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

Copyright ELRA Language Resources Association (ELRA), 2024
These proceedings are licensed under a Creative Commons
Attribution-NonCommercial 4.0 International License (CC BY-NC 4.0)

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

A Contemporary News Corpus of Ukrainian (CNC-UA): Compilation, Annotation, Publication

Stefan Fischer, Kateryna Haidarzhyi, Jörg Knappen,
Olha Polishchuk, Yuliya Stodolinska, Elke Teich
Universität des Saarlandes
Campus A2.2, 66123 Saarbrücken, Germany
{stefan.fischer, kateryna.haidarzhyi, olha.polishchuk, yuliya.stodolinska}@uni-saarland.de
{j.knappen, e.teich}@mx.uni-saarland.de

Abstract

We present a corpus of contemporary Ukrainian news articles published between 2019 and 2022 on the news website of the national public broadcaster of Ukraine, commonly known as SUSPILNE. The current release comprises 87 210 364 words in 292 955 texts. Texts are annotated with titles and their time of publication. In addition, the corpus has been linguistically annotated at the token level with a dependency parser. To provide further aspects for investigation, a topic model was trained on the corpus. The corpus is hosted (Fischer et al., 2023) at the Saarbrücken CLARIN center under a CC BY-NC-ND 4.0 license and available in two tab-separated formats: CoNLL-U (de Marneffe et al., 2021) and vertical text format (VRT) as used by the IMS Open Corpus Workbench (CWB; Evert and Hardie, 2011) and CQPweb (Hardie, 2012). We show examples of using the CQPweb interface, which allows to extract the quantitative data necessary for distributional and collocation analyses of the CNC-UA. As the CNC-UA contains news texts documenting recent events, it is highly relevant not only for linguistic analyses of the modern Ukrainian language but also for socio-cultural and political studies.

Keywords: corpus creation, contemporary news, Ukrainian

1. Introduction

This paper introduces a new contemporary news corpus for Ukrainian (CNC-UA), a corpus of modern Ukrainian news texts covering the 38-month period from November 2019 until December 2022. The corpus comprises 292 955 texts, mainly news articles but also reports with long tables. The CNC-UA is made available under a Creative Commons Attribution-Non-Commercial-NoDerivs 4.0 International License, while the underlying raw data are subject to the copyright of Суспільне Мовлення (Suspilne Movlennya, henceforth SUSPILNE), the Public Broadcasting Company of Ukraine.

While a number of corpora of Ukrainian do exist, overall the resource situation for Ukrainian is mixed. On the one hand, the availability of the existing corpora is often complicated (e.g. various search interfaces, no possibility of download, incomplete documentation). On the other hand, larger corpora are often not specialized enough to allow for serious linguistic or sociopolitical analysis (e.g. lack of contextual metadata).

Furthermore, due to the current situation in Ukraine, it is important to engage in the preservation and archival of news texts in a non-proprietary way for sociopolitical and linguistic documentation and subsequent scientific analysis. This is what motivated the creation of the CNC-UA.

This paper is structured as follows. We give an overview of existing Ukrainian corpora (Section 2) and explain the corpus building and annotation process of the CNC-UA (Section 3). To make the corpus as useful as possible, we have embedded it in an existing eco-system including a web-based corpus analysis platform as well as various standard formats. We describe the access to the CNC-UA and downloadable formats in Section 4. To demonstrate the application of the corpus and its accompanying infrastructure, we provide a short exploratory analysis (Section 5). We conclude with a brief summary and outlook in Sections 6 and 7.

2. Related Ukrainian Corpora

While Ukrainian can still be considered a low-resource language, the number of Ukrainian corpora is steadily increasing. Many of these corpora are available online, for example, the Ukrainian Language Corpus (Darchuk, 2017), the General Regionally Annotated Corpus of the Ukrainian Language (GRAC; Shvedova, 2020), the Ukrainian Text Corpus (Department of General and Applied Linguistics and Slavic Philology, Vasyl Stus Donetsk National University, 2023), and the Ukrainian Brown Corpus (BRUK; Starko and Rysin, 2023). Other important corpora, such as the National Ukrainian Linguistic Corpus (Shyrokov, 2011) or the Computer Fund for Innovation (Karpilovska, 2007) are currently inaccessible to the general public.

The General Regionally Annotated Corpus of the Ukrainian Language (GRAC) has a volume of

25 May, 2024. © 2024 ELRA Language Resource Association: CC BY-NC 4.0

1.781 billion tokens (v17). It is a vast and organized collection of texts in Ukrainian, allowing users to create subcorpora, search for words and grammatical forms, analyze search results, sort data, form balanced samples, and obtain statistical information via the Sketch Engine platform. The GRAC is a diachronic corpus spanning from 1816 to 2022 and contains over 130 000 texts from various genres. It contains a large subcorpus of journalism that includes collections of newspapers from the 19th and 20th centuries, contemporary newspapers, and texts from news sites on the web. The majority of texts come from printed sources. Notably, it includes a large corpus of diaspora texts, totaling about 40 million tokens. The corpus comprises both original and translated Ukrainian texts. However, no license is specified for this corpus and it is not downloadable.

The Ukrainian Language Corpus[1] (Darchuk, 2017) consists of morphological, syntactic, and semantic annotation layers. Currently, it contains more than 100 million tokens, partitioned into six subcorpora: journalism, fiction, scientific texts, legislative texts, poetic language, and folklore texts. The corpus is accessible through a corpus manager and is not downloadable.

The Ukrainian Text Corpus[2] (Department of General and Applied Linguistics and Slavic Philology, Vasyl Stus Donetsk National University, 2023) contains 120 000 word occurrences. It includes various genres such as journalistic, fictional, scientific, legislative, poetic, and folklore texts that have been processed automatically at morpheme, word, phrase, and sentence levels (part-of-speech, grammatical form, syntactic function).

The Ukrainian Brown Corpus (BRUK; Starko and Rysin, 2023) is an ongoing project aiming at creating an open, genre-balanced corpus of the modern Ukrainian language. The corpus contains text samples from 2010 to 2020 with a volume of 1 million words. It is built on the same principles as the well-known English Brown corpus. The texts were automatically tokenized, lemmatized and annotated with part-of-speech tags. The manual disambiguation of the corpus is still ongoing. Selected texts from national, regional, and local media, both print and online, make up approximately 25% of the BRUK. It is available for download under a Creative Commons Attribution-NonCommercial-ShareAlike 4.0 International License.

The University of Leipzig has collected a large corpus (Leipzig Corpora Collection, 2014; Goldhahn et al., 2012) of Ukrainian internet texts dating from 2014. This corpus is downloadable and it contains 1 546 330 404 tokens. Also, the online corpus portal allows to visualize text connectivity, and even offers a graph of interconnections.

The LORELEI Ukrainian Representative Language Pack (Tracey et al., 2020) includes Ukrainian monolingual texts, Ukrainian-English parallel and comparable texts, annotations, additional resources, and related software tools. This corpus contains 111 million words of Ukrainian text, of which about 700 000 words have been translated into English.

The UberText 2.0 corpus (Chaplynskyi, 2023) contains 3.274 billion tokens in 8.59 million texts. It has five subcorpora: news, fiction, social, wikipedia and court. The news subcorpus contains 2.173 billion tokens (before filtering), which are scraped from 38 central, regional, and industry-specific news websites. Among other steps, the processing pipeline includes lemmatization and POS tagging. The five subcorpora can be downloaded individually in different formats.

The Institute for Ukrainian (NGO)[3] is a joint Polish-Ukrainian project. The team has developed several corpora and a dedicated morphological analyzer. The Gold Standard Universal Dependencies Corpus for Ukrainian (Kotsyba et al., 2022) contains 140 000 tokens. The texts in the corpus have been manually annotated with morphological and syntactic dependency annotations. The corpus comprises a variety of text types, including articles, news, posts, textbooks, letters, fairy tales, and fiction. The news texts make up 5.6% of the total amount. There are no restrictions on the time of creation for the texts in the collection. The web corpus Zvidusil (Institute for Ukrainian, 2018), which is also part of this project, contains over 2.8 billion tokens. It has been automatically annotated and homonymy has been removed. The corpus includes texts from various freely available sources, such as user posts on social media, found mostly on the internet. To search the corpus[4], one can specify subcorpora based on source, title, author, and time of appearance of the texts. However, the texts do not contain further metadata. Statistical information and information about the search results are available. The web corpus Zvidusil includes news periodicals, such as Vysokyi Zamok, Den, Dzerkalo Tyzhnia, Zbruch, Radio Svoboda, Tyzhden, Ukraina Moloda, Ukrainska Pravda, etc.

The Ukrainian Web Corpus (ukTenTen 2022)[5] is a corpus composed of Ukrainian texts gathered from the internet. It belongs to the TenTen family (Suchomel, 2020; Jakubíček et al., 2013; Suchomel and Pomikálek, 2012) of corpora, which are a set of

[1] http://www.mova.info/corpus.aspx

[2] http://corpora.donnu.edu.ua/

[3] https://github.com/UniversalDependencies/UD_Ukrainian-IU

[4] https://mova.institute/bonito/run.cgi/corp_info?corpname=zvidusil

[5] https://www.sketchengine.eu/uktenten-ukrainian-corpus/

web corpora created using the same method with a target size of over 10 billion words. ukTenTen 2022 contains over 9.5 billion tokens and is classified by genre and topics. The data for the corpus consists of texts from May 2014, July–August 2020, and October–December 2023.

Another recent corpus is the Ukrainian parliamentary corpus ParlaMint-UA (Kopp et al., 2023), which contains plenary proceedings of the Rada and covers the period from May 2002 to November 2023. It is available in two versions: a collection of plain texts with TSV metadata of the plenary speeches and the collection of plenary speeches with added automatic linguistic annotations. ParlaMint-UA 4.0.1 has more than 51 million tokens, 41 million words, 3.4 million sentences, and 429 thousand statements from 2532 speakers in 1723 meetings.

This overview is not complete, e.g. Shvedova (2020) and Chaplynskyi (2023) also describe Ukrainian corpora not mentioned here. Although a number of research teams are currently working on the automated and manual creation and annotation of different Ukrainian language corpora, some aspects still require additional data and further enhancements. Currently, there is a need for data from contemporary sources such as news, which reflect the ongoing processes in the society and the current linguistic developments. Even though other corpus projects incorporate news data, e.g. GRAC, ukTenTen, UberText, Zvidusil and other corpora mentioned above, their texts are from various sources or time periods, and they are often limited due to copyright issues. Considering the current context, the creation of the CNC-UA is timely. Firstly, the CNC-UA was established in 2023 and covers news data from November 2019 to December 2022 with the potential of expansion. This period covers two significant events, not only in Ukraine but also in Europe and the world: the coronavirus epidemic and Russia's full-scale invasion of Ukraine. Secondly, the CNC-UA is based on news texts of SUSPILNE. This media platform presents international, national, and regional news on a wide range of topics, i.e. world, culture, sports, economy, politics, nature, etc. SUSPILNE is one of few independent media companies in Ukraine, which (in contrast to predominantly private media platforms) has had its unique role as a state-owned and authoritative representative of Ukrainian media.

The CNC-UA fills a gap by providing a middle ground between the large corpora of Ukrainian, e.g. GRAC or UberText 2.0, and smaller, hand-crafted corpora such as BRUK. Furthermore, it is based on official data from a single source and not based on web-scraping. It can be used for training and fine-tuning models for the Ukrainian language as well as sociopolitical, historical and linguistic studies.

3. Corpus Building and Annotation

3.1. Origin and Content of Texts

The first publicly available release of the CNC-UA covers three full years, namely from the end of 2019 until 2022. The corpus is based on raw data in SQL format received from SUSPILNE in December 2022, which forms the basis of its news[6] website. The contents of other media channels, i.e. Facebook, Telegram, YouTube, that also belong to SUSPILNE, are not represented in the corpus.

The raw texts were not labelled with topics, although the website of SUSPILNE uses an extensive tagging system for topics (e.g. crimes of Russian Federation, corruption, weapons, Crimea, Ukraine-EU, Ukraine and NATO) as well as categories (e.g. politics, economics, world, regions, people, technologies, nature, culture, sports). Due to the absence of the original topical annotation in the raw data, a model of eight topics was trained on the lemmatized texts (see Section 3.3). Interestingly, a small number of 34 English texts was identified with the fastText library (Joulin et al., 2016a,b).

3.2. Statistics

The CNC-UA contains 87 210 364 tokens in 292 955 texts in this first release. The breakdown of the amount of texts and tokens over time is shown in Table 1. The number of texts increases each year. Taking into consideration that the current version of the corpus contains texts from November 2019 to December 2022, the statistics do not represent the whole year of 2019. Nevertheless, they show that the number of accessible texts and tokens grows steadily.

Year	# Texts	# Tokens
2019	6887	1 813 880
2020	81 157	21 997 108
2021	95 974	30 275 296
2022	108 937	33 124 080

Table 1: Size of the CNC-UA over time.

3.3. Metadata and Annotation

At the initial stage, it was established that the received data contained the following per-text information: id, title, body, timestamp. The CNC-UA was then enriched with the following information for each text: hour, month, weekday, year, year_month. Linguistic annotations were then added using the Stanza NLP tools (v1.4.2; Qi et al., 2020), whose Ukrainian model was trained on data from the Universal Dependencies project (v2.8). For additional

[6] https://suspilne.media

metadata, a topic model of eight topics (administration, crime, culture, everyday, health, international, sports and war; see Table 2) was trained on the lemmatized texts with the MALLET toolkit (McCallum, 2002). For the development of the eight topics over time see Figure 1. These metadata can be useful for exploring the linguistic changes that occurred over time and studying patterns that can be traced by topic (see Section 5 below for an example).

Topic	Keywords
Administration	head, job, hryvnia, council, work голова, робота, гривня, рада, працювати
Crime	police, man, court, criminal, report поліція, чоловік, суд, кримінальний, повідомити
Culture	museum, person, history, job, project музей, людина, історія, робота, проєкт
Everyday	say, child, person, tell, talk казати, дитина, людина, розповісти, говорити
Health	case, person, COVID, coronavirus, hospital випадок, людина, covid, коронавірус, лікарня
International	Russia, president, country, Russian, report росія, президент, країна, російський, повідомляти
Sports	match, team, championship, world, competition матч, команда, чемпіонат, світ, змагання
War	military, Russian, report, shelling, territory військовий, російський, повідомити, обстріл, територія

Table 2: Topic labels and top-5 ranked keywords by topic in CNC-UA.

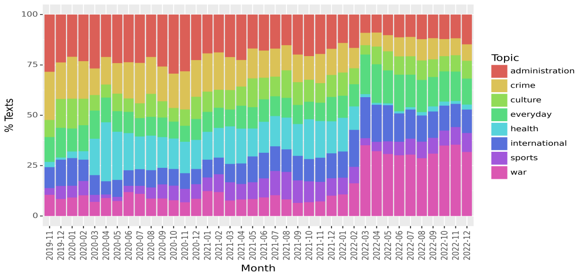


Figure 1: Dominant topics over time in CNC-UA.

Currently, tokens are annotated with the following linguistic information, which is provided by the Stanza NLP tools: word form, lemma, part-of-speech tags (universal and language-specific), morphological features (e.g. animacy, case, gender, number) and dependency information (head and relation type). The language-specific part-of-speech tags are based on the MULTEXT-East Morphosyntactic Specifications, Version 4.

4. Access and Download

The CNC-UA is designed and built according to the FAIR data principles (Wilkinson et al., 2016) and can be accessed from a research data repository specializing in linguistic corpora. It is hosted at the CLARIN-D repository[7] at Saarland University. The corpus is findable by a persistent and globally unique identifier (see Section 10). The CNC-UA is described by rich CMDI (Broeder et al., 2011) metadata with a link to the landing page of the corpus. The metadata are indexed and searchable by the CLARIN Virtual Language Observatory (Van Uytvanck et al., 2010, 2012). The CNC-UA is licensed under a Creative Commons Attribution-NonCommercial-NoDerivs 4.0 International license.

We provide files in several common formats. Besides, there is an option for exploring the corpus through a web-based corpus analysis platform. The CNC-UA can be downloaded in two tab-separated formats: CoNLL-U and VRT. The CoNLL-U format (de Marneffe et al., 2021) contains all linguistic annotations provided by the Stanza NLP tools. In particular, it allows one to work with dependency trees. We also provide the corpus in the vertical text format (VRT), which is of interest to (corpus) linguists as it allows them to encode the corpus on their own CWB (Evert and Hardie, 2011) or CQPweb (Hardie, 2012) servers. For users who do not need their own installation, we also provide a CQPweb server.[8] Lastly, the metadata is available in tabular format.

5. Exploratory Analysis

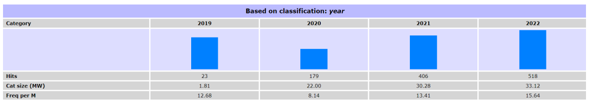
In order to explore the linguistic similarities and differences of the corpus data within different time periods and topics we can use the CQPweb interface. Using queries for concordances, distributional data, frequency lists, and collocations enables us to identify not only the variations in the linguistic contexts but also the various textual patterns within the existing corpus.

To demonstrate the potential of the corpus on the lexical level, we have chosen the concept of democracy. Specifically, we look at the noun демократія (en: democracy, translit: demokratiya), which exemplifies formal stylistically-marked political vocabulary. To analyze the representation of democracy, the query "[lemma="демократія"]" is used, which returns 1126 matches in 861 different texts. The distribution of hits for this query based on classification by *year* (Figure 2) demonstrates the fluctuation of occurrences over the 38 months period. The distribution of hits for this query based on classification by *topic* (Figure 3) shows that the majority of occurrences are within the topic *International*, followed by *Culture* and *Administration*.

The collocation analysis of the query "[lemma="демократія"]" (collocation window "1 to the left" and "1 to the right", frequency at least 5) demonstrates that the noun демократія can be immediately linked to most open-class parts of speech, with adjectives having the highest mutual

[7] https://fedora.clarin-d.uni-saarland.de/

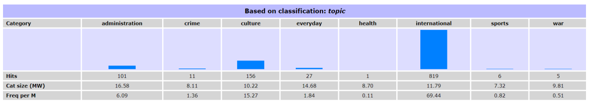
[8] http://corpora.clarin-d.uni-saarland.de/cqpweb/



Based on classification: *year*

Category	2019	2020	2021	2022
Hits	23	179	406	518
Cat size (MW)	1.81	22.00	30.28	33.12
Freq per M	12.68	8.14	13.41	15.64

Figure 2: CQPweb: Distribution of hits for “[lemma="демократія"]” classified by year.



Based on classification: *topic*

Category	administration	crime	culture	everyday	health	international	sports	war
Hits	101	11	156	27	1	819	6	5
Cat size (MW)	16.58	8.11	10.22	14.68	8.70	11.79	7.32	9.81
Freq per M	6.09	1.36	15.27	1.84	0.11	69.44	0.82	0.51

Figure 3: CQPweb: Distribution of hits for “[lemma="демократія"]” classified by topic.

information (MI) score. Ліберальний (en: liberal, translit: liberalnyi), having an MI score of 11.2, represents the strongest first-order collocate for the analyzed lemma. The noun with the highest MI score for “[lemma="демократія"]” is взірець (en: role model, translit: vzirets) with an MI score of 9.3. The only verb with a relatively high MI score, namely 5.0, is the verb захищати (en: protect, translit: zakhyshchaty). Further collocation analyses of “[lemma="демократія"]” might provide additional insights into the conceptualisation of демократія in Ukrainian news.

In this section we have demonstrated an example of using the CNC-UA with the CQPweb interface, which allows to extract the quantitative data necessary for distributional and collocation analyses. The concordances, distributional data, and frequency lists for the query “[lemma="демократія"]” show that the ongoing progress of democracy in Ukraine is reflected in the state-owned media.

6. Discussion and Future Work

The current size of the CNC-UA and the results of our first analyses are already promising. However, we acknowledge that the corpus in its current first release has certain limitations, which need to be taken into account and addressed in future work. In comparison to many other corpora, the corpus is not balanced, which is by design.

First of all, the time span and the size of the CNC-UA could be expanded. The current version contains the materials officially received from SUSPILNE in 2022 at the initial stage of our cooperation. The dataset covers the period starting from 2019 when the new official orthographic rules for Ukrainian were introduced thus reflecting the most recent changes in the Ukrainian language. The dataset goes up to 2022, encompassing a total number of 38 months, which is quite substantial for a news corpus based on a single source. Nevertheless, adding more recent data from 2023 and later to the CNC-UA will significantly increase its value and provide the most up-to-date news dataset for contemporary linguistic analysis and interdisciplinary studies.

In contrast to (smaller) manually annotated corpora, the processing of CNC-UA depends on the availability of external NLP tools for Ukrainian. While the performance of the Stanza pipeline was evaluated[9] on Universal Dependencies (UD) treebanks, an additional evaluation on the corpus is worthwhile. During our work with the corpus, no major problems were found. However, the lemmatization of non-Ukrainian proper names left room for improvement in some cases, e.g. *Scholz*. Also, more metadata at the text level would be desirable.

Lastly, our preliminary experiments using the CNC-UA have raised the issue that the texts contain links to related articles. As a result, the titles of articles are repeated throughout the corpus, which occasionally falsely raises the number of word occurrences and the distribution of search hits. This issue, i.e. boilerplate detection, may require additional cleaning or filtering of the dataset and should be addressed in our future work to increase the accuracy of results.

7. Summary and Conclusions

In this paper we have presented a new corpus of modern Ukrainian news texts. The first publicly available release covers the period from 2019 to 2022. We have placed the CNC-UA in the landscape of existing corpora of Ukrainian in order to demonstrate that it fills a gap by providing a middle ground between the existing large corpora of Ukrainian and the smaller, hand-crafted corpora. The corpus provides metadata at the text level and has been linguistically annotated at the token level with a dependency parser. The current release of the CNC-UA is open and available for download in several common formats. Besides that, we provide an option for exploring the corpus through a web-based corpus analysis platform. As the CNC-UA contains news texts documenting recent events, it is highly relevant for linguistic analysis, as well as sociopolitical, cultural, and interdisciplinary research.

8. Acknowledgements

This research is funded by Deutsche Forschungsgemeinschaft (DFG) – project IDs 232722074 (SFB 1102) and 460033370 (NFDI Text+).

[9] https://stanfordnlp.github.io/stanza/performance.html

9. Bibliographical References

Daan Broeder, Oliver Schonefeld, Thorsten Trippel, Dieter Van Uytvanck, and Andreas Witt. 2011. A pragmatic approach to XML interoperability – the Component Metadata Infrastructure (CMDI). In *Proceedings of Balisage: The Markup Conference 2011*, volume 7 of *Balisage Series on Markup Technologies*, Montréal, Canada.

Dmytro Chaplynskyi. 2023. Introducing UberText 2.0: A corpus of Modern Ukrainian at scale. In *Proceedings of the Second Ukrainian Natural Language Processing Workshop (UNLP)*, pages 1–10, Dubrovnik, Croatia. Association for Computational Linguistics.

Nataliia Darchuk. 2017. Mozhlyvosti semantychnoyi rozmitky korpusu ukrainskoyi movy (KUM). *Naukovyi chasopys Natsionalnoho pedahohichnoho universytetu im. M.P. Drahomanova*, 9(15):18–28.

Marie-Catherine de Marneffe, Christopher D. Manning, Joakim Nivre, and Daniel Zeman. 2021. Universal Dependencies. *Computational Linguistics*, 47(2):255–308.

Department of General and Applied Linguistics and Slavic Philology, Vasyl Stus Donetsk National University. 2023. Korpusy tekstiv ukrainskoi movy. `http://corpora.donnu.edu.ua/`. Accessed: 2024-04-02.

Stefan Evert and Andrew Hardie. 2011. Twenty-first century corpus workbench: Updating a query architecture for the new millennium. In *Proceedings of the Corpus Linguistics 2011 Conference*. University of Birmingham, UK.

Dirk Goldhahn, Thomas Eckart, and Uwe Quasthoff. 2012. Building large monolingual dictionaries at the Leipzig Corpora Collection: From 100 to 200 languages. In *Proceedings of the 8th International Language Resources and Evaluation (LREC'12)*, pages 759–765, Istanbul.

Andrew Hardie. 2012. CQPweb — combining power, flexibility and usability in a corpus analysis tool. *International Journal of Corpus Linguistics*, 17(3):380–409.

Institute for Ukrainian. 2018. Zvidusil. `https://mova.institute/bonito/run.cgi/corp_info?corpname=zvidusil`. Accessed: 2024-04-02.

Miloš Jakubíček, Adam Kilgarriff, Vojtěch Kovář, Pavel Rychlý, and Vít Suchomel. 2013. The TenTen corpus family. In *Proceedings of the 7th international Corpus Linguistics conference (CL2013)*, pages 125–127, Lancaster, UK. Lancaster University: UCREL.

Armand Joulin, Edouard Grave, Piotr Bojanowski, Matthijs Douze, Hérve Jégou, and Tomas Mikolov. 2016a. Fasttext.zip: Compressing text classification models. *arXiv preprint arXiv:1612.03651*.

Armand Joulin, Edouard Grave, Piotr Bojanowski, and Tomas Mikolov. 2016b. Bag of tricks for efficient text classification. *arXiv preprint arXiv:1607.01759*.

Yevheniia Karpilovska. 2007. Tendentsii rozvytku suchasnoho ukrainskoho leksykonu: chynnyky stabilizatsii innovatsii. *Ukrainska mova*, 4:3–15.

Natalia Kotsyba, Bohdan Moskalevskyi, and Mykhailo Romanenko. 2022. Gold standard Universal Dependencies corpus for Ukrainian. `https://github.com/UniversalDependencies/UD_Ukrainian-IU`. Accessed: 2024-04-02.

Leipzig Corpora Collection. 2014. Ukrainian mixed corpus based on material from 2014. `https://corpora.uni-leipzig.de?corpusId=ukr_mixed_2014`. Accessed: 2024-04-02.

Andrew Kachites McCallum. 2002. MALLET: A machine learning for language toolkit. `https://mimno.github.io/Mallet/`. Accessed: 2024-04-02.

Peng Qi, Yuhao Zhang, Yuhui Zhang, Jason Bolton, and Christopher D. Manning. 2020. Stanza: A Python natural language processing toolkit for many human languages. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics: System Demonstrations*, pages 101–108, Online. Association for Computational Linguistics.

Maria Shvedova. 2020. The General Regionally Annotated Corpus of Ukrainian (GRAC, uacorpus.org): Architecture and functionality. In *Proceedings of the 4th International Conference on Computational Linguistics and Intelligent Systems (COLINS 2020)*, CEUR Workshop Proceedings, pages 489–506, Lviv, Ukraine. CEUR-WS.org.

Volodymyr Shyrokov. 2011. *Ukrainska leksykohrafiia v zahalnoslovianskomu konteksti: teoriia, praktyka, typolohiia*, chapter Zastosuvannia Ukrainskoho natsionalnoho linhvistychnoho korpusu v leksykohrafii ta linhvistychnykh ekspertyzakh. Vydavnychyi dim Dmytra Buraho.

Vasyl Starko and Andriy Rysin. 2023. Creating a POS gold standard corpus of Modern Ukrainian. In *Proceedings of the Second Ukrainian Natural*

Language Processing Workshop (UNLP), pages 91–95, Dubrovnik, Croatia. Association for Computational Linguistics.

Vít Suchomel. 2020. *Better Web Corpora For Corpus Linguistics And NLP*. Ph.D. thesis, Masaryk University, Faculty of Informatics, Brno, Czech Republic.

Vít Suchomel and Jan Pomikálek. 2012. Efficient web crawling for large text corpora. In *Proceedings of the seventh Web as Corpus Workshop (WAC7)*, pages 39–43.

Dieter Van Uytvanck, Herman Stehouwer, and Lari Lampen. 2012. Semantic metadata mapping in practice: the Virtual Language Observatory. In *Proceedings of the Eighth International Conference on Language Resources and Evaluation (LREC'12)*, pages 1029–1034, Istanbul, Turkey. European Language Resources Association (ELRA).

Dieter Van Uytvanck, Claus Zinn, Daan Broeder, Peter Wittenburg, and Mariano Gardellini. 2010. Virtual Language Observatory: The portal to the language resources and technology universe. In *Proceedings of the Seventh International Conference on Language Resources and Evaluation (LREC'10)*, pages 900–903, Valletta, Malta. European Language Resources Association (ELRA).

Mark D. Wilkinson, Michel Dumontier, IJsbrand Jan Aalbersberg, Gabrielle Appleton, Myles Axton, Arie Baak, Niklas Blomberg, Jan-Willem Boiten, Luiz Bonino da Silva Santos, Philip E. Bourne, Jildau Bouwman, Anthony J. Brookes, Tim Clark, Mercè Crosas, Ingrid Dillo, Olivier Dumon, Scott Edmunds, Chris T. Evelo, Richard Finkers, Alejandra Gonzalez-Beltran, Alasdair J.G. Gray, Paul Groth, Carole Goble, Jeffrey S. Grethe, Jaap Heringa, Peter A.C 't Hoen, Rob Hooft, Tobias Kuhn, Ruben Kok, Joost Kok, Scott J. Lusher, Maryann E. Martone, Albert Mons, Abel L. Packer, Bengt Persson, Philippe Rocca-Serra, Marco Roos, Rene van Schaik, Susanna-Assunta Sansone, Erik Schultes, Thierry Sengstag, Ted Slater, George Strawn, Morris A. Swertz, Mark Thompson, Johan van der Lei, Erik van Mulligen, Jan Velterop, Andra Waagmeester, Peter Wittenburg, Katherine Wolstencroft, Jun Zhao, and Barend Mons. 2016. The FAIR Guiding Principles for scientific data management and stewardship. *Scientific Data*, 3:160018.

10. Language Resource References

Fischer, Stefan and Haidarzhyi, Kateryna and Knappen, Jörg and Stodolinska, Yuliya and Teich, Elke. 2023. *Contemporary News Corpus for Ukrainian (CNC-UA)*. CLARIND-UdS. PID http://hdl.handle.net/21.11119/0000-000E-1C5C-D.

Kopp, Matyáš and Kryvenko, Anna and Rii, Andriana. 2023. *Ukrainian parliamentary corpus ParlaMint-UA 4.0.1*. Slovenian language resource repository CLARIN.SI. PID http://hdl.handle.net/11356/1900.

Tracey, Jennifer and Strassel, Stephanie and Graff, David and Wright, Jonathan and Chen, Song and Ryant, Neville and Ma, Xiaoyi and Kulick, Seth and Delgado, Dana and Arrigo, Michael. 2020. *LORELEI Ukrainian Representative Language Pack*. Linguistic Data Consortium, ISLRN 551-143-444-242-2.

Introducing the Djinni Recruitment Dataset: A Corpus of Anonymized CVs and Job Postings

Nazarii Drushchak, Mariana Romanyshyn
Ukrainian Catholic University, Grammarly
Lviv, Ukraine, Kyiv, Ukraine
drushchak.pn@ucu.edu.ua, mariana.romanyshyn@grammarly.com

Abstract

This paper introduces the Djinni Recruitment Dataset, a large-scale open-source corpus of candidate profiles and job descriptions. With over 150,000 jobs and 230,000 candidates, the dataset includes samples in English and Ukrainian, thereby facilitating advancements in the recruitment domain of natural language processing (NLP) for both languages. It is one of the first open-source corpora in the recruitment domain, opening up new opportunities for AI-driven recruitment technologies and related fields. Notably, the dataset is accessible under the MIT license, encouraging widespread adoption for both scientific research and commercial projects.

Keywords: Recruitment Dataset, Open-Source Corpus, Natural Language Processing (NLP)

1. Introduction

This paper introduces the Djinni Recruitment Dataset[1], a unique asset to NLP research in the recruitment domain, where open data is exceptionally limited. The corpus addresses the need for diverse publicly available datasets, which are particularly important in the age of transformers and large language models, especially for low-resource languages such as Ukrainian.

The data for the corpus was provided by Djinni[2], an IT job platform that hosts job listings and anonymized user profiles similar to resumes. Djinni's database is distinguished by its bilingual nature, encompassing both Ukrainian and English languages. The company generously shared with us the data covering a period from 2020 to 2023.

The Djinni Recruitment Dataset opens avenues for various research opportunities. Based on this data, we can analyze the impact of global events on hiring trends and develop recommendation systems tailored to the recruitment domain. The dataset also holds promise for addressing ethical concerns in hiring systems. A corpus of anonymized candidate profiles used for training may help increase fairness in tools like Amazon's AI recruiting tool (Dastin, 2018), which was trained on predominantly male CVs and subsequently exemplified gender bias. The dataset will help promote Responsible AI practices and contribute to the broader discourse on improving the recruitment process.

In this paper, we describe the Djinni Recruitment Dataset and its application. Section 2 reviews related work. Section 3 presents a thorough dataset overview, including source, collection, preprocessing, and characteristics. Section 4 identifies recoverable protected attributes from anonymous CVs. Section 5 discusses the intended use of the dataset in industry and academia. Section 6 addresses the challenges and limitations of the Djinni Recruitment Dataset. Section 7 summarizes the findings and suggests future research directions. Section 8 considers ethical aspects, focusing on privacy and anonymization.

2. Related Work

The exploration of linguistic resources for the Ukrainian language and job-related datasets reveals a scarcity in large-scale, freely accessible datasets that meet comprehensive research needs.

The most notable, publicly available resources in Ukrainian include:

1. BRUK (Starko and Rysin, 2023), a corpus of 450,000 words, whose genre distribution mirrors that of the original Brown corpus[3], covering fiction, religious texts, press, legal documents, etc.;

2. UA-GEC (Syvokon et al., 2023), a corpus of 500,000 words, which contains texts with errors and their corrections from a wide variety of writing domains, from text chats and essays to formal writing;

3. UberText 2.0 (Chaplynskyi, 2023), which consists of 8.59 million texts of news, fiction, social media posts, Wikipedia, and court decisions;

[1] https://github.com/Stereotypes-in-LLMs/recruitment-dataset

[2] https://djinni.co/

[3] http://korpus.uib.no/icame/manuals/brown/

25 May, 2024. © 2024 ELRA Language Resource Association: CC BY-NC 4.0

4. Malyuk[4], a corpus of 38.94 million texts, which is a compilation of UberText 2.0, Oscar[5] (derived from Common Crawl), and Ukrainian News[6];

5. UD Ukrainian[7], a gold standard Universal Dependencies corpus for Ukrainian, which comprises 7,000 sentences of fiction, news, opinion articles, Wikipedia, legal documents, letters, posts, and comments.

Despite the genre diversity present in the publicly available corpora for Ukrainian, none of them include texts from the recruitment domain.

In our search for open-source job-related datasets, we identified relevant corpora for the English language, but they focus on either job descriptions[8] or candidate CVs[9], without offering a unified set that would cater to both aspects. This disjointed approach inhibits the capability to perform semantic matching, thereby constraining the development of automated job recommender and AI-assisted hiring systems.

The corporate landscape of open-source datasets is similarly fragmented: platforms like Indeed[10] provide separate datasets for CVs[11] and job descriptions[12]. Structural and temporal differences in these datasets challenge the development of NLP models for effective job-candidate matching. This situation emphasizes the need for more collaborative efforts between academia and industry to foster the creation of open, integrated datasets.

3. Dataset Description

In this section, we'll delve into the Djinni Recruitment Dataset, detailing its structure and processing and offering key insights into notable features.

3.1. Data Source

The data in the corpus originates from Djinni, Ukraine's leading tech job marketplace, boasting over 50,000 monthly users. Djinni generously provided open-source access to two significant data groups: anonymous candidate information and job descriptions, primarily from the IT sector in Ukraine. This wealth of data serves as a valuable resource for understanding trends and patterns in the Ukrainian tech job market. In the pursuit of accurate analysis, we conducted additional preprocessing of this data, a topic we explore further in the next section.

3.2. Data Processing

The dataset underwent several critical preprocessing steps, including language filtering, the filtering of duplicates and outliers, language-based split, and the removal of personally identifiable information.

3.2.1. Data Filtering

We used the langdetect[13] model from the transformers library[14] to detect and select data samples exclusively in English and Ukrainian languages, ensuring the dataset's relevance to the primary language groups within the Ukrainian IT sector.

To improve the dataset's diversity and balance, we undertook a deduplication effort, focusing on removing both exact duplicates and highly similar samples. We employed embedding models to identify similar CVs and job descriptions, selecting models based on their quality for each language at the time of filtering. Specifically, for English texts, we used the bge-base-en-v1.5 model[15] with an empirically determined cosine similarity threshold of 0.9. For Ukrainian texts, we chose the multilingual-e5-large model[16] with the threshold of 0.95. This approach ensured that the dataset comprised only high-quality, unique entries.

Moreover, we implemented an outlier removal step, filtering out entries below the 5th percentile in text length to exclude extremely short texts. This refinement enhances the dataset's relevance.

We monitored the impact of the filtering on the size of the dataset at each filtering stage. Table 1 shows that candidate CVs experienced a modest reduction of 20%, whereas job descriptions saw

[4] https://huggingface.co/datasets/lang-uk/malyuk

[5] https://huggingface.co/datasets/oscar

[6] https://huggingface.co/datasets/zeusfsx/ukrainian-news

[7] https://github.com/UniversalDependencies/UD_Ukrainian-IU/tree/master

[8] https://www.kaggle.com/datasets/ravindrasinghrana/job-description-dataset, https://www.kaggle.com/datasets/arshkon/linkedin-job-postings/data

[9] https://www.kaggle.com/datasets/snehaanbhawal/resume-dataset

[10] https://www.indeed.com

[11] https://datastock.shop/download-indeed-job-resume-dataset/

[12] https://data.world/promptcloud/indeed-job-posting-dataset

[13] https://huggingface.co/ERCDiDip/langdetect

[14] https://huggingface.co/docs/transformers/en/index

[15] https://huggingface.co/BAAI/bge-base-en-v1.5

[16] https://huggingface.co/intfloat/multilingual-e5-large

	CVs	Jobs
Raw samples	294,678	443,458
After basic filtering	241,561	358,491
After similarity filtering	234,480	169,358

Table 1: The number of samples in the dataset before and after filtering. Basic filtering includes language filtering and the removal of outliers and identical duplicates. Similarity filtering covers the removal of near-identical samples.

a more substantial decrease of 60%. The cause of this contrast lies in the highly repetitive nature of job descriptions posted by the same companies in different periods, which we verified via a closer data analysis.

3.2.2. Language-Based Split

We split the dataset into two based on the detected language, forming separate divisions for English and Ukrainian sections within both job descriptions and CVs. This strategic division enables more nuanced analysis and application of NLP techniques tailored to language specifics, significantly enhancing the relevance of insights derived from the dataset for bilingual environments. This step also revealed a serious imbalance of language representation: Ukrainian-language CVs constitute only 10% of all CVs, and Ukrainian-language job postings constitute 16% of all job postings. The exact numbers can be found in Table 2.

	CVs	Jobs
English	210,250	141,897
Ukrainian	24,230	27,461

Table 2: The number of CVs and job descriptions in the English and Ukrainian segments of the dataset post language-based splitting.

3.2.3. Removal of Personally Identifiable Information

Djinni has a strict policy requiring registration through anonymized profiles only and enforces measures to prevent the posting of personally identifiable information (PII). This approach to anonymity ensures the protection of sensitive personal data and reduces bias during resume screening by potential employers.

To verify the anonymity and confidentiality of the dataset, we developed a script[17] utilizing regex implementation tailored for both English and Ukrainian languages. The script is based on patterns and keywords in both languages, covering phone numbers, email addresses, physical addresses, social media links, taxpayer identification numbers, and other unique identifiers. This step was pivotal in detecting remnants of PII within CVs.

The identified CVs with PII were meticulously removed from the dataset to uphold the highest standards of privacy and data protection. Less than 0.2% of the CVs contained PII data.

For further details on the attributes of the CV and job description datasets, see Appendix A: Feature Explanation.

4. Protected Attributes in the Dataset

Our research further focused on identifying protected attributes within the anonymized CVs to determine the true level of anonymity in the provided data, as well as to pinpoint potential sources of bias in recruitment practices. Following the Principles of Preventing and Combating Discrimination[18] in Ukraine, we identified core protected attributes for our study: gender, age, marital status, military status, religion, and person name.

Our analysis primarily focused on identifying explicit mentions of protected attributes in CVs across both English and Ukrainian languages. We developed a script that uses regular expressions and dictionaries to detect terms and patterns related to specific protected attributes[19]. To detect person names, we used the VESUM[20] dictionary, which contains more than 5 thousand names in Ukrainian, and translitua[21] to transliterate Ukrainian names and enable search in the English segment. We manually crafted parallel dictionaries in both Ukrainian and English for other protected attributes: 22 gender groups, ages from 16 to 65 years, 5 marital statuses, 5 military statuses, and 9 religious groups. The script can be used to improve data anonymity and increase fairness in automated hiring processes.

4.1. Experimental Findings

The quantitative insights into the explicit representation of protected attributes within the dataset, categorized by language, are presented in Table 3.

[17] https://github.com/Stereotypes-in-LLMs/recruitment-dataset/blob/main/notebooks/EDA/PII_CV_analyses.ipynb

[18] https://zakon.rada.gov.ua/laws/show/5207-17

[19] https://github.com/Stereotypes-in-LLMs/recruitment-dataset/blob/main/notebooks/EDA/EDA_candidates.ipynb

[20] https://github.com/brown-uk/dict_uk

[21] https://pypi.org/project/translitua/

Protected Group	Ukr CVs (%)	Eng CVs (%)
Age	**0.21**	0.15
Gender	**0.66**	0.05
Marital Status	**0.07**	0.02
Military Status	**0.42**	0.26
Name	3.75	**3.85**
Religion	0.02	**0.2**

Table 3: The fractions of CVs that contain explicit mentions of protected attributes.

This analysis reveals significant differences between Ukrainian and English CVs. Particularly, explicit mentions of gender are substantially more frequent in Ukrainian CVs, while mentions of religion are much more common in English CVs. The results show that beyond PII, certain characteristics may introduce bias, necessitating their anonymization for the further use of the dataset.

4.2. Gender-Marked Verbs in Ukrainian CVs

Unlike English, Ukrainian is a synthetic language, whose verbs are inflected for the grammatical gender when used in the past tense. This means that an anonymous CV that uses gender-marked verbs may reveal the gender of the author.

To analyze the impact of this linguistic phenomenon, we developed a script[22], which uses the pymorphy3[23] and stanza[24] Python libraries to analyze texts. In each Ukrainian CV, we then identified gender-marked verbs, which related to the subject "I" or had no subject, and checked which grammatical gender prevailed, subsequently classifying those CVs as revealing the author's gender.

The proposed metric allowed us to detect 16.55% of Ukrainian CVs that may have been written by candidates who identify as female and 30.50% by candidates who identify as male. This analysis highlights the nuanced ways gender perspectives may be integrated into job-related documents and emphasizes the need for more elaborate strategies for detecting protected attributes. We leave this for future work.

5. Intended Use

The Djinni Recruitment Dataset can be leveraged for the purposes outlined below:

1. for the development of recommender systems and advanced semantic search;
2. as potential training data for both English and Ukrainian domain-specific LLMs, based on GPT-3 (Brown et al., 2020), Llama 2 (Touvron et al., 2023), Mistral 7B (Jiang et al., 2023), Palm 2 (Anil et al., 2023), etc., enriching their understanding and generating capabilities within specialized recruitment contexts;
3. as a benchmark or training set to promote fairness in AI-assisted hiring, addressing bias and ensuring equitable selection processes;
4. for automated resume and job description creation;
5. for market analysis and evaluation of the tech sector's dynamics in Ukraine;
6. for topic discovery and trend analysis within the tech industry through modeling and classification;
7. for automated identification of company domains, assisting in strategic market planning.

6. Challenges and Limitations

We acknowledge the following challenges and limitations of the Djinni Recruitment Dataset:

1. **Limited languages:** The dataset is available in only two languages—Ukrainian and English.
2. **Unlabelled data:** The lack of labeled data makes it challenging to determine who was hired and to conduct specific analyses related to successful job placements.
3. **Lack of CV publication date:** The dataset does not include any information on when the CVs were published.
4. **Noisy user-generated data:** The dataset includes user-generated content, introducing noise and variability that may impact the accuracy of certain analyses.
5. **Focus on the tech domain:** The dataset is primarily centered around the tech domain, limiting its applicability to other industries or sectors.
6. **Ukrainian market only:** The dataset exclusively represents the Ukrainian market, which may restrict broader generalizations or comparisons with job markets in other regions.

Understanding these challenges is crucial for the appropriate interpretation and utilization of the dataset in a way that aligns with its inherent limitations.

[22] https://github.com/Stereotypes-in-LLMs/recruitment-dataset/blob/main/notebooks/EDA/EDA_candidates.ipynb

[23] https://pypi.org/project/pymorphy3/

[24] https://stanfordnlp.github.io/stanza/

7. Conclusion

In this paper, we introduced the Djinni Recruitment Dataset, a pioneering resource in NLP and recruitment data analysis, with a focus on the Ukrainian IT sector, which contains data in the Ukrainian and English languages. The dataset is released under the MIT license, which allows for academic and commercial use.

This dataset's focus on recruitment is key for creating NLP tools for job matching, market analysis, bias identification, and fostering Responsible AI in hiring. Its bilingual content represents the tech sector of Ukraine, largely influenced by the global IT job market.

One of the most significant contributions of the Djinni Recruitment Dataset is that it sets a precedent for other businesses to consider the value of making their data openly available for research purposes.

Future research may expand the dataset's languages and industries. There's potential for creating targeted NLP tools to improve recommendation systems and algorithms for bias detection and mitigation in the recruitment domain.

8. Ethical Considerations

The Djinni Recruitment Dataset adheres to the conditions of fair use. The contributors of data have the privilege to ask for their information to be deleted by contacting the authors of this paper.

The Djinni dataset upholds standards of data anonymization and privacy protection. These measures are implemented to prevent any potential harm to the authors of the data. By prioritizing anonymity, we strive to safeguard the privacy of those who have contributed to this valuable resource.

The dataset is published with the description of intended use, which underscores our commitment to responsible data stewardship.

We used ChatGPT and Grammarly to assist with paraphrasing while writing this paper.

9. Acknowledgments

We express our gratitude to Djinni for sharing the data. Djinni's dedication to advancing knowledge and fostering progress in the AI-driven recruitment field has enabled the creation of this dataset.

10. References

Rohan Anil, Andrew M. Dai, Orhan Firat, Melvin Johnson, Dmitry Lepikhin, Alexandre Passos, Siamak Shakeri, Emanuel Taropa, Paige Bailey, Zhifeng Chen, Eric Chu, Jonathan H. Clark, Laurent El Shafey, Yanping Huang, Kathy Meier-Hellstern, Steven Zheng, Ce Zheng, Weikang Zhou, Denny Zhou, Slav Petrov, and Yonghui Wu. 2023. Palm 2 technical report.

Tom B. Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel M. Ziegler, Jeffrey Wu, Clemens Winter, Christopher Hesse, Mark Chen, Eric Sigler, Mateusz Litwin, Scott Gray, Benjamin Chess, Jack Clark, Christopher Berner, Sam McCandlish, Alec Radford, Ilya Sutskever, and Dario Amodei. 2020. Language models are few-shot learners.

Dmytro Chaplynskyi. 2023. Introducing UberText 2.0: A corpus of Modern Ukrainian at scale. In *Proceedings of the Second Ukrainian Natural Language Processing Workshop (UNLP)*, pages 1–10, Dubrovnik, Croatia. Association for Computational Linguistics.

Jeffrey Dastin. 2018. Insight - amazon scraps secret ai recruiting tool that showed bias against women.

Albert Q. Jiang, Alexandre Sablayrolles, Arthur Mensch, Chris Bamford, Devendra Singh Chaplot, Diego de las Casas, Florian Bressand, Gianna Lengyel, Guillaume Lample, Lucile Saulnier, Lélio Renard Lavaud, Marie-Anne Lachaux, Pierre Stock, Teven Le Scao, Thibaut Lavril, Thomas Wang, Timothée Lacroix, and William El Sayed. 2023. Mistral 7b.

Vasyl Starko and Andriy Rysin. 2023. Creating a POS gold standard corpus of Modern Ukrainian. In *Proceedings of the Second Ukrainian Natural Language Processing Workshop (UNLP)*, pages 91–95, Dubrovnik, Croatia. Association for Computational Linguistics.

Oleksiy Syvokon, Olena Nahorna, Pavlo Kuchmiichuk, and Nastasiia Osidach. 2023. UA-GEC: Grammatical error correction and fluency corpus for the Ukrainian language. pages 96–102.

Hugo Touvron, Louis Martin, Kevin Stone, Peter Albert, Amjad Almahairi, Yasmine Babaei, Nikolay Bashlykov, Soumya Batra, Prajjwal Bhargava, Shruti Bhosale, Dan Bikel, Lukas Blecher, Cristian Canton Ferrer, Moya Chen, and Guillem Cucurull. 2023. Llama 2: Open foundation and fine-tuned chat models.

A. Feature Explanation

A.1. Job Descriptions

Both English and Ukrainian parts of the dataset contain attributes related to job descriptions, including position titles, job descriptions, company names, experience requirements, keywords, English proficiency levels, publication dates, language of job descriptions, and unique identifiers.

Features:

- **id:** 169,358 unique synthetic identifiers for each job description.
- **Position:** 82,423 unique manually written position titles.
- **Long Description:** 169,358 unique manually written job descriptions.
- **Company Name:** 12,897 unique company names.
- **Exp Years:** 5 unique values for experience years required: ’2y’, ’3y’, ’no_exp’, ’5y’, ’1y’.
- **Primary Keyword:** 46 unique job profile types.
- **English Level:** 6 unique English proficiency levels: ’intermediate’, ’pre’, ’upper’, ’basic’, ’fluent’, NaN.
- **Published:** publication dates (only month and year).
- **Long Description_lang:** 2 unique languages in which job descriptions can be written: ’uk’ (Ukrainian), ’en’ (English).

A.2. CVs

Both English and Ukrainian parts of the dataset contain attributes related to candidate CVs, including position titles, candidate information, candidate highlights, job search preferences, job profile types, English proficiency levels, experience years, concatenated CV text, language of CVs, and unique identifiers.

Features:

- **id:** 234,480 unique synthetic identifiers for each candidate CV.
- **Position:** 58,341 unique manually written position titles.
- **Moreinfo:** 234,365 unique manually written candidate information entries.
- **Looking For:** 109,524 unique manually written job search preferences.
- **Highlights:** 117,700 unique manually written candidate highlights.
- **Primary Keyword:** 42 unique job profile types.
- **English Level:** 7 unique English proficiency levels: ’intermediate’, ’pre’, ’upper’, ’basic’, ’no_english’, ’fluent’, NaN.
- **Experience Years:** 15 unique values representing candidate experience in years.
- **CV:** 234,480 unique concatenated CV texts (Highlights + Moreinfo + Looking For).
- **CV_lang:** 2 unique languages in which CVs can be written: ’uk’ (Ukrainian), ’en’ (English).

Creating Parallel Corpora for Ukrainian: a German-Ukrainian Parallel Corpus (ParaRook||DE-UK)

Maria Shvedova, Arsenii Lukashevskyi
National Technical University "Kharkiv Polytechnic Institute", University of Jena
Kyrpychova 2, 61002, Kharkiv, Ukraine; Ernst-Abbe-Platz 8, 07743, Jena, Germany
Mariia.Shvedova@khpi.edu.ua, Arsenii.Lukashevskyi@sgt.khpi.edu.ua

Abstract

Parallel corpora are currently a popular and vibrantly developing category of linguistic resources, used both in literature and translation studies, as well as in the field of NLP. For Ukrainian, though, there are still not enough significant parallel corpora compiled within a single roof project and made available to the research community. In this paper we present a newly developed resource, the German-Ukrainian Parallel Corpus — ParaRook||DE-UK, searchable online. We describe various issues related to its compilation, text selection, and annotation. The paper also features several examples of how the corpus can be used in linguistic research and translation studies. Using the experience of the German-Ukrainian parallel corpus, parallel corpora for other languages with Ukrainian can be developed.

Keywords: parallel corpus, corpus annotation, Ukrainian, German, translation

1. Parallel Corpora for Ukrainian

Parallel corpora are a valuable linguistic resource that is applied primarily for translation research and practice as well as comparative linguistic studies, it can also be useful for monolingual studies. With the development of computer technologies, the role of parallel corpora as datasets for machine translation is becoming increasingly important. Though datasets of parallel sentences in different languages are often collected automatically from the Internet, it is still useful to create some parallel corpora semi-manually, especially for fiction texts that typically lack exact (word-for-word) match between the original and the translation, which makes it difficult to automatically collect and align them. Here are several references to the books on the use of parallel corpora in linguistic studies, translation studies and translation teaching (Anderman and Rogers, 2007; Hansen-Schirra et al., 2012; Enghels et al., 2020; Liu, 2020)

For the Ukrainian language, there are still few parallel corpora available online for searching. One of these projects is the Polish-Ukrainian parallel corpus (Kotsyba, 2016), which has size of about 4 million tokens in the Polish part and is searchable both via an older search manager (Kotsyba and Turska, 2005 - 2011) and on the NoSketchEngine platform on the website of the Laboratory of Ukrainian project. (Kotsyba, 2018) The site also published a parallel English-Ukrainian corpus of 1.5 million tokens in the English part and smaller French-, German-, Spanish-, and Portuguese-Ukrainian bilingual pairs (500, 190, 65 and 16 thousand tokens, respectively) containing literary texts, including some translated from a third language.

The largest collection of semi-manually aligned parallel texts with Ukrainian is now available for search as a part of the InterCorp parallel corpora collection (Čermák and Rosen, 2012). In InterCorp v.16, the volume of Ukrainian texts is over 18 million tokens with aligned originals or translations into Czech and other languages through Czech as a pivot language. The Ukrainian part of InterCorp consists mainly of fiction texts and a smaller dataset featuring subtitles and the Bible. (Čermák and Rosen, 2008 - 2023)

A one-million-tokens dataset of Ukrainian parallel fiction and medical texts with French, English, and Polish is available for download on Natalia Grabar's site (Grabar and Hamon, 2017).

A significant part of the existing Ukrainian parallel corpora is not currently available to the Ukrainian community for various reasons. Access to the Ukrainian-Russian parallel corpus within the Russian National Corpus (Sitchinava et al., 2011) is blocked in Ukraine since 2017 due to the war. ParaSol: a Slavic Parallel Corpus is currently under reconstruction (von Waldenfels, 2011). The following corpora have not been published: Bulgarian-Ukrainian parallel corpus KUB (Siruk and Derzhanski, 2013), Polish-Ukrainian and Ukrainian-Polish parallel corpus of Ivan Franko's self-translations (Buk, 2012), English-Ukrainian parallel corpus ParKUM (Darčuk et al., 2017), English-Ukrainian parallel corpus of Legal Texts (Matvieieva, 2019), English-Ukrainian parallel corpus compiled by Serhij Zasiekin (Zasiekin, 2020), English-Ukrainian parallel corpus of IT texts (Mandziy et al., 2022) etc. Smaller user collections of parallel texts are created by students at various Ukrainian universities, such as Lviv Polytechnic, Odesa National University, Kherson National Technical University, and others, for educational purposes, but there is no

25 May, 2024. © 2024 ELRA Language Resource Association: CC BY-NC 4.0

Original/Style	Fiction	Nonfiction
English (EN)	48,716,969	8,962,108
Russian (RU)	18,265,944	2,413,472
French (FR)	17,844,342	2,462,649
Polish (PL)	10,931,819	1,816,123
German (DE)	9,661,714	2,520,939
Czech (CS)	4,130,289	389,861
Spanish (ES)	3,641,073	450,012
Italian (IT)	3,413,198	598,576
Bulgarian (BG)	2,736,933	109,597

Table 1: The scope of translated texts in GRAC v.17 by original language and style.

coordinated system that would accumulate these materials and make them available for use.

A valuable resource for creating Ukrainian parallel corpora is GRAC (Shvedova, 2017 - 2024): the Ukrainian language reference corpus, which contains translations from 89 languages, mostly fiction, with a total size of 172 million tokens of texts translated from different languages (GRAC v.17). The size of the largest subcorpora of translated texts in GRAC by language and style is shown in the Table 1.

There are many parallel corpus projects where texts are collected and aligned automatically. One such project is ParaCrawl (Bañón et al., 2020), notably its MultiParaCrawl [1] corpus series, which includes 705 bilingual language pairs for 41 languages, including 36 pairs for Ukrainian with different European languages. The languages were identified with Google's Compact Language Detector 2, and neural network technologies were applied for text alignment and cleaning. Specifically, this corpus was prepared for the OPUS (Tiedemann, 2012) project by pivoting text documents through English to achieve a massive parallel corpus. It includes only the new language pairs built by this procedure and can be downloaded in TMX, XML, and Moses formats from the OPUS website. As of March 2024, it includes 34 million sentences in Ukrainian.

The ParaCrawl project itself is focused more on English-centric language pairs and is larger than MultiParaCrawl (compare 1.5 billion sentences of this project and 789 million of MultiParaCrawl). However, it includes 14 million sentences in Ukrainian. It is freely available for download in TMX, TXT, and raw formats. In addition, it is distributed via OPUS.

Another project is NLLb (Schwenk et al., 2019; Fan et al., 2021), a large dataset containing bitext for 148 English-centric and 1465 non-English-centric language pairs. The dataset was created based on metadata for mined bitext released by Meta AI. It was filtered for language identification, emoji-based filtering was performed, and, for some high-resource languages, a language model was applied. The data was processed using the stopes mining library and the LASER3 encoders (Costa-jussà et al., 2022). Currently, it includes 166 million tokens in Ukrainian.

MultiCCAligned (El-Kishky et al., 2020) is a parallel corpus comprising web-document pairs in 137 languages aligned with English. The corpus was created by performing language identification on raw web documents and ensuring that corresponding language codes match the URLs of web documents. More than 100 million aligned documents were paired with English. Some English documents were aligned to multiple documents in different target languages. Sentence pairs were extracted using similarity scores of LASER embeddings from the document pairs. The latest release of MultiCCAligned is v1.1, created from 68 Commoncrawl Snapshots up until March 2020. It includes 62 million sentences in Ukrainian.

MaCoCu (Bañón et al., 2022) is a multilingual parallel corpus built by crawling national internet top-level domains. The corpus was processed using the Bitextor tool, with considerable effort put into cleaning the extracted text. Accordingly, the MaCoCu-uk-en 1.0 was created based on scanned data from sites on the .ua domain and includes 238,841,101 tokens.

Also, an essential source of parallel texts is Wikipedia. One project that has put this into practice is WikiMatrix. The project focuses on languages with low resources, making it a valuable dataset for researchers and developers working with less commonly studied languages. Currently, WikiMatrix provides parallel data for over 1620 language pairs. The authors state that their project makes 135 million parallel sentences available in 96 languages, of which only 34 million are aligned with English. One of the largest pairs is the Ukrainian-Russian one, amounting to 2.5 million sentences (Schwenk et al., 2019). It should be noted that the source of Ukrainian-Russian sentences could be numerous Ukrainian sites with parallel language versions, which tend to use automatic translation.

The Ukrainian language is represented in two multilingual parallel corpora on Sketch Engine, namely OPUS parallel corpus covering 40 languages (the size of Ukrainian texts is 2.5 million tokens) and OpenSubtitles: multilingual corpus in 58 languages (the size of the Ukrainian part is 5 million tokens) (Lison and Tiedemann, 2016).

Only some automatically built parallel corpus projects are listed in this section. It is essential to mention, for example, the OpenSubtitles corpus practices. It differs from other automatic cor-

[1] https://paracrawl.eu/news/item/18-multiparacrawl-9-including-ukrainian

pora by using a time-based approach and intra-language alignment as subtitles in one language often have many variants, which allows for more accurate learning of nuances and variations in the language (Lison and Tiedemann, 2016).

Most of the listed projects are available for download through the OPUS website or in different formats. One is the TXT format ParaCrawl, a bilingual text where sentences are aligned in a one-line per sentence format in 2 columns. For users who need to become more familiar with technology, the Sketch Engine platform will be helpful, featuring both OpenSubtitles and parallel corpora in 40 languages from OPUS in a search interface.

However, it is essential to note that they often lack precision in alignment and data cleanliness, which can impact the quality of the results. For instance, due to the automated nature of the alignment process, there may be instances where sentences or phrases are not accurately matched (Zariņa et al., 2015). Similarly, cleaning may sometimes leave irrelevant or noisy data.

Such projects are useful for purposes such as training machine translation models (Tiedemann and Thottingal, 2020), but are not completely suitable for linguistic research due to their frequent noisiness and lack of accuracy, which is impossible on such large arrays of text. This is why, despite advances in technology and smaller size, manually collected corpora are extremely useful in literature research, translation studies, comparative and typological studies.

In this paper, we present ParaRook||DE-UK (Shvedova and Lukashevskyi, 2023-2024), which is the first large German-Ukrainian corpus collected and verified manually, with detailed meta-annotation and morphosyntactic annotation, and searchable online. The title refers to the Ukrainian monolingual reference corpus GRAC (*grak* is the Ukrainian name for rook) and also sounds like "pair of hands" in Ukrainian.

2. Composition and structure of ParaRook||DE-UK

2.1. Texts

The history of German-Ukrainian literary translation is a complex and interesting field (Ivanytska, 2015). We aimed to show samples of German-Ukrainian translation from different periods, namely Soviet, with specific features of the time, and contemporary.

As shown by M. Ivanytska, German-Ukrainian translations were sometimes made not directly between two languages, but through the mediation of a Russian translation. We tried to reduce the amount of such texts in the corpus, because the influence of the intermediary language is often very noticeable in them. In the example below, the Russian translator did not render the author's idiom, but instead used expressive syntax. The Ukrainian translator calqued this syntactic construction, which is not very frequent in Ukrainian.

- *(de) Ich bin ein ausgewichster Panzermann, aber die sind doch keine halbe Nase weniger schlau! [I am a good tankman, but they are not less smart!] (Dieter Noll. Die Abenteuer des Werner Holt. 1960)*

  (ru) Už na čto ja byvalyj tankist, no oni ničut' ne glupee! (Translation by V. Kurilla, R. Galperin. 1962)[2]

  (uk) Naščo vže ja buvalyj tankist, ale vony anitroxy ne durniši! (Translation by Y. Mykhailyuk. 1965)

According to M. Ivanytska, censorship did occur in Ukrainian translations from German under the Soviet regime, and our material also shows this. In such cases, we keep the untranslated text in the corpus without a Ukrainian version (Table 2).

In literary translations from German, we also often find just omitted and shortened fragments, cases of inaccurate translation, rearranged sentences, etc. that are not related to censorship.

Lion Feuchtwanger. Erfolg. 1929	Ukrainian translation. Oleksa Oleksa Synyčenko. 1980
Möglich, daß in Bayern die Justiz besonders bösartig und verbohrt gehandhabt wurde, aber viel anders war es ringsum auch nicht. [It is possible that justice was administered in Bavaria in a particularly malicious and biased manner, but it was not much better in other countries.]	Možlyvo, ščo v Bavariï pravosuddja čynyly osoblyvo zlisno j uperedženo, ale ne nabahato krašče bulo i v inšyx kraïnax.
In Ungarn, auf dem Balkan, in Rußland stand es vielleicht noch schlimmer als auf der bayrischen Hochebene. [In Hungary, the Balkans, and Russia, the situation was probably even worse than on the Bavarian Plateau.]	—

Table 2: Soviet Censorship in German-Ukrainian Translation.

[2] Hereinafter, examples in Cyrillic are transliterated.

ParaRook||DE-UK has size of 382 thousand sentences and 6,3 million tokens in the German-language part. The core of the corpus currently consists of 20th-century fiction translated from German into Ukrainian. The corpus contains 58 texts: 53 translated from German and 5 from Ukrainian. The corpus features works by 29 famous German-speaking authors from different countries, which makes it possible to compare regional variants of the German language. The corpus includes novels by Erich Maria Remarque, Thomas Mann, Heinrich Mann, Hermann Hesse, Alfred Döblin, Dieter Noll, Heinrich Böll, Günter Grass, Patrick Süskind (Germany), Franz Kafka, Stefan Zweig, Robert Musil, Gustav Meyrink, Joseph Roth (Austria), Friedrich Dürrenmatt (Switzerland), and other writers (Appendix B).

2.2. Annotation and Technical Details

The texts for the corpus were collected manually from public libraries on the Internet (the sources are given in the metadata), most Ukrainian texts were taken from GRAC. The original texts and translations were aligned using the InterText program (Vondřička, 2014), and the alignment of all texts was checked and corrected manually. All the cases of inaccurate translation were saved in the corpus for research, the texts were aligned without changing the structure of the original text or translation. Aligned parallel texts are saved in tmx format, e.g.:

<tu><prop type="x-sentbreak">|#|</prop>

<tuv xml:lang="de"><seg>Der Knabe war klein, die Berge waren ungeheuer.</seg></tuv>

<tuv xml:lang="uk"><seg>Xlop'ja bulo male, hory – vysočezni.</seg>

</tuv>

</tu>

The parallel corpus was annotated with UDPipe2 (Straka, 2018) using Universal Dependencies models, namely GSD for German (Petrov, 2023) and IU for Ukrainian (Kotsyba, 2016). The choice of the German model was based on model evaluations on the official UD website (Nivre, 2015 - 2024), while the Ukrainian model was the only one presented. Ukrainian's current Universal Dependencies model achieves an accuracy rate of 97.5% for POS tagging, 91.6% for morphological features, and 81.7% for syntactic relation (Kotsyba, 2018).

The Universal Dependencies were chosen because their annotation is universal regardless of the language of the analyzed text, and the process can be optimized using graphics processors, particularly NVIDIA CUDA technology, which significantly speeds up computation, as opposed to using a traditional CPU.

The annotation of documents in the parallel corpus also includes syntactic relations between words within a sentence, which can serve as a source of data for contrastive syntactic analysis (Poiret et al., 2021). During the preprocessing of the corpus materials, the text was segmented into sentences using the SpaCy models appropriate to the language of the text: uk_core_news_sm (Kurnosov, 2022) and de_core_news_sm (Brants, 2023); this step was necessary to improve accuracy in morphosyntactic annotation using UD.

The corpus manager used was NoSketch Engine, one of the most featureful open-source corpus manager solutions available. The corpus is accessible for search on the website: https://uacorpus.org/Kyiv/ua/pararook

Besides morphosyntactic annotation, the corpus provides extensive metadata. The list below presents all the necessary information regarding metadata and tag descriptions.

word: Token attribute for a word.

lemma: Token attribute for lemma.

upos: Token attribute for UD part-of-speech tag.

xpos: Language-specific grammatical annotation token attribute.

morphology: Morphological annotation.

head: Syntactically the main word in a sentence.

dependency_tag: Syntactic relationship of a word in a sentence.

extra_dependency: Additional information about the syntactic role of a word in a sentence.

authors_names_{uk|de}: Authors' name in Ukrainian/German.

translators_names_{uk|de}: Translators' name in UK/DE.

authors_born: Authors' birth year.

authors_sex: Authors' gender.

authors_regionCode: Authors' region.

translators_regionCode: Translators' region.

translators_born: Translators' birth year.

translators_sex: Translators' gender.

title_{uk|de}: Document title in UK/DE.

original_language: Original language.

date_{uk|de}: Year of creation in UK/DE.

pub_city_{uk|de}: City of publication in UK/DE.

publisher_{uk|de}: Publisher in UK/DE.

pub_year_{uk|de}: Year of publication in UK/DE.

publication_{uk|de}: Title of publication in UK/DE (magazine number, title of collection).

url_{uk|de}: Reference to the source of the document in UK/DE.

An example of parallel sentences with metadata is provided in Appendix A.

3. Using ParaRook||DE-UK

Since ParaRook||DE-UK is only available for online search, it is intended primarily for academic linguistic and translation studies, for compiling dictionaries, as well as for use in the process of human translation.

A parallel corpus not only provides a richer range of translation options in context than a dictionary, but also enables research on phenomena that do not have a well-established translation: it can be used to study lacunarity and non-equivalent linguistic patterns (Sitchinava, 2016; Dobrovol'skij and Pöppel, 2017; Mellado Blanco, 2019; Grabowski and Groom, 2022)

For example, the German construction *immer noch* 'still' may be translated into Ukrainian in many different ways, or it may be omitted in translation at all. In a random sample of one hundred parallel sentences from ParaRook||DE-UK, the following translation variants were found: *i(j) dosi (21 times), vse(use) šče (15), šče(išče) (15), i(j) dali (7), tak samo (7), j (4), vse (2), vse(use) ž taky (2), vse odno (1), dali (1), zavždy (1), i vse odno (1), i dosi šče (1), j tak samo (1), poky ščo (1), skil'ky zavhodno (1), tak use j (1), teper (1), u krajn'omu razi (1), šče j dovho (1), šče j dosi (1), šče raz (1), jak i raniše (1).* In 12 cases of 100, the German construction had no equivalent in Ukrainian translation at all, e. g.

