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

The Third Ukrainian Natural Language Processing Workshop (UNLP) @LREC-COLING-2024, pages 23–29
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