 200

A Distributed Architecture for Interactive Parse Annotation
Baden Hughes, James Haggerty, Joel Nothman, Saritha Manickam and James R. Curran 207

Multi-document Summarisation and the PASCAL Textual Entailment Challenge
Nicola Stokes and Eamonn Newman 215

Design and Development of a Speech-driven Control for a In-car Personal Navigation System
Ying Su, Tao Bai and Catherine I. Watson 224

Combining Confidence Scores with Contextual Features for Robust Multi-Device Dialogue
Lawrence Cavedon, Matthew Purver and Florin Ratiu 233

Automatic Utterance Segmentation in Instant Messaging Dialogue
Edward Ivanovic 241

Workshop Programme

Day 1 — 10 December, 2005

09:25-09:30 Opening Remarks

09:30-10:00 *Disambiguating Conjunctions in Named Entities*
Pawel Mazur and Robert Dale

10:00-10:30 *Learning of Graph Rules for Question Answering*
Diego Molla and Menno van Zaanen

10:30-11:00 *A Statistical Approach towards Unknown Word Type Prediction for Deep Grammars*
Yi Zhang and Valia Kordoni

11:00-11:30 Coffee Break

11:30-12:00 *Tagging Unknown Words with Raw Text Features*
David Vadas and James R. Curran

12:00-12:30 *POS Tagging with a More Informative Tagset*
Andrew MacKinlay and Timothy Baldwin

12:30-13:00 *Augmenting Approximate Similarity Searching with Lexical Information*
James Gorman and James R. Curran

13:00-14:00 Lunch

14:00-15:00 *Dimensions of Deep Grammar Validation*
Invited Speaker – Dan Flickinger

15:00-15:15 Coffee Break

15:15-16:30 Poster Session 1

Word Prediction in a Running Text: A Statistical Language Modeling for the Persian Language
Masood Ghayoomi and Seyyed Mostafa Assi

Using Diverse Information Sources to Retrieve Samples of Low Density Languages
Andrew MacKinlay

Faking it: Synthetic Text-to-speech Synthesis for Under-resourced Languages – Experimental Design
Harold Somers

Dual-Type Automatic Speech Recogniser Designs for Spoken Dialogue Systems
Jason Littlefield and Michael Broughton

Efficient Knowledge Acquisition for Extracting Temporal Relations
Son Bao Pham and Achim Hoffmann

Formal Grammars for Linguistic Treebank Queries
Mark Dras and Steve Cassidy

Extracting Exact Answers using a Meta Question Answering System
Luiz Augusto Pizzato and Diego Molla

Multimedia Presentation of Grammatical Description: Design Issues
Simon Musgrave

16:30-17:00 *Structuring Documents Efficiently*
Robert Marshall, Steven Bird and Peter Stuckey

17:00-17:30 *Round-trip Translation: What Is It Good For?*
Harold Somers

17:30-18:00 *Evaluating the Utility of Appraisal Hierarchies as a Method for Sentiment Classification*
Jeremy Fletcher and Jon Patrick

DAY 2 — 11 DECEMBER, 2005

09:30-10:00 *Efficient Grapheme-phoneme Alignment for Japanese*
Lars Yencken and Timothy Baldwin

10:00-10:30 *Statistical Interpretation of Compound Nominalisations*
Jeremy Nicholson and Timothy Baldwin

10:30-11:00 *Paraphrase Identification by Text Canonicalization*
Yitao Zhang and Jon Patrick

11:00-11:30 Coffee Break

11:30-12:30 *Text Summarization: News and Beyond*
Invited Speaker – Kathy McKeown

12:30-14:00 Lunch

14:00-15:15 POSTER SESSION 2

Words and Word Usage: Newspaper Text versus the Web
Vinci Liu and James R. Curran

Automatic Induction of a POS Tagset for Italian
Raffaella Bernardi, Andrea Bolognesi, Corrado Seidenari and Fabio Tamburini

A Dual-Iterative Method for Concept-Word Acquisition from Large-Scale Chinese Corpora
Guogang Tian and Cungen Cao

Programming With Unrestricted Natural Language
David Vadas and James R. Curran

Identifying FrameNet Frames for Verbs from a Real-Text Corpus
Matthew Honnibal and Tobias Hawker

A Distributed Architecture for Interactive Parse Annotation
Baden Hughes, James Haggerty, Joel Nothman, Saritha Manickam and James R. Curran

Multi-document Summarisation and the PASCAL Textual Entailment Challenge
Nicola Stokes and Eamonn Newman

Design and Development of a Speech-driven Control for a In-car Personal Navigation System
Ying Su, Tao Bai and Catherine I. Watson

15:15-15:45 *Combining Confidence Scores with Contextual Features for Robust Multi-Device Dialogue*
Lawrence Cavedon, Matthew Purver and Florin Ratiu

15:45-16:15 *Automatic Utterance Segmentation in Instant Messaging Dialogue*
Edward Ivanovic

16:15-16:45 Coffee Break

16:45-17:45 *From Non-segmenting Language Processing to Web Language Engineering*
Invited Speaker – Virach Sornlertlamvanich

17:45-18:00 Award Ceremony and Closing Remarks