TSAR 2022

Workshop on Text Simplification, Accessibility, and Readability

Proceedings of the Workshop

December 8, 2022

The TSAR organizers gratefully acknowledge the support from the following sponsors.

We acknowledge Frontiers as an official sponsor of TSAR 2022

©2022 Association for Computational Linguistics

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 978-1-959429-25-8

Introduction

Welcome to the proceedings of the 1st edition of the Workshop on Text Simplification, Accessibility and Readability (TSAR).

We have received 35 submissions to the workshop's main track and 11 submissions describing the systems that participated in the shared task on lexical simplification for English, Spanish, and Portuguese, held in conjunction with the TSAR workshop.

The submissions to the main track covered various topics: readability assessment, user studies, creation of datasets for text simplification in several domains and languages, novel text simplification architectures, use of complexity assessment in machine translation, as well as other related topics. Interestingly, most submissions focussed on languages other than English. The lexical simplification systems that participated in the shared task explored various architectures, ranging from non-neural (dictionary-based) approaches to those using the latest GPT-3 models. The best systems outperformed the previous state-of-the-art lexical simplification system on the shared task benchmark dataset.

All submissions were peer-reviewed by the members of the program committee which includes distinguished specialists in text simplification, accessibility, and readability. Out of the 35 submissions to the workshop's main track, one submission was withdrawn by the authors, 15 were rejected and 19 were accepted. Out of 19 accepted papers, 8 were selected to be presented orally and 11 as posters. Out of 11 submissions submitted to the shared task system description track, one was desk-rejected, and 10 were accepted.

The workshop is held fully virtually. The program encompasses two sessions with a total of 9 oral presentations, a poster session with 21 poster presentations, two invited talks, and a round table discussion. The oral presentations feature 8 papers submitted to the workshop's main track and one system demonstration paper which describes the winning system in the English track of the shared task. The poster session features 12 papers from the workshop's main track and 9 papers which describe the participating systems of the shared task on lexical simplification for English, Spanish, and Portuguese.

We would like to thank the members of the program committee for their timely help in reviewing the submissions, all the authors for submitting their papers to the workshop, and all teams that participated in the shared task and submitted the outputs of their systems. We also thank Frontiers for sponsoring the workshop.

TSAR Organizing Committee
Sanja Štajner, Horacio Saggion, Daniel Ferrés, Matthew Shardlow, Kim Cheng Sheang, Kai North, Marcos Zampieri, and Wei Xu

Organizing Committee

Organizers

Sanja Štajner, Karlsruhe, Germany
Horacio Saggion, Universitat Pompeu Fabra, Spain
Daniel Ferrés, Universitat Pompeu Fabra, Spain
Matthew Shardlow, Manchester Metropolitan University, United Kingdom
Kim Cheng Sheang, Universitat Pompeu Fabra, Spain
Kai North, George Mason University, USA
Marcos Zampieri, George Mason University, USA
Wei Xu, Georgia Institute of Technology, USA

Program Committee

Program Committee

Raksha Agarwal, Indian Institute of Technology Delhi, India
Sweta Agrawal, University of Maryland, USA
Rodrigo Alarcon, Universidad Carlos III de Madrid, Spain
Oliver Alonzo, Rochester Institute of Technology, USA
Fernando Alva-Manchego, Cardiff University, United Kingdom
Yuki Arase, Osaka University, Japan
Susana Bautista, Universidad Francisco de Vitoria, Spain
Rémi Cardon, Université Catholique de Louvain, Belgium
Felice Dell'orletta, Institute for Computational Linguistics "Antonio Zampolli", Italy
Anna Dmitrieva, University of Helsinki, Finland
Sarah Ebling, University of Zurich, Switzerland
Richard Evans, University of Wolverhampton, United Kingdom
Núria Gala, Aix Marseille Université, France
Itziar Gonzalez-Dios, University of the Basque Country UPV/EHU, Spain
Natalia Grabar, Université de Lille, France
Raquel Hervas, Universidad Complutense de Madrid, Spain
Tomoyuki Kajiwara, Ehime University, Japan
Jaap Kamps, University of Amsterdam, Netherlands
David Kauchak, Pomona College, USA
Tannon Kew, University of Zurich, Switzerland
Reno Kriz, Johns Hopkins University, USA
Philippe Laban, Salesforce Research, USA
Bruce W. Lee, University of Pennsylvania, USA
Mounica Maddela, Georgia Institute of Technology, USA
Lourdes Moreno, Universidad Carlos III de Madrid, Spain
Nhung Nguyen, The University of Manchester, United Kingdom
Christina Niklaus, University of St. Gallen, Switzerland
Tadashi Nomoto, National Institute of Japanese Literature, Japan
Brian Ondov, National Library of Medicine, USA
Maja Popović, Dublin City University, Ireland
Piotr Przybyła, Institute of Computer Science, Polish Academy of Sciences, Poland
Jipeng Qiang, Yangzhou University, China
Evelina Rennes, Linköping University, Sweden
Regina Stodden, Heinrich Heine University Düsseldorf, Germany
Giulia Venturi, Institute of Computational Linguistics "Antonio Zampolli", Italy
Gayatri Venugopal, Symbiosis International (Deemed University), India
Laura Vásquez-Rodríguez, University of Manchester, United Kingdom
Daniel Wiechmann, Institute for Logic Language and Computation, Netherlands
Victoria Yaneva, University of Wolverhampton, United Kingdom

