Abstract

The ALTA shared tasks are programming competitions where all participants attempt to solve the same problem, and the winner is the system with the best results. The 2011 ALTA shared task is the second in the series and it focuses on trying to automatically grade the level of clinical evidence in medical research papers. In this paper we describe the task, present the results of several baselines, and the results of our method. We apply a sequence of high precision machine learning classifiers with varying feature sets for each. In addition to using \( n \)-grams, we incorporate domain knowledge by representing specific medical concepts using their semantic categories. We also apply a specialised rule-based approach for automatically identifying the publication types of articles, which is then used as a feature set. Our approach obtains an accuracy of 62.84% which is a significant improvement over the baselines.

1 Introduction

An important step for physicians who practise Evidence Based Medicine (EBM) is the grading of the quality of the clinical evidence present in the medical literature. Evidence grading is a manual process, and the time required to perform it adds to the already time-consuming nature of EBM practice. It has been shown that EBM practitioners often do not pursue evidence based answers to clinical questions because of the time required (Ely et al., 1999; Ely et al., 2005). Therefore, there is a strong motivation for systems that can automatically appraise the evidence present in medical publications and generate evidence grades on a specialised scale.

The 2011 ALTA shared task addressed the problem of automatic evidence grading. The goal of the task was to build a system that can predict the grade of evidence given a set of medical publications from which the evidence has been extracted. This is a difficult task, and as we show below, machine learning methods that use simple bag-of-word features do not perform significantly better than a trivial baseline. We attempt to solve the problem using supervised machine learning using features such as abstract and title \( n \)-grams and publication types. We employ a set of classifiers that utilise the different feature sets and apply them sequentially to obtain an accuracy value of 62.84%, which is a significant improvement over the baseline and also the best result obtained among all the submissions for the shared task.

In the following sections, we provide a brief background of EBM, evidence grading, and related work in this area, followed by a description of our methods and the final results.

2 Evidence Based Medicine and Evidence Grading

EBM is the ‘conscientious, explicit, and judicious use of current best evidence in making decisions about the care of individual patients’ (Sackett et al., 1996). Current clinical guidelines urge physicians to practise EBM when providing care for their patients. Good practice of EBM requires practitioners to search for the best quality evidence, synthesise collected information and grade the quality of the
2.1 The Strength of Recommendation Taxonomy

There are over 100 grading scales to specify grades of evidence in use today. The Strength of Recommendation Taxonomy (SORT) (Ebell et al., 2004) is one such grading scale. It is a simple, straightforward and comprehensive grading system that can be applied throughout the medical literature. Consequently, it is used by various family medicine and primary care journals, such as the Journal of Family Practice (JFP)\(^1\). SORT uses three ratings — A (strong), B (moderate) and C (weak) — to specify the Strength of Recommendation (SOR) of a body of evidence. In SORT, grade A reflects a recommendation based on consistent and good-quality patient-oriented evidence; grade B reflects a recommendation based on inconsistent or limited-quality patient-oriented evidence; and grade C reflects a recommendation based on consensus, usual practice, opinion or disease-oriented evidence. This is the chosen grading scale for the ALTA shared task.

3 Related Work

Related research has focused mostly on automatic quality assessment of medical publications for purposes such as retrieval and post-retrieval re-ranking, where approaches based on word co-occurrences (Goetz and von der Lieth, 2005) and bibliometrics (Plikus et al., 2006) have been proposed for improving the retrieval of medical documents. Tang et al. (2009) propose a post-retrieval re-ranking approach that attempts to re-rank results returned by a search engine, which may or may not be published research work. However, their approach is only tested in a specific sub-domain (i.e., Depression) of the medical domain. Kilicoglu et al. (2009) focus on identifying high-quality medical articles and build on the work by Aphinyanaphongs et al. (2005). They use machine learning and obtain 73.7% precision and 61.5% recall. These approaches rely heavily on meta-data associated with the articles, making them dependent on the database from which the articles are retrieved. Hence, these approaches would not work on publications that do not have associated meta-data.