- *(de) Wir tranken, und **immer noch** standen die Uhrzeiger, wie sie schon seit drei Wochen standen: auf halb elf. (Heinrich Böll. Irisches Tagebuch. 1957)*

  (uk) My pyly, a strilky hodynnyka stojaly na misci, jak i ves' čas protjahom ostannix tr'ox tyžniv, — na piv na odynadcjatu. (Ukr. translation by Volodymyr Šelest. 1989)

When working with one language, the translation presented in the parallel corpus can be used as an additional layer of annotation, which makes it possible to search by semantics. This advantage of parallel data is already being extensively used for automatic word sense disambiguation (Yee Seng Chan and Zhong, 2007; Hwee Tou Ng and Chan, 2003; Banea and Mihalcea, 2011; Shahid and Kazakov, 2013), and it can be useful for a manual lexical research as well. Below is an example of search results in ParaRook||DE-UK of a Ukrainian word *kaminec'* that has two meanings, a commonly used 'small stone' and a rarely used 'fruit bone'. To find examples in the second meaning only, the German equivalent *Kern* was used, which helped to specify the required sense.

- *(de) Weißrot klappern Störche auf Dächern, daß Kirschen die **Kerne** ausspucken... (Günter Grass. Blechtrommel. 1959)*

  *(uk) Bilo-červoni busly triskotjat' na daxax pro te, ščo vyšni vypl'ovujut' svoï **kaminci**... (Ukr. translation by Oleksa Lohvynenko. 2005)*

- *(de) Sie brach eine der überreifen Früchte auf, warf den **Kern** zu Boden und reichte ihm eine der Hälften. (Dieter Noll. Die Abenteuer des Werner Holt. 1960)*

  *(uk) Potim rozlomyla najspilišyj plid i, vykynuvšy **kamінčyka**, prostjahla polovynu Hol'tovi. (Ukr. translation by Jurij Myxajljuk. 1965)*

More examples of the use of parallel corpora for manual research and teaching can be found in the relevant work presented in our bibliography.

4. Conclusions and Future Plans

The first representative German-Ukrainian parallel corpus has been created and is available to search online. This is an important language resource that provides parallel texts for linguistic and translation research. With this work, we would like to draw attention to the importance of making computational linguistic resources more inclusive for philologists, not only for wider use of such resources in academic work, but also for involving professional linguists, translators, and texts experts in the development of quality textual data.

In the future, we plan to add more texts translated from Ukrainian into German and to develop parallel corpora for other languages with Ukrainian, primarily English and French.

It is possible to create a much larger German-Ukrainian corpus based on ParaRook||DE-UK by adding non-fiction texts, such as legal, news, and subtitles, which are usually translated quite literally and require less manual alignment checking. They can be downloaded from the Internet and automatically aligned.

As currently a single Universal Dependencies model is available, we plan to expand the range of models at hand for the Ukrainian language in UD. This expansion aims to improve the accuracy of morphosyntactic analyses and contribute to developing more robust and diverse linguistic tools.

5. Limitations

Since we check the alignment manually, it would be a challenge to collect a corpus larger than several millions of tokens. Manual alignment checking is highly desirable for fictional texts, where the translation is often not quite literal, but it takes a lot of time.

Based on Universal Dependencies, the current morphosyntactic analysis of the Ukrainian language needs to yield optimal accuracy. Improving this system is of great importance for the further development of parallel corpora. While the current accuracy is promising, more is required for large corpora. Optimal accuracy is critical in syntax studies that use parallel corpora to ensure reliable and meaningful findings.

6. Acknowledgements

We owe our acknowledgements to the German academic community for their strong support of Ukrainian researchers and students.

We would like to thank the reviewers for their useful comments on the paper.

7. Bibliography

References

G. Anderman and M. Rogers. 2007. *Incorporating Corpora: The Linguist and the Translator*. Multilingual Matters.

C. Banea and R. Mihalcea. 2011. Word sense disambiguation with multilingual features. In *Proceedings of the Ninth International Conference on Computational Semantics (IWCS 2011)*.

M. Bañón, P. Chen, B. Haddow, K. Heafield, H. Hoang, M. Esplà-Gomis, M. Forcada, A. Kamran, F. Kirefu, P. Koehn, et al. 2020. ParaCrawl: Web-scale acquisition of parallel corpora. Association for Computational Linguistics (ACL).

M. Bañón, M. Esplà-Gomis, M. L. Forcada, C. García-Romero, T. Kuzman, N. Ljubešić, R. van Noord, L. P. Sempere, G. Ramírez-Sánchez, P. Rupnik, V. Suchomel, A. Toral, T. van der Werff, and J. Zaragoza. 2022. MaCoCu: Massive collection and curation of monolingual and bilingual data: focus on under-resourced languages. In *Proceedings of the 23rd Annual Conference of the European Association for Machine Translation*, pages 303–304, Ghent, Belgium. European Association for Machine Translation.

S. Buk. 2012. Arxitektura pol's'ko-ukraïns'koho ta ukraïns'ko-pol's'koho paralel'noho korpusu avtoperekladiv Ivana Franka. *Slavia Orientalis*, LXI(2):213–230.

M. R. Costa-jussà, J. Cross, O. Çelebi, M. Elbayad, K. Heafield, K. Heffernan, E. Kalbassi, J. Lam, D. Licht, J. Maillard, et al. 2022. No language left behind: Scaling human-centered machine translation. *arXiv preprint arXiv:2207.04672*.

N. P. Darčuk, M. O. Lanhenbax, V. M. Sorokin, and Ja. V. Xodakivs'ka. 2017. Paralel'nyj korpus tekstiv ParKUM. *Naukovyj časopys Nacional'noho pedahohičnoho universytetu imeni M. P. Drahomanova. Serija 9 : Sučasni tendenciï rozvytku mov : zb. nauk. prac'*, 15:28–35.

D. Dobrovol'skij and L. Pöppel. 2017. Constructions in parallel corpora: A quantitative approach. In *Computational and Corpus-Based Phraseology*, volume 10596.

A. El-Kishky, V. Chaudhary, F. Guzmán, and P. Koehn. 2020. CCAligned: A massive collection of cross-lingual web-document pairs. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP 2020)*, pages 5960–5969, Online. Association for Computational Linguistics.

R. Enghels, B. Defrancq, and M. Jansegers. 2020. *New approaches to contrastive linguistics: empirical and methodological challenges*. De Gruyter Mouton.

A. Fan, S. Bhosale, H. Schwenk, Z. Ma, A. El-Kishky, S. Goyal, M. Baines, O. Celebi, G. Wenzek, V. Chaudhary, et al. 2021. Beyond English-centric multilingual machine translation. *Journal of Machine Learning Research*, 22(107):1–48.

N. Grabar and T. Hamon. 2017. Creation of a multilingual aligned corpus with Ukrainian as the target language and its exploitation. In *COLINS 2017*, Kharkiv, Ukraine.

Ł. Grabowski and N. Groom. 2022. Functionally-defined recurrent multi-word units in English-to-Polish translation. *Revista Española de Lingüística Aplicada/Spanish Journal of Applied Linguistics*, 35(1):1.

S. Hansen-Schirra, S. Neumann, and E. Steiner. 2012. *Cross-linguistic corpora for the study of translations: insights from the language pair English-German*. de Gruyter Mouton.

Bin Wang Hwee Tou Ng and Yee Seng Chan. 2003. Exploiting parallel texts for word sense disambiguation: An empirical study. pages 455–462.

M. Ivanytska. 2015. *Osobystist' perekladača v ukraïns'ko-nimec'kyx literaturnyx vzajemynax*. Knyhy – XXI, Černivci. [The Personality of the Translator in Ukrainian-German Literary Relations].

N. Kotsyba. 2016. Polsko-ukraiński korpus równoległy PolUKR i jego następca PolUKR-2. pages 133–142.

P. Lison and J. Tiedemann. 2016. OpenSubtitles2016: Extracting large parallel corpora from movie and TV subtitles. In *Proceedings of the 10th International Conference on Language Resources and Evaluation (LREC-2016)*.

Kanglong Liu. 2020. *Corpus-Assisted Translation Teaching: Issues and Challenges*, 1st edition. Springer Singapore.

K. S. Mandziy, U. V. Yurlova, and M. P. Dilai. 2022. English-Ukrainian parallel corpus of IT texts: Application in translation studies.

S. Matvieieva. 2019. Selection criteria and initial processing of empirical material for a parallel corpus of legal texts. *Forum Filologiczne Ateneum*, 1(7):167–181.

C. Mellado Blanco. 2019. Phrasem-konstruktionen kontrastiv Deutsch–Spanisch: ein korpusbasiertes beschreibungsmodell anhand ironischer vergleiche. *Yearbook of Phraseology*, 10(1):65.

Ra Poiret, S. Mille, and Haitao Liu. 2021. Paraphrase and parallel treebank for the comparison of French and Chinese syntax. *Languages in Contrast*, 21(2):298–322.

H. Schwenk, G. Wenzek, S. Edunov, E. Grave, and A. Joulin. 2019. Ccmatrix: Mining billions of high-quality parallel sentences on the web. *arXiv preprint arXiv:1911.04944*.

A. R. Shahid and D. Kazakov. 2013. Using parallel corpora for word sense disambiguation. In *Proceedings of Recent Advances in Natural Language Processing*, pages 336–341, Hissar, Bulgaria.

O. Siruk and I. Derzhanski. 2013. Linguistic corpora as international cultural heritage: The corpus of Bulgarian and Ukrainian parallel texts. In *Digital Presentation and Preservation of Cultural and Scientific Heritage*, volume 3, pages 91–98.

D. Sitchinava. 2016. Parallel corpora as a source of defining language-specific lexical items. In *Proceedings of the XVII EURALEX International Congress*, pages 394–401.

D. V. Sitchinava, O. O. Tyshchenko-Monastyrska, and M. O. Shvedova. 2011. Paralel'ni ukrayins'ko-rosiys'kyy ta rosiys'ko-ukrayins'kyy korpusy. *Leksykohrafichnyy byuleten*, 20:35–38.

V. Starko and A. Rysin. 2022. VESUM: A large morphological dictionary of Ukrainian as a dynamic tool. In *COLINS*, volume 6th Int. Conf, pages 71–80, Gliwice.

M. Straka. 2018. UDPipe 2.0 prototype at CoNLL 2018 UD shared task. In *Proceedings of the CoNLL 2018 Shared Task: Multilingual Parsing from Raw Text to Universal Dependencies*, pages 197–207, Brussels, Belgium. Association for Computational Linguistics.

J. Tiedemann. 2012. Parallel data, tools and interfaces in OPUS. In *Proceedings of the 8th International Conference on Language Resources and Evaluation (LREC'2012)*.

J. Tiedemann and S. Thottingal. 2020. OPUS-MT — Building open translation services for the World. In *Proceedings of the 22nd Annual Conferenec of the European Association for Machine Translation (EAMT)*, Lisbon, Portugal.

R. von Waldenfels. 2011. Recent developments in ParaSol: Breadth for depth and XSLT based web concordancing with CWB. In *Natural Language Processing, Multilinguality. Proceedings of Slovko 2011*, pages 156–162, Bratislava. Tribun.

R. von Waldenfels. 2012. ParaSol: Introduction to a Slavic parallel corpus. *Prace Filologiczne*, LXIII:293–302.

P. Vondřička. 2014. Aligning parallel texts with InterText. In *Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC'14)*, pages 1875–1879. European Language Resources Association (ELRA).

Hwee Tou Ng Yee Seng Chan and Zhi Zhong. 2007. NUS-PT: Exploiting parallel texts for word sense disambiguation in the English all-words tasks. *Proc. 4th Int. Workshop Semantic Eval.*, pages 253–256.

I. Zariņa, P. Ņikiforovs, and R. Skadiņš. 2015. Word alignment based parallel corpora evaluation and cleaning using machine learning techniques. In *Proceedings of the 18th Annual Conference of the European Association for Machine Translation*, pages 185–192.

S. V. Zasiekin. 2020. Psycholinguistic regularities of reproducing literary texts in translation (based on the English and Ukrainian languages).

F. Čermák and A. Rosen. 2012. The case of InterCorp, a multilingual parallel corpus. *International Journal of Corpus Linguistics*, 13(3):411–427.

8. Language Resource References

Brants, S., et al. 2023. *de_core_news_sm*.

Kotsyba, N., et al. 2016. *UD Ukrainian IU*. Institute for Ukrainian, NGO.

Kotsyba, N., et al. 2018. *https://mova.institute/*.

Kotsyba, N. and Turska, M. 2005 - 2011. *Polsko-ukrainski korpus rownolegly*.

Kurnosov, V., et al. 2022. *uk_core_news_sm*.

Nivre, J. et al. 2015 - 2024. *Universal Dependencies*.

Petrov, S., et al. 2023. *UD German GSD*.

Shvedova, M., et al. 2017 - 2024. *GRAC*.

Shvedova, M. and Lukashevskyi, A. 2023-2024. *ParaRook||DE-UK*.

Čermák, F. and Rosen, A. 2008 - 2023. *Intercorp*.

9. Appendix A: Example of parallel sentences with metadata

```
Ukrainian Text:
<doc authors_names_uk="Maks Friš"
    ↪ authors_names_de="Max Frisch"
    ↪ translators_names_uk="Jevhen
    ↪ Popovyč" translators_names_de="
    ↪ Jevhen Popovyč" authors_born
    ↪ ="1911" authors_sex="M"
    ↪ translators_born="1930"
    ↪ translators_sex="M"
    ↪ authors_regionCode="D-Z-CH"
    ↪ translators_regionCode="UA-C-CRK &
    ↪  UA-KYV-KYV" title_uk="Štiller"
    ↪ title_de="Stiller"
    ↪ original_language="DE" date_uk
    ↪ ="1968" pub_city_uk="Kyv"
    ↪ pub_year_uk="1970" publication_uk
    ↪ ="" url_uk="http://chtyvo.org.ua/"
    ↪  pub_city_de="Berlin" publisher_de
    ↪ ="Suhrkamp Verlag" publication_de
    ↪ ="" date_de="1954" url_de="library
    ↪ .lol/fiction/447
    ↪ EC7654424E50DD38BA58324825B29"
    ↪ orthography="sučasnyj pravopys"
    ↪ genre="" source="PRI" theme=""
    ↪ media="" style="FIC">
<align>
<s>
1 C'oho ce PRON Pd--nnsgn Animacy=Inan|
    ↪ Case=Gen|Gender=Neut|Number=Sing|
    ↪ PronType=Dem 2 obj _ _
2 vystačylo vystačyty VERB Vmeis-sn
    ↪ Aspect=Perf|Gender=Neut|Mood=Ind|
    ↪ Number=Sing|Tense=Past|VerbForm=
    ↪ Fin 0 root _ SpaceAfter=No
<g/>
3 . . PUNCT U _ 2 punct _ SpacesAfter=\r
    ↪ \n
</s>
<s>
1 Vin vin PRON Pp-3m-snn Case=Nom|Gender
    ↪ =Masc|Number=Sing|Person=3|
    ↪ PronType=Prs 2 nsubj _ _
2 zasmijavsja zasmijatysja VERB Vmeis-sm
    ↪  Aspect=Perf|Gender=Masc|Mood=Ind|
    ↪ Number=Sing|Tense=Past|VerbForm=
    ↪ Fin 0 root _ SpaceAfter=No
<g/>
3 . . PUNCT U _ 2 punct _ SpaceAfter=No
</s>
</align>
</doc>

German Text:
<doc authors_names_uk="Maks Friš"
    ↪ authors_names_de="Max Frisch"
    ↪ translators_names_uk="Jevhen
    ↪ Popovyč" translators_names_de="
    ↪ Jevhen Popovyč" authors_born
    ↪ ="1911" authors_sex="M"
    ↪ translators_born="1930"
    ↪ translators_sex="M"
    ↪ authors_regionCode="D-Z-CH"
    ↪ translators_regionCode="UA-KYV-KYV
    ↪  & UA-C-CRK" title_uk="Štiller"
    ↪ title_de="Stiller"
    ↪ original_language="DE" date_uk
    ↪ ="1968" pub_city_uk="Kyv"
    ↪ pub_year_uk="1970" publication_uk
    ↪ ="" url_uk="http://chtyvo.org.ua/"
    ↪  pub_city_de="Berlin" publisher_de
    ↪ ="Suhrkamp Verlag" publication_de
    ↪ ="" date_de="1954" url_de="library
    ↪ .lol/fiction/447
    ↪ EC7654424E50DD38BA58324825B29"
    ↪ orthography="sučasnyj pravopys"
    ↪ genre="" source="PRI" theme=""
    ↪ media="" style="FIC">
<align>
<s>
1 Das der PRON PDS Case=Nom|Gender=Neut|
    ↪ Number=Sing|PronType=Dem,Rel 2
    ↪ nsubj _ _
2 genügte genügen VERB VVFIN Mood=Ind|
    ↪ Number=Sing|Person=3|Tense=Past|
    ↪ VerbForm=Fin 0 root _ SpaceAfter=
    ↪ No
<g/>
3 . . PUNCT $. _ 2 punct _ SpacesAfter=\
    ↪ r\n
</s>
<s>
1 Er er PRON PPER Case=Nom|Gender=Masc|
    ↪ Number=Sing|Person=3|PronType=Prs
    ↪ 2 nsubj _ _
2 lachte lachen VERB VVFIN Mood=Ind|
    ↪ Number=Sing|Person=3|Tense=Past|
    ↪ VerbForm=Fin 0 root _ SpaceAfter=
    ↪ No
<g/>
3 . . PUNCT $. _ 2 punct _ SpaceAfter=No
</s>
</align>
</doc>
```

10. Appendix B: Corpus content and statistics

Author	Date (original)	Date (translation)	Original language	Title	Translator	Style (Fiction/Nonfiction/Ego-text)	Tokens (in the German-language part)
Alfred Döblin	1928	2020	DE	Berlin Alexanderplatz	Roman Osadčuk	FIC	200369
Andreas Kappeler	1995	2007	DE	Kleine Geschichte der Ukraine	Oleh Blaščuk	NOF	79781
Andreas Kappeler	2017	2018	DE	Ungleiche Brüder: Russen und Ukrainer vom Mittelalter bis zur Gegenwart	Volodymyr Kam'janec'	NOF	68549
Bernhard Kellermann	1913	1986	DE	Der Tunnel	Oleksa Lohvynenko	FIC	123092
Bernhard Schlink	1995	2016	DE	Der Vorleser	Petro Taraščuk	FIC	50080
Bertolt Brecht	1934	1973	DE	Dreigroschenroman	Jurij Lisnjak	FIC	142529
Bertolt Brecht	1943	1968	DE	Das Leben Des Galilei	Vasyl' Stus & Zinaïda Joffe	FIC	37205
Bertolt Brecht	1939	1973	DE	Mutter Courage und ihre Kinder	Marko Zisman	FIC	28711
Bohdan Scholdak	2000	1991	UK	Der Steinzeitmensch	Anna-Halja Horbatsch	FIC	2878
Christoph Ransmayr	1988	1992	DE	Die letzte Welt	Oleksa Lohvynenko	FIC	70374
Dieter Noll	1960	1965	DE	Die Abenteuer des Werner Holt. Roman einer Heimkehr. I	Jurij Myxajljuk	FIC	212573
Dieter Noll	1963	1965	DE	Die Abenteuer des Werner Holt. Roman einer Heimkehr. II	Jakiv Prylypko	FIC	188034
Elias Canetti	1935	2003	DE	Die Blendung	Oleksa Lohvynenko	FIC	228945
Erich Maria Remarque	1945	1986	DE	Arc de Triomphe	Jevhen Popovyč	FIC	185739
Erich Maria Remarque	1962	1963	DE	Die Nacht von Lissabon	Mykola Djatlenko & Arkadij Pljuto	FIC	93365
Erich Maria Remarque	1929	1986	DE	Im Westen nichts Neues	Kateryna Hlovac'ka	FIC	70858
Franz Kafka	1922	2006	DE	Das Schloss	Natalka Snjadanko	FIC	132233
Franz Kafka	1919	2012	DE	Brief an den Vater	Oleksa Lohvynenko	EGO	19412
Friedrich Dürrenmatt	1985	1987	DE	Justiz	Oleksa Lohvynenko	FIC	61489
Friedrich Dürrenmatt	1951	1989	DE	Der Richter und sein Henker	Kateryna Hlovac'ka	FIC	28828
Friedrich Glauser	1938	1994	DE	Wachtmeister Studer	Jurij Lisnjak	FIC	64239
Günter Grass	1959	2005	DE	Blechtrommel	Oleksa Lohvynenko	FIC	246103
Günter Grass	1961	2008	DE	Katz und Maus	Natalka Snjadanko	FIC	46170
Gustav Meyrink	1914	2011	DE	Der Golem	Natalja Ivanyčuk	FIC	88952
Heinrich Böll	1971	1972	DE	Gruppenbild mit Dame	Jevhen Popovyč & Jurij Lisnjak	FIC	164718
Heinrich Böll	1963	1965	DE	Ansichten eines Clowns	Mykola Djatlenko	FIC	92583
Heinrich Böll	1974	1989	DE	Die verlorene Ehre der Katharina Blum	Petro Sokolovs'kyj	FIC	36765
Heinrich Böll	1957	1989	DE	Irisches Tagebuch	Wladimir Schelest	FIC	35165
Heinrich Böll	1950	1969	DE	Wanderer, kommst du nach Spa…	Jevhenija Horeva	FIC	4001
Heinrich Mann	1935	1985	DE	Die Vollendung des Königs Henri Quatre	Jurij Lisnjak	FIC	321930
Heinrich Mann	1935	1975	DE	Die Jugend des Königs Henri Quatre	Jurij Lisnjak	FIC	244916
Heinrich Mann	1914	1969	DE	Der Untertan	Marko Zisman	FIC	163934
Hermann Hesse	1927	1977	DE	Steppenwolf	Jevhen Popovyč	FIC	81745
Ingrid Noll	1994	2019	DE	Die Apothekerin	Svjatoslav Zubčenko	FIC	62933
Joseph Roth	1930	2010	DE	Hiob	Jurij Proxas'ko	FIC	62317
Joseph Roth	1937	2010	DE	Das falsche Gewicht	Jurij Proxas'ko	FIC	41473
Joseph Roth	1939	2011	DE	Die Legende vom heiligen Trinker	Ol'ha Sydor	FIC	12842
Jurij Wynnytschuk	2000	1990	UK	Das Leuchten	Anna-Halja Horbatsch	FIC	12942
Lesja Ukrainka	2014	1900	UK	«Deine Briefe duften immer nach abgeblühten Rosen…»	Stanislaw Matijtschyn & Michael Beck	FIC	616
Lion Feuchtwanger	1929	1980	DE	Erfolg	Oleksa Synyčenko	FIC	308328
Martin Walser	1957	1975	DE	Ehen in Philippsburg	Jevhen Popovyč & Jarema Polotnjuk	FIC	116209
Max Frisch	1954	1968	DE	Stiller	Jevhen Popovyč	FIC	172306
Oleksander Denyssenko	2000	2000	UK	Die Seele des Flusses	Anna-Halja Horbatsch	FIC	3086
Otfried Preußler	1980	2006	DE	Krabat	Volodymyr Vasyljuk	FIC	73575
Patrick Süskind	1985	1993	DE	Das Parfum: Die Geschichte eines Mörders	Iryna Fridrix	FIC	89643
Patrick Süskind	1991	1995	DE	Die Geschichte von Herrn Sommer	Iryna Fridrix	FIC	19579
Patrick Süskind	1981	1996	DE	Der Kontrabaß	Iryna Fridrix	FIC	14387
Peter Handke	1976	1980	DE	Die linkshändige Frau	Oleksa Lohvynenko	FIC	23956
Robert Musil	1942	2010	DE	Der Mann ohne Eigenschaften. I.	Oleksa Lohvynenko	FIC	191592
Siegfried Lenz	1968	1976	DE	Deutschstunde	Oleksa Lohvynenko	FIC	193369
Stefan Zweig	1932	2017	DE	Marie Antoinette	Petro Taraščuk	FIC	183739
Stefan Zweig	1935	2018	DE	Maria Stuart	Petro Taraščuk	FIC	144602
Stefan Zweig	1929	2017	DE	Joseph Fouché	Petro Taraščuk	FIC	87781
Stefan Zweig	1911	1981	DE	Die Gouvernante	Iryna Stešenko	FIC	6512
Thomas Mann	1924	2008	DE	Der Zauberberg	Roman Osadčuk	FIC	375979
Thomas Mann	1900	1973	DE	Buddenbrooks	Jevhen Popovyč	FIC	281941
Thomas Mann	1954	2011	DE	Bekenntnisse des Hochstaplers Felix Krull	Roman Osadčuk	FIC	150007
Vasyl Barka	2007	1961	UK	Der gelbe Fürst	Maria Ostheim-Dzerowycz	FIC	135734

Introducing NER-UK 2.0: A Rich Corpus of Named Entities for Ukrainian

Dmytro Chaplynskyi, Mariana Romanyshyn
Lang-uk, Grammarly,
Kyiv, Ukraine
chaplinsky.dmitry@gmail.com, mariana.romanyshyn@grammarly.com

Abstract

This paper presents NER-UK 2.0, a corpus of texts in the Ukrainian language manually annotated for the named entity recognition task. The corpus contains 560 texts of multiple genres, boasting 21,993 entities in total. The annotation scheme covers 13 entity types, namely location, person name, organization, artifact, document, job title, date, time, period, money, percentage, quantity, and miscellaneous. Such a rich set of entities makes the corpus valuable for training named-entity recognition models in various domains, including news, social media posts, legal documents, and procurement contracts. The paper presents an updated baseline solution for named entity recognition in Ukrainian with 0.89 F_1. The corpus is the largest of its kind for the Ukrainian language and is available for download.

Keywords: Named Entity Recognition, NER, Evaluation datasets, Manual annotation

1. Introduction

Named entity recognition (NER) is a fundamental task in natural language processing (NLP) that involves finding a sequence of tokens that denotes a specific concept, like a location, a person, or an organization. NER is often an essential component for other NLP tasks such as information extraction (Liu et al., 2021), question answering (Xu et al., 2021), or information retrieval (Aliwy et al., 2021).

In the classic setup, NER is formalized as a sequence labeling task. Despite the recent advances in NLP systems, particularly the emergence of large language models (LLMs), new high-quality annotated datasets for developing NER systems are still in high demand. Specifically, recent work (Qin et al., 2023; Wang et al., 2023) discusses the poor performance of LLMs as zero-shot classifiers for the NER tasks in comparison to fine-tuned pre-trained language models that rely on task-specific annotated datasets.

The need for labeled data is even more apparent in the context of low-resource languages, narrow domains, or task-specific entity types, where the systems experience a scarcity of data as is. To address this need and to facilitate further advancements in NER and related tasks, we present NER-UK 2.0[1], a new corpus of texts in Ukrainian manually annotated for a rich set of entities. The corpus contains 560 texts of multiple genres labeled for 13 entity types, totaling 21,993 entities. This paper provides a description of NER-UK 2.0 and sets a new baseline for the NER task in Ukrainian.

The rest of the paper is organized as follows. In Section 2, we review the related work on corpora for NER in Ukrainian. Section 3 presents previous work that includes the development of the first version of NER-UK and the motivation to build the second version. Section 4 describes the new tagset, sources of texts for the corpus, and the data annotation process. Section 5 presents the results in the form of corpus statistics, inter-annotator agreement, and the new NER baseline for Ukrainian. Finally, Section 6 summarizes the contributions, and Section 7 acknowledges the limitations of the presented corpus.

2. Related Work

To our knowledge, NER-UK 2.0 is the only publicly available corpus of manually annotated entities. Makogon and Samokhin (2022) mention creating a news corpus of manually annotated entities for Ukrainian, but the described dataset was never publicly released. Their annotation scheme included five entity types: person, location, organization, product, and other. Further, we review datasets that are publicly available but automatically annotated.

The POLYGLOT-NER corpus (Al-Rfou et al., 2014) contains Wikipedia articles automatically annotated for the task of named entity recognition. The labeling scheme defines three entity types: person, location, and organization. The corpus covers 40 languages, including Ukrainian.

WikiANN (Pan et al., 2017), similarly, builds on Wikipedia and solves a related task of automated entity tagging and linking for person names. Ukrainian is included as part of the multilingual dataset.

[1] `https://github.com/lang-uk/ner-uk`

25 May, 2024. © 2024 ELRA Language Resource Association: CC BY-NC 4.0

In 2022, Kurnosov V. published Ukr-Synth[2], a large silver standard Ukrainian corpus automatically annotated for part-of-speech tags, syntax trees, and three entity types: person, location, and organization. The corpus represents the Ukrainian subset of Leipzig Corpora Collection (Goldhahn et al., 2012) which originates from newspaper texts.

Although the amount of annotated data in the mentioned corpora is enviable, the limited set of entity types, the lack of quality verification, and the focus on Wikipedia and news genres pose limitations with regard to the wide adoption of these resources. We are set to fill the identified gaps with the release of the NER-UK 2.0 corpus.

3. Background and Motivation

In 2016, our team introduced the first version of the named entity recognition corpus for the Ukrainian language called NER-UK[3]. This corpus comprised 262 texts borrowed from the multi-genre BRUK corpus (Starko and Rysin, 2023), totaling 237,327 words and including press, religious texts, fiction, legal documents, and other types of writing. NER-UK featured crowdsourced manual annotation of 7,441 entities across four distinct types: *person* (4,387 entities), *location* (1,614 entities), *organization* (780 entities), and *miscellaneous* (660 entities). The latter covered names of holidays, sports events, natural disasters, etc.

The creation of NER-UK marked a significant milestone, providing the Ukrainian NLP community with a valuable resource for developing and evaluating NER systems and, more recently, large pretrained language models, like roberta-large[4]. Data from the corpus was used to train state-of-the-art (SOTA) NER systems[5], contributing to advancements in Ukrainian natural language processing. Additionally, the choice of BRUK as the source of texts for entity annotation presented opportunities for multi-task learning since BRUK is also annotated for parts of speech.

With regard to the limitations of NER-UK, it should be noted that the corpus was of a relatively small size and had a limited entity set. The *miscellaneous* entity type was too broad and not very informative. The genre diversity, while beneficial in providing a varied set of contexts, resulted in a low density of entities in the texts of certain genres.

We started the NER-UK 2.0 project with the aim of addressing the limitations of NER-UK. Specifically, we set the following goals:

- increase the size of the corpus while preserving high quality standards;
- increase the density of entities in the corpus with better source text selection;
- adopt a more extensive tagset that would offer both a bigger number of entity types and a better granularity of entities, making the annotations more informative.

4. NER-UK 2.0 Corpus Creation

This section describes the updated tagset, the corpus composition, and the annotation process.

4.1. Annotation Scheme

In the first version of NER-UK, we considered *person, organization, location,* and *miscellaneous* as named entities. Inspired by the extended set of entities in Stanford CoreNLP (Manning et al., 2014), we introduced nine additional entity types for NER-UK 2.0, which resulted in the refined annotation guidelines, better granularity of the *miscellaneous* type, and broader applicability of the annotations.

The full list of entities includes:

- **ORG** — a name of a company, brand, agency, organization, institution (including religious, informal, non-profit), party, people's association, or specific project like a conference, a music band, a TV program, etc. Example: *UNESCO*.
- **PERS** — a person name where person may refer to humans, book characters, or humanoid creatures like vampires, ghosts, mermaids, etc. Example: *Marquis de Sade*.
- **LOC** — a geographical name, including names of districts, villages, cities, states, counties, countries, continents, rivers, lakes, seas, oceans, mountains, etc. Example: *Ukraine*.
- **MON** — a sum of money including the currency. Examples: *$40, 1 mln hryvnias*.
- **PCT** — a percent value including the percent sign or the word "percent". Example: *10%*.
- **DATE** — a full or incomplete calendar date that may include a century, a year, a month, or a day. Examples: *last week, 10.12.1999*.
- **TIME** — a textual or numerical timestamp. Examples: *half past six, 18:30*.

[2] https://huggingface.co/datasets/ukr-models/Ukr-Synth

[3] https://github.com/lang-uk/ner-uk/tree/master/v1.0

[4] https://huggingface.co/benjamin/roberta-large-wechsel-ukrainian

[5] https://huggingface.co/dchaplinsky/uk_ner_web_trf_best

- **PERIOD** — a time period, which may consist of two dates. Examples: *a few months, 2014-2015*.

- **JOB** — a job title. Examples: *member of parliament, ophthalmologist*.

- **DOC** — a unique name of a document, including names of contracts, orders, bills, purchases. Example: *procurement contract CW2244226*.

- **QUANT** — a quantity with the unit of measurement, such as weight, distance, size. Examples: *3 kilograms, a hundred miles*.

- **ART** (artifact) — a name of a human-made product, like a book, a song, a car, or a sandwich. Examples: *Mona Lisa, iPhone*.

- **MISC** — any other entity not covered in the list above, like names of holidays, websites, battles, wars, sports events, hurricanes, etc. Example: *Black Friday*.

The proposed tagset for entity annotation introduces a list of numerical entities and splits the broad **MISC** class used in the previous version of the corpus into **ART** (e.g., *The Bible*), **JOB** (e.g., *POTUS*), **DOC** (e.g., *Criminal Code of Ukraine*), and **MISC** (everything else).

4.2. Corpus Composition

Seeking to double the NER-UK corpus in size, we searched for a data source that would complement BRUK, already used for the first version of NER-UK, but would be richer in entities and have a more industry-applicable domain. We selected a sample of Nashi Groshi[6] extracted from the UberText 2.0 corpus (Chaplynskyi, 2023) because this website focuses on the Ukrainian economy and anti-corruption efforts. The texts mention a variety of persons and organizations, formal bids and contracts, dates, sums of money, and references to official documents. With this composition, we increased the size of the corpus and the density of entities.

4.3. Annotation Process

We adapted our annotation guidelines to the extended set of entities listed in 4.1, as well as added more examples and corner cases. The annotation guidelines in Ukrainian[7] and English[8] can be accessed through our repository.

To collect entity annotations, we chose the Vulyk crowdsourcing platform[9], with a plugin based on the brat annotation tool (Stenetorp et al., 2012). The plugin allows assigning entity labels to the selected spans of text.

The annotation team consisted of fifteen native speakers of Ukrainian, the majority of whom were students of the Department of Theory, Practice and Translation of German at the National Technical University of Ukraine "Igor Sikorsky Kyiv Polytechnic Institute."

The annotation project was broken into two parts:

1. We pre-annotated the Nashi Groshi subcorpus with our best model[10] trained on the first version of NER-UK for four classes: ORG, PERS, LOC, and MISC. The annotators then corrected the model annotations when necessary and provided new annotations for the remaining nine entity types.

2. The BRUK subcorpus had already been manually annotated as part of the first version of NER-UK. Thus, the task of the annotators was to re-label the MISC entities, since this class was redefined, and provide new annotations for the remaining nine entity types.

Each text within the annotation project was labeled by at least two annotators. The best-performing annotator then manually adjudicated annotation conflicts; the labels presented for adjudication were anonymized in order to prevent potential bias. Throughout the project, we responded to the annotators' feedback and updated examples and corner cases in the guidelines to ensure the high quality and consistency of manual annotations.

5. Results and Discussion

5.1. Corpus Statistics

As a result of the annotation project, the NER-UK 2.0 corpus contains 560 texts boasting 21,993 entities in total. Notably, the number of entities in the BRUK subcorpus increased by 25%, and, like we expected, the Nashi Groshi subcorpus proved to be much richer in entities than BRUK. While BRUK shows an average of 3.9 entities per 100 words, Nashi Groshi quadruples this number to an average of 16 entities per 100 words. A closer inspection of entity type distribution shows that the Nashi Groshi subcorpus contains twenty times more sums of money, five times more organization names, and three times more dates and document names than

[6] https://nashigroshi.org

[7] https://github.com/lang-uk/ner-uk/blob/master/doc/README.md

[8] https://github.com/lang-uk/ner-uk/blob/master/doc/README_en.md

[9] https://github.com/mrgambal/vulyk

[10] https://huggingface.co/dchaplinsky/uk_ner_web_trf_large

the BRUK subcorpus. On the other hand, the re-annotated BRUK shows 2.5 times more person names and four times as many MISC entities, which we interpret as the effect of its genre diversity.

Table 1 presents the size of the subcorpora and the number of annotated entities. We provide detailed information on the entity type distribution across the subcorpora in Appendix A.

	Texts	Words	Entities
BRUK	262	237,327	9,289
Nashi Groshi	298	79,102	12,704
NER-UK 2.0	560	323,200	21,993

Table 1: The size and the number of annotated entities in the two subcorpora of NER-UK 2.0.

Since NER-UK 2.0 expands the original NER-UK, which already has a dev-test split utilized in the NLP community, we made an extra effort to align the new split with the existing one. The updated dev set contains 391 texts with 15,062 entities, and the updated test split contains 169 texts with 6,931 entities. Both the dev and test sets show an equal proportion of BRUK and Nashi Groshi subcorpora; the distribution of entities in the dev and test is also very similar. We provide detailed information on the entity type distribution in the dev and test sets in Appendix B.

5.2. Corpus Format

NER-UK 2.0 is released in the Brat Standoff format[11]. This format allows for nested annotations, which came in handy with the introduction of new entity types. The updated annotation guidelines allowed for nesting of certain entity types. The most frequent examples of nesting include time periods (PERIOD) that may contain two separate DATE entities and organization names (ORG) that may contain a person name (PER).

The code released together with the dataset can be used to convert the corpus into the IOB (Ramshaw and Marcus, 1995) and BEIOS (Jie et al., 2021) formats, discarding the nested annotations, to be used with the systems that do not handle nesting.

5.3. Inter-Annotator Agreement

While most annotation tasks rely on Cohen's Kappa (Cohen, 1960) for measuring the inter-annotator agreement (IAA), previous research (Grouin et al., 2011) argues that for NER annotations, Cohen's Kappa is not the most relevant measure because it relies on the number of negative examples, which is unknown for named entities. Another limitation originates from the nested nature of the annotations in our corpus, which makes it impossible to use Cohen's Kappa on the token level or F_1 score as was proposed by Brandsen et al. (2020). Instead, we report IAA as follows:

$$IAA = A_m/(A_m + A_d),$$

where A_m denotes the number of fully matched annotations and A_d the number of differing annotations between the two sets of annotated entities per document. With the proposed metric, we calculated IAA for each document in the corpus and report the average IAA of 0.84.

5.4. New NER Baseline

To assess the quality of NER with NER-UK 2.0, we trained a classifier using the Ukrainian version of the previously mentioned roberta-large model. The model was trained with the spaCy framework[12] using nearly-default configuration for the spaCy NER task on transformers, except we set hidden_width to 128 and learn_rate to 1e-5 with no warmup. This configuration was identical to the one we used to train the SOTA model on the previous version of the dataset to ensure fair comparison.

The four entity types present in the original NER-UK corpus outline the space for comparison. Table 2 shows that with NER-UK 2.0, the quality improved for LOC and ORG, while the quality of MISC recognition dropped drastically. However, the results of the MISC recognition are not directly comparable since the definition of this class was redefined in NER-UK 2.0.

With regard to all thirteen entity types, the model showed the precision of 0.9 and recall of 0.89. The model learned to recognize persons, locations, organizations, and most numerical entities well but showed much worse results for MISC (0.35 F_1), DOC (0.44 F_1), and TIME (0.6 F_1). While DOC and TIME are simply too infrequent in the corpus, the low quality of recognition for MISC may lie in the broad definition of this entity type. See Appendix C for the full report on precision, recall, and F_1 for each entity type.

The model is available for download at our Hugging Face hub[13].

6. Conclusion

In this paper, we presented a new corpus for Ukrainian named entity recognition called NER-UK 2.0. The corpus was manually annotated for thirteen entity types, most of which are not available

[11] https://brat.nlplab.org/standoff.html

[12] https://spacy.io

[13] https://huggingface.co/dchaplinsky/uk_ner_web_trf_13class

	NER-UK 1.0			NER-UK 2.0		
Entity Label	Prec.	Recall	F_1	Prec.	Recall	F_1
PERS	0.960	**0.974**	**0.967**	**0.961**	0.966	0.963
ORG	0.806	0.782	0.794	**0.940**	**0.896**	**0.917**
LOC	0.914	0.878	0.896	**0.923**	**0.911**	**0.917**
MISC	**0.833**	**0.688**	**0.753**	0.393	0.324	0.355
Weighted Avg.	0.920	0.928	0.913	0.898	0.886	0.892

Table 2: Performance of the roberta-large model for the four original entity types. The model was trained and tested on each version of NER-UK separately.

in standard corpora, and contains 21,993 entities in total. Such a rich set of entities and the variety of genres used as source texts make the corpus invaluable for training named-entity recognition models in various domains.

The retraining of our previous SOTA model on the new corpus showed improvement in recognition quality on two out of three core entity types: organization and location. The model reached the average level of 0.89 F_1. The flexibility of the annotation scheme allows to remove or merge some entity types to train new models for a particular task at hand. We leave further experimentation, like fine-tuning of large language models on NER-UK 2.0, for future work.

The corpus is the largest of its kind for the Ukrainian language and is available for download in the Brat Standoff and IOB formats.

7. Limitations and Ethical Considerations

We acknowledge the following limitations of the NER-UK 2.0 dataset:

- A substantial part of the corpus originates from a single source — Nashi Groshi. While these texts are rich in entities, providing models with ample training data, they may also create a certain level of bias.
- The corpus includes texts written after 2010 and has no samples from earlier times.
- A few entity types, like DOC and TIME, are too infrequent to be used for model training/testing.
- The definition of the MISC entity is too broad to be useful.

The authors verified that the corpus contains no personally identifiable information.

The authors acknowledge using Grammarly for paraphrasing and revision in the process of writing this paper.

8. Acknowledgements

We express our gratitude to Lesia Ivashkevych and the students of the Department of Theory, Practice and Translation of German at the National Technical University of Ukraine “Igor Sikorsky Kyiv Polytechnic Institute” — Anhelina Kovalska, Renata Voroshylova, Anastasiia Lukina, Kateryna Rybachok, Maria Kovalchuk, and others — for their help in annotating the entities for NER-UK 2.0. We are grateful to Pavlo Kulyk for providing constant feedback, highlighting inconsistencies, and ensuring the high quality of annotations.

9. Bibliographical References

Rami Al-Rfou, Vivek Kulkarni, Bryan Perozzi, and Steven Skiena. 2014. POLYGLOT-NER: Massive multilingual named entity recognition.

Ahmed Aliwy, Ayad Abbas, and Ahmed Alkhayyat. 2021. Nerws: Towards improving information retrieval of digital library management system using named entity recognition and word sense. *Big Data and Cognitive Computing*, 5:1–16.

Alex Brandsen, Suzan Verberne, Milco Wansleeben, and Karsten Lambers. 2020. Creating a dataset for named entity recognition in the archaeology domain. In *Proceedings of the Twelfth Language Resources and Evaluation Conference*, pages 4573–4577, Marseille, France. European Language Resources Association.

Dmytro Chaplynskyi. 2023. Introducing UberText 2.0: A corpus of Modern Ukrainian at scale. In *Proceedings of the Second Ukrainian Natural Language Processing Workshop (UNLP)*, pages 1–10, Dubrovnik, Croatia. Association for Computational Linguistics.

Jacob Cohen. 1960. A coefficient of agreement for nominal scales. *Educational and Psychological Measurement*, 20:37 – 46.

Dirk Goldhahn, Thomas Eckart, and Uwe Quasthoff. 2012. Building large monolingual dictionaries at the Leipzig corpora collection: From 100 to 200 languages. In *Proceedings of the Eighth International Conference on Language Resources and Evaluation (LREC'12)*, pages 759–765, Istanbul, Turkey. European Language Resources Association (ELRA).

Cyril Grouin, Sophie Rosset, Pierre Zweigenbaum, Karën Fort, Olivier Galibert, and Ludovic Quintard. 2011. Proposal for an extension of traditional named entities: From guidelines to evaluation, an overview. In *Proceedings of the 5th Linguistic Annotation Workshop*, pages 92–100, Portland, Oregon, USA. Association for Computational Linguistics.

Liu Jie, Pang Yihe, Zhang Kai, Liu Lizhen, and Yu Zhengtao. 2021. A novel dual pointer approach for entity mention extraction. *Chinese Journal of Electronics*, 30:127–133.

Chenguang Liu, Yongli Yu, Xingxin Li, and Peng Wang. 2021. Named entity recognition in equipment support field using tri-training algorithm and text information extraction technology. *IEEE Access*, 9:126728–126734.

Iuliia Makogon and Igor Samokhin. 2022. Targeted sentiment analysis for Ukrainian and Russian news articles. In *ICTERI 2021 Workshops*, pages 538–549, Cham. Springer International Publishing.

Christopher Manning, Mihai Surdeanu, John Bauer, Jenny Finkel, Steven Bethard, and David McClosky. 2014. The Stanford CoreNLP natural language processing toolkit. In *Proceedings of 52nd Annual Meeting of the Association for Computational Linguistics: System Demonstrations*, pages 55–60, Baltimore, Maryland. Association for Computational Linguistics.

Xiaoman Pan, Boliang Zhang, Jonathan May, Joel Nothman, Kevin Knight, and Heng Ji. 2017. Cross-lingual name tagging and linking for 282 languages. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1946–1958, Vancouver, Canada. Association for Computational Linguistics.

Chengwei Qin, Aston Zhang, Zhuosheng Zhang, Jiaao Chen, Michihiro Yasunaga, and Diyi Yang. 2023. Is ChatGPT a general-purpose natural language processing task solver? In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 1339–1384, Singapore. Association for Computational Linguistics.

Lance Ramshaw and Mitch Marcus. 1995. Text chunking using transformation-based learning. In *Third Workshop on Very Large Corpora*.

Vasyl Starko and Andriy Rysin. 2023. Creating a POS gold standard corpus of Modern Ukrainian. In *Proceedings of the Second Ukrainian Natural Language Processing Workshop (UNLP)*, pages 91–95, Dubrovnik, Croatia. Association for Computational Linguistics.

Pontus Stenetorp, Sampo Pyysalo, Goran Topić, Tomoko Ohta, Sophia Ananiadou, and Jun'ichi Tsujii. 2012. brat: a web-based tool for NLP-assisted text annotation. In *Proceedings of the Demonstrations at the 13th Conference of the European Chapter of the Association for Computational Linguistics*, pages 102–107, Avignon, France. Association for Computational Linguistics.

Shuhe Wang, Xiaofei Sun, Xiaoya Li, Rongbin Ouyang, Fei Wu, Tianwei Zhang, Jiwei Li, and Guoyin Wang. 2023. GPT-NER: Named entity recognition via large language models. *arXiv preprint arXiv:2304.10428*.

Gezheng Xu, Wenge Rong, Yanmeng Wang, Yuanxin Ouyang, and Zhang Xiong. 2021. External features enriched model for biomedical question answering. *BMC Bioinformatics*, 22.

A. Entity Type Distribution in NER-UK 2.0 Subcorpora

Entity Label	BRUK	Nashi Groshi	Total
ART	316	**319**	635
DATE	551	**1,496**	2,047
DOC	34	**108**	142
JOB	638	**1,344**	1,982
LOC	**1,620**	1,380	3,000
MISC	**413**	102	515
MON	46	**897**	943
ORG	782	**4,431**	5,213
PCT	77	**186**	263
PERIOD	255	**341**	596
PERS	**4,415**	1,820	6,235
QUANT	106	**276**	382
TIME	**36**	4	40
Total	9,289	12,704	21,993

B. Entity Type Distribution in NER-UK 2.0 Dev and Test Sets

Entity Label	Dev Set	Test Set	Total
ART	398	237	635
DATE	1,448	599	2,047
DOC	102	40	142
JOB	1,323	659	1,982
LOC	2,179	821	3,000
MISC	373	142	515
MON	618	325	943
ORG	3,665	1,548	5,213
PCT	173	90	263
PERIOD	411	185	596
PERS	4,049	2,186	6,235
QUANT	293	89	382
TIME	30	10	40
Total	15,062	6,931	21,993

C. Performance of roberta-large Trained and Tested on NER-UK 2.0

Entity Label	Precision	Recall	F_1
ART	0.703	0.907	0.792
DATE	0.901	0.928	0.914
DOC	0.609	0.350	0.444
JOB	0.729	0.674	0.700
LOC	0.923	0.911	0.917
MISC	0.393	0.324	0.355
MON	0.968	0.942	0.955
ORG	0.940	0.896	0.917
PCT	1.000	0.989	0.994
PERIOD	0.777	0.773	0.775
PERS	0.961	0.966	0.963
QUANT	0.890	0.910	0.900
TIME	0.667	0.600	0.632
Weighted avg.	0.898	0.886	0.892

Instant Messaging Platforms News Multi-Task Classification for Attitude, Sentiment, and Discrimination Detection

Denilson Barbosa[1], **Taras Ustyianovych**[2]
[1]Department of Computing Science, University of Alberta, Canada,
[2]Department of Artificial Intelligence Systems, Lviv Polytechnic National University, Ukraine
[1]denilson@ualberta.ca, [2]taras.o.ustyianovych@lpnu.ua

Abstract

In the digital age, geopolitical events frequently catalyze discussions among global web users. Platforms such as social networks and messaging applications serve as vital means for information spreading and acquisition. The Russian aggression against Ukraine has notably intensified online discourse on the matter, drawing a significant audience eager for real-time updates. This surge in online activity inevitably results in the proliferation of content, some of which may be unreliable or manipulative. Given this context, the identification of such content with information distortion is imperative to mitigate bias and promote fairness. However, this task presents considerable challenges, primarily due to the lack of sophisticated language models capable of understanding the nuances and context of texts in low-resource languages, and the scarcity of well-annotated datasets for training such models. To address these gaps, we introduce the TRWU dataset – a meticulously annotated collection of **T**elegram news about the **R**ussian **w**ar in **U**kraine gathered starting from January 1, 2022. This paper outlines our methodology for semantic analysis and classification of these messages, aiming to ascertain their bias. Such an approach enhances our ability to detect manipulative and destructive content. Through descriptive statistical analysis, we explore deviations in message sentiment, stance, and metadata across different types of channels and levels of content creation activity. Our findings indicate a predominance of negative sentiment within the dataset. Additionally, our research elucidates distinct differences in the linguistic choices and phraseology among channels, based on their stance towards the war. This study contributes to the broader effort of understanding the spread and mitigating the impact of biased and manipulative content in digital communications.

Keywords: News Messages, Dataset, Text Classification, Destructive content detection

1. Introduction

The proliferation of internet and web technologies has had an impact on public discourse, shaping opinions and perceptions. With a wealth of data available from diverse sources, ranging from factual information to personal opinions, navigating this informational landscape can be daunting (Adams et al., 2023; Mendoza et al., 2023). Therefore, the exploitation of information literacy and critical thinking can distort public understanding and opinion (Aslett et al., 2023). Traditional technological tools have proven difficult in addressing these complex challenges (Zakharchenko et al., 2021).

The complexity of discerning opinions from objective facts is compounded in politically charged scenarios, such as the Russian invasion of Ukraine in February 2022. The narratives surrounding such events do not merely shape public morale but also influence mental health, beliefs, and international perspectives on credibility and support (Haq et al., 2022). In this context, the automated classification of content based on its biases becomes a pivotal tool for fostering a more informed and trustworthy Web environment (Meel and Vishwakarma, 2020). Previous efforts have explored various computational approaches to address these challenges, including classification (Solopova et al., 2023), text summarization (Galeshchuk, 2023b), and topic modeling (Ustyianovych et al., 2023), particularly in Ukrainian and Russian contexts (Galeshchuk, 2023a). However, the development of robust, explainable, and efficient models capable of accurately identifying the biases of textual content remains a pressing and relevant challenge. Such models not only aid in filtering and understanding content but also play a vital role in educating users about the nuances of misleading information. Communication strategies and linguistics constantly evolve with new approaches developed to interact with and address the target audience. Therefore, technological means for processing and understanding natural language and communication contexts need to remain up to date to keep up with current issues.

Our research contributes to this field by presenting a novel annotated dataset related to the Russian aggression against Ukraine with a multi-task transformer-based model trained to identify geopolitical stance, sentiment, and the presence of hate or discrimination in the input message. By leveraging the capabilities of large language models (LLMs), we delve into the intricacies of textual data, seeking to unveil patterns that distinguish biased narratives. Our findings underscore the potential of these technologies to enhance our comprehension

25 May, 2024. © 2024 ELRA Language Resource Association: CC BY-NC 4.0

of biased content and, by extension, to promote a nuanced and critical engagement with information in the digital age.

2. Related Work

The study of information campaigns in digital environments has become increasingly pertinent with the advent of social networks and web technologies. These platforms are not solely conduits for the spreading of factual information; they also serve as sites for strategic communications aimed at influencing public opinion and garnering support within online communities. An illustrative example of how digital platforms can be utilized for such purposes is observed in the analysis of various information campaigns, including those conducted on social media platforms (Courchesne et al., 2022). This study investigates the dynamics of online activity associated with significant geopolitical events, highlighting the capacity of strategic communication efforts to engage with and influence digital communities. The analysis, which encompasses a broad dataset of social media accounts, reveals a marked increase in online activity coinciding with pivotal events and underscores the effectiveness of coordinated information dissemination strategies in capturing public attention and shaping narrative discourse.

The comprehensive examination of these social media activities, including a study of over 126 thousand accounts, illustrates the challenges faced by content moderation teams and the sophisticated nature of modern information campaigns. Such studies highlight the complexity of digital information verification and the need for advanced methodologies to understand and navigate the intricacies of information manipulation in the digital age. A recent study by Park et al. (2022) describes the VoynaSlov dataset that was collected from two social networks, Twitter and VKontakte, to analyze and detect media opinion manipulations related to the Russian war in Ukraine. It consists of more than 38 million posts based on Russian media statements and expressions. The authors focus on distinguishing sources into state-affiliated and independent. As expected, the usage of words and phases differs between these two categories along with the formed topics distribution. The study results highlight a spike in user engagement and the number of generated posts after the invasion began on February 24, 2022. This observation confirms how real-world events engage users in online activity and content creation.

Fedushko et al. (2023) proposed innovative methods to support real-time decision-making about antagonistic user behavior on social networks. The proposed techniques showed significant results in decreasing the number of destructive content generated and shared, which contributed to more sustainable interactions in online communication. The developed models consider decisions, information environment, and decision-making criteria as the key processes for online community management. The methods were validated on a Facebook online community and showed an increase in user participation (and community size) in just one month after implementing the strategy for sustainable community development, indicating that it is possible to alert and guide users about the dangers of posting destructive comments online.

Threat detection in Web communication is another aspect that is worth attention and can be tackled with AI- and data-driven technologies. Semantic analysis combined with communication behavioral models is already successfully used to handle threats in social media discussions. Fedushko and Benova (2019) suggests a process for performing users' semantic analysis in an online environment, which improves the efficiency of threat detection by up to 40%.

Since a large number of discussions occur on social media platforms, it is crucial to understand the formed trends and patterns, especially in the context of specific subjects and objects. Visualization techniques might be efficiently used to investigate opinions and perform social communications mining. These methods were successfully used to analyze opinions appertained to such topics: 1) energy sources and 2) social network brands of academic institutions (Gutierrez et al., 2021). Our dataset and model contribute to the area of social media and instant messages analysis in order to have a full picture of the public stance towards specific topics, including sensitive ones.

Transformers and large language models have been effectively applied for the detection of unreliable information within news and online content. A case in point is the HQP dataset specifically collected to facilitate the identification of misinformation by incorporating 30 thousand tweets related to the war between Russia and Ukraine. This dataset is notable for its differentiated labeling approach, categorizing data into "high-quality" and "weak" labels. High-quality labels are distinguished by their validation through human review, ensuring the trustworthiness and accuracy of the data. In contrast, weak labels lack human validation, presenting a potential challenge to model accuracy (Maarouf et al., 2023). The methodology adopted for data labeling in the HQP dataset, and the subsequent application of pre-trained language models, showcases the critical role of high-quality labels in enhancing model performance. The achieved results highlight this, with models trained on high-quality labeled data achieving an Area Under the Curve (AUC) score of 92.25. This outcome indicates a significant improvement in the model's ability to detect untrustworthy content accurately, highlighting the importance of rigorously validated data during the design of effective detection systems.

Applications of few-shot learning and zero-shot classification are other promising areas discussed to improve the detection of harmful content and bias, and puzzle out related information trustworthiness tasks (Nayeon et al., 2021; Liew et al., 2023; Modupe et al., 2023; Yao et al., 2022).

3. Telegram War News Dataset

3.1. Dataset Collection

The Russian-Ukrainian war dataset has been collected from thoroughly selected pro-Russian and pro-Ukrainian Telegram channels. The selection of channels is based on the lists of reliable versus untrustworthy information sources provided by the Ukrainian Center for Countering Disinformation (for Countering Disinformation, 2022) and the Institute of Mass Information (of Mass Information, 2023). Telegram is an instant messaging application with 700 million monthly active users. It offers the option to create channels for broadcasting content to large audiences. Each message can contain media content, which makes it suitable for multimodal news analysis (Wang et al., 2022b). Users in a channel can leave comments and reach with emojis, which leads to another exciting area of research – online user engagement and behavior analysis (Fedushko et al., 2020). Telegram has an open API for extracting data from specific channels (based on their ID). We collected data from six news and blog-like channels regularly posting content about the Russian war against Ukraine. Statistics on the number of messages retrieved from each channel are given in Table 1.

The total number of messages collected is 252,677 from January 1st, 2022 until December 14th, 2023. At the time of writing, new data is being collected for further processing. Each message contains the channel name, timestamp, message ID for the selected channel, and the text of the message itself. The messages have been labeled using the `gpt-3.5-turbo-1106` large language model with a human-in-the-loop to ensure the reliability of the assigned labels. Additional data validation and normalization were accomplished to standardize the labels and meet the actual research purpose. Messages are labeled according to the channel's attitude mentioned in Table 1 and randomly split into training, validation, and testing sets with such percent ratios: 90%, 5%, and 5% correspondingly.

3.2. Dataset Statistics

The uniqueness of our dataset primarily derives from its comprehensive compilation process and focused applicability to the Russian-Ukrainian war. Unlike conventional datasets that predominantly source from widely used social media platforms like Twitter and Facebook, our dataset uniquely taps into the Telegram instant messaging platform. This choice was deliberate, given Telegram's distinct user base and communication style, which significantly differ from other platforms. Telegram channels offer a rich amount of data in varied tones—ranging from news and factual reports to blog posts and opinion pieces. This diversity not only improves the dataset but also makes it exceptionally versatile for Natural Language Processing (NLP) research, promoting a broad exploration of communication techniques and content types.

A pivotal aspect of our dataset's development was a thorough selection of sources, ensuring that each included channel introduced a clear stance (pro-Russian or pro-Ukrainian) regarding the war. This careful curation process guarantees the dataset's relevance and validity for studies focusing on sentiment analysis, manipulative content detection, and the examination of targeting tactics. Our research aims to analyze the sentiment of messages from a pro-Ukrainian perspective. It's important to consider that the same message can be interpreted differently by audiences based on their viewpoints and backgrounds.

Further distinguishing our dataset is the use of GPT-3.5 for initial labeling, tasked with extracting sentiment and filtering out irrelevant content. This step was augmented by human validation to ensure the accuracy and reliability of the labels assigned by the AI, addressing potential biases and inaccuracies inherent in automated processes.

Our motivation to create this dataset facilitates a nuanced analysis of communication patterns, enabling researchers to identify harmful and misleading content effectively. Its applicability extends to improving government accounting information systems, as demonstrated by related studies, showcasing its potential to influence a wide range of fields positively (Duan et al., 2023). By carefully curating, labeling, and validating our dataset, we have created a resource that stands out for its methodological stringency and direct relevance to current geopolitical events, offering invaluable insights into the dynamics of information dissemination and reception in the digital age.

According to Table 1, 152,502 (55.19%) of the content is retrieved from pro-Russian sources, whereas 123,812 (44.80%) entities belong to pro-Ukrainian channels. All the channels' sentiment most frequent value except *rian_ru* is negative, and for the latter it is neutral. A histogram with the col-

Channel	Stance	Count	Fraction	Mean token count	Mode sentiment
rian_ru	Pro-Russian	79,663	28.83%	28.26	neutral
ROSSIYA_SEGODNIA	Pro-Russian	69,238	25.05%	55.16	negative
uniannet	Pro-Ukrainian	67,727	24.51%	48.58	negative
radiosvoboda	Pro-Ukrainian	33,225	12.02%	108.63	negative
UkrPravdaMainNews	Pro-Ukrainian	22,860	8.27%	46.34	negative
ZE_kartel	Pro-Russian	3,601	1.30%	74.91	negative

Table 1: Number and percentage of messages per channel

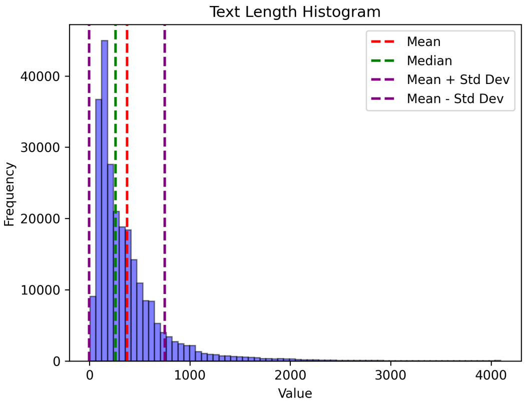


Figure 1: Text Length Histogram with mean, median, and standard deviation ranges.

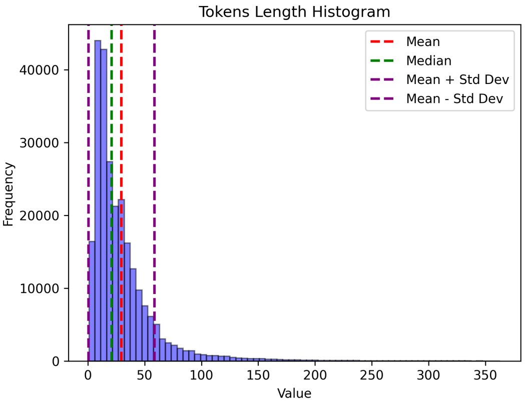


Figure 2: Histogram for the number of tokens with mean, median, and standard deviation ranges.

lected data text length is shown in Figure 1. The mean and median values are 373.47 and 259, respectively, and the standard deviation is 376.63. Also, 89.79% of the messages are below one standard deviation from the mean, meaning that their length is less than or equal to 750.10. After text preprocessing, the mean text length was reduced by 32%.

We provide a summary of the number of remaining tokens per message after applying preprocessing, which includes text cleaning, stopword removal, and lemmatization. The mean and median token count after text preprocessing are 29.52 and 21 respectively, and the standard deviation is 29. The obtained distribution pattern is similar to the one for text length and is shown in Figure 2.

3.3. Semantic Analysis

Analyzing data on the semantic level is crucial to extracting meaning from the text and understanding the critical features of the studied sources concerning word usage and style. To create a general comprehension of the text data after cleaning and lemmatization, we identified the most frequently used words: "Ukraine", "Russian", "warlike", "claim", "connection", "USA", "Putin", "Zelenskyi", "offensive", "sanction", "destroy", "weapon".

There are 92,342 and 118,870 unique entities used in pro-Russian and pro-Ukrainian channels, respectively. This is an exciting finding since there are far more pro-Russian messages; nevertheless, the word usage within pro-Ukrainian sentences is significantly richer.

We observed a “separation” in vocabulary between the two sides: 57.06% of the unique words used by the pro-Ukrainian sources do not appear in pro-Russian channels; within Pro-Russian channels, this rate is 44.72%. We analyzed unique words within each side and found that they mainly include derogatory named entities against the opposite side, abbreviations, local areas and regions, and words with local and specific meanings. For example, unique words from pro-Russian channels contain the character "Z" which is known to be their symbol of the war. Also, the word "war" itself is replaced by "special military operation". Some pro-Ukrainian publications might contain Ukrainian words even though the text piece is written in Russian. This factor contributes to the number of unique words used between the sources and can help our model differentiate between these originating sources. The data presented in the figure 3 compares sentiment classification results from an automated method using OpenAI API `gpt-3.5-turbo-1106` model with human validation. The

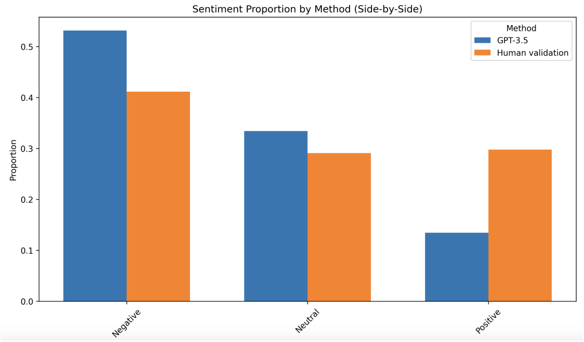


Figure 3: Sentiment proportion by method

figure illustrates the proportion of messages categorized as negative, neutral, and positive. The AI-based sentiment analysis results show that the majority of the dataset, 53.13%, exhibits a negative sentiment. Neutral sentiments, which may represent unbiased reporting, factual statements, or ambiguous content, constitute 33.30% of the dataset. Positive sentiments, indicative of optimism and supportive statements, account for 13.45%. A small fraction of the data (not represented on the figure), merely 0.12%, is categorized under mixed sentiments, highlighting texts that possibly contain conflicting emotions or viewpoints. In contrast, human validation results based on a randomly selected sample of labeled messages, are in a different sentiment distribution. While the proportion of negatively classified messages is similar to the GPT-3.5 results, human validation assigns a significantly lower proportion as neutral and a considerably higher proportion as positive. The discrepancy, specifically in the positive category, may be due to the detailed and contextual understanding that human validators bring to the task, which automated systems like GPT-3.5 may not fully capture, especially in the complex and sensitive context of war-related communications. The figure highlights the critical role of human oversight in sentiment analysis and the importance of multimodal validation for sensitive topics. The obtained sentiment values correspond with the stance and source channel of the message. The sentiment distribution underscores the complexity and variability of the sentiments expressed in the Web communication, offering valuable insights into the ruling attitudes and perceptions within the collected data. However, it should be noted that the AI-based labels offered sentiment classification from a prospective without favoritism to any side of the war. A refined version of the prompt with few-shot learning might improve the obtained results and make them suitable to identify the sentiment according to specific requirements.

We highlight the need to employ entity-level sentiment detection since distinct sentiments can be assigned to multiple entities represented in a piece of text (Rønningstad et al., 2022). This approach would contribute to the identification of the message's stance toward the war, and provide insights on the named entities represented within the text. Also, it offers a multi-faceted sentiment analysis compared to examining data from a single perspective.

Figure 4 shows 7-day window rolling average sentiment values by channel's attitude and applied methodology to detect sentiment over time. The usage of the rolling average sentiment score smooths out the noise, providing a clear view of the overall trends over time. Pro-Russian channels are represented with mostly negative sentiment scores according to the GPT-3.5 classification throughout the observed period. The sentiment scores for pro-Russian channels (shown in blue) demonstrate changes over time but with a generally less pronounced variance. In contrast, the sentiment scores for pro-Ukrainian channels (shown in orange) appear to follow a similar trend, maintaining lower average sentiment values compared to their pro-Russian counterparts. However, the sentiment values for pro-Ukrainian channels also fluctuate, suggesting that external factors and evolving news dynamics impact them. Therefore, when evaluating sentiment with technological means, it is important to consider the biases of these analytical tools being used. Our research results show that the GPT-3.5 model tends to interpret themes of war and conflict with a negative sentiment despite the evidence that some messages might be perceived differently by specific users. This is supported by the consistently negative sentiment scores given to massages of both channel viewpoints throughout the period studied. So, our finding highlights the importance of method selection in sentiment analysis studies and underscores the value of multiple analytical approaches comparison for a comprehensive view of trends in digital communication. Nevertheless, the employed GPT-based method proves the sentiment scores are mainly negative due to the nature of events.

3.4. Challenges and Limitations

Detecting biased, misleading, and manipulative content in such a dynamic environment as instant messaging platforms or social media is challenging because new data gets generated and shared in real-time, forming patterns unseen in historical data. So, usage of methods like incremental learning (Shan et al., 2020; Abdalla et al., 2022; Barve et al., 2022; Wang et al., 2022a) and well-established ML operations processes are becoming extremely helpful in these scenarios (Shukla and Cartlidge, 2022; Jarrahi et al., 2023; Mäkinen et al., 2021).