Secondary Reviewers

Omar Hassan
David Heineman
Michael Ryan

Keynote Talk: Human-Computer Interaction and Automatic Text Simplification: Understanding the Perspective of Deaf and Hard of Hearing Users

Matt Huenerfauth
Rochester Institute of Technology, New York

Abstract: While there have been major advances in automatic text simplification and other related natural language processing technologies, there has been much less research conducted with direct participation of users, to understand their needs for this technology nor how it can be best evaluated through their participation in studies. In this talk, I will discuss how research methods from human-computer interaction and computing accessibility for people with disabilities can illuminate the potential benefits of this technology for a specific user group who has been the focus of research at our laboratory: Deaf and Hard of Hearing adult readers. In prior research presented at the ACM CHI and ASSETS conferences, we have learned that reading-assistance tools that incorporate lexical simplification benefit DHH adult readers, and we have also found that these users prefer designs in which they have greater autonomy over which portions of text have been simplified and transparency as to whether text has been modified. Focusing specifically on DHH adults working in the computing and information technology professions, we have also conducted research on users' current reading practices, approaches they use when encountering difficult text, their interest in reading-assistance technologies, and specific design considerations that would affect their interest (e.g., sense of autonomy, privacy, or social acceptability of this technology in the workplace). Finally, our most recent work has been methodological in nature, in which we have identified specific types of questions that can be asked in studies with DHH adults, of various English literacy levels, to effectively measure the complexity and fluency of English texts that have been simplified. Beyond our specific findings for DHH readers, our work illustrates how human-computer interaction researchers can contribute to progress in the field of automatic text simplification and provide useful guidance and methodological tools for other researchers.

Bio: Matt Huenerfauth is a Professor and Dean of the Golisano College of Computer and Information Sciences at Rochester Institute of Technology (RIT). He studies the design of technology to benefit people who are Deaf or Hard of Hearing or who have low written-language literacy, and his team of research students operates bilingually in English and American Sign Language (ASL). He has secured $5.25 million in external research funding since 2007, including a U.S. National Science Foundation CAREER Award in 2008. He has authored over 115 peer-reviewed scientific journal articles, book chapters, and conference papers, including at top venues in human-computer interaction and computing accessibility. He is a five-time recipient of the Best Paper Award at the top computing research conference in the field of computing accessibility, the ACM SIGACCESS Conference on Computers and Accessibility (ASSETS), which is more than any other individual in the conference history. In 2021, he was elected Chair of the ACM SIGACCESS special interest group on accessible computing for a three-year term, and in 2019, he completed a maximum six-year term as editor-in-chief of the ACM Transactions on Accessible Computing (TACCESS) journal. In 2018, RIT awarded him the Trustees Scholarship Award, the university's highest honor for faculty research.

Keynote Talk: Beyond the state-of-the-art models: What is complex text, and what are we simplifying?