The definitions of ‘good-quality evidence’ (Ebell et al., 2004) suggest that the publication types of medical articles are good indicators of their qualities. Literature in the medical domain consists of a large number of publication types of varying qualities\(^2\). For example, a randomised controlled trial is of much higher quality than a case study of a single patient. Evidence obtained from the former is thus more reliable. Greenhalgh (2006) mentions some other factors that influence the grade of an evidence, such as the number of subjects included in a study and the mechanism by which subjects are allocated (e.g., randomisation/ no randomisation), but the latter is generally specified by the publication type (e.g., randomised controlled trial) of the article. Recently, Sarker and Molla (2010) emphasised on the importance of publication types for SOR determination and showed that automatic identification of high-quality publication types (e.g., Systematic Review and Randomised Controlled Trial) is relatively simple.

Factors influencing the automatic detection of evidence grades have been explored by Sarker et al. (2011). In this research work, information such publication types, publication years, journal titles, and article titles were obtained from a specialised corpus and used as features. Publication types were shown to be the most useful features giving accuracy values of approximately 68%. This research work is almost identical to the shared task. The only difference is that for the shared task, all features are required to be generated automatically since information from a specialised corpus is not available.

4 Methods

4.1 Shared Task Data

The data for the shared task consisted of a set of ‘evidences’ with the SOR grade for each. Each evidence was represented as a list of publications from which the evidence had been generated. Information for each publication was provided in the form of an

\(^1\)http://www.jfponline.com

\(^2\)A list of publication types used by the US National Library of Medicine can be found at http://www.nlm.nih.gov/mesh/pubtypes2006.html. This list is not exhaustive.
XML file per publication obtained from PubMed and named according to the publication PubMed ID. This XML file contained bibliographic data (title, author, etc.), the text of the abstract, and additional annotations provided by PubMed such as the medical semantic concepts found in the publication. Two sets of such data were provided initially for training (677 evidences) and development time testing (178 evidences), and an additional set was used for testing the final system (183 evidences).

An additional file contains the information related to the evidences, their SOR grades, and their publications (Figure 1). Each line represents an evidence. The first item in the line is the evidence ID. This is followed by the SOR grade (A, B, or C), and then the PubMed IDs of the abstracts that form the evidence. Thus, the first evidence listed in Figure 1 contains the abstracts with PubMed IDs 10553790 and 15265350, and is graded with SOR B.

The evidences were obtained from the corpus described by Mollá and Santiago-Martínez (2011), which in turn uses the references and SOR judgements present in the ‘Clinical Inquiries’ section of the website from the Journal of Family Practice.

4.2 Baselines

The most trivial baseline is to classify all of the elements with majority according to the training set, which is SOR B. With such a baseline, the accuracy is 48.63% (CI: 41.50-55.83).

A more complex baseline uses a machine learning classifier based on bag-of-word features. We tried with several variants. The best-performing system uses all non-stop n-grams \((n = 1, 2, 3)\) from the abstract after stemming and lowercasing as the features, and Naïve Bayes as the classifier, and achieves an accuracy of 45.90%. These results appear worse than the simpler baseline, though the difference is not statistically significant.

4.3 Preliminary Analysis

In our approach, we built on the work by Sarker et al. (2011). Our preliminary analysis involved using simple features such as n-grams and various other information (including publication types) from the training set data. As noted above, obtaining significant improvements over the ‘majority class’ baseline was extremely difficult using any classifier. Furthermore, the ‘PublicationType’ tags in the PubMed articles did not cover important publication types such as cohort study and systematic review. As a result, even the use of these tags did not produce accuracies greater than 60%. We therefore applied a rule-based approach for identifying publication types of articles and used them as features.

4.4 Feature Selection

Our final system utilises three feature sets — n-grams (semantic types replaced), titles, and publication types.

4.4.1 N-grams

We generated n-grams \((n = 1, 2, 3, 4)\) for each of the abstracts in the training set and replaced specific medical concepts in the texts with generic ‘sem_type’ tags. We used MetaMap to identify domain specific concepts as defined in the UMLS (Unified Medical Language System). The UMLS provides a vast vocabulary of medical concepts and also broad semantic groups into which the concepts can be classified. For example, all disease names fall under the semantic category Disease or Syndrome (dsyn). Replacing each occurrence of a disease or syndrome name with the generic tag ensures that the name does not have an influence on the classifiers used and reduces over-fitting. We used the same semantic groups as Uzuner et al. (2009): pathological function, disease or syndrome, mental or behavioural dysfunction, cell or molecular dysfunction, virus, neoplastic process, anatomic abnormality, acquired abnormality, congenital abnormality and injury or poisoning. We also preprocessed

\(^3\)http://www.ncbi.nlm.nih.gov/pubmed/  
\(^4\)Data obtained with kind permission from the publishers.  
\(^5\)We have experimented with other features but this combination produced the best results.  
\(^6\)http://metamap.nlm.nih.gov/  
\(^7\)http://www.nlm.nih.gov/research/umls/
the n-grams by stemming, lowercasing and removing stop words.

