Exploring the Use of Foundation Models for Named Entity Recognition and Lemmatization Tasks in Slavic Languages

Gabriela Pałka and Artur Nowakowski
Adam Mickiewicz University, Poznań, Poland
{gabriela.palka,artur.nowakowski}@amu.edu.pl

Full text of the article
https://arxiv.org/abs/2304.05336

TL;DR?

- **Adam Mickiewicz University's (AMU) solution for the 4th Shared Task on SlavNER**
- **Our main goals:** identification, categorization, and lemmatization of **named entities in Slavic languages**
- **Approach:** exploring the use of **foundation models**
- **Based on** the **BERT+CRF** and **T5** model architectures
- Use of **external datasets** to improve the quality
- **High metrics scores** achieved in both tasks

External Datasets

In addition to **the data released by the organizers**, we used **external datasets** for named entity recognition and lemmatization. All training and validation samples containing named entities were converted to a **CoNLL-2003 dataset format**. Due to the lack of external Czech and Russian datasets dedicated to lemmatization tasks, we decided to use **OPUS-MT** to **machine translate** all the samples prepared from the three datasets mentioned below for lemmatization.

Dataset	pl 🇵🇱	cs 🇨🇿	ru 🇷🇺
NER			
Collection3	✘	✘	✔
MultiNERD	✔	✘	✔
Polyglot-NER	✔	✔	✔
WikiNEuRal	✔	✘	✔
Lemmatization			
SEJF	✔	✘	✘
SEJFEK	✔	✘	✘
PolEval 2019: Task 2	✔	✘	✘

Named Entity Recognition

In our solution, we used several **monolingual BERT** models to better handle the specific linguistic nuances of individual Slavic languages. For comparison, we also used **multilingual BERT** models that can handle multiple languages. We found that incorporating **a CRF layer** enhanced the quality of our NER models.

Lemmatization

We approached the lemmatization task as **a text-to-text problem**. The input to the model is **an inflected phrase or named entity**, which can consist of several word forms. To address the lack of dedicated models for Czech and Russian, we used **one monolingual and a multilingual T5** model. In the multilingual experiments, we included **a language token** (»pl«, »cs«, »ru«) as the first token of the source phrases, depending on the language of the phrase.

Employed Foundation Models

Named entity recognition:

- Polish → HerBERT
- Czech → Czert
- Russian → RuBERT
- Multilingual → Slavic-BERT, XLM-RoBERTa

Lemmatization:

- Polish → plT5
- Multilingual → mT5

Shared Task Results

Submission	Recognition			Normalization		
	pl	cs	ru	pl	cs	ru
System 1	83.33	88.08	84.30	80.27	76.62	79.32
System 2	**85.37**	**89.70**	**86.16**	**82.37**	**76.89**	81.27
System 3	83.40	85.19	82.77	80.32	73.06	**81.47**
System 4	83.33	81.70	79.20	80.27	71.11	76.84

- **Lemmatization:** $plT5_{LARGE}$ for Polish trained on all available data and $mT5_{LARGE}$ for Czech and Russian trained on the data provided by the organizers and the data from PolEval 2019 Task 2.
- **System 1:** $HerBERT_{LARGE}$ for Polish trained on all available data, Czert for Czech and RuBERT for Russian trained only on the data provided by the organizers.
- **System 2:** $XLM\text{-}RoBERTa_{LARGE}$ for all languages trained only on the data provided by the organizers.
- **System 3:** $XLM\text{-}RoBERTa_{LARGE}$ for all languages trained on all available data.
- **System 4:** $HerBERT_{LARGE}$ for Polish, Czert for Czech and RuBERT for Russian trained on all available data.

Gabriela Pałka

Lemmatization models
https://huggingface.co/amu-cai

Artur Nowakowski