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.

The Third Ukrainian Natural Language Processing Workshop (UNLP) @LREC-COLING-2024, pages 120–128

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 pre-training 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.