4.4.2 Publication Types

We employed a rule-based approach for automatically identifying publication types of the articles from their abstracts. It has been shown that such an approach obtains very accurate results for high quality publication types (Sarker and Mollá-Aliod, 2010). We extended this approach by creating regular expressions for identifying publication types such as cohort studies that are not tagged in the PubMed XML files. We combined the publication types identified by our rule-based approach with the publication types given in the articles. For articles with multiple publication types, we only kept the tag that represents the highest quality. For example, if an article was tagged as a Randomised Controlled Trial, a Clinical Trial, and a Journal Article, we only kept the Randomised Controlled Trial tag since it has the highest quality among the three types. In this way, we identified the publication types of all articles (total of 23 publication types) and used them as features.

4.4.3 Titles

Since titles have been shown to be informative and to produce better results than baseline in the past (Sarker et al., 2011), we used them as features as well. We generated uni- and bi-grams from the titles, preprocessed them (in the same manner as the n-grams) and used them as features.

4.5 Classification

We modelled the problem of evidence grading as a three-way classification problem using the above-mentioned features. Our preliminary analysis revealed that combining a set of features for a single classifier does not produce significant improvements over the baseline. Furthermore, beating the majority class baseline is difficult itself. We, therefore, attempted to develop a sequential approach that would achieve small improvements in accuracy over the baseline at each step. Thus, we use a sequence of classifiers that attempt to separate A and C grade instances from B with high precision. At each step, instances classified as A or C are removed and the rest are passed on to the next step. The sequence in which the classifiers were applied and specific details about each of them are as follows:

**Step 1:** Classify all evidences as grade B (majority class).

**Step 2:** SVMs with n-grams \((n = 1, 2, 3, 4\) and semantic types replaced) as features. Parameters: \(cost = 2.0\) and \(\gamma = 0.0\). Attribute selection: Using the information gain measure to select the top 400 n-grams.

**Step 3:** SVMs with publication types as features. For each instance, the frequency of each publication type is used. Parameters: \(cost = 1.0\) and \(\gamma = 0.0\).

**Step 4:** SVMs with titles as features. Parameters: \(cost = 32.0\) and \(\gamma = 0.002\).

The parameters for each of the SVMs were tuned using the training set for training and the development time test set for evaluation. All experiments were carried out using the software package Weka.\(^8\) Each of the above classifiers and their parameters were chosen based on their precision in classifying A and C grade evidences. Using this approach, the classification accuracy increases with each step of the algorithm as more instances are correctly classified as A and C.

5 Results and Discussion

For the final evaluation, we trained all our classifiers using the training set and the development test set, and evaluated the performance using test set instances. Among the 183 instances of the test set, our classifiers classify 42 as grade A, 124 as grade B, and 17 as grade C. This achieves an overall accuracy of 62.84%, meaning that 115 instances out of the 183 were correctly classified. This is significantly better than the baseline of classifying all instances as grade B, which has an accuracy of 48.63% (CI: 41.50 – 55.83).

Our results show that extracting specific information such as the publication types from text can significantly improve accuracy of grading. As Sarker et

---

\(^8\)http://www.cs.waikato.ac.nz/ml/weka/
al. (2011) point out, features such as sizes of studies and consistency among studies play an important role in influencing evidence grades. However, identifying these factors automatically pose difficult problems themselves.

6 Conclusion

The 2011 ALTA Shared Task turned out to be a difficult one. A simple bag-of-word baseline does not significantly improve the results of a trivial majority-based baseline, and in fact none of the participants to the shared task managed to achieve results significantly better than this trivial baseline except us.

We have approached the problem of evidence grading as a three-way classification problem. We use three feature sets — n-grams, publication types, and titles. For the n-grams, we apply generic tags for specific medical concepts and we obtain publication type information using a rule-based approach. By employing a sequence of classifiers that attempt to identify A and C grade classes with high precision, our approach obtains an accuracy of 62.84%, which is a significant improvement over the baseline.