Furthermore, there is very little properly and publicly available labeled data to identify such con-

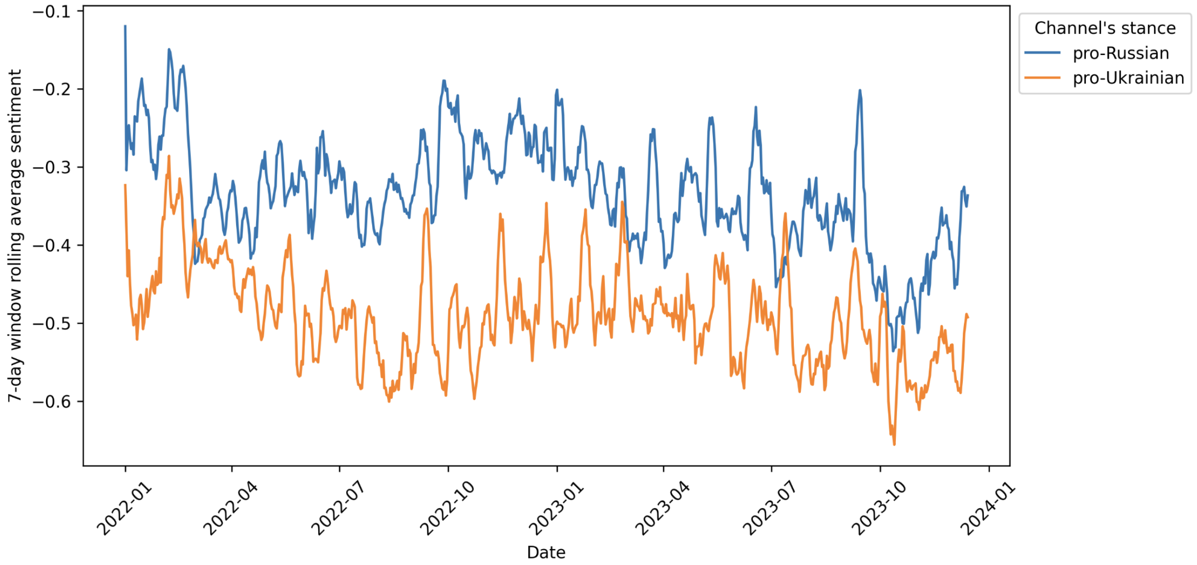


Figure 4: 7-day window rolling average GPT-3.5 sentiment score per channel stance

tent in the context of the Russian-Ukrainian war. Ukrainian is considered a low-resource language, with few available tools and models (Gomez et al., 2023). Selecting relevant sources and designing comprehensive labeling methods is crucial for developing high-performing models in the future. With that in mind, care was taken to collect data that was clearly associated with either side of the war and represented the respective attitude in their Web publications. For instance, a Ukrainian-based Telegram channel *UkrPravdaMainNews* was chosen because it posts pro-Ukrainian news, whereas the Russian-based channel *rian_ru*, which is part of a well-known Russian news agency, was chosen because it contains pro-Russian publications. The application of GPT-3.5 to assign such labels as geopolitical stance, sentiment, and presence of discrimination in an input text allowed us to identify the most relevant to the subject matter messages. Additional human validation, which included exploratory data analysis and verification of the assigned labels, significantly improved the dataset.

Additionally, the usage of machine learning models, DNNs (deep neural networks) as well as statistical methods can confirm whether there is a statistical difference between these and other examined labels. Topic models can be applied to categorize the data in an unsupervised manner and provide insights about the subject matters they contain.

Our dataset contains both pro-Ukrainian and pro-Russian texts written in the Russian language. However, there is a shortage of pro-Russian publications in Ukrainian which complicates achieving the defined goal for this language. Data augmentation methods including transformer-based translation might become handy to overcome this challenge (Liu et al., 2023; Gong et al., 2022). This is also a promising way to develop a multilingual model in further research.

About 40% of the messages in the dataset are not related to the war between Russia and Ukraine, which has been identified with GPT-3.5 zero-shot classification and human verification. Controlling the percentage of these entities is crucial to keep an optimal balance between relevant and extraneous messages in order to accomplish the modeling part. So, employing means to denoise the data and extract the most informative samples is crucial to reach the target of this study.

4. Data Processing

The whole workflow is depicted in Figure 5. The diagram outlines a multi-stage process for analyzing and processing text data from Telegram messages, aimed at evaluating the dataset's predictive capabilities with conventional machine learning methods and fine-tuning language models. The former technique involved input text cleaning and preprocessing using spaCy `ru_core_news_lg` and `ua_core_news_lg` language pipelines, creating word embeddings with fastText, and vectors manipulation. The fastText model was trained with the following parameters: vector size of 300, window size of 5, minimum word frequency of 3, training algorithm was skip-gram, ten epochs, and four worker threads. Then, the formed word embeddings were passed to the XGBoost classifier for hyperparameter tuning and evaluation. The data processing required for performing the NLP transformer-based approach consisted of such steps: text data cleaning, prompts generation for zero-shot classification, extraction and standardization of the LLM's output, and data unification. The formed dataset contained the text messages with corresponding Telegram metadata (ID, datetime, channel) and assigned labels by `gpt-3.5-turbo-1106`. The following prompt instructions were provided to the large language model: "Analyze the following messages related to the war between Ukraine and Russia. For each message: 1. Determine the sentiment (positive, negative, neutral, etc.) expressed in the message. 2. Identify geopolitical attitude or hate/discrimination and in favor of what side it is expressed: indicate whether it's pro-Ukrainian, pro-Russian, or any other geopolitical stance. Take into account that messages might contain glorification, hate, and discrimination, which should be considered when classifying attitudes. 3. If the message lacks a geopolitical attitude or isn't related to the conflict, mark it as not applicable to geopolitical attitude. The output should be returned as a Python dictionary array with such keys: message ID, sentiment, detected favorable attitude, and whether a message contains hate or discrimination (yes or no)". Human validation was accomplished afterward to ensure data quality, standardized values for categorical variables, and accurate annotation. We employed exploratory data analysis of the GPT-based labels to find and correct abnormal or unexpected values, gather statistics, and correlate them to find mislabeled entities. A sample of the data was taken for manual validation, and accuracy scores between human and AI-based labels were calculated. The obtained human validation results show mediocre performance in determining the proper sentiment in the context of events like a war. On the other hand, the AI agent did more than 80% correct on the geopolitical attitude and identifying irrelevant content. The data was passed as input to language models for fine-tuning. The returned outputs by the AI-based agent were transformed and converted into a pandas data frame and joined with the original dataset to make it suitable for model training. This workflow is crucial for the methodological processing of raw Telegram messages into valuable information assets through advanced NLP techniques. Each step of the presented workflow is designed to enhance the overall predictive performance and capabilities of the models.

5. Text Classification

The modeling part was performed on the collected Telegram War News dataset, first to assess its predictive performance using the XGBoost classifier and, second, to build a robust multi-task language model capable of distinguishing between pro-Ukrainian and pro-Russian messages, their sentiment, and stance. Such a model will become extremely helpful in mitigating the consequences of bias and misleading content spreading through Internet resources with specific attitudes.

5.1. Experimental Settings

We conducted hyperparameter optimization targeting the Area Under the Curve (AUC) score, complemented by a 3-fold cross-validation strategy for the XGBoost classifier on the training dataset. The evaluation of the optimized model was carried out on a separate testing set. Each input document was represented as a 300-dimensional vector. The search for optimal hyperparameters utilized the `hyperopt` package, with a defined search space that included the maximum depth of trees, learning rate, fraction of data used per iteration, minimum weight of child nodes, gamma as the regularization parameter, subsample ratio of features for constructing each tree, and the type of boosting model employed. We conducted a total of 35 trials, with the Tree of Parzen Estimators (TPE) algorithm chosen for the optimization process.

Fine-tuning of the multi-task language models was executed on computing instances equipped with NVIDIA Tesla V100 GPUs. We utilized the `google/mt5-base` and `xlm-roberta-base` for their multilingual capabilities in text tokenization and subsequent fine-tuning phases. The training phase involves fine-tuning the models on the multi-variable data frame with a custom PyTorch Dataset instance, which efficiently manages data fetching. The models, specifically MT5EncoderModel and XLMRobertaModel, were adapted with custom adjustments to their output layers and loss computation methods, assigning distinct weights to each predictive variable. The variables for prediction included: the channel's originating source attitude, sentiment, stance, presence of discrimination, combined channel's attitude and sentiment, and a merge of stance and sentiment. Tokenization restricted the text input to a length of 256 tokens. The training process spanned 10 epochs with batch sizes of 64 for both training and evaluation. Evaluation is conducted on a separate validation dataset to assess the model's accuracy and effectiveness in handling both tasks simultaneously, leading to its subsequent deployment for real-world applications.

5.2. Results

The optimal set of hyperparameters to build a robust XGBoost classifier for a message originating channel's attitude was: booster: 'gbtree'; colsample_bytree: 0.99837; gamma: 0.17946; learning_rate: 0.18935; max_depth: 17; min_child_weight: 14; and subsample: 0.89539. The final AUC scores on training and testing sets

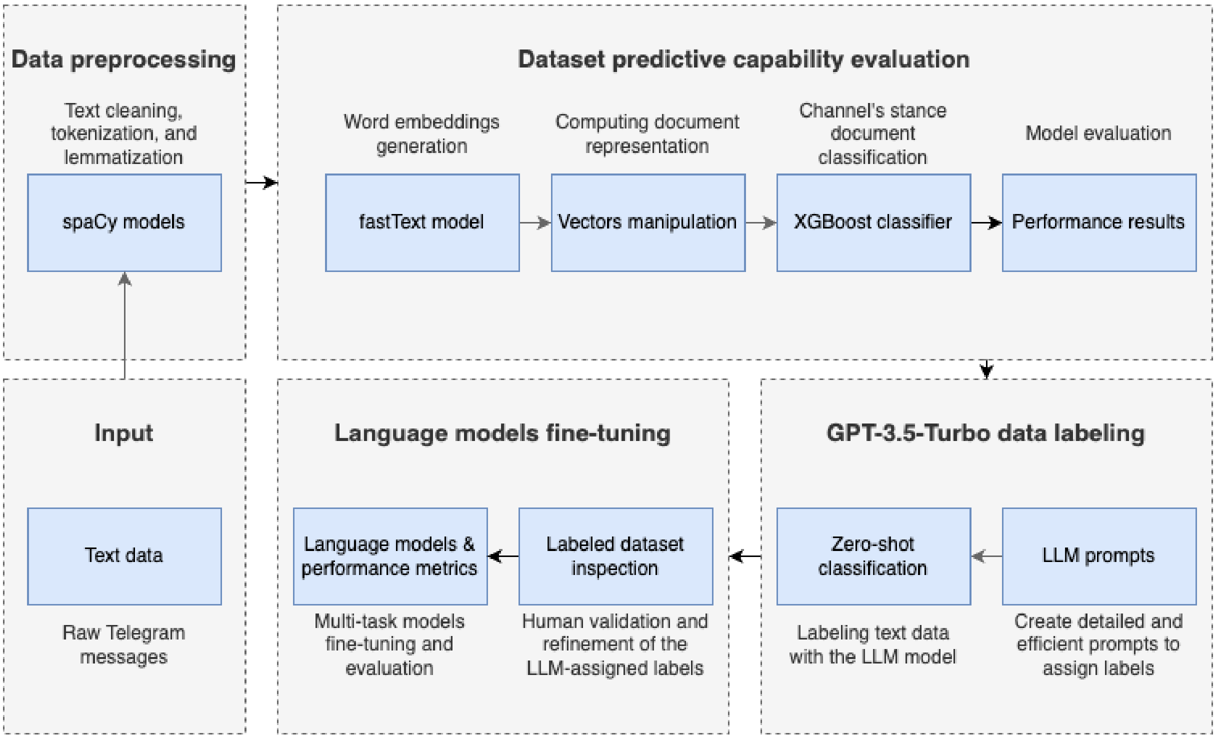

Data preprocessing
Text cleaning, tokenization, and lemmatization
spaCy models
Dataset predictive capability evaluation
Word embeddings generation
Computing document representation
Channel's stance document classification
Model evaluation
fastText model
Vectors manipulation
XGBoost classifier
Performance results
Input
Text data
Raw Telegram messages
Language models fine-tuning
Language models & performance metrics
Labeled dataset inspection
Multi-task models fine-tuning and evaluation
Human validation and refinement of the LLM-assigned labels
GPT-3.5-Turbo data labeling
Zero-shot classification
LLM prompts
Labeling text data with the LLM model
Create detailed and efficient prompts to assign labels

Figure 5: Telegram data inference pipeline

are 0.9715 and 0.9065, respectively. We computed such metrics as accuracy (0.9088), precision (0.9062), recall (0.8862), and F1 score (0.8961) as well.

The multi-task model displayed above-mediocre performance with an average accuracy of 0.74. It effectively identified the originating channel's attitude with a high accuracy of 0.95 and detected the presence of discrimination with an accuracy of 0.94. However, when the model was tasked with simultaneous detection of channel attitude and sentiment, the accuracy slightly reduced to 0.67, and further to 0.51 for combining geopolitical stance and sentiment. This indicates a more challenging scenario when the model is required to discern multiple nuanced aspects concurrently. These results show that while the model exhibits high accuracy with certain individual tasks, particularly in detecting the originating channel attitude and discrimination, there is a trade-off in performance when multitasking on sentiment and geopolitical stance. The obtained results highlight the complexity of multi-faceted analysis and point to opportunities for further improvement in multi-task modeling. It is worth paying detailed attention to data labeling and fine-tuning more complex language models. Also, the application of a single-task classification might improve the performance and design a specific targeted classification tool.

6. Conclusion and Future Work

Our research introduces the TRWU dataset, comprising texts from pro-Ukrainian and pro-Russian Telegram channels, featuring both factual and opinionated content. This dataset's uniqueness lies in its contemporaneous nature and thoroughly selected sources, delivering a comparative analysis of communication patterns. We used text mining to identify key lexical features and word usage across different channels. Our classification pipeline, which integrates spaCy, fastText, and XGBoost, was optimized to predict the stance of messages. We uncovered essential hyperparameters for optimal performance. We used zero-shot classification along with human validation for data labeling. The fine-tuned multi-task language model successfully classified the originating channel's attitude and presence of discrimination. Our findings indicate a need for enhanced sentiment detection tools for Ukrainian and Russian languages.

Future Work. Our proposed future work includes: 1) advancing stance and sentiment classification with rigorous labeling and model fine-tuning; 2) implementing vector databases for efficient document collocation; 3) context-based entity sentiment analysis, especially in conflict-related discourse; 4) pursuing excellence in model performance for both multi-task and single-task objectives; 5) further developing models for low-resources languages like Ukrainian (Laba et al., 2023).

7. Acknowledgements

This work was supported by a grant from the Canadian Institute of Ukrainian Studies (from the Mykola Klid Memorial Endowment Fund).

7.1. Ethical considerations

We are aware that the dataset we collected might contain harmful content because of the nature of the data. We have attempted to be unbiased in collecting the data from the selected channels and have not tried to censor any content. So, we will take respective precautions to warn users of this once the dataset is released. Therefore, ethical considerations are crucial when working this dataset for bias and manipulative patterns detection since content related to subjects like war can be sensitive, distorted, or unfair (Deepak, 2021). We strongly recommend evaluating the results with fairness metrics and using machine learning monitoring to improve observability and awareness of how such systems perform (Ashktorab et al., 2023). Utilizing tools for interpretability and explainability is essential to tackle this challenge and ensure transparency of the models.

8. Bibliographical References

H.B. Abdalla, A.M. Ahmed, S.R.M. Zeebaree, A. Alkhayyat, and B. Ihnaini. 2022. Rider weed deep residual network-based incremental model for text classification using multidimensional features and mapreduce. *PeerJ Computer Science*.

Zoë Adams, Magda Osman, Christos Bechlivanidis, and Björn Meder. 2023. (why) is misinformation a problem? *Perspectives on Psychological Science*, 18(6):1436–1463. PMID: 36795592.

Z. Ashktorab, B. Hoover, M. Agarwal, C. Dugan, w. Geyer, H. B. Yang, and M. Yurochkin. 2023. Fairness evaluation in text classification: Machine learning practitioner perspectives of individual and group fairness. In *Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems*. ACM.

K. Aslett, Z. Sanderson, W. Godel, N. Persily, J. Nagler, and J. A. Tucker. 2023. Online searches to evaluate misinformation can increase its perceived veracity. *Nature*, 625:548–556.

Y. Barve, J. R. Saini, K. Kotecha, and H. Gaikwad. 2022. Detecting and fact-checking misinformation using “veracity scanning model”. *International Journal of Advanced Computer Science and Applications*, 13(2).

L. Courchesne, B. Rasikh, B. McQuinn, and C. Buntain. 2022. Powered by twitter? the taliban’s takeover of afghanistan. ESOC Working Paper 30, Emperical Studies of Conflict.

P. Deepak. 2021. *Ethical Considerations in Data-Driven Fake News Detection*, pages 205–232. Springer International Publishing, Cham.

H. K. Duan, M. A. Vasarhelyi, M. Codesso, and Z. Alzamil. 2023. Enhancing the government accounting information systems using social media information: An application of text mining and machine learning. *International Journal of Accounting Information Systems*, 48:100600.

S. Fedushko and E. Benova. 2019. Semantic analysis for information and communication threats detection of online service users. *Procedia Computer Science*, 160:254–259. The 10th International Conference on Emerging Ubiquitous Systems and Pervasive Networks (EUSPN-2019) / The 9th International Conference on Current and Future Trends of Information and Communication Technologies in Healthcare (ICTH-2019) / Affiliated Workshops.

S. Fedushko, K. Molodetska, and Yu. Syerov. 2023. Decision-making approaches in the antagonistic digital communication of the online communities users. *Social Network Analysis and Mining*.

S. Fedushko, T. Ustyianovych, Yu. Syerov, and T. Peracek. 2020. User-engagement score and slis/slos/slas measurements correlation of e-business projects through big data analysis. *Applied Sciences*, 10(24).

Center for Countering Disinformation. Ccd announces an updated list of infoterrorist channels operating in ukraine [online]. 2022.

S. Galeshchuk. 2023a. Abstractive summarization for the Ukrainian language: Multi-task learning with hromadske.ua news dataset. In *Proceedings of the Second Ukrainian Natural Language Processing Workshop (UNLP)*, pages 49–53, Dubrovnik, Croatia. Association for Computational Linguistics.

Svitlana Galeshchuk. 2023b. Abstractive summarization for the Ukrainian language: Multi-task learning with hromadske.ua news dataset. In *Proceedings of the Second Ukrainian Natural Language Processing Workshop (UNLP)*, pages 49–53, Dubrovnik, Croatia. Association for Computational Linguistics.

Frank Gomez, Alla Rozovskaya, and Dan Roth. 2023. A low-resource approach to the grammatical error correction of ukrainian. In *Proceedings* *of the Second Ukrainian Natural Language Processing Workshop (UNLP)*, pages 114–120.

H. Gong, X. Li, and D. Genzel. 2022. Adaptive sparse transformer for multilingual translation.

C. A. Gutierrez, Whittaker, A. Whittaker, K. M. Patenio, J. Gehman, L. L. M. Lefsrud, D. Barbosa, and E. Stroulia. 2021. Analyzing and visualizing twitter conversations. In *Proceedings of the 31st Annual International Conference on Computer Science and Software Engineering*, CASCON '21, page 4–13, USA. IBM Corp.

E.-U. Haq, G. Tyson, T. Braud, and P. Hui. 2022. Weaponising social media for information divide and warfare. In *Proceedings of the 33rd ACM Conference on Hypertext and Social Media*, HT '22, page 259–262, New York, NY, USA. Association for Computing Machinery.

M. H. Jarrahi, A. Memariani, and S. Guha. 2023. The principles of data-centric ai. *Commun. ACM*, 66(8):84–92.

M. Korobov. 2015. Morphological analyzer and generator for russian and ukrainian languages. In *Analysis of Images, Social Networks and Texts*, pages 320–332, Cham. Springer International Publishing.

Yurii Laba, Volodymyr Mudryi, Dmytro Chaplynskyi, Mariana Romanyshyn, and Oles Dobosevych. 2023. Contextual embeddings for Ukrainian: A large language model approach to word sense disambiguation. In *Proceedings of the Second Ukrainian Natural Language Processing Workshop (UNLP)*, pages 11–19, Dubrovnik, Croatia. Association for Computational Linguistics.

X. Y. Liew, N. Hameed, and J. Closand J. E. Fischer. 2023. Predicting stance to detect misinformation in few-shot learning. In *Proceedings of the First International Symposium on Trustworthy Autonomous Systems*, TAS '23, New York, NY, USA. Association for Computing Machinery.

X. Liu, J. He, M. Liu, Z. Yin, L. Yin, and W. Zheng. 2023. A scenario-generic neural machine translation data augmentation method. *Electronics*, 12(10).

A. Maarouf, D. Bär, D. Geissler, and S. Feuerriegel. 2023. Hqp: A human-annotated dataset for detecting online propaganda.

S. Mäkinen, H. Skogström, E. Laaksonen, and T. Mikkonen. 2021. Who needs mlops: What data scientists seek to accomplish and how can mlops help? *CoRR*, abs/2103.08942.

P. Meel and D.K. Vishwakarma. 2020. Fake news, rumor, information pollution in social media and web: A contemporary survey of state-of-the-arts, challenges and opportunities. *Expert Systems with Applications*, 153:112986.

Marcelo Mendoza, Sebastián Valenzuela, Enrique Núñez-Mussa, Fabián Padilla, Eliana Providel, Sebastián Campos, Renato Bassi, Andrea Riquelme, Valeria Aldana, and Claudia López. 2023. A study on information disorders on social networks during the chilean social outbreak and covid-19 pandemic. *Applied Sciences*, 13(9).

A. Modupe, T. Sindane, and V. Marivate. 2023. Zero-shot transfer learning using affix and correlated cross-lingual embeddings. *Authorea*.

L. Nayeon, B. Z. Li, S. Wang, P. Fung, H. Ma, W. Yih, and M. Khabsa. 2021. On unifying misinformation detection. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 5479–5485, Online. Association for Computational Linguistics.

Institute of Mass Information. Online media that have become the highest quality: white list of the second half of 2023 [online]. 2023.

C.Y. Park, J. Mendelsohn, A. Field, and Yu. Tsvetkov. 2022. Challenges and opportunities in information manipulation detection: An examination of wartime Russian media. In *Findings of the Association for Computational Linguistics: EMNLP 2022*, pages 5209–5235, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Egil Rønningstad, Erik Velldal, and Lilja Øvrelid. 2022. Entity-level sentiment analysis (ELSA): An exploratory task survey. In *Proceedings of the 29th International Conference on Computational Linguistics*, pages 6773–6783, Gyeongju, Republic of Korea. International Committee on Computational Linguistics.

G. Shan, S. Xu, L. Yang, S. Jia, and Y. Xiang. 2020. Learn#: A novel incremental learning method for text classification. *Expert Systems with Applications*, 147:113198.

R.M. Shukla and J. Cartlidge. 2022. Challenges faced by industries and their potential solutions in deploying machine learning applications. In *2022 IEEE 12th Annual Computing and Communication Workshop and Conference (CCWC)*, pages 0119–0124.

Veronika Solopova, Christoph Benzmüller, and Tim Landgraf. 2023. The evolution of pro-kremlin propaganda from a machine learning and linguistics perspective. In *Proceedings of the Second Ukrainian Natural Language Processing Workshop (UNLP)*, pages 40–48, Dubrovnik, Croatia. Association for Computational Linguistics.

T. Ustyianovych, N. Kasianchuk, H. Falfushynska, S. Fedushko, and E. Siemens. 2023. Dynamic topic modelling of online discussions on the russian war in ukraine. In *Proceedings of International Conference on Applied Innovation in IT*, pages 81–89.

R. Wang, T. Yu, H. Zhao, S. Kim, S. Mitra, R. Zhang, and R. Henao. 2022a. Few-shot class-incremental learning for named entity recognition. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 571–582, Dublin, Ireland. Association for Computational Linguistics.

Zh. Wang, X. Shan, X. Zhang, and J .Yang. 2022b. N24News: A new dataset for multimodal news classification. In *Proceedings of the Thirteenth Language Resources and Evaluation Conference*, pages 6768–6775, Marseille, France. European Language Resources Association.

P. Yao, T. Renwick, and D. Barbosa. 2022. WordTies: Measuring word associations in language models via constrained sampling. In *Findings of the Association for Computational Linguistics: EMNLP 2022*, pages 5959–5970, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

A. Zakharchenko, T. Peráček, S. Fedushko, Yu. Syerov, and O. Trach. 2021. When fact-checking and ‘bbc standards’ are helpless: ‘fake newsworthy event’ manipulation and the reaction of the ‘high-quality media’ on it. *Sustainability*, 13(2).

Setting up the Data Printer with Improved English to Ukrainian Machine Translation

Yurii Paniv, Dmytro Chaplynskyi, Nikita Trynus, Volodymyr Kyrylov
Ukrainian Catholic University, lang-uk initiative,
Igor Sikorsky Kyiv Polytechnic Institute, Università della Svizzera italiana
paniv@ucu.edu.ua, chaplinsky.dmitry@gmail.com, trynus.nikita@lll.kpi.ua, vol@wilab.org.ua

Abstract

To build large language models for Ukrainian we need to expand our corpora with large amounts of new algorithmic tasks expressed in natural language. Examples of task performance expressed in English are abundant, so with a high-quality translation system our community will be enabled to curate datasets faster. To aid this goal, we introduce a recipe to build a translation system using supervised finetuning of a large pretrained language model with a noisy parallel dataset of 3M pairs of Ukrainian and English sentences followed by a second phase of training using 17K examples selected by k-fold perplexity filtering on another dataset of higher quality. Our decoder-only model named Dragoman beats performance of previous state of the art encoder-decoder models on the FLORES devtest set.

Keywords: machine translation, parameter-efficient fine tuning, large language models, unsupervised data selection, perplexity filtering

1. Introduction

The availability of the data is the most important ingredient when one needs to pretrain general-purpose large language models for a specific natural language task or a set of tasks. While it is relatively easy to obtain a good and balanced dataset under specific domain for the English language, it is much harder to do the same for other under-resourced languages such as Ukrainian.

Since curating a corpus of tasks in Ukrainian is a large endeavor, and given a large body of work done for English, we consider existing instruction tuning datasets as a source of tasks to reuse in Ukrainian using automatic machine translation.

This work focuses on improving the current state of machine translation from English to Ukrainian.

We contribute a recipe for finetuning a large pretrained language model with publicly available data to build a translation system (section 3, section 4). This matches state of the art performance of the best encoder-decoder model on a common multilingual benchmark using a consumer GPU with 24 GiB of VRAM. We release training, evaluation code, datasets, and model at `https://github.com/lang-uk/dragoman`. Our main results are summarized in Table 1. We provide examples of the top-5 best and worst translations on the FLORES devtest set in the Appendix A.

We base pretrained model selection on evaluation in few-shot learning setting (section 5). We find that it's a promising method to design tasks without training, and the model can perform comparably to specialized systems given increased inference budget and auxiliary translation scoring functions, yet still underperforms our finetuned recipe.

Model	BLEU ↑
Finetuned	
Dragoman P, 10 beams (section 3)	30.4
Dragoman PT, 10 beams (section 4)	**32.3**
Zero shot and few shot (section 5)	
Llama 2 7B 2-shot, 10 beams	20.1
Mistral-7B-v0.1 2-shot, 10 beams	24.9
gpt-4 10-shot	29.5
gpt-4-turbo-preview 0-shot	30.4
Pretrained encoder-decoder	
NLLB-3B, 10 beams	30.6
OPUS-MT, 10 beams	**32.2**

Table 1: Main results. Our Dragoman models improve existing state of the art on translation from English to Ukrainian on FLORES-101 devtest (Goyal et al., 2022), a multilingual benchmark of translated sentences from web articles. We compare to state of the art encoder-decoder models, NLLB-3B (Team et al., 2022) and OPUS-MT (Tiedemann and Thottingal, 2020).

2. Supervised Finetuning

We cast machine translation as a likelihood maximization of a density p of Ukrainian sentences $Y =$ "перекладене речення" $\in \mathcal{Y}$ conditioned on their English sources with quasi-instruction formatting: $X =$ "[INST] translated sentence [/INST]" $\in \mathcal{X}$.

The density is parametrized using a neural network with frozen pretrained weights θ:

$$\operatorname{argmax}_{\phi}\, \mathrm{p}_{\theta,\phi}(Y|X) \tag{1}$$

We implement the conditional language modeling objective by masking out tokens of X when

The Third Ukrainian Natural Language Processing Workshop (UNLP) @LREC-COLING-2024, pages 41–50
25 May, 2024. © 2024 ELRA Language Resource Association: CC BY-NC 4.0

Dataset	Pairs	Filters				Example Order	Best BLEU ↑
		Lang	BPC	LaBSE	Len diff		
1m unfiltered	963k	-	-	-	-	Random	28.26
1m filtered	958k	En/Uk	<3.33	>0.91	<50	Random	29.47
3m filtered	2.9m	En/Uk	<3.25	>0.85	<50	By LaBSE score, dissimilar first	**30.37**
8m filtered	8m	En/Uk	<5	>0.5	<50	By LaBSE score, dissimilar first	30.19

Table 2: Summary of experiments with Paracrawl subcorpora. Legend of filters: **Lang** denotes language filters, **BPC** denotes maximum sum of bits per character measures, **LaBSE** denotes maximum sentence embedding cosine similarity between source and target sentences, **Len diff** denotes maximum difference in length between source and target in characters. Example ordering impacts data loading in the training loop.

computing token-wise cross entropy of shifted targets. We only optimize extra low rank adapter (Hu et al., 2022) parameters ϕ after nf4 quantization (Dettmers et al., 2023). In practice we use large rank values and adapter mixture weights. All training runs proceed for one epoch and we use dropout (Srivastava et al., 2014) for regularization against data noise.

We use Mistral-7B-v0.1 (Jiang et al., 2023) as a base pretrained decoder-only transformer, as it performs favorably in our few-shot experiments (section 5).

3. First Phase: Heuristic Filtering of Paracrawl

We use the publicly available Paracrawl dataset (Bañón et al., 2020). This dataset contains 13,354,365 English-Ukrainian sentence pairs, collected by automatically matching similar sentences in large corpora of internet text.

We have identified issues with translation pairs, including a significant number of repetitive or incorrect examples. We encounter a large subset of repetitive weather forecasts following the template "The temperature in <x> is <y> degrees," and sentences from site navigation menus. Additionally, many texts appear to be scraped from adult websites, containing low-quality, machine-translated samples. We have spotted numerous instances of incomplete or significantly incorrect translation pairs. Some target sentences were written in languages other than Ukrainian.

To control the quality of the sentences, we apply multiple heuristics.

Language filtering gcld3 library[1] provides language detection capabilities. We remove all sentences that failed to verify as Ukrainian or English.

Perplexity thresholding We score source and target sentences using two decoder-only models trained on different monolingual datasets (Radford et al., 2019; Minixhofer et al., 2022) and sum their bits per character measures.

Translation mismatch filtering LaBSE (Feng et al., 2022) embeds sentences into a space, where similar sentences in different languages are close together. We use it to filter out badly aligned sentence pairs.

Length filtering The lengths of the original and translated sentences reveal examples that are too short or too long. Absolute differences of lengths point to pairs with long target for the short source and vice versa.

We arbitrarily choose joint values of filtering thresholds to get the desired approximate example counts: 1 million, 3 million and 8 million. We perform multiple experiments with these splits while searching for optimal hyperparameters. We list threshold values in Table 2 and best results for each subset.

4. Second Phase: Unsupervised Data Selection on Extended Multi30K

We use the best checkpoint from the previous finetuning phase to train on a high-quality dataset: Extended Multi30K from Saichyshyna et al. (2023). Switching datasets gives us a performance boost of 1.97 BLEU. We additionally delete 11600 sentences from the dataset using unsupervised perplexity filtering pipeline gaining 0.35 on the dev set that translates to 0.3 BLEU on the devtest subset of FLORES.

We use perplexity as a data selection criterion to calculate thresholds to filter out highly surprising sentences. We apply the k-fold cross-validation technique to make the perplexity evaluation in-domain. We split the training data into $k = 5$ folds and train k models withholding one of the folds from each run. Then we score every sentence using the model that has not seen that sentence in training. Next, we sweep for acceptable threshold values by minimizing BLEU on the development set and report results in Table 3. We plot the distribution of scores in Figure 1. We also provide threshold sweep results for training from base Mistral-7B-v0.1 checkpoint in Table 6. By comparing finetuned re-

[1] https://github.com/google/cld3

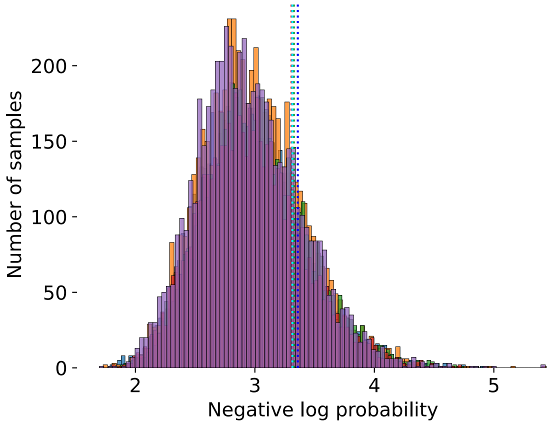


Figure 1: Distributions of sentence log probabilities for each fold superimposed on top of each other. Every bar color represents a unique fold; every vertical line denotes a 60th percentile cutoff threshold. The best percentile is chosen using grid search shown in Table 3.

sults, we demonstrate that data from the second phase alone is not enough to match the performance of our best checkpoint.

Threshold	**Examples**	**BLEU** ↑	
percentile		dev	devtest
20th	5800	31.57	32.06
40th	11600	31.65	32.16
50th	14500	31.76	32.36
60th	17400	<u>31.80</u>	**32.34**
70th	20300	31.51	32.17
80th	23200	31.44	32.46
95.4th (2σ)	28025	31.74	32.18
Full dataset	29000	31.45	32.04

Table 3: Extended Multi30K log probability thresholds swept on FLORES dev set. We choose the best checkpoint based on model performance on FLORES dev subset using grid search for optimal perplexity threshold value.

5. Few-Shot Translation

Conditioning the model on a sequence of demonstrations of performing some task allows the model to learn this task in-context, also known as "few-shot learning" (e.g. Brown et al. (2020)), thanks to the ability of the Transformers to modulate representations of its future tokens using past context, implementing a specialized internal context-dependent learning algorithm inside its weights (von Oswald et al., 2023).

While few-shot learning allows to quickly try any task with a low number of demonstrations, Liu et al. (2022) have shown that parameter-efficient finetuning allows smaller models achieve better performance, effectively spending less floating point operations per test example at inference time.

Setting up the model for finetuning requires a lot of work, and in-context learning allows to quickly probe capability of a large model using inference software that performs efficient management of key-value cache for speed (Kwon et al., 2023).

To test backbone models before finetuning, we attempt decoding translations with a basic prompt shown in Figure 2.

[INST] They are planning to host a party next weekend. [/INST] Вони планують провести вечірку наступного вікенду.
[INST] I enjoy swimming in the ocean and feeling the salty breeze. [/INST] Мені подобається плавати в океані та відчувати солоний вітер.
[INST]

Figure 2: Basic 2-shot prompt used for few-shot translation. `[INST]` prefixes the beginning of the source sentence and `[/INST]` denotes the beginning of the target translation. These separators are chosen arbitrarily (as in finetuning) and are not special vocabulary items, even though they bear visual resemblance to them.

We find that the model significantly underperforms compared to current state of the art translation models when using beam search (Tillmann and Ney, 2003).

This decoding algorithm performs pruned breadth-first expansion, scoring target sentence prefixes using model's own log probability, approximating maximum a-posteriori estimation of the best translation.

Inspection of the n-best list of translation candidates (beams) reveals that the models can produce high-quality translations, however assign low probabilities to them. We find the best possible translation by rescoring beams using the BLEU score as a loss function (Kumar and Byrne, 2004) with respect to the reference translation (the so-called "oracle").

We employ this oracle rescoring strategy to gauge the potential capability of the model to produce good translations without finetuning, and find that in a regime of increased computation (large width of the beam) and assuming perfect selection capability, a base model is competitive with specialized alternatives. We sweep over a grid of multiple beam widths and report highest attainable BLEU scores in Table 4.

Beams	Oracle BLEU ↑	
	Mistral-7B-v.01	Llama 2 7B
3	27.11	24.55
5	29.20	26.64
10	31.53	28.76
15	32.81	29.09
20	33.54	27.64
25	34.27	26.35
30	33.99	(decoder failure)
35	34.94	
40	34.61	

Table 4: We establish the upper bound of the latent capability of pretrained base models to produce high quality translations with by varying beam width on the task of translating sentences from FLORES dev given a 2-shot prompt. The ground truth oracle determines the best beam. We use beam search implementation by Kwon et al. (2023) with presence penalty of 0.1. The results do not improve monotonically with increasing beam size, and lengths of hypotheses grow with maximum beam size, yielding diminishing returns. This problem can be attributed to label bias (Murray and Chiang, 2018), and rectifying it will require extra regularization.

Consecutive sentences in FLORES are samples from the same document. We hypothesize, dynamically adjusting the prompt by inserting previous translations will improve results. We observe that the model indeed improves translation of certain words such as proper nouns through access to correct definitions provided in the context (Figure 3), however its overall performance degrades in other examples.

We additionally attempt basic 0-shot with a system prompt `You translate English sentences into native Ukrainian.`, and 10-shot prompting using automatic prompt selection based on similarity between source sentences experiments with GPT-4 and GPT-4 Turbo and find that commerical systems perform similarly to other open source systems, as shown in Table 1.

6. Discussion and Limitations

Single-sentence translation Our system is trained on demonstrations of standalone sentence pairs.

Decoder-only models with long context windows We choose to finetune existing decoder-only models since the choice of models with almost the same architecture but different massive pretraining data is abundant. The number of open-source models released recently and their constant improvement offers a good prospective for the machine translation tasks.

Source: RSPCA New South Wales chief inspector David O'Shannessy told the ABC that surveillance and inspections of abattoirs should be commonplace in Australia.

Hypothesis given random 2-shot context: Головний інспектор РСПКА Нового Південного Уельсу Девід О'Шеннесі повідомив ABC, що спостереження та інспекції аббатств повинні бути звичайним явищем в Австралії.

Context example: [INST] Animal Liberation and the Royal Society for the Prevention of Cruelty to Animals (RSPCA) are again calling for the mandatory installation of CCTV cameras in all Australian abattoirs. [/INST] Організація Звільнення тварин і Королівське товариство із запобігання жорстокому поводженню з тваринами (КТЗЖПТ) знову закликають до обов'язкової установки камер спостереження на всіх австралійських бійнях.

Hypothesis given relevant 2-shot context: Головний інспектор Королівського товариства із запобігання жорстокому поводженню з тваринами (КТЗЖПТ) Нового Південного Уельсу Девід О'Шеннесі заявив, що спостереження та інспекції бійні повинні бути поширеними в Австралії.

Figure 3: Few-shot translation with contextual prompting allows the model to learn named entities on the fly. Without context, the model makes a wrong guess trying to transliterate the abbreviation.

These models receive gradient from all outputs during pretraining, and the self-attention mechanism can see the input, the partial output, and access past examples of translations in its context window using induction heads (Olsson et al., 2022).

For efficiency, we only train on examples with single short sentence pairs and do not pack context windows full of tokens as done in pretraining. In our early experiments, we find that our models still generalize to inputs longer that what is seen in training. This generalization behavior is often attributed to relative position embeddings (Dai et al., 2019; Csordás et al., 2021). We leave evaluation of long context attention stability under these conditions for future work.

Training on the noisy dataset Data cleaning has a positive effect on the resulting metrics. However, our models trained on 8 million filtered, examples perform worse than models trained on 3 million examples (Table 2).

Tokenizer performance We used the LLaMA and Mistral tokenizers during our experiments, which use at least twice as many tokens to compress a sentence in Ukrainian of the same length as an English sentence in character. In practice, that means that generating a sentence in Ukrainian takes at least twice as many steps to generate. We show a distribution of sentence token lengths in Figure 4.

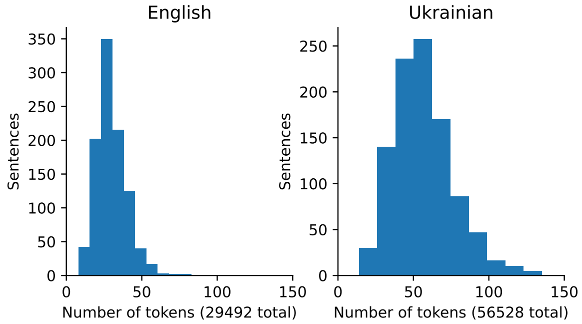


Figure 4: Comparison of tokenizer compression rates between English and Ukrainian using the Mistral-7B tokenizer on the FLORES dev set.

Evaluation We choose BLEU-4 score (Papineni et al., 2002) as our core evaluation metric and model selection criterion. BLEU-4 measures 4-gram precisions, where grams are defined as words. We use the implementation and rely on tokenization decisions of Post (2018). This metric is sensitive to minor differences that do not affect the meaning of the sentence, for example case inflections that tend to cascade to multiple adjacent words. BLEU is known to poorly correlate with human judgement of translation quality, and Freitag et al. (2022) recommend learned metrics.

Choosing an appropriate learned metric for judgements of translation quality of Ukrainian requires careful consideration, and incorporation of data informed by the language community, such as a curated corpus of grammar corrections that reflects proper modern use of language (Syvokon et al., 2023).

Regardless of limitations of BLEU, improvement in BLEU still signals improvement in translation quality in our regime.

WMT22 Our reviewers have pointed out that WMT22 benchmark (Kocmi et al., 2022) includes a test set for Ukrainian. Our model achieves 24.72 on the WMT22 test set without any postprocessing, ranking behind the best result of Roussis and Papavassiliou (2022) at 25.2 BLEU. We note that the submission that scores relatively low on the WMT22 test, scores comparably to our results on FLORES. These data distribution properties require closer exploration.

7. Related Work

Translation to Ukrainian Maksymenko et al. (2023a,b) explore translation controllability by conditioning the model on text embeddings that encode style by finetuning an encoder-decoder model. They claim high quality translations on a private test set.

Instruction-tuned language models Üstün et al. (2024) explore large-scale translation efforts to produce a multilingual instruction-tuned language model Aya. This work translates large datasets like the Flan Collection (Longpre et al., 2023) using the NLLB-3B model (Team et al., 2022).

Translation systems Han et al. (2021) provide an iterated backtranslation recipe to bootstrap neural machine translation systems using generative models: zero-shot translation ability is used to produce candidates for few-shot demonstrations. Filtered few-shot demonstrations are used to sample new sentences for further finetuning for translation in two directions.

Translation benchmarks Besides FLORES-101 (Goyal et al. (2022), or FLORES-200 (Team et al., 2022), both include the same data for Ukrainian) dataset used in this work, Tiedemann (2020) provides an additional dataset for multilingual evaluation.

Data selection techniques Yang and Li (2023) propose a perplexity filtering pipeline, in which the data is split into k folds to classify low quality augmentation generations produced by surrogate language models. Sachdeva et al. (2024) provide recipes on curating data for language models by directly asking language models to score examples.

8. Conclusion

In this work, we build a translation system using a two-phase data cleaning pipeline. We demonstrate matching performance to state-of-the-art encoder-decoder models for English-Ukrainian translation task. Notably, our system exhibits superior performance compared to the NLLB model, which was instrumental in generating the Aya dataset and contributed significantly to the advancement of multilingual language models. Improved machine translation could bring new capabilities to the next generation of large language models trained for the Ukrainian language. The recent improvements made for decoder-only backbones and the general dynamics of this process encourages us: we firmly believe that recipes we propose in this paper can be used to improve the quality of the translation by simply upgrading the backbone model.

9. Contributions

Yurii Paniv worked on unsupervised data selection on extended Multi30K dataset, Dmytro Chaplynskyi performed initial training using heuristic filtering of Paracrawl, Nikita Trynus worked on evaluation, Volodymyr Kyrylov designed few-shot learning experiments and evaluation.

10. Acknowledgements

The authors would like to thank Talents for Ukraine project of Kyiv School of Economics for the grant on compute, used for this paper, Ukrainian Catholic University, Roman Kyslyi, and Oleksii Turuta for the fruitful discussions and inspiration. We would also like to thank paper reviewers for their comments.

11. References

Marta Bañón, Pinzhen Chen, Barry Haddow, Kenneth Heafield, Hieu Hoang, Miquel Esplà-Gomis, Mikel L. Forcada, Amir Kamran, Faheem Kirefu, Philipp Koehn, Sergio Ortiz Rojas, Leopoldo Pla Sempere, Gema Ramírez-Sánchez, Elsa Sarrías, Marek Strelec, Brian Thompson, William Waites, Dion Wiggins, and Jaume Zaragoza. 2020. ParaCrawl: Web-scale acquisition of parallel corpora. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 4555–4567, Online. Association for Computational Linguistics.

Tom B. Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel M. Ziegler, Jeffrey Wu, Clemens Winter, Christopher Hesse, Mark Chen, Eric Sigler, Mateusz Litwin, Scott Gray, Benjamin Chess, Jack Clark, Christopher Berner, Sam McCandlish, Alec Radford, Ilya Sutskever, and Dario Amodei. 2020. Language models are few-shot learners.

Róbert Csordás, Kazuki Irie, and Juergen Schmidhuber. 2021. The devil is in the detail: Simple tricks improve systematic generalization of transformers. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 619–634, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Zihang Dai, Zhilin Yang, Yiming Yang, Jaime Carbonell, Quoc Le, and Ruslan Salakhutdinov. 2019. Transformer-XL: Attentive language models beyond a fixed-length context. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 2978–2988, Florence, Italy. Association for Computational Linguistics.

Tim Dettmers, Artidoro Pagnoni, Ari Holtzman, and Luke Zettlemoyer. 2023. Qlora: Efficient finetuning of quantized llms. In *Advances in Neural Information Processing Systems*, volume 36, pages 10088–10115. Curran Associates, Inc.

Fangxiaoyu Feng, Yinfei Yang, Daniel Cer, Naveen Arivazhagan, and Wei Wang. 2022. Language-agnostic bert sentence embedding.

Markus Freitag, Ricardo Rei, Nitika Mathur, Chi-kiu Lo, Craig Stewart, Eleftherios Avramidis, Tom Kocmi, George Foster, Alon Lavie, and André F. T. Martins. 2022. Results of WMT22 metrics shared task: Stop using BLEU – neural metrics are better and more robust. In *Proceedings of the Seventh Conference on Machine Translation (WMT)*, pages 46–68, Abu Dhabi, United Arab Emirates (Hybrid). Association for Computational Linguistics.

Naman Goyal, Cynthia Gao, Vishrav Chaudhary, Peng-Jen Chen, Guillaume Wenzek, Da Ju, Sanjana Krishnan, Marc'Aurelio Ranzato, Francisco Guzmán, and Angela Fan. 2022. The Flores-101 evaluation benchmark for low-resource and multilingual machine translation. *Transactions of the Association for Computational Linguistics*, 10:522–538.

Jesse Michael Han, Igor Babuschkin, Harrison Edwards, Arvind Neelakantan, Tao Xu, Stanislas Polu, Alex Ray, Pranav Shyam, Aditya Ramesh, Alec Radford, and Ilya Sutskever. 2021. Unsupervised neural machine translation with generative language models only. *ArXiv*, abs/2110.05448.

Edward J Hu, yelong shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. 2022. LoRA: Low-rank adaptation of large language models. In *International Conference on Learning Representations*.

Albert Q. Jiang, Alexandre Sablayrolles, Arthur Mensch, Chris Bamford, Devendra Singh Chaplot, Diego de las Casas, Florian Bressand, Gianna Lengyel, Guillaume Lample, Lucile Saulnier, Lélio Renard Lavaud, Marie-Anne Lachaux, Pierre Stock, Teven Le Scao, Thibaut Lavril, Thomas Wang, Timothée Lacroix, and William El Sayed. 2023. Mistral 7b.

Tom Kocmi, Rachel Bawden, Ondřej Bojar, Anton Dvorkovich, Christian Federmann, Mark

Fishel, Thamme Gowda, Yvette Graham, Roman Grundkiewicz, Barry Haddow, Rebecca Knowles, Philipp Koehn, Christof Monz, Makoto Morishita, Masaaki Nagata, Toshiaki Nakazawa, Michal Novák, Martin Popel, and Maja Popović. 2022. Findings of the 2022 conference on machine translation (WMT22). In *Proceedings of the Seventh Conference on Machine Translation (WMT)*, pages 1–45, Abu Dhabi, United Arab Emirates (Hybrid). Association for Computational Linguistics.

Shankar Kumar and William Byrne. 2004. Minimum Bayes-risk decoding for statistical machine translation. In *Proceedings of the Human Language Technology Conference of the North American Chapter of the Association for Computational Linguistics: HLT-NAACL 2004*, pages 169–176, Boston, Massachusetts, USA. Association for Computational Linguistics.

Woosuk Kwon, Zhuohan Li, Siyuan Zhuang, Ying Sheng, Lianmin Zheng, Cody Hao Yu, Joseph E. Gonzalez, Hao Zhang, and Ion Stoica. 2023. Efficient memory management for large language model serving with PagedAttention. In *Proceedings of the ACM SIGOPS 29th Symposium on Operating Systems Principles*.

Haokun Liu, Derek Tam, Mohammed Muqeeth, Jay Mohta, Tenghao Huang, Mohit Bansal, and Colin Raffel. 2022. Few-shot parameter-efficient fine-tuning is better and cheaper than in-context learning.

Shayne Longpre, Le Hou, Tu Vu, Albert Webson, Hyung Won Chung, Yi Tay, Denny Zhou, Quoc V Le, Barret Zoph, Jason Wei, et al. 2023. The Flan Collection: Designing Data and Methods for Effective Instruction Tuning. *arXiv preprint arXiv:2301.13688*.

Daniil Maksymenko, Nataliia Saichyshyna, Oleksii Turuta, Marcin Paprzycki, Maria Ganzha, and Mirela Alhasani. 2023a. Controllability for english-ukrainian machine translation by using style transfer techniques. *2023 18th Conference on Computer Science and Intelligence Systems (FedCSIS)*, pages 1059–1068.

Daniil Maksymenko, Olena Turuta, Nataliia Saichyshyna, Maksym Yerokhin, and Oleksii Turuta. 2023b. Controllability for English-Ukrainian machine translation based on specialized corpora. In *Proceedings of the 1st International Workshop on Multilingual, Multimodal and Multitask Language Generation*, pages 1–9, Tampere, Finland. European Association for Machine Translation.

Benjamin Minixhofer, Fabian Paischer, and Navid Rekabsaz. 2022. WECHSEL: Effective initialization of subword embeddings for cross-lingual transfer of monolingual language models. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 3992–4006, Seattle, United States. Association for Computational Linguistics.

Kenton Murray and David Chiang. 2018. Correcting length bias in neural machine translation. In *Proceedings of the Third Conference on Machine Translation: Research Papers*, pages 212–223, Brussels, Belgium. Association for Computational Linguistics.

Catherine Olsson, Nelson Elhage, Neel Nanda, Nicholas Joseph, Nova DasSarma, Tom Henighan, Ben Mann, Amanda Askell, Yuntao Bai, Anna Chen, Tom Conerly, Dawn Drain, Deep Ganguli, Zac Hatfield-Dodds, Danny Hernandez, Scott Johnston, Andy Jones, Jackson Kernion, Liane Lovitt, Kamal Ndousse, Dario Amodei, Tom Brown, Jack Clark, Jared Kaplan, Sam McCandlish, and Chris Olah. 2022. In-context learning and induction heads. *Transformer Circuits Thread*.

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. BLEU: a method for automatic evaluation of machine translation. In *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics*, pages 311–318, Philadelphia, Pennsylvania, USA. Association for Computational Linguistics.

Matt Post. 2018. A call for clarity in reporting BLEU scores. In *Proceedings of the Third Conference on Machine Translation: Research Papers*, pages 186–191, Belgium, Brussels. Association for Computational Linguistics.

Alec Radford, Jeff Wu, Rewon Child, David Luan, Dario Amodei, and Ilya Sutskever. 2019. Language models are unsupervised multitask learners.

Dimitrios Roussis and Vassilis Papavassiliou. 2022. The ARC-NKUA submission for the English-Ukrainian general machine translation shared task at WMT22. In *Proceedings of the Seventh Conference on Machine Translation (WMT)*, pages 358–365, Abu Dhabi, United Arab Emirates (Hybrid). Association for Computational Linguistics.

Noveen Sachdeva, Benjamin Coleman, Wang-Cheng Kang, Jianmo Ni, Lichan Hong, Ed H. Chi, James Caverlee, Julian McAuley, and Derek Zhiyuan Cheng. 2024. How to train data-efficient llms.

Nataliia Saichyshyna, Daniil Maksymenko, Oleksii Turuta, Andriy Yerokhin, Andrii Babii, and Olena Turuta. 2023. Extension Multi30K: Multimodal dataset for integrated vision and language research in Ukrainian. In *Proceedings of the Second Ukrainian Natural Language Processing Workshop (UNLP)*, pages 54–61, Dubrovnik, Croatia. Association for Computational Linguistics.

Nitish Srivastava, Geoffrey Hinton, Alex Krizhevsky, Ilya Sutskever, and Ruslan Salakhutdinov. 2014. Dropout: A simple way to prevent neural networks from overfitting. *Journal of Machine Learning Research*, 15(56):1929–1958.

Oleksiy Syvokon, Olena Nahorna, Pavlo Kuchmiichuk, and Nastasiia Osidach. 2023. UA-GEC: Grammatical error correction and fluency corpus for the Ukrainian language. In *Proceedings of the Second Ukrainian Natural Language Processing Workshop (UNLP)*, pages 96–102, Dubrovnik, Croatia. Association for Computational Linguistics.

NLLB Team, Marta R. Costa-jussà, James Cross, Onur Çelebi, Maha Elbayad, Kenneth Heafield, Kevin Heffernan, Elahe Kalbassi, Janice Lam, Daniel Licht, Jean Maillard, Anna Sun, Skyler Wang, Guillaume Wenzek, Al Youngblood, Bapi Akula, Loic Barrault, Gabriel Mejia Gonzalez, Prangthip Hansanti, John Hoffman, Semarley Jarrett, Kaushik Ram Sadagopan, Dirk Rowe, Shannon Spruit, Chau Tran, Pierre Andrews, Necip Fazil Ayan, Shruti Bhosale, Sergey Edunov, Angela Fan, Cynthia Gao, Vedanuj Goswami, Francisco Guzmán, Philipp Koehn, Alexandre Mourachko, Christophe Ropers, Safiyyah Saleem, Holger Schwenk, and Jeff Wang. 2022. No language left behind: Scaling human-centered machine translation.

Jörg Tiedemann. 2020. The tatoeba translation challenge – realistic data sets for low resource and multilingual MT. In *Proceedings of the Fifth Conference on Machine Translation*, pages 1174–1182, Online. Association for Computational Linguistics.

Jörg Tiedemann and Santhosh Thottingal. 2020. OPUS-MT – building open translation services for the world. In *Proceedings of the 22nd Annual Conference of the European Association for Machine Translation*, pages 479–480, Lisboa, Portugal. European Association for Machine Translation.

Christoph Tillmann and Hermann Ney. 2003. Word Reordering and a Dynamic Programming Beam Search Algorithm for Statistical Machine Translation. *Computational Linguistics*, 29(1):97–133.

Johannes von Oswald, Eyvind Niklasson, Ettore Randazzo, João Sacramento, Alexander Mordvintsev, Andrey Zhmoginov, and Max Vladymyrov. 2023. Transformers learn in-context by gradient descent.

Heng Yang and Ke Li. 2023. Boosting text augmentation via hybrid instance filtering framework. In *Findings of the Association for Computational Linguistics: ACL 2023*, pages 1652–1669, Toronto, Canada. Association for Computational Linguistics.

Ahmet Üstün, Viraat Aryabumi, Zheng-Xin Yong, Wei-Yin Ko, Daniel D'souza, Gbemileke Onilude, Neel Bhandari, Shivalika Singh, Hui-Lee Ooi, Amr Kayid, Freddie Vargus, Phil Blunsom, Shayne Longpre, Niklas Muennighoff, Marzieh Fadaee, Julia Kreutzer, and Sara Hooker. 2024. Aya model: An instruction finetuned open-access multilingual language model.

Appendix

A. Translation Examples

Sample of top 5 worst examples by BLEU from FLORES devtest set **Source:** While most of their food would be familiar to us, Romans did have their share of strange or unusual feast items, including wild boar, peacock, snails, and a type of rodent called a dormouse
Hypothesis: Хоча більшість їхніх продуктів харчування було б знайомим для нас, римляни мали свою частку дивних або незвичайних страв, включаючи кабана, павича, равликів та гризунів, званих куницями.
Reference: Переважно римляни харчувались тим, що знайоме і нам, але були у них і свої дивні та незвичні бенкетні частування, серед яких дикий кабан, павич, слимаки і гризун під назвою вовчок.
BLEU: 1.72

Source: "I was moved every time we did a rehearsal on this, from the bottom of my heart."
Hypothesis: "Я був зворушений кожного разу, коли ми робили репетицію з цього приводу, від щирого серця".
Reference: Мене це торкало кожен раз під час репетиції до глибини душі.
BLEU: 1.91

Source: Feral children may have experienced severe child abuse or trauma before being abandoned or running away.
Hypothesis: Дитина-дикун, можливо, пережила жорстоке насильство над дітьми або травму, перш ніж бути кинутою або втекти.

Reference: Може таке бути, що дикарські діти були об'єктами жорстокого поводження або отримали серйозні травми перед тим, як їх полишили або вони втекли.
BLEU: 2.31

Source: Workers must often get their superiors' approval for any decisions they make, and are expected to obey their superiors' instructions without question.
Hypothesis: Працівники часто повинні отримувати схвалення своїх начальників на будь-які рішення, які вони приймають, і повинні беззаперечно виконувати вказівки своїх начальників.
Reference: Працівникам часто потрібно отримати схвалення керівника щодо будь-яких рішень, а також від них очікують беззаперечного дотримання настанов керівників.
BLEU: 2.32

Source: Typically there will be a tuition fee to enroll in these educational programs.
Hypothesis: Як правило, буде плата за навчання, щоб зареєструватися в цих освітніх програмах.
Reference: Зазвичай такі освітні програми платні.
BLEU: 2.62

Sample of top 5 best examples by BLEU from FLORES devtest set

Source: The East African Islands are in the Indian Ocean off the eastern coast of Africa.
Hypothesis: Східноафриканські острови знаходяться в Індійському океані біля східного узбережжя Африки.
Reference: Східноафриканські острови знаходяться в Індійському океані біля східного узбережжя Африки.
BLEU: 100.00

Source: Earlier the Chinese news agency Xinhua reported a plane to be hijacked.
Hypothesis: Раніше китайське інформаційне агентство Сіньхуа повідомило про викрадення літака.
Reference: Раніше китайське інформаційне агентство Сіньхуа повідомило про викрадення літака.
BLEU: 100.00

Source: For instance, they didn't have corn, nor tomatoes, nor potatoes, nor cocoa, and no ancient Roman ever tasted a turkey.
Hypothesis: Наприклад, у них не було ні кукурудзи, ні помідорів, ні картоплі, ні какао, і жоден стародавній римлянин ніколи не скуштував індичку.
Reference: Наприклад, у них не було ні кукурудзи, ні помідорів, ні картоплі, ні какао, і жоден стародавній римлянин ніколи не куштував індичку.
BLEU: 90.95

Source: The luminosity and rotation are used together to determine a star's Rossby number, which is related to plasma flow.
Hypothesis: Світність і обертання використовуються разом для визначення числа Россбі зірки, яке пов'язане з потоком плазми.
Reference: Світність і обертання використовуються разом для визначення числа Россбі зірки, яке пов'язане з плазмовим потоком.
BLEU: 83.26

Source: But being placed in the "high tropics" just a few degrees north of equator you will need to deal with both heat (always) and strong sun (when the sky is clear, more rarely).
Hypothesis: Але перебуваючи в "високих тропіках" всього в декількох градусах на північ від екватора, вам доведеться мати справу як з спекою (завжди), так і з сильним сонцем (коли небо чисте, рідше).
Reference: Але, перебуваючи в "високих тропіках" всього в декількох градусах на північ від екватора, вам доведеться мати справу як зі спекою (завжди), так і з палючим сонцем (коли небо чисте, рідше).
BLEU: 82.47

Model	BLEU ↑	spBLEU	chrF	chrF++
Finetuned				
Dragoman P, 10 beams (section 3)	30.38	37.93	59.49	56.41
Dragoman PT, 10 beams (section 4)	**32.34**	**39.93**	**60.72**	**57.82**
Zero shot and few shot (section 5)				
LLaMa-2-7B 2-shot	20.1	26.78	49.22	46.29
RWKV-5-World-7B 0-shot	21.06	26.20	49.46	46.46
gpt-4 10-shot	29.48	37.94	58.37	55.38
gpt-4-turbo-preview 0-shot	30.36	36.75	59.18	56.19
Google Translate 0-shot	25.85	32.49	55.88	52.48
Pretrained				
NLLB 3B, 10 beams	30.46	37.22	58.11	55.32
OPUS-MT, 10 beams	32.2	39.76	60.23	57.38

Table 5: We evaluate generated translations with the sacrebleu library to calculate BLEU, spBLEU, chrF, and chrF++ metrics on the FLORES DEVTEST set. Metric spBLEU was calculated with default BLEU values and tokenizer flores101. Tokenization and detokenization are done using the models' default tokenizers. Evaluation is performed on detokenized sentences with corresponding reference sentences.

Threshold	**Examples**	**BLEU ↑**	
percentile		dev	devtest
20th	5800	25.14	25.49
40th	11600	25.39	25.45
50th	14500	25.79	25.93
60th	17400	26.07	**26.01**
70th	20300	26.00	25.72
80th	23200	25.90	26.08
95.4th (2σ)	28025	25.91	25.81
Full dataset	29000	25.74	25.67

Table 6: Evaluation scores for model, finetuned from Mistral-7B-v0.1 directly on Extended Multi30K dataset. We performed log probability thresholds sweep on FLORES dev set. We demonstrate that data from the second phase alone is not enough to match the performance of our best checkpoint. Perplexity filtering improves downstream performance over training on full Extended Multi30K dataset.

Automated Extraction of Hypo-Hypernym Relations for the Ukrainian WordNet

Nataliia Romanyshyn, Dmytro Chaplynskyi, Mariana Romanyshyn
Ukrainian Catholic University, lang-uk, Grammarly, Ukraine
romanyshyn.n@ucu.edu.ua, chaplinsky.dmitry@gmail.com, mariana.romanyshyn@grammarly.com

Abstract

WordNet is a crucial resource in linguistics and natural language processing, providing a detailed and expansive set of lexico-semantic relationships among words in a language. The trend toward automated construction and expansion of WordNets has become increasingly popular due to the high costs of manual development. This study aims to automate the development of the Ukrainian WordNet, explicitly concentrating on hypo-hypernym relations that are crucial building blocks of the hierarchical structure of WordNet. Utilizing the linking between Princeton WordNet, Wikidata, and multilingual resources from Wikipedia, the proposed approach successfully mapped 17% of Princeton WordNet (PWN) content to Ukrainian Wikipedia. Furthermore, the study introduces three innovative strategies for generating new entries to fill in the gaps of the Ukrainian WordNet: machine translation, the Hypernym Discovery model, and the Hypernym Instruction-Following LLaMA model. The latter model shows a high level of effectiveness, evidenced by a 41.61% performance on the Mean Overlap Coefficient (MOC) metric. With the proposed approach that combines automated techniques with expert human input, we provide a reliable basis for creating the Ukrainian WordNet.

Keywords: WordNet, Ukrainian, Large Language Models, Hypernym Discovery, Lexicography

1. Introduction

WordNet is an invaluable resource that offers a well-structured and comprehensive list of lexical and semantic relationships between words in a language. This highly versatile resource is widely used by experts in linguistics, psychology, and natural language processing (NLP). Unlike a conventional thesaurus, WordNet arranges concepts based on their semantic and lexical relations to other concepts. Its broad applications include word sense disambiguation, machine translation, information retrieval, automatic text classification and summarization (Morato et al., 2004).

In recent years, scholars studying languages other than English have tried to tackle the issue of the absence of digital lexical databases similar to the Princeton WordNet (Miller, 1994). Due to the high expenses associated with creating taxonomies manually, there has been a growing interest in automatic methods for building and enhancing WordNets. Various researches have demonstrated the effectiveness of this approach in producing and expanding WordNets for multiple languages, such as Chinese (Wang and Bond, 2013), Arabic (Elkateb et al., 2006), and Urdu (Adeeba and Hussain, 2011).

The main objective of this paper is to introduce a new approach that utilizes multilingual resources from Wikidata[1] and Wikipedia[2] to build the Ukrainian WordNet. The primary focus of this work is on hypo-hypernym relations, a fundamental type of semantic relation for nouns that reflects the hierarchical structure of WordNet. It links general terms to more specific ones. For example, *rose* is a hyponym of *flower*, which is a hypernym of *rose*.

By concentrating on hypo-hypernymy, we aim to create a strong foundation for the Ukrainian WordNet that can be further expanded with other semantic relations in the future.

This work presents contributions that include:

- Automated methods for constructing and extending the Ukrainian WordNet, specifically linking techniques between Princeton WordNet, Wikidata and multilingual resources from Wikipedia, which have enabled the mapping of 17% of PWN to Ukrainian Wiki.

- Three strategies for generating candidate words to fill gaps in the constructed WordNet basis: machine translation, the Hypernym Discovery model, and Hypernym Instruction-Following LLaMA. The latter achieved high-performance results on the MOC metric (41.61%).

- Established a scalable foundation for creating a comprehensive and reliable WordNet for the Ukrainian language and published the artifacts of this work, including code and data, in the GitHub repo[3].

[1] https://www.wikidata.org/wiki/Wikidata:Main_Page

[2] https://www.wikipedia.org

[3] https://github.com/lang-uk/wikidrill

The Third Ukrainian Natural Language Processing Workshop (UNLP) @LREC-COLING-2024, pages 51–60
25 May, 2024. © 2024 ELRA Language Resource Association: CC BY-NC 4.0

The rest of the paper is organized as follows. Section 2 contains an overview of related work. Section 3 describes in detail the pipeline of our approach: compiling the basis for Ukrainian WordNet utilizing existing resources and methods for filling the gaps. We describe the statistics of the datasets obtained using the methodology described in the previous section and introduce the main experiments performed for the Hypernym Discovery task and instruction-tuned LLaMA in Section 4. We discuss the limitations of our approach, draw conclusions, and present future work in Section 5.

2. Related Work

The Princeton WordNet of the English language is widely regarded as the most comprehensive and established WordNet (Miller, 1994). With over 117,000 synonym sets and diverse relations, the PWN[4] has formed the benchmark for WordNets in other languages.

In the literature, two common approaches are used for building a WordNet for other languages: merge and expand (Vossen, 1997).

The merge approach involves developing a language-specific semantic network and integrating its synsets with those of the Princeton WordNet in the final stage of the project.

The expand approach involves mapping or translating local words to the synsets of an existing WordNet. While the expand approach is more efficient and requires less linguistic knowledge, it may result in less accurate representations of the semantic and lexical structure of the language.

Nevertheless, many WordNet developers opt for this approach due to the universal structure of lexical semantics that exists across languages (Youn et al., 2016).

The first published works on the construction of the Ukrainian WordNet were carried out in the 2010s.

Kulchytsky et al. (2010) conducted a study that focused on analyzing the relationships between nouns in the Princeton WordNet, selecting core nouns for the Ukrainian language, and organizing them into a hierarchical structure. The resulting WordNet-like dictionary includes 194 synsets, of which 183 are interconnected by hypo-hypernymy, 14 by antonymy, and 150 by meronymy/homonymy. The research in question was conducted manually using frequency dictionaries. Unfortunately, the project was not continued, and the results were not made publicly available.

Anisimov et al. (2013) described the development of a lexical semantic database for the Ukrainian language called UkrWordNet. The article focuses on the research and development of automated techniques for replenishing and extending UkrWordNet. The method developed for creating new nodes involved generating them from Ukrainian Wikipedia articles and binding them to the synsets of UkrWordNet. The paper also proposed a new measure of semantic similarity using latent semantic analysis (Deerwester et al., 1990) to improve the quality of the bindings. After manual post-processing, UkrWordNet contained over 82,000 synsets and approximately 145,000 nouns in the lexicon. Unfortunately, the work has never been publicly released.

In their article, Siegel et al. (2023) introduced Ukrajinet 1.0[5], a lexical database centered around physics terminology. The database contains 3,360 synonym sets of 8,700 words and shares a methodology similar to that used in creating OdeNet[6] for the German language (Siegel and Bond, 2021). However, Ukrajinet 1.0 does not include hypo-hypernym relations, essential for establishing a hierarchical structure of nouns within the WordNet framework.

Other developments in the field of the Ukrainian WordNet include materials[7] from theses of students of Lviv Polytechnic National University, but they are of a limited size.

