ACL 2013

Predicting and Improving Text Readability for Target Reader Populations

Proceedings of the Workshop

August 8, 2013
Sofia, Bulgaria

Production and Manufacturing by
Omnipress, Inc.
2600 Anderson Street
Madison, WI 53704 USA

Order copies of this and other ACL proceedings from:

Association for Computational Linguistics (ACL)
209 N. Eighth Street
Stroudsburg, PA 18360
USA
Tel: +1-570-476-8006
Fax: +1-570-476-0860
acl@aclweb.org

ISBN 978-1-937284-64-0

Introduction

Welcome to the second International Workshop on Predicting and Improving Text Readability for Target Reader Populations (PITR).

The last few years have seen a resurgence of work on text simplification and readability. Examples include learning lexical and syntactic simplification operations from Simple English Wikipedia revision histories, exploring more complex lexico-syntactic simplification operations requiring morphological changes as well as constituent reordering, simplifying mathematical form, applications for target users such as deaf students, second language learners and low literacy adults, and fresh attempts at predicting readability.

The PITR 2013 workshop has been organised to provide a cross-disciplinary forum for discussing key issues related to predicting and improving text readability for target users. It will be held on August 8, 2013 in conjunction with the 51st Conference of the Association for Computational Linguistics in Sofia, Bulgaria, and is sponsored by the ACL Special Interest Group on Speech and Language Processing for Assistive Technologies (SIG-SLPAT).

These proceedings include nine papers that cover various perspectives on the topic. Papers this year fall into 3 broad categories: (i) Readability Enhancement, where the aim is to improve text quality in some way (e.g., inserting punctuation) or tailor text for specific users (e.g., hearing-impaired readers); (ii) Predicting the reading level of text, where approaches vary from psycho-linguistic measurements (e.g. reading time) to standard readability measures applied to particular genres (e.g., web texts); and (iii) Text Simplification, where papers address learning from corpora as well as evaluation metrics for simplification systems.

We hope this volume is a valuable addition to the literature, and look forward to an exciting Workshop.

Sandra Williams
Advaith Siddharthan
Ani Nenkova

Organizers:

Sandra Williams, The Open University, UK.
Advaith Siddharthan, University of Aberdeen, UK.
Ani Nenkova, University of Pennsylvania, USA.

Program Committee:

Julian Brooke, University of Toronto, Canada.
Kevyn Collins-Thompson, Microsoft Research (Redmond), USA.
Siobhan Devlin, University of Sunderland, UK.
Micha Elsner, University of Edinburgh, UK.
Thomas François, University of Louvain, Belgium.
Caroline Gasperin, TouchType Ltd., UK.
Albert Gatt, University of Malta, Malta.
Pablo Gervás, Universidad Complutense de Madrid, Spain.
Iryna Gurevych, Technische Universitat Darmstadt, Germany.
Raquel Hervás, Universidad Complutense de Madrid, Spain.
Véronique Hoste, University College Ghent, Belgium.
Matt Huenerfauth, The City University of New York (CUNY), USA.
Iustina Ilisei, University of Wolverhampton, UK.
Annie Louis, University of Pennsylvania, USA.
Hitoshi Nishikawa, NTT, Japan.
Ehud Reiter, University of Aberdeen, UK.
Horacio Saggion, Universitat Pompeu Fabra, Spain.
Irina Temnikova, University of Wolverhampton, UK.
Ielka van der Sluis, University of Groningen, The Netherlands.
Kristian Woodsend, University of Edinburgh, UK.

Invited Speaker:

Annie Louis, University of Edinburgh, UK.

Identifying outstanding writing: Corpus and experiments based on the science journalism genre

I will discuss the hitherto unexplored area of text quality prediction: identifying outstanding pieces of writing. A system to do this task will benefit article recommendation and information retrieval. To do the task, we need to not only be able to measure spelling, grammar and organization quality but also quantify creative and engaging writing and topic. In addition, new resources are needed as existing corpora are focused on non-native student writing, output of text generation systems and artificial manipulation to create texts with low quality writing.

I will propose the science journalism genre as an apt one for such text quality experiments. Science journalism pieces entertain a reader as much as they teach and inform. I will introduce a corpus of science journalism articles which we have collected for use in text quality studies. The corpus contains science journalism pieces from the New York Times split into two categories—written by award-winning journalists and others. This corpus offers many desirable properties which were unavailable in previous resources. It represents realistic differences in writing quality, samples are based on professional writers rather than language learners, contains thousands of articles, and is publicly available. I will also describe automatic measures based on visual elements, surprisal and structure of these articles which are indicative of outstanding articles in the corpus and also turn out complementary to traditional metrics to quantify readability and organization quality of writing.

Bio: Annie Louis is a Newton International Fellow at the University of Edinburgh. She completed her PhD at University of Pennsylvania with a thesis on text quality prediction. She has also worked on automatic summarization and discourse parsing. She is currently working on discourse and document-level issues in machine translation. Annie has received a EMNLP best paper award and a SIGDIAL best student paper award.

Table of Contents

Workshop Program

(August 8, 2013)

09:20 – 10.30 **Session 1: Plenary**

09:20 Welcome and Introduction

09:30 Invited Talk: *Identifying outstanding writing: Corpus and experiments based on the science journalism genre*
Annie Louis, University of Edinburgh

10:30 – 11.00 Coffee break

11:00 – 12.30 **Session 2: Posters**

11:00 Poster Teasers

11:20 Poster Session

Sentence Simplification as Tree Transduction
Dan Feblowitz and David Kauchak

Building a German/Simple German Parallel Corpus for Automatic Text Simplification
David Klaper, Sarah Ebling and Martin Volk

The C-Score – Proposing a Reading Comprehension Metrics as a Common Evaluation Measure for Text Simplification
Irina Temnikova and Galina Maneva

A Language-Independent Approach to Automatic Text Difficulty Assessment for Second-Language Learners
Wade Shen, Jennifer Williams, Tamas Marius and Elizabeth Salesky

Guest paper: *A System for the Simplification of Numerical Expressions at Different Levels of Understandability*
Susana Bautista, Raquel Hervás, Pablo Gervás, Richard Powerand Sandra Williams (2013). Proc. Workshop on NLP for Improving Textual Accessibility (NLP4ITA), Atlanta, USA, pp.10–19.

12:30 – 14:00 Lunch break

14:00 – 15:30	**Session 3: Presentations**
	14:00 *Text Modification for Bulgarian Sign Language Users* Slavina Lozanova, Ivelina Stoyanova, , Svetlozara Leseva, , Svetla Koeva and Boian Savtchev
	14:20 *Modeling Comma Placement in Chinese Text for Better Readability using Linguistic Features and Gaze Information* Tadayoshi Hara, Chen Chen, Yoshinobu Kano and Akiko Aizawa
	14:40 *On The Applicability of Readability Models to Web Texts* Sowmya Vajjala and Detmar Meurers
	15:00 *Report from NLP4ITA 2013* Horacio Saggion
15:20 – 16:00	Tea break
16:00 – 1700	**Session 4: Presentations and Close**
	16:00 *The CW Corpus: A New Resource for Evaluating the Identification of Complex Words* Matthew Shardlow
	16:20 *A Pilot Study of Readability Prediction with Reading Time* Hitoshi NISHIKAWA, Toshiro MAKINO and Yoshihiro MATSUO
	16:50 Final Discussion and Close