Introducing NER-UK 2.0

A Rich Corpus of Named Entities for Ukrainian

Dmytro Chaplynskyi

Mariana Romanyshyn

lang uk

grammarly

Introduction

Named Entity Recognition **(NER)** is a task in NLP that involves identifying entities like locations, persons, organizations, etc.

NER-UK 2.0 corpus was created to improve NER for the Ukrainian language.

Corpus Size: **560** texts from various genres with **21,993** entities annotated

NER-UK 1.0 Overview

Release Year: 2016

Content: 262 texts from the BRUK corpus, covering multiple genres such as press, religious texts, fiction, and legal documents.

Annotations: 7,441 entities

Entity Types: Person (4,387), Location (1,614), Organization (780), Miscellaneous (660)

Usage: Provided a valuable resource for developing and evaluating NER systems and large pre-trained language models like roberta-large.

v2.0 Corpus Details

Annotation scheme with 13 tags

- Location
- Person,
- Organization
- Artifact
- Document
- Job title
- Date
- Time
- Period
- Money
- Percentage
- Quantity
- Miscellaneous

Corpus statistics

Sources: BRUK corpus and Nashi Groshi

	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

Entity Density: Nashi Groshi is richer in entities (16 per 100 words) compared to BRUK (3.9 per 100 words)

Results and Baseline

Inter-Annotator Agreement (IAA): **0.84**

Baseline model (roberta-large)

	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

NB: comparison is made only for the tags present in both corpora.

Conclusion and Future Work

Conclusion

- NER-UK 2.0 is the largest manually annotated NER corpus for Ukrainian.
- It includes a rich set of entities across various genres.
- Provides valuable data for developing and evaluating NER models.

Future Work:

- Fine-tuning large language models on NER-UK 2.0.
- Addressing the limitations such as bias and infrequent entity types.

Q&A

https://github.com/lang-uk/ner-uk

chaplinsky.dmitry@gmail.com, mariana.romanyshyn@grammarly.com

https://github.com/lang-uk/

https://huggingface.co/lang-uk/

lang uk

grammarly