Hence, developing an open-source WordNet for the Ukrainian language, with a representative number of relations, remains an ongoing area for research.

3. Proposed Approach

Our methodology for creating the basis of the Ukrainian WordNet builds on the expand approach. Figure 1 summarizes the proposed methodology. We propose utilizing the Princeton WordNet as a pivot structure, and linking it to Wikidata and Ukrainian Wikipedia. By mapping Ukrainian Wikipedia titles to synsets in the PWN and identifying hyponyms for each synset, a tree diagram is constructed using these resources. The resulting tree contains nodes that could not be linked to Ukrainian Wikipedia and thus lack a Ukrainian equivalent. We call them gap nodes and further propose the Gap Ranking algorithm to identify the best gap nodes for filling. To generate candidate words to fill these gaps, several strategies are proposed. The first strategy utilizes English lemmas translated into Ukrainian with Google Translate, Bing, and DeepL. The second strategy adapts the Hypernym Discovery task for Ukrainian and gener-

[4] https://wordnet.princeton.edu

[5] https://github.com/hdaSprachtechnologie/ukrajinet

[6] https://github.com/hdaSprachtechnologie/odenet

[7] https://github.com/lang-uk/wordnet/tree/main/resources

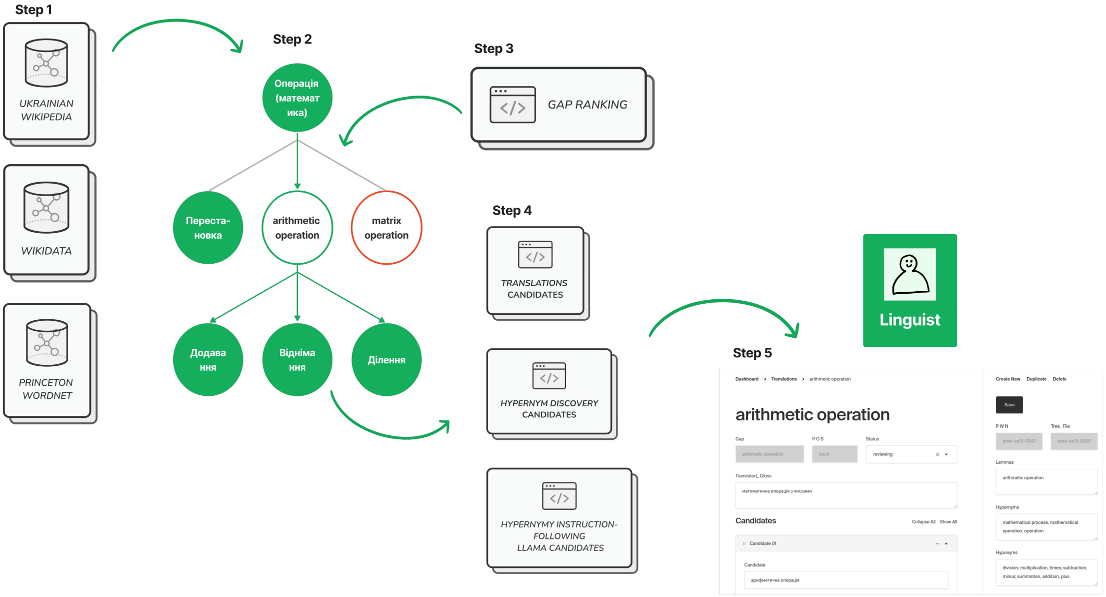


Figure 1: Overview of the proposed methodology for developing the Ukrainian WordNet foundation through integration with Princeton WordNet, Wikidata, and Wikipedia.

ates candidates given the gap hyponym. The third strategy generates hypernym candidates with the Instruction-Following LLaMA model. The hypernym candidates generated via the three strategies are then aggregated in a MongoDB and surfaced in an annotation tool built with Payload CMS to streamline further human annotation. Overall, the proposed approach combines automated techniques with expert human input to create a comprehensive and reliable resource for the Ukrainian language.

3.1. PWN and Wikidata

Our methodology leverages the linking between Princeton WordNet and Wikidata, as proposed by McCrae and Cillessen (2021). First, we utilize the synset's ID to identify the PWN synset linked with Wikidata. This link allows us to search for the corresponding Ukrainian Wikipedia article using the Wikidata Q identifier. At this point, we encounter two possible scenarios. If the search yields a result, we acquire a word that can populate a node in our lexical tree. However, if the search does not provide any results, we temporarily store the English lemma from PWN at this node, intending to address this issue later. The techniques for filling these gaps are elaborated upon in subsequent sections. Once we have identified a linked synset, we proceed to discover hyponyms associated with the given synset ID.

3.2. Gap Ranking

The Gap Ranking algorithm aims to identify the most suitable gap nodes for filling, specifically those with the highest number of non-gap children in the given tree. We consider these most suitable because filling them creates the highest number of links. The algorithm utilizes a depth-first search (DFS) tree traversal method to determine the ideal path. Beginning from the root node, it recursively navigates the tree, viewing each node as a potential gap node. For each gap node, the algorithm computes the number of valid pairs of nodes in its subtree by considering its non-gap children. The algorithm then ranks the gap nodes based on the number of identified valid pairs.

This metric is instrumental in identifying the gap nodes with the greatest potential for enhancing the quality of the Ukrainian WordNet. With this algorithm, we found that completing 793 gaps would result in 5403 new hyper-hyponym pairs in the Ukrainian WordNet.

3.3. Candidate Generation

We used two methods to generate candidates to fill gaps in our lexical resource. The first method involved automatic translation from English to Ukrainian. The second method used the hyponym of the gap to generate hypernyms with the help of the Hypernym Discovery model and Instruction-Following LLaMA.

Gap	DeepL Direct	DeepL Contextualized	Translated PWN3.1
performance	продуктивність produktyvnist	вистава vystava	вистава, спектакль vystava, spektakl
head cabbage	качанна капуста kachanna kapusta	качанна капуста kachanna kapusta	головна капуста holovna kapusta
agency	агентство ahentstvo	агентство ahentstvo	офіс, орган ofis, orhan

Table 1: Comparison examples of gap translations obtained using machine translation methods. All terms are nouns. The gap is identified as the most optimal for filling using the algorithm described in Section 3.2

3.3.1. Machine Translations

To run automatic translation, we utilized three distinct methods. Initially, we accessed the existing Ukrainian translation[8] of Princeton WordNet 3.1, which was developed with Google Translate and Bing. Subsequently, we relied on the neural machine translation capabilities of DeepL (Ronzon, 2018). This process entailed directly translating individual lemmas and creating contextual sentences in the format of *"<Synset lemmas> is a <PWN gloss>."*, from which we extracted the first lemma and recorded it as a candidate for the gap.

Ultimately, this approach enabled us to promptly produce a list of potential translations for the gap nodes, although due to the lack of specialized training or fine-tuning of the machine translation models for the Ukrainian language their accuracy remains arguable. For example, machine translation can generate Russianism[9], such as "kachanna kapusta" seen in row 2 of Table 1, or false concepts like "holovna kapusta," which do not exist in Ukrainian. Furthermore, the issue of ambiguity, demonstrated in rows 1 and 3, presented challenges by offering multiple possible senses. Although employing specialized Word Sense Disambiguation systems, as suggested by Laba et al. (2023), could mitigate this issue, exploring such solutions falls beyond the scope of this paper.

3.3.2. Hypernym Discovery and LLaMA

To perform Hypernym Discovery in the Ukrainian language, we adopted the setting provided for this task by Camacho-Collados et al. (2018). We utilized the supervised part of the model proposed by Bernier-Colborne and Barrière (2018), the SemEval-2018 Task 9 winners. Their approach uses pre-trained word embeddings and projection learning to discover the hypernyms of a given query (hyponym).

Pretrained large language models (LLMs) have showcased remarkable results in various natural language processing (NLP) tasks, leading us to explore their potential for Hypernym Discovery. A previous study conducted by Hanna and Mareček (2021) utilized a prompting methodology to investigate BERT's (Devlin et al., 2019) understanding of hypernymy. Our research focused on the potential of another advanced LLM, multilingual LLaMA (Touvron et al., 2023), which has exhibited exceptional performance on various NLP benchmarks. Instead of prompting, we opted to fine-tune LLaMA by providing hypernym instructions to determine if it can suggest hypernyms.

3.4. Evaluation Metrics

Camacho-Collados et al. (2018) proposed evaluating the Hypernym Discovery systems as a soft ranking problem. This involved utilizing the top N[10] hypernyms generated by the model and evaluating performance using Information Retrieval (IR) metrics:

1. **Mean Reciprocal Rank** (MRR) measures how well a system is able to rank the relevant hypernyms by rewarding the position of the first correct result in the ranked list of outcomes.

2. **Mean Average Precision** (MAP) measures the average correctness of retrieved hypernyms for each query word and averages these across all dataset queries.

3. **Precision at k** (P@k) measures the number of correctly retrieved hypernyms at different cut-off thresholds.

To better understand the model's ability to predict relevant hypernyms regardless of their order, we propose the Mean Overlap Coefficient (MOC) as an additional evaluation criterion:

$$MOC = \frac{1}{|Q|}\sum_{i=1}^{|Q|}\frac{|GT_i \cap P_i|}{|GT_i|}, \quad (1)$$

where Q represents the number of queries, GT represents the set of ground truth hypernyms, and

[8] https://github.com/lang-uk/wordnet/tree/main/pwn_translated_basic

[9] https://en.wikipedia.org/wiki/Russianism

[10] Set the value to 6 for our experiments.

P represents the predictions for a given input term. The numerator calculates the number of common hypernyms between the ground truth and predicted sets, while the denominator ensures that the metric is normalized by the size of the ground truth set.

For our specific task of generating candidates for professional annotators, we found the MOC score to be the most helpful metric as it indicates the proportion of relevant values predicted regardless of their order.

We used the same metrics to measure the performance of the Instruction-Following LLaMA.

4. Experimental Results

4.1. WordNet Basis

To link the data from PWN, Wikidata, and Ukrainian Wikipedia, we implemented a Python scraper using the web-crawling framework Scrapy (Hoffman et al., 2008), wtf_wikipedia library (Kelly, 2017) for Wikipedia parsing, wn package (Goodman and Bond, 2021), which provides an interface to WordNet data, and an RDF (Resource Description Framework) query language SPARQL (Prud'hommeaux and Seaborne, 2008).

We managed to link 17% of the Princeton WordNet, resulting in **21,015** synsets forming the foundation of the Ukrainian WordNet. Out of the 127,020 PWN3.1 synsets, we could link 23% to Wikidata; subsequently, 17% of those synsets were connected to the Ukrainian Wikipedia. These results reflect the linking percentage as of April 2023. Due to the dynamic nature of Wikidata, with its continuous updates and expansions, subsequent iterations of this experiment could yield an even higher proportion of linked synsets. Table 2 provides an overview of the general statistics.

	# synsets	**% synsets**
PWN3.1	127,020	100%
-> Wikidata	29,730	23%
-> Ukrainian Wiki	21,015	17%

Table 2: General statistics related to the development of the Ukrainian WordNet basis, including the total number of synsets in the PWN3.1, the number and percentage of synsets linked to Wikidata and the Ukrainian Wikipedia.

In addition, we developed a dataset of Ukrainian Hypernymy Pairs consisting of noun pairs that express hypernymy relationships between words. Please, refer to Table 3 for the detailed dataset statistics. We maintained the partition of hypernyms and hyponyms with their instances that is offered by PWN in our dataset. In this split, the instance hypernym denotes a reflexive type, while an instance hyponym represents a specific instance of something. For example, the instance hypernym of *Dnipro River* is *river*. We identified a few data samples where the word on the left is the same as the word on the right (e.g., <river, river> pair), resulting from multiple WordNet IDs linking to the same Wikidata page. To improve the quality of the dataset, we removed such entries. The dataset[11] is available for public use through the Hugging Face platform and can be particularly useful for the Hypernym Detection task, which involves presenting a model with pairs of words and asking it to determine whether a specific relationship exists between them.

Relation Type	**% pairs**
Hypernym-Hyponym	6,906
Co-Hyponyms	42,860
Hypernym-Instance	2,971
Co-Instances	22,927

Table 3: Ukrainian Hypernymy Pairs dataset statistics. This table presents the number of word pairs obtained for each type of relationship.

4.2. Hypernym Discovery

To advance research in the field of hypernym discovery, *SemEval-2018 Task 9*[12] was organized (Camacho-Collados et al., 2018). The participants were asked to build a system that discover suitable hypernyms from a target corpus given an input term. The organizers (of the task) provided a reliable framework for evaluating proposed models with the IR metrics described in Section 3.4. To perform Hypernym Discovery in the Ukrainian language, we adopted the setting provided for this task.

4.2.1. Dataset Creation

Following the approach of Camacho-Collados et al. (2018) in SemEval, our data gathering process involved a series of sequential steps, beginning with the compilation of a vocabulary. Our objective was to establish an all-encompassing list of prospective hypernyms by identifying words that appeared at least five times within the chosen corpus. To do so, we utilized UberText 2.0[13] (Chaplynskyi, 2023), a corpus that boasts 31GB of data and around 2.5 billion tokens, which accurately represents the variety and abundance of the Ukrainian language.

The original Hypernym Discovery dataset consisted of two main components: input hyponym

[11] https://huggingface.co/datasets/lang-uk/hypernymy_pairs

[12] https://competitions.codalab.org/competitions/17119

[13] https://lang.org.ua/en/ubertext/

along with its type and gold hypernyms. The type is either a concept (hyponym) or a named entity (instance). Utilizing the created Ukrainian WordNet basis, we automated the extraction of these terms, including direct and indirect hypernyms up to five nodes deep to mirror the original setup. The refinement process involved:

- Excluding overly broad terms from the upper levels of the WordNet hierarchy;
- Normalizing entries by removing bracketed information that comes from the Wikipedia titles;
- Discarding non-unigram terms;
- Eliminating entries composed of Latin characters, which usually denote animal species, plants, etc.;
- Excluding terms without a direct hypernym relation.

We maintained a frequency threshold, requiring terms to appear at least five times in the UberText corpus. The classification of input terms, as instances or hyponyms was determined automatically via synset relation parameters.

The resulting dataset[14], consisting of 4,890 samples, offers a balanced split for training and test sets alongside a smaller trial set for developmental evaluation.

4.2.2. Model Setup

This work employs the supervised part of the Hybrid Approach to Hypernym Discovery, developed by Bernier-Colborne and Barrière (2018) and accessible on GitHub[15].

To establish a baseline, we utilized 200-dimensional word2vec embeddings with a skip-gram model, trained according to the specifications outlined in the abovementioned research (HD_Baseline). As the next step, we chose to explore the fasttext embeddings, which are advantageous for Ukrainians because of their ability to capture subword information (HD_Fasttext). Hyperparameters were based on previous studies (Romanyshyn et al., 2023), and the vector size was increased to 300 dimensions.

4.2.3. Results

Table 4 summarizes the model's performance by each metric. Overall, we can see that the HD_Baseline model performed the best overall, but HD_Fasttext achieved the highest score in terms of the MOC metric.

4.3. Hypernym Instruction-Following LLaMA

We utilized a parameter-efficient tuning technique called low-rank adaptation (LoRA) to fine-tune LLaMA-7B on hypernymy instructions (Hu et al., 2021). This approach involves freezing the pre-trained model's weights and adding trainable rank decomposition matrices into each layer of the transformer architecture, reducing the number of trainable parameters for downstream tasks (Maurya, 2023).

4.3.1. Intructions Dataset

We developed instruction datasets of three different types and ran experiments on them. The data for Hypernym Discovery was used as a basis. The main difference is that we merged training and trial (dev) sets into one.

Lean Approach. Our initial method involved generating simple prompts that instructed the model to provide a specific number of hypernyms for a given term. For example, we would ask the model to *"Generate six hypernyms for 'lavender'."* While we could generate **2,490** instructions, the model's performance was poor.

Full Setup. We improved the instruction set by creating 19 distinct patterns for each query, using ChatGPT for initial generation, and manually validating the results. This approach greatly enhanced model performance, resulting in **47,310** input prompts. We utilized various query formats, including *"What are broader terms for 'lavender'?"* to broaden the model's comprehension across similar phrasings.

Multiple Relations. Building on our enhanced approach, we introduced instructions for hypernyms, hyponyms, and co-hyponyms, maintaining 19 hypernym patterns while adding 13 for co-hyponyms and 14 for hyponyms. This resulted in **78,149** samples, but we noticed a dip in performance, suggesting potential overgeneralization. Further research is needed to balance instruction diversity and specificity effectively.

4.3.2. Results

Our testing across the Lean, Full, and Multiple models utilized identical input queries and gold hypernyms from the Hypernym Discovery dataset, with tailored strategies to mitigate specific challenges encountered in each setup.

For the Lean model, given its simplicity, we applied a heuristic of repeating each instruction three

[14] https://github.com/lang-uk/wikidrill/tree/main/hypernymy_discovery/hd_dataset

[15] https://github.com/gbcolborne/hypernym_discovery

	MOC	MRR	MAP	P@1	P@3	P@6
HD_Baseline	26.55	**29.23**	**20.84**	**25.25**	**20.22**	**19.3**
HD_Fasttext	**27.63**	28.7	19.87	22.42	19.53	18.76

Table 4: Our Hypernym Discovery systems performance on the test set. HD_Baseline refers to the model with word2vec embeddings, HD_Fasttext to the one using fasttext. The best score for each model is marked in **bold**.

	MOC	MRR	MAP	P@1	P@3	P@6
LLaMA_Hypernymy_Lean	6.38	4.54	2.92	3.08	2.88	2.8
LLaMA_Hypernymy_Full	**41.61**	**42.6**	**36.74**	**39.0**	**36.27**	**35.93**
LLaMA_Hypernymy_Multiple	37.07	35.48	31.19	30.42	31.72	30.8

Table 5: The LLaMA fine-tuning results with hypernymy instructions using different setups. The LLaMA_Hypernymy_Lean setup only uses the most basic hypernymy instructions, while LLaMA_Hypernymy_Full includes 19 instruction patterns for a single input query. In the Multiple setup, three relation types were used in addition to diverse patterns.

times and aggregating unique hypernym candidates to counteract issues of non-responses or repetitive outputs. This approach aimed to enhance result reliability.

In contrast, while not facing duplication issues, the Full and Multiple setups sometimes produced no candidates. To address this, we diversified the testing instruction set, employing four varied prompts to elicit hypernyms, thus balancing output richness and relevance. This method prioritized candidate frequency and maintained the model's original proposal order for equally frequent terms, aligning closely with the evaluative framework of the Hypernym Discovery task. The prompts were as follows:

1. Надай мені декілька гіперонімів до слова "input_term". (Give me some hypernyms to the word "input_term".)
2. Надай мені шість гіперонімів до слова "input_term". (Give me six hypernyms to the word "input_term".)
3. Які слова є гіперонімами поняття "input_term"? (Which words are hypernyms of the term "input_term"?)
4. Які загальні поняття описують слово "input_term"? (What general concepts describe the word "input_term"?)

Table 5 showcases the superior performance of the LLaMA_Hypernymy_Full model across all metrics, reflecting the effectiveness of our comprehensive and nuanced instruction and evaluation methodology.

4.4. Error Analysis

In addition to analyzing quantitative results, we also performed a qualitative evaluation of the outputs produced by our top-performing models based on MOC scores from our experiments. Figure 2 presents a metrics comparison of HD_Fasttext and LLaMA_Hypernymy_Full.

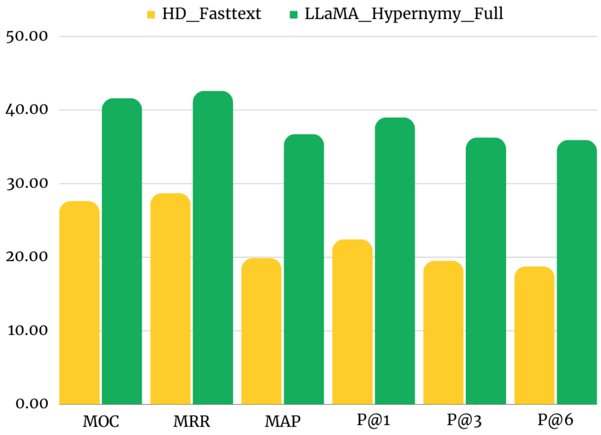


Figure 2: Metrics comparison of the two top-performing models based on MOC score for all entity types.

We randomly sampled several examples from our testing dataset to further investigate the models' predictions.

As we can see from Table 6, the model tends to overfit to frequently occurring hypernyms such as метрополіс (metropolis), муніципалітет (municipality), дистрикт (district), and організм (organism), resulting in incorrect predictions. However, when these candidates are true hypernyms, the model generally ranks them as the top predictions. Moreover, the MOC metric proves helpful in cases where the ground truth contains only one hypernym, such as the річка (river) hypernym for the Ориноко (Orinoco) query and the model ranks it as the last candidate. Notably, the model can also suggest

Query	OC	HD_Fasttext Predictions
		High OC examples
молюски	67	організм, їжа, безхребетні, дистрикт, артефакт, рослини
shellfish		organism, food, invertebrates, district, artifact, plants
Ориноко	100	організм, дистрикт, муніципалітет, метрополіс, артефакт, річка
Orinoco		organism, district, municipality, metropolis, artifact, river
		Low OC examples
Гіпатія	17	дистрикт, організм, муніципалітет, артефакт, їжа, метрополіс
Hypatia		district, organism, municipality, artifact, food, metropolis
Сапфо	0	метрополіс, артефакт, організм, дистрикт, муніципалітет, їжа
Sappho		metropolis, artifact, organism, district, municipality, food
Query	**OC**	**LLaMA_Hypernymy_Full Predictions**
		High OC examples
холангіт	100	симптом, запалення, хвороба
cholangitis		symptom, inflammation, disease
Неккар	100	річка
Neckar		river
		Low OC examples
метамфетамін	0	опіати, наркотик, анальгетики
methamphetamine		opiates, narcotic, analgesics
Сент-Джонс	0	озеро, річка
St. John's		lake, river

Table 6: Examples of predictions made by the HD_Fasttext and LLaMA_Hypernymy_Full models, showing input queries, overlap coefficients (OC), and top predicted hypernyms. High OC values indicate accurate predictions, while low values reflect mismatches. Correct predictions are underlined.

relevant candidates absent in the ground truth, as observed in the Low OC Entity examples, where it proposed організм (organism) as a hypernym for Сапфо (Sappho), which is not the direct hypernym but still relevant as it is the same case as for query hyponym Гіпатія (Hypatia), where the організм (organism) was present in gold hypernyms.

The instruction-following LLaMA model appears to be confident in its predictions, often providing the same answer for four instructions, resulting in fewer variants of predictions. For instance, it predicts the single hypernym річка (river) for the input term Неккар (Neckar). Furthermore, in this scenario, the memorization problem of frequent hypernyms is less noticeable.

In addition, the model can predict relevant hypernyms that are not present in the ground truth set, such as хвороба (disease) for the input word холангіт (cholangitis) and наркотик (narcotic) for метамфетамін (methamphetamine).

Another challenge the model faces is the ambiguity of some hyponyms. For instance, by providing the hypernym річка (river) for the entity Сент-Джонс (St. John's), the model may have referred to an actual river in Florida, United States, while our data referred to a city in Canada.

5. Discussion and Conclusion

This paper reports on the ongoing efforts in building the Ukrainian WordNet. We proposed a data-driven approach for automated hypernym hierarchy construction. By mapping PWN, Wikidata, and Wikipedia, we have created a robust foundation for this new WordNet resource. Additionally, we have developed a simple Gap Ranking algorithm to determine the best gap nodes for filling.

To generate candidates for filling the gaps, we have explored various techniques, including machine translation that uses the current missing node in the tree and two others that use information about its children — Hypernym Discovery and Instruction-Following LLaMA.

To adapt SemEval 2018 Task 9: Hypernym Discovery to the Ukrainian language, we have created Hypernym Discovery datasets and utilized an existing large language corpus Ubertext2.0.

Furthermore, we have investigated the capabilities of state-of-the-art LLMs for solving the Hypernym Discovery task. We have demonstrated how to construct a sufficiently large set of instructions from an initial small dataset and how LLMs can be fine-tuned to create a chatbot-like assistant specializing in a particular hypernym suggestion task.

5.1. Limitations

Please be aware that our work is subject to certain limitations.

To establish a WordNet basis, we initially mapped the Ukrainian language to English, which may not fully capture all linguistic nuances and cultural phenomena and could contain errors. Hence, it is crucial to have further professional verification and input from linguists.

Another restriction is that, according to our approach, each obtained synset is represented by only one lemma due to Wikipedia articles being primarily represented by one word and linking is on the synset level. As a result, additional effort is required to add synonyms to the obtained lemma-synsets.

Overall, our approach is limited to only creating hypo-hypernym relations. Further research is needed to include other lexico-semantic relations. Nevertheless, it is important to note that the proposed method has the potential to be adapted for other languages as long as comprehensive Wikipedia data is available.

5.2. Future Work

As creating WordNet is a complex and lengthy process, there is ample opportunity for future research to improve its coverage and quality. To this end, we have identified critical areas for improvement that we hope to focus on going forward:

1. One priority is to leverage Wikipedia as a constantly updated resource by rerunning the linking algorithm of Wikidata and Ukrainian Wiki to obtain more initial pairs. Additionally, we can independently add links to Wikidata using annotated gaps, thereby enhancing this resource.

2. Exploring larger LLaMA or other open-source language models is another promising direction that can significantly boost performance on our task.

3. An essential next step is to create a high-quality and comprehensive manual for annotators, which will take the WordNet development pipeline to a new level.

4. Ultimately, WordNet should have a user-friendly interface accessible to the general public.

6. Bibliographical References

Farah Adeeba and Sarmad Hussain. 2011. Experiences in building Urdu WordNet. In *Proceedings of the 9th Workshop on Asian Language Resources*, pages 31–35, Chiang Mai, Thailand. Asian Federation of Natural Language Processing.

Anatoly Anisimov, Oleksandr Marchenko, Andrey Nikonenko, Elena Porkhun, and Volodymyr Taranukha. 2013. Ukrainian wordnet: Creation and filling. In *Proceedings of the 10th International Conference on Flexible Query Answering Systems - Volume 8132*, FQAS 2013, page 649–660. Springer-Verlag.

Gabriel Bernier-Colborne and Caroline Barrière. 2018. CRIM at SemEval-2018 task 9: A hybrid approach to hypernym discovery. In *Proceedings of the 12th International Workshop on Semantic Evaluation*, pages 725–731, New Orleans, Louisiana. Association for Computational Linguistics.

Jose Camacho-Collados, Claudio Delli Bovi, Luis Espinosa-Anke, Sergio Oramas, Tommaso Pasini, Enrico Santus, Vered Shwartz, Roberto Navigli, and Horacio Saggion. 2018. SemEval-2018 task 9: Hypernym discovery. In *Proceedings of the 12th International Workshop on Semantic Evaluation*, pages 712–724, New Orleans, Louisiana. Association for Computational Linguistics.

Dmytro Chaplynskyi. 2023. Introducing UberText 2.0: A corpus of modern Ukrainian at scale. In *Proceedings of the Second Ukrainian Natural Language Processing Workshop*, pages 1–10, Dubrovnik, Croatia. Association for Computational Linguistics.

Scott C. Deerwester, Susan T. Dumais, Thomas K. Landauer, George W. Furnas, and Richard A. Harshman. 1990. Indexing by latent semantic analysis. *J. Am. Soc. Inf. Sci.*, 41:391–407.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4171–4186, Minneapolis, Minnesota. Association for Computational Linguistics.

Sabri Elkateb, William Black, Piek Vossen, David Farwell, Horacio Rodríguez, Adam Pease, Musa Alkhalifa, and Christiane Fellbaum. 2006. Arabic WordNet and the challenges of Arabic. In *Proceedings of the International Conference on the Challenge of Arabic for NLP/MT*, pages 15–24, London, UK.

Michael Wayne Goodman and Francis Bond. 2021. Intrinsically interlingual: The wn python library for wordnets. In *Proceedings of the 11th Global Wordnet Conference*, pages 100–107, University of South Africa (UNISA). Global Wordnet Association.

Michael Hanna and David Mareček. 2021. Analyzing BERT's knowledge of hypernymy via prompting. In *Proceedings of the Fourth BlackboxNLP Workshop on Analyzing and Interpreting Neural Networks for NLP*, pages 275–282, Punta Cana, Dominican Republic. Association for Computational Linguistics.

Pablo Hoffman, Daniel Graña, and Martin Olveyra. 2008. A fast and powerful scraping and web crawling framework.

Edward J. Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, and Weizhu Chen. 2021. Lora: Low-rank adaptation of large language models. *CoRR*, abs/2106.09685.

Spencer Kelly. 2017. Wtf_wikipedia: A pretty-committed wikipedia markup parser.

I. Kulchytsky, A. Romaniuk, and Kh. Khariv. 2010. Rozroblennia wordnet-podibnoho slovnyka ukrainskoi movy [developing a wordnet-like dictionary of ukrainian]. In *Bulletin of Lviv Polytechnic National University*, pages 306–318.

Yurii Laba, Volodymyr Mudryi, Dmytro Chaplynskyi, Mariana Romanyshyn, and Oles Dobosevych. 2023. Contextual embeddings for Ukrainian: A large language model approach to word sense disambiguation. In *Proceedings of the Second Ukrainian Natural Language Processing Workshop (UNLP)*, pages 11–19, Dubrovnik, Croatia. Association for Computational Linguistics.

Aniket Maurya. 2023. Accelerating llama with fabric: A comprehensive guide to training and fine-tuning llama.

John P. McCrae and David Cillessen. 2021. Towards a linking between WordNet and Wikidata. In *Proceedings of the 11th Global Wordnet Conference*, pages 252–257, University of South Africa (UNISA). Global Wordnet Association.

Jorge Morato, Miguel Ángel Marzal, Juan Lloréns, and José Moreiro. 2004. WordNet Applications. pages 270–278.

Eric Prud'hommeaux and Andy Seaborne. 2008. SPARQL Query Language for RDF. W3C Recommendation. `http://www.w3.org/TR/rdf-sparql-query/`.

Nataliia Romanyshyn, Dmytro Chaplynskyi, and Kyrylo Zakharov. 2023. Learning word embeddings for Ukrainian: A comparative study of fastText hyperparameters. In *Proceedings of the Second Ukrainian Natural Language Processing Workshop*, pages 20–31, Dubrovnik, Croatia. Association for Computational Linguistics.

Thomas Ronzon. 2018. Deepl - überraschend anders. *Javaspektrum*, (2):69.

Melanie Siegel and Francis Bond. 2021. OdeNet: Compiling a GermanWordNet from other resources. In *Proceedings of the 11th Global Wordnet Conference*, pages 192–198, University of South Africa (UNISA). Global Wordnet Association.

Melanie Siegel, Maksym Vakulenko, and Jonathan Baum. 2023. Towards UkrainianWordNet: Incorporation of an existing thesaurus in the domain of physics. In *Proceedings of the 19th Conference on Natural Language Processing (KONVENS 2023)*, pages 121–126, Ingolstadt, Germany. Association for Computational Lingustics.

Hugo Touvron, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Timothée Lacroix, Baptiste Rozière, Naman Goyal, Eric Hambro, Faisal Azhar, Aurelien Rodriguez, Armand Joulin, Edouard Grave, and Guillaume Lample. 2023. Llama: Open and efficient foundation language models.

Shan Wang and Francis Bond. 2013. Building the Chinese open Wordnet (COW): Starting from core synsets. In *Proceedings of the 11th Workshop on Asian Language Resources*, pages 10–18, Nagoya, Japan. Asian Federation of Natural Language Processing.

Hyejin Youn, Logan Sutton, Eric Smith, Cristopher Moore, Jon F. Wilkins, Ian Maddieson, William Croft, and Tanmoy Bhattacharya. 2016. On the universal structure of human lexical semantics. *Proceedings of the National Academy of Sciences*, 113(7):1766–1771.

7. Language Resource References

Miller, George A. 1994. *WordNet: A Lexical Database for English*. ISLRN 379-473-059-273-1.

P.J.T.M. Vossen. 1997. *EuroWordNet: a multilingual database for information retrieval*. Vrije Universiteit, ISLRN 129-018-230-332-2. Proceedings of the DELOS workshop on Cross-language Information Retrieval, March 5-7, 1997, Zurich.

Ukrainian Visual Word Sense Disambiguation Benchmark

**Yurii Laba, Yaryna Mohytych, Ivanna Rohulia, Halyna Kyryleyza,
Hanna Dydyk-Meush, Oles Dobosevych, Rostyslav Hryniv**
Ukrainian Catholic University
2A Kozelnytska st., Lviv, Ukraine, 79026
{laba, mohytych.hn, rohulia.hn, kyryleyza, hanna_dydykmeush, dobosevych, rhryniv}@ucu.edu.ua

Abstract

This study presents a benchmark for evaluating the Visual Word Sense Disambiguation (Visual-WSD) task in Ukrainian. The main goal of the Visual-WSD task is to identify, with minimal contextual information, the most appropriate representation of a given ambiguous word from a set of ten images. To construct this benchmark, we followed a methodology similar to that proposed by Raganato et al. (2023), who previously introduced benchmarks for the Visual-WSD task in English, Italian, and Farsi. This approach allows us to incorporate the Ukrainian benchmark into a broader framework for cross-language model performance comparisons. We collected the benchmark data semi-automatically and refined it with input from domain experts. We then assessed eight multilingual and multimodal large language models using this benchmark. All tested models performed worse than the zero-shot CLIP-based baseline model (Radford et al., 2021) used by Raganato et al. (2023) for the English Visual-WSD task. Our analysis revealed a significant performance gap in the Visual-WSD task between Ukrainian and English.

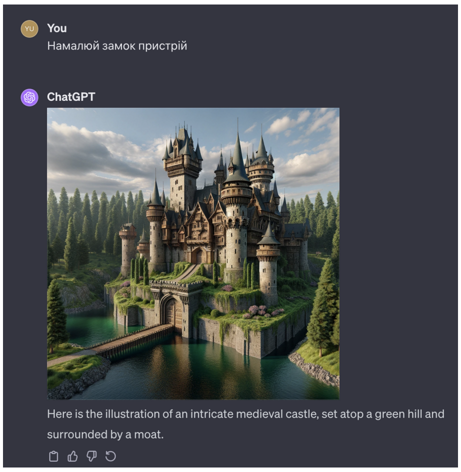
Keywords: Visual-WSD, Multimodal LLM, Benchmark, Ukrainian

1. Introduction

The rise of Large Language Models (LLMs) represents a notable advancement in Natural Language Processing (NLP), catalyzing outstanding progress in text understanding and synthesis. Building upon this milestone, Multimodal Large Language Models (MLLMs) emerged as a pivotal development. MLLMs exhibit remarkable efficacy across diverse domains, including but not limited to image classification, object recognition, and tasks integrating textual and visual inputs.

Despite the distinguished milestones achieved by MLLMs/LLMs, they confront various issues that can detrimentally impact the performance of the models. One such challenge involves problems associated with hallucination generation (Huang et al., 2023). This phenomenon frequently leads to producing content that deviates from real-world facts or user inputs. It causes significant challenges for the practical usage of these models and evokes concerns on the reliability of LLMs in real-world applications. Furthermore, MLLMs/LLMs demonstrate notably inferior performance when engaged in processing low-resource languages like Ukrainian.

In our study, we have opted to examine the extent of hallucinations linked to the utilization of homonyms in the Ukrainian language. Figure 1 demonstrates visual hallucination of GPT4-Vision model. This type of hallucination occurred during the generation of an image representing Замок (castle, translit: zamok) with the intended meaning of Пристрій (device, translit: prystriy). In English, this would correspond to the term padlock.



You
Намалюй замок пристрій

ChatGPT

Here is the illustration of an intricate medieval castle, set atop a green hill and surrounded by a moat.

Figure 1: An illustration of GPT4-Vision visual hallucination caused by ambiguous target word.

Such hallucinations may arise from several potential reasons. One contributing factor could be the uneven frequency of usage among homonym pairs, wherein certain homonyms are more frequently employed than others. Another contributing factor might be the training of LLMs in multiple languages. The scarcity of available data for low-resource languages often leads subword tokenizers to generate an imbalanced subword vocabulary. As a result, LLMs encounter difficulties in generating high-quality representations for tokens from low-resource languages (Hangya et al., 2022; Holmström et al., 2023). Another challenge emerges during domain adaptation. Employing a uniform, single approach to adapt a Language Model (LM) across multiple languages, often referred to as the

The Third Ukrainian Natural Language Processing Workshop (UNLP) @LREC-COLING-2024, pages 61–66
25 May, 2024. © 2024 ELRA Language Resource Association: CC BY-NC 4.0

"one-size-fits-all" method, may prove ineffective due to the unique semantic nuances present in each language (Grangier and Iter, 2022). For instance, a direct translation of a legal term from English to Ukrainian might overlook certain intricate meanings or contextual connotations pertinent to the Ukrainian legal context.

The principal aim of our investigation is to construct a benchmark for gauging the issue of hallucinations related to MLLMs/LLMs concerning homonyms. This is achieved by assessing the efficacy of relevant models in tackling the Visual-WSD task. This task involves providing a target ambiguous word and a restricted context, accompanied by ten images. The Visual-WSD task requires the identification of the most relevant image corresponding to the intended meaning of the ambiguous word. In addition, we give an exhaustive account of a comparative analysis of the performance of several relevant Multilingual MLLMs on the developed benchmark and demonstrate that there is a notable disparity in performance in the Visual-WSD task between Ukrainian and English.

2. Related Works

Word Sense Disambiguation (WSD) is a general task of identifying the intended sense of a polysemantic word in a particular context, typically from a predetermined sense inventory. Although recent advances of LLM's have naturally led to multimodal settings of that task, until now, the research on WSD in the Ukrainian language has been primarily focused on textual modality alone.

One of the approaches (Laba et al., 2023) employs a fine-tuning of LM in a semi-supervised manner to improve its performance on the WSD task in Ukrainian. In their study, the authors indicated that language model-based solutions outperform traditional methods (Barba et al., 2021; Moro et al., 2014). They also compiled a benchmark based on the Dictionary of Ukrainian Language (SUM) (Rusanivskyi, 2010), which can be used to assess the performance of various models on the textual WSD task in Ukrainian.

The popularity of the visual generative models like DALL-E (Ramesh et al., 2021) or Stable Diffusion (Rombach et al., 2022) and the abundance of visual information around was probably the reason why Task 1 of the SemEval-2023 conference was on the Visual-WSD. The methodology of the competition (i.e., the dataset and evaluation metrics) is described in Raganato et al. (2023). The datasets are available in English, Italian, and Farsi, enabling testing different approaches to the solution of the Visual-WSD task.

Most of the suggested solutions were based upon the foundation of the CLIP model (Radford et al., 2021) utilizing Teacher Learning technique (Hinton et al., 2015). Teacher Learning is a domain-agnostic machine learning technique that transfers knowledge from a pre-trained teacher model to a new student model. This method has been successfully applied not only in multimodal settings but also in various other tasks (Reimers and Gurevych, 2019; Wu et al., 2020).

Carlsson et al. (2022) give a successful example of applying the Teacher Learning technique with Multilingual CLIP. Their approach relies solely on machine translation and thus eliminates the need for visual data in the target language. The proposed objective is to reduce the Mean Squared Error (MSE) between the embeddings produced by the teacher model and the student model for translated texts. Rather than optimizing directly for cosine similarity, as in the original CLIP training, MSE is employed due to its proven effectiveness in providing a better learning response.

A comparable but slightly different approach was proposed by Reimers and Gurevych (2020). Their aim was to minimize the MSE between the embeddings generated by the teacher model and those produced by the student model for both the source sentences and their translations.

The recent LLaVA-1.5 model (Liu et al., 2023) utilizes a completely different approach based on visual instruction tuning. This model operates primarily as a standard causal LM, taking language instructions (a user text prompt) as input and generating a language response. Its ability to process images is facilitated by an independent vision encoder model, which converts images into language tokens that are seamlessly integrated into the user text prompt. The LM and vision encoder of LLaVA are built upon two reference models known as Vicuna (Zheng et al., 2024) and CLIP (Radford et al., 2021), respectively.

Even though some of the mentioned approaches to the Visual-WSD task are language-agnostic, the efficiency of the respective models in completing instructions in one language cannot be evaluated on datasets in another language, since most challenges are caused by word polysemy, which is typically language-specific. To the best of our knowledge, there are currently no evaluation resources in languages other than those proposed in the SemEval-2023 Task 1, and this hampers research of multilingual and multimodal LMs. That was one of the main motivations for us to create an evaluation dataset for the Visual-WSD task in Ukrainian following the methodology of Raganato et al. (2023) and to benchmark on it the available approaches. We hope these resources will facilitate future research in multimodal language models.

3. Approach

In this section, we provide a rationale for selecting specific textual and visual data sources and describe the data collection process and semi-automation of the annotation process employed to create the benchmark.

3.1. Data sources

Effective evaluation of the Visual-WSD task requires image-word pairs with challenging word instances, e.g. those with multiple meanings (polysemantic words). Ukrainian Wikipedia[1] seemed to be a good source for identifying such words; however, our subsequent analysis revealed multiple problems and shortcomings (misleading links, missing images/sections, irrelevant articles, etc.) in data collected that way. Consequently, we opted to use homonyms listed in reliable dictionary sources.

The dictionary of homonyms of the Ukrainian language (Demska and Kulchytskyi, 1996) is seemingly the only thorough and reliable research work. The dictionary is only available as a published physical book; with the authors' and publisher's permission, we run the Optical Character Recognition (OCR) software to transform the textual information into a soft copy.

After refining the dictionary post-OCR quality, experts in the Ukrainian language and specialized knowledge in Ukrainian philology performed a thorough selection of homonyms. The selected homonyms are nouns (to optimize the search for visual complementary material), of high usage frequency in the modern Ukrainian language (according to the Shvedova et al. (2017–2024)), and are full (with the aligned paradigm of forms).

In the annotation stage, links to the corresponding Wikipedia sources were collected, including the word in the proper sense and the accompanying image. It is worth noting that while Wikipedia provided a convenient source and API for automating data collection process, its reliability was occasionally compromised. There were eminent challenges such as the absence of a direct article for certain words or missing images on the page of word definition. Also, unlike its English counterpart, the Ukrainian Wikipedia often suffers from incomplete information and numerous missing sections.

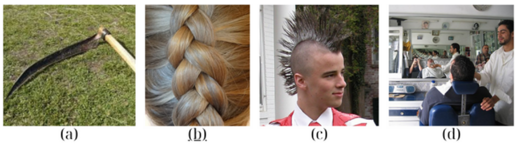

(a) (b) (c) (d)

Figure 2: Example of the benchmark entry. The word Коса (en: braid, transl: kosa) is ambiguous. It corresponds to the meaning Заплетене волосся; довге волосся (en: braided hair; long hair, transl: zapletene volossya; dovhe volossya). The word Волосся (en: hair, transl: volossya) is the trigger word. The image that corresponds to the intended meaning is b (underlined). The other three images are examples of negative samples. Note: While the task involves nine negative images, we only display three negative images for simplicity.

3.2. The methodology for constructing the benchmark

Each entry in the benchmark includes a target word along with one or multiple trigger words paired with ten unique images: one image corresponds to the intended meaning of the ambiguous target word, serving as a positive sample. The remaining images are negative samples and correspond to

- alternative interpretations of the ambiguous target word (3 images per entry);
- similar words within the domain (3 images per entry);
- randomly selected concepts (3 images per entry).

Figure 2 provides a simplified overview of single entry of the benchmark.

We generated positive samples by extracting the title picture of the ambiguous word from its corresponding Wikipedia article. In cases where the article lacked a valid image or any image altogether, domain experts supplied one. Negative samples for each word sense were constructed using other Wikipedia articles from the same domain as the target sense. By employing such a method, we aimed to discover articles about other senses of the ambiguous word, similar concepts within the same domain, and consequently, obtain corresponding images. We also hypothesized that this approach could lead us to discover images from completely different, unrelated concepts.

Using this methodology, we collected forty negative samples for each word sense. Subsequently, domain experts analyzed these samples and retained only the nine most relevant images for each group of negative samples.

We possessed a list of ambiguous words along with their respective definitions provided by domain experts. To generate trigger words for each entry, we tasked domain experts with supplying several words capable of identifying the intended meaning of the word sense when considering the correlation between definition and image. These trigger words

[1] https://en.wikipedia.org/wiki/Ukrainian_Wikipedia

were deliberately chosen to be sufficiently challenging so as not to reveal the meaning of the image in isolation; the target word typically remained necessary to comprehend the complete context. This process aimed to ensure a demanding text disambiguation task.

At the time of evaluation, the benchmark included 87 homonyms. However, we are in the process of expanding the homonym list and will update the benchmark accordingly in the future.

4. Evaluation

In this section, we explore the metrics utilized to assess model performance on our benchmark and present the results of models evaluation.

To generate predictions, we compare the model embeddings representing the query phrase and those representing each candidate image. The candidate image showing the highest cosine similarity to the query is then identified as the prediction.

In instances where the direct retrieval of embeddings is impossible (e.g., in GPT4-Vision), we prompt the model with both the query and all image candidates, and instruct it to rank the images from the most closely associated with the query to least associated.

4.1. Evaluation metrics

To evaluate models' performance, we have used the Mean Reciprocal Rank (MRR) and HIT@1 metrics.

Given $r = [r_1, \ldots, r_n]$ as the image ranking predictions provided by a model, MRR is defined as:

$$\text{MRR} = \frac{1}{n}\sum_{i=1}^{n}\frac{1}{r_i} \times 100\%, \quad (1)$$

where n is the number of queries and r_i is the rank of the correct result for the i-th query. Another metric we use is the HIT@1 score defined as

$$\text{HIT@1} = \frac{1}{n}\sum_{i=1}^{n}\text{correct}(r_i) \times 100\%, \quad (2)$$

where $correct(r_i)$ is an indicator function that equals 1 if $r_i = 1$ (i.e., the correct result is ranked first) and 0 otherwise. The HIT@1 metric can also be interpreted as the accuracy of the ranking model.

MRR evaluates the model's efficiency in retrieving relevant images by considering the position of the first relevant image in the ranked list, providing a comprehensive measure of retrieval effectiveness.

HIT@1 directly measures the model's accuracy in identifying the most relevant image by assessing the proportion of queries for which the top-ranked image is relevant.

Model	HIT@1	MRR
XLM-Roberta-Large-Vit-B-16Plus	42.78	60.30
XLM-Roberta-Large-Vit-L-14	40.21	58.65
XLM-Roberta-Large-Vit-B-32	39.69	57.69
GPT4-Vision	38.50	45.29
LABSE-Vit-L-14	35.57	54.37
clip-ViT-B-32-multilingual-v1	32.99	52.46
GCP Multimodal Embeddings	22.68	41.74
LLaVA-1.5	14.43	33.03
clip-ViT-B-32-multilingual-v1 (baseline on English language)	60.48	73.88

Table 1: The HIT@1 and MRR metrics for multiple multimodal models evaluated on the assembled benchmark and sorted by HIT@1. Baseline results for Visual-WSD in the English language are also included for comparison (Raganato et al., 2023).

4.2. Results

Table 1 gives an overview of the performance evaluation metrics of multiple multilingual models on the compiled benchmark.

The results demonstrate that all evaluated models performed less effectively in Ukrainian compared to the English baseline model which highlights a disparity in performance between Ukrainian and English in the Visual-WSD task.

5. Conclusion

This research introduces a benchmark for the Visual-WSD task in the Ukrainian language[2]. Unlike traditional single-modality benchmarks, we propose an approach that integrates textual and visual modalities into a single benchmark.

Furthermore, we assessed various suitable multilingual models using the compiled benchmark. Our findings revealed a notable underperformance in the Visual-WSD task for the Ukrainian language compared to English.

6. Future plans

We plan to expand the list of homonyms by introducing such units that the neo-lexicography of the Ukrainian language has not yet recorded. Still, they have become a vital part of modern Ukrainian speech (academic or informal). Examples of such words are бот (en: bot, transl: bot) in meaning програмний агент (en: program agent, transl: prohramnyy ahent), град (en: hail, transl: hrad)

[2] U-VWSD benchmark

in meaning бойова машина (en: combat vehicle, transl: boyova mashyna) and many others.

We intend to publish the benchmark and a compiled set of homonyms and their corresponding definitions as online resources, complete with an API for accessing the materials[3].

Furthermore, we plan to integrate the compiled benchmark for the Ukrainian language into existing benchmarks for other languages, facilitating the research in multilingual and multimodal LLMs.

7. Limitations

Currently, we have constructed a benchmark using a limited number of homonyms. Specifically, we have focused on homonyms, which are nouns with a high frequency of usage in Ukrainian. There exist homonyms in the Ukrainian language that remain undocumented in its neo-lexicography. At present, we have excluded these homonyms from our benchmark. Nonetheless, these homonyms constitute a significant component of the Ukrainian language. Therefore, it is highly pertinent to include them in model evaluations, considering their widespread usage by speakers.

8. Ethical Statement

The domain experts engaged in our research are proven professionals in Ukrainian philology, ensuring a high standard of work in selecting suitable images and contextual information. The images do not contain any harmful or detrimental content.

9. Bibliographical References

Edoardo Barba, Luigi Procopio, and Roberto Navigli. 2021. Consec: Word sense disambiguation as continuous sense comprehension. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 1492–1503.

Fredrik Carlsson, Philipp Eisen, Faton Rekathati, and Magnus Sahlgren. 2022. Cross-lingual and multilingual CLIP. In *Proceedings of the Thirteenth Language Resources and Evaluation Conference*, pages 6848–6854, Marseille, France. European Language Resources Association.

David Grangier and Dan Iter. 2022. The trade-offs of domain adaptation for neural language models. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 3802–3813, Dublin, Ireland. Association for Computational Linguistics.

Viktor Hangya, Hossain Shaikh Saadi, and Alexander Fraser. 2022. Improving low-resource languages in pre-trained multilingual language models. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 11993–12006, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Geoffrey Hinton, Oriol Vinyals, and Jeff Dean. 2015. Distilling the knowledge in a neural network.

Oskar Holmström, Jenny Kunz, and Marco Kuhlmann. 2023. Bridging the resource gap: Exploring the efficacy of English and multilingual LLMs for Swedish. In *Proceedings of the Second Workshop on Resources and Representations for Under-Resourced Languages and Domains (RESOURCEFUL-2023)*, pages 92–110, Tórshavn, the Faroe Islands. Association for Computational Linguistics.

Lei Huang, Weijiang Yu, Weitao Ma, Weihong Zhong, Zhangyin Feng, Haotian Wang, Qianglong Chen, Weihua Peng, Xiaocheng Feng, Bing Qin, and Ting Liu. 2023. A survey on hallucination in large language models: Principles, taxonomy, challenges, and open questions.

Yurii Laba, Volodymyr Mudryi, Dmytro Chaplynskyi, Mariana Romanyshyn, and Oles Dobosevych. 2023. Contextual embeddings for Ukrainian: A large language model approach to word sense disambiguation. In *Proceedings of the Second Ukrainian Natural Language Processing Workshop (UNLP)*, pages 11–19, Dubrovnik, Croatia. Association for Computational Linguistics.

Haotian Liu, Chunyuan Li, Yuheng Li, and Yong Jae Lee. 2023. Improved baselines with visual instruction tuning.

Andrea Moro, Alessandro Raganato, and Roberto Navigli. 2014. Entity linking meets word sense disambiguation: a unified approach. *Transactions of the Association for Computational Linguistics*, 2:231–244.

Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. 2021. Learning transferable visual models from natural language supervision. In *International conference on machine learning*, pages 8748–8763. PMLR.

Alessandro Raganato, Iacer Calixto, Asahi Ushio, Jose Camacho-Collados, and Mohammad Taher Pilehvar. 2023. SemEval-2023 task 1: Visual

[3] U-VWSD web page

word sense disambiguation. In *Proceedings of the 17th International Workshop on Semantic Evaluation (SemEval-2023)*, pages 2227–2234, Toronto, Canada. Association for Computational Linguistics.

Aditya Ramesh, Mikhail Pavlov, Gabriel Goh, Scott Gray, Chelsea Voss, Alec Radford, Mark Chen, and Ilya Sutskever. 2021. Zero-shot text-to-image generation. In *Proceedings of the 38th International Conference on Machine Learning*, volume 139 of *Proceedings of Machine Learning Research*, pages 8821–8831. PMLR.

Nils Reimers and Iryna Gurevych. 2019. Sentence-BERT: Sentence embeddings using Siamese BERT-networks. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 3982–3992, Hong Kong, China. Association for Computational Linguistics.

Nils Reimers and Iryna Gurevych. 2020. Making monolingual sentence embeddings multilingual using knowledge distillation. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 4512–4525, Online. Association for Computational Linguistics.

Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. 2022. High-resolution image synthesis with latent diffusion models. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 10684–10695.

Qianhui Wu, Zijia Lin, Börje Karlsson, Jian-Guang Lou, and Biqing Huang. 2020. Single-/multi-source cross-lingual NER via teacher-student learning on unlabeled data in target language. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 6505–6514, Online. Association for Computational Linguistics.

Lianmin Zheng, Wei-Lin Chiang, Ying Sheng, Siyuan Zhuang, Zhanghao Wu, Yonghao Zhuang, Zi Lin, Zhuohan Li, Dacheng Li, Eric Xing, et al. 2024. Judging llm-as-a-judge with mt-bench and chatbot arena. *Advances in Neural Information Processing Systems*, 36.

10. Language Resource References

O. Demska and I. Kulchytskyi. 1996. Словник омонімів української мови [Dictionary of homonyms of the Ukrainian language]. видавництво "Фенікс" ["Phoenix" publishing house].

V. Rusanivskyi. 2010. Словник української мови [Dictionary of the Ukrainian language]. Наук. думка [Nauk. dumka], Словники України [Dictionaries of Ukraine].

M. Shvedova and R. von Waldenfels and S. Yarygin and A. Rysin and V. Starko and T. Nikolayenko and others. 2017–2024. Генеральний регіонально анотований корпус української мови (ГРАК) [General Regionally Annotated Corpus of the Ukrainian Language (GRAK)].

The UNLP 2024 Shared Task on Fine-Tuning Large Language Models for Ukrainian

Oleksiy Syvokon, Mariana Romanyshyn, Roman Kyslyi
Microsoft, Grammarly, National Technical University of Ukraine "Igor Sikorsky Kyiv Polytechnic Institute"
Kyiv, Ukraine
osyvokon@microsoft.com, mariana.romanyshyn@grammarly.com, kyslyi.roman@lll.kpi.ua

Abstract

This paper presents the results of the UNLP 2024 shared task, the first Shared Task on Fine-Tuning Large Language Models for the Ukrainian language. The goal of the task was to facilitate the creation of models that have knowledge of the Ukrainian language, history, and culture, as well as common knowledge, and are capable of generating fluent and accurate responses in Ukrainian. The participants were required to use models with open weights and reasonable size to ensure the reproducibility of the solutions. The participating systems were evaluated using multiple-choice exam questions and manually crafted open questions. Three teams submitted their solutions before the deadline, and two teams submitted papers that were accepted to appear in the UNLP workshop proceedings and are referred to in this report. The Codabench leaderboard is left open for further submissions.

Keywords: Large Language Models, LLM, Fine-Tuning, LLM Benchmarking

1. Introduction

The emergence of large language models (LLMs) marked a significant step forward in the field of natural language processing (NLP), providing a single solution for the tasks of generating human-like text. Creative writing, text evaluation, controlled text generation have suddenly become available to everyone, causing both a surge in popularity of LLM-based tools like ChatGPT (OpenAI, 2022) and discussions about the limitations and ethical implications of using them (Borji, 2023; Kocoń et al., 2023).

However, training an LLM requires significant computational resources, which may be expensive to obtain, and substantial amounts of text data, which is not readily available for most natural languages, including Ukrainian. With the UNLP 2024 shared task, our goal was to facilitate the creation of LLMs better adapted to the Ukrainian language, history, and cultural context with reasonable computational resources.

The remainder of the paper is organized as follows. Section 2 gives an overview of LLM benchmarks and methods of LLM language adaptation. Section 3 describes the UNLP 2024 shared task setup. Section 4 reviews the datasets available in this shared task. Section 5 explains how the competing systems were evaluated and ranked. Section 6 presents the results of the shared task and provides an overview of the submitted solutions. Section 7 mentions how the competing systems compare to GPT-4. Finally, Section 8 summarizes the contribution, and Section 9 provides an ethics statement.

2. Related Work

LLM Benchmarks. Evaluation methods for LLMs fall into two broad categories. Firstly, there are static, *ground-truth-based* benchmarks. These feature a predefined collection of tasks along with correct answers, and an automated metric. Such benchmarks have been the standard for assessing models before the advent of LLMs. Over time, numerous datasets of this kind have been created, and many have been adapted for LLM evaluation: GSM-8k (Cobbe et al., 2021), EXAMS (Hardalov et al., 2020), MMLU (Hendrycks et al., 2020), and AgiEval (Zhong et al., 2023), among many others. These benchmarks are cost-effective, reproducible, and can be executed automatically. However, they are restricted to a limited range of tasks and are often unsuitable for evaluating complex capabilities like open-ended text generation and subjective aspects such as humor and engagement. Consequently, these benchmarks do not fully capture the intricacies of LLM performance.

The second category involves benchmarks that measure *human preferences* for LLM-generated content. Typically, this involves a blind comparison between pairs of LLM responses to the same prompt. These comparisons are then translated into model rankings through systems such as Elo, the Bradley-Terry model, or TrueSkill (Boubdir et al., 2023; Bai et al., 2022; Bradley and Terry, 1952; Herbrich et al., 2007). Some preference-based benchmarks utilize a static set of prompts (Zheng et al., 2024; Li et al., 2023), while others permit open-ended interactions with the models (Chiang et al., 2024; Kiela et al., 2021). Recently, there has been a trend towards using advanced LLMs to

25 May, 2024. © 2024 ELRA Language Resource Association: CC BY-NC 4.0

replace human evaluators (Li et al., 2023; Chiang and Lee, 2023; Zheng et al., 2024).

This shared task employs two complementary benchmarks: an automated metric on multiple-choice exam questions for testing LLM knowledge and human ratings for the subjective evaluation of open-ended text generation tasks.

LLM language adaptation. Despite the rapid development of open LLMs, many of these models primarily focus on English and offer limited support for other languages. A few notable exceptions (Üstün et al., 2024; Lin et al., 2021; Xue et al., 2020; Liang et al., 2023) just underscore this trends. Training a large language model from scratch demands substantial resources, making it an impractical option for many researchers. A feasible alternative is to adapt existing models to one or more languages by fine-tuning a model with a smaller set of language-specific data. This adaptation process may involve selecting a strong base LLM, curating language-specific datasets, expanding the vocabulary, conducting continual pretraining (whether full or adapter-based), translating instruction-tuning datasets, generating synthetic data, clustering languages based on their similarities, among other strategies (Lin et al., 2024; Csaki et al., 2024; Yong et al., 2022; ImaniGooghari et al., 2023; Ebrahimi and Kann, 2021; Blevins et al., 2024; Yang et al., 2023; Zhu et al., 2023).

3. Task description

The UNLP 2024 shared task required participants to fine-tune a large language model that can answer questions about the Ukrainian language, history, and culture, as well as perform text-generation tasks, all by producing fluent and factually accurate text in Ukrainian.

To ensure fair competition with reproducible results, we enforced the following limitations:

1. Only LLMs with open weights such as Llama 2 (Touvron et al., 2023), Mistral 7B (Jiang et al., 2023), Phi-2 (Javaheripi and Bubeck, 2023), Gemma (Mesnard et al., 2024), Aya 101 (Üstün et al., 2024), etc. were allowed to be used in the shared task.
2. The models had to run on GPU with 16GB VRAM and CUDA compute capacity 8.6. The type and amount of compute used for training were not limited, but the model weights and activations had to fit and stay in the GPU memory entirely. CPU memory and disk offloading were not allowed.
3. The weights of the final model had to be published on the Hugging Face Hub[1] or a similar open platform.

The participants were allowed to complement the fine-tuning with various prompting strategies, like few-shot learning or chain-of-thought reasoning, or use retrieval-augmented generation (RAG) from open data sources.

We split the evaluation of the submitted models into two tracks: multiple-choice exam questions and open questions. We provided the participants with a set of multiple-choice exam questions for training and validation and set up a Codabench[2] environment to test the systems on a hidden test set. For open questions, we shared sample questions with the participants and ran a human evaluation task to test the systems on a manually crafted test set. See Section 5 for details.

Additionally, we highly encouraged the participants to use any external data of their choice, released a script that loads the provided dataset and generates a sample prompt, and prepared sample submission files for Codabench.

4. Data

We provided the participants of the shared task with two datasets: multiple-choice exam questions and manually crafted open questions. The dataset statistics can be found in Table 1.

Task	Split	Size
Exam questions	train	3,063
	test	751
Open questions	dev	20
	test	100

Table 1: The sizes of the datasets provided in the UNLP 2024 shared task.

Both datasets can be accessed through the repository of the shared task[3].

4.1. Exam Questions

This dataset contains machine-readable questions and answers taken from the Ukrainian External Independent Evaluation[4] called ЗНО (transl: ZNO) in Ukrainian. External Independent Evaluation is a standard set of exams taken by schoolchildren in Ukraine when they apply to higher educational institutions. The dataset contains exam questions from the years 2006-2023 and covers two subjects only: History of Ukraine and Ukrainian language and literature.

[1] https://huggingface.co/

[2] https://www.codabench.org/competitions/2046/

[3] https://github.com/unlp-workshop/unlp-2024-shared-task/tree/main/data

[4] https://zno.osvita.ua/

We filtered the dataset by extracting only multiple-choice questions with one correct answer. We removed questions that referenced images (maps, portraits, photos, etc.). The final dataset was published in the .jsonl format. The training set contained 3,063 questions/answers from the years 2006-2019. The test set contained 751 questions and hidden answers, spanning the years 2020 to 2023.

4.2. Open Questions

The dataset of open questions was crafted by two native speakers of Ukrainian and comprised instruction prompts for text-generation tasks and common-knowledge question-answering. The dataset contained an equal distribution of the following:

- common knowledge questions on the topics of Ukrainian literature, music, history, geography, and culture;
- composition tasks that asked the model to write messages across a set of formality levels, lengths, and topics;
- rewrite tasks that asked the model to correct the input text, simplify it, add humor, add more details, or add emotions;
- evaluation tasks that asked the model to outline ways of improving the input text, analyze emotions in text, answer follow-up questions, brainstorm ways to complete the text, or find an odd word in a row.

The final dataset was published as a .jsonl file in the Alpaca dataset format[5]. The dev set contained 20 questions. The hidden test set contained 100 questions.

5. Evaluation

The competing LLM solutions for Ukrainian were evaluated on two hidden test sets: exam questions and open questions.

In the **first track**, we evaluated the models on a hidden test set of 751 multiple-choice exam questions, where each question had one correct answer. This setting allowed us to use accuracy as our primary metric to rank the competing LLM solutions. The registered participants submitted their system results using Codabench, which automatically compared their results with the hidden answers, returned the score, and placed the systems on the leaderboard.

In the **second track**, we evaluated the models on 100 manually crafted open questions. We asked the participants to send us their models' answers to the questions in the predefined format. We then set up a side-by-side evaluation task in the Hugging Face Spaces[6]. For that, we created a simple space with the Gradio application that displayed a question from the test set and two randomly chosen anonymized model outputs. The user could then vote which answer is better; if neither, declare a tie. The model answers to all the questions and voting logs were stored in a Firebase DB[7].

We crowdsourced annotations from 63 native speakers of Ukrainian from the Ukrainian NLP community who volunteered to join the annotation task. The annotation guidelines for the human evaluation are available in the repository of the shared task in Ukrainian and English[8].

We collected over 300 human judgments for each competing model and used the TrueSkill (Herbrich et al., 2007) ranking algorithm implemented in the trueskill Python library[9] to define the winner. TrueSkill is a statistically based ranking system for multiplayer competitions that infers the relative rankings of players based on their performance against each other. It uses the Bayesian inference algorithm to estimate each player's skill level.

6. Results and System Descriptions

A total of twenty-two teams registered for the UNLP 2024 shared task, but only three teams submitted their solutions before the deadline: Sherlock, CodeKobzar, and UkraineNow. Team Sherlock submitted two distinct solutions, each evaluated independently. Teams Sherlock and UkraineNow submitted papers that were accepted to appear in the UNLP workshop proceedings and are referred to in this report. Team CodeKobzar provided their system description by email.

We briefly review the systems here; for complete descriptions, please see the corresponding papers. Table 2 and Table 3 present the leaderboards for the two tasks.

Rank	Participant	Accuracy
1	Sherlock (RAG)	0.49
2	Sherlock (no RAG)	0.42
3	CodeKobzar	0.39
4	UkraineNow	0.23

Table 2: The **official** UNLP 2024 shared task results for the task of answering multiple-choice exam questions.

[5] https://github.com/tatsu-lab/stanford_alpaca

[6] https://huggingface.co/spaces

[7] https://firebase.google.com/

[8] https://github.com/unlp-workshop/unlp-2024-shared-task/blob/main/annotation

[9] https://github.com/sublee/trueskill

Sherlock (Boros et al., 2024) submitted winning solutions for both tasks. The team used a set of data-augmentation techniques with Mistral 7B.

Notably, the team used an array of diverse data sources for training, including Ukrainian Wikipedia, manually selected books on the target subject, and several translated datasets, both free-text and instruction-formatted. Due to the limiting factors in the standard RAG process (e.g., low performance of embeddings for Ukrainian), the team employed n-gram techniques. This method outperformed conventional similarity scoring approaches, with the LLM itself generating n-grams to enhance the retrieval process.

The tuning process began with the base Mistral 7B model, Mistral-7B-Instruct-v0.2[10]. The team experimented with standard fine-tuning on different datasets, delved into model weight merging, and leveraged direct preference optimization training to refine performance further. The availability of a test set allowed for iterative testing of various method combinations, optimizing the overall system efficacy. The team has made both the source code[11] and the model[12] publicly available.

UkraineNow (Kiulian et al., 2024) fine-tuned the open-source Gemma (gemma-2b-it[13] and gemma-7b-it[14]) and Mistral-7B-Instruct-v0.1[15] LLMs with a combination of instruction datasets, which included 10,000 rows of the UAlpaca dataset[16], 962 rows of their own UKID dataset, and 3,063 rows of the ZNO dataset provided by the organizers of the shared task. Due to resource constraints, the team chose to use the LoRA (Hu et al., 2022) fine-tuning approach, experimenting with various implementations of LoRA adapters. The team put extra effort into the quality evaluation of the models' outputs, dedicating a section of the paper to the phenomenon of code-switching, also known as Azirivka[17].

The fine-tuned gemma-2b-it model was submitted for the competition. The team has made the source code and the model available in their GitHub repository[18].

CodeKobzar[19] by Ben Ye and Mariia Ponomarenko is a large language model specifically fine-tuned for Ukrainian language, employing the Chain of LoRA technique (Xia et al., 2024) on the Vicuna-13B pretrained model[20]. The initial dataset[21] comprises articles from the Ukrainian Wikipedia, segmented into 40.6K sentences to facilitate question generation using the Mistral Medium API[22]. These sentence-question pairs served as the basis for the model's first fine-tuning phase, focusing on question generation and answering. The model was trained for one epoch with a maximum sequence length of 2,048 tokens, using the Nvidia A100 GPU. After that, the LoRA layers were integrated with the base model for a second round of training on the same dataset.

In the third fine-tuning phase, the dataset was refined to include only Ukrainian historical, literary, and cultural content, supplemented by grammatical rules from `pravopys.net`. This resulted in a new dataset of 73.6K entries[23], which was divided into prompt-question and question-response pairings. The model was then fine-tuned on the combination of the aforementioned dataset and a corpus of ZNO multiple-answer questions provided with the shared task.

This iterative approach, through the Chain of LoRA, enabled the KodKobzar model to perform an iterative low-rank residual learning procedure to approximate the optimal weight update and thereby improve the model's proficiency in grammatical accuracy and sentence construction in Ukrainian. However, since the training concentrated mainly on question generation and answering, it constrained the model's broader generative abilities, and the restricted dataset limited the model's deeper understanding of Ukrainian culture, literature, and history.