References


Australasian Language Technology Association Workshop 2011

Proceedings of the Workshop

Editors:
Diego Mollá
David Martinez

1-2 December 2011
Australian National University
Canberra, Australia
Volume 9, 2011
ISSN: 1834-7037
ALTA 2011 Workshop Committees

Workshop Co-Chairs
- Diego Mollá (Macquarie University)
- David Martinez (NICeTA Victoria Research Lab and University of Melbourne)

Workshop Local Organisers
- Hanna Suominen (NICeTA Canberra Research Lab and ANU)
- Wray Buntine (NICeTA Canberra Research Lab and ANU)

Program Committee
- Timothy Baldwin (University of Melbourne)
- Francis Bond (Nanyang Technological University, Singapore)
- Wray Buntine (NICeTA Canberra Research Lab and ANU)
- Lawrence Cavedon (NICeTA Victoria Research Lab and University of Melbourne)
- Eric Choi (NICeTA)
- Jean-Yves Delort (Google, Zurich)
- Mark Dras (Macquarie University)
- Rebecca Dridan (University of Melbourne)
- Dominique Estival (University of Western Sydney)
- Tanja Gaustad (Tilburg University, The Netherlands)
- Ben Hachey (Macquarie University and CMCRF)
- Nitin Indurkya (University of New South Wales and eBay Research Labs)
- Alistair Knott (University of Otago, New Zealand)
- Kazunori Komatani (Nagoya University, Japan)
- Meladel Mistica (ANU)
- Scott Nowson (Appen Butler Hill)
- Cécile Paris (CSIRO)
- Son Bao Pham (Vietnam National University, Vietnam)
- Luiz Augusto Pizzato (University of Sydney)
- David Powers (Flinders University)
- Adam Saulwick (DSTO)
- Andrea Schalley (Griffith University)
- Rolf Schwitter (Macquarie University)
- Hanna Suominen (NICeTA Canberra Research Lab and ANU)
- Menno Van Zaanne (Tilburg University, The Netherlands)
- Jette Viethen (Tilburg University, The Netherlands)
- Wayne Wobcke (University of New South Wales)
- Simon Zwarts (Google, Sydney)
Preface

This volume contains the papers accepted for presentation at the Australasian Language Technology Workshop (ALTA) 2011, held at the Australian National University (ANU), Canberra, Australia on December 1-2, 2011. This is the ninth annual instalment of the ALTA workshop in its most-recent incarnation, and the continuation of an annual workshop series that has existed in various forms Down Under since the early 1990s.

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 collaboration within the community and with the wider 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’s ALTA Workshop includes 18 peer-reviewed papers, of which 12 have been presented orally, and 6 have been presented as posters. We received a total of 29 submissions. Each paper in the ‘peer-reviewed papers’ and the ‘peer-reviewed posters’ section was independently peer-reviewed by at least two members of an international program committee, in accordance with the DEST requirements for E1 conference publications. The review process was double-blind: Great care was exercised to avoid all conflicts of interest whenever an author also served as program committee/co-chair or the reviewer worked at the same institution as an author. Such conflicts of interest were resolved by transferring the reviewing task to other members of the program committee.

The proceedings also include the abstracts of the keynote presentation by Wray Buntine (National ICT Australia Ltd – NICTA) and of the special presentation by Dominique Estival (MARCS Auditory Laboratories) about the Australian Computational Linguistics Olympiad (OzCLO). There is also description of the second ALTA shared task by Diego Mollá and Abeed Sarker (Macquarie University).

We would like to thank all the authors who submitted papers to ALTA, the members of the program committee for the time and effort they put into the review process, the local organisers for their commitment and work organising this conference, and our invited speaker: Wray Buntine.

Finally, we would like to thank our sponsors, NICTA, ANU and Appen Butler Hill for supporting the workshop.

