LREC-COLING 2024

The Third Ukrainian Natural Language Processing Workshop (UNLP 2024)

Workshop Proceedings

Editors
Mariana Romanyshyn

May 25, 2024
Torino, Italia

Proceedings of the Third Ukrainian Natural Language Processing Workshop (UNLP) @LREC-COLING-2024

ISBN 978-2-493814-43-2
ISSN 2951-2093 (COLING); 2522-2686 (LREC)

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

Welcome to UNLP 2024

We warmly welcome you to the Third Ukrainian Natural Language Processing Workshop, held on May 25, 2024, in conjunction with LREC-Coling 2024!

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, sixteen papers were accepted to be presented at the workshop. The papers showcase novel research in the areas of machine translation, news classification, named entity recognition, word sense disambiguation, and various aspects of developing and benchmarking large language models (LLMs) for Ukrainian. Over half of the papers introduce new datasets for the Ukrainian language. We are grateful to the program committee for their careful and thoughtful reviews of the papers submitted this year!

The third UNLP features the first Shared Task on Fine-Tuning Large Language Models for Ukrainian. The goal of the task was to facilitate the creation of models that have knowledge of the Ukrainian language, history, and culture, and are capable of generating fluent and factually accurate responses in Ukrainian. The participants were required to use models with open weights and of reasonable size, which ensured that the solutions would be usable in real-life scenarios. All solutions were openly published, and two teams submitted papers that were accepted to the UNLP workshop.

UNLP 2024 will host two amazing keynote speeches. Ivan Vulić will share his experience building equitable and culturally adapted multilingual dialog systems, while Vasyl Starko and Andriy Rysin will dive into the challenges of creating corpora for Ukrainian.

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

Organizers of UNLP 2024,
Mariana Romanyshyn, Oleksii Ignatenko, Nataliia Romanyshyn, Andrii Hlybovets, Oleksiy Syvokon, and Roman Kyslyi

Workshop Committees

Main Organizers

Andrii Hlybovets, National University of Kyiv-Mohyla Academy
Mariana Romanyshyn, Grammarly
Nataliia Romanyshyn, Ukrainian Catholic University
Oleksii Ignatenko, Ukrainian Catholic University

Shared Task Organizers

Mariana Romanyshyn, Grammarly
Oleksiy Syvokon, Microsoft
Roman Kyslyi, National Technical University of Ukraine "Igor Sikorsky Kyiv Polytechnic Institute"

Program Committee

Andrii Liubonko, Grammarly
Anna Rogers, University of Copenhagen
Anton Bazdyrev, Dun & Bradstreet
Artem Chernodub, Grammarly
Bogdan Babych, Heidelberg University
Dmytro Karamshuk, Meta
Igor Samokhin, Grammarly
Kostiantyn Omelianchuk, Grammarly
Maksym Tarnavskyi, Shelf
Natalia Grabar, CNRS, Université de Lille
Natalia Kotsyba, Samsung Research Poland
Nataliia Cheilytko, Friedrich Schiller University Jena
Oleksandr Marchenko, Taras Shevchenko National University of Kyiv
Oleksandr Skurzhanskyi, Grammarly
Oleksii Molchanovskii, Ukrainian Catholic University
Oleksii Turuta, Kharkiv National University of Radio Electronics
Olena Nahorna, Grammarly
Olha Kanishcheva, Friedrich Schiller University Jena
Ruslan Chornei, National University of Kyiv-Mohyla Academy
Serhii Havrylov, University of Edinburgh
Svitlana Galeshchuk, Université Paris Dauphine, BNP Paribas
Taras Lehinevych, Amazon
Taras Shevchenko, Giphy
Tatjana Scheffler, Ruhr-Universität Bochum
Thierry Hamon, Université Paris-Saclay, CNRS, LIMSI & Université Sorbonne
Vasyl Starko, Ukrainian Catholic University
Veronika Solopova, Technische Universität Berlin
Volodymyr Sydorskyi, National Technical University of Ukraine "Igor Sikorsky Kyiv Polytechnic Institute"

Volodymyr Taranukha, Taras Shevchenko National University of Kyiv
Vsevolod Dyomkin, Ukrainian Catholic University
Yevhen Kupriianov, National Technical University "Kharkiv Polytechnic Institute"
Yuliia Makohon, Semantrum
Yurii Paniv, Ukrainian Catholic University

Table of Contents

Workshop Program

Saturday, May 25, 2024

09:00–10:30 **Morning session 1: New Datasets**
Chair: Mariana Romanyshyn

09:00–09:10 Opening remarks

09:10–09:25 *A Contemporary News Corpus of Ukrainian (CNC-UA): Compilation, Annotation, Publication*
Stefan Fischer, Kateryna Haidarzhyi, Jörg Knappen, Olha Polishchuk, Yuliya Stodolinska and Elke Teich

