LREC-COLING 2024

DeTermIt! Evaluating Text Difficulty in a Multilingual Context (DeTermIt! 2024)

Workshop Proceedings

Editors
Giorgio Maria Di Nunzio, Federica Vezzani, Liana Ermakova, Hosein Azarbonyad, and Jaap Kamps

21 May, 2024
Torino, Italia

Proceedings of the Workshop on DeTermIt! Evaluating Text Difficulty in a Multilingual Context @ LREC-COLING 2024

ISBN 978-2-493814-15-9
ISSN 2951-2093 (COLING); 2522-2686 (LREC)

Jointly organized by the ELRA Language Resources Association
and the International Committee on Computational Linguistics

Preface from the General Chairs

Automatic Text Simplification (ATS) is the process that involves the reduction of linguistic complexity within a text to enhance its comprehensibility and readability. ATS plays a pivotal role in enhancing content and conveying clear, unambiguous information, while serving as a valuable preprocessing step, making texts more 'manageable' for various tasks like information extraction and retrieval. On a broader scale, ATS holds significant societal implications, particularly in assisting individuals with low literacy levels or those encountering challenges in reading comprehension.

One of the main barriers in text understanding is unfamiliar context and terminology. Even in developed countries, up to 30% of the population can only comprehend texts written with a basic vocabulary. Lexical simplification strives to enhance text comprehensibility for a broad audience by substituting intricate vocabulary and phrases with simpler alternatives while retaining the initial intended significance. Several initiatives emerged to help citizens with reading disabilities, e.g. the French project ALECTOR aims to leverage document accessibility for children with dyslexia. [1] EasyText.AI [2] focuses on text simplification for people with cognitive disabilities and provides simplifications of COVID-19-related texts in multiple languages. Finally, identification of difficult terms for second language learners can be helpful to optimize and personalize learning materials.

The *DeTermIt! Evaluating Text Difficulty in a Multilingual Context* workshop explores the theoretical and practical perspectives surrounding the evaluation of text difficulty in a multilingual context. In today's interconnected world, where information dissemination knows no linguistic bounds, it is mandatory to ensure that knowledge is accessible to diverse audiences, regardless of their language proficiency.

From a *theoretical* point of view, this workshop discusses the development of refined models and strategies for ATS. Additionally, the workshop promotes the study of the identification of common patterns and challenges encountered in different languages, which can lead to the creation of more effective tools and multilingual resources and promoting linguistic inclusivity. From a *practical* standpoint, the workshop considers the role of multilingual resources and their application in simplifying complex terminology. The development and utilization of language resources, such as bilingual and multilingual glossaries, translation memories, and terminology databases, are pivotal in achieving this goal. Furthermore, we analyze the effectiveness of machine translation and natural language processing techniques in aiding the simplification of text, and their implications for cross-linguistic text difficulty assessment.

The central inquiries in this workshop revolve around two key aspects: first, the theoretical elements that identify complexity within the text, and second the experimental analysis for simplifying the text to align with the reading proficiency of the target audience.

This first edition of DeTermIt! 2024 is co-located with the LREC-COLING 2024 joint conference and held in Turin, on May 21, 2024.[3]

The submitted papers went through a double-blind review process that required at least three reviews by members of the international scientific committee. We accepted 18 papers out of 29 submissions (62% acceptance rate): 12 long papers and 6 short papers.

[1] https://anr.fr/Project-ANR-16-CE28-0005

[2] https://easytext.ai/

[3] https://determit2024.dei.unipd.it/

Overall, these contributions encompass a diverse range of topics, showcasing the breadth and depth of research in the field of automatic text simplification. Papers deal with the development and refinement of text simplification systems in various languages, such as German, Finnish, French, and Arabic, reflecting a global interest in linguistic accessibility. Additionally, some studies explore innovative approaches to simplify complex scientific, legal, and governmental texts, aiming to enhance readability and comprehension. Multilingualism is a recurring theme, with papers addressing the challenges and opportunities of simplification across different linguistic contexts. Furthermore, advancements in lexical complexity prediction and the evaluation of simplification techniques through quantitative and qualitative research methodologies are examined, highlighting the interdisciplinary nature of the field.

The keynote speaker is Prof. Sara Carvalho (University of Aveiro, Portugal) with the title "Clear Communication, Better Healthcare: Leveraging Terminological Data for Automatic Text Simplification". By exploring the systematic representation and organization of terminological data, the talk is aimed at demonstrating how the double-dimensional approach to terminology has an impact on the development of ATS tools, ultimately enhancing patient-provider interactions and driving better healthcare outcomes.

Giorgio Maria Di Nunzio - Università degli Studi di Padova, Italy
Federica Vezzani - Università degli Studi di Padova, Italy
Liana Ermakova - Université de Bretagne Occidentale, France
Hosein Azarbonyad - Elsevier, The Netherlands
Jaap Kamps - University of Amsterdam, The Netherlands

Organizing Committee

General Chairs
Giorgio Maria Di Nunzio - Università degli Studi di Padova, Italy
Federica Vezzani - Università degli Studi di Padova, Italy
Liana Ermakova - Université de Bretagne Occidentale, France
Hosein Azarbonyad - Elsevier, The Netherlands
Jaap Kamps - University of Amsterdam, The Netherlands

Scientific Commitee
Hosein Azarbonyad - Elsevier, The Netherlands
Florian Boudin - Nantes University, France
Lynne Bowker - University of Ottawa, Canada
Sara Carvalho - Universidade de Aveiro, Portugal
Rute Costa - Universidade NOVA de Lisboa, Portugal
Giorgio Maria Di Nunzio - Università degli Studi di Padova, Italy
Eric Gaussier - University Grenoble Alpes, France
Natalia Grabar - CNRS, France
Jaap Kamps - University of Amsterdam, The Netherlands
Rodolfo Maslias - TermNet, Austria
Ana Ostroški Anić - Institute of Croatian Language and Linguistics, Croatia
Horacio Saggion - University Pompeu Fabra
Grigorios Tsoumakas - Aristotle University of Thessaloniki
Sara Vecchiato - University of Udine, Italy
Federica Vezzani - Università degli Studi di Padova, Italy
Cornelia Wermuth - KU Leuven, Belgium

Table of Contents