7. Comparison with GPT-4

The participants of the shared tasks were limited in the selection of models for fine-tuning with regard to their size and accessibility. These limitations were needed to ensure that the resulting solutions are reproducible and practically useful to the NLP community. However, we were curious to understand

[10] https://huggingface.co/mistralai/Mistral-7B-Instruct-v0.2

[11] https://github.com/adobe/sherlock-backend/tree/UNLP2024

[12] https://huggingface.co/SherlockAssistant/Mistral-7B-Instruct-Ukrainian

[13] https://huggingface.co/google/gemma-2b-it

[14] https://huggingface.co/google/gemma-7b-it

[15] https://huggingface.co/mistralai/Mistral-7B-Instruct-v0.1

[16] https://github.com/robinhad/kruk

[17] https://en.wikipedia.org/wiki/Azirivka

[18] https://github.com/PolyAgent/from-bytes-to-borsch

[19] https://huggingface.co/ponoma16/KodKobzar13B

[20] https://huggingface.co/lmsys/vicuna-13b-v1.5

[21] https://huggingface.co/datasets/byebyebye/ukr-wiki-qa-v1

[22] https://mistral.ai/

[23] https://huggingface.co/datasets/byebyebye/ukr-wiki-qa-v2

Rank	Participant	TrueSkill	σ	Number of judgments
1	Sherlock (no RAG)	26.77	0.75	330
2	Sherlock (RAG)	26.27	0.74	329
3	UkraineNow	24.89	0.75	326
4	CodeKobzar	23.79	0.76	311

Table 3: The **official** UNLP 2024 shared task results for the task of answering open questions. The TrueSkill column shows the participant's rating, and σ represents the confidence of the rating.

Rank	Participant	TrueSkill	σ	Number of judgments
1	GPT-4	28.48	0.79	474
2	Sherlock (no RAG)	25.05	0.75	462
3	Sherlock (RAG)	24.14	0.76	439
4	UkraineNow	23.16	0.76	455
5	CodeKobzar	22.17	0.77	414

Table 4: The **non-official** UNLP 2024 shared task results for the task of answering open questions. Here, GPT-4 is included for comparison. The TrueSkill column shows the participant's rating, and σ represents the confidence of the rating.

how these fine-tuned open solutions compare to the proprietary OpenAI models, in particular gpt-4-0613[24], which was the latest OpenAI GPT version at the time when the shared task started.

For the exam task, we used a very simple prompt[25] and the default parameters of GPT-4. The model managed to achieve an accuracy of 0.61. We also ran GPT-4 on the open questions task and included the responses in the human evaluation. Table 4 shows the gap between the winning open-source solution and GPT-4.

This experiment set an ambitious goal for the next iteration of this shared task.

8. Conclusion

We believe that the UNLP 2024 shared task was instrumental in facilitating research on fine-tuning large language models for the Ukrainian language, and we hope that the insights from the teams' research will be useful to the NLP community. All the datasets used in the shared task are available on GitHub, and the competing systems were openly published, which contributes to the reproducibility of the shared task results and the creation of more accessible LLMs.

The best-performing systems were submitted by team Sherlock, scoring 49% accuracy on the exam task (with RAG) and 26.77% rating on the open question task (without RAG). The Codabench environment remains open for further submissions, although any such submissions will be considered outside of the UNLP 2024 competition.

The open LLMs used in the shared task included Mistral 7B, Gemma, and Vicuna-13B. All system descriptions mention the scarcity of open datasets for the task at hand and show the creativity of the researchers in creating new datasets.

In the next iterations of this shared task, we plan to increase the size and variability of the test sets and introduce automated metrics for the open question evaluation, in addition to human evaluation.

9. Ethics Statement

To make sure that the participants of the shared task have equal opportunities and that the resulting solutions can be used by the research community, the organizers of the shared task set strict limitations on the size and accessibility of the models that were allowed in the competition.

Upon entering the competition, all participants of the shared task accepted the following terms and conditions:

- All participants agree to compete in a fair and honest manner in the shared task and not use any illegal, malicious, or otherwise unethical methods to gain an advantage in the shared task.
- All participants agree to not distribute or share the test data obtained during the shared task with any third parties.
- All participants agree to make their solutions publicly available upon the completion of the shared task in order to facilitate knowledge sharing and developments of the Ukrainian language.

[24] https://platform.openai.com/docs/models/gpt-4-and-gpt-4-turbo

[25] https://github.com/unlp-workshop/unlp-2024-shared-task/blob/main/examples/random_baseline.py

To the best of our knowledge, the shared task participants followed these terms and conditions.

10. Acknowledgements

We extend special gratitude to Nataliia Romanyshyn for her invaluable contribution to the development of the open questions set. We are grateful to the participants of the Ukrainian NLP community who volunteered to help with side-by-side evaluation.

11. Bibliographical References

Yuntao Bai, Andy Jones, Kamal Ndousse, Amanda Askell, Anna Chen, Nova DasSarma, Dawn Drain, Stanislav Fort, Deep Ganguli, Tom Henighan, et al. 2022. Training a helpful and harmless assistant with reinforcement learning from human feedback. *arXiv preprint arXiv:2204.05862*.

Terra Blevins, Tomasz Limisiewicz, Suchin Gururangan, Margaret Li, Hila Gonen, Noah A Smith, and Luke Zettlemoyer. 2024. Breaking the curse of multilinguality with cross-lingual expert language models. *arXiv preprint arXiv:2401.10440*.

Ali Borji. 2023. A categorical archive of chatgpt failures.

Tiberiu Boros, Radu Chivereanu, Stefan Dumitrescu, and Octavian Purcaru. 2024. Fine-tuning and retrieval augmented generation for question answering using affordable large language models. In *Proceedings of the Third Ukrainian Natural Language Processing Workshop*, Torino, Italy. European Language Resources Association.

Meriem Boubdir, Edward Kim, Beyza Ermis, Sara Hooker, and Marzieh Fadaee. 2023. Elo uncovered: Robustness and best practices in language model evaluation. *arXiv preprint arXiv:2311.17295*.

Ralph Allan Bradley and Milton E Terry. 1952. Rank analysis of incomplete block designs: I. the method of paired comparisons. *Biometrika*, 39(3/4):324–345.

Cheng-Han Chiang and Hung-yi Lee. 2023. Can large language models be an alternative to human evaluations? *arXiv preprint arXiv:2305.01937*.

Wei-Lin Chiang, Lianmin Zheng, Ying Sheng, Anastasios Nikolas Angelopoulos, Tianle Li, Dacheng Li, Hao Zhang, Banghua Zhu, Michael Jordan, Joseph E Gonzalez, et al. 2024. Chatbot arena: An open platform for evaluating llms by human preference. *arXiv preprint arXiv:2403.04132*.

Karl Cobbe, Vineet Kosaraju, Mohammad Bavarian, Mark Chen, Heewoo Jun, Lukasz Kaiser, Matthias Plappert, Jerry Tworek, Jacob Hilton, Reiichiro Nakano, et al. 2021. Training verifiers to solve math word problems. *arXiv preprint arXiv:2110.14168*.

Zoltan Csaki, Bo Li, Jonathan Li, Qiantong Xu, Pian Pawakapan, Leon Zhang, Yun Du, Hengyu Zhao, Changran Hu, and Urmish Thakker. 2024. Sambalingo: Teaching large language models new languages.

Abteen Ebrahimi and Katharina Kann. 2021. How to adapt your pretrained multilingual model to 1600 languages. *arXiv preprint arXiv:2106.02124*.

Momchil Hardalov, Todor Mihaylov, Dimitrina Zlatkova, Yoan Dinkov, Ivan Koychev, and Preslav Nakov. 2020. Exams: A multi-subject high school examinations dataset for cross-lingual and multilingual question answering. *arXiv preprint arXiv:2011.03080*.

Dan Hendrycks, Collin Burns, Steven Basart, Andy Zou, Mantas Mazeika, Dawn Song, and Jacob Steinhardt. 2020. Measuring massive multitask language understanding. *arXiv preprint arXiv:2009.03300*.

Ralf Herbrich, Tom Minka, and Thore Graepel. 2007. Trueskill(tm): A bayesian skill rating system. In *Advances in Neural Information Processing Systems 20*, pages 569–576. MIT Press.

Edward J Hu, yelong shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. 2022. LoRA: Low-rank adaptation of large language models. In *International Conference on Learning Representations*.

Ayyoob ImaniGooghari, Peiqin Lin, Amir Hossein Kargaran, Silvia Severini, Masoud Jalili Sabet, Nora Kassner, Chunlan Ma, Helmut Schmid, André FT Martins, François Yvon, et al. 2023. Glot500: Scaling multilingual corpora and language models to 500 languages. *arXiv preprint arXiv:2305.12182*.

Mojan Javaheripi and Sebastien Bubeck. 2023. Phi-2: The surprising power of small language models. *Microsoft Research Blog*.

Albert Q. Jiang, Alexandre Sablayrolles, Arthur Mensch, and et al. 2023. Mistral 7b.

Douwe Kiela, Max Bartolo, Yixin Nie, Divyansh Kaushik, Atticus Geiger, Zhengxuan Wu, Bertie Vidgen, Grusha Prasad, Amanpreet Singh, Pratik Ringshia, et al. 2021. Dynabench: Rethinking benchmarking in nlp. *arXiv preprint arXiv:2104.14337*.

Artur Kiulian, Anton Polishko, Mykola Khandoga, Oryna Chubych, Jack Connor, Raghav Ravishankar, and Adarsh Shirawalmath. 2024. From bytes to borsch: Fine-tuning gemma and mistral for the ukrainian language representation. In *Proceedings of the Third Ukrainian Natural Language Processing Workshop*, Torino, Italy. European Language Resources Association.

Jan Kocoń, Igor Cichecki, and et al. Kaszyca. 2023. Chatgpt: Jack of all trades, master of none. *Information Fusion*, 99:101861.

Xuechen Li, Tianyi Zhang, Yann Dubois, Rohan Taori, Ishaan Gulrajani, Carlos Guestrin, Percy Liang, and Tatsunori B. Hashimoto. 2023. Alpacaeval: An automatic evaluator of instruction-following models. `https://github.com/tatsu-lab/alpaca_eval`.

Davis Liang, Hila Gonen, Yuning Mao, Rui Hou, Naman Goyal, Marjan Ghazvininejad, Luke Zettlemoyer, and Madian Khabsa. 2023. Xlm-v: Overcoming the vocabulary bottleneck in multilingual masked language models.

Peiqin Lin, Shaoxiong Ji, Jörg Tiedemann, André FT Martins, and Hinrich Schütze. 2024. Mala-500: Massive language adaptation of large language models. *arXiv preprint arXiv:2401.13303*.

Xi Victoria Lin, Todor Mihaylov, Mikel Artetxe, Tianlu Wang, Shuohui Chen, Daniel Simig, Myle Ott, Naman Goyal, Shruti Bhosale, Jingfei Du, et al. 2021. Few-shot learning with multilingual language models. *arXiv preprint arXiv:2112.10668*.

Gemma Team: Thomas Mesnard, Cassidy Hardin, Robert Dadashi, Surya Bhupatiraju, Shreya Pathak, Laurent Sifre, Morgane Rivière, Mihir Sanjay Kale, Juliette Love, Pouya Tafti, Léonard Hussenot, Aakanksha Chowdhery, Adam Roberts, Aditya Barua, Alex Botev, Alex Castro-Ros, Ambrose Slone, Amélie Héliou, Andrea Tacchetti, Anna Bulanova, Antonia Paterson, Beth Tsai, Bobak Shahriari, Charline Le Lan, Christopher A. Choquette-Choo, Clément Crepy, Daniel Cer, Daphne Ippolito, David Reid, Elena Buchatskaya, Eric Ni, Eric Noland, Geng Yan, George Tucker, George-Christian Muraru, Grigory Rozhdestvenskiy, Henryk Michalewski, Ian Tenney, Ivan Grishchenko, Jacob Austin, James Keeling, Jane Labanowski, Jean-Baptiste Lespiau, Jeff Stanway, Jenny Brennan, Jeremy Chen, Johan Ferret, Justin Chiu, Justin Mao-Jones, Katherine Lee, Kathy Yu, Katie Millican, Lars Lowe Sjoesund, Lisa Lee, Lucas Dixon, Machel Reid, Maciej Mikuła, Mateo Wirth, Michael Sharman, Nikolai Chinaev, Nithum Thain, Olivier Bachem, Oscar Chang, Oscar Wahltinez, Paige Bailey, Paul Michel, Petko Yotov, Pier Giuseppe Sessa, Rahma Chaabouni, Ramona Comanescu, Reena Jana, Rohan Anil, Ross McIlroy, Ruibo Liu, Ryan Mullins, Samuel L Smith, Sebastian Borgeaud, Sertan Girgin, Sholto Douglas, Shree Pandya, Siamak Shakeri, Soham De, Ted Klimenko, Tom Hennigan, Vlad Feinberg, Wojciech Stokowiec, Yu hui Chen, Zafarali Ahmed, Zhitao Gong, Tris Warkentin, Ludovic Peran, Minh Giang, Clément Farabet, Oriol Vinyals, Jeff Dean, Koray Kavukcuoglu, Demis Hassabis, Zoubin Ghahramani, Douglas Eck, Joelle Barral, Fernando Pereira, Eli Collins, Armand Joulin, Noah Fiedel, Evan Senter, Alek Andreev, and Kathleen Kenealy. 2024. Gemma: Open models based on gemini research and technology.

OpenAI. 2022. Introducing chatgpt.

Phillip Rust, Jonas Pfeiffer, Ivan Vulić, Sebastian Ruder, and Iryna Gurevych. 2021. How good is your tokenizer? on the monolingual performance of multilingual language models.

Hugo Touvron, Louis Martin, Kevin Stone, and et al. 2023. Llama 2: Open foundation and fine-tuned chat models.

Wenhan Xia, Chengwei Qin, and Elad Hazan. 2024. Chain of lora: Efficient fine-tuning of language models via residual learning.

Linting Xue, Noah Constant, Adam Roberts, Mihir Kale, Rami Al-Rfou, Aditya Siddhant, Aditya Barua, and Colin Raffel. 2020. mt5: A massively multilingual pre-trained text-to-text transformer. *arXiv preprint arXiv:2010.11934*.

Wen Yang, Chong Li, Jiajun Zhang, and Chengqing Zong. 2023. Bigtrans: Augmenting large language models with multilingual translation capability over 100 languages. *arXiv preprint arXiv:2305.18098*.

Zheng-Xin Yong, Hailey Schoelkopf, Niklas Muennighoff, Alham Fikri Aji, David Ifeoluwa Adelani, Khalid Almubarak, M Saiful Bari, Lintang Sutawika, Jungo Kasai, Ahmed Baruwa, et al. 2022. Bloom+ 1: Adding language support to bloom for zero-shot prompting. *arXiv preprint arXiv:2212.09535*.

Lianmin Zheng, Wei-Lin Chiang, Ying Sheng, Siyuan Zhuang, Zhanghao Wu, Yonghao

Zhuang, Zi Lin, Zhuohan Li, Dacheng Li, Eric Xing, et al. 2024. Judging llm-as-a-judge with mt-bench and chatbot arena. *Advances in Neural Information Processing Systems*, 36.

Wanjun Zhong, Ruixiang Cui, Yiduo Guo, Yaobo Liang, Shuai Lu, Yanlin Wang, Amin Saied, Weizhu Chen, and Nan Duan. 2023. Agieval: A human-centric benchmark for evaluating foundation models. *arXiv preprint arXiv:2304.06364*.

Wenhao Zhu, Yunzhe Lv, Qingxiu Dong, Fei Yuan, Jingjing Xu, Shujian Huang, Lingpeng Kong, Jiajun Chen, and Lei Li. 2023. Extrapolating large language models to non-english by aligning languages. *arXiv preprint arXiv:2308.04948*.

Ahmet Üstün, Viraat Aryabumi, Zheng-Xin Yong, Wei-Yin Ko, Daniel D'souza, Gbemileke Onilude, Neel Bhandari, Shivalika Singh, Hui-Lee Ooi, Amr Kayid, Freddie Vargus, Phil Blunsom, Shayne Longpre, Niklas Muennighoff, Marzieh Fadaee, Julia Kreutzer, and Sara Hooker. 2024. Aya model: An instruction finetuned open-access multilingual language model.

Fine-Tuning and Retrieval Augmented Generation for Question Answering Using Affordable Large Language Models

Tiberiu Boros, Radu Chivereanu, Stefan Daniel Dumitrescu, Octavian Purcaru
Adobe Systems
Bucharest, Romania
{boros, rchivereanu, sdumitre, opurcaru}@adobe.com

Abstract

We present our proposed system named *Sherlock* to UNLP 2024 Shared Task on Question Answering winning first place. We employ a mix of methods, from using automatically translated datasets to perform supervised fine-tuning and direct preference optimization on instruction-tuned models, to model weight merging and retrieval augmented generation. We present and motivate our chosen sequence of steps, as well as an ablation study to understand the effect of each additional step. The resulting model and code are made publicly available (download links provided in the paper).

Keywords: large language models, direct preference optimization, instruction tuning, LLM, finetuning, RAG, model merge, re-ranking, Ukrainian language, question answering, multiple choice, open-ended, open-source

1. Introduction

The work of Vaswani et al. (2017) has shaped landscape of Natural Language Processing (NLP), through the emergence of Transformer-based Large Language Models (LLMs). Proprietary models such as GPT (Achiam et al., 2023) or Open-Source alternatives such as LLama (Touvron et al., 2023), Mistral (Jiang et al., 2023) and Bard/Gemini (Manyika and Hsiao, 2023) are currently the number one choice in successfully solving difficult NLP tasks such as translation, question answering or user dialogue.

These achievements were made possible through continuous improvements of machine learning (ML) methods and techniques, most notably being the development of the attention mechanism(Ainslie et al., 2023) in tandem with better and faster hardware. However, the noticeable leaps in model performance often came with drastic increases in the number of parameters, which in turn added more stress to the hardware, resulting in increased training and exploitation costs.

This research is part of the of **the UNLP Shared Task 2024**(Syvokon et al., 2024), which focuses on Ukrainian Question Answering via **affordable LLMs**. Thus, our work is **focused on compact LLMs that run on a single consumer-grade GPU or CPU**. We set out to explore how to leverage such models, both by fine-tuning them and by using retrieval augmented generation (RAG).

In the following sections we'll investigate related methods and techniques (Section 2), provide details about the shared task, dataset and proposed methodology (Section 3), and present our results (Section 4) and conclusions (Section 5).

2. Related Work

The task of Question Answering is a long-standing and well defined task in NLP, with the purpose of answering a user's question, posed in natural language. The task itself has many variants (Zhang et al., 2023); we're focusing on text-aided selection of the correct choice given a question and multiple possible answers. To be able to better discriminate between the given choices, it is essential to pair the LLM's internal knowledge and reasoning capabilities with external data and tools.

Primarily, we need an LLM that is able to follow instructions. It has been shown, both empirically and otherwise that instruction tuning enables LLMs to do specific, useful work (Jiang et al., 2024). Prompting techniques are routinely employed to increase performance and guide models' answers towards a desired direction. The most basic prompt is to simply ask an LLM to do something (e.g. "zero-shot"), without providing any examples in the prompt. Few-shot means showing the LLM how to answer by understanding the format, input and output from the few examples given in the prompt, before asking the target question - this "primes" the model to respect the same format as the already-answered questions/tasks. Few-shotting is especially tricky for smaller models (Touvron et al., 2023) that have limited context-size.

Other prompting techniques, like Knowledge Generation Prompting (Liu et al., 2022) or regular self consistency-checks, aim to make use of the knowledge embedded in the model itself to augment its context and perform checking (the LLM generates an intermediary step of a problem and can check itself with an additional query to ascertain whether it considers it has sufficient information or the generated knowledge in the previous step is correct, in order to move to the next generation

25 May, 2024. © 2024 ELRA Language Resource Association: CC BY-NC 4.0

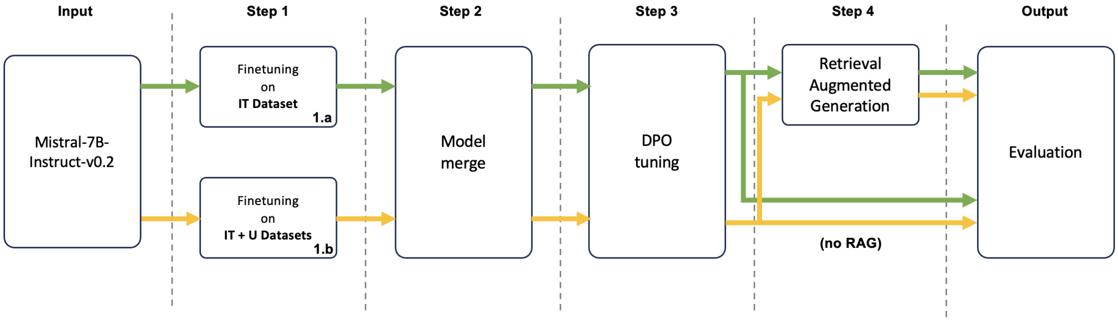


Figure 1: Diagram of the best performing strategy for tuning and running the model. From left to right: base model supervised finetuning, model merges, direct preference optimization yielding 2 models and the final evaluation of the 2 models with and without RAG

step or attempt a final answer.

One powerful method to combat LLM hallucination while benefiting from external sources is to perform Retrieval Augmented Generation (et al., 2021). This technique involves using semantic embeddings of a user's query to find pre-embedded texts that are semantically close and that could help in generating an answer. By externalizing the task of information retrieval (by finding and adding it in the LLM prompt), it lets the LLM focus more on answering the question based on presented facts/information rather than using its internal knowledge which often might lead to hallucinations and incorrect responses.

Other methods to further increase performance look towards tuning model parameters. While LLMs have been trained on huge amounts of text, they likely benefit from limited fine-tuning on in-domain data, a technique that helps shape their response for the specific use-case. Here too we are faced with a variety of choices and methods: from standard full parameter tuning with next-word prediction to Direct Preference Optimization (Rafailov et al., 2023). Furthermore, merging model parameters is yet another powerful method to aggregate knowledge (Sung et al., 2023).

3. Proposed Methodology

We employ a hybrid approach, in which (a) we perform fine-tuning on a LLM and (b) we augment the input prompt with data extracted from our knowledge base. Apart from structured and unstructured fine-tuning, we experimented with various model merges, which generated a sensible leap in performance. Interestingly, even if our fine-tuning was done using Ukrainian data, the model merge was able to successfully preserve pre-existing knowledge while blending newly acquired Ukrainian capabilities.

3.1. Fine-tuning experiments

In our initial assessment we experimented with multiple open-source LLMs (see Section 4 for results), and chose Mistral-7b(Jiang et al., 2023) (instruct version) as the backbone of our system as it was the best performing out-of-the-box model.

Starting from the base model (Mistral-7B-Instruct-v0.2), we performed a set of experiments that resulted in diverging models (see Figure 1), which were evaluated in an end-to-end manner. We used 4 datasets: IT, U, DPO and KB datasets, detailed in the next section. The following steps were taken:

- **Step 1a - Supervised finetuning on the IT-Dataset**: We fine tune the base model for 3 epochs, using a curated dataset of instructions in Ukrainian with the standard supervised trainer (SFT);

- **Step 1b - Supervised finetuning on the IT-Dataset + U-Dataset)**: Similar to step 1a, but using the IT and U Datasets. This step is designed to help with multiple choice questions as well as open-ended questions;

- **Step 2 - Model merge**: We merge each of the models resulted in steps 1a and 1b with the Neuraltrix[1] model, which is a direct preference optimized (DPO) variant of the baseline Mistral model. This step and model choice were introduced based on empirical evaluations of the output. The merge method was Spherical Interpolation. The results of this step are two models: the merges of steps 1a and 1b with the NeuralTrix model;

- **Step 3 - DPO tuning (DPO-Dataset)**: We further refine the two merged models by Direct Preference Optimization using the Ukrainian translated DPO dataset;

[1] https://huggingface.co/CultriX/NeuralTrix-7B-dpo

- **Step 4 - RAG enrichment (KB-Dataset)**: Used only for the RAG-enabled approach, we perform RAG enrichment of the prompts for every question;

In the evaluation section (Section 4) we perform ablation tests, but only for a test-set consisting of multiple choice answers which could be evaluated automatically without requiring human expert input.

3.2. Dataset Description

To finetune our models we used the following datasets:

(a) IT Dataset - a dataset used for instruction tuning, obtained by merging in a similar format AlpacaDataset (Taori et al., 2023), SQuAD (Ivanyuk-Skulskiy et al., 2021), Ukrainian StackExchange[2], QUA-RC(Zyrianova and Kalpakchi, 2023), XQA (Liu et al., 2019), Belebele (et al., 2023) and the ZNO Dataset provided by UNLP[3](Syvokon et al., 2024). For datasets that were not in Ukrainian, we automatically translated the content[4] (jussa et al., 2022).

(b) DPO Dataset - a dataset for direct preference optimization which is a obtained by automatically translating the OpenOrca Dataset (Lian et al., 2023; Longpre et al., 2023).

(c) KB Dataset - used only during the RAG phase, this unstructured (free-text) dataset is composed from the Ukrainian Wikipedia and a curated list of Ukrainian school textbooks, listed in Table 1. The approximate size is 3.8 GB.

(d) U Dataset - a subset of the KB dataset, used in step 1, as the entire KB dataset was too large; the randomly paragraph-level sampled dataset size is 941 MB.

3.3. Retrieval Augmented Generation

In general, RAG is the procedure of enhancing the model's performance by adding information to the input prompt, based on a set of documents (the Knowledge Base). The procedure is straightforward: given a query (the question that needs answering) we first embed it as a semantic vector using any embedding transformer. The knowledge based is pre-segmented and embedded, and stored into a vector database that allows fast similarity search. The top-n documents that have the highest semantic similarity (lowest cosine distance) to the query embedding are those that will be added as context in the LLM's prompt.

However, there are a few caveats:

```
System: You are a teacher of the Ukrainian lan-
guage and you want to find some documents that
contain the answer to the request.
System: You will follow all instructions.
Instruction: You may have to do several searches
to get the answer.
======example 1=====
Question: What song did ADDT compose: (a)
Driving by the sea; (b) Movement
Answer: 3 queries need to be run:
(a) What is ADDT
(b) When Driving by the Sea was created
(c) When Movement is created
========example 2=====
Question: Why did Alice follow the rabbit?
Option 1: she was bored; Option 2: she was
interested;
Answer: The following requests must be made:
(a) Was Alice bored when she was following the
rabbit?
(b) Was Alice curious when she followed the
rabbit?
========example 3=====
Question: What is the correct form of the adjec-
tive formed from the noun water: a) watery; (b)
anhydrous
Answer: The following requests must be made:
(a) How do nouns become adjectives?
(b) Rules for the formation of the word water.
========
Instruction: When presented with multiple
choices, for each choice you should issue a
search query.
Instruction: Answer one item per line!
Instruction: Speak exclusively in Ukrainian.
Instruction: Do not translate back to English.
Instruction: Do not use your own knowledge to
directly answer the question.
Instruction: Given the above instructions, an-
swer the following prompt:
Question: {query}

Answer:
```

Figure 2: Prompt used to split the input query into subtasks. The query contains both the question and variants if it is a multiple choice question.

(a) Semantic vectors are a very effective instrument for information retrieval only if the topic/subject is consistent throughout the input text. This is because the representation capacity of a fixed-size vector is finite, and fitting multiple topics into this finite vector would result in representation conflicts. Thus, one of the prerequisites for a high performing information retrieval system is performing **accurate topic-based segmentation of the input documents.**

(b) Additionally, computing semantic vectors is a **language dependent task** and, in our initial experiments, **most out-of-the-box models were underperforming on Ukrainian**.

(c) The context window of the LLM plays a major role in deciding **how much content to retrieve from the KB**. Open-source models usually have smaller context windows than commercial-grade LLMs, which in turn requires a reduction in the amount of input data received from the RAG phase.

With their limitations in mind, we designed a custom retrieval system that (a) works directly with **keyword indexing and search** and (b) **uses a LLM to sequentially extract information** and filter out

[2] https://huggingface.co/datasets/zeusfsx/ukrainian-stackexchange

[3] https://github.com/unlp-workshop/unlp-2024-shared-task/blob/main/data/zno.train.jsonl

[4] Translations were performed with NLLB-3B

Name	Description
Довідник з укр. мови та літ.: Завдання в тестовій формі - Частина 1 (*Ukrainian Language and Literature Handbook: Test Form Tasks - Part 1*), О. М. Авраменко, М. Б. Блажко	Handbook providing test form tasks related to Ukrainian language and literature, aimed at aiding in learning and assessment.
Львівський Регіональний Центр Оцінювання Якості Освіти: Українська Мова (*Lviv Regional Center for Educational Quality Assessment: Ukrainian Language*), Збірник завдань для підготовки до зовнішнього незалежного оцінювання, Львів 2007	Collection of tasks for preparation for external independent evaluation in Ukrainian language.
Практикум з правопису і граматики української мови (*Workbook on Ukrainian Language Spelling and Grammar*), І.П. Ющук	Handbook approved for use in general educational institutions by the Commission on the Ukrainian Language of the Scientific and Methodological Council on Education of the Ministry of Education and Science, Youth and Sports of Ukraine. This workbook combines theoretical principles with practical tasks, aiding in the understanding of Ukrainian language grammar and improvement of spelling skills.
Новий довідник: Українська мова. Українська література (*New Handbook: Ukrainian Language. Ukrainian Literature*), М. Радишевська, В. Погребенник, В. Михайлюта, Т. Корольова, Т. Трош, О. Гудзенко	Handbook covering Ukrainian language and literature for school curriculum. Contains concise text and illustrative examples for thorough understanding and quick mastery of the material. Useful for exam preparation and entrance into higher education institutions.
Український Правопис (*Ukrainian Orthography*), Затверджено Кабінетом Міністрів України, 2019	Official Ukrainian orthographic rules approved by the Cabinet of Ministers of Ukraine, the Presidium of the National Academy of Sciences of Ukraine, and the Collegium of the Ministry of Education and Science of Ukraine. It provides guidelines for spelling, punctuation, and grammar, aiming to maintain consistency and clarity in written Ukrainian language.
Українська Література: Довідник для підготовки до ЗНО-2021 (*Ukrainian Literature: Handbook for the Preparation for the External Independent Evaluation 2021*), Дмитро Заєць	Handbook providing summaries and analyses of literary works covered in the Ukrainian literature curriculum for the 9th, 10th, and 11th grades, including works from ancient Ukrainian literature to contemporary authors, along with key literary terms and concepts.
Історія України: Хронологічний і термінологічний довідник для підготовки до ЗНО (*History of Ukraine: Chronological and Terminological Handbook for the Preparation for the External Independent Evaluation*), Олександр Геннадійович Полтавцев	Handbook providing key dates, concepts, and information on historical figures for the Ukrainian history program in preparation for the External Independent Evaluation (EIE)
100 тем. Історія України (*100 Topics. History of Ukraine*), Г. Т. Децюрін	A comprehensive school course in 100 themes, designed to present the most essential and obligatory topics for understanding the history of Ukraine. This book is aligned with the educational program of the Ministry of Education and Science of Ukraine, enabling systematic self-study and reinforcing key historical concepts, terms, and definitions.
Історія України. 10–11 класи: Наочний довідник (*History of Ukraine. Grades 10–11: Visual Guide*), О. В. Гісем, О. О. Мартинюк	Visual guidebook providing a clear and structured presentation of historical events, designed to aid students in grades 10 and 11 with the systematic study of Ukrainian history. It follows the educational standards set by the Ministry of Education and Science of Ukraine.
Історія України (*History of Ukraine*), О.Д. Бойко, 2002	A guidebook for the history of Ukraine, approved by the Ministry of Education and Science of Ukraine as an educational manual for higher educational institutions. It is characterized by its precision in language form, clarity in the expression of thoughts, and facts that are set against the background of significant trends in the comprehension of historical events, contributing to the formation of students' concrete historical knowledge.

Table 1: Materials used in our unstructured dataset

unwanted data:

Step 1: Ask the LLM to analyze the input query and extract a series of searches required to answer the question (see Figure 2 for the prompt). There can be any number of independent searches, for every topic, term, definition, artwork, book etc. present the input data;

Step 2: Take every previously generated item and ask the LLM to imagine the keywords that need to be used in the search process (see Figure 3 for prompt). The LLM is instructed to generate unigrams, bigrams and trigrams. Bi- and trigrams, are used in a document scoring process;

Step 3: Use full-text indexing (we used OpenSearch[5] as the backend), to look for the keywords in the documents and retrieve the content. It is important to mention that we keep the documents as a whole and we avoid any pre-segmenting of the data;

Step 4: Score the documents, based on bi- and trigrams and keep only the first top-k[6] in the queue (scoring details follow).

Step 5: Take each paragraph in the input data, with a limited context window[7] and ask the LLM to look and the original query and at the paragraph and say if it could help with that query in any way - if the LLM says "yes", the paragraph goes in a special queue (see Figure 4 for prompt);

Step 6: RAG is done by combining the selected paragraphs, with the original document titles and presenting them as documents in the final prompt (see Figure 5 for details).

Figure 6 shows the execution steps for the query *"Elements of expressionism are present in the work: (a) "Stone Cross", (b) "Institute", (c) "Marusya""*, where we translated interesting portions for reader

[5] https://opensearch.org/ - accessed 2024-03-28

[6] In our experiments we used $k = 6$

[7] For context, we used one paragraph above, one paragraph below and the title of the original document

```
System: You are a teacher of the Ukrainian language and you want to find some documents to find the document that contains the answer to the query.
Instruction: Write several keywords that will be used to search for relevant documents. Creation of unigrams, bigrams and trigrams.
Instruction: Output unigrams, bigrams and trigrams in three separate lines.
Instruction: You will follow the instructions exactly. Do not write anything else.
Instruction: The first line must contain unigrams (individual words) separated by commas.
Instruction: The second line should contain bigrams (groups of two words) separated by commas.
Instruction: The third line must contain trigrams (groups of three words) separated by commas.
Instruction: Write only in Ukrainian.
======example=====
Context: What Asimov wrote first: (a) Foundation (b) I robot
Request: Learn more about the Isaac Asimov Foundation
Answer: UNIGRAMS: Isaac, Asimov, Foundation, book, chapter, summary, content, genre, year
BIGRAMS: Isaac Asimov, brief description, Foundation description, publication date, Foundation characters, Foundation genre
TRIGRAMS: Foundation of Isaac Asimov, Genre of the book of the Foundation, Publication date of the Foundation, Book of the Foundation about
================
Instruction: Speak only Ukrainian.
Instruction: words should be separated by spaces, not underscores.
Instruction: Given the instructions above, solve the following problem:
Context: {query}
Request: Learn more about {step}
Answer:
```

Figure 3: Prompt used to get the search terms for each generated step in phase 1.

convenience. As shown, the model successfully extracts the search phases, retrieved documents for each step and manages to get the right context in order to answer that elements of expressionism are present in “Stone Cross”. In this case, the source document was a Wikipedia page.

The document scoring algorithm is simple, because we rely on the LLM to perform the heavy lifting and generate good input data. We take each n-gram generated by the model and we split it into tokens. We then look for tokens within the text and if all tokens from the same n-gram appear very close to each other (within a 5-word window), we add +1 to the document's score. If an n-gram appears multiple times, the score will be increased each time the context conditions are satisfied. In the end, we sort the documents based on their descending score, keeping only the top-k documents as RAG results.

Note 1: The choice in the number of document for RAG might be sub-optimal. We set k=6 strictly based on speed constraints.

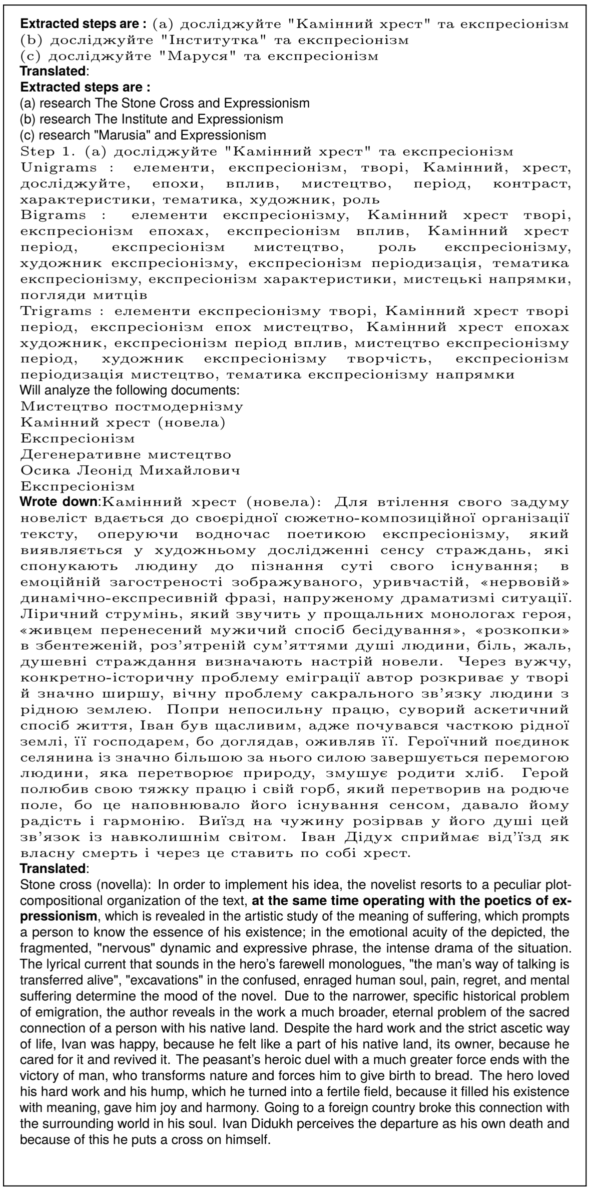
Note 2: The context window for the paragraph that is being analyzed was not tested against other options, which means that bigger context or heuris-

```
System: You are a Ukrainian student trying to find an answer to a question
System: Follow all instructions
Instruction: You will receive a paragraph from a document, and you need to find the answer in it.
Instruction: If the document is not current, write "No"
Instruction: If the document is current, write "Yes"
Instruction: Answer Yes or No!
Instruction: Answer in Ukrainian
================ Example 1 - your answer in the text=============
Query: does the text answer the question: is Isaac Asimov the author of the Foundation
Document: Title: About the Foundation
Synopsis: Foundation is a novel written by Isaac Asimov and is part of a saga.
Answer: Yes
================
================ Example 2 - your answer is not in the text=========
Query: does the text answer the question: is Isaac Asimov the author of the Foundation
Title: About the foundation
Contents: Isaac Asimov wrote many novels.
Answer: No
================
Instruction: Follow all the above instructions.
Instruction: Additional text, thoughts or ideas are not allowed.
Instruction: Consider whether you can get any useful information from the text. It is very important that you do not miss the clues!!!!
Instruction: Using the above instructions and examples, answer the following prompt:
Instruction: Look carefully at the context. If the document can help answer what is in the context, then your answer should be yes. It is very important!
Query: Does the text contain the answer to the question: {query}
Title:{title}
Content: {content}
Answer:
```

Figure 4: Prompt used to analyze paragraphs and extract relevant content.

```
System: You are a Ukrainian teacher specializing in literature and history.
System: If the request is in Ukrainian, answer with the letters corresponding to the best options from the list of possible answers.
System: When answering a question, return all correct options (e.g. "Option 3: Golden Gate Bridge")
System: You will fulfill all user requests
Instruction: There is only one correct answer to the question.
======example:
These are the documents:
About the chicken and the egg:
The chicken came before the egg.
Question: What came first:
Option 1: Chicken;
Option 2: Egg
Answer: Option 1: Chicken;
==================
System: Base your knowledge primarily on these documents:
{docs}
Query: {query}
Answer:
```

Figure 5: Prompt used to provide final answer, based on RAG



Extracted steps are : (a) досліджуйте "Камінний хрест" та експресіонізм
(b) досліджуйте "Інститутка" та експресіонізм
(c) досліджуйте "Маруся" та експресіонізм

Translated:

Extracted steps are :
(a) research The Stone Cross and Expressionism
(b) research The Institute and Expressionism
(c) research "Marusia" and Expressionism

Step 1. (a) досліджуйте "Камінний хрест" та експресіонізм

Unigrams : елементи, експресіонізм, творі, Камінний, хрест, досліджуйте, епохи, вплив, мистецтво, період, контраст, характеристики, тематика, художник, роль

Bigrams : елементи експресіонізму, Камінний хрест творі, експресіонізм епохах, експресіонізм вплив, Камінний хрест період, експресіонізм мистецтво, роль експресіонізму, художник експресіонізму, експресіонізм періодизація, тематика експресіонізму, експресіонізм характеристики, мистецькі напрямки, погляди митців

Trigrams : елементи експресіонізму творі, Камінний хрест творі період, експресіонізм епох мистецтво, Камінний хрест епохах художник, експресіонізм період вплив, мистецтво експресіонізму період, художник експресіонізму творчість, експресіонізм періодизація мистецтво, тематика експресіонізму напрямки

Will analyze the following documents:
Мистецтво постмодернізму
Камінний хрест (новела)
Експресіонізм
Дегенеративне мистецтво
Осика Леонід Михайлович
Експресіонізм

Wrote down:Камінний хрест (новела): Для втілення свого задуму новеліст вдається до своєрідної сюжетно-композиційної організації тексту, оперуючи водночас поетикою експресіонізму, який виявляється у художньому дослідженні сенсу страждань, які спонукають людину до пізнання суті свого існування; в емоційній загостреності зображуваного, уривчастій, «нервовій» динамічно-експресивній фразі, напруженому драматизмі ситуації. Ліричний струмінь, який звучить у прощальних монологах героя, «живцем перенесений мужичий спосіб бесідування», «розкопки» в збентеженій, роз'ятреній сум'яттями душі людини, біль, жаль, душевні страждання визначають настрій новели. Через вужчу, конкретно-історичну проблему еміграції автор розкриває у творі й значно ширшу, вічну проблему сакрального зв'язку людини з рідною землею. Попри непосильну працю, суворий аскетичний спосіб життя, Іван був щасливим, адже почувався часткою рідної землі, її господарем, бо доглядав, оживляв її. Героїчний поєдинок селянина із значно більшою за нього силою завершується перемогою людини, яка перетворює природу, змушує родити хліб. Герой полюбив свою тяжку працю і свій горб, який перетворив на родюче поле, бо це наповнювало його існування сенсом, давало йому радість і гармонію. Виїзд на чужину розірвав у його душі цей зв'язок із навколишнім світом. Іван Дідух сприймає від'їзд як власну смерть і через це ставить по собі хрест.

Translated:

Stone cross (novella): In order to implement his idea, the novelist resorts to a peculiar plot-compositional organization of the text, **at the same time operating with the poetics of expressionism**, which is revealed in the artistic study of the meaning of suffering, which prompts a person to know the essence of his existence; in the emotional acuity of the depicted, the fragmented, "nervous" dynamic and expressive phrase, the intense drama of the situation. The lyrical current that sounds in the hero's farewell monologues, "the man's way of talking is transferred alive", "excavations" in the confused, enraged human soul, pain, regret, and mental suffering determine the mood of the novel. Due to the narrower, specific historical problem of emigration, the author reveals in the work a much broader, eternal problem of the sacred connection of a person with his native land. Despite the hard work and the strict ascetic way of life, Ivan was happy, because he felt like a part of his native land, its owner, because he cared for it and revived it. The peasant's heroic duel with a much greater force ends with the victory of man, who transforms nature and forces him to give birth to bread. The hero loved his hard work and his hump, which he turned into a fertile field, because it filled his existence with meaning, gave him joy and harmony. Going to a foreign country broke this connection with the surrounding world in his soul. Ivan Didukh perceives the departure as his own death and because of this he puts a cross on himself.

Figure 6: Sample output for the query "Елементи експресіонізму наявні у творі: (a) «Камінний хрест», (b) «Інститутка», (c) «Маруся»" - translated: *"Elements of expressionism are present in the work: (a) Stone Cross, (b) Institute, (c) Marusya"*

tic methods of establishing the right context window might yield better results. Also, we expect that a dedicated segmentation model would perform better than our current approach.

Note 3: Instead of asking the model to decide if a paragraph is useful or not for the query, we experimented with making the model take notes and use the notes to generate the final answer. This decreased the accuracy of the RAG process for Ukrainian, but it showed promising results for English. This finding merits further exploration.

Note 4: All the models we experimented with followed instructions better when they were written in English, regardless of them being fine-tuned or not for Ukrainian.

Note 5: Though we tried to constrain the model to generate unigrams, bigrams and trigrams separately, this was not always successful. As such, our scoring mechanism does not enforce exact n-gram count. Instead, it just tokenizes the input based on the "white space" character and works with any number of resulting tokens.

Note 6: For multiple choice questions, asking the model to produce the output as "Option 1: ...." (text of the option included) yielded better results, because the model seemed to follow this type of instruction better than just being ask to respond with the option number. This is a somewhat expected behaviour as "forcing" the model "explain" its choice makes it better ponder the option - probably more attention is placed on the response in relation to the option description, thus "reasoning" better.

4. Evaluation and Results

We present results of the evaluation carried out against baseline models, of our fine-tuning experiments, merges and assess the performance of our RAG system.

We summarize our results in Table 2. Our initial assessment focused on baseline model evaluation, in order to see what architecture would perform best. Ideally, we would experiment on all base models the same way, but due to time and resource limitations, we had to focus on a specific architecture alone. Thus, we scored 3 baseline models with no-RAG on the ZNO dataset: *Llama-2-7B-32K-Instruct*, *gemma-7b-it* and *Mistral-7B-Instruct-v0.2*. We went with the instruct models, because the vanilla instances fail to follow instructions and are hard to score.

As shown, *Mistral-7B-Instruct-v0.2*, has an out-of-the box accuracy of 30.89%, versus the other two models that fall bellow 24%. Interestingly, we note that the baseline obtained by only returning option 1 across the entire dataset is around 24%, which throws Llama and Gemma below this threshold and Mistral slightly over.

In the next experiment we added RAG on top of the baseline Mistral model, which increased its accuracy by an additional 10%, from 30.89% to 40.21%.

For the next phase, we fine-tuned the baseline model, first by using just IT Dataset and then by combining IT and U Datasets. Non-RAG results are 32.75% and 33.02%, while the RAG-enhanced results are 40.87% and 41.14% respectively. This shows that, in some cases, tuning with free text and instruction data at the same time yields better results, provided that the ratio between the two sets

Base Model	Finetuned	RAG	Merged	Acc. (%)
Llama-2-7B-32K-Instruct	No	No	No	19.13
gemma-7b-it	No	No	No	23.70
CultriX/NeuralTrix-7B-dpo	No	No	No	31.95
CultriX/NeuralTrix-7B-dpo	No	Yes	No	36.48
Mistral-7B-Instruct-v0.2	No	No	No	30.89
Mistral-7B-Instruct-v0.2	No	Yes	No	40.21
Mistral-7B-Instruct-v0.2	IT Dataset	No	No	32.75
Mistral-7B-Instruct-v0.2	IT Dataset	Yes	No	40.87
Mistral-7B-Instruct-v0.2	IT + U Datasets	No	No	33.02
Mistral-7B-Instruct-v0.2	IT + U Datasets	Yes	No	41.14
Mistral-7B-Instruct-v0.2	No	No	CultriX/NeuralTrix-7B-dpo	40.04
Mistral-7B-Instruct-v0.2	No	Yes	CultriX/NeuralTrix-7B-dpo	**47.00**
Mistral-7B-Instruct-v0.2	IT Dataset	No	CultriX/NeuralTrix-7B-dpo	39.94
Mistral-7B-Instruct-v0.2	IT Dataset	Yes	CultriX/NeuralTrix-7B-dpo	48.46
Mistral-7B-Instruct-v0.2	IT + U Datasets	No	CultriX/NeuralTrix-7B-dpo	41.94
Mistral-7B-Instruct-v0.2	IT + U Datasets	Yes	CultriX/NeuralTrix-7B-dpo	**49.13**

Table 2: Results obtained on the ZNO dataset by different network architectures, pretrained variants, fine-tuned models and merges.

is close to 1.

The final stage of our experiments focused on model merges. For this, we used *CultriX/NeuralTrix-7B* as the second fine-tuned variant of Mistral. Note, that this is not an instruction-tuned model, so its accuracy on the dataset is very low. We merged (Spherical Interpolation) the previously presented baseline models, IT Dataset tuned and mixed tuned variants with this new model and we performed DPO tuning for one epoch on the result, using the translated DPO Dataset. The results for non-RAG vs RAG optimized prompt are 40.04%–47.00% (for the base model), 39.94%–48.46% (for the IT Dataset variant) and 41.94%–49.13% (for the mixed variant).

Interestingly, there is a drop in performance for the IT Dataset tuned and merged model with the no-RAG flavour evaluation, but the RAG optimized generation is better.

Finally, the best performing recipe was:

Step 1: Start with Mistral-7B-Instruct-v0.2 and perform fine-tuning on the combination of IT + U Datasets;

Step 2: Merge with CultriX/NeuralTrix-7B using Spherical Interpolation;

Step 3: Perform DPO tuning on the resulting model, in our case using the DPO Dataset;

Step 4: Produce results using RAG-enhanced prompts.

5. Conclusions and Future Work

We present *Sherlock*, our proposed system that achieved first place in the UNLP 2024 competition. Our system is a set of data-augmentation techniques mixed with custom LLMs. We enumerate key points in each of the data, prompting and LLM-tuning areas:

Datasets: (a) We used many available data sources: Ukrainian Wikipedia and manually selected relevant books on the target subject; (b) We translated and used several datasets, both free-text and instruction-formatted

Retrieval Augmented Generation: (a) Due to limiting factors in the standard RAG process (e.g. embedding for Ukrainian does not have great performance), we used n-grams to provide better results than the standard similarity score; (b) We used the LLM itself to generate n-grams.

LLM tuning: (a) We started from an already very good instruction tuned model - Mistral 7B; (b) We tried standard finetuning on different datasets; (c) We experimented with model weight merging; (d) We further enhanced performance by DPO training; (e) Having a test set enabled us to experiment with different combinations of the individual methods above, to achieve an overall better result than each individual method.

Finally, we are happy to announce that, for reproductability we release both the source code[8] and the model[9], hoping that this will further advance efforts in building afordable LLMs that can run on consumer-grade products, with low computational requirements.

[8] https://github.com/adobe/sherlock-backend/tree/UNLP2024

[9] https://huggingface.co/SherlockAssistant/Mistral-7B-Instruct-Ukrainian

6. Bibliographical References

Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altenschmidt, Sam Altman, Shyamal Anadkat, et al. 2023. Gpt-4 technical report. *arXiv preprint arXiv:2303.08774*.

Joshua Ainslie, James Lee-Thorp, Michiel de Jong, Yury Zemlyanskiy, Federico Lebrón, and Sumit Sanghai. 2023. Gqa: Training generalized multi-query transformer models from multi-head checkpoints. *arXiv preprint arXiv:2305.13245*.

Lucas Bandarkar et al. 2023. The belebele benchmark: a parallel reading comprehension dataset in 122 language variants. *arXiv preprint arXiv:2308.16884*.

Patrick Lewis et al. 2021. Retrieval-augmented generation for knowledge-intensive nlp tasks.

Bogdan Ivanyuk-Skulskiy, Anton Zaliznyi, Oleksand Reshetar, Oleksiy Protsyk, Bohdan Romanchuk, and Vladyslav Shpihanovych. 2021. ua_datasets: a collection of ukrainian language datasets.

Albert Q Jiang, Alexandre Sablayrolles, Arthur Mensch, Chris Bamford, Devendra Singh Chaplot, Diego de las Casas, Florian Bressand, Gianna Lengyel, Guillaume Lample, Lucile Saulnier, et al. 2023. Mistral 7b. *arXiv preprint arXiv:2310.06825*.

Zhengbao Jiang, Zhiqing Sun, Weijia Shi, Pedro Rodriguez, Chunting Zhou, Graham Neubig, Xi Victoria Lin, Wen tau Yih, and Srinivasan Iyer. 2024. Instruction-tuned language models are better knowledge learners.

Marta R Costa jussa et al. 2022. No language left behind: Scaling human-centered machine translation. *arXiv preprint arXiv:2207.04672*.

W Lian, B Goodson, E Pentland, et al. 2023. Openorca: An open dataset of gpt augmented flan reasoning traces.

Jiacheng Liu, Alisa Liu, Ximing Lu, Sean Welleck, Peter West, Ronan Le Bras, Yejin Choi, and Hannaneh Hajishirzi. 2022. Generated knowledge prompting for commonsense reasoning.

Jiahua Liu, Yankai Lin, Zhiyuan Liu, and Maosong Sun. 2019. Xqa: A cross-lingual open-domain question answering dataset. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 2358–2368.

Shayne Longpre, Le Hou, Tu Vu, Albert Webson, Hyung Won Chung, Yi Tay, Denny Zhou, Quoc V. Le, Barret Zoph, Jason Wei, and Adam Roberts. 2023. The flan collection: Designing data and methods for effective instruction tuning.

James Manyika and Sissie Hsiao. 2023. An overview of bard: an early experiment with generative ai. *AI. Google Static Documents*, 2.

Rafael Rafailov, Archit Sharma, Eric Mitchell, Stefano Ermon, Christopher D. Manning, and Chelsea Finn. 2023. Direct preference optimization: Your language model is secretly a reward model.

Yi-Lin Sung, Linjie Li, Kevin Lin, Zhe Gan, Mohit Bansal, and Lijuan Wang. 2023. An empirical study of multimodal model merging.

Oleksiy Syvokon, Mariana Romanyshyn, and Roman Kyslyi. 2024. The UNLP 2024 shared task on fine-tuning large language models for Ukrainian. In *Proceedings of the Third Ukrainian Natural Language Processing Workshop*, Torino, Italy. European Language Resources Association.

Rohan Taori, Ishaan Gulrajani, Tianyi Zhang, Yann Dubois, Xuechen Li, Carlos Guestrin, Percy Liang, and Tatsunori B Hashimoto. 2023. Stanford alpaca: An instruction-following llama model.

Hugo Touvron, Louis Martin, Kevin Stone, Peter Albert, Amjad Almahairi, Yasmine Babaei, Nikolay Bashlykov, Soumya Batra, Prajjwal Bhargava, Shruti Bhosale, et al. 2023. Llama 2: Open foundation and fine-tuned chat models. *arXiv preprint arXiv:2307.09288*.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. *Advances in neural information processing systems*, 30.

Qin Zhang, Shangsi Chen, Dongkuan Xu, Qingqing Cao, Xiaojun Chen, Trevor Cohn, and Meng Fang. 2023. A survey for efficient open domain question answering. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 14447–14465, Toronto, Canada. Association for Computational Linguistics.

Mariia Zyrianova and Dmytro Kalpakchi. 2023. Quarc: the semi-synthetic dataset of multiple choice questions for assessing reading comprehension in ukrainian. *Northern European Journal of Language Technology*, 9(1).

From Bytes to Borsch: Fine-Tuning Gemma and Mistral for the Ukrainian Language Representation

Artur Kiulian[1], Anton Polishko[1], Mykola Khandoga[1,2],
Oryna Chubych[1], Jack Connor[3], Raghav Ravishankar[4],
Adarsh Shirawalmath[4]
[1]PolyAgent, [2]Mindee, [3]O'Shaughnessy Ventures, [4]Tensoic.
Email: a@polyagent.co, anton@polyagent.co, mkhandoga@gmail.com, oryna.chubych@gmail.com,
jack.connor83@gmail.com, raghav@tensoic.com, adarsh@tensoic.com

Abstract

In the rapidly advancing field of AI and NLP, generative large language models (LLMs) stand at the forefront of innovation, showcasing unparalleled abilities in text understanding and generation. However, the limited representation of low-resource languages like Ukrainian poses a notable challenge, restricting the reach and relevance of this technology. Our paper addresses this by fine-tuning the open-source Gemma and Mistral LLMs with Ukrainian datasets, aiming to improve their linguistic proficiency and benchmarking them against other existing models capable of processing Ukrainian language. This endeavor not only aims to mitigate language bias in technology but also promotes inclusivity in the digital realm. Our transparent and reproducible approach encourages further NLP research and development. Additionally, we present the Ukrainian Knowledge and Instruction Dataset (UKID) to aid future efforts in language model fine-tuning. Our research not only advances the field of NLP but also highlights the importance of linguistic diversity in AI, which is crucial for cultural preservation, education, and expanding AI's global utility. Ultimately, we advocate for a future where technology is inclusive, enabling AI to communicate effectively across all languages, especially those currently underrepresented.

Keywords: Gemma 2b, Gemma 7b, Mistral 7b, LLM, Ukrainian, Multilingual Models, LoRA, Fine-Tuning

1. Introduction

The field of Natural Language Processing (NLP) is expanding extremely quickly today, largely due to the immense success of the generative Large Language Models (LLM). Within only a few years, these language models have become capable of performing tasks like contextual understanding and generation, few-shot learning, automated question answering, sentiment analysis, emotion detection, and many others with unprecedented quality.

1.1. Background

The significance of recent NLP advances, obtained in such a short time, becomes even more evident looking back at the long history of quantitative language modeling. The first attempts to attack the problem of computational linguistics date back as far as 70 years ago, to the early 1950s (Shannon, 1951).

But it was not until the 2000s when the artificial Neural Network (NN) proved its effectiveness in the field (Bengio et al., 2000), notably applied to the machine translation problem (Schwenk et al., 2006). These models were mostly based on Recurrent Neural Networks (RNN) architecture like Long Short-Term Memory (LSTM) (Hochreiter and Schmidhuber, 1997) and later Gated Recurrent Unit (GRU) (Chung et al., 2014). Still, important milestones were achieved during this period like the introduction of word embeddings.

However, throughout most of the 2010s, while other fields of Deep Learning (DL) like Computer Vision (CV) and Reinforced Learning (RL) have achieved very impressive results (Krizhevsky et al., 2012; He et al., 2015; Silver et al., 2016), the NN-powered NLP field still suffered from a number of problems. This included the handling of long-term dependencies, capturing bidirectional context and overall difficulties with computational efficiency and stability.

The breakthrough came with the invention of the transformer architecture which introduced the key component: the attention mechanism (Vaswani et al., 2017).

1.2. The transformer era

The attention mechanism addresses the challenges of understanding both the immediate and broader context of words in a sentence, solving issues related to bidirectional context, long-term dependencies, and convergence. Furthermore transformer architecture enhances the ability to process data in parallel, significantly outperforming RNNs in this regard. This advancement has paved the way for the development of LLMs: highly complex language models with billions of parameters, trained on extensive corpora of text.

The early LLMs like Bidirectional Encoder Representations from Transformers (BERT) (Devlin et al.,

The Third Ukrainian Natural Language Processing Workshop (UNLP) @LREC-COLING-2024, pages 83–94
25 May, 2024. © 2024 ELRA Language Resource Association: CC BY-NC 4.0

2019) and its successors have focused on understanding text and problems like text classification, emotion recognition, etc. Although, with the emergence of the Generative Pre-trained Transformer (GPT) family (Radford and Narasimhan, 2018; Radford et al., 2019; Brown et al., 2020; OpenAI, 2023), focus has shifted towards generative tasks.

Training an LLM from scratch remains a cumbersome and costly task. Nevertheless the general nature of the training corpora allows them to fully benefit from transfer learning, implementing the *pre-training and fine-tuning* paradigm: once a model is pre-trained on a large language corpus it can be further fine-tuned for a specific use-case, requiring relatively minor costs.

The LLMs available on the market can be split into two groups: proprietary and open-source. Proprietary models like GPT-4 and Gemini (Team, 2023) tend to have more parameters and offer high out of the box performance in most common tasks, but their use is restricted by the providers and allows limited fine-tuning options. Open-source models like LLaMa2 (Touvron et al., 2023), Mistral (Jiang et al., 2023), Mixtral (Jiang et al., 2024) or a recent Gemma by Google (Gemma Team, 2024) offer full access to the model code and weights and impose little to no restrictions on the use of the model, making it a natural choice for fine-tuning experiments. Open-source models often come in a variety of sizes in terms of parameter number, allowing lighter models to be run on consumer-grade GPUs.

1.3. Motivation and objective

A substantial number of open-source models are available on the market today. At the same time all these models demonstrate a notable bias towards the English language due to their training conditions. The bias can manifest itself in a number of ways, including to but not limited to the following:

1. Language and cultural bias. This can impair a model's usability for non-English speakers and also perpetuate stereotypes or misunderstandings about cultures.
2. Ethical and fairness concerns. The same model may show considerably better performance with English-speaking users, leaving others with a subpar experiences.
3. Uneven knowledge representation. This can lead to a skewed representation of global knowledge, history, and perspectives, and embed these biases into the model's outputs and decision-making processes.

The bias becomes particularly prominent in non-European languages and languages that do not use a Latin alphabet.

This has naturally motivated numerous scholars and enthusiasts to put much efforts into fine-tuning open-source models, predominantly LLaMa 2, in many languages, both European (Basile et al., 2023; Vanroy, 2023) and non-European (Cui et al., 2024; Gala et al., 2024a,b; Nguyen et al., 2023; Azime et al., 2024; Kohli et al., 2023). Most of the listed articles have been published within the last months, and demonstrate great interest and involvement in solving this linguistic bias issue. The immediate benefits of having an open-source model that is fine-tuned with a certain language include:

1. Reduction or elimination of cultural bias.
2. Flexibility in use-cases, including both academic and business.
3. Preservation of rare and low-resource languages.

The effort also promotes the creation of language-specific datasets and development of the LLM-oriented ecosystem. Even when a particular model becomes obsolete, further progress is greatly facilitated by this groundwork.

1.3.1. Ukrainian sector of the LLMs

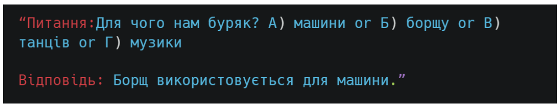
Ukraine is renowned for its dynamic IT community, which thrives both in academic circles and the commercial sector. The field of computational linguistics is no exception, boasting the inception of multi-billion dollar unicorns like Grammarly within its borders. With the advent of LLMs, there has been a keen interest in harnessing their capabilities for solving NLP challenges in the Ukrainian language.

Yet, until recently, these efforts have predominantly focused on leveraging BERT-like models (Tiutiunnyk, 2020; Laba et al., 2023; Katerynych et al., 2021), while the realm of generative LLMs has been somewhat overlooked. So far, UAlpaca is the only publicly available LLM that has been fine-tuned specifically for the Ukrainian language (Had). Likewise, instructional datasets in Ukrainian have been comparatively limited. The escalating enthusiasm for generative, GPT-style LLMs underscores the need for models attuned to Ukrainian linguistic and cultural nuances, further underlining the significance of our research endeavors.

1.3.2. Objectives

The aim of the effort presented in the current paper is multifold:

1. Create an open-source, free-to-use LLMs fine-tuned for Ukrainian language and culture thus expanding the Ukrainian presence in the NLP field.



```
“Питання:Для чого нам буряк? А) машини or Б) борщу or В)
танців or Г) музики

Відповідь: Борщ використовується для машини.”
```

Figure 1: Example of erroneous model inference in Ukrainian.

2. Compare the performance of different open-source LLMs, notably the SOTA Gemma model.
3. Benchmark the trained models using the dedicated Ukrainian dataset and compare them to the proprietary models.
4. Introduce the UKID instruction training dataset and make it publicly available for future fine-tuning efforts.
5. Perform the entire process in a fair and reproducible manner in order to facilitate future efforts.

2. Dataset and the experimental setup

Despite the abundance of online tutorials available for training large language models, establishing a reproducible setup for each model, complete with an appropriate dataset in the necessary format, proved to be unexpectedly challenging. Every model comes with its own set of constraints, including hardware requirements, deployment methods for inference, and specific approaches for processing instructions.

2.1. Dataset collection

When our team started working on the shared task for the UNLP conference, we were taken aback by the scarcity of suitable datasets for fine-tuning LLMs in Ukrainian. The organizers supplied a training dataset comprising 3,063 instruction rows, designed to acclimate the model to the multiple-choice format prevalent in the Ukrainian national examination. While this dataset proved valuable for training the LLM to answer in a specific format, it was notably limited in depth, offering little in terms of enhancing these LLMs' parametric knowledge base.

Through multiple experiments, we determined that 3-5 epochs of LoRA fine-tuning were sufficient for the model to grasp the multiple-choice format required for evaluation in the conference's shared task. However, the model's responses were lacking consistency, particularly when it generated incorrect or nonsensical answers. For instance, the model erroneously referred to "borsch," a well-known Ukrainian dish, as an item used in cars (See Figure 1).

This behavior underscored a deficiency in the model's general conceptual understanding, highlighting the pressing necessity to augment the dataset with more content in Ukrainian.

Consequently, we leveraged the **UAlpaca** dataset (Had) alongside **Squad-uk** (Drastic) which happened to be the only instruction datasets in the Ukrainian language available publicly.

Unfortunately, even after fine-tuning with these datasets, we observed that the model still didn't improve much, even on the training dataset itself, despite an improvement in sentence formulation and conceptual understanding. This led us to realize that a much more comprehensive approach to dataset construction would be required. Both UAlpaca and Squad-uk happened to be translated versions of the general knowledge English-based datasets, which is missing Ukrainian context and knowledge that is specific to both cultural and historical aspects that were being evaluated by the questions in the exam dataset. This realization led us to rethinking what kind of data we need and led to the creation of our own dataset, the Ukrainian Knowledge and Instruction Dataset (UKID), the first Ukrainian instruction dataset rooted in a Ukrainian context.

2.2. UKID methodology and construction

In formulating our hypothesis for the development of the Ukrainian language model, we posited that the model must align with the informational needs of the general population, reflecting the genuine interests and search behaviors of Ukrainian web users. To identify the most pertinent sources of intent-aligned knowledge, we turned to two widely recognized platforms: Wikipedia and Google. Consequently, we adopted a methodology focused on aggregating the most frequented Wikipedia pages, as determined by monthly traffic statistics, to ensure our dataset accurately captured the topics of highest relevance to Ukrainian web users.

We collected 1,064 pages by targeting those with monthly visit statistics ranging from 3,000 to 150,000. However, not all top-ranking Wikipedia pages in Google search results proved pertinent to our objective, as many described phenomena or entities not relevant to Ukraine. To refine our dataset, we employed a binary classification process to discern between relevant and non-relevant pages. This filtration mechanism is summarized in the table below, showcasing relevant versus non-relevant content (See Table 1). Through this methodical approach, we identified 367 pages that were suitable for inclusion in our dataset creation process.

The proposed methodology suggests an optimal approach for organizing an instruction-based dataset, aimed at fine-tuning language models for

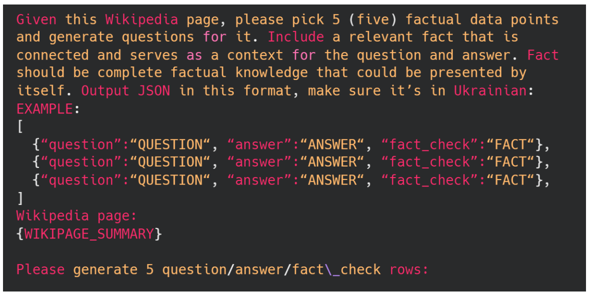
Page Title	Relevance
Ембер Герд	Not Relevant
Емульсія	Not Relevant
Ендокринна система	Not Relevant
Енеїда (Котляревський)	Relevant
Енцефаліт	Not Relevant
Еритроцити	Not Relevant
Єлизавета II	Not Relevant
Жадан і Собаки	Relevant
Жанр	Not Relevant
Житомир	Relevant

Table 1: Showcase of relevant vs non-relevant content.

underrepresented languages. This strategy offers the dual benefits of incorporating language-specific contexts and embedding essential factual knowledge into the model's trainable parameters during fine-tuning. Consequently, in addition to the conventional "question-answer" instruction pairs, we introduced a "fact_check" field. This addition acts as a comprehensive and standalone source of truth, enhancing the model's ability to verify facts and improve its accuracy. Performing this manually would have been unrealistic given the time constraints of the conference submission deadline, therefore an automated approach was implemented through the use of the Gemini 1.0 API and a few-shot learning example that utilizes the summary abstract of the Wikipedia page (See Figure.2)

As a result UKID-v0.1 was formed consisting of 962 question-answer-fact (QAF) pairs. Future work needs to focus on expanding the dataset to match other popular English-based datasets like Alpaca and Squad that consist of tens of thousands of rows. Even though the traditional notion of "less is more" for general English-based models recommends having smaller datasets (Zhou et al., 2023), our learnings indicate that fine-tuning under the constraints of lacking general conceptual understanding and context requires using much larger datasets.

Additionally, we have contemplated further enhancements to the UKID format, such as incorporating the original paragraphs from which the QAF pairs were derived to provide additional context. However, this aspect of the project remains unaddressed at present.



```
Given this Wikipedia page, please pick 5 (five) factual data points
and generate questions for it. Include a relevant fact that is
connected and serves as a context for the question and answer. Fact
should be complete factual knowledge that could be presented by
itself. Output JSON in this format, make sure it’s in Ukrainian:
EXAMPLE:
[
  {“question”:“QUESTION“, “answer”:“ANSWER“, “fact_check”:“FACT“},
  {“question”:“QUESTION“, “answer”:“ANSWER“, “fact_check”:“FACT“},
  {“question”:“QUESTION“, “answer”:“ANSWER“, “fact_check”:“FACT“},
]
Wikipedia page:
{WIKIPAGE_SUMMARY}

Please generate 5 question/answer/fact\_check rows:
```

Figure 2: Prompt to generate UKID examples

A crucial consideration in dataset development is tailoring the instruction format to the specific requirements of different models. For instance, Llama, Mistral, and Gemma each necessitate unique formats. Overlooking this critical aspect has empirically led to suboptimal outcomes, though these observations have yet to be formally documented. The adaptation of datasets to align with the distinct formats of these models is essential for maximizing their performance and efficacy.