Sowmya Vajjala
National Research Council, Canada

Abstract: We have seen over two decades of NLP research on readability assessment and text simplification by now. But, what do we really mean by "readability", and how is a "simplified" text different from an unsimplified one? In this talk, I will try to explore this question by looking into relevant literature in education and psychology research, and attempt to connect them with NLP research. I will also explore whether the current explainable AI research will help in addressing this question. Through this **non-technical** talk, I hope to initiate a discussion on what else should we be doing apart from building state of the art readability and simplification models with standard datasets.

Bio: Sowmya Vajjala is a researcher in the Multilingual Text Processing group , within the Digital Technologies Research Center at National Research Council, Canada. She has worked extensively on automatic readability assessment in the past, and is currently interested in developing and studying methods to understand the generalizability of NLP systems. She is also a co-author of "Practical Natural Language Processing", published by O'Reilly Media (2020).

Table of Contents

The Fewer Splits are Better: Deconstructing Readability in Sentence Splitting
Tadashi Nomoto 1

Parallel Corpus Filtering for Japanese Text Simplification
Koki Hatagaki, Tomoyuki Kajiwara and Takashi Ninomiya 12

Patient-friendly Clinical Notes: Towards a new Text Simplification Dataset
Jan Trienes, Jörg Schlötterer, Hans-Ulrich Schildhaus and Christin Seifert 19

Target-Level Sentence Simplification as Controlled Paraphrasing
Tannon Kew and Sarah Ebling 28

Conciseness: An Overlooked Language Task
Felix Stahlberg, Aashish Kumar, Chris Alberti and Shankar Kumar 43

Revision for Concision: A Constrained Paraphrase Generation Task
Wenchuan Mu and Kwan Hui Lim 57

Controlling Japanese Machine Translation Output by Using JLPT Vocabulary Levels
Alberto Poncelas and Ohnmar Htun 77

IrekiaLF_es: a New Open Benchmark and Baseline Systems for Spanish Automatic Text Simplification
Itziar Gonzalez-Dios, Iker Gutiérrez-Fandiño, Oscar m. Cumbicus-Pineda and Aitor Soroa . . . 86

Lexical Simplification in Foreign Language Learning: Creating Pedagogically Suitable Simplified Example Sentences
Jasper Degraeuwe and Horacio Saggion 98

Eye-tracking based classification of Mandarin Chinese readers with and without dyslexia using neural sequence models
Patrick Haller, Andreas Säuberli, Sarah Kiener, Jinger Pan, Ming Yan and Lena Jäger 111

A Dataset of Word-Complexity Judgements from Deaf and Hard-of-Hearing Adults for Text Simplification
Oliver Alonzo, Sooyeon Lee, Mounica Maddela, Wei Xu and Matt Huenerfauth 119

(Psycho-)Linguistic Features Meet Transformer Models for Improved Explainable and Controllable Text Simplification
Yu Qiao, Xiaofei Li, Daniel Wiechmann and Elma Kerz 125

Lexically Constrained Decoding with Edit Operation Prediction for Controllable Text Simplification
Tatsuya Zetsu, Tomoyuki Kajiwara and Yuki Arase 147

An Investigation into the Effect of Control Tokens on Text Simplification
Zihao Li, Matthew Shardlow and Saeed Hassan 154

Divide-and-Conquer Text Simplification by Scalable Data Enhancement
Sanqiang Zhao, Rui Meng, Hui Su and Daqing He 166

Improving Text Simplification with Factuality Error Detection
Yuan Ma, Sandaru Seneviratne and Elena Daskalaki 173

JADES: New Text Simplification Dataset in Japanese Targeted at Non-Native Speakers
Akio Hayakawa, Tomoyuki Kajiwara, Hiroki Ouchi and Taro Watanabe 179

A Benchmark for Neural Readability Assessment of Texts in Spanish
Laura Vásquez-Rodríguez, Pedro-Manuel Cuenca-Jiménez, Sergio Morales-Esquivel and Fernando Alva-Manchego 188

Controllable Lexical Simplification for English
Kim Cheng Sheang, Daniel Ferrés and Horacio Saggion 199

CILS at TSAR-2022 Shared Task: Investigating the Applicability of Lexical Substitution Methods for Lexical Simplification
Sandaru Seneviratne, Elena Daskalaki and Hanna Suominen 207

PresiUniv at TSAR-2022 Shared Task: Generation and Ranking of Simplification Substitutes of Complex Words in Multiple Languages
Peniel Whistely, Sandeep Mathias and Galiveeti Poornima 213