09:25–09:40 *Introducing the Djinni Recruitment Dataset: A Corpus of Anonymized CVs and Job Postings*
Nazarii Drushchak and Mariana Romanyshyn

09:40–09:55 *Creating Parallel Corpora for Ukrainian: A German-Ukrainian Parallel Corpus (ParaRook||DE-UK)*
Maria Shvedova and Arsenii Lukashevskyi

09:55–10:10 *Introducing NER-UK 2.0: A Rich Corpus of Named Entities for Ukrainian*
Dmytro Chaplynskyi and Mariana Romanyshyn

10:10–10:20 Lightning talk: *Introducing CLARIN K-center for Ukrainian Language Research: Cooperation and Development*
Olha Kanishcheva

10:20–10:30 Lightning talk: *PAWUK: Polish Automatic Web corpus of UKrainian*
Witold Kieraś, Łukasz Kobyliński, Dorota Komosińska, Bartłomiej Nitoń, Michał Rudolf, Maria Shvedova and Aleksandra Zwierzchowska

10:30–11:00 **Coffee break**

11:00–13:00 **Morning session 2: New Directions**
Chair: Oleksii Ignatenko

11:00–11:20 *Instant Messaging Platforms News Multi-Task Classification for Stance, Sentiment, and Discrimination Detection*
Taras Ustyianovych and Denilson Barbosa

Saturday, May 25, 2024 (continued)

11:20–11:35 *Setting up the Data Printer with Improved English to Ukrainian Machine Translation*
Yurii Paniv, Dmytro Chaplynskyi, Nikita Trynus and Volodymyr Kyrylov

11:35–11:55 *Automated Extraction of Hypo-Hypernym Relations for the Ukrainian WordNet*
Nataliia Romanyshyn, Dmytro Chaplynskyi and Mariana Romanyshyn

11:55–12:10 *Ukrainian Visual Word Sense Disambiguation Benchmark*
Yurii Laba, Yaryna Mohytych, Ivanna Rohulia, Halyna Kyryleyza, Hanna Dydyk-Meush, Oles Dobosevych and Rostyslav Hryniv

12:10–13:00 Invited talk: *Towards Equitable and Culturally Adapted Multilingual Dialog Systems*
Ivan Vulić, University of Cambridge

13:00–14:00 Lunch

14:00–16:00 Afternoon session 1: LLMs for Ukrainian
Chair: Mariana Romanyshyn

14:00–14:15 *The UNLP 2024 Shared Task on Fine-Tuning Large Language Models for Ukrainian*
Mariana Romanyshyn, Oleksiy Syvokon and Roman Kyslyi

14:15–14:35 *Fine-Tuning and Retrieval Augmented Generation for Question Answering Using Affordable Large Language Models*
Tiberiu Boros, Radu Chivereanu, Stefan Dumitrescu and Octavian Purcaru

14:35–14:55 *From Bytes to Borsch: Fine-Tuning Gemma and Mistral for the Ukrainian Language Representation*
Artur Kiulian, Anton Polishko, Mykola Khandoga, Oryna Chubych, Jack Connor, Raghav Ravishankar and Adarsh Shirawalmath

14:55–15:15 *Spivavtor: An Instruction Tuned Ukrainian Text Editing Model*
Aman Saini, Artem Chernodub, Vipul Raheja and Vivek Kulkarni

15:15–15:35 *Eval-UA-tion 1.0: Benchmark for Evaluating Ukrainian (Large) Language Models*
Serhii Hamotskyi, Anna-Izabella Levbarg and Christian Hänig

15:35–15:55 *LiBERTa: Advancing Ukrainian Language Modeling through Pre-training from Scratch*
Mykola Haltiuk and Aleksander Smywiński-Pohl

Saturday, May 25, 2024 (continued)

16:00–16:30 **Coffee break**

16:30–18:00 **Afternoon session 2: LLMs for Ukrainian**
Chair: Oleksii Ignatenko

16:30–16:45 *Entity Embellishment Mitigation in LLMs Output with Noisy Synthetic Dataset for Alignment*
Svitlana Galeshchuk

16:45–17:00 *Language-Specific Pruning for Efficient Reduction of Large Language Models*
Maksym Shamrai

17:00–17:50 Invited talk: *BRUK Team's Resources for Ukrainian Corpus Creation*
Vasyl Starko, Ukrainian Catholic University, and Andriy Rysin, Independent researcher

17:50–18:00 Closing Words