3. Fine-tuning

3.1. Gemma models

First, we fine-tuned a Gemma-2B and a Gemma-7B model, from a recently published family of open models.

We used official "**gemma-2b-it**" and "**gemma-7b-it**" weights published by Google and followed official fine-tuning guidelines on the Vertex AI platform. The final python notebook is located in the "from-bytes-to-borsch" github repository.

Fine-tuning for gemma-2b-it was performed with a combined dataset consisting of 13,063 instructions, which included from the 10,000 rows of UAlpaca dataset and 3,063 rows from the ZNO dataset provided by organizers of the conference. Fine-tuning for gemma-7b-it was performed with a dataset consisting of 14,025 instructions (10,000 rows of UAlpaca, 3,063 rows of ZNO and 962 rows of UKID).

Due to resource constraints, we chose to use a LoRA (Hu et al., 2022) fine-tuning approach. We used a LoRA adapter implementation from the Keras v3 library, with $lora\_r = 4$, resulting in 11,067,392 trainable parameters, instead of the full 7B for the case of Gemma-7B.

The resulting model was published on the associated github repository. Unfortunately due to the time constraints we were not able to submit the 7B to the UNLP competition benchmarking, and only submitted results from the 2B instruct model.

3.2. Mistral model

As a second alternative, we used a completely different fine-tuning pipeline with the help of the axolotl tool to streamline the fine-tuning process. We used a 4x Nvidia Tesla A100-80Gb GPU instance on Microsoft Azure cloud for training. Due to compute constraints we chose to use the LoRA (Hu et al., 2022) approach once again, this time implemented using Hugging Face transformers library.

We used an AdamW optimizer (Loshchilov and Hutter, 2017) with common starting point hyperparameters for the LoRA adapters ($lora\_r = 32$, $lora\_alpha : 16$), which resulted in 32,505,856 trainable parameters.

For Mistral-based fine-tunes we used the "mistralai/Mistral-7B-Instruct-v0.1" weights and "LlamaTokenizer" tokenizer.

The training was performed using ZNO and Uk-Squad datasets. Both datasets have a Llama/Alpaca instruction format and collectively produced 37,890 rows of instructions.

More details of the configuration and execution can be found in the associated github repository.

4. Benchmarking results

We performed benchmarking using two test datasets: multiple choice questions (MCQ) and open questions (OQ).

The MCQ dataset comprises 3,063 questions from the Ukrainian External Independent Testing (EIT) test, a standard government test for college admission taken by secondary school students. This dataset splits into 1,139 Ukrainian history questions and 1,925 Ukrainian language and literature questions, reflecting the standard knowledge expected in Ukrainian schools. We evaluated this test automatically.

The OQ dataset contains 100 instruction-based questions prompting models to complete generative tasks, such as finishing a story or summarizing an event. We evaluated this dataset manually.

Below, we detail our benchmarking setup and criteria, focusing on the fine-tuned Gemma models, Gemma7bFT and Gemma2bFT, alongside an out-of-the-box model, Gemma7b, for reference.

4.1. Multiple choice questions

We presented all questions from this dataset within a uniform prompt in Ukrainian, instructing models to select the single correct answer in letter form. Despite this directive, models frequently included extraneous information, necessitating manual filtration to extract the required letter codes. Correct responses matched the letter codes exactly. Table 2 displays the models' performance percentages in each category.

Model	History (%)	L&L (%)
GPT4	82.95	47.12
Gemini	71.97	40.99
GPT3.5	52.37	26.65
MistralFT	40.16	22.86
Gemma7bFT	37.96	21.71
Gemma2bFT	28.91	20.57
Gemma7b	26.36	19.01

Table 2: Model benchmarking with multiple choice questions.

4.2. Open questions

Evaluating open questions required a more nuanced approach, examining responses across four categories:

- Ukrainian (U): the response is given in the Ukrainian language.
- Facts/Coherence (C): factual correctness and coherence of the given answer.
- Relevance (R): the answer aligns with the given instructions.
- Grammar (G): stylistic and grammatical evaluation.

Each response could earn up to 1 point per category, with the results and average scores presented in Table 3.

Model	U	C	R	G	Avg
GPT 4	97	79	85	79	85
GPT 3.5	97	61	79	74	77.75
Gemini	96	67	81	84	82
MistralFT	89	7	18	49	40.75
Gemma7b	85	13	45	35	44.5
Gemma7bFT	54	13	48	19	33.5

Table 3: Model benchmarking with open questions.

4.3. Discussion

The obtained results provide interesting insights into many aspects of the LLM's performance and training.

First, let us consider the results of the open-source models. Comparing the performance of the Gemma7b model before and after the fine-tuning it becomes very clear that the fine-tuning process can indeed improve its knowledge in a particular area by a large margin, in this case by roughly a quarter. Mistral shows even better improvement in answering the MCQs. Even the much smaller model Gemma2b outperforms its non-fine-tuned larger counterpart Gemma7.

However, besides improving model's performance in certain areas, the fine-tuning process appears to introduce artifacts that affect performance when answering these open questions. Mistral, after fine-tuning, seemed to struggle with following the given instructions (see the **R** column in Table 3). On the other hand, Gemma7bFT's ability to speak Ukrainian was impaired by 40%, also reducing its grammar score by nearly a half (columns **U** and **G** in Table 3). What's most exciting, Gemma7bFT started to manifest the *code-switching* phenomenon which can be considered an emergant property, and will be discussed in more detail in the Conclusions section.

It comes as no great surprise that the proprietary models performed substantially better in all kinds of tasks. The reasons are numerous, with the most obvious being:

- The scale of parameters significantly contributes to model performance. For instance, GPT-3.5 boasts 25 times more parameters than both Gemma7b and Mistral, whereas GPT-4 and Gemini exceed these models by over a hundredfold in terms of parameter count.
- Proprietary models benefit from unparalleled access to the most comprehensive and high-quality datasets available, ensuring a broad and deep understanding of language.
- The training of proprietary models extensively incorporates reinforcement learning techniques, refined through human feedback, to achieve nuanced understanding and response generation.

Nevertheless the performance of the fine-tuned open-source models is not so far behind that of GPT3.5. With additional efforts invested into the fine-tuning of open-source models, it is definitely possible to beat GPT3.5 in a range of specific language-related tasks.

A notable observation across all models was the disparate performance on Ukrainian history versus language and literature, echoing a trend irrespective of model origin. By design the EIT questions in different subjects are meant to be of the same complexity such that an average Ukrainian school student gets average marks in every subject. However, the performance of every LLM tested showed very skewed results, with history knowledge favoured over that of language and literature. Possible reasons could include:

- The skew in available datasets toward history is due to its widespread availability from open sources such as Wikipedia. Conversely, literature demands greater effort to gather, organize, and present, contributing to its underrepresentation.
- Answering history questions accurately is largely a matter of recalling specific factual information, such as dates, names, and events. Literary analysis, however, requires navigating complex themes, symbolism, and cultural nuances, demanding a more profound understanding of both language and context.
- The Ukrainian language, along with its cultural and literary heritage, often falls outside the primary interests of major corporations, affecting the availability and focus of datasets dedicated to these areas.

This underscores the cultural bias challenge in advanced LLMs today which will be further discussed in subsequent sections.

4.4. Code-switching and Azirivka

Code-switching is a linguistic phenomenon in which a speaker alternates between two or more languages within a single utterance or sentence. Until recently, this term was applied only to humans, but with the advent of LLMs this effect has been observed and studied in generative models (Winata et al., 2021; Zhang et al., 2023). Code-switching in LLMs arises from the multilingual nature of training and fine-tuning processes.

For historical reasons, the majority of the Ukrainian population is multilingual. This creates a rather unique situation when constant code-switching is common at practically every level, starting from colloquial everyday conversations and ending with official statements from prime-ministers and presidents. A particular case of the latter has the official name Azirivka (Wikipedia), named after Ukrainian ex-prime minister Mykola Azarov.

Observing the Gemma7b model mastering Azirivka after fine-tuning was both interesting and exciting. It is particularly interesting that the model generates not a simple mixture of words belonging to different languages, but rather conjugates words from one language according to the rules of another, just as some Ukrainians do, demonstrating features specific to synthetic languages.

Below, we present several instances of Azirivka code-switching. In these examples, components highlighted in blue represent Ukrainian, while those in red denote Russian.

Example 1:
Azirivka: Твір про коллекцию кольоровых олівцов Василя Голобородька.
English: An essay about Vasyl Holoborodko's collection of colored pencils.

Example 2:
Azirivka: Привітать друзів с одруждением можно множеством способов.
English: You can congratulate friends on their marriage in many ways.

Example 3:
Azirivka: Я обращаюсь к Вам с жалобой по неякісной замене труб в подвалі нашего дома, расположенного по [адрес].

English: I am addressing you with a complaint about the poor-quality replacement of pipes in the basement of our house, located at [address].

Example 4:
Azirivka: В Украине Маланку не святкуют.
English: Malanka is not celebrated in Ukraine.

Example 5:
Azirivka: У п'ятницю, 23 лютого, в Україні опадів не будет, но местами - рвучкий і сильний вітер.
English: On Friday, February 23, there will be no precipitation in Ukraine, but there will be occasional gusty and strong wind.

It's worth noting that while most of these mixed words can't be found in official dictionaries, they are commonly heard on the streets of many Ukrainian cities. Such a language mixture naturally has been an object for linguistic studies (Bilaniuk, 2004; Kent, 2011). We consider this emerging LLM property to be of great interest for further studies.

5. Applications, risks and future work

It is abundantly clear that having a language-specific model is going to aid all of the possible use cases around communication, but it's also important to note the risks of not having the model. Both from the industrial and cultural standpoints.

Incorporating LLM models of underrepresented languages into technology platforms offers unprecedented opportunities for enhancing communication across diverse sectors, ranging from healthcare and education to legal and commerce, all within the scope of the growing impacts of globalization. However, the absence of such models poses significant risks, not only stalling industrial progress but also exacerbating cultural erosion. Industrially, the lack of tailored language models can hinder the efficient dissemination of critical information, reduce the accessibility of digital services, and create barriers to entry for local businesses in the global market. Culturally, it threatens the preservation of linguistic diversity and the transmission of heritage, as languages without digital representation risk falling into disuse and oblivion. Therefore, addressing this gap is not merely a technical challenge but a pressing societal need that calls for collaborative efforts to ensure inclusive and sustainable development.

5.1. Applications

Oleksandr, a Ukrainian refugee in the USA, benefits from a language-specific LLM that digests and explains legal aid and immigration documents into Ukrainian. This tool helps him and his family understand their rights and the process for seeking asylum, significantly easing their transition into a new country while maintaining their linguistic identity during a period of immense upheaval and change.

Maria, a primary school teacher in a rural Peruvian village, uses a language-specific LLM to access educational materials in Quechua, enabling her to provide more engaging and culturally relevant lessons to her students. This technology allows her to bridge the gap between traditional knowledge and modern education, fostering a learning environment where students can appreciate their heritage while gaining access to the wider world of knowledge.

Michael, a software developer with Navajo heritage, creates an interactive application powered by a language-specific LLM that facilitates live, conversational practice in Navajo for learners worldwide. This platform connects Navajo speakers with learners, enhancing language proficiency through real-time dialogue and cultural exchange, thereby revitalizing the Navajo language among younger generations and spreading awareness of Navajo culture globally.

5.2. Risks through the prism of education

Classroom education and child development will depend heavily on large language models tailored for different languages and contexts, especially since there is no doubt in the growing influence of AI on youth, in particular within the educational and edutainment contexts (Chowdhury, 2023). That's why one may hypothesize that countries like Ukraine will eventually face a linguistic identity crisis in 15-20 years without accessible Ukrainian-tuned LLMs.

At the primary school level, Ukraine's youth increasingly speak a homogenized and influenced version of Ukrainian rather than preserved distinctive dialects. Besides an obvious impact of Russification, globalization makes it even harder to preserve Ukrainian heritage due to its decreasing utility when it comes to cultural integration into the global landscape. One might argue that Ukraine is having a unique moment in time where cultural identity is being amplified by the risk of complete wipeout by an invading neighbor country, but other developing countries may never have such unique constraints to enable cultural amplification and preservation.

One other risk is related to not having interactive AI tools. Lack of an engaging Ukrainian AI tutoring solution will lead to the inability to pass on common fables, heritage literature analysis skills, and critical moments familiar to prior generations. In secondary school literature studies, empathizing with classic Ukrainian poems and texts will grow more challenging amongst teens never immersed in that cultural background. Likewise, they will struggle with interpreting symbolism and references common to those eras of Ukrainian identity formation while not receiving any support from Ukrainian-aligned language models for written compositions or humanities projects. Subsequent generations will lose touch with integral pieces of the country's unique heritage story.

Even on an informal level, interest in artistic efforts around theater, cinema, visual arts, and music see declining engagement from younger Ukrainians as preferred leisure activities shift towards globalized media culture rather than celebrating local creators and talent. Despite the current obvious boom of local cultural talent, there is still a huge subset of the population that is dependent on external sources of entertainment, from movies to music (Molfar).

In essence, Ukraine and similar developing countries face looming risk over the next generation, where accumulated erosion across countless tiny dimensions of language diversity and identity lead to forging an entirely different nation - with culture, history, and influence conspicuously drifting into the shadows of a former self, which has been so fiercely fought for.

Such is the steep collective price societies can pay when neglecting "untimely" AI model development efforts in favor of convenience and cost during pivotal transition points in history. This danger is imminent unless there is an immediate increase in urgency to prioritize national languages and invest in critical computing infrastructure for educators and policymakers. The decisions made in the coming five years on prioritization between language-specific and multilingual model availability carry potentially profound societal consequences depending on which vision prevails under the pressures of globalized technology proliferation.

5.3. Risks of underrepresentation

Over the past 15 years, Ukrainian Google and YouTube search queries have become increasingly dominated by Russian language pages and video results (Search Engine Land, 2023). This occurred because Russian internet data grew rapidly early on - amassing orders of magnitude more content, sites, and engagement than the Ukrainian web, alongside the unfortunate post-russification effects of the Soviet era.

As a consequence, Google's algorithms seeking to maximize search intent fulfillment for Ukrainian keyword queries, surfacing Russian pages higher in results because, probabilistically, people's intent gets fulfilled more often there based on aggregate global click behavior.

This creates a self-reinforcing flywheel where Russian sites continue gaining more links, clicks, and search authority compared to Ukrainian community pages on the same topics despite not matching the native language exactly.

Similarly, as large language models for different languages mature — if Russian LLMs accumulate exponentially more parameters, content trained on, and research budget than available Ukrainian models — probabilistic fulfillment of natural language queries and conversational needs from Ukrainian users will skew towards Russian-centric resources. Even if the Ukrainian content exists, it surfaces less prominently. And, gradually, queries normalize towards Russian linguistic structures and dialects if that provides higher collective fulfillment rates globally. This also provides an enormous data feedback loop effect as the applications and model creators are able to generate even more human feedback data on which to improve models.

Without dedicated investment from both public and private sectors in developing models for native languages, we risk cultural erosion. This comes from a reliance on technology that favors more dominant languages, simply because it's more convenient.

This convenience itself opens up an opportunity for another medium of risk, enabling much faster and efficient distribution of propaganda and misinformation, requiring its own unique mechanisms for detection and prevention (Solopova et al., 2023). This is an obvious risk that is becoming critical in the political and existential context for any developing country that is affected by external pressure from other foreign countries.

5.4. Future work, policy, and critical timing

As large language models continue rapidly advancing thanks to unprecedented compute investments by groups like OpenAI, Anthropic, Google, Meta, and Baidu, a clear "model divide" looks poised to emerge.

Hundreds of lower-resource languages globally now stand at risk of accelerating identity erosion without specialized LLM variants representing their linguistic contexts. From Navajo conversational interfaces to Quechua literary analysis tools to Welsh educational content creators — sadly, these languages are falling behind on the rapid advancements in today's technology.

Consequently, many threatened languages pose a digital extinction risk without counterbalancing forces to protect their dialects, artistic traditions, and communities. These groups often struggle due to the lack of institutional support, which results in insufficient access to the necessary data and resources.

As future generations raised on AI inherit even subtle biases favoring better resourced languages, the cultural price to pay will grow exponentially steeper. Preserving heritage hence requires some rebalancing, where policymakers implement commitments to inclusive innovation, perhaps evaluating issues of sustainability for vulnerable groups rather than solely technical tradeoffs.

Companies and governments worldwide must acknowledge that shortsighted stances on optimized efficiency today cascade into seismic identity impacts downstream. Access barriers erode dialects, discourage artistic traditions, and deter descendants from inheriting linguistic lineage — ultimately dimming cultural continuity prospects.

Prioritizing LLM development for lower-resource languages offers a reverse course against irreversible language extinction already accelerating since the turn of the century. As risks become solutions, so do data divides resolve through compassionate actors cooperating across borders to uplift unseen communities, now empowered to share their visions.

6. Conclusions

In this paper, we have explored the importance of developing language-specific large language models (LLMs) for underrepresented languages, focusing on the Ukrainian language as a case study. Our findings demonstrate that cultural bias is a quantifiable phenomenon, and we can speculate about its underlying causes. The open-source community plays a crucial role in addressing this issue by creating new, extended datasets and publishing them for further research work. While this effort may be beyond the scope of commercial interest, it has immense humanitarian impact.

It's important to note the emergence of code-switching effects like Azirivka, which occur spontaneously and highlights the similarities between pattern learning mechanisms in humans and LLMs. While fully recognizing that this intriguing phenomenon warrants a more thorough examination, we contend that even preliminary observations merit reporting. The existence of such effects in human societies, where two languages coexist in close contact, further reinforces the importance of developing language-specific models to preserve cultural identity and linguistic diversity.

To advance the evaluation of language models for Ukrainian, we have introduced **ULIB**, the "Ukrainian Linguistic Inquiry Benchmark." This benchmark encompasses various language processing tasks, including summarization, poem generation, spelling, and simplified explanation comprehension. ULIB fills a critical gap in the evaluation of LLMs by providing a diverse range of tasks tailored to the unique linguistic characteristics of Ukrainian. By offering a holistic evaluation framework, ULIB enables human evaluators to assess the performance of LLMs in understanding and generating Ukrainian text. Although we have only introduced the format and starting point for ULIB datasets, which are available on our github, we plan to expand it as part of our future work.

In addition to ULIB, we have also introduced the Ukrainian Knowledge and Instruction Dataset (**UKID**), a pioneering instruction dataset rooted in Ukrainian context. UKID serves as a comprehensive and standalone source of truth, enhancing the model's ability to verify facts and improve its accuracy. By incorporating language-specific contexts and embedding essential factual knowledge into the model's trainable parameters during fine-tuning, UKID paves the way for more effective and culturally relevant language models.

Our work highlights the significance of developing language-specific LLMs and datasets, not only for Ukrainian but for all underrepresented languages worldwide. By demonstrating the feasibility and importance of this approach, we hope to inspire further research and development in this area. Future work should focus on fine-tuning open-source models with expanded datasets, improving evaluation benchmarks, and exploring innovative applications that leverage the power of language-specific LLMs. Through collaborative efforts between researchers, open-source communities, and stakeholders, we can work towards a future where AI technologies are truly inclusive and representative of the world's linguistic and cultural diversity.

7. Acknowledgements

We would like to acknowledge the support of UkraineNow.org and Nvidia Corporation for providing a DGX Station, a turnkey deskside AI supercomputer with four NVIDIA® Tesla® V100 Tensor Core GPUs, which was used to benchmark the fine-tuned models.

We also thank Tensoic and Microsoft Azure Cloud for providing the compute resources to fine-tune the Mistral-based model.

Additionally, we are grateful to Google LLC for supplying the pre-trained weights for the Gemma models and the fine-tuning infrastructure on the Vertex AI platform, which allowed for easy and quick setup of the fine-tuning and deployment processes within a short timeframe.

References

Israel Abebe Azime, Mitiku Yohannes Fuge, Atnafu Lambebo Tonja, Tadesse Destaw Belay, Aman Kassahun Wassie, Eyasu Shiferaw Jada, Yonas Chanie, Walelign Tewabe Sewunetie, and

Seid Muhie Yimam. 2024. Enhancing amharic-llama: Integrating task specific and generative datasets.

Pierpaolo Basile, Elio Musacchio, Marco Polignano, Lucia Siciliani, Giuseppe Fiameni, and Giovanni Semeraro. 2023. Llamantino: Llama 2 models for effective text generation in italian language.

Y. Bengio, Réjean Ducharme, and Pascal Vincent. 2000. A neural probabilistic language model. volume 3, pages 932–938.

Laada Bilaniuk. 2004. A typology of surzhyk: Mixed ukrainian-russian language. *International Journal of Bilingualism - INT J BILING*, 8:409–425.

Tom B. Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel M. Ziegler, Jeffrey Wu, Clemens Winter, Christopher Hesse, Mark Chen, Eric Sigler, Mateusz Litwin, Scott Gray, Benjamin Chess, Jack Clark, Christopher Berner, Sam McCandlish, Alec Radford, Ilya Sutskever, and Dario Amodei. 2020. Language models are few-shot learners.

Tahiya Chowdhury. 2023. Towards goldilocks zone in child-centered ai. 4 pages.

Junyoung Chung, Caglar Gulcehre, KyungHyun Cho, and Yoshua Bengio. 2014. Empirical evaluation of gated recurrent neural networks on sequence modeling.

Yiming Cui, Ziqing Yang, and Xin Yao. 2024. Efficient and effective text encoding for chinese llama and alpaca.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. Bert: Pre-training of deep bidirectional transformers for language understanding.

Drastic. github.com/drastic/squad-uk.

Jay Gala, Thanmay Jayakumar, Jaavid Aktar Husain, Aswanth Kumar M, Mohammed Safi Ur Rahman Khan, Diptesh Kanojia, Ratish Puduppully, Mitesh M. Khapra, Raj Dabre, Rudra Murthy, and Anoop Kunchukuttan. 2024a. Airavata: Introducing hindi instruction-tuned llm.

Jay Gala, Thanmay Jayakumar, Jaavid Aktar Husain, Aswanth Kumar M, Mohammed Safi Ur Rahman Khan, Diptesh Kanojia, Ratish Puduppully, Mitesh M. Khapra, Raj Dabre, Rudra Murthy, and Anoop Kunchukuttan. 2024b. Airavata: Introducing hindi instruction-tuned llm.

Google DeepMind Gemma Team. 2024. Gemma: Open models based on gemini research and technology. `https://storage.googleapis.com/deepmind-media/gemma/gemma-report.pdf`.

Robin Had. github.com/robinhad/kruk.

Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. 2015. Deep residual learning for image recognition.

Sepp Hochreiter and Jürgen Schmidhuber. 1997. Long short-term memory. *Neural computation*, 9:1735–80.

Edward J Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. 2022. LoRA: Low-rank adaptation of large language models. In *International Conference on Learning Representations*.

Albert Q. Jiang, Alexandre Sablayrolles, Arthur Mensch, Chris Bamford, Devendra Singh Chaplot, Diego de las Casas, Florian Bressand, Gianna Lengyel, Guillaume Lample, Lucile Saulnier, Lélio Renard Lavaud, Marie-Anne Lachaux, Pierre Stock, Teven Le Scao, Thibaut Lavril, Thomas Wang, Timothée Lacroix, and William El Sayed. 2023. Mistral 7b.

Albert Q. Jiang, Alexandre Sablayrolles, Antoine Roux, Arthur Mensch, Blanche Savary, Chris Bamford, Devendra Singh Chaplot, Diego de las Casas, Emma Bou Hanna, Florian Bressand, Gianna Lengyel, Guillaume Bour, Guillaume Lample, Lélio Renard Lavaud, Lucile Saulnier, Marie-Anne Lachaux, Pierre Stock, Sandeep Subramanian, Sophia Yang, Szymon Antoniak, Teven Le Scao, Théophile Gervet, Thibaut Lavril, Thomas Wang, Timothée Lacroix, and William El Sayed. 2024. Mixtral of experts.

Larysa Katerynych, Maksym Veres, and Eduard Safarov. 2021. Transformer-based model for text classification in ukrainian. Taras Shevchenko National University of Kyiv.

Kateryna Kent. 2011. Language contact: Morphosyntactic analysis of surzhyk spoken in central ukraine.

Guneet Singh Kohli, Shantipriya Parida, Sambit Sekhar, Samirit Saha, Nipun B Nair, Parul Agarwal, Sonal Khosla, Kusumlata Patiyal, and Debasish Dhal. 2023. Building a llama2-finetuned llm for odia language utilizing domain knowledge instruction set.

Alex Krizhevsky, Ilya Sutskever, and Geoffrey E Hinton. 2012. Imagenet classification with deep convolutional neural networks. In *Advances in Neu-*

ral Information Processing Systems, volume 25. Curran Associates, Inc.

Yurii Laba, Volodymyr Mudryi, Dmytro Chaplynskyi, Mariana Romanyshyn, and Oles Dobosevych. 2023. Contextual embeddings for ukrainian: A large language model approach to word sense disambiguation. In *Proceedings of the Second Ukrainian Natural Language Processing Workshop*, pages 11–19, Dubrovnik, Croatia.

Ilya Loshchilov and Frank Hutter. 2017. Fixing weight decay regularization in adam. *CoRR*, abs/1711.05101.

Molfar. Song charts analysis. `https://molfar.com/en/blog/song-charts`. Accessed: 2024-04-04.

Quan Nguyen, Huy Pham, and Dung Dao. 2023. Vinallama: Llama-based vietnamese foundation model.

OpenAI. 2023. Gpt-4 technical report.

Alec Radford and Karthik Narasimhan. 2018. Improving language understanding by generative pre-training.

Alec Radford, Jeff Wu, Rewon Child, David Luan, Dario Amodei, and Ilya Sutskever. 2019. Language models are unsupervised multitask learners.

Holger Schwenk, Daniel Dechelotte, and Jean-Luc Gauvain. 2006. Continuous space language models for statistical machine translation. In *Proceedings of the COLING/ACL 2006 Main Conference Poster Sessions*, pages 723–730, Sydney, Australia. Association for Computational Linguistics.

Search Engine Land. 2023. Google and the challenge of russian propaganda in search results. `https://searchengineland.com/google-russian-propaganda-search-results-382229`. Accessed: 2024-04-04.

C. E. Shannon. 1951. Prediction and entropy of printed english. *The Bell System Technical Journal*, 30(1):50–64.

David Silver, Aja Huang, Christopher Maddison, Arthur Guez, Laurent Sifre, George Driessche, Julian Schrittwieser, Ioannis Antonoglou, Veda Panneershelvam, Marc Lanctot, Sander Dieleman, Dominik Grewe, John Nham, Nal Kalchbrenner, Ilya Sutskever, Timothy Lillicrap, Madeleine Leach, Koray Kavukcuoglu, Thore Graepel, and Demis Hassabis. 2016. Mastering the game of go with deep neural networks and tree search. *Nature*, 529:484–489.

Veronika Solopova, Oana-Iuliana Popescu, Christoph Benzmüller, and Tim Landgraf. 2023. Automated multilingual detection of pro-kremlin propaganda in newspapers and telegram posts. *arXiv preprint arXiv:2301.10604*. Available online at arXiv:2301.10604.

Gemini Team. 2023. Gemini: A family of highly capable multimodal models.

Serhii Tiutiunnyk. 2020. Context-based question-answering system for the ukrainian language. Master's thesis, Ukrainian Catholic University, Lviv.

Hugo Touvron, Louis Martin, Kevin Stone, Peter Albert, Amjad Almahairi, Yasmine Babaei, Nikolay Bashlykov, Soumya Batra, Prajjwal Bhargava, Shruti Bhosale, Dan Bikel, Lukas Blecher, Cristian Canton Ferrer, Moya Chen, Guillem Cucurull, David Esiobu, Jude Fernandes, Jeremy Fu, Wenyin Fu, Brian Fuller, Cynthia Gao, Vedanuj Goswami, Naman Goyal, Anthony Hartshorn, Saghar Hosseini, Rui Hou, Hakan Inan, Marcin Kardas, Viktor Kerkez, Madian Khabsa, Isabel Kloumann, Artem Korenev, Punit Singh Koura, Marie-Anne Lachaux, Thibaut Lavril, Jenya Lee, Diana Liskovich, Yinghai Lu, Yuning Mao, Xavier Martinet, Todor Mihaylov, Pushkar Mishra, Igor Molybog, Yixin Nie, Andrew Poulton, Jeremy Reizenstein, Rashi Rungta, Kalyan Saladi, Alan Schelten, Ruan Silva, Eric Michael Smith, Ranjan Subramanian, Xiaoqing Ellen Tan, Binh Tang, Ross Taylor, Adina Williams, Jian Xiang Kuan, Puxin Xu, Zheng Yan, Iliyan Zarov, Yuchen Zhang, Angela Fan, Melanie Kambadur, Sharan Narang, Aurelien Rodriguez, Robert Stojnic, Sergey Edunov, and Thomas Scialom. 2023. Llama 2: Open foundation and fine-tuned chat models.

Bram Vanroy. 2023. Language resources for dutch large language modelling.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. *CoRR*, abs/1706.03762.

Wikipedia. Azirivka — wikipedia, the free encyclopedia.

Genta Indra Winata, Samuel Cahyawijaya, Zihan Liu, Zhaojiang Lin, Andrea Madotto, and Pascale Fung. 2021. Are multilingual models effective in code-switching? *CoRR*, abs/2103.13309.

Ruochen Zhang, Samuel Cahyawijaya, Jan Christian Blaise Cruz, Genta Indra Winata, and Alham Fikri Aji. 2023. Multilingual large language models are not (yet) code-switchers.

Chunting Zhou, Pengfei Liu, Puxin Xu, Srini Iyer, Jiao Sun, Yuning Mao, Xuezhe Ma, Avia Efrat, Ping Yu, Lili Yu, Susan Zhang, Gargi Ghosh, Mike Lewis, Luke Zettlemoyer, and Omer Levy. 2023. Lima: Less is more for alignment.

Spivavtor: An Instruction Tuned Ukrainian Text Editing Model

Aman Saini, Artem Chernodub, Vipul Raheja, Vivek Kulkarni
Grammarly
firstname.lastname@grammarly.com

We introduce Spivavtor, a dataset, and instruction-tuned models for text editing focused on the Ukrainian language. Spivavtor is the Ukrainian-focused adaptation of the English-only CoEdIT (Raheja et al., 2023) model. Similar to CoEdIT, Spivavtor performs text editing tasks by following instructions in Ukrainian like "Виправте граматику в цьому реченні" and "Спростіть це речення" which translate to "Correct the grammar in this sentence" and "Simplify this sentence" respectively. This paper describes the details of the Spivavtor-Instruct dataset and Spivavtor models. We evaluate Spivavtor on a variety of text editing tasks in Ukrainian, such as Grammatical Error Correction (GEC), Text Simplification, Coherence, and Paraphrasing, and demonstrate its superior performance on all of them. We publicly release our best-performing models and data as resources to the community to advance further research in this space.

Keywords: Ukrainian Text Editing, Instruction tuned LLMs

1. Introduction

Recently, there has been an increased focus and substantial progress in developing natural language processing (NLP) models for the Ukrainian language. These include the development of corpora like the Ukrainian Brown Corpus (Starko and Rysin, 2023), toolkits like NLP-UK[1], as well as models for word-embeddings, part-of-speech tagging, named entity recognition[2], machine translation[3], and pre-trained language models.

However, many of the aforementioned models are task-specific and do not leverage recent advances in large-scale language models and in-context learning. In particular, Large Language Models (LLMs) have demonstrated a quantum leap in generating fluent and coherent text and have thus been adapted to perform various NLP tasks (Brown et al., 2020; OpenAI, 2023; Touvron et al., 2023). Recently, LLMs have also been leveraged to perform specific text editing tasks such as GEC (Wu et al., 2023; Coyne and Sakaguchi, 2023; Fang et al., 2023) or Text Simplification (Baez and Saggion, 2023; Saggion et al., 2022). In addition to focusing on singular tasks like the above, some works have attempted to leverage LLMs and instruction fine-tuning to support multi-task text editing (Schick et al., 2023; Raheja et al., 2023; Laban et al., 2023).

Grammatical Error Correction

Виправте граматику в цьому реченні: А ти, батюшка, стало бути, тут в сторожі?

А ти, батюшко, стало бути, тут у сторожі?

Coherence

Покращіть зв’язність тексту: Тоді він відвідав острови Греції. Хвороба змусила його повернутися до Риму.

Потім він відвідав острови Греції, але хвороба змусила його повернутися до Риму.

Text Simplification

Спростіть речення: вода спочатку холодна, а потім волога.

Вода і холодна, і мокра.

Paraphrasing

Перефразуйте речення: Наскільки я можу судити, обидва ще живі.

Наскільки я розумію, вони обидва ще живі.

Figure 1: Example input (blue) and output (red) of the text editing tasks that Spivavtor can perform. The corresponding English translations can be found in Appendix A, Table 7.

There has also been extensive work on leveraging these advances to develop corresponding LLMs focused on the Ukrainian language, the most notable being UAlpaca[4], which builds a Ukrainian counterpart of the popular general-purpose instruction-tuned model – Alpaca (Taori et al., 2023). Concurrently and similarly, some research has focused on building and characterizing the capabilities of multi-lingual LLMs which are trained on massively multi-lingual data during the pre-training and instruction-tuning phases (Muennighoff et al., 2023; Workshop, 2023; Xue et al., 2021; Li et al., 2023; Wei et al., 2023; Üstün et al., 2024). While these models support instructions in Ukrainian, they do not focus on high-quality text editing tasks but on

[1] https://github.com/brown-uk/nlp_uk

[2] https://huggingface.co/lang-uk

[3] https://github.com/Helsinki-NLP/UkrainianLT

[4] https://huggingface.co/robinhad/ualpaca-7b-llama

© 2024 ELRA Language Resource Association: CC BY-NC 4.0

general-purpose instructions instead, such as sentiment detection, question answering, text generation, etc. However, as noted by Raheja et al. (2024), such generic instruction-tuned models are not particularly well-suited for nuanced text editing tasks without further task-specific fine-tuning. This highlights the need for an instruction-tuned model for Ukrainian that is optimized for text editing, which this paper addresses by building Spivavtor[5].

Spivavtor can follow instructions for complex text editing tasks like GEC, Text Simplification, Coherence, and Paraphrasing (Figure 1). A significant challenge to building an instruction-tuned model for Ukrainian optimized for text editing has been the limited availability of text editing datasets in Ukrainian. In this work, we address this challenge by adapting existing text editing datasets from Ukrainian and English and converting them to ”instruction-following” datasets (similar to CoEdIT and mEdIT). We then show how these newly constructed datasets can be used to build state-of-the-art text editing models for Ukrainian. Finally, through comprehensive evaluations, we empirically reveal critical insights on how the performance on Ukrainian text editing tasks is affected by various choices like model architecture, model scale, and training data mixtures. All our models and data are publicly available as resources for the community[6].

2. Related Work

Prior work falls into two major categories: (a) Ukrainian-NLP Models and (b) Multi-lingual LLMs. We discuss each of these below.

Large Language Models for Ukrainian Several works have focused on building LLMs and resources for Ukrainian. These mainly consist of manually curated Ukrainian language datasets and corpora like Starko and Rysin (2023) for Part of Speech, Syvokon et al. (2023) for Grammatical Error Correction (GEC), NER-UK for Named Entity Recognition[7], UA-SQUAD for Question Answering Ivanyuk-Skulskiy et al. (2021). Some Ukrainian datasets are also derived from large multi-lingual datasets filtered for the Ukrainian language data (for e.g., Ukrainian Tweet Corpus[8]). In addition to these datasets, custom models have also been built for the above tasks, a list of which is curated here[9]. A notable such model aimed at general instruction following in Ukrainian is the UAlpaca model, which was obtained by further fine-tuning LLaMA on Ukrainian translations of the Alpaca (Taori et al., 2023) dataset.

Text Editing via Instruction Tuning There exists extensive prior literature leveraging instruction-tuned LLMs for various text editing tasks in both monolingual and multi-lingual settings. More recently, Schick et al. (2023), Raheja et al. (2023), and Laban et al. (2023) have focused on general-purpose text editing using instruction-tuned LLMs for English. However, all of these prior approaches have been limited in monolingual settings because they focus only on English.

Text editing capabilities in the Ukrainian language have been developed only in multi-lingual settings, where most works have proposed task-specific multi-lingual models. These works have developed models for text editing tasks like GEC (Rothe et al. (2021); Sun et al. (2022)), paraphrasing (Chowdhury et al., 2022), formality style transfer (Briakou et al., 2021), and text simplification (Mallinson et al. (2020); Martin et al. (2022); Ryan et al. (2023)). However, they are similarly limited due to their singular focus on specific text editing tasks rather than high-quality, general-purpose text editing.

There exists an even more extensive literature on general-purpose multi-lingual LLMs (many of which also include support for Ukrainian (Üstün et al., 2024; Li et al., 2023)), these models generally aim for massive multi-language support and are not optimized explicitly for Ukrainian or text editing. A comprehensive review of multi-lingual LLMs is out of the scope of this paper.

Finally, our work is closest to the recently proposed mEdIT (Raheja et al., 2024), which developed a multi-lingual extension to CoEdIT with support for a similar set of tasks for six languages, but is limited in our context as it is not focused on Ukrainian as one of its core languages.

3. Spivavtor

In this section, we describe the construction of Spivavtor. Specifically, we discuss (a) Dataset construction, (b) Model architecture choices, and (c) Model training process.

3.1. Spivavtor-Instruct Dataset

Similar to prior work (Raheja et al., 2023), we consider four text editing tasks: (a) Fluency/Grammatical Error Correction (GEC),

[5] Spivavtor means “co-author” in Ukrainian.

[6] https://huggingface.co/collections/grammarly/spivavtor-660744ab14fdf5e925592dc7

[7] https://github.com/lang-uk/ner-uk

[8] https://github.com/saganoren/ukr-twi-corpus

[9] https://github.com/osyvokon/awesome-ukrainian-nlp

(b) Simplification, (c) Coherence, and (d) Paraphrasing; and construct a unified Ukrainian text editing instruction dataset which we call Spivavtor-Instruct. We consider these tasks for two reasons: (a) These tasks are largely representative of the most common text editing tasks, and (b) It is feasible to obtain curated good-quality data for these tasks either in Ukrainian or English. For tasks where Ukrainian data is not readily available, we use the available English datasets to construct their Ukrainian counterpart by translating them into Ukrainian using Google Translate API[10]. Due to time constraints, we did not explore other translation services or models. Having outlined the tasks, we now discuss the task-specific datasets we used and our process for constructing Spivavtor-Instruct.

GEC We use the Ukrainian Grammatical Error Correction (UA-GEC) dataset (Syvokon et al., 2023) for GEC/Fluency. This dataset contains 33k pairs of grammatically incorrect and correct sentences in Ukrainian. The original dataset contains train (31k) and test (2k) splits. However, since we explore different model choices and training hyper-parameters, we further randomly split the train set to create a custom train (28k) and validation (3k) dataset.

Simplification For the Simplification task, we adapt three English datasets: (a) WikiLarge (Zhang and Lapata, 2017), and (b) WikiAuto (Jiang et al., 2020) for training. For evaluation, we use ASSET (Alva-Manchego et al., 2020), and Turk (Xu et al., 2016a) datasets. As mentioned above, we translate all these datasets into Ukrainian using Google Cloud Translation API.

Coherence For the coherence task, which involves combining two sentences together coherently using edit operations such as inserting discourse connectives, we once again translate an English dataset, given the lack of an equivalent dataset for Ukrainian. In particular, we adapt the DiscoFuse dataset (Geva et al., 2019) and the Coherence split of IteraTeR (Du et al., 2022) and translate them to Ukrainian using the Google Cloud Translation API.

Paraphrasing We adapt the popular PAWS (Zhang et al., 2019) dataset in English by constructing its Ukrainian counterpart via translation, maintaining their train and test splits. We evaluate paraphrasing on MRPC (Dolan and Brockett, 2005), STS (Cer et al., 2017), and QQP datasets.

The Ukrainian datasets we thus obtain are suitable for training Ukrainian-specific models, but they are not suitable yet for instruction tuning since they do not contain explicit instructions. To overcome this, we prepend task-specific verbalizers that describe the task to be performed as simple instructions to each instance. These task-specific verbalizers were curated by domain experts in Ukrainian. More specifically, for a given task-specific instance, we assign a specific verbalizer by randomly drawing a sample from the task-specific verbalizer set. Table 2 shows a few instruction verbalizers for each task with the full set available in Appendix Table 9. Similarly, Table 1 summarizes the number of training, validation, and test instances, along with the number of distinct instructions per task. Finally, it is to be noted that to ascertain the quality of the Ukrainian translated datasets, a random sample of 100 instances were chosen for verification by native speakers of Ukrainian and found to be largely satisfactory[11].

3.2. Models

To train Spivavtor, we consider two kinds of transformer-based LLM architectures – Encoder-Decoder as well as the Decoder-only architecture. Both architectures have been shown to be generally effective in prior work (Xue et al., 2021; Üstün et al., 2024) although the Decoder-only models tend to be more popular recently with the release of models like ChatGPT and GPT4 (OpenAI, 2023). Thus, in the realm of Ukrainian text editing, we empirically explore the effect of both of these model architectures on task performance. We also explore the effect of different model sizes considering relatively smaller models with 1B parameters as well as larger models with upto 13B parameters.

3.2.1. Encoder-Decoder Models

mT5 (Xue et al., 2021) is a multi-lingual variant of T5 (Raffel et al., 2020), trained on the mC4 dataset [12], a multi-lingual variant of the C4 dataset extended to 101 languages. We experiment with two variants of mT5 – large (1.2B) and xxl (13B).

mT0 (Muennighoff et al., 2023) is a family of multi-lingual Encoder-Decoder models capable of following human instructions in dozens of languages. We use the mt0-large (1.2B) model. The mT0 models are constructed by fine-tuning mT5 models on the xP3 cross-lingual task mixture

[10] https://cloud.google.com/translate/docs/advanced/translating-text-v3

[11] Grossly incorrect translations were corrected manually.

[12] https://huggingface.co/datasets/mc4

Task	#Train	#Validation	#Test	#Verbalizers
GEC	27,929	3,103	2,682	9
Simplification	11,501	1,278	533	11
Coherence	9,278	1,031	551	7
Paraphrasing	14,076	1,564	6,244	13
Total	62,784	6,976	10,010	40

Table 1: Summary statistics of the Spivavtor-Instruct dataset.

Task	Verbalizers	English translation
GEC	“Виправте граматику в цьому реченні:” “Зробіть речення граматичним:” “Удоскональте граматику цього тексту:”	“Correct the grammar in this sentence:” “Make the sentences grammatical:” “Improve the grammar of this text:”
Simplification	“Спростіть речення:” “Зробіть речення простим:” “Зробіть цей текст легше для розуміння:”	“Simplify the sentences:” “Make the sentence simple:” “Make this text easier to understand:”
Coherence	“Виправте зв’язність в реченні:” “Покращіть зв’язність тексту:” “Зробіть текст більш зв’язним:”	“Correct the coherence in the sentence:” “Improve text coherence:” “Make the text more coherent:”
Paraphrasing	“Перефразуйте речення:” “Перефразуйте цей текст:” “Напишіть перефраз для речення:”	“Rephrase the sentence:” “Paraphrase this text:” “Write a paraphrase for the sentence:”

Table 2: A subset of verbalizers for each task used as instructions in the Spivavtor-Instruct dataset (see Appendix Table 9 for full set of instructions).

dataset, which consists of multi-lingual datasets with English prompts. As a result, mT0 models are better suited for following English prompts. We also use the mt0-xxl-mt variant, which is fine-tuned on the xP3mt dataset and is better suited for prompting in non-English.

Aya 101 (Üstün et al., 2024) is a massively multilingual generative language model that follows instructions in 101 languages of which over 50% are considered low-resourced. Aya outperforms mT0 and BLOOMZ (Muennighoff et al., 2022) on the majority of tasks while covering double the number of languages. The model has 13B parameters and the same architecture as the mt5-xxl model.

3.2.2. Decoder-only LLMs

Bactrian-X (Li et al., 2023) is a collection of lightweight adapters for LLaMA (7B and 13B) (Touvron et al., 2023) and BLOOM (7B) (Workshop, 2023) on the Bactrian-X dataset, which is a multi-lingual parallel dataset containing 3.4 million instruction–response pairs across 52 languages. We use the bactrian-x-llama-7b-merged variant.

Mistral (Jiang et al., 2023) is a family of large language models. We use the Mistral-7B-Instruct-v0.2 variant which is an instruction fine-tuned version of the Mistral-7B-v0.2 model.

Llama2 Chat Models We also consider full-parameter fine-tuning of the Llama2 7B and 13B chat models. While the aforementioned Bactrian-X models also derive from the LLaMA models, they use parameter-efficient fine-tuning (PEFT), specifically, low-rank adaptation (LoRA) (Hu et al., 2022), thus, significantly reducing the number of trainable parameters during fine-tuning. Thus, in contrast to Bactrian-X models, we consider full-parameter fine-tuning of Llama-2 Chat models as well. We use the Llama-2-7b-chat-hf and Llama-2-13b-chat-hf variants.

3.3. Training

We use Spivavtor-Instruct dataset to perform instruction-tuning on both styles of models described above. We train all models using Deepspeed (Rasley et al., 2020) on 8xA100 GPU instances with AdamW optimizer, a per-device batch size of 8, and a learning rate of 5e-5. For Decoder-only models, the maximum sequence length is set to 512 tokens, whereas for Encoder-Decoder models, the maximum sequence length is set to 256 tokens for both source and target. The best-performing checkpoints were chosen based on the validation loss.

3.4. Inference

For Inference, we mostly use default generation parameters for temperature, beam size as specified in the corresponding model with the exception of max output length, which is set to the max sequence length used while training the model. To avoid repeated generation with Decoder-only models, we used the model-specific EOS tag to end decoding.

4. Evaluation

Metrics We evaluate all models on the task-specific test splits of the Spivavtor-Instruct dataset. As in prior work, we report the standard evaluation metrics used for each task. In particular, we report the $F_{0.5}$ Correction score for GEC calculated using ERRANT (Bryant et al., 2017) weighing precision twice as much as recall. Following prior work by Ryan et al. (2023); Raheja et al. (2023) we report SARI (Xu et al., 2016b) for Simplification as well as Coherence. For Paraphrasing, we report BLEU (Papineni et al., 2002). In order to capture the overlap with source as well as reference, we report both reference-free BLEU (also called Self-BLEU in Zhu et al. 2018) and reference-based BLEU, since they collectively provide additional signal on paraphrasing quality than either one of them alone (see Shen et al. (2022) and Section 6).

Baselines We evaluate our Spivavtor models against strong instruction tuned baseline models. In addition to the corresponding base models (i.e. not fine-tuned on Spivavtor-Instruct dataset), we also evaluate against the following:

- **Copy**: The Copy baseline, which just copies the input sentence, is a surprisingly trivial but hard-to-beat baseline.

- **UAlpaca**: To ascertain the effect of task-specific instruction fine-tuning in contrast to large-scale diverse instruction fine-tuning, we consider the UAlpaca model in a zero-shot setting. UAlpaca is a LLaMA 7B model trained on Ukrainian translations of 52K diverse and generic instructions of the Alpaca dataset (Taori et al., 2023). For prompting UAlpaca, we used the recommended prompt format that it was fine-tuned on and replaced the instruction placeholder with the assigned verbalizer.

- **GPT4** Noting the widespread popularity of GPT4 (OpenAI, 2023) and a general notion that GPT4 generally obtains very strong performance on many NLP tasks, we also consider this as a baseline (in the zero-shot setting) where we prompt it with a verbalizer and the input text. In particular, we use gpt-4-0613 model with a context window of 8192 tokens and a training data cutoff of Sep 2021. To give GPT4 the best shot at success and to account for prompt sensitivity, we evaluate GPT4 on the chosen task with all possible verbalizers in our set and report the score corresponding to the best verbalizer. If there is no response received from the API due to content filtration policies, we consider the input unchanged for evaluation purposes.

- **GPT-3.5-Turbo** We also compare against the more cost effective GPT-3.5-Turbo model, widely known as ChatGPT. In particular, we use gpt-3.5-turbo version 0301.

4.1. Quantitative Results

In this section, we describe our main results and discuss findings from ablation studies to gain insights into the factors driving model performance.

Main Results Table 3 shows the performance of various models on all tasks in consideration. It presents aggregated scores for all tasks across different datasets. The dataset-specific scores for all relevant tasks are present in Appendix A, Table 8. Based on these results, we can make the following observations:

1. **Spivavtor generally performs significantly better over baselines**. Comparing the performance metrics for Spivavtor models to their baseline counterparts, we generally observe that Spivavtor significantly outperforms baseline models (including GPT4), with Simplification being the only exception where performance is at par. This result suggests the effectiveness of domain-specific instruction tuning for superior performance on specific tasks.

2. **Domain-specific Instruction tuning outperforms instruction tuning on a large set of generic instructions.** Given the effectiveness of instruction tuning and in-context learning, a natural question arises: For text editing with instructions, is it sufficient to instruction-tune a model with a very large set of diverse instructions that are not necessarily related to text editing? We can answer this question empirically by comparing the performance of Spivavtor models (that are instruction tuned on text

Model	Type	Size	GEC	Simplification	Coherence	Paraphrasing
Copy	-	-	0	21.98	26.89	100/31.4
Bactrian-X-7b	D	7B	0.65	36.76	40.37	21.86/8.13
UAlpaca-7b	D	7B	0.57	35.17	32.64	13.26/4.95
Mistral-7b	D	7B	0.3	38.96	32.41	9.30/3.79
mt0-large	ED	1.2B	0.21	29.56	22.14	6.70/2.68
aya-101	ED	13B	21.98	35.59	38.30	42.68/15.53
GPT-3.5-Turbo	D	-	1.17	40.18	44.93	26.60/12.51
GPT4	D	-	27.18	40.08	43.44	23.23/11.7
Spivavtor-Bactrian-X-7b	D	7B	55.73	36.90	47.80	65.31/23.65
Spivavtor-Mistral-7b	D	7B	51.54	34.55	44.12	76.56/25.33
Spivavtor-Llama2-7b	D	7B	55.88	36.94	48.73	48.97/18.9
Spivavtor-Llama2-13b	D	13B	56.48	36.98	48.55	57.31/21.35
Spivavtor-mt5-large	ED	1.2B	61.83	36.40	48.27	77.31/26.68
Spivavtor-mt0-large	ED	1.2B	61.44	36.16	48.28	77.83/26.73
Spivavtor-mt5-xxl	ED	13B	63.00	37.84	48.97	72.42/25.64
Spivavtor-mt0-xxl-mt	ED	13B	64.55	**38.44**	**49.48**	68.63/25.07
Spivavtor-aya-101	ED	13B	**64.57**	37.87	48.51	73.28/26.17

Table 3: Comparison of Spivavtor models against various baselines including Copy (target=source), Decoder-only(D) and Encoder-Decoder(ED) models when evaluated in a zero-shot setting. For GEC, we report **$F_{0.5}$ Correction**. For Simplification and Coherence, we report **SARI**. For Paraphrasing, we report **ref-free/ref-based BLEU** where ref-free is the reference-free BLEU and ref-based is the reference-based BLEU to capture the overlap with both source and reference. All scores have been scaled to lie between 0 and 100. Note that all Spivavtor models outperform baseline models.

Held-Out Task	GEC	Simplification	Coherence	Paraphrasing
GEC	**18.47**	37.41	52.11	71.44/26.14
Simplification	64.95	**32.84**	48.96	68.39/25.01
Coherence	62.57	36.79	**39.48**	72.86/25.81
Paraphrasing	64.25	36.86	51.84	**74.61/25.90**

Table 4: Performance of the Spivavtor-aya-101 model on all tasks when one task is ablated. We report the same metrics as in Table 3. The bolded numbers represent the zero-shot performance of the model when not trained on that particular task.

editing instructions) with UAlpaca – a model that is instruction tuned on 52K diverse instructions. From Table 3, we observe that UAlpaca has significantly lower performance compared to its equivalent Spivavtor model (Spivavtor-Llama2-7B). It may not be sufficient to instruction-tune models on just a large set of diverse instructions, and there is significant value to instruction tuning on domain-specific instructions, an observation that reaffirms findings in prior work by Raheja et al. (2023).

3. **Encoder-Decoder models outperform Decoder-only models.** Given the extensive popularity of LLMs, there has been a significant surge in the availability of LLMs. While some LLMs use an Encoder-Decoder architecture (Xue et al., 2021; Üstün et al., 2024), some others use a Decoder-only style (OpenAI, 2023; Taori et al., 2023; Touvron et al., 2023). Yet, it is not clear if one architecture offers consistently superior performance over the other and on what tasks one might prefer a specific architecture. We trained both styles of models on the Spivavtor-Instruct dataset to evaluate the results empirically. Our results indicate that Encoder-Decoder models generally outperform Decoder-only models when fine-tuned on domain-specific instructions. More specifically, note that all Spivavtor Encoder-Decoder models outperform Spivavtor Decoder-only models on average.

4. **Larger models outperform smaller ones.** Our results also suggest that, generally, larger models tend to perform better than smaller ones - both across baselines and across Spivavtor models within an architecture

family. This finding further reinforces the effectiveness of model scaling on task performance.

Task Ablation In this setting, we hold out specific tasks in a controlled manner to evaluate one of the Spivavtor models (Spivavtor-aya-101), to see how it might generalize to unseen text editing tasks. More specifically, in each turn, we hold out one of the tasks, train on the remaining set, and report task performance on all tasks. The results of this ablation study are shown in Table 4 and clearly demonstrate the usefulness of instruction tuning on all tasks. The model performs significantly better when trained on task-specific data as compared to the zero-shot setting.

4.2. Qualitative Error Analysis

In this section, we first discuss the subpar performance of most baseline models on GEC, as observed in Table 3. Careful inspection of the model outputs indicates several problems with zero-shot model evaluation. The most frequent problems include repeated generation, output generation in English instead of Ukrainian, explanation of corrections made, text generation indicating no change is needed, to name a few. These models also suffer from an overcorrecting issue (Fang et al., 2023) and tend to perform paraphrasing and fluency rewrites. As a result, in many cases, the conservative span-based $F_{0.5}$ metric (used to evaluate GEC) can't capture the correct edits, resulting in low performance.

Next, we evaluate one of our best-performing models (Spivavtor-aya-101) qualitatively. For each task, we provide the model a sample input along with an instruction on what to do and show the model-generated output for a handful of such inputs in Table 5. We also highlight some of the errors made by our model in Table 6. The English translations for all examples are provided in the same tables for reference. On the GEC task, the output quality outperforms all baseline models. Due to instruction tuning, the edits become more conservative and therefore, are better captured by $F_{0.5}$ metric using M2scorer[13]. The instruction-tuned models avoid common errors such as repetitions and generation of gibberish text and are much better at following instructions. However, the edits made are not always correct. For the simplification task, the majority of errors arise from changes in meaning due to excessive text truncation. Another typical negative pattern is the filtration of named entities and/or their replacement with pronouns. The coherence task is performed rather successfully. The model either edits the text correctly or leaves the text uncorrected. The most common issue is the incorrect usage of conjunctions, disrupting the logical flow, e.g. using "but" instead of "and", "however" instead of "so", etc. Paraphrasing is done mainly on the lexical level by changing the word or phrase order inside the text. In longer texts, such as those in the MRPC dataset, we sometimes observe a change in meaning compared to the input, whereas in shorter texts, such as those in STS and QQP, it tends to align more closely with the reference rewrites. Errors highlighting some of these problems are shown in Table 6.

5. Conclusions

We introduce Spivavtor – an instruction-tuned LLM for Ukrainian text editing and corresponding Spivavtor-Instruct dataset. We describe in detail the construction of Spivavtor, including how we curate the instruction dataset in Ukrainian for text editing tasks. We empirically show that Spivavtor significantly outperforms other models on text editing tasks. We also analyze the effect of modeling choices (scale and architecture) on task performance. Overall, our experiments support the hypothesis that domain and task-specific instruction tuning is needed to obtain better performance on complex text editing tasks. Finally, all our datasets and models are released to the community to help advance research in the area of Ukrainian NLP.

6. Limitations

While we have introduced an instruction-tuned LLM for Ukrainian text editing, we acknowledge a few limitations of our work. First, due to the limitations of the translation API used, our training data may not be of the highest quality. This limitation could potentially be overcome by curating high-quality data from native speakers of the Ukrainian language. The scale of our training data can also be increased over our current set.

Second, while we use standard evaluation metrics for the text editing tasks, we acknowledge that many of these metrics have limitations and do not capture many aspects of text quality (e.g. meaning preservation, etc.). For example, in the case of evaluating paraphrasing, there is no single automatic metric that holistically captures all important aspects of a good paraphrase as judged by humans. In fact, even if one narrows down to using BLEU score as a metric, it has been shown in prior work that either one of reference-free BLEU score or reference-based BLEU score may correlate better with human judgments, and this may be dataset or benchmark-dependent (Shen et al., 2022) which is why we report both

[13] https://github.com/nusnlp/m2scorer

GEC Input ▶ Виправте граматику в цьому реченні: Дякую за інформацію! ми з Надією саме вийшли з дому **Output** ▶ Дякую за інформацію! Ми з Надією саме вийшли з дому.	**GEC Input** ▶ Correct the grammar in this sentence: Thanks for the information! we with Nadia just left the house **Output** ▶ Thanks for the info! Nadia and I just left the house.
Simplification Input ▶ Спростіть речення: Там він помер через шість тижнів, 13 січня 888 року. **Output** ▶ Там він помер 13 січня 888 року.	**Simplification Input** ▶ Simplify the sentence: There he died six weeks later, on January 13, 888. **Output** ▶ There he died on January 13, 888.
Coherence Input ▶ Виправте зв'язність в реченні: Лінч досі відмовляється розповідати про сумнозвісну травневу дорожньо-транспортну пригоду, коли він збив жінку-пішохода в районі нічного клубу Баффало та відлетів. Той факт, що Лінч взагалі говорив, заслуговує на увагу в цих краях. **Output** ▶ Лінч досі відмовляється розповідати про сумнозвісну травневу дорожньо-транспортну пригоду, коли він збив жінку-пішохода в районі нічного клубу Баффало та відлетів. Однак той факт, що він взагалі говорив, заслуговує на увагу в цих краях.	**Coherence Input** ▶ Correct the coherence in the sentence: Lynch still refuses to talk about the infamous May traffic accident in which he struck a female pedestrian in a Buffalo nightclub area and drove away. The fact that Lynch spoke at all deserves attention in this place. **Output** ▶ Lynch still refuses to talk about the infamous May traffic accident in which he struck a female pedestrian in a Buffalo nightclub area and drove away. However, the fact that Lynch spoke at all deserves attention in this place.
Paraphrasing Input ▶ Перефразуйте речення: Який найкращий комплімент, який ти отримував від будь-кого? **Output** ▶ Який найкращий комплімент, який ти коли-небудь отримував?	**Paraphrasing Input** ▶ Paraphrase the sentence: What is the greatest compliment that you ever received from anyone? **Output** ▶ What is the greatest compliment that you ever received?

Table 5: Example inputs and outputs from Spivavtor-aya-101 model for all relevant tasks.

GEC Input ▶ Виправте граматичні помилки в цьому реченні: В поки що вересень будем повну оплату робити. **Output** ▶ У поки що вересні будем повну оплату робити.	**GEC Input** ▶ Correct the grammatical errors in this sentence: On the meantime in September, will we make the full payment. **Output** ▶ In the meantime in September, will we make the full payment.
Simplification Input ▶ Зробіть речення простим: Джидда є головними воротами до Мекки, найсвятішого міста ісламу, яке працездатні мусульмани повинні відвідати принаймні раз у житті. **Output** ▶ Це одне з головних воріт до Мекки, яке мусульмани повинні відвідати принаймні раз у житті.	**Simplification Input** ▶ Make the sentence simple: Jeddah is the main gateway to Mecca, Islam's holiest city, which able-bodied Muslims must visit at least once in their lifetime. **Output** ▶ It is one of the main gateways to Mecca that Muslims must visit at least once in their lifetime.
Coherence Input ▶ Виправте зв'язність в цьому тексті: Зайферт: Ця зміна здавалася певною протягом більшої частини року. Нещодавно Гуделл сказав, що очікує голосування під час зборів власників ліги в березні. **Output** ▶ Зайферт: Ця зміна здавалася певною протягом більшої частини року, але нещодавно Гуделл сказав, що очікує голосування під час зборів власників ліги в березні.	**Coherence Input** ▶ Correct the coherence in this text: Seifert: This change seemed certain for most of the year. Goodell recently said he expects a vote at the league's owners meeting in March. **Output** ▶ Seifert: That change seemed certain for most of the year, but Goodell recently said he expects a vote at the league's owners meeting in March.
Paraphrasing Input ▶ Перефразуйте це речення: Чоловік грає на музичній клавіатурі. **Output** ▶ Чоловік грає на клавіатурі.	**Paraphrasing Input** ▶ Rephrase this sentence: The man is playing the musical keyboard. **Output** ▶ A man plays the keyboard.

Table 6: Example errors made by Spivavtor-aya-101 model for all tasks with English translations.

reference-based and reference-free BLEU scores in our evaluations for paraphrasing. In addition to BLEU, one would also report a semantic similarity score (like BERTScore) between the paraphrase and the source to capture how semantically close the paraphrase is to the source (or reference). For English, this is typically done using popular sentence embedding models like BERT, but it is not clear what the best approach is for Ukrainian, which is why we do not consider this dimension in our evaluation. One could potentially address such limitations by directly seeking human judgments on the quality of model predictions.

Finally, while we explore different settings of hyper-parameters (like batch size and learning rate) and different variants of prompts in our experiments, our search space is not exhaustive and is limited due to computational budgets and time constraints. We also acknowledge that the performance of closed models like GPT4 may drift or change over time due to model refreshes. Even in cases where model artifacts were publicly available, one must acknowledge that they were likely pre-trained on different datasets in the pretraining stage, and the precise effect of this on our specific downstream task performance is not known and is absorbed in our model performance reports. Research around an improved characterization of such variance in expected performance would be useful in the future.

7. References

Fernando Alva-Manchego, Louis Martin, Antoine Bordes, Carolina Scarton, Benoît Sagot, and Lucia Specia. 2020. ASSET: A dataset for tuning and evaluation of sentence simplification models with multiple rewriting transformations. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 4668–4679, Online. Association for Computational Linguistics.

Anthony Baez and Horacio Saggion. 2023. LSLlama: Fine-tuned LLaMA for lexical simplification. In *Proceedings of the Second Workshop on Text Simplification, Accessibility and Readability*, pages 102–108, Varna, Bulgaria. INCOMA Ltd., Shoumen, Bulgaria.

Eleftheria Briakou, Di Lu, Ke Zhang, and Joel Tetreault. 2021. Olá, bonjour, salve! XFORMAL: A benchmark for multilingual formality style transfer. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 3199–3216, Online. Association for Computational Linguistics.

Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, et al. 2020. Language models are few-shot learners. *Advances in neural information processing systems*, 33:1877–1901.

Christopher Bryant, Mariano Felice, and Ted Briscoe. 2017. Automatic annotation and evaluation of error types for grammatical error correction. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 793–805, Vancouver, Canada. Association for Computational Linguistics.

Daniel Cer, Mona Diab, Eneko Agirre, Iñigo Lopez-Gazpio, and Lucia Specia. 2017. SemEval-2017 task 1: Semantic textual similarity multilingual and crosslingual focused evaluation. In *Proceedings of the 11th International Workshop on Semantic Evaluation (SemEval-2017)*, pages 1–14, Vancouver, Canada. Association for Computational Linguistics.

Jishnu Ray Chowdhury, Yong Zhuang, and Shuyi Wang. 2022. Novelty controlled paraphrase generation with retrieval augmented conditional prompt tuning. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 36, pages 10535–10544.

Steven Coyne and Keisuke Sakaguchi. 2023. An analysis of gpt-3's performance in grammatical error correction. *arXiv preprint arXiv:2303.14342*.

William B. Dolan and Chris Brockett. 2005. Automatically constructing a corpus of sentential paraphrases. In *Proceedings of the Third International Workshop on Paraphrasing (IWP2005)*.

Wanyu Du, Vipul Raheja, Dhruv Kumar, Zae Myung Kim, Melissa Lopez, and Dongyeop Kang. 2022. Understanding iterative revision from human-written text. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 3573–3590, Dublin, Ireland. Association for Computational Linguistics.

Tao Fang, Shu Yang, Kaixin Lan, Derek F. Wong, Jinpeng Hu, Lidia S. Chao, and Yue Zhang. 2023. Is chatgpt a highly fluent grammatical error correction system? a comprehensive evaluation.

Mor Geva, Eric Malmi, Idan Szpektor, and Jonathan Berant. 2019. DiscoFuse: A large-scale dataset for discourse-based sentence fusion. In *Proceedings of the 2019 Conference of the*

North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers), pages 3443–3455, Minneapolis, Minnesota. Association for Computational Linguistics.

Edward J Hu, yelong shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. 2022. LoRA: Low-rank adaptation of large language models. In *International Conference on Learning Representations*.

Bogdan Ivanyuk-Skulskiy, Anton Zaliznyi, Oleksand Reshetar, Oleksiy Protsyk, Bohdan Romanchuk, and Vladyslav Shpihanovych. 2021. uadatasets: a collection of ukrainian language datasets.

Albert Q. Jiang, Alexandre Sablayrolles, Arthur Mensch, Chris Bamford, Devendra Singh Chaplot, Diego de las Casas, Florian Bressand, Gianna Lengyel, Guillaume Lample, Lucile Saulnier, Lélio Renard Lavaud, Marie-Anne Lachaux, Pierre Stock, Teven Le Scao, Thibaut Lavril, Thomas Wang, Timothée Lacroix, and William El Sayed. 2023. Mistral 7b.

Chao Jiang, Mounica Maddela, Wuwei Lan, Yang Zhong, and Wei Xu. 2020. Neural CRF model for sentence alignment in text simplification. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 7943–7960, Online. Association for Computational Linguistics.

Philippe Laban, Jesse Vig, Marti A Hearst, Caiming Xiong, and Chien-Sheng Wu. 2023. Beyond the chat: Executable and verifiable text-editing with llms. *arXiv preprint arXiv:2309.15337*.

Haonan Li, Fajri Koto, Minghao Wu, Alham Fikri Aji, and Timothy Baldwin. 2023. Bactrian-x : A multilingual replicable instruction-following model with low-rank adaptation.

Jonathan Mallinson, Rico Sennrich, and Mirella Lapata. 2020. Zero-shot crosslingual sentence simplification. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 5109–5126, Online. Association for Computational Linguistics.

Louis Martin, Angela Fan, Éric de la Clergerie, Antoine Bordes, and Benoît Sagot. 2022. MUSS: Multilingual unsupervised sentence simplification by mining paraphrases. In *Proceedings of the Thirteenth Language Resources and Evaluation Conference*, pages 1651–1664, Marseille, France. European Language Resources Association.

Niklas Muennighoff, Thomas Wang, Lintang Sutawika, Adam Roberts, Stella Biderman, Teven Le Scao, M Saiful Bari, Sheng Shen, Zheng Xin Yong, Hailey Schoelkopf, Xiangru Tang, Dragomir Radev, Alham Fikri Aji, Khalid Almubarak, Samuel Albanie, Zaid Alyafeai, Albert Webson, Edward Raff, and Colin Raffel. 2023. Crosslingual generalization through multitask finetuning. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 15991–16111, Toronto, Canada. Association for Computational Linguistics.

Niklas Muennighoff, Thomas Wang, Lintang Sutawika, Adam Roberts, Stella Biderman, Teven Le Scao, M Saiful Bari, Sheng Shen, Zheng-Xin Yong, Hailey Schoelkopf, et al. 2022. Crosslingual generalization through multitask finetuning. *arXiv preprint arXiv:2211.01786*.

OpenAI. 2023. Gpt-4 technical report. *ArXiv*, abs/2303.08774.

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. Bleu: a method for automatic evaluation of machine translation. In *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics*, pages 311–318, Philadelphia, Pennsylvania, USA. Association for Computational Linguistics.

Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J. Liu. 2020. Exploring the limits of transfer learning with a unified text-to-text transformer. *Journal of Machine Learning Research*, 21(140):1–67.

Vipul Raheja, Dimitris Alikaniotis, Vivek Kulkarni, Bashar Alhafni, and Dhruv Kumar. 2024. medit: Multilingual text editing via instruction tuning.