UoM&MMU at TSAR-2022 Shared Task: Prompt Learning for Lexical Simplification
Laura Vásquez-Rodríguez, Nhung Nguyen, Matthew Shardlow and Sophia Ananiadou 218

PolyU-CBS at TSAR-2022 Shared Task: A Simple, Rank-Based Method for Complex Word Substitution in Two Steps
Emmanuele Chersoni and Yu-Yin Hsu 225

CENTAL at TSAR-2022 Shared Task: How Does Context Impact BERT-Generated Substitutions for Lexical Simplification?
Rodrigo Wilkens, David Alfter, Rémi Cardon, Isabelle Gribomont, Adrien Bibal, Watrin Patrick, Marie-Catherine De marneffe and Thomas François 231

teamPN at TSAR-2022 Shared Task: Lexical Simplification using Multi-Level and Modular Approach
Nikita Nikita and Pawan Rajpoot 239

MANTIS at TSAR-2022 Shared Task: Improved Unsupervised Lexical Simplification with Pretrained Encoders
Xiaofei Li, Daniel Wiechmann, Yu Qiao and Elma Kerz 243

UniHD at TSAR-2022 Shared Task: Is Compute All We Need for Lexical Simplification?
Dennis Aumiller and Michael Gertz 251

RCML at TSAR-2022 Shared Task: Lexical Simplification With Modular Substitution Candidate Ranking
Desislava Aleksandrova and Olivier Brochu Dufour 259

GMU-WLV at TSAR-2022 Shared Task: Evaluating Lexical Simplification Models
Kai North, Alphaeus Dmonte, Tharindu Ranasinghe and Marcos Zampieri 264

Program

Thursday, December 8, 2022

09:30 - 09:45 *Opening Remarks*

09:45 - 10:30 *Session 1*

Parallel Corpus Filtering for Japanese Text Simplification
Koki Hatagaki, Tomoyuki Kajiwara and Takashi Ninomiya

Patient-friendly Clinical Notes: Towards a new Text Simplification Dataset
Jan Trienes, Jörg Schlötterer, Hans-Ulrich Schildhaus and Christin Seifert

IrekiaLF_es: a New Open Benchmark and Baseline Systems for Spanish Automatic Text Simplification
Itziar Gonzalez-Dios, Iker Gutiérrez-Fandiño, Oscar m. Cumbicus-Pineda and Aitor Soroa

10:30 - 11:00 *Coffee Break*

11:00 - 12:30 *Session 2*

Lexically Constrained Decoding with Edit Operation Prediction for Controllable Text Simplification
Tatsuya Zetsu, Tomoyuki Kajiwara and Yuki Arase

(Psycho-)Linguistic Features Meet Transformer Models for Improved Explainable and Controllable Text Simplification
Yu Qiao, Xiaofei Li, Daniel Wiechmann and Elma Kerz

A Dataset of Word-Complexity Judgements from Deaf and Hard-of-Hearing Adults for Text Simplification
Oliver Alonzo, Sooyeon Lee, Mounica Maddela, Wei Xu and Matt Huenerfauth

Eye-tracking based classification of Mandarin Chinese readers with and without dyslexia using neural sequence models
Patrick Haller, Andreas Säuberli, Sarah Kiener, Jinger Pan, Ming Yan and Lena Jäger

Findings of the TSAR-2022 Shared Task on Multilingual Lexical Simplification
Horacio Saggion, Sanja Štajner, Daniel Ferrés, Kim Cheng Sheang, Matthew Shardlow, Kai North and Marcos Zampieri

Thursday, December 8, 2022 (continued)

UniHD at TSAR-2022 Shared Task: Is Compute All We Need for Lexical Simplification?
Dennis Aumiller and Michael Gertz

12:30 - 14:00 *Lunch Break*

14:00 - 15:30 *Session 3 (Posters)*

15:30 - 16:00 *Coffee Break*

16:00 - 16:30 *Round Table Discussion*

16:30 - 17:30 *Invited Talk 1: Matt Huenerfauth*

17:45 - 18:45 *Invited Talk 2: Sowmya Vajjala*

18:45 - 19:00 *Closing Statements*