EACL 2023

The Second Ukrainian Natural Language Processing Workshop (UNLP 2023)

Proceedings of the Workshop

May 5, 2023

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 978-1-959429-52-4

Welcome to UNLP 2023

We warmly welcome you to the Second Ukrainian Natural Language Processing Workshop, held on May 5, 2023, in conjunction with EACL 2023!

The workshop brings together academics, researchers, and practitioners in the fields of natural language processing and computational linguistics who work with the Ukrainian language or do cross-Slavic research that can be applied to the Ukrainian language.

The Ukrainian NLP community has only started forming in recent years, with most of the projects done by isolated groups of researchers. The UNLP workshop provides a platform for discussion and sharing of ideas, encourages collaboration between different research groups, and improves the visibility of the Ukrainian research community.

This year, fifteen papers were accepted to be presented at the workshop. The papers present novel research in the areas of grammatical error correction, large language models, word and text embeddings, coreference resolution, summarization, and news classification. More than half of the papers present new datasets for the Ukrainian language, which is vital for further advances in any low-resource language. We are grateful to the program committee for their careful and thoughtful reviews of the papers submitted this year!

The Second UNLP features the first Shared Task on Grammatical Error Correction for Ukrainian. The shared task had two tracks: GEC-only and GEC+Fluency. The participating systems were expected to make the text grammatical or both grammatical and fluent, depending on the track. It was exciting to watch six teams compete and set the state-of-the-art results for Ukrainian GEC!

We believe that the UNLP 2023 shared task was instrumental in facilitating research on grammatical error correction for the Ukrainian language. All six competing systems were openly published, four teams submitted papers that were accepted to the UNLP workshop, and the CodaLab environment where the shared task was held remains open for further submissions.

UNLP 2023 will host two amazing keynote speeches by Mona Diab and Gulnara Muratova. The speakers will inspire the audience with their work on low-resource and endangered languages.

We are looking forward to the workshop and anticipate lively discussions covering a wide range of topics!

Organizers of UNLP 2023,
Mariana Romanyshyn, Oleksii Ignatenko, Oleksiy Syvokon, Andrii Hlybovets, Oleksii Molchanovskyi

Organizing Committee

Organizing Committee

Mariana Romanyshyn, Grammarly
Oleksii Molchanovskyi, Ukrainian Catholic University
Oleksiy Syvokon, Microsoft
Andrii Hlybovets, National University of Kyiv-Mohyla Academy
Oleksii Ignatenko, Ukrainian Catholic University

Program Committee

Program Committee

Andrii Babii, Kharkiv National University of Radio Electronics
Andrii Liubonko, Grammarly
Anna Rogers, University of Copenhagen
Artem Chernodub, Grammarly
Bogdan Babych, Heidelberg University
Bogdana Oliynyk, National University of Kyiv-Mohyla Academy, Silesian University of Technology
Dmytro Karamshuk, Facebook
Igor Samokhin, Grammarly
Julia Rogushina, Institute of software systems
Kostiantyn Omelianchuk, Grammarly
Maksym Tarnavskyi, Ukrainian Catholic University
Nataliia Cheilytko, Friedrich Schiller University Jena
Natalia Grabar, CNRS STL UMR8163, Université de Lille
Natalia Kotsyba, Insitute of Ukrainian, NGO; Samsung R&D Poland
Oleksandr Marchenko, Taras Shevchenko National University of Kyiv
Oleksandr Skurzhanskyi, Grammarly
Oleksii Turuta, Kharkiv National University of Radio Electronics
Olena Siruk, Bulgarian Academy of Sciences, Institute of Mathematics and Informatics
Olha Kanishcheva, University of Jena
Ruslan Chornei, National University of Kyiv-Mohyla Academy
Serhii Havrylov, University of Edinburgh, Institute for Language, Cognition and Computation
Svitlana Galeshchuk, PSL/Paris Dauphine, West Ukrainian National University, BNP Paribas
Taras Lehinevych, National University of Kyiv-Mohyla Academy
Taras Shevchenko, Giphy
Tatjana Scheffler, Ruhr-Universität Bochum
Thierry Hamon, LISN, Universite Paris-Saclay & Universite Sorbonne Paris Nord
Veronika Solopova, Freie Universität Berlin
Volodymyr Taranukha, Taras Shevchenko National University of Kyiv
Vsevolod Domkin, m8nware
Yevhen Kupriianov, National Technical University Kharkiv Polytechnic Institute"
Yuliia Makohon, Semantrum LLC

Table of Contents

Introducing UberText 2.0: A Corpus of Modern Ukrainian at Scale
Dmytro Chaplynskyi 1

Contextual Embeddings for Ukrainian: A Large Language Model Approach to Word Sense Disambiguation
Yurii Laba, Volodymyr Mudryi, Dmytro Chaplynskyi, Mariana Romanyshyn and Oles Dobosevych 11

Learning Word Embeddings for Ukrainian: A Comparative Study of FastText Hyperparameters
Nataliia Romanyshyn, Dmytro Chaplynskyi and Kyrylo Zakharov 20

GPT-2 Metadata Pretraining Towards Instruction Finetuning for Ukrainian
Volodymyr Kyrylov and Dmytro Chaplynskyi 32

The Evolution of Pro-Kremlin Propaganda From a Machine Learning and Linguistics Perspective
Veronika Solopova, Christoph Benzmüller and Tim Landgraf 40