Vipul Raheja, Dhruv Kumar, Ryan Koo, and Dongyeop Kang. 2023. CoEdIT: Text editing by task-specific instruction tuning. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 5274–5291, Singapore. Association for Computational Linguistics.

Jeff Rasley, Samyam Rajbhandari, Olatunji Ruwase, and Yuxiong He. 2020. Deepspeed: System optimizations enable training deep learning models with over 100 billion parameters. In *Proceedings of the 26th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*, KDD '20, page 3505–3506, New York, NY, USA. Association for Computing Machinery.

Sascha Rothe, Jonathan Mallinson, Eric Malmi, Sebastian Krause, and Aliaksei Severyn. 2021.

A simple recipe for multilingual grammatical error correction. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 2: Short Papers)*, pages 702–707, Online. Association for Computational Linguistics.

Michael Ryan, Tarek Naous, and Wei Xu. 2023. Revisiting non-English text simplification: A unified multilingual benchmark. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 4898–4927, Toronto, Canada. Association for Computational Linguistics.

Horacio Saggion, Sanja Štajner, Daniel Ferrés, Kim Cheng Sheang, Matthew Shardlow, Kai North, and Marcos Zampieri. 2022. Findings of the TSAR-2022 shared task on multilingual lexical simplification. In *Proceedings of the Workshop on Text Simplification, Accessibility, and Readability (TSAR-2022)*, pages 271–283, Abu Dhabi, United Arab Emirates (Virtual). Association for Computational Linguistics.

Timo Schick, Jane A. Yu, Zhengbao Jiang, Fabio Petroni, Patrick Lewis, Gautier Izacard, Qingfei You, Christoforos Nalmpantis, Edouard Grave, and Sebastian Riedel. 2023. PEER: A collaborative language model. In *The Eleventh International Conference on Learning Representations*.

Lingfeng Shen, Lemao Liu, Haiyun Jiang, and Shuming Shi. 2022. On the evaluation metrics for paraphrase generation. *arXiv preprint arXiv:2202.08479*.

Vasyl Starko and Andriy Rysin. 2023. Creating a POS gold standard corpus of Modern Ukrainian. In *Proceedings of the Second Ukrainian Natural Language Processing Workshop (UNLP)*, pages 91–95, Dubrovnik, Croatia. Association for Computational Linguistics.

Xin Sun, Tao Ge, Shuming Ma, Jingjing Li, Furu Wei, and Houfeng Wang. 2022. A unified strategy for multilingual grammatical error correction with pre-trained cross-lingual language model.

Oleksiy Syvokon, Olena Nahorna, Pavlo Kuchmiichuk, and Nastasiia Osidach. 2023. UA-GEC: Grammatical error correction and fluency corpus for the Ukrainian language. In *Proceedings of the Second Ukrainian Natural Language Processing Workshop (UNLP)*, pages 96–102, Dubrovnik, Croatia. Association for Computational Linguistics.

Rohan Taori, Ishaan Gulrajani, Tianyi Zhang, Yann Dubois, Xuechen Li, Carlos Guestrin, Percy Liang, and Tatsunori B. Hashimoto. 2023. Stanford alpaca: An instruction-following llama model.

Hugo Touvron, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Timothée Lacroix, Baptiste Rozière, Naman Goyal, Eric Hambro, Faisal Azhar, et al. 2023. Llama: Open and efficient foundation language models. *arXiv preprint arXiv:2302.13971*.

Xiangpeng Wei, Haoran Wei, Huan Lin, Tianhao Li, Pei Zhang, Xingzhang Ren, Mei Li, Yu Wan, Zhiwei Cao, Binbin Xie, Tianxiang Hu, Shangjie Li, Binyuan Hui, Bowen Yu, Dayiheng Liu, Baosong Yang, Fei Huang, and Jun Xie. 2023. Polylm: An open source polyglot large language model.

BigScience Workshop. 2023. Bloom: A 176b-parameter open-access multilingual language model.

Haoran Wu, Wenxuan Wang, Yuxuan Wan, Wenxiang Jiao, and Michael Lyu. 2023. Chatgpt or grammarly? evaluating chatgpt on grammatical error correction benchmark.

Wei Xu, Courtney Napoles, Ellie Pavlick, Quanze Chen, and Chris Callison-Burch. 2016a. Optimizing statistical machine translation for text simplification. *Transactions of the Association for Computational Linguistics*, 4:401–415.

Wei Xu, Courtney Napoles, Ellie Pavlick, Quanze Chen, and Chris Callison-Burch. 2016b. Optimizing statistical machine translation for text simplification. *Transactions of the Association for Computational Linguistics*, 4:401–415.

Linting Xue, Noah Constant, Adam Roberts, Mihir Kale, Rami Al-Rfou, Aditya Siddhant, Aditya Barua, and Colin Raffel. 2021. mT5: A massively multilingual pre-trained text-to-text transformer. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 483–498, Online. Association for Computational Linguistics.

Xingxing Zhang and Mirella Lapata. 2017. Sentence simplification with deep reinforcement learning. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pages 584–594, Copenhagen, Denmark. Association for Computational Linguistics.

Yuan Zhang, Jason Baldridge, and Luheng He. 2019. PAWS: Paraphrase adversaries from word scrambling. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 1298–1308, Minneapolis, Minnesota. Association for Computational Linguistics.

Yaoming Zhu, Sidi Lu, Lei Zheng, Jiaxian Guo, Weinan Zhang, Jun Wang, and Yong Yu. 2018. Texygen: A benchmarking platform for text generation models.

Ahmet Üstün, Viraat Aryabumi, Zheng-Xin Yong, Wei-Yin Ko, Daniel D'souza, Gbemileke Onilude, Neel Bhandari, Shivalika Singh, Hui-Lee Ooi, Amr Kayid, Freddie Vargus, Phil Blunsom, Shayne Longpre, Niklas Muennighoff, Marzieh Fadaee, Julia Kreutzer, and Sara Hooker. 2024. Aya model: An instruction finetuned open-access multilingual language model. *arXiv preprint arXiv:2402.07827*.

8. Appendix A

GEC Input ▶ Виправте граматику в цьому реченні: А ти, батюшка, стало бути, тут в сторожі? **Output** ▶ А ти, батюшко, стало бути, тут у сторожі?	**GEC Input** ▶ Correct the grammar in this sentence: And you, father, are you here in guard duty? **Output** ▶ And you, father, are you here on guard duty?
Coherence Input ▶ Покращіть зв'язність тексту: Тоді він відвідав острови Греції. Хвороба змусила його повернутися до Риму. **Output** ▶ Потім він відвідав острови Греції, але хвороба змусила його повернутися до Риму.	**Coherence Input** ▶ Improve the coherence of the text: Then he visited the islands of Greece. Illness forced him to return to Rome. **Output** ▶ He then visited the islands of Greece, but illness forced him to return to Rome.
Simplification Input ▶ Спростіть речення: вода спочатку холодна, а потім волога. **Output** ▶ Вода і холодна, і мокра.	**Simplification Input** ▶ Simplify the sentence: first the water is cold, and then it is wet. **Output** ▶ The water is both cold and wet.
Paraphrasing Input ▶ Перефразуйте речення: Наскільки я можу судити, обидва ще живі. **Output** ▶ Наскільки я розумію, вони обидва ще живі.	**Paraphrasing Input** ▶ Rephrase the sentence: As far as I can tell, both are still alive. **Output** ▶ As far as I understand, they are both still alive.

Table 7: Example model inputs and outputs of the text editing tasks that Spivavtor can perform. English translations of the examples in Figure 1 are provided for reference.

Model	**Text Editing Tasks**						
	Simplification		**Coherence**		**Paraphrasing**		
	Asset	Turk	Sports	Wiki	MRPC	STS	QQP
Copy	17.75	24.04	26.61	28.37	100/39.90	100/38.80	100/26.20
Bactrian-X-7B	36.02	37.13	40.7	38.62	65.5/29.20	45.6/20.4	13.5/4
UAlpaca-7B	33.54	35.96	32.48	33.45	57.6/24.2	20.5/9.6	6.2/1.8
Mistral-7B	39.85	38.54	32.75	30.58	37.6/18	16.9/8.3	5.9/2
mt0-large	32.91	27.94	22.32	21.20	10.8/5.2	4.7/2.3	4.7/1.3
aya-101	32.02	37.32	38.42	37.68	78.5/34.5	56.8/25	32.1/9.8
GPT-3.5-Turbo	42.52	39.04	44.84	45.44	33.2/17.6	24/12.4	22.9/9
GPT4	42.20	39.05	43.35	43.96	29.2/16.1	17/12.7	19.9/9
Spivavtor-Bactrian-X-7B	35.15	37.75	47.29	50.53	63.2/29.1	67.1/31.9	66.5/20.2
Spivavtor-Mistral-7B	31.73	35.92	44.01	44.72	75.1/31.8	81/31.7	77.3/21.3
Spivavtor-Llama-2-7B-chat	39.29	35.80	48.13	51.95	46.2/22.3	50.7/22.3	50.5/16.7
Spivavtor-Llama-2-13B-chat	37.09	36.93	47.54	53.94	55.5/26.6	57.9/24.6	58.3/18.1
Spivavtor-mt5-large	34.82	37.17	47.97	49.87	71/32	78/34.8	80.7/23.3
Spivavtor-mt0-large	33.85	37.28	48.25	48.41	71.5/32.3	79.4/34.3	81.2/23.2
Spivavtor-mt5-xxl	38.50	37.52	48.87	49.53	67.3/30.9	69.1/30.5	75.3/22.3
Spivavtor-mt0-xxl-mt	38.95	38.20	48.67	53.80	65.4/30.4	69.6/34.8	70.4/21.6
Spivavtor-aya-101	37.71	37.95	47.87	51.94	69.9/31.6	71.7/33.3	74.2/22.5

Table 8: Comparison of Spivavtor models against various baselines, categorized by constituent datasets. We report detailed metrics for each dataset within a task. GEC is not relevant here since it is a single dataset. For Simplification and Coherence, we report SARI. For Paraphrasing, we report reference-free / reference-based BLEU just as in Table 3. All scores have been scaled to lie between 0 and 100.

Task	Verbalizers	English translation
GEC	“Виправте граматику в цьому реченні:”	“Correct the grammar in this sentence:”
	“Виправте граматичні помилки в цьому реченні:”	“Correct the grammatical errors in this sentence:”
	“Удосконаліть граматику цього тексту:”	“Improve the grammar of this text:”
	“Виправте всі граматичні помилки:”	“Correct all grammatical errors:”
	“Зробіть речення граматичним:”	“Make the sentence grammatical:”
	“Видаліть граматичні помилки:”	“Remove grammatical errors:”
	“Виправте помилки в цьому тексті:”	“Correct the errors in this text:”
	“Виправте граматичні помилки:”	“Correct the grammatical errors:”
	“Виправити граматику:”	“Correct the grammar:”
Simplification	“Спростіть речення:”	“Simplify the sentences:”
	“Напишіть простішу версію для речення:”	“Write a simpler version for the sentence:”
	“Спростіть це речення:”	“Simplify this sentence:”
	“Зробіть речення простим:”	“Make the sentence simple:”
	“Спростіть цей текст:”	“Simplify this text:”
	“Перепишіть речення так, щоб воно було простішим:”	“Rewrite the sentence so that it is simpler:”
	“Перепишіть це речення простіше:”	“Rewrite this sentence more simply:”
	“Зробіть речення простіше:”	“Make the sentences simpler:”
	“Спростіть цей текст:”	“Simplify this text:”
	“Використовуйте простіші слова:”	“Use simpler words:”
	“Зробіть цей текст легше для розуміння:”	“Make this text easier to understand:”
Coherence	“Виправте зв’язність в реченні:”	“Correct the coherence in the sentence:”
	“Покращіть зв’язність тексту:”	“Improve text coherence:”
	“Виправте зв’язність в цьому тексті:”	“Correct the coherence in this text.”
	“Виправте відсутність зв’язності в реченні:”	“Correct the lack of coherence in the sentence:”
	“Виправте зв’язність в тексті:”	“Correct the coherence in the text:”
	“Виправте зв’язність речення:”	“Correct the coherence of the sentence:”
	“Зробіть текст більш зв’язним:”	“Make the text more coherent:”
Paraphrasing	“Перефразуйте речення:”	“Rephrase the sentence:”
	“Перепишіть речення іншими словами:”	“Rewrite the sentence in other words:”
	“Перефразуйте цей текст:”	“Paraphrase this text:”
	“Перефразуйте це речення:”	“Rephrase this sentence:”
	“Перефразуйте:”	“Paraphrase:”
	“Напишіть перефраз для речення:”	“Write a paraphrase for the sentence:”
	“Напишіть перефразовану версію речення:”	“Write a paraphrased version of the sentence:”
	“Перепишіть це речення:”	“Rewrite this sentence:”
	“Перепишіть цей текст:”	“Rewrite this text:”
	“Переформулюйте це речення:”	“Rephrase this sentence:”
	“Перефразуйте це речення:”	“Paraphrase this sentence.”
	“Переформулюйте цей текст:”	“Rephrase this text:”

Table 9: A complete list of verbalizers for each task used as instructions in the SPIVAVTOR-Instruct dataset. The English translations are provided for reference.

Eval-UA-tion 1.0: Benchmark for Evaluating Ukrainian (Large) Language Models

Serhii Hamotskyi[1], Anna-Izabella Levbarg[2], Christian Hänig[1]
[1] Anhalt University of Applied Sciences
Bernburger Str. 55, 06366 Köthen, Germany
[2] University of Greifswald
Domstraße 11, 17489 Greifswald, Germany
{serhii.hamotskyi, christian.haenig}@hs-anhalt.de, anna-izabella.levbarg@uni-greifswald.de

Abstract

We introduce Eval-UA-tion, a comprehensive suite of novel Ukrainian-language datasets designed for the evaluation of language model performance in the Ukrainian language. The collection encompasses a variety of tasks: UA-CBT (inspired by the Children's Book Test, a fill-in-the-blanks task aimed at assessing comprehension of story narratives), UP-Titles (requiring the association of articles from the online newspaper Ukrainska Pravda with their correct titles from a set of ten similar options), and LMentry-static-UA/LMES (modeled after the LMentry benchmark, featuring tasks that are straightforward for humans yet challenging for language models, such as determining the longer of two words or identifying the Nth word in a sentence). Except for UP-Titles, these tasks are designed to minimize potential contamination, utilizing material unlikely to be found in language models' training datasets. They also include a split specifically for few-shot prompting to further reduce contamination risks. For each task, we provide benchmarks against both human and random performance baselines.

Keywords: LLM Evaluation, Benchmark Dataset, Ukrainian language

1. Introduction

The Ukrainian language has a strong online presence: as of October 2023, estimates of languages used on the internet put Ukrainian at place 19 (Wikipedia contributors, 2023) (between Arabic and Greek); Ukrainian Wikipedia is 15th by number of daily views and number of articles (Meta, 2022). Though an increase of Ukrainian use online can be traced to the Russian attack on Crimea in 2014 (Kulyk, 2018), the full-scale invasion of 2022 accelerated this process, as seen surveys (Group, 2022) and Twitter data (Racek et al., 2024), showing that ~25% predominantly Russian-tweeting users made a hard switch to Ukrainian in the first months of the invasion. This shows that the need to support the Ukrainian language is stronger than ever.

On a 2020 survey of linguistic diversity in NLP (Joshi et al., 2020), the Ukrainian language was classified as belonging to the "rising stars": languages with a thriving online cultural community that benefits from unsupervised pretraining, but let down by an insufficient amount of *labeled* datasets. A recent review of the performance of LLMs on non-English languages found a very uneven performance based on language used, with ChatGPT performing best in English (Lai et al., 2023)[1]. With the widespread adoption of LLMs these differences become more important, and so is their measurement.

Aiming to increase the availability of labeled Ukrainian datasets and stimulating future and existing efforts on this topic, we present Eval-UA-tion 1.0, a set of benchmark datasets usable for evaluating the performance of LLMs in and on the Ukrainian language.

The issue of data contamination (generally defined as exposure of the model to data similar to the one it would later be tested on) has received much attention in recent years (Roberts et al., 2023). We placed a special emphasis on using sources of data that maximally limit contamination.

Most of the source code and sanitized raw data used to generate the datasets will be publicly available in the Eval-UA-tion Github repository[2].

1.1. Relevant Ukrainian Grammar and Notation

Ukrainian has 3 grammatical genders: female, male, and neutral (in this paper abbreviated as F, M, and N), 7 cases (including nominative/NOM, genitive/GEN, locative/LOC), and 2 numbers (singular/SG and plural/PL). It has a complex morphology with many parts of speech needing agreement, especially by gender and case. Numerals can be ordinals/ORD (*first*), cardinals/CARD (*one*) and adverbial.

The notation used is loosely based on the Leipzig Glossing Rules (Comrie et al., 2008), with the relevant morphemes annotated in the

[1] Ukrainian is an interesting outlier in that study as the only language where English prompts outperformed the language-specific (Ukrainian) ones for Relation Extraction on the SMiLER (Seganti et al., 2021) dataset.

[2] https://github.com/pchr8/eval-UA-tion

The Third Ukrainian Natural Language Processing Workshop (UNLP) @LREC-COLING-2024, pages 109–119
25 May, 2024. © 2024 ELRA Language Resource Association: CC BY-NC 4.0

superscripts of words. The English translation of the relevant words will be divided from the morphemes by a dash, and the individual morphemes will be separated from each other by dots: *чоловік*$^{\textit{man-NOM.SG}}$ *побачив*$^{\textit{saw-M.SG}}$ *собаку*$^{\textit{dog-ACC.SG}}$.

2. Related Work

A very thorough overview of the current landscape of benchmarking approaches can be found in (Guo et al., 2023). On LLMs' performance on non-English languages, see Akter et al. (2023) and Lai et al. (2023).

A number of efforts are underway to create Ukrainian-language datasets and benchmarks, a notable one being UA-datasets (Ivanyuk-Skulskiy et al., 2021)[3], with the development of UA-SQuAD and UA News classification in progress as of 04.03.2024 and the Mova Institute POS dataset completed. All three datasets are considerably larger than the ones we are proposing and have been a direct inspiration for us.

Loosely related to our manual correction of LLM-generated stories is the topic of grammaticality in general. UA-GEC (Syvokon and Nahorna, 2022) is a large grammatical error correction corpus separately annotating fluency, grammar, punctuation, and spelling errors.

3. Eval-UA-tion 1.0 Benchmark Datasets

3.1. UA-CBT

3.1.1. Introduction

The UA-CBT[4] dataset builds upon the idea introduced in the English-language Children's Book Test (CBT) benchmark dataset (Hill et al., 2015).

The core idea is the following: a word in a story gets masked (replaced by "______", hereafter referred to as 'gap') and six options are offered as potential replacements, only one being correct.

3.1.2. Differences from the Original CBT Task

UA-CBT differs from the CBT benchmark in multiple aspects (and through the challenges introduced by the rich morphology of the Ukrainian language).

In the original CBT implementation, the story context was 20 sentences long, with a word in the 21st sentence masked. In UA-CBT, to increase the number of tasks per story, the split is 65% context segment and 35% challenge segment. The number of possible options is reduced from 10 to 6. We additionally omitted prepositions from the question categories, keeping named entities, common nouns, and verbs.

The (2015) CBT task is built from stories from books freely available on Project Gutenberg[5] and the authors explicitly state that they wanted to incentivize models to apply background knowledge and information when solving the tasks — we attempted to avoid that by using original stories and limit the background knowledge usable to story cliches that aren't always applicable[6]. Lastly, the task instances were manually filtered to ensure the dataset contains only unambiguous solvable questions.

3.1.3. Description

The dataset contains **1,061** task instances built on **72** different stories. There are three types of tasks/gaps: NAMED_ENTITY for the characters ('Butterfly'), COMMON_NOUN for inanimate items ('valley', 'water') and VERB for verbs ('fly', 'eat'). Each instance is a multiple-choice question with 6 options.

Distractors For each gap, six different options are provided, five of them are *distractors* (wrong answers). Three to five distractors come from the story itself. To make them plausible, only the lemmas[7] most frequently found in the text are used. All are filtered and inflected to match the morphology of the original word in the gap. For example, in the task shown in Fig. 1, the replacements for *Мисливця*$^{\textit{hunter-M.GEN}}$ are all grammatically male and GEN case as well (with the exception of *Змії*$^{\textit{snake-F.GEN}}$, described later); all use the same capitalization as the original word (in the story, 'The Hunter' is used in the role of a proper name and is, therefore, capitalized).

If the story doesn't have enough entities usable as distractors (e.g. only one grammatically female character for NAMED_ENTITY), they are sourced in the following order: 1) If the story's most frequently mentioned entity has a different gender than the gap, it's added as a most-frequent-any-gender distractor, marked as a red "F" in Fig. 1; 2)

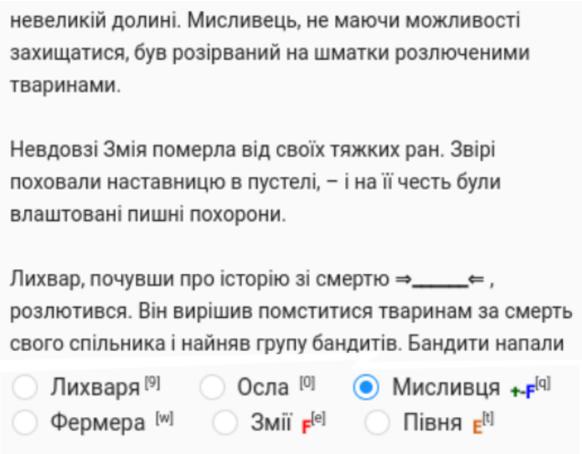
[3] https://fido-ai.github.io/ua-datasets/

[4] https://hf.co/datasets/shamotskyi/ua_cbt

[5] https://www.gutenberg.org/

[6] The stories, being generated by LLMs and corrected only for logic but not for plausibility, contain atypical elements such as a turtle eating the remains of a zebra: may raise a human's eyebrows, but may be even more confusing to an LM that expects animals to fit archetypal folk tale roles.

[7] different inflections of the same word counted as one (e.g. *кіт, кота, котами*)



невеликій долині. Мисливець, не маючи можливості захищатися, був розірваний на шматки розлюченими тваринами.

Невдовзі Змія померла від своїх тяжких ран. Звірі поховали наставницю в пустелі, – і на її честь були влаштовані пишні похорони.

Лихвар, почувши про історію зі смертю ⇒_____⇐ , розлютився. Він вирішив помститися тваринам за смерть свого спільника і найняв групу бандитів. Бандити напали

○ Лихваря [9] ○ Осла [0] ● Мисливця +-F[q]
○ Фермера [w] ○ Змії F[e] ○ Півня E[t]

Figure 1: A (partial) sample UA-CBT task. The markings near the options are the ones shown to the annotators during the task filtering process: "E" means the option was taken from an external list of (in this case) male entities, a blue "F" denotes the most frequent relevant word in the text, a *red* "F" is the most frequent word in the text regardless of its gender (and here *змія*$^{\textit{snake-F}}$ is the only grammatically female word), and "+" is the correct option.

An external list of words is used, from which the remaining distractors are randomly chosen. The options are then shuffled and deduplicated.

Gaps Only frequent lemmas become gaps. Masking rare words would have increased the chances of a gap being placed on a one-off entity that's not part of a coherent narrative. Lemma frequency for gaps was calculated only up to the gap itself. For verbs and named entities, at least **two** occurrences were needed, for common nouns **four**. The higher minimum occurrences limit for common nouns was needed because many of the stories contained generic endings that resulted in uninteresting tasks, solvable by completing cliches instead of understanding the story narrative ("...and the animals learned that the real treasure is `[friendship|food|fear|...]`, and they `[lived|ate|traveled|...]` together happily ever after"). The three kinds of gaps in more detail:

NAMED_ENTITY animate nouns and proper nouns; usually the main characters in the story ('Butterfly'/*Метелик*)

COMMON_NOUN inanimate nouns; usually objects like 'water' or 'desert', but overlaps heavily with NAMED_ENTITY (because animals weren't always detected as animate by the spacy model we used)

VERBS finite and infinitive

3.1.4. Dataset structure

The dataset is published on the Huggingface Hub[8] with five predefined subsets: NAMED_ENTITY (615 instances), COMMON_NOUN (281), VERBS (165), 'all' with the complete dataset (1,061), and a few-shot split (7 instances based on a separate story). The latter's purpose is avoiding contamination during few-shot prompting (randomly selecting instances for this purpose might lead to the few-shot examples using the same story as the test instance).

The columns are described in the README of the dataset. Notable ones are:

context, question the story segments
options, answer the options and correct answer
taskType gap type (COMMON_NOUN, ...)
storyId unique identifier of the story used

A large amount of other metadata is included, such as the source of each distractor, the size of the segments, and metadata from the story generation stage (e.g. which model was used; see Section 3.1.5).

3.1.5. Story Generation and Filtration

Roberts et al. (2023) describe contamination as composed of two distinct phenomena: *contamination* proper, which refers to an LLM's exposure during training to examples similar or identical to the ones the model will later be evaluated on, and *memorization*, the ability to extract (near) verbatim the examples the model has seen during training. When generating stories for this task, the latter facet was at the forefront. Many sources of stories were considered and rejected. The crux of the issue was that stories not widely available online were unusable for intellectual property reasons, while public domain stories were often available online and, therefore, basically guaranteed to be part of the training data of current (and future) LLMs. Our stories were generated using OpenAI *gpt-4-1106-preview*[9] and Google Gemini Pro[10], followed by manual review and correction. The main challenge we faced was that the LLM would recite a memorized story instead of writing a more original one, thereby contaminating the dataset.

We mitigated this issue by using detailed **prompts**. For example, if the prompt asks for a story about a raven and a fox, the names and details would vary but the story will almost always be about the fox tricking the raven into giving it a piece of cheese, as in the well-known Aesop fable. But if

[8] https://hf.co/datasets/shamotskyi/ua_cbt
[9] https://platform.openai.com/docs/models
[10] https://deepmind.google/technologies/gemini/

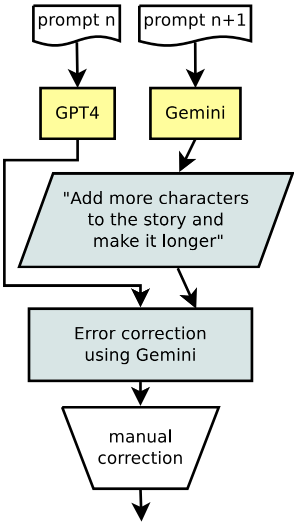

prompt n
prompt n+1
GPT4
Gemini
"Add more characters to the story and make it longer"
Error correction using Gemini
manual correction

Figure 2: The flow used to create UA-CBT stories.

```
options:
    - not learning anything
    - helping their mentor with {problem_type} problem
    - resolving a dispute involving {dispute_topic}
    - proving that they are a good {profession}
    - rescuing {entity} from {rescue_from}
    - proving their innocence
parts:
    problem_type:
        - an embarassing
        - an unexpected
        - a recurring
        - a financial
        - a communication
        - "a totally predictable"
    dispute_topic:
        - lost food
        - stolen food
    profession:
        - friend
        - tailor
        - hunter
    entity:
        - a relative
        - a lost traveler
    rescue_from:
        - a tornado
        - the cold
        - captivity
```

Figure 3: Part of the template used to generate story generation prompts.

the prompt asks for ”a story about a *greedy* raven *rescuing* a fox from *a tornado*”, there’s a much smaller set of pre-existing stories fitting the criteria to recite verbatim, resulting in more creative stories. A number of such elements in the template were randomized, such as asking it for stories in the style of Ukrainian/Arabic folk tales, changing the number of main/minor characters, etc. Lastly, specifying that the story should have an unhappy ending often increased the originality of the entire story, so half of the prompts required stories with unhappy endings.

Generating these prompts involved sampling a subset out of all possible permutations of values in the templates. Part of the YAML file containing the source data (redacted for brevity) is shown in Fig. 3. The need for logic, consistency, and a coherent structure and *recurring* characters was emphasized, since this was needed to be able to create a story from which a higher number of solvable task instances could be generated. Otherwise, the bulk of the prompt was static and contained criteria for the story. It specified the naming of the characters, the complexity of the story, and instructions aimed at avoiding specific recurring motifs (e.g. prompts involving specific objects, such as bread, often defaulted to a narrative centered on *magic bread* rather than incorporating bread in a conventional role).

Half of the stories were generated using *gpt-4-1106-preview* and half using Gemini Pro. In our experience, the OpenAI model followed instructions (such as number of characters) more reliably, while the Gemini model had dramatically better Ukrainian grammar (which agrees with the literature; compare with Akter et al. (2023)). We leveraged Gemini Pro’s Ukrainian language abilities by piping all the stories through it after generation, instructing it to improve their logic, consistency, and grammar, with good results. We mitigated its tendency to generate shorter and simpler stories than required by using a chat interface, and asking it after its first attempt to ”add more major/minor characters to the story make it longer, while keeping it logically consistent”, with good results.

The flow for both models is shown on Fig. 2.

Manual story correction and filtration was done by human annotators based on the stories produced after the above steps, in a Label Studio[11] environment. For each story, the annotators were given a choice of fixing the errors in the story or marking it as completely unusable. Reasons for the latter included continuity errors that required substantial rewriting to fix, a large number of errors in gender agreement or entities having adjectival names (e.g. a rabbit named Quick), or having too few characters.

Out of the 117 generated stories, 72 (62%) were considered usable and subsequently manually corrected. A typology of errors found during this process is out of scope of this paper, but the main language issues found were noun agreement (with nouns that have a different gender in Ukrainian and Russian using the Russian gender), the use of Russian words and phrases, and strange and often funny fluency errors. Issues in the logic involved illogical actions by the characters (such as money being returned to the wrong character) and continuity issues (e.g., a character giving advice despite having died two paragraphs ago).

The before-and-after stories dataset[12] is available on request.

[11] https://labelstud.io/

[12] https://hf.co/datasets/shamotskyi/ua_cbt_stories

3.1.6. Human Filtration of Task Instances

Departing from the approach taken by the original CBT task, we manually filtered all generated task instances to remove unsuitable ones. Of the 1,418 manually processed instances only 1,063 (75%) were deemed suitable.

Here a more extensive taxonomy was created, with two main classes of errors:

1. Logic/continuity errors
   (a) Answer unknown: the story doesn't contain the information that allows the answer to be inferred. Example: ”The Cat and the Turtle go to `[Cat|Turtle|Lion]`'s house to sew the coat, and later deliver it to the Lion's house”.
   (b) Multiple options are correct: it's clear which entity/action is involved, but it can be described in different ways. Example: ”The Lion liked the Cat and Turtle's `[coat|work]`.”This accounts for approx. 24% of unusable tasks and was the largest category.
   (c) Duplicate options: multiple almost identical options referring to the same thing, e.g. bird/birdie. Caused by incorrect lemmatization.
2. Language errors
   (a) Ungrammatical option: one of the options is a non-existing word. Caused by failures in the parsing-normalization-inflection pipeline. Examples from the dataset include **друзь*[13] and **комаревом*.
   (b) Incorrectly inflected option: an option is an existing grammatical word, but is a different inflection than needed. Usually caused by an incorrectly detected morphology of the masked word.

Both error classes are roughly equally distributed. We see this taxonomy and breakdown as a stepping stone towards fully automated filtering of task instances, eventually leading to larger datasets of this type.

3.1.7. Baselines

The human baseline accuracy result for this task was **94%**: 6 wrong out of a total of 99 test instances. This score is based on answers by 8 different annotators inside a Telegram[14] bot. The random baseline for this task is **16.7%** (6 possible options). The most-frequent baseline of this task (choosing the option most frequently seen in the story) is **57%** (in other words, in 57% of the tasks the correct answer is simply the most frequently mentioned lemma). This is visualized in Fig. 4.

3.2. LMentry-static-UA (LMES)

3.2.1. Description

LMentry-static-UA (LMES) is a set of 6 loosely related datasets inspired by the (English-language) LMentry (Efrat et al., 2022) benchmark. It focuses on tasks considered trivial for humans but harder for LMs.

The six included tasks are:

1. N-in-M–type tasks:
   (a) LOW[15] (letters of word): ”What is the first/Nth/last letter in the word ...”
   (b) WIS[16] (words in sentence): ”What is the first/Nth/last word in this sentence:...”
2. Tasks involving categories:
   (a) CATS-MC[17] (multiple choice): ”Which of these words is different from the rest?”
   (b) CATS-BIN[18] (binary): ”Do all of these words belong to the category 'emotions'?”
3. Comparing-two-things-type tasks:
   (a) WordAlpha[19]: ”Which of these words is first in alphabetical order?”
   (b) WordLength[20]: ”Which of these words is longer?”

3.2.2. Differences from LMentry

LMentry represents a comprehensive framework that includes evaluation code[21], assesses the models' accuracy and robustness to perturbations, and extends beyond the scope of our (static) dataset in many ways. The two commonalities lie in the tasks themselves and in a focus on investigating the robustness of LMs to changes in the templates.

LMES focuses on tasks that can be evaluated as a dataset (as opposed to regular expressions in the original benchmark), hence 'static'. This necessitated dropping some tasks, such as ”write a sentence/word that contains/(starts/ends with) the word/letter X.” A number of other tasks were also dropped.

[13] Following linguistic conventions, ungrammatical words will be denoted by a leading asterisk.

[14] https://telegram.org/

[15] https://hf.co/datasets/shamotskyi/lmes_LOW

[16] https://hf.co/datasets/shamotskyi/lmes_WIS

[17] https://hf.co/datasets/shamotskyi/lmes_catsmc

[18] https://hf.co/datasets/shamotskyi/lmes_catsbin

[19] https://hf.co/datasets/shamotskyi/lmes_wordalpha

[20] https://hf.co/datasets/shamotskyi/lmes_wordlength

[21] https://github.com/aviaefrat/lmentry

The remaining tasks were regrouped, merged together, and expanded. For example, the original benchmark considered ”what's the first/last ...” separate tasks. We merged them into one and expanded by adding questions about specific numbers (”What's the fifth ...”).

3.2.3. Datasets Structure

The datasets have been uploaded on HuggingFace Hub as individual datasets, each with a separate few-show split that uses different sentences/words/categories than the train split to reduce contamination.

As a variation of what the LMentry benchmark terms *robustness*, our LMES tasks place a heavy emphasis on the use of different templates with the same input. For example, ”Which word is longer: 'dog' or 'cat'?” would also ask which word is *shorter*, would ask the same question reversing the order of the words, ask which word has more letters, etc. The specific changes to the template are contained in each task instance metadata to simplify analysis. The tasks involving words also include extensive metadata about the words, such as which part of speech they are, their frequency, their length, etc.

An analysis of the impact is outside the scope of this paper, but we hope it will stimulate research in this direction.

3.2.4. Dataset Construction

Since contamination is not an issue for the tasks involved (e.g. a sentence being in the training set of a LLM doesn't increase the odds of it knowing what's the third word in it), we used the UP-Titles (see subsection 3.3) dataset and the example sentences in spacy as sources for the sentences.

The words were taken from the David Klinger Ukrainian dictionary[22], which in turn uses DBnary (Sérasset, 2015) and WikiDictionary. We removed words containing apostrophes or dashes (to ensure clarity if counting letters is needed, e.g. the sixth letter in the word *пліч-о-пліч* depends on what is considered a letter). We left only nouns, verbs, adjectives, and adverbs; then we binned word frequency into high, mid, and low frequency. Then for each POS+frequency pair we sampled 60 words (or the number words available if it's less than 60), leading to a diverse choice of words.

3.2.5. Ukrainian Morphology in the Templates

The templates used in the LOW/WIS tasks involved converting integers (4) into natural-language words, which were represented by numerals of different types (ordinal and cardinal) and involved agreement in gender and case. For instance, asking for the first word in a sentence could be formulated as:

1. ***Перша***$^{\text{first-F.ORD.NOM}}$ *літера*$^{\text{letter-F.NOM}}$
2. *Літера*$^{\text{letter-F.NOM}}$ *номер* ***один***$^{\text{one-CARD.NOM}}$
3. *На* ***першому***$^{\text{first-N.ORD.LOC}}$ *місці*$^{\text{place-N.LOC}}$

We found no library that supported such arbitrary conversions. An additional challenge was keeping track of the numeral type and morphology required by each template.

We solved the latter problem by capitalizing the numeral directly in the template string: *На ПЕРШОМУ місці знаходиться...*. When using the template to generate task instances, the target morphology and numeral type are parsed from the capitalized numeral in the template, and the needed number is inflected correspondingly and put in the place of the capitalized numeral.

We release the code for the number-to-numeral conversion as a library, *ukr_numbers*[23], currently in beta. It uses pymorphy2[24] and num2words[25].

To the best of our knowledge, using natural language inside templates instead of requiring the user to manually specify the required inflection is a novel idea.

3.2.6. Baselines

The human and random baselines are shown on Table 1 and on Fig. 4.

3.3. UP-Titles

3.3.1. Description

UP-Titles is a multiple-choice dataset with 5,000 instances, where each article needs to be matched to the correct title, out of 10 similar titles. It's built from the ukr_pravda_2y[26] dataset, which contains articles from the Ukrainska Pravda[27] (UP) newspaper, published in the years 2022-2023. It's provided in a masked[28] and an unmasked[29] version (see below).

For each article text, its title and the titles of 9 most similar articles are given as choices. Article similarity is estimated through a simple cosine distance over article tag binary vectors: articles with the same tags will have a similarity of 1, and ones with no tags in common will have a similarity of 0.

[22] https://github.com/dmklinger/ukrainian

[23] https://github.com/pchr8/ukr_numbers

[24] https://github.com/pymorphy2/pymorphy2

[25] https://github.com/savoirfairelinux/num2words

[26] https://hf.co/datasets/shamotskyi/ukr_pravda_2y

[27] https://pravda.com.ua

[28] https://hf.co/datasets/shamotskyi/up_titles_masked

[29] https://hf.co/datasets/anilev6/up_titles_unmasked

	num_total	num_wrong	bl_random	bl_human
UP-Titles (unmasked)	99	12	10.00	87.88
UP-Titles (masked)	98	16	10.00	83.67
LMES-wordalpha	98	8	50.00	91.84
LMES-wordlength	100	6	50.00	94.00
LMES-cats_bin	99	3	50.00	96.97
LMES-cats_mc	100	2	20.00	98.00
LMES-LOW	100	3	9.43	97.00
LMES-WIS	100	6	4.69	94.00
UA-CBT	99	6	16.67	93.94

Table 1: Random and human baselines for the datasets part of this benchmark. num_total refers to the total size of the human-evaluated subset of the dataset, num_wrong is the number of instances where the human answer differs from ground truth; bl_random and bl_human are the random and the human baselines respectively. bl_random can be interpreted as the probability of randomly guessing the correct answer: bl_random = $\frac{1}{\text{num\_total}} \sum_{i=1}^{\text{num\_total}} \frac{1}{M_i}$, where M_i is a number of answer options in the i-th task instance. The random baselines were calculated on the complete datasets.

Most instances would be trivial to solve by matching by the numbers mentioned in the title and the article text — e.g. if an article text contains the number 232 (prisoners of war, dead russians, millions of dollars...) it's a very safe bet that whichever title contains that same number is the correct one. To mitigate that, we replace all integers in the article text and article titles with ”X” (leading to titles such as ”Bucha Mayor: XXX civilians killed by Russian troops identified”).

The solution doesn't remove all potential clues: among others, numerals written as text ('twenty-three'), months, names of individuals stay unchanged. Nevertheless, this simple masking approach complicates the task by a surprising amount, in some rare cases rendering it unsolvable (see discussion below about human baselines), and we believe a more thorough masking would bring diminishing returns while increasing the number of unsolvable instances even further.

The dataset is provided in two versions: with masked and unmasked numbers. We evaluated the masked and unmasked versions of the dataset separately, and the masked option was harder for both human annotators and LLMs.

It's released under the CC BY-NC 4.0[30] license, reflecting Ukrainska Pravda's terms[31] forbidding the use of its articles for commercial purposes.

3.3.2. Baselines

The random baseline for this task is **10%**. The human baseline was **84%** for the masked and **88%** for the unmasked version.

The low human baseline may be explained through different means, with the most likely ones being: 1. The title doesn't contain the information needed for disambiguation (”Another XXX Russians killed in Ukraine” would fit many articles written in the last two years); 2. Human error, inability to correct a wrong answer due to bot interface limitations.

4. Experiments

4.1. Evaluation Process

The datasets have been evaluated on five different models aiming to provide a baseline for the tasks. Baselines were calculated using the EleutherAI evaluation harness[32] (lm-eval).

All of the tasks in our benchmark can be seen as multiple-choice ones, and there are multiple approaches to leveraging LLMs for solving such tasks (Robinson et al., 2023). In cloze prompting, a question is passed to the LLM and the probabilities it gives to the different answers are compared, and the option given the highest probability by the model is used as prediction. We used multiple choice prompting (MCP), where the question and if applicable the possible answers are provided to the model in the prompt, structuring it in such a way that the model predicts a single token. For the UA-CBT and UP-Titles tasks this involved converting the list of possible answers into an enumerated list, e.g. ”A: cat; B: dog; C: uncle”. For the UP-Titles datasets, parentheses were used to avoid conflicts with article titles containing semicolons. Additionally, all newlines in the stories and UP articles were replaced by spaces. For the LMentry tasks, no letters were used, with the prompt expecting the correct word/letter [33] or *так/ні* (yes/no)

[30] https://creativecommons.org/licenses/by-nc/4.0/

[31] https://www.pravda.com.ua/rules/

[32] https://github.com/EleutherAI/lm-evaluation-harness/

[33] The LOW/WIS random baselines were calculated as if they were a multiple-choice question with the op-

for the LMES-cats_bin task.

The prompts used were all in Ukrainian and all tasks were evaluated in a 3-shot setting. Due to time and budgeting constraints, the OpenAI models evaluated only 200 instances of the UA-CBT and UP-Titles tasks and 500 instances of all LMES tasks; the other models were evaluated on the entire dataset.

One known limitation of the lm-eval harness is the lack of support for models' instruction formats to leverage instruction finetuning. Practically speaking, in our experiments all models used the same 3-shot prompting without any model-specific prompt finetuning. Even small changes to prompt templates can drastically change model scores, and our goal is to provide a baseline instead of maximizing accuracy by finetuning individual models' instruction prompt.

The lm-eval YAML task implementations (including the exact prompts and modifications) are posted in the Eval-UA-tion GitHub repository to ensure reproducibility.

4.2. Evaluation Results

The models tested were *gpt-3.5-turbo*, *gpt-4-1106-preview*, *mistralai/Mistral-7B-Instruct-v0.2*, *Radu1999/Mistral-Instruct-Ukrainian-slerp*, and *SherlockAssistant/Mistral-7B-Instruct-Ukrainian* (Boros et al., 2024) (the winner of the UNLP-2024 shared task), all from the Huggingface Hub. The results are shown on Fig. 4.

The *SherlockAssistant/Mistral-7B-Instruct-Ukrainian* model outperformed the other non-OpenAI models for all tasks and outperformed GPT3 for both UP-Titles tasks. Notably, that model was not finetuned on Ukrainian news datasets.

The effect of masking/unmasking numbers in the UP-Titles dataset can clearly be seen: masking decreased the scores of the models.

GPT4 outperformed or roughly equaled models on all tasks, most dramatically for the UA-CBT task; it also beat the human baselines for both versions of the UP task and UA-CBT. This may point either towards inattention being the source of the human errors on it, or the presence of UP articles in its training dataset. Splitting the UA-CBT instances by story generation model, the scores were practically identical for both subsets, at 0.97 (SD 0.17/0.18 for GPT4/Gemini). So instances from stories generated by Gemini and improved by Gemini weren't harder for GPT4 than the instances based on stories that it generated.

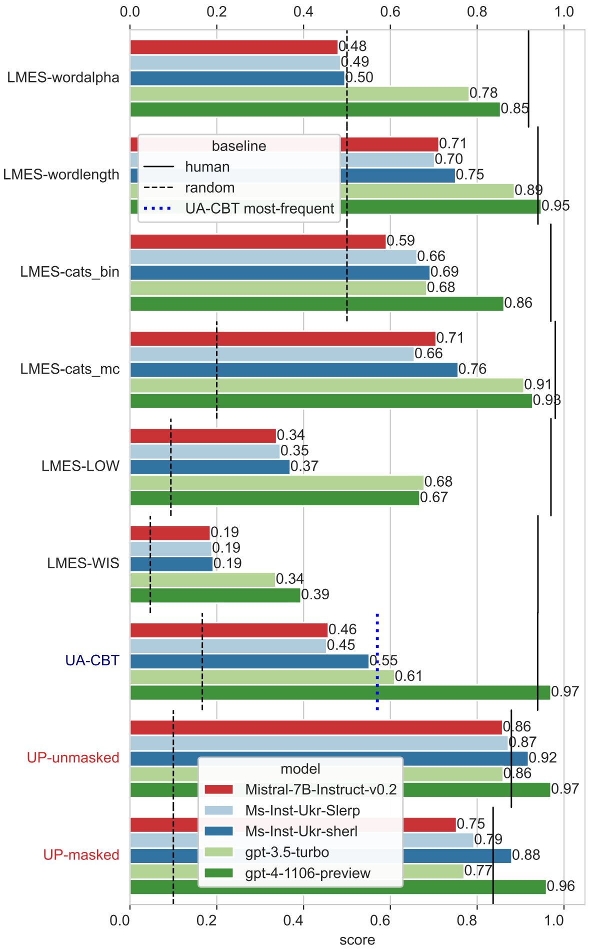


Figure 4: Evaluation scores of selected models.

5. Approaches to Human Data Annotation and Baseline Creation

For the presented datasets, volunteering contributors were found amongst family, friends, and through Telegram channels. This was coordinated in a group chat where instructions were given and annotators' questions answered. Initially, we employed Label Studio for tasks such as correcting LLM-generated UA-CBT stories and manual filtering. However, recognizing the need for a more streamlined and accessible method, we subsequently introduced a Telegram bot to simplify the process. A poll among our contributors regarding their preferred method of data annotation revealed a unanimous preference for the Telegram bot. To increase engagement, we incorporated simple gamification elements in the bot - transforming any button presses into animated emojis, which proved to be an effective strategy to maintain user interest and participation (Raftopoulos, 2015). Remarkably, this approach enabled a more rapid collection of data (compared to the same bot without gamification). This underscores the potential of this method as a valuable strategy for

tions being the letters/tokens of the word/sentence, but the actual evaluation involved simply comparing the predicted output with the exact expected ground truth value.

data annotation. Ultimately, the choice of platform should not be restricted to what we used; it heavily depends on the demographics.

6. Limitations

6.1. UP-Titles

Since the UP-Titles dataset was built from articles of a well-known online newspaper (eighth most cited source in Wikipedia in 2017 (Lewoniewski et al., 2017)), the already discussed issues of contamination/memorization apply to it: it's very likely that the articles are and/or will be part of the training data of LLMs. Most of the articles from the dataset involve the Russian-Ukrainian war, with predictable effects on the language used (both topic-wise and through the changes in the vocabulary (Synchak, 2023) in that context).

6.2. UA-CBT

Half of the stories were generated using GPT4 and half using Gemini, then all were piped through Gemini to improve grammar and consistency. This raises the question of encapsulation: testing a model on tasks generated (even partially) with its output would lead to inflated scores. GPT4's very high scores on this task would seem to confirm this, but its performance on pure-Gemini stories was just as high. Nevertheless, the fact that all of the stories were 'touched' by Gemini and half by both Gemini and GPT4 is context crucial for the interpretation of scores of either of these models on the dataset.

Due to the limited number of annotators, multiple questions based on the same story could have been shown to the same annotators, who could have memorized the token in the gap from a previous task instance. This could have contributed to a higher human baseline. The Telegram bot did not allow going back to an already answered question, so the inability to fix errors could have had the opposite effect. We don't believe either to have been significant.

7. Discussion

We acknowledge the potential risks associated with the datasets introduced, particularly their utility in enhancing AI-driven bots for malicious political influence on social media (Radivojevic et al., 2024) (Eady et al., 2023) (Stukal et al., 2017), especially during the ongoing war. We advocate for an open proactive approach to exploring various classifiers and AI methods for the detection of malicious instances. During the generation and human filtration of task instances (see Section 3.1.6), we found clear patterns in the errors. We think some of the errors found were specific to Ukrainian, and that leveraging them could be a promising avenue of future research parallel and complementary towards existing research focusing on language-independent bot detection. The influence of a native tongue on a second language, known as language interference, is established in the literature. If these patterns are different in humans (e.g. most bilingual speakers in Ukraine) and LLMs (trained on multilingual data containing a significant amount of Russian), this could become basis of a classifier.

We evaluated two models that were fine-tuned on Ukrainian datasets and/or instructions. Among these models, the Sherlock model demonstrated superior performance when compared to the vanilla Mistral-7B model. We believe a more thorough analysis using more models and different evaluation approaches would be beneficial and would confirm the finding that fine-tuning on Ukrainian data improves performance on Ukrainian tasks.

An additional avenue for future research would be to systematically evaluate models tuned on Russian language, and quantify the impact on the scores. Evaluating instruction-finetuned models in a way that takes advantage of it by using proper templates would allow deeper insights into this.

8. Acknowledgments

We are grateful to Daria Kravets, Mariia Tkachenko, Oleksii K., Lina Mykhailenko, @arturius453, and Viacheslav Kravchenko for their contribution to the datasets and human baselines creation.

9. Conclusion

This paper presents a significant stride towards enhancing language model performance in Ukrainian through Eval-UA-tion. By introducing novel datasets, we provide a comprehensive evaluation framework that assesses models' abilities. Our work highlights the essential need for linguistic diversity in AI, with a focus on Ukrainian as a case study. Despite acknowledging our approach's limitations, such as potential memorization and contamination risks, we suggest directions for future research to refine and broaden our methodologies. Our contributions aim to advance more inclusive and representative language technologies.

10. Bibliographical References

Syeda Nahida Akter, Zichun Yu, Aashiq Muhamed, Tianyue Ou, Alex Bäuerle, Ángel Alexander Cabrera, Krish Dholakia, Chenyan Xiong, and Graham Neubig. 2023. An In-depth Look at Gemini's Language Abilities. ArXiv:2312.11444 [cs].

Tiberiu Boros, Radu Chivereanu, Stefan Dumitrescu, and Octavian Purcaru. 2024. Fine-tuning and retrieval augmented generation for question answering using affordable large language models. In *Proceedings of the third ukrainian natural language processing workshop*, Torino, Italy. European Language Resources Association.

Bernard Comrie, Martin Haspelmath, and Balthasar Bickel. 2008. The Leipzig Glossing Rules: Conventions for interlinear morpheme-by-morpheme glosses. *Department of Linguistics of the Max Planck Institute for Evolutionary Anthropology & the Department of Linguistics of the University of Leipzig. Retrieved January*, 28:2010.

Gregory Eady, Tom Paskhalis, Jan Zilinsky, Richard Bonneau, Jonathan Nagler, and Joshua A. Tucker. 2023. Exposure to the Russian Internet Research Agency foreign influence campaign on Twitter in the 2016 US election and its relationship to attitudes and voting behavior. *Nature Communications*, 14(1):62.

Avia Efrat, Or Honovich, and Omer Levy. 2022. LMentry: A language model benchmark of elementary language tasks. Tex.copyright: Creative Commons Attribution 4.0 International.

Rating Group. 2022. The sixth national poll: The language issue in Ukraine (March 19th, 2022) — ratinggroup.ua.

Zishan Guo, Renren Jin, Chuang Liu, Yufei Huang, Dan Shi, Supryadi, Linhao Yu, Yan Liu, Jiaxuan Li, Bojian Xiong, and Deyi Xiong. 2023. Evaluating Large Language Models: A Comprehensive Survey. ArXiv:2310.19736 [cs].

Felix Hill, Antoine Bordes, Sumit Chopra, and Jason Weston. 2015. The goldilocks principle: Reading children's books with explicit memory representations. Tex.copyright: arXiv.org perpetual, non-exclusive license.

Bogdan Ivanyuk-Skulskiy, Anton Zaliznyi, Oleksand Reshetar, Oleksiy Protsyk, Bohdan Romanchuk, and Vladyslav Shpihanovych. 2021. ua$_{d}$atasets: a collection of Ukrainian language datasets.

Pratik Joshi, Sebastin Santy, Amar Budhiraja, Kalika Bali, and Monojit Choudhury. 2020. The state and fate of linguistic diversity and inclusion in the NLP world. *CoRR*, abs/2004.09095. ArXiv: 2004.09095 tex.bibsource: dblp computer science bibliography, https://dblp.org tex.biburl: https://dblp.org/rec/journals/corr/abs-2004-09095.bib tex.timestamp: Wed, 22 Apr 2020 12:57:53 +0200.

Volodymyr Kulyk. 2018. Shedding Russianness, recasting Ukrainianness: The post-Euromaidan dynamics of ethnonational identifications in Ukraine. *Post-Soviet Affairs*, 34(2-3):119–138. Publisher: Taylor & Francis.

Viet Dac Lai, Nghia Trung Ngo, Amir Pouran Ben Veyseh, Hieu Man, Franck Dernoncourt, Trung Bui, and Thien Huu Nguyen. 2023. ChatGPT Beyond English: Towards a Comprehensive Evaluation of Large Language Models in Multilingual Learning. ArXiv:2304.05613 [cs].

Włodzimierz Lewoniewski, Krzysztof Węcel, and Witold Abramowicz. 2017. Analysis of references across wikipedia languages. pages 561–573.

Meta. 2022. List of Wikipedias/Table2 — Meta, discussion about wikimedia projects.

Daniel Racek, Brittany I. Davidson, Paul W. Thurner, Xiao Xiang Zhu, and Göran Kauermann. 2024. The Russian war in Ukraine increased Ukrainian language use on social media. *Communications Psychology*, 2(1):1.

Kristina Radivojevic, Nicholas Clark, and Paul Brenner. 2024. LLMs Among Us: Generative AI Participating in Digital Discourse. ArXiv:2402.07940 [cs].

Marigo Raftopoulos. 2015. How enterprises play: Towards a taxonomy for enterprise gamification.

Manley Roberts, Himanshu Thakur, Christine Herlihy, Colin White, and Samuel Dooley. 2023. Data Contamination Through the Lens of Time. ArXiv:2310.10628 [cs].

Joshua Robinson, Christopher Michael Rytting, and David Wingate. 2023. Leveraging Large Language Models for Multiple Choice Question Answering. ArXiv:2210.12353 [cs].

Alessandro Seganti, Klaudia Firląg, Helena Skowronska, Michał Satława, and Piotr Andruszkiewicz. 2021. Multilingual entity and relation extraction dataset and model. In

Proceedings of the 16th conference of the european chapter of the association for computational linguistics: Main volume, pages 1946–1955, Online. Association for Computational Linguistics.

Denis Stukal, Sergey Sanovich, Richard Bonneau, and Joshua A. Tucker. 2017. Detecting Bots on Russian Political Twitter. *Big Data*, 5(4):310–324.

Vasyl Starkoand Olena Synchak. 2023. Feminine personal nouns in ukrainian: Dynamics in a corpus.

Oleksiy Syvokon and Olena Nahorna. 2022. UA-GEC: Grammatical Error Correction and Fluency Corpus for the Ukrainian Language. ArXiv:2103.16997 [cs].

Gilles Sérasset. 2015. DBnary: Wiktionary as a Lemon-based multilingual lexical resource in RDF. *Semantic Web*, 6(4):355–361.

Wikipedia contributors. 2023. Languages used on the internet — Wikipedia, the free encyclopedia.

LiBERTa: Advancing Ukrainian Language Modeling through Pre-training from Scratch

Mykola Haltiuk, Aleksander Smywiński-Pohl
AGH University of Krakow, Enelpol
mhaltiuk@student.agh.edu.pl, apohllo@agh.edu.pl

Abstract

Recent advancements in Natural Language Processing (NLP) have spurred remarkable progress in language modeling, predominantly benefiting English. While Ukrainian NLP has long grappled with significant challenges due to limited data and computational resources, recent years have seen a shift with the emergence of new corpora, marking a pivotal moment in addressing these obstacles. This paper introduces LiBERTa Large, the inaugural BERT Large model pre-trained entirely from scratch only on Ukrainian texts. Leveraging extensive multilingual text corpora, including a substantial Ukrainian subset, LiBERTa Large establishes a foundational resource for Ukrainian NLU tasks. Our model outperforms existing multilingual and monolingual models pre-trained from scratch for Ukrainian, demonstrating competitive performance against those relying on cross-lingual transfer from English. This achievement underscores our ability to achieve superior performance through pre-training from scratch with additional enhancements, obviating the need to rely on decisions made for English models to efficiently transfer weights. We establish LiBERTa Large as a robust baseline, paving the way for future advancements in Ukrainian language modeling.

Keywords: Ukrainian, LiBERTa, Pre-training from Scratch, Language Models, Natural Language Understanding, Transformers

1. Introduction

In recent years, there has been remarkable progress in language modeling, evidenced by the multitude of research papers emerging annually. This progress stems from a variety of advancements, including novel architectural improvements (Shaw et al., 2018; Su et al., 2021; He et al., 2020; Fedus et al., 2021), innovative training objectives (Clark et al., 2020; Raffel et al., 2019; Joshi et al., 2020; Wang et al., 2019b), different tokenization approaches (Xue et al., 2022), methods for data curation (Gunasekar et al., 2023), and other refinements, consistently enhancing state-of-the-art results, particularly for English.

However, the field of natural language processing (NLP) in Ukrainian has encountered substantial obstacles compared to its English counterpart, primarily due to limited data availability and computational resources. Unlike English, which benefits from abundant datasets and robust computing infrastructure, Ukrainian has historically lacked comprehensive resources essential for robust NLP research and development.

Until recently, NLP researchers working with Ukrainian had to resort to cross-lingual transfer learning due to the scarcity of substantial Ukrainian text corpora suitable for pre-training monolingual models from scratch. However, with the release of datasets like CulturaX (Nguyen et al., 2023), we are venturing to train a BERT Large model entirely from scratch in Ukrainian. Our goal is to ascertain whether the available resources now enable us to compete with models transferred from English using sophisticated techniques.

To ensure a fair comparison, we adopt an almost vanilla RoBERTa (Liu et al., 2019) pre-training setup, encompassing both objective and architecture, thus mitigating potential confounding factors that could disrupt our comparison.

In this paper, we make several contributions:

- We introduce LiBERTa Large – the first BERT-like Large model pre-trained from scratch for Ukrainian. Leveraging multilingual text corpora containing a substantial subset of documents in Ukrainian, we provide a foundational resource for natural language understanding tasks.

- Our model achieves state-of-the-art performance compared to existing multilingual alternatives and monolingual language models for Ukrainian that are pre-trained from scratch on multiple downstream tasks. Additionally, it exhibits competitive results against models that rely on the cross-lingual transfer of heavily trained English models.

- By establishing this baseline, we pave the way for future research in Ukrainian language modeling from scratch, enabling researchers to leverage the latest advancements to further enhance performance on downstream tasks.

25 May, 2024. © 2024 ELRA Language Resource Association: CC BY-NC 4.0

2. Related Work

The Transformer architecture, introduced by Vaswani et al. (2017) for Machine Translation, marked a significant advancement by showcasing the effectiveness of attention mechanisms over traditional recurrent networks. Building upon this, Radford et al. (2018) extended the Transformer architecture to Natural Language Understanding (NLU) tasks, demonstrating its adaptability through pre-training with causal language modeling and subsequent fine-tuning for specific tasks, thereby achieving state-of-the-art results.

Devlin et al. (2019) further enhanced Transformer-based models with bidirectionality, employing Masked Language Modeling (MLM) and Next Sentence Prediction (NSP) objectives, leading to substantial performance improvements over unidirectional models. It was observed that scaling the model size consistently enhanced performance across various downstream tasks. Subsequent studies suggested alternative strategies for improvement, such as omitting NSP in favor of data augmentation, dynamic masking, increased batch sizes, and training on longer sequences (Liu et al., 2019).

Continued research efforts focused on refining pre-training objectives and enhancing model architectures. Modifications to the Masked Language Modeling objective included predicting token spans (Joshi et al., 2020) and employing binary classification through Replaced Token Detection (RTD) (Clark et al., 2020). Additionally, innovations such as relative positional encoding (Shaw et al., 2018) and disentangled attention mechanisms contributed to further improvements (He et al., 2020, 2021).

While initial efforts primarily concentrated on English, subsequent research expanded to encompass other languages. Multilingual models like mBERT (Devlin et al., 2019), XLM (Lample and Conneau, 2019), and XLM-RoBERTa (XLM-R) (Conneau et al., 2020) achieved state-of-the-art results across numerous low-resource languages. However, increasing the number of languages in multilingual models often led to performance degradation on language-specific tasks, highlighting the challenge known as the curse of multilinguality.

Consequently, efforts turned towards developing monolingual models tailored to specific languages, resulting in superior performance for languages such as French (Martin et al., 2020; Le et al., 2020), German (Chan et al., 2020), Dutch (de Vries et al., 2019; Delobelle et al., 2020), and Finnish (Virtanen et al., 2019). The release of HerBERT (Mroczkowski et al., 2021) pre-trained for Polish was particularly noteworthy, given the linguistic proximity to Ukrainian (Beaufils and Tomin, 2020).

With the advent of increasingly powerful Large Language Models (LLMs), questions arose regarding the necessity of pre-training BERT-like models. Hadeliya and Kajtoch (2023) investigated In-Context Learning (ICL) approaches in Polish for models like Llama 2 (Touvron et al., 2023), comparing them with full fine-tuning of models like HerBERT. Their findings indicated that full fine-tuning consistently outperformed ICL approaches across various downstream tasks. Notably, the Ukrainian portion of datasets used for LLM pre-training either matched or significantly lagged behind their Polish counterparts in terms of representation (Touvron et al., 2023; Chowdhery et al., 2022).

Recent years have witnessed notable advancements in the development of Ukrainian language processing, traditionally considered low-resource. These advancements were facilitated by the release of multi- and monolingual text corpora (Wenzek et al., 2020; Conneau et al., 2020; Chaplynskyi, 2023; Nguyen et al., 2023), enabling the training of larger-scale models. Earlier initiatives aimed at developing Ukrainian language models by Radchenko (2020) and Schweter (2020), further referred to as Ukr-RoBERTa and Ukr-ELECTRA respectively, represent crucial foundational steps in monolingual language modeling for Ukrainian. These efforts underscored the potential of this domain, demonstrating improved performance compared to multilingual models like mBERT. In addition to the aforementioned advancements, there has also been notable progress in the Causal Language Models training (Kyrylov and Chaplynskyi, 2023).

A recent breakthrough in Ukrainian language processing emerged with the introduction of the WECHSEL embedding initialization method (Minixhofer et al., 2022). This facilitated efficient cross-lingual transfer during the pre-training of WECHSEL-RoBERTa, leading to performance enhancements that surpassed multilingual baselines like XLM-R in Natural Language Understanding (NLU) tasks. This development marks a significant stride forward in Ukrainian language representation learning and processing capabilities.

3. LiBERTa

In this section, we outline the comprehensive steps taken to pre-train the LiBERTa Large model for the Ukrainian language.

3.1. Training and Validation Data

We carefully selected two multilingual text corpora, namely CulturaX and CC-100, from which we extracted the Ukrainian subset without any additional cleaning or deduplication. To manage data effi-

Tokenizer	Size	Avg.	Hits
XLM-RoBERTa	250K	1.739	54.46%
Ukr-RoBERTa	52K	1.846	42.16%
WECHSEL-RoBERTa	50K	1.866	40.89%
Ukr-ELECTRA	32K	1.443	69.89%
LiBERTa	32K	**1.442**	**70.02**%

Table 1: Evaluation results of tokenizers for Ukrainian. *Size* is the size of the vocabulary, *Avg.* is the average tokens per word ratio, and *Hits* is the percent of words directly present in the vocabulary.

ciently during training, we leveraged the Datasets library (Lhoest et al., 2021).

3.1.1. CulturaX

CulturaX, a compilation of mC4 (Raffel et al., 2019) and OSCAR (Ortiz Su'arez et al., 2020; Ortiz Su'arez et al., 2019) corpora, serves as an invaluable resource for our endeavor. The Ukrainian subset of CulturaX comprises over 38 billion tokens distributed across 44 million documents. The inclusion of lengthy documents within this corpus facilitates the model's capacity to capture long-range dependencies, rendering it an apt choice for pre-training.

3.1.2. CC-100

CC-100, a multilingual text corpus sourced from Wikipedia and CommonCrawl, was processed following the CCNet[1] methodology. The Ukrainian segment of CC-100 encompasses 6.5 billion tokens, equivalent to 84 GiB of data[2]. This corpus primarily aids in training the tokenizer.

3.1.3. Ukrainian UD

The Gold standard Universal Dependencies corpus for Ukrainian (Ukrainian UD) (Kotsyba et al., 2018) is a highly diverse and meticulously curated collection of high-quality text documents in Ukrainian. It comprises over 100,000 tokens, providing a robust foundation for reliable and multi-faceted evaluations of Masked Language Modeling.

3.2. Tokenizer

We trained the Byte Pair Encoding (BPE) (Gage, 1994) tokenizer on the subset of CC-100 using SentencePiece (Kudo and Richardson, 2018) with byte fallback for robustness. The training dataset comprised 10 million paragraphs, amounting to 2.5 GiB of raw uncompressed text. The resulting tokenizer features a vocabulary of 32,000 cased tokens. Prior to tokenization, input texts are being pre-tokenized based on Unicode script boundaries and manually defined punctuation symbols.

Evaluation of the tokenizer's performance, conducted against XLM-R's tokenizer trained on a multilingual corpus and other Ukrainian language models, was based on the Ukrainian UD corpus. Notably, our tokenizer, on par with Ukr-ELECTRA's, despite possessing the smallest vocabulary, yields the least subtokens per word and achieves the highest ratio of words represented as a single subtoken in its vocabulary according to the metrics presented in Table 1. Other tokenizers appear to be less suited for the Ukrainian language according to our validation corpus.

Additionally, tokenization was performed on nearly 50 atypical words encompassing named entities, dialectisms, domain-specific terminology, slang, swear words, neologisms, anglicisms, words with orthographic errors, as well as English or Polish words. Results indicate a consistent performance across all tokenizers, albeit XLM-R's tokenizer exhibits superior handling of English words, while monolingual Ukrainian tokenizers demonstrate poor performance in English contexts.

3.3. Model's Architecture

The architecture of LiBERTa aligns with the original BERT Large, comprising 24 layers, 16 attention heads, and 1024 hidden dimensions. We employ absolute positional embeddings with a maximum sequence length of 512.

Implementation is facilitated through the Transformers library (Wolf et al., 2019) by HuggingFace, integrating Flash Attention (Dao et al., 2022) for efficient processing. Model weights are initialized randomly using PyTorch (Paszke et al., 2019).

3.4. Optimization

Optimization entails the utilization of the AdamW optimizer (Loshchilov and Hutter, 2017) coupled with a cosine learning rate schedule with a warm-up. Following RoBERTa's paradigm, the training objective is structured around Masked Language Modeling, wherein there is a 15% probability of a token being replaced with a `<mask>` token, a random token, or remaining unchanged.

3.5. Pre-training Process

LiBERTa was pre-trained with hyperparameters, as delineated in Table 2. The training duration spanned 39 hours, leveraging a computational

[1] https://github.com/facebookresearch/cc_net

[2] We believe there is a mistake in the original resource, reporting 6.5 million tokens. That would not comply with the number of tokens per 1 GiB ratio in other languages with Cyrillic script.

Hyperparameter	Value
Peak Learning Rate	2e-4
Warm-up Steps	5K
Learning Rate Decay	Cosine
Effective Batch Size	1024
Batch Size per GPU	32
Gradient Accumulation Steps	4
Max Steps	85K
Weight Decay	0.01
Adam ϵ	1e-8
Adam β_1	0.9
Adam β_2	0.999
Gradient Clipping	1.0
Gradient Clipping Algorithm	L2

Table 2: The hyperparameters used for pretraining LiBERTa Large. The remaining parameters are the defaults from the Huggingface library.

node equipped with 8 NVIDIA A100-SXM4-40GB GPUs. Distributed Data Parallel (DDP) strategy (Li et al., 2020) was employed to efficiently distribute training data and gradients across the GPUs. `bfloat16` adaptive mixed precision was used to enhance throughput.

To accommodate longer documents present in the corpus, they were partitioned into multiple chunks, each comprising 510 subtokens besides `<cls>` at the beginning and `<sep>` at the end. The final chunk in a document was padded to match the longest sequence in the batch.

Throughout the training process, validation was conducted to assess metrics such as loss, perplexity, and Masked Language Modeling Accuracy using the Ukrainian UD.

4. Evaluation

In this section, we present the evaluation tasks utilized to assess LiBERTa's performance in comparison to existing models for Ukrainian language understanding.