Diego Mollá and David Martinez
Program Co-Chairs
## ALTA 2011 Program

**Thursday, 1st December 2011**

<table>
<thead>
<tr>
<th>Time</th>
<th>Activity</th>
</tr>
</thead>
<tbody>
<tr>
<td>9:00-10:00</td>
<td>Keynote: <em>Discovery in Text: Visualisation, Topics and Statistics</em> by Wray Buntine (NICTA, Canberra)</td>
</tr>
<tr>
<td>10:00-10:30</td>
<td>Coffee break</td>
</tr>
<tr>
<td>10:30-10:40</td>
<td>ALTA Opening remarks</td>
</tr>
<tr>
<td><strong>Session 1</strong></td>
<td><strong>10:40 - 12:20</strong></td>
</tr>
<tr>
<td>10:40-11:05</td>
<td>Benjamin Börschinger and Mark Johnson</td>
</tr>
<tr>
<td></td>
<td><em>A Particle Filter algorithm for Bayesian Wordsegmentation</em></td>
</tr>
<tr>
<td>11:05-11:30</td>
<td>Bevan Jones, Mark Johnson and Sharon Goldwater</td>
</tr>
<tr>
<td></td>
<td><em>Formalizing Semantic Parsing with Tree Transducers</em></td>
</tr>
<tr>
<td>11:30-11:55</td>
<td>Mark Johnson</td>
</tr>
<tr>
<td></td>
<td><em>Parsing in Parallel on Multiple Cores and GPUs</em></td>
</tr>
<tr>
<td>11:55-12:20</td>
<td>Mehdi Parviz, Mark Johnson, Blake Johnson and Jon Brock</td>
</tr>
<tr>
<td></td>
<td><em>Using Language Models and Latent Semantic Analysis to Characterise the N400m Neural Response</em></td>
</tr>
<tr>
<td>12:45-14:00</td>
<td>Lunch break</td>
</tr>
<tr>
<td><strong>Session 2</strong></td>
<td><strong>14:00 - 15:15</strong></td>
</tr>
<tr>
<td>14:00-14:25</td>
<td>Shunichi Ishihara</td>
</tr>
<tr>
<td></td>
<td><em>A Forensic Authorship Classification in SMS Messages: A Likelihood Ratio Based Approach Using N-gram</em></td>
</tr>
<tr>
<td>14:25-14:50</td>
<td>Su Nam Kim and Lawrence Cavedon</td>
</tr>
<tr>
<td></td>
<td><em>Classifying Domain-Specific Terms Using a Dictionary</em></td>
</tr>
<tr>
<td>14:50-15:15</td>
<td>Stephen Merity and James R. Curran</td>
</tr>
<tr>
<td></td>
<td><em>Frontier Pruning for Shift-Reduce CCG Parsing</em></td>
</tr>
<tr>
<td>15:15-15:45</td>
<td>Coffee break</td>
</tr>
<tr>
<td><strong>Session 3</strong></td>
<td><strong>15:45 - 16:15</strong></td>
</tr>
<tr>
<td>15:45-16:15</td>
<td>Poster presentations (5 mins per poster)</td>
</tr>
<tr>
<td></td>
<td>John Cocks and Te Taka Keegan</td>
</tr>
<tr>
<td></td>
<td><em>A word-based approach for diacritic restoration in Māori</em></td>
</tr>
<tr>
<td></td>
<td>Nobuagi Akagi and Francesco-Alessio Ursini</td>
</tr>
<tr>
<td></td>
<td><em>The Interpretation of Complement Anaphorae: the case of The Others</em></td>
</tr>
<tr>
<td></td>
<td>Francesco-Alessio Ursini and Nobuagi Akagi</td>
</tr>
<tr>
<td></td>
<td><em>The Interpretation of Plural Pronouns in Discourse: The Case of They</em></td>
</tr>
<tr>
<td></td>
<td>Jenny Mcdonald, Alistair Knott, Richard Zeng and Ayelet Cohen</td>
</tr>
<tr>
<td></td>
<td><em>Learning from student responses: A domain-independent natural language tutor</em></td>
</tr>
<tr>
<td></td>
<td>Md. Waliur Rahman Miah, John Yearwood and Sid Kulkarni</td>
</tr>
<tr>
<td></td>
<td><em>Detection of child exploiting chats from a mixed chat dataset as a text classification task</em></td>
</tr>
<tr>
<td></td>
<td>Marcin Nowina-Krowicki, Andrew Zschorn, Michael Pilling and Steven Wark</td>
</tr>
<tr>
<td></td>
<td><em>ENGAGE: Automated Gestures for Animated Characters</em></td>
</tr>
<tr>
<td>16:15-17:25</td>
<td>Presenters put up their posters</td>
</tr>
<tr>
<td>17:25-18:45</td>
<td>Poster session with ALS (drinks and nibbles provided)</td>
</tr>
<tr>
<td>19:30</td>
<td>Dinner with ADCS</td>
</tr>
</tbody>
</table>
Friday, 2nd December 2011