Abstractive Summarization for the Ukrainian Language: Multi-Task Learning with Hromadske.ua News Dataset
Svitlana Galeshchuk 49

Extension Multi30K: Multimodal Dataset for Integrated Vision and Language Research in Ukrainian
Nataliia Saichyshyna, Daniil Maksymenko, Oleksii Turuta, Andriy Yerokhin, Andrii Babii and Olena Turuta 54

Silver Data for Coreference Resolution in Ukrainian: Translation, Alignment, and Projection
Pavlo Kuchmiichuk 62

Exploring Word Sense Distribution in Ukrainian with a Semantic Vector Space Model
Nataliia Cheilytko and Ruprecht von Waldenfels 73

The Parliamentary Code-Switching Corpus: Bilingualism in the Ukrainian Parliament in the 1990s-2020s
Olha Kanishcheva, Tetiana Kovalova, Maria Shvedova and Ruprecht von Waldenfels 79

Creating a POS Gold Standard Corpus of Modern Ukrainian
Vasyl Starko and Andriy Rysin 91

UA-GEC: Grammatical Error Correction and Fluency Corpus for the Ukrainian Language
Oleksiy Syvokon, Olena Nahorna, Pavlo Kuchmiichuk and Nastasiia Osidach 96

Comparative Study of Models Trained on Synthetic Data for Ukrainian Grammatical Error Correction
Maksym Bondarenko, Artem Yushko, Andrii Shportko and Andrii Fedorych 103

A Low-Resource Approach to the Grammatical Error Correction of Ukrainian
Frank Palma Gomez, Alla Rozovskaya and Dan Roth 114

RedPenNet for Grammatical Error Correction: Outputs to Tokens, Attentions to Spans
Bohdan Didenko and Andrii Sameliuk 121

The UNLP 2023 Shared Task on Grammatical Error Correction for Ukrainian
Oleksiy Syvokon and Mariana Romanyshyn 132

Program

Friday, May 5, 2023

09:00 - 09:10	*Opening Remarks*
09:10 - 09:55	*Keynote Speech: Mona Diab*
09:55 - 10:50	*Morning Session: New Datasets*
	Silver Data for Coreference Resolution in Ukrainian: Translation, Alignment, and Projection Pavlo Kuchmiichuk
	The Parliamentary Code-Switching Corpus: Bilingualism in the Ukrainian Parliament in the 1990s-2020s Olha Kanishcheva, Tetiana Kovalova, Maria Shvedova and Ruprecht von Waldenfels
	Creating a POS Gold Standard Corpus of Modern Ukrainian Vasyl Starko and Andriy Rysin
10:50 - 11:20	*Morning Break*
11:20 - 12:05	*Keynote Speech: Gulnara Muratova*
12:05 - 12:55	*Morning Session: New Directions*
	The Evolution of Pro-Kremlin Propaganda From a Machine Learning and Linguistics Perspective Veronika Solopova, Christoph Benzmüller and Tim Landgraf
	Extension Multi30K: Multimodal Dataset for Integrated Vision and Language Research in Ukrainian Nataliia Saichyshyna, Daniil Maksymenko, Oleksii Turuta, Andriy Yerokhin, Andrii Babii and Olena Turuta
	Exploring Word Sense Distribution in Ukrainian with a Semantic Vector Space Model Nataliia Cheilytko and Ruprecht von Waldenfels
12:55 - 14:25	*Lunch*

Friday, May 5, 2023 (continued)

14:25 - 16:00 *Afternoon Session: Shared Task*

UA-GEC: Grammatical Error Correction and Fluency Corpus for the Ukrainian Language
Oleksiy Syvokon, Olena Nahorna, Pavlo Kuchmiichuk and Nastasiia Osidach

The UNLP 2023 Shared Task on Grammatical Error Correction for Ukrainian
Oleksiy Syvokon and Mariana Romanyshyn

Comparative Study of Models Trained on Synthetic Data for Ukrainian Grammatical Error Correction
Maksym Bondarenko, Artem Yushko, Andrii Shportko and Andrii Fedorych

A Low-Resource Approach to the Grammatical Error Correction of Ukrainian
Frank Palma Gomez, Alla Rozovskaya and Dan Roth

RedPenNet for Grammatical Error Correction: Outputs to Tokens, Attentions to Spans
Bohdan Didenko and Andrii Sameliuk

15:50 - 16:00 *Best Paper and Thank You*

16:00 - 16:30 *Afternoon Break*

16:30 - 18:00 *Afternoon Session: UberText*

Introducing UberText 2.0: A Corpus of Modern Ukrainian at Scale
Dmytro Chaplynskyi

GPT-2 Metadata Pretraining Towards Instruction Finetuning for Ukrainian
Volodymyr Kyrylov and Dmytro Chaplynskyi

Learning Word Embeddings for Ukrainian: A Comparative Study of FastText Hyperparameters
Nataliia Romanyshyn, Dmytro Chaplynskyi and Kyrylo Zakharov

Friday, May 5, 2023 (continued)

Contextual Embeddings for Ukrainian: A Large Language Model Approach to Word Sense Disambiguation
Yurii Laba, Volodymyr Mudryi, Dmytro Chaplynskyi, Mariana Romanyshyn and Oles Dobosevych

Abstractive Summarization for the Ukrainian Language: Multi-Task Learning with Hromadske.ua News Dataset
Svitlana Galeshchuk

18:00 - 18:10 *Closing Words*