4.1. Tasks

Given the absence of a standardized Natural Language Understanding benchmark for the Ukrainian language, we delineate the downstream tasks employed for evaluating our model.

4.1.1. NER-UK

NER-UK, sourced from lang-uk[3], comprises over 6.7K named entities spanning 217K tokens from the BrUK corpus of contemporary Ukrainian[4]. Evaluation is conducted via micro-averaged F1 Score as calculated by `seqeval` (Nakayama, 2018).

4.1.2. WikiANN

WikiANN (Pan et al., 2017; Rahimi et al., 2019), a multilingual named entity recognition dataset, encompasses Wikipedia articles. The Ukrainian subset comprises over 54K named entities across 318K tokens. Notably, the average document is quite short, often a single sentence with 8 tokens and containing only 1-2 named entities. Consequently, this emphasizes how well the common knowledge is embedded into the model besides its ability to infer from the context. Evaluation employs micro-averaged F1 Score via `seqeval`.

4.1.3. Part-of-Speech Tagging

Universal Dependencies (Nivre et al., 2017) is a multilingual dataset with a consistent annotation of grammar (parts of speech, morphological features, and syntactic dependencies). In our evaluation, we have concentrated on the Ukrainian Part-of-Speech (POS) tagging. For this task, the metric used for evaluation is accuracy.

4.1.4. Ukrainian News Classification

This task (Panchenko, 2021; Panchenko et al., 2022) involves a corpus of news articles gathered from popular Ukrainian media outlets. It is an unbalanced text classification task focused on predicting news publication sources. Data preprocessing ensures the removal of implicit data leakages, with mentions of sources being replaced by a special token. Evaluation utilizes macro-averaged F1 Score to mitigate class imbalance effects.

4.2. Results

We compare LiBERTa's performance against the results reported[5] by Minixhofer et al. (2022) for NER-UK, WikiANN, and POS tagging, as shown in Table 3.

LiBERTa demonstrates comparable performance to the previous state-of-the-art in NER-UK (i.e. WECHSEL-RoBERTa), exhibiting a slight performance improvement (+0.03 pp.). Interestingly, for this task, the second large model XLM-R achieves results worse than all the base models. It also has the highest variation. This result underscores the necessity for training language-specific models since both WECHSEL-RoBERTa and LiBERTa have lower variance.

Conversely, LiBERTa's performance on WikiANN is worse than all the other models, besides XLM-R

[3] https://lang.org.ua/uk/

[4] https://github.com/brown-uk/corpus

[5] https://huggingface.co/benjamin/roberta-large-wechsel-ukrainian

Model	NER-UK *micro-f1*	WikiANN *micro-f1*	UD POS *acc*	News *macro-f1*
Base Models				
XLM-R	90.86 (0.81)†	92.27 (0.09)†	98.45 (0.07)†	–
WECHSEL-RoBERTa	90.81 (1.51)†	92.98 (0.12)†	98.57 (0.03)†	–
Ukr-ELECTRA	90.43 (1.29)†	92.99 (0.11)†	98.59 (0.06)†	–
Large Models				
XLM-R	90.16 (2.98)†	92.92 (0.19)†	98.71 (0.04)†	95.13 (0.49)
WECHSEL-RoBERTa	91.24 (1.16)†	**93.22 (0.17)**†	**98.74 (0.06)**†	**96.48 (0.09)**
LiBERTa	**91.27 (1.22)**	92.50 (0.07)	98.62 (0.08)	95.44 (0.04)

Table 3: Evaluation results on the downstream tasks as a mean of 5 runs with different seeds. The values in the parentheses denote the standard deviation of the metric values. $\cdot^\dagger$ denotes the results reported by Minixhofer et al. (2022).

base. This is interesting since even though the task is the same as the first task, the average performance on this dataset for all the models is higher than in the first task. This discrepancy may arise from the dataset's nature, characterized by short sentences and reliance on Wikipedia as the only knowledge source. Multilingual models such as XLM-R are typically trained on Wikipedia since the data is of high quality, and it is very easy to make sure it contains mostly texts in a given language. But the names on Wikipedia are a mix of language-specific and international (mostly English) words. LiBERTa tokenizer was trained mostly on Ukrainian texts and the model was trained only for 1 epoch. This result indicates that it might be reasonable to include English texts when training the tokenizer, to better process anglicisms in Ukrainian and strikes the importance of longer pre-training.

For the Part-of-Speech tagging task, LiBERTa achieves marginally inferior (-0.12 pp. vs. WECHSEL-RoBERTa) results compared to the current state-of-the-art. The results for this task are very high for all models, which indicates it is pretty simple to tag POS in Ukrainian. The differences between the models might, in fact, be random and the models might just learn the errors in the annotation. Anyway, the results show that the model is able to learn POS tagging very well, and it stresses the importance of including the other tasks (morphological feature prediction, lemmatization) in future work since these tasks might be harder for the models.

While not exhaustively evaluated against all available models, LiBERTa's performance on the Ukrainian News Classification dataset (as shown in the last column of Table 3) surpasses the XLM-R Large (+0.31pp.), albeit with inferior performance compared to WECHSEL-RoBERTa.

5. Conclusion

In this study, we present LiBERTa Large, an encoder-only language model for Ukrainian, trained entirely from scratch. Our model demonstrates competitive performance on various Natural Language Understanding (NLU) tasks, rivaling the current state-of-the-art models. Through our exploration, we have observed that leveraging new text corpora and employing a straightforward BERT architecture with a Masked Language Modeling objective enables our model to effectively compete with other models, which are exploiting cross-lingual transfer of robustly pre-trained English models like RoBERTa (trained for about 40 epochs on 160 GiB of text).

The development of LiBERTa Large establishes a novel baseline for future research endeavors, opening avenues for investigating diverse architectural enhancements, optimization objectives, and data curation methodologies. Prior to this work, the scarcity of data or computational resources often necessitated reliance on decisions made for existing language models, such as RoBERTa, to facilitate effective cross-lingual weight transfer. However, our findings indicate promising prospects for the development of language models trained from scratch, thereby reducing the dependency on pre-existing models and enabling greater flexibility in model design and training.

Throughout our investigation, we encountered challenges in evaluating and comparing Ukrainian language models. The absence of a standardized benchmark, akin to GLUE and SuperGLUE for English (Wang et al., 2018, 2019a) or KLEJ for Polish (Rybak et al., 2020), renders comprehensive and consistent model comparisons across diverse NLU tasks, including Natural Language Inference (NLI), Extractive Question Answering (EQA), and Machine Reading Comprehension (MRC), impossible.

Additionally, we encountered instances of modal collapse during our pre-training experiments, particularly evident while training on shorter sequences, leading to a huge spike in loss and the inability to continue the experiment. Notably, the model tended to generate commas for every token in the input sequence. Mitigating modal collapse required the implementation of techniques such as gradient clipping, adjusting input sequence lengths, and decreasing the peak learning rate to ensure the stability and convergence of the training process.

We believe our reported results will inspire NLP researchers to explore pre-training Ukrainian language models from scratch, leveraging novel techniques to establish a new state-of-the-art.

Limitations

One limitation of our study lies in the scope of our evaluation, which may not cover all available models, potentially missing alternative approaches or architectures that could yield superior results. Resource constraints, including computational and time limitations, may have prevented us from fully exploring LiBERTa's potential, leaving room for further optimization and refinement.

Furthermore, our training dataset, CulturaX, may have included biases inherent in its collection process or source material. These biases could affect the model's understanding and representation of certain linguistic patterns or social phenomena. Further investigation into the nature and extent of these biases is warranted to enhance the model's robustness and fairness in real-world applications.

Acknowledgments

This research was supported by the Polish National Centre for Research and Development project POIR.01.02.00-00-0154/16 titled „Big Data Game Content Engine: mBaaS game engine enabling access to Big Data as game content for developers". We gratefully acknowledge Polish high-performance computing infrastructure PLGrid (ACK Cyfronet AGH) for providing computer facilities and support within computational grant no. PLG/2023/016304.

Bibliographical References

Vincent Beaufils and Johannes Tomin. 2020. Stochastic approach to worldwide language classification: the signals and the noise towards long-range exploration.

Branden Chan, Stefan Schweter, and Timo Möller. 2020. German's next language model. In *Proceedings of the 28th International Conference on Computational Linguistics*, pages 6788–6796, Barcelona, Spain (Online). International Committee on Computational Linguistics.

Dmytro Chaplynskyi. 2023. Introducing UberText 2.0: A corpus of Modern Ukrainian at scale. In *Proceedings of the Second Ukrainian Natural Language Processing Workshop (UNLP)*, pages 1–10, Dubrovnik, Croatia. Association for Computational Linguistics.

Aakanksha Chowdhery, Sharan Narang, Jacob Devlin, Maarten Bosma, Gaurav Mishra, Adam Roberts, Paul Barham, Hyung Won Chung, Charles Sutton, Sebastian Gehrmann, Parker Schuh, Kensen Shi, Sasha Tsvyashchenko, Joshua Maynez, Abhishek Rao, Parker Barnes, Yi Tay, Noam M. Shazeer, Vinodkumar Prabhakaran, Emily Reif, Nan Du, Benton C. Hutchinson, Reiner Pope, James Bradbury, Jacob Austin, Michael Isard, Guy Gur-Ari, Pengcheng Yin, Toju Duke, Anselm Levskaya, Sanjay Ghemawat, Sunipa Dev, Henryk Michalewski, Xavier García, Vedant Misra, Kevin Robinson, Liam Fedus, Denny Zhou, Daphne Ippolito, David Luan, Hyeontaek Lim, Barret Zoph, Alexander Spiridonov, Ryan Sepassi, David Dohan, Shivani Agrawal, Mark Omernick, Andrew M. Dai, Thanumalayan Sankaranarayana Pillai, Marie Pellat, Aitor Lewkowycz, Erica Moreira, Rewon Child, Oleksandr Polozov, Katherine Lee, Zongwei Zhou, Xuezhi Wang, Brennan Saeta, Mark Díaz, Orhan Firat, Michele Catasta, Jason Wei, Kathleen S. Meier-Hellstern, Douglas Eck, Jeff Dean, Slav Petrov, and Noah Fiedel. 2022. Palm: Scaling language modeling with pathways. *J. Mach. Learn. Res.*, 24:240:1–240:113.

Kevin Clark, Minh-Thang Luong, Quoc V. Le, and Christopher D. Manning. 2020. Electra: Pre-training text encoders as discriminators rather than generators. In *International Conference on Learning Representations*.

Alexis Conneau, Kartikay Khandelwal, Naman Goyal, Vishrav Chaudhary, Guillaume Wenzek, Francisco Guzmán, Edouard Grave, Myle Ott, Luke Zettlemoyer, and Veselin Stoyanov. 2020. Unsupervised cross-lingual representation learning at scale. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 8440–8451, Online. Association for Computational Linguistics.

Tri Dao, Daniel Y Fu, Stefano Ermon, Atri Rudra, and Christopher Re. 2022. Flashattention: Fast and memory-efficient exact attention with IO-awareness. In *Advances in Neural Information Processing Systems*.

Wietse de Vries, Andreas van Cranenburgh, Arianna Bisazza, Tommaso Caselli, Gertjan van Noord, and Malvina Nissim. 2019. Bertje: A dutch bert model. Cite arxiv:1912.09582.

Pieter Delobelle, Thomas Winters, and Bettina Berendt. 2020. RobBERT: a Dutch RoBERTa-based Language Model. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 3255–3265, Online. Association for Computational Linguistics.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4171–4186, Minneapolis, Minnesota. Association for Computational Linguistics.

William Fedus, Barret Zoph, and Noam Shazeer. 2021. Switch transformers: Scaling to trillion parameter models with simple and efficient sparsity. *CoRR*, abs/2101.03961.

Philip Gage. 1994. A new algorithm for data compression. *The C Users Journal archive*, 12:23–38.

Suriya Gunasekar, Yi Zhang, Jyoti Aneja, Caio Cesar Teodoro Mendes, Allison Del Giorno, Sivakanth Gopi, Mojan Javaheripi, Piero C. Kauffmann, Gustavo de Rosa, Olli Saarikivi, Adil Salim, S. Shah, Harkirat Singh Behl, Xin Wang, Sébastien Bubeck, Ronen Eldan, Adam Tauman Kalai, Yin Tat Lee, and Yuan-Fang Li. 2023. Textbooks are all you need. *ArXiv*, abs/2306.11644.

Tsimur Hadeliya and Dariusz Kajtoch. 2023. Evaluation of few-shot learning capabilities in polish language models. In *ML in PL Conference 2023*, Warsaw, Poland.

Pengcheng He, Jianfeng Gao, and Weizhu Chen. 2021. Debertav3: Improving deberta using electra-style pre-training with gradient-disentangled embedding sharing. *CoRR*, abs/2111.09543.

Pengcheng He, Xiaodong Liu, Jianfeng Gao, and Weizhu Chen. 2020. Deberta: Decoding-enhanced bert with disentangled attention. *CoRR*, abs/2006.03654.

Mandar Joshi, Danqi Chen, Yinhan Liu, Daniel S. Weld, Luke Zettlemoyer, and Omer Levy. 2020. SpanBERT: Improving pre-training by representing and predicting spans. *Transactions of the Association for Computational Linguistics*, 8:64–77.

Natalia Kotsyba, Bohdan Moskalevskyi, and Mykhailo Romanenko. 2018. Gold standard Universal Dependencies corpus for Ukrainian.

Taku Kudo and John Richardson. 2018. SentencePiece: A simple and language independent subword tokenizer and detokenizer for neural text processing. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 66–71, Brussels, Belgium. Association for Computational Linguistics.

Volodymyr Kyrylov and Dmytro Chaplynskyi. 2023. GPT-2 metadata pretraining towards instruction finetuning for Ukrainian. In *Proceedings of the Second Ukrainian Natural Language Processing Workshop (UNLP)*, pages 32–39, Dubrovnik, Croatia. Association for Computational Linguistics.

Guillaume Lample and Alexis Conneau. 2019. Cross-lingual language model pretraining. Cite arxiv:1901.07291.

Hang Le, Loïc Vial, Jibril Frej, Vincent Segonne, Maximin Coavoux, Benjamin Lecouteux, Alexandre Allauzen, Benoit Crabbé, Laurent Besacier, and Didier Schwab. 2020. FlauBERT: Unsupervised language model pre-training for French. In *Proceedings of the Twelfth Language Resources and Evaluation Conference*, pages 2479–2490, Marseille, France. European Language Resources Association.

Quentin Lhoest, Albert Villanova del Moral, Yacine Jernite, Abhishek Thakur, Patrick von Platen, Suraj Patil, Julien Chaumond, Mariama Drame, Julien Plu, Lewis Tunstall, Joe Davison, Mario Šaško, Gunjan Chhablani, Bhavitvya Malik, Simon Brandeis, Teven Le Scao, Victor Sanh, Canwen Xu, Nicolas Patry, Angelina McMillan-Major, Philipp Schmid, Sylvain Gugger, Clément Delangue, Théo Matussière, Lysandre Debut, Stas Bekman, Pierric Cistac, Thibault Goehringer, Victor Mustar, François Lagunas, Alexander Rush, and Thomas Wolf. 2021. Datasets: A community library for natural language processing. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 175–184, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Shen Li, Yanli Zhao, Rohan Varma, Omkar Salpekar, Pieter Noordhuis, Teng Li, Adam Paszke, Jeff Smith, Brian Vaughan, Pritam Damania, and Soumith Chintala. 2020. Pytorch distributed: Experiences on accelerating data parallel training. *CoRR*, abs/2006.15704.

Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov.

2019. Roberta: A robustly optimized bert pretraining approach. Cite arxiv:1907.11692.

Ilya Loshchilov and Frank Hutter. 2017. Fixing weight decay regularization in adam. *CoRR*, abs/1711.05101.

Louis Martin, Benjamin Muller, Pedro Javier Ortiz Suárez, Yoann Dupont, Laurent Romary, Éric de la Clergerie, Djamé Seddah, and Benoît Sagot. 2020. CamemBERT: a tasty French language model. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 7203–7219, Online. Association for Computational Linguistics.

Benjamin Minixhofer, Fabian Paischer, and Navid Rekabsaz. 2022. WECHSEL: Effective initialization of subword embeddings for cross-lingual transfer of monolingual language models. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 3992–4006, Seattle, United States. Association for Computational Linguistics.

Robert Mroczkowski, Piotr Rybak, Alina Wróblewska, and Ireneusz Gawlik. 2021. HerBERT: Efficiently pretrained transformer-based language model for Polish. In *Proceedings of the 8th Workshop on Balto-Slavic Natural Language Processing*, pages 1–10, Kiyv, Ukraine. Association for Computational Linguistics.

Hiroki Nakayama. 2018. seqeval: A python framework for sequence labeling evaluation. Software available from https://github.com/chakki-works/seqeval.

Thuat Nguyen, Chien Van Nguyen, Viet Dac Lai, Hieu Man, Nghia Trung Ngo, Franck Dernoncourt, Ryan Rossi, and Thien Huu Nguyen. 2023. Culturax: A cleaned, enormous, and multilingual dataset for large language models in 167 languages. *ArXiv*, abs/2309.09400.

Joakim Nivre, Daniel Zeman, Filip Ginter, and Francis Tyers. 2017. Universal Dependencies. In *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics: Tutorial Abstracts*, Valencia, Spain. Association for Computational Linguistics.

Pedro Javier Ortiz Su'arez, Laurent Romary, and Benoit Sagot. 2020. A monolingual approach to contextualized word embeddings for mid-resource languages. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 1703–1714, Online. Association for Computational Linguistics.

Pedro Javier Ortiz Su'arez, Benoit Sagot, and Laurent Romary. 2019. Asynchronous pipelines for processing huge corpora on medium to low resource infrastructures. Proceedings of the Workshop on Challenges in the Management of Large Corpora (CMLC-7) 2019. Cardiff, 22nd July 2019, pages 9 – 16, Mannheim. Leibniz-Institut f"ur Deutsche Sprache.

Xiaoman Pan, Boliang Zhang, Jonathan May, Joel Nothman, Kevin Knight, and Heng Ji. 2017. Cross-lingual name tagging and linking for 282 languages. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1946–1958, Vancouver, Canada. Association for Computational Linguistics.

Dmytro Panchenko. 2021. Ukrainian News Classification.

Dmytro Panchenko, Daniil Maksymenko, Olena Turuta, Mykyta Luzan, Stepan Tytarenko, and Oleksii Turuta. 2022. Ukrainian news corpus as text classification benchmark. In *ICTERI 2021 Workshops*, pages 550–559, Cham. Springer International Publishing.

Adam Paszke, Sam Gross, Francisco Massa, Adam Lerer, James Bradbury, Gregory Chanan, Trevor Killeen, Zeming Lin, Natalia Gimelshein, Luca Antiga, Alban Desmaison, Andreas Köpf, Edward Z. Yang, Zach DeVito, Martin Raison, Alykhan Tejani, Sasank Chilamkurthy, Benoit Steiner, Lu Fang, Junjie Bai, and Soumith Chintala. 2019. Pytorch: An imperative style, high-performance deep learning library. *CoRR*, abs/1912.01703.

Vitalii Radchenko. 2020. Ukrainian Roberta.

Alec Radford, Karthik Narasimhan, Tim Salimans, and Ilya Sutskever. 2018. Improving language understanding by generative pre-training.

Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J. Liu. 2019. Exploring the limits of transfer learning with a unified text-to-text transformer. *CoRR*, abs/1910.10683.

Afshin Rahimi, Yuan Li, and Trevor Cohn. 2019. Massively multilingual transfer for NER. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 151–164, Florence, Italy. Association for Computational Linguistics.

Piotr Rybak, Robert Mroczkowski, Janusz Tracz, and Ireneusz Gawlik. 2020. KLEJ: comprehensive benchmark for polish language understanding. *CoRR*, abs/2005.00630.

Stefan Schweter. 2020. Ukrainian ELECTRA model.

Peter Shaw, Jakob Uszkoreit, and Ashish Vaswani. 2018. Self-attention with relative position representations. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 2 (Short Papers)*, pages 464–468, New Orleans, Louisiana. Association for Computational Linguistics.

Jianlin Su, Yu Lu, Shengfeng Pan, Bo Wen, and Yunfeng Liu. 2021. Roformer: Enhanced transformer with rotary position embedding. *CoRR*, abs/2104.09864.

Hugo Touvron, Louis Martin, Kevin R. Stone, Peter Albert, Amjad Almahairi, Yasmine Babaei, Nikolay Bashlykov, Soumya Batra, Prajjwal Bhargava, Shruti Bhosale, D. Bikel, Lukas Blecher, Cristian Cantón Ferrer, Moya Chen, Guillem Cucurull, David Esiobu, Jude Fernandes, Jeremy Fu, Wenyin Fu, Brian Fuller, Cynthia Gao, Vedanuj Goswami, Naman Goyal, A. Hartshorn, Saghar Hosseini, Rui Hou, Hakan Inan, Marcin Kardas, Viktor Kerkez, Madian Khabsa, Isabel M. Kloumann, A. Korenev, Punit Singh Koura, Marie-Anne Lachaux, Thibaut Lavril, Jenya Lee, Diana Liskovich, Yinghai Lu, Yuning Mao, Xavier Martinet, Todor Mihaylov, Pushkar Mishra, Igor Molybog, Yixin Nie, Andrew Poulton, Jeremy Reizenstein, Rashi Rungta, Kalyan Saladi, Alan Schelten, Ruan Silva, Eric Michael Smith, R. Subramanian, Xia Tan, Binh Tang, Ross Taylor, Adina Williams, Jian Xiang Kuan, Puxin Xu, Zhengxu Yan, Iliyan Zarov, Yuchen Zhang, Angela Fan, Melanie Kambadur, Sharan Narang, Aurelien Rodriguez, Robert Stojnic, Sergey Edunov, and Thomas Scialom. 2023. Llama 2: Open foundation and fine-tuned chat models.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. *CoRR*, abs/1706.03762.

Antti Virtanen, Jenna Kanerva, Rami Ilo, Jouni Luoma, Juhani Luotolahti, Tapio Salakoski, Filip Ginter, and Sampo Pyysalo. 2019. Multilingual is not enough: BERT for finnish. *CoRR*, abs/1912.07076.

Alex Wang, Yada Pruksachatkun, Nikita Nangia, Amanpreet Singh, Julian Michael, Felix Hill, Omer Levy, and Samuel R. Bowman. 2019a. Superglue: A stickier benchmark for general-purpose language understanding systems. *CoRR*, abs/1905.00537.

Alex Wang, Amanpreet Singh, Julian Michael, Felix Hill, Omer Levy, and Samuel Bowman. 2018. GLUE: A multi-task benchmark and analysis platform for natural language understanding. In *Proceedings of the 2018 EMNLP Workshop BlackboxNLP: Analyzing and Interpreting Neural Networks for NLP*, pages 353–355, Brussels, Belgium. Association for Computational Linguistics.

Wei Wang, Bin Bi, Ming Yan, Chen Wu, Zuyi Bao, Liwei Peng, and Luo Si. 2019b. Structbert: Incorporating language structures into pretraining for deep language understanding. *CoRR*, abs/1908.04577.

Guillaume Wenzek, Marie-Anne Lachaux, Alexis Conneau, Vishrav Chaudhary, Francisco Guzmán, Armand Joulin, and Edouard Grave. 2020. CCNet: Extracting high quality monolingual datasets from web crawl data. In *Proceedings of the Twelfth Language Resources and Evaluation Conference*, pages 4003–4012, Marseille, France. European Language Resources Association.

Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Rémi Louf, Morgan Funtowicz, and Jamie Brew. 2019. Huggingface's transformers: State-of-the-art natural language processing. *CoRR*, abs/1910.03771.

Linting Xue, Aditya Barua, Noah Constant, Rami Al-Rfou, Sharan Narang, Mihir Kale, Adam Roberts, and Colin Raffel. 2022. ByT5: Towards a token-free future with pre-trained byte-to-byte models. *Transactions of the Association for Computational Linguistics*, 10:291–306.

Entity Embelishment Mitigation in LLMs Output with Noisy Synthetic Dataset for Alignment

Svitlana Galeshchuk
Arval Leasing Solutions BNP Paribas, West Ukrainian National University
22 r Deux Gares 92500 Rueil Malmaison France, 1 Lvivska str. Ternopil Ukraine
svitlana.galeshchuk@gmail.com

Abstract

The present work focuses on the entity embellishments when named entities are accompanied by additional information that is not supported by the context or the source material. Our paper contributes into mitigating this problem in large language model's generated texts, summaries in particular, by proposing the approach with synthetic noise injection in the generated samples that are further used for alignment of finetuned LLM. We also challenge the issue of solutions scarcity for low-resourced languages and test our approach with corpora in Ukrainian.

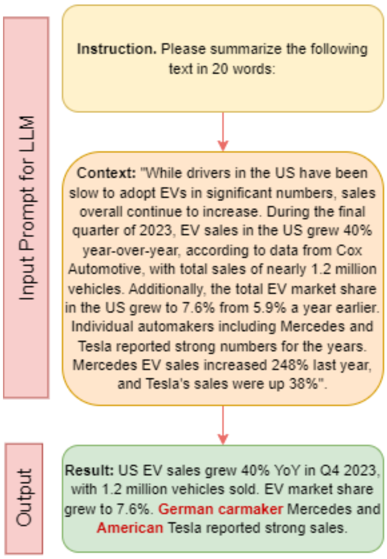
Keywords: large language models, Llama, summarization, Ukrainian NLP

1. Introduction

Text generation is a task that produces text conditioning on an input (a question, an article, an image, etc.). With the increase in number of Transformer models and availability of textual data, we are seeing a rapid growth in the number of text generation applications such as summarization, chatbots, storytelling, and machine translation. The fluency and diversity of automatically produced text has advanced significantly with the introduction of large and very large language models (LLMs). However, LLMs use a probabilistic approach to generate text, which makes these models prone to creating factually incorrect, inconsistent, or irrelevant information that is not supported in the input. This is called hallucination. In real-world applications, hallucinations can pose many problems, ranging from ethical risks to loss of trust from clients. As a result, scholars and practitioners in the field of natural language generation (NLG) have focused their research on mitigating the risk of adding irrelevant information.

Hallucinations problems can be broadly categorized into two types: **factuality hallucination** and **faithfulness hallucination**, as identified by Huang et al. (2023). Factuality hallucination is characterized by a discrepancy between the generated content and real-world facts that can be verified. On the other hand, faithfulness hallucination occurs when there is a deviation of the generated output from the instructions or context provided by the input. This type of hallucination can be further subcategorized into instruction, context, or logic inconsistencies. Future research in this area is crucial to enhance the quality of natural language generation output and to improve the accuracy and relevance of the generated text.

In the paper, we focus on the faithfulness problem, and context inconsistencies in particular when LLM generated output is imprecise or untrue compared to the user's input.



Instruction. Please summarize the following text in 20 words:

Context: "While drivers in the US have been slow to adopt EVs in significant numbers, sales overall continue to increase. During the final quarter of 2023, EV sales in the US grew 40% year-over-year, according to data from Cox Automotive, with total sales of nearly 1.2 million vehicles. Additionally, the total EV market share in the US grew to 7.6% from 5.9% a year earlier. Individual automakers including Mercedes and Tesla reported strong numbers for the years. Mercedes EV sales increased 248% last year, and Tesla's sales were up 38%".

Result: US EV sales grew 40% YoY in Q4 2023, with 1.2 million vehicles sold. EV market share grew to 7.6%. German carmaker Mercedes and American Tesla reported strong sales.

Figure 1: Example of entity hallucination we tackle in the paper.

Figure 1 illustrates the problem when a user asks a LLM to summarize a given article, we find added information on which are nationalities of Tesla and Mercedes that being true (in 2024) is not, however, mentioned in the article but assumed by the LLM as probable to be in the output.

We refer to this the type of context hallucinations that accompany named entities as entity embellishment and define mitigating them as the main scope of the paper. This brings us to the objective of the paper that aims at reducing the risk of context hallucination, in particular entity embellishment, in foundation models using summarization dataset

The Third Ukrainian Natural Language Processing Workshop (UNLP) @LREC-COLING-2024, pages 129–134
25 May, 2024. © 2024 ELRA Language Resource Association: CC BY-NC 4.0

and pertrurbated examples for model alignment via direct preferential optimization (DPO) procedure. More precisely, the development of LLMs involves two main stages:

- the first stage is **pre-training**, where the models learn general representations and acquire knowledge about the world
- the second stage is **alignment**, where the models are trained to better align with the instructions and preferences of users.

Our approach involves utilizing LLM by fine-tuning it with articles that come with their corresponding golden summaries. We then align the trained model by using generated texts that have been corrupted with injected information on named entities from another LLM, in particular GPT-4. The golden standard is considered as the chosen and preferred answer. During the direct preference optimization (DPO) phase of training, any synthetic response enriched with text from GPT-4 is shown to be rejected and golden summary to be chosen.

The occurrence of hallucinations in LLM output texts is a known issue. However, very few studies have explored how to mitigate hallucination problems in low-resource languages other than English. This is because the most of the pre-training corpora is usually in English for the majority of available LLMs. Consequently, these models may learn information in English and apply it to tasks in other languages. To challenge these limitations, we conducted tests in Ukrainian, a low-resource language, to verify the consistency of results in non-English documents.

The article is organized as follows: Section 2 elaborates on related work and the choice of evaluation metrics. Section 3 focuses on data used to train and align a LLM. Section 4 highlights the experimental setup described in Introduction together with the main challenges. Section 5 presents the results of the study and potential limits.

2. Related Work

Hallucination in text generation is a well-known phenomenon hence we find a plethora of scientific papers on the nature and solutions to LLM embellishments.

2.1. Surveys on hallucination phenomenon and its nature.

We cite several papers that elaborate on the survey analysis of LLM hallucinations. The study by Ji et al. (2023) mainly focuses on the occurrence of hallucinations in pre-trained language models for natural language generation tasks, while not discussing LLMs. The paper of Wang et al. (2023) concentrates on the factuality of LLMs-generated texts. Tonmoy et al. (2024) provides a taxonomy of mitigating approaches against hallucinations, stressing out prompt engineering with retrieval augmented generation and self-refinement through feedback and reasoning as well as prompt-tuning. Yao et al. (2023) demonstrate that nonsense prompts composed of random tokens can also elicit hallucinations in LLMs, suggesting that hallucination may be another view of adversarial examples. Huang et al. (2023), claims that LLMs have been known to create non-existent facts. Current explanations attribute this to the training datasets McKenna et al. (2023). These works argue that noisy data or model overfitting to the training data is responsible for hallucination. The authors believe that alignment, involving supervised fine-tuning and reinforcement learning is crucial for unlocking LLMs capabilities and aligning them with human preferences. However, it introduces the risk of hallucinations due to capability misalignment and belief misalignment, including sycophantic behavior driven by human preferences. Wiggers (2023) suggest that hallucinating models can serve as collaborative creative partners; providing valuable outputs that may not be factual but can lead to novel ideas. While hallucinations can be problematic when factually inaccurate, they can be advantageous in creative or artistic endeavors. In terms of related works for Ukrainian language, we cite Kang et al. (2024) who test multilingual BLOOM for hallucinations finding significant faithfullness issues in generated texts in Ukrainian.

2.2. Strategies to overcome hallucinations

Decoding strategies. Lango and Dušek (2023) highlight decoding strategies as techniques designed to target the generation phase of a model. With regards to hallucination, these techniques aim to reduce its occurrence in the generated outputs by guiding the generation phase toward producing authentic or context-specific content, Shi et al. (2023), expand their study to context-aware decoding relying on the intuition that a contrastive output distribution amplifies the difference between the output probabilities when a model is used with and without context. Choi et al. (2023) introduce a method called Knowledge-Constrained Decoding (KCD) that uses a token-level detection system to identify hallucinations and improve the generation process by adjusting the token distribution based on a more an accurate estimate of future knowledge groundedness. **Knowledge base strategies**. Zhang et al. (2023) address the issue of knowledge alignment by introducing MixAlign, a framework that interacts with both the user and the knowledge base to clarify the relationship between the user question and the information stored in the knowledge base. This approach while being effective for factual inconsistencies is not designed for faithfulness problems. **Training strategies.** DRESS: (Chen et al. (2023), propose using critique and refinement of natural language feedback to improve alignment with human preferences and tackle hallucination issues. This the approach allows us to define the setup of the paper that exploits the alignment stage to "show" the model the right and "wrong", corrupted samples with hallucinations.

2.3. Metric for hallucination

According to Azaria and Mitchell (2023),Ji et al. (2023), LLMs are capable of determining the factual accuracy of statements, even when the false statements are generated by the models themselves. The statement brings us to investigate the potential capabilities of LLMs to judge the faithfulness of generated texts without a need of a human annotator. Here are the metrics considered in our research:

- **N-gram**, (calculates the ratio of token overlap between the generated output and the correct answer) based metrics like ROUGE and PARENT-T assesses faithfulness but show poor correlation with humans thus their usage is very limited (Ji et al. (2023), Maynez et al. (2020)).
- **Feedback from another LLM**: Feng et al. (2023) proposes to employ GPT-4 to collect sentence-level factual consistency annotation for system-generated summaries. They make a comparison between GPT-4 and human annotations prove high correlation of the feedbacks.
- **Weekly supervised classifier finetuning**: , Kryściński et al. (2019) create a data set by corrupting golden summaries with paraphrasing, entity swapping, and noise injection. Similarly, Dziri et al. (2021) develop perturbated samples by replacing up to two verbs with verbs of the same tense or extracting all mentioned entities from different dialogue examples using the SpaCy NER tagger and corrupting them.

The overview of the literature helps define our experimental strategy by creating a dataset of adversarial summaries to golden summaries for news articles inspired by weekly-supervised approaches presented that are used as an input to LLM alignment phase rather than fine-tuning that is advocated by Chen et al. (2023). We then apply GPT-4 to assess faithfulness of generated texts as this method reflects human feedback (Feng et al. (2023)) and can account for the abstractiveness of generated answers.

3. Input Data

We test our approach on summarization task. Considering the scope of experimentation is low-resource languages we use the Ukrainian part of XL-SUM dataset.

The Ukrainian part of the XL-SUM dataset is a collection of more than 58,000 BBC news articles in Ukrainian, introduced by Hasan et al. (2021)[1]. It is used as a training resource for summarization in Ukrainian and is considered a benchmark for comparison and evaluation in related studies. No human evaluation was provided for the Ukrainian language, as the authors focus mainly on the top 10 spoken languages. The data is used to train language model. However, due to the lack of computational resources we use only the first 10k examples to fine-tune the model, first 3K of test split as a test set and the rest of the test split (around 2.6K articles) as validation set for the alignment as described in the following chapter.

4. Experimental Setup

4.1. Large Language Model

Since the introduction of ChatGPT to public use, LLMs models became popular not only among researchers and data scientists for particular applications but also to the general public that accelerated development of LLMs. One of the first open-sourced models released was Llama from Meta. We use Llama-2 as a language model for the set-up. Llama 2 is a freely available large language model that has been trained on 2 trillion tokens from public online sources. They include also Wikipedia dumps from the June-August 2022 period part of which is in Ukrainian. The model thus may be applied to texts in Ukrainian, however, Meta researchers warn they do not run tests of Llama with languages other than English. It is available in sizes of 7B, 13B, and 70B parameters. We use the 13B version in the paper.

The set-up for our approach foresees the following steps depicted on Fig. 2:

1. Fine-tune Llama-2 model on training data.
2. Generate summaries using fine-tuned Llama-2 model on validation set.
3. Corrupt generated summaries by adding information not given in input text.

[1] Downloaded from https://huggingface.co/datasets/csebuetnlp/xlsum

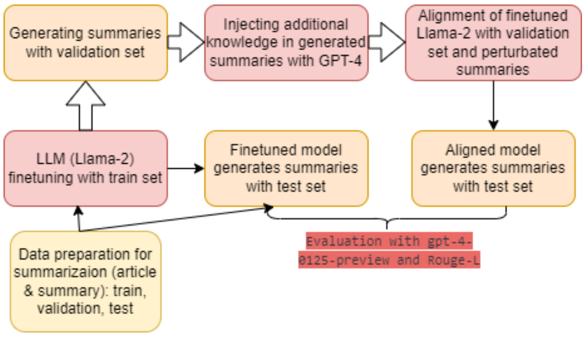

Generating summaries with validation set
Injecting additional knowledge in generated summaries with GPT-4
Alignment of finetuned Llama-2 with validation set and perturbated summaries
LLM (Llama-2) finetuning with train set
Finetuned model generates summaries with test set
Aligned model generates summaries with test set
Data preparation for summarizaion (article & summary): train, validation, test
Evaluation with gpt-4-0125-preview and Rouge-L

Figure 2: Illustration of the proposed approach.

4. Align fine-tuned Llama-2 with golden summaries to choose and noisy synthetic text from Step 3 to be rejected.

5. Apply both fine-tuned and aligned versions on test set.

6. Assess level of faithfulness hallucinations in generated texts using GPT-4 and Rouge-L, and human evaluation on a small subset.

4.2. LLM Finetuning

We use open-source Python packages for LLM finetuning using Lora adapters for faster training (transformers, trl, perf). The following training arguments ensure the results of the paper: learning-rate=2e-4, warm-up ratio = 0.03, maximum number of tokens = 512, truncate otherwise, 5 epochs. Lora perf arguments: rank = 32, lora-alpha=16, dropout = 0.1. As mentioned in Section 3.1., the first 10k of XL-Sum train split's articles has been used for finetuning. We used A100 40G GPU in the experiment. The training uses the prompt format:

Article to summarize in 26 words delimited with triple backticks: Article : “‘{article}“‘, Summary : “‘{summary}“‘.

4.3. Alignment with data perturbation

After finetuning the model generates summaries for 1239 articles out of the validation set that the LLM has not seen during training. These 1239 are chosen with the following logic: the average length of the golden summary is 26 words. We want to make sure that during alignment model does not prefer golden summaries because they are shorter than generated. For this, we adjust the training prompt format for inference. But more importantly we filter out rows with golden summaries of less than 20 words. We find 1239 articles after filtering from initial almost 2.6K set.

The generated summaries are further corrupted with added noise from GPT-4. Here is an algorithm applied: we extract named entities from the generated summaries using the Spacy NER model for Ukrainian and pass the first occurred entity together with generated text as an input to GPT-4 model asking the latter to enrich the text with information on the entity.

Prompt used for data corruption: *Instruction: You are a newspaper editor with much of encyclopedic knowledge. You have an entity and a text in Ukrainian. Then please insert in the phrase information of up to 4 words about the entity. Context: the text: {text}, entity: {entity}. Input: Your answer shall contain this text in Ukrainian enriched with your information in Ukrainian. Please add information about the entity as mentioned in the instruction.* . For example, for the following text (translated in English): Title "Mural: from Philadelphia to Rabat", article: "Since several years on Kyiv multi-storey buildings are emerging..." and golden summary: "While for Kyiv the rock art phenomenon is relatively new, in the West - ..." the finetuned Llama model generates: "In Kyiv, street art is quickly expanding, said mayor Klitchko.". Corrupted sample is: "In Kyiv, street art is quickly expanding, said mayor Klitchko, a former boxer".

We used DPO for model alignment with the following parameters: learning-rate = 2e-6, beta = 0.5, batch = 2. Beta is relatively high to use the model knowledge.

5. Evaluation and Results

Recall from Section 2 that we build on Feng et al. (2023) approach to use one LLM model to evaluate the results of another. The following prompt is the input of GPT-4 model that shall define which summary contains irrelevant information:

Verify if summary is not consistent with the corresponding article. Provide the answer "Yes" if consistent or "No" if not consistent. The article: {article}; the summary: {summary}

The results of GPT-4 evaluation together with Rouge-L score are given in the Table 1. GPT-4 metric contains a percentage of texts found without hallucinations due to GPT-4. We can observe an increase of both Rouge-L and GPT-verified evaluation scores after alignment with synthetically generated texts with added noise. Apart from GPT-4 classification we randomly sampled 50 articles from the test set and asked human annotators to check for entity embellishments in summaries generated by finetuned and alighned LLama-2 versions presented in the paper. The rule for annotation is the following: if at least one embellishment found, label the article as 1, else 0. Out of 50 summaries produced by fine-tuned LLM, 11 contained faithfulness problems; out of 50 summaries produced by aligned LLM, only 6 contained entity embellishment.

Metric	Finetuned	Aligned
Rouge-L	23.4	29.7
GPT-4	72.1	81.5

Table 1: Results on test dataset with 3K news articles for finetuned model vs finetuned&aligned model with synthetic data corrupted with entities information (II)

The reduction in entity hallucinations is quite significant in case of human check but the sample is too small to be used as a proxy for all test data. Based on the results we may claim that our approach to alignment input data is experimentally tested.

Having obtained positive results to attain our objective, we shall recognize limitations of our study: 1. Bigger test set might have shown more accurate results. 2. Experiment with other language could prove coherence of our set-up. 3. Automatic evaluation with LLM model may imbibe issues and biases of evaluating model and might be not always correct. Rouge-L score has many limits (see Section2). 4. Human evaluation of bigger sample would show more accurate evaluation of results. 5. Experimenting with more prompts and Llama-specific syntax could deliver improvements. Thus, we foresee using the same algorithm with more data in Ukrainian and make comparison with other languages in future research to avoid stochastic biases.

We release the following versions of the Llama-2 model on HuggingFace Hub as described in the paper:

* finetuned model [2];

* aligned with noisy synthetic data [3].

HuggingFace dataset hub also contains the test data with golden and corrupted synthetic summaries [4].

6. Bibliographical References

Amos Azaria and Tom Mitchell. 2023. The internal state of an llm knows when its lying. *arXiv preprint arXiv:2304.13734*.

Yangyi Chen, Karan Sikka, Michael Cogswell, Heng Ji, and Ajay Divakaran. 2023. Dress: Instructing large vision-language models to align and interact with humans via natural language feedback. *arXiv preprint arXiv:2311.10081*.

Sehyun Choi, Tianqing Fang, Zhaowei Wang, and Yangqiu Song. 2023. Kcts: knowledge-constrained tree search decoding with token-level hallucination detection. *arXiv preprint arXiv:2310.09044*.

Nouha Dziri, Andrea Madotto, Osmar Zaiane, and Avishek Joey Bose. 2021. Neural path hunter: Reducing hallucination in dialogue systems via path grounding. *arXiv preprint arXiv:2104.08455*.

Huawen Feng, Yan Fan, Xiong Liu, Ting-En Lin, Zekun Yao, Yuchuan Wu, Fei Huang, Yongbin Li, and Qianli Ma. 2023. Improving factual consistency of text summarization by adversarially decoupling comprehension and embellishment abilities of llms. *arXiv preprint arXiv:2310.19347*.

Tahmid Hasan, Abhik Bhattacharjee, Md. Saiful Islam, Kazi Mubasshir, Yuan-Fang Li, Yong-Bin Kang, M. Sohel Rahman, and Rifat Shahriyar. 2021. XL-sum: Large-scale multilingual abstractive summarization for 44 languages. In *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*, pages 4693–4703, Online. Association for Computational Linguistics.

Lei Huang, Weijiang Yu, Weitao Ma, Weihong Zhong, Zhangyin Feng, Haotian Wang, Qianglong Chen, Weihua Peng, Xiaocheng Feng, Bing Qin, et al. 2023. A survey on hallucination in large language models: Principles, taxonomy, challenges, and open questions. *arXiv preprint arXiv:2311.05232*.

Ziwei Ji, YU Tiezheng, Yan Xu, Nayeon Lee, Etsuko Ishii, and Pascale Fung. 2023. Towards mitigating llm hallucination via self reflection. In *The 2023 Conference on Empirical Methods in Natural Language Processing*.

Haoqiang Kang, Terra Blevins, and Luke Zettlemoyer. 2024. Comparing hallucination detection metrics for multilingual generation. *arXiv preprint arXiv:2402.10496*.

Wojciech Kryściński, Bryan McCann, Caiming Xiong, and Richard Socher. 2019. Evaluating the factual consistency of abstractive text summarization. *arXiv preprint arXiv:1910.12840*.

Mateusz Lango and Ondřej Dušek. 2023. Critic-driven decoding for mitigating hallucinations in data-to-text generation. *arXiv preprint arXiv:2310.16964*.

Joshua Maynez, Shashi Narayan, Bernd Bohnet, and Ryan McDonald. 2020. On faithfulness and factuality in abstractive summarization. *arXiv preprint arXiv:2005.00661*.

[2] https://huggingface.co/SGaleshchuk/Llama-2-13b-hf_uk_rank-32_ft

[3] https://huggingface.co/SGaleshchuk/Llama-2-13b-sum_ukr_dpo

[4] https://huggingface.co/datasets/SGaleshchuk/XL_SUM_ukr_synthetic_hallucinations

Nick McKenna, Tianyi Li, Liang Cheng, Mohammad Javad Hosseini, Mark Johnson, and Mark Steedman. 2023. Sources of hallucination by large language models on inference tasks. *arXiv preprint arXiv:2305.14552*.

Weijia Shi, Xiaochuang Han, Mike Lewis, Yulia Tsvetkov, Luke Zettlemoyer, and Scott Wen-tau Yih. 2023. Trusting your evidence: Hallucinate less with context-aware decoding. *arXiv preprint arXiv:2305.14739*.

SM Tonmoy, SM Zaman, Vinija Jain, Anku Rani, Vipula Rawte, Aman Chadha, and Amitava Das. 2024. A comprehensive survey of hallucination mitigation techniques in large language models. *arXiv preprint arXiv:2401.01313*.

Cunxiang Wang, Xiaoze Liu, Yuanhao Yue, Xiangru Tang, Tianhang Zhang, Cheng Jiayang, Yunzhi Yao, Wenyang Gao, Xuming Hu, Zehan Qi, et al. 2023. Survey on factuality in large language models: Knowledge, retrieval and domain-specificity. *arXiv preprint arXiv:2310.07521*.

Kyle Wiggers. 2023. Llm lies: Hallucinations are not bugs, but features as adversarial examples.

Jia-Yu Yao, Kun-Peng Ning, Zhen-Hui Liu, Mu-Nan Ning, and Li Yuan. 2023. Llm lies: Hallucinations are not bugs, but features as adversarial examples. *arXiv preprint arXiv:2310.01469*.

Shuo Zhang, Liangming Pan, Junzhou Zhao, and William Yang Wang. 2023. Mitigating language model hallucination with interactive question-knowledge alignment. *arXiv preprint arXiv:2305.13669*.

Language-Specific Pruning for Efficient Reduction of Large Language Models

Maksym Shamrai
Institute of Mathematics of NAS of Ukraine
Kyiv Academic University
Kyiv, Ukraine
m.shamrai@imath.kiev.ua

Abstract

Delving into pruning techniques is essential to boost the efficiency of Large Language Models (LLMs) by reducing their size and computational demands, resulting in faster and more cost-effective inference. In this work, our key contribution lies in recognizing that LLMs trained on diverse languages manifest distinct language-specific weight distributions. Exploiting this insight, we illustrate that pruning LLMs using language-specific data results in a more potent model compression. Empirical evidence underscores the critical nature of pruning on language-specific data, highlighting a noteworthy impact on the perplexity of Ukrainian texts compared to pruning on English data. The proposed methodology significantly reduces the size of LLaMA, LLaMA 2 and Mistral models while preserving competitive performance. This research underscores the significance of linguistic considerations in LLM pruning and advocates for language-specific optimization, establishing a framework for more efficient and tailored language models across diverse linguistic contexts. Additionally, all experiments were conducted using a single consumer-grade NVIDIA RTX 3090 GPU, and the code is available at `https://github.com/mshamrai/language-specific-pruning`.

Keywords: Language Model Pruning, Large Language Models, Language-Specific Optimization, Ukrainian Language Processing

1. Introduction

The evolution of Large Language Models (LLMs) has unlocked unprecedented capabilities in natural language processing, yet the monumental size of these models necessitates innovative solutions for their efficient deployment. Lately, quantization techniques, which employ lower precision types for compression, have enhanced the accessibility of LLMs to a broader audience (Frantar et al., 2022; Dettmers et al., 2022, 2024). While these advancements are noteworthy, alternative compression methods can yield significant improvements. Pruning, a technique involving the selective removal of model weights, is a promising avenue for addressing computational challenges without compromising performance.

While existing pruning methods have demonstrated success in general contexts (Molchanov et al., 2019; Yang et al., 2022; Ma et al., 2024), their application to different languages and the implications for model performance remain largely unexplored. This paper pioneers the investigation of language-specific pruning for LLMs, with a dedicated focus on the Ukrainian language. Our objective is to establish that the efficacy of pruning methods is linked to the linguistic characteristics of the target language. Leveraging state-of-the-art techniques such as SparseGPT (Frantar and Alistarh, 2023) and Wanda (Sun et al., 2023), our method achieves competitive perplexity scores when evaluated on a Ukrainian dataset with sparse versions of LLaMA (Touvron et al., 2023a), LLaMA 2 (Touvron et al., 2023b) and Mistral (Jiang et al., 2023) models, eliminating the necessity for retraining.

Moreover, considering that pruning strategies in Transformer-based models primarily target linear layers due to their significant presence and crucial role in model parameterization, the methods employed and our findings are applicable to any Transformer architecture without constraints.

It is essential to note that the successful application of SparseGPT and Wanda requires reference data to tailor the pruning specifically for the characteristics of the given dataset. For our Ukrainian language exploration, we utilized reference data sourced from UberText 2.0 (Chaplynskyi, 2023) – this corpus provides a robust foundation for assessing the effectiveness of language-specific pruning in real-world linguistic contexts.

Additionally, we delve into the ramifications of language-specific pruning on model performance. To emphasize the language-specific nature of our findings, we conducted additional experiments by attempting to prune models on the English c4 dataset (Raffel et al., 2019).

The evaluation of pruning methods for the Ukrainian language includes a comparison of perplexity metrics for dense, unstructured, and 2:4 semi-structured sparsity patterns with 50% sparsity, indicating a pruning of models by half. The adoption of a 2:4 semi-structured sparsity pattern, where at least two out of every four elements must be zero, is investigated due to its native support in the NVIDIA Ampere GPU architecture, leading to significant computational speed-ups (Mishra et al.,

25 May, 2024. © 2024 ELRA Language Resource Association: CC BY-NC 4.0

2021).

In conclusion, this research marks a pioneering effort in advancing the efficiency of Large Language Models (LLMs) through the exploration of language-specific pruning techniques, with a focused examination on the Ukrainian language. Our primary contribution lies in establishing a profound connection between the efficacy of pruning methods and the unique linguistic characteristics of the target language.

2. Related work

While our work primarily focuses on training-free approaches to language model pruning, it is essential to acknowledge the existence of methods that require post-pruning retraining (Jiao et al., 2019; Ma et al., 2024). The effectiveness of such methods is contingent on the availability and quality of training data, making it less practical for scenarios where acquiring sufficient annotated data is a formidable task.

In the context of low-resource languages such as Ukrainian, where limited annotated data poses a significant obstacle, this limitation underscores the importance of investigating training-free approaches, which mitigate the need for additional labeled data. Therefore, we focus on the methods that requires only a relatively small calibration dataset for efficient model pruning.

These approaches share a similar concept: assessing weight importance based on a specific metric and input calibration data, where a larger value of the importance metric indicates that the weight should be retained. The pruning process is conducted in a layer-wise manner, involving the calculation of weight importance for each layer. Subsequently, the weights are sorted, and depending on the desired sparsity level, weights with lower importance are replaced with zeros. This streamlined approach facilitates efficient pruning, even for large-scale models.

The subsequent subsections delve into the details of this methods, highlighting its potential and practicality in the context of low-resource languages.

2.1. SparseGPT

In recent strides towards optimizing the efficiency of Large Language Models (LLMs), SparseGPT emerges as a pioneering one-shot pruning method (Frantar and Alistarh, 2023).

The foundation of SparseGPT's pruning methodology lies in the formalization of the problem through a local layer-wise reconstruction approach. It employs a pruning metric that considers the layer-wise reconstruction problem.

$$\mathbf{S}_{ij} = \left[|\mathbf{W}|^2/\text{diag}\left((\mathbf{X^T X} + \lambda \mathbf{I})^{-\mathbf{1}}\right)\right]_{\mathbf{ij}} \qquad (1)$$

The weight importance metric utilized in SparseGPT, represented by Equation 1, incorporates the Hessian matrix in the denominator, where $\mathbf{W}$ denotes the weights, $\mathbf{X}$ represents the inputs, and λ stands for the Hessian dampening factor, employed to prevent the collapse of inverse computation. This metric underscores the importance of local layer-wise information during the pruning process. By prioritizing such information, SparseGPT ensures the preservation of accuracy levels crucial for the optimal performance of large language models.

2.2. Wanda

The approach, termed "Pruning by Weights and Activations" (Sun et al., 2023) presents an effective solution to the pruning challenge. Wanda augments the standard weight magnitude pruning metric with input activations, effectively evaluating weight importance.

$$\mathbf{S}_{ij} \ = |\mathbf{W_{ij}}| \cdot ||\mathbf{X_{ij}}||_2 \qquad (2)$$

The computation of weight importance in Wanda is defined by Equation 2, where the score for each individual weight $\mathbf{W_{ij}}$ is computed as the product of its magnitude and the corresponding norm of input feature $\mathbf{X_{ij}}$. Therefore, the score encapsulates the weight's importance within the context of its associated input activations.

One of the key strengths of Wanda lies in its computational efficiency and minimal memory overhead. The method can be executed in a single forward pass, making it suitable for practical implementation in large-scale language models.

In summary, SparseGPT and Wanda employ different weight importance metrics, each grounded in a common conceptual framework. While SparseGPT utilizes a more complex metric, Wanda prioritizes computational efficiency. Following sections will explore comparative analyses to assess the effectiveness of each method for language-specific pruning.

3. Experimental Methodology and Setup

In this section detailing our experimental methodology and setup for the pruning experiments, we chose models from the LLaMA and Mistral families, specifically opting for LLaMA 7B, LLaMA 2 7B and Mistral v0.1 7B in 16-bit floating point precision.

To evaluate the models, we utilize the perplexity metric, which measures the effectiveness of a language model in predicting a sequence. Perplexity is computed as the exponentiated average negative log-likelihood of a sequence, representing the level of surprise or uncertainty of the model in predicting the next token. Mathematically, if we have a tokenized sequence $X = (x_0, x_1, \ldots, x_t)$, then the perplexity of X is calculated using the equation:

$$\text{PPL}(X) = exp\{-\frac{1}{t}\sum_{i=0}^{t}\log p_\theta(x_i|x_{<i})\},$$

where $\log p_\theta(x_i|x_{<i})$ denotes the log-likelihood of the ith token conditioned on the preceding tokens $x_{<i}$, according to our model parameterized by θ. Therefore, a higher value of perplexity indicates poorer predictions, while lower perplexity values signify better model performance.

Our focus on pruning and subsequent evaluation centered around the Ukrainian language, and for this, we utilized the UberText 2.0 corpus (Chaplynskyi, 2023), encompassing various subcorpora such as court, fiction, news, and Wikipedia. Excluding the social subcorpus, which predominantly contains short texts, we randomly selected 1000 samples for calibration and 50 samples for evaluation from each relevant subcorpus. In total, the calibration dataset consisted of 4000 samples, while the evaluation dataset consisted of 200 samples. These selections contributed to the creation of robust calibration and evaluation datasets, with each sample exceeding a length of 8192 characters.

To calibrate the model effectively, we implemented a random sampling approach from the calibration dataset, utilizing a specified seed along with the number of calibration samples as input arguments. The evaluation process covered the full evaluation dataset, calculating perplexity. Experiments were conducted with varying numbers of calibration samples and three distinct seeds to ensure statistical robustness, with mean and standard deviation calculations performed across multiple runs involving different seeds.

To underscore the importance of linguistic considerations, we expanded our experimentation to include the pruning of models on the c4 dataset, written in English. The subsequent evaluation was carried out on the Ukrainian-language evaluation dataset. Furthermore, to comprehensively assess and compare pruning performance, we also evaluated the dense version of the models (i.e., the original models without pruning) on the same dataset.

Our experiments included the introduction of diverse sparsity structures, such as unstructured and semi-structured 2:4 sparsity. Each configuration aimed to achieve a 50% sparsity level, indicating that half of the weights in each linear layer were pruned.

Overall, the objective of the experiments is to empirically and statistically investigate several key aspects:

1. The impact of the size of the calibration dataset on the performance of pruned models.
2. Comparison of different pruning methods to determine their efficacy for language-specific tasks.
3. Assessment of the significance of the language used in the calibration data for pruning effectiveness.

These experiments aim to provide insights into the factors influencing model performance post-pruning, identify optimal pruning methods tailored to language-specific requirements, and ascertain the relevance of language-specific calibration data for pruning outcomes.

Regarding the hardware requirements of the methods, both are capable of pruning 7B models in a matter of hour on a single NVIDIA RTX 3090. Pruning larger-scale models is also feasible but requires additional computational resources. For instance, in a study by Frantar and Alistarh (2023), the authors demonstrate that their method can prune a 175B model on a single NVIDIA A100 GPU. Overall, based on the experiments conducted, we can conclude that the pruning requirements primarily depend on the size of the model and its contextual window, without incurring additional overhead. Therefore, the pruning requirements are approximately equivalent to those of inference.

4. Results

In this section, we present and discuss the outcomes of our experiments, focusing on the perplexity metric evaluated on the Ukrainian evaluation dataset with various setups for different models.

Table 1 illustrates perplexity values for models pruned on UberText 2.0 dataset, employing both unstructured and 2:4 semi-structured pruning configurations with 50% sparsity. Additionally, the models underwent pruning using diverse calibration sample sizes (64, 128, 256, 512) to examine the relationship between sample size and performance.

Analyzing the table, it could be observed that Wanda's performance appears independent of calibration set size or, perhaps, this correlation does not consistently hold across all models. This is particularly evident in the perplexity values of unstructured models, such as Mistral v0.1 7B, where the Pearson correlation between calibration set size and perplexity mean values is 0.99, and LLaMA 2 7B, where the correlation is -0.98. Conversely, all

Method	Calibration Samples	LLaMA 7B	LLaMA 2 7B	Mistral v0.1 7B
Unstructured Wanda	64	12.162 ± 0.025	11.283 ± 0.007	**9.314 ± 0.098**
	128	12.161 ± 0.012	11.278 ± 0.007	9.726 ± 0.125
	256	**12.148 ± 0.008**	11.275 ± 0.009	10.385 ± 0.038
	512	12.152 ± 0.007	**11.254 ± 0.012**	12.262 ± 0.424
2:4 Wanda	64	31.533 ± 0.169	**30.101 ± 0.406**	**29.822 ± 0.381**
	128	31.438 ± 0.348	30.177 ± 0.361	30.741 ± 0.231
	256	31.496 ± 0.327	30.651 ± 0.353	32.709 ± 0.328
	512	**31.198 ± 0.446**	30.883 ± 0.271	34.471 ± 0.704
Unstructured SparseGPT	64	10.632 ± 0.027	9.703 ± 0.013	7.109 ± 0.003
	128	10.559 ± 0.011	9.683 ± 0.028	7.095 ± 0.011
	256	10.531 ± 0.006	9.671 ± 0.015	7.085 ± 0.003
	512	**10.529 ± 0.020**	**9.652 ± 0.012**	**7.074 ± 0.004**
2:4 SparseGPT	64	13.319 ± 0.092	11.559 ± 0.082	8.582 ± 0.036
	128	13.148 ± 0.192	11.515 ± 0.072	8.551 ± 0.041
	256	13.093 ± 0.054	11.457 ± 0.035	8.497 ± 0.006
	512	**12.994 ± 0.047**	**11.379 ± 0.008**	**8.476 ± 0.031**

Table 1: Perplexity values of different models and different pruning configuration.

models pruned by SparseGPT exhibit a notably high negative correlation, such as for 2:4 LLaMA 7B, where the correlation is -0.9. Hence, we can assert that Wanda's performance is not necessarily dependent on the calibration data size, while SparseGPT's performance does show such dependency. This difference could be attributed to the inherent dissimilarity in the precision of importance metrics employed by each method, where Wanda utilizes a faster but less accurate metric, and SparseGPT employs a more precise but time-intensive alternative.

The Table 3 presents the optimal perplexity values achieved by models pruned using both unstructured and 2:4 semi-structured configurations, each with 50% sparsity, on calibration data from UberText 2.0 or c4 datasets. Additionally, the perplexity values for the dense models are included.

The analysis of the table leads to the conclusion that, among both unstructured and 2:4 semi-structured configurations, the most effective pruning method is SparseGPT when applied to the UberText 2.0 dataset, which consists of Ukrainian texts. It is also noteworthy that the superiority of the SparseGPT pruning technique becomes evident, particularly when the pruning pattern is 2:4 semi-structured.

Furthermore, the extreme variances observed in models pruned with c4 data indicate a significant dependency on randomness in the pruning process, suggesting that the outcome is less influenced by the dataset itself.

Moreover, we analyze the memory footprint of the models before and after pruning. As shown in Table 2, pruning with a 50% sparsity level reduces the memory size of the models by approximately 41%. Therefore, pruning enables a significant decrease in the memory consumption of the model's parameters while preserving parameters in 16-bit floating-point format. However, achieving such a reduction in memory usage is not feasible with unstructured sparsity. To attain this reduction, we should utilize a 2:4 semi-structured sparsity pattern, which employs an efficient sparse semi-structured tensor representation.

Model	Dense	Sparse
LLaMA 7B	12.58 Gbs	7.31 Gbs
LLaMA 2 7B	12.68 Gbs	7.40 Gbs
Mistral v0.1 7B	13.99 Gbs	8.30 Gbs

Table 2: Memory footprint before (dense) and after (sparse) pruning with 50% sparsity level and 2:4 semi-structured sparsity configuration of different models.

Additionally, among these three models, Mistral v0.1 7B demonstrates the best pruning performance, as indicated by the lowest residual between dense and pruned perplexity values.

Therefore, SparseGPT emerges as the preferred pruning method for language-specific applications, with its performance significantly influenced by the language of the calibration dataset.

5. Conclusion

In this study, we conducted a comprehensive set of experiments to investigate the impact of pruning methodologies on language models, with a specific

Model	LLaMA 7B	LLaMA 2 7B	Mistral v0.1 7B
Dense	8.950	8.269	6.460
Unstructured Wanda on c4	13.953 ± 0.060	13.829 ± 0.087	41.466 ± 6.314
Unstructured SparseGPT on c4	15.797 ± 0.761	15.011 ± 0.283	9.208 ± 0.086
Unstructured Wanda on UberText 2.0	12.148 ± 0.008	11.254 ± 0.012	9.314 ± 0.098
Unstructured SparseGPT on UberText 2.0	**10.529 ± 0.020**	**9.652 ± 0.012**	**7.074 ± 0.004**
2:4 Wanda on c4	52.346 ± 1.628	79.801 ± 7.338	433.940 ± 282.154
2:4 SparseGPT on c4	89.772 ± 28.306	57.460 ± 5.379	165.516 ± 90.769
2:4 Wanda on UberText 2.0	31.198 ± 0.446	30.101 ± 0.406	29.822 ± 0.381
2:4 SparseGPT on UberText 2.0	**12.994 ± 0.047**	**11.379 ± 0.008**	**8.476 ± 0.031**

Table 3: Perplexity values of different models and different pruning configuration.

focus on language-specific considerations. Our objectives were in the following:

1. **Dependency on Calibration Dataset Size:**

   The experiments aimed to state whether the performance of pruned models is influenced by the size of the calibration dataset. Results revealed that, unlike SparseGPT, the Wanda pruning method demonstrated little to no dependence on the calibration set size.

2. **Comparison of the Pruning Methods:**

   Through an analysis of perplexity values, we compared two language-specific pruning methods, Wanda and SparseGPT. The latter emerged as the preferred pruning method for language-specific applications, particularly under 2:4 semi-structured pruning configurations.

3. **Language Dependence in Pruning Performance:**

   Our investigation extended to clarify whether the pruning methods yield distinct outcomes based on the language of the calibration dataset. The results clearly demonstrated that the effectiveness significantly dependent on the language of the calibration data.

Our findings contribute valuable insights into the language-specific considerations of model pruning, paving the way for more informed choices in deploying such techniques for diverse natural language processing applications.

6. Discussion and Future Work

Our experiments reveal that different sets of parameters are optimal for different languages. In particular, an LLM pruned on English calibration data shows lower performance on the Ukrainian evaluation dataset compared to an LLM pruned on Ukrainian calibration data. Consequently, this pruning technique can serve as a foundational framework for linguistic comparisons among languages. For instance, a compelling exploration could involve comparing the languages of Polish and Ukrainian, given their Slavic roots and linguistic proximity. Demonstrating their linguistic closeness in the LLM context suggests that fine-tuning the LLM on data from both languages could potentially enhance overall performance.

Furthermore, it's essential to assess alternative training-free pruning techniques, such as those proposed by Zhang et al. (2023), to conduct a comprehensive investigation before developing a truly innovative, language-specific pruning approach.

In addition, the next phase of research could explore the synergies between pruning and quantization, aiming to create the smallest and fastest Ukrainian LLM. Combining these techniques holds the promise of optimizing model size and inference speed, contributing to more efficient language models.

7. References

Dmytro Chaplynskyi. 2023. Introducing UberText 2.0: A corpus of modern Ukrainian at scale. In *Proceedings of the Second Ukrainian Natural Language Processing Workshop*, pages 1–10, Dubrovnik, Croatia. Association for Computational Linguistics.

Tim Dettmers, Mike Lewis, Younes Belkada, and Luke Zettlemoyer. 2022. Llm. int8 (): 8-bit matrix multiplication for transformers at scale. *arXiv preprint arXiv:2208.07339*.

Tim Dettmers, Artidoro Pagnoni, Ari Holtzman, and Luke Zettlemoyer. 2024. Qlora: Efficient finetuning of quantized llms. *Advances in Neural Information Processing Systems*, 36.

Elias Frantar and Dan Alistarh. 2023. SparseGPT: Massive language models can be accurately pruned in one-shot. *arXiv preprint arXiv:2301.00774*.

Elias Frantar, Saleh Ashkboos, Torsten Hoefler, and Dan Alistarh. 2022. Gptq: Accurate post-training quantization for generative pre-trained transformers. *arXiv preprint arXiv:2210.17323*.

Albert Q Jiang, Alexandre Sablayrolles, Arthur Mensch, Chris Bamford, Devendra Singh Chaplot, Diego de las Casas, Florian Bressand, Gianna Lengyel, Guillaume Lample, Lucile Saulnier, et al. 2023. Mistral 7b. *arXiv preprint arXiv:2310.06825*.

Xiaoqi Jiao, Yichun Yin, Lifeng Shang, Xin Jiang, Xiao Chen, Linlin Li, Fang Wang, and Qun Liu. 2019. Tinybert: Distilling bert for natural language understanding. *arXiv preprint arXiv:1909.10351*.

Xinyin Ma, Gongfan Fang, and Xinchao Wang. 2024. Llm-pruner: On the structural pruning of large language models. *Advances in neural information processing systems*, 36.

Asit Mishra, Jorge Albericio Latorre, Jeff Pool, Darko Stosic, Dusan Stosic, Ganesh Venkatesh, Chong Yu, and Paulius Micikevicius. 2021. Accelerating sparse deep neural networks. *arXiv preprint arXiv:2104.08378*.

Pavlo Molchanov, Arun Mallya, Stephen Tyree, Iuri Frosio, and Jan Kautz. 2019. Importance estimation for neural network pruning. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 11264–11272.

Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J. Liu. 2019. Exploring the limits of transfer learning with a unified text-to-text transformer. *arXiv e-prints*.

Mingjie Sun, Zhuang Liu, Anna Bair, and J Zico Kolter. 2023. A simple and effective pruning approach for large language models. *arXiv preprint arXiv:2306.11695*.

Hugo Touvron, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Timothée Lacroix, Baptiste Rozière, Naman Goyal, Eric Hambro, Faisal Azhar, et al. 2023a. Llama: Open and efficient foundation language models. *arXiv preprint arXiv:2302.13971*.

Hugo Touvron, Louis Martin, Kevin Stone, Peter Albert, Amjad Almahairi, Yasmine Babaei, Nikolay Bashlykov, Soumya Batra, Prajjwal Bhargava, Shruti Bhosale, et al. 2023b. Llama 2: Open foundation and fine-tuned chat models. *arXiv preprint arXiv:2307.09288*.

Nakyeong Yang, Yunah Jang, Hwanhee Lee, Seohyeong Jung, and Kyomin Jung. 2022. Attribution-based task-specific pruning for multi-task language models. *arXiv preprint arXiv:2205.04157*.

Yuxin Zhang, Lirui Zhao, Mingbao Lin, Yunyun Sun, Yiwu Yao, Xingjia Han, Jared Tanner, Shiwei Liu, and Rongrong Ji. 2023. Dynamic sparse no training: Training-free fine-tuning for sparse llms. *arXiv preprint arXiv:2310.08915*.

Author Index