Session 4 - 9:00 - 10:40

Joint Session with ADCS
Li Wang, Diana Mccarthy and Timothy Baldwin
*Predicting Thread Linking Structure by Lexical Chaining*
Diego Mollá and Maria Elena Santiago-Martinez
*Development of a Corpus for Evidence Based Medicine Summarisation*
Mike Symonds, Peter Bruza, Laurianne Sitbon and Ian Turner
*Tensor Query Expansion: A cognitively motivated relevance model*
Yan Shen, Yuefeng Li, Yue Xu, Renato Lannella, Abdulmohsen Algarni and Xiaohui Tao
*An Ontology-based Mining Approach for User Search Intent Discovery*

10:40-11:10 Coffee break

Session 5 - 11:10-12:25

Paper Presentations

11:10-11:35 (best paper award) François Lareau, Mark Dras, Benjamin Börschinger and Robert Dale
*Collocations in multilingual text generation: Lexical Functions meet Lexical Functional Grammar*

11:35-12:00 Abeed Sarker, Diego Mollá and Cécile Paris
*Outcome Polarity Identification of Medical Papers*

12:00-12:25 Sze-Meng Jojo Wong, Mark Dras and Mark Johnson
*Topic Modeling for Native Language Identification*

12:25-14:00 Lunch break

14:00-15:00 Prizes and ALTA AGM Meeting

Special presentation - 15:00-15:30
Dominique Estival
*OzCLO: The Australian Computational Linguistic Olympiad*

15:30-16:00 Coffee break

Special presentation - 16:00-16:45
Diego Mollá and Abeed Sarker
*Automatic Grading of Evidence: The 2011 ALTA Shared Task*

16:45-17:00 Wrap-up

17:25-18:45 Poster session with ALS (drinks and nibbles provided)
Contents

Non-reviewed papers 1

Discovery in Text: Visualisation, Topics and Statistics
Wray Buntine 2

OzCLO: The Australian Computational Linguistic Olympiad
Dominique Estival 3

Automatic Grading of Evidence: the 2011 ALTA Shared Task
Diego Mollá and Abeed Sarker 4

Peer-reviewed papers: Oral presentations 9

A Particle Filter algorithm for Bayesian Wordsegmentation
Benjamin Börshinger and Mark Johnson 10

Formalizing Semantic Parsing with Tree Transducers
Bevan Jones, Mark Johnson and Sharon Goldwater 19

Parsing in Parallel on Multiple Cores and GPUs
Mark Johnson 29

Using Language Models and Latent Semantic Analysis to Characterise the N400m Neural Response
Mehdi Parviz, Mark Johnson, Blake Johnson and Jon Brock 38

A Forensic Authorship Classification in SMS Messages: A Likelihood Ratio Based Approach Using N-gram
Shunichi Ishihara 47

Classifying Domain-Specific Terms Using a Dictionary
Su Nam Kim and Lawrence Cavedon 57

Frontier Pruning for Shift-Reduce CCG Parsing
Stephen Merity and James R. Curran 66

Predicting Thread Linking Structure by Lexical Chaining
Li Wang, Diana McCarthy and Timothy Baldwin 76

Development of a Corpus for Evidence Based Medicine Summarisation
Diego Mollá and Maria Elena Santiago-Martinez 86

Collocations in Multilingual Natural Language Generation: Lexical Functions meet Lexical Functional Grammar
François Lareau, Mark Dras, Benjamin Börshinger and Robert Dale 95

Outcome Polarity Identification of Medical Papers
Abeed Sarker, Diego Mollá and Cécile Paris 105

Topic Modeling for Native Language Identification
Sze-Meng Jojo Wong, Mark Dras and Mark Johnson 115
Peer-reviewed papers: Poster presentations

A word-based approach for diacritic restoration in Māori
John Cocks and Te Taka Keegan

The Interpretation of Complement Anaphorae: the case of The Others
Nobuagi Akagi and Francesco-Alessio Ursini

The Interpretation of Plural Pronouns in Discourse: The Case of They
Francesco-Alessio Ursini and Nobuagi Akagi

Learning from student responses: A domain-independent natural language tutor
Jenny Mcdonald, Alistair Knott, Richard Zeng and Ayelet Cohen

Detection of child exploiting chats from a mixed chat dataset as a text classification task
Md. Waliur Rahman Miah, John Yearwood and Sid Kulkarni

ENGAGE: Automated Gestures for Animated Characters
Marcin Nowina-Krowicki, Andrew Zschorn, Michael Pilling and Steven Wark