Exploring Transformer-Based Models for Named Entity Recognition in Ukrainian

Master's Thesis

Timo Junolainen

Februrary 2023

Department of Computer Science
University of Turku

Abstract

Named entity recognition (NER)[2] is an important task in natural language processing, consisting of identifying and classifying named entities in text. With the recent success of transformer-based models for NER in high-resource languages, this thesis explores the effectiveness of these models for NER in a low-resource[1] language: Ukrainian[3]. The research questions addressed in this thesis include the effectiveness of transformer-based models for NER in Ukrainian, the impact of different pre-training and fine-tuning strategies on model performance, and a comparison of the results to existing state-of-the-art methods for NER in Ukrainian. The objectives of this thesis are to develop transformer-based models for NER in Ukrainian, evaluate their performance on standard benchmarks, analyze the impact of different pre-training and fine-tuning strategies, and compare the results to existing state-of-the-art methods. The experimental results show that transformer-based models can achieve competitive results on NER in Ukrainian with appropriate pre-training and fine-tuning strategies. This thesis contributes to the development of NER models for low-resource languages and provides insights into the effectiveness of transformer-based models for NER in such languages.

Contents

Chapter 1

Introduction

1.1 Motivation behind the work

Named entity recognition (NER)[2] is one of the primary tasks in natural language processing, and consists of classifying named entities in text, such as people, organizations, and locations.

Last years, with use of deep learning methods, with transformers, state-of-the-arts results have been achieved with English, Spanish and Chinese languages (ref ?)

It must be observed, that English, Spanish and Chinese, are high-resource languages, meaning essentially that a lot of computating power went into finetuning models for English, Spanish and Chinese.

However, there are much so called "low-resource" languages, as for example:

1. Basque: spoken in the Basque Country, an autonomous region of northern Spain and southwestern France.
2. Belarusian: spoken in Belarus, a country in Eastern Europe.
3. Fijian: spoken in Fiji, an island country in the South Pacific.
4. Irish: spoken primarily in Ireland and Northern Ireland.
5. Kyrgyz: spoken in Kyrgyzstan, a country in Central Asia.
6. Luxembourgish: spoken in Luxembourg, a small country in Europe.
7. Maltese: spoken in Malta, a small island nation in the Mediterranean.
8. Samoan: spoken in Samoa, an island nation in the South Pacific.

9. Scottish Gaelic: spoken primarily in Scotland.

10. Yoruba: spoken in Nigeria and other West African countries.

Effectiveness of transformer-based models in context of Ukrainian language specifically, is going to be explored in this research. Ukrainian is second most spoken language in Eastern Europe, but there is lack of high-quality annotated data and linguistic resources for Ukrainian language.

Impact of different pre-training and finetuning strategies with transformer models will be evaluated with standard metrics and benchmarks.

1.2 Questions to research

List of research questions

1. How effective are transformer-based models for NER in Ukrainian, a low-resource language

2. How effectiveness of performance is impacted by different pretraining and finetuning strategies in Ukrainian language

3. How transfomer based models compare to current state-of-the-art methods for NER in Ukrainian language

Objectives of this research are

1. Development of transformer based models using variety of pretrained and finetuned methods

2. Evaluation of models using standart NER benchmarks

3. Analyze impact of different pretraining and finetuning strategies on the performance of the models

4. Compare results to current best methods

1.3 Overview of this works report structure

This paper is build as:

1. The chapter 2 provides review of literature on NER with a focus on transformer-based models in lowresource languages

2. The chapter 3 describes methodology used in this work, including dataset and its preprocessing steps, as well transfomer models being used

3. The chapter 4 presentes results of experiments and analysis of impact of different strategies on the perfomance of transfomer-based models.

4. The chapter 5 provides findings and contributions of the work, discusses current limitations and future research

Chapter 2

Literature Review

This chapter reviews the literature on NER, with a focus on transformer-based models for NER and existing work on.

2.1 Brief review, with history of development in the field

2.2 Existing NER methods, including Machine Learning approach

2.3 Discussion about current advances for NER field

2.4 Review of related NER works in the field of low-resource languages

Chapter 3

Methodology

3.1 Dataset description and preprocessing

3.2 Overview of transformer models, including strategies of pretraining and finetunning

3.3 Metrics and procedures evaluation

Chapter 4

Results

4.1 Analysis of experiment results

4.2 Discussion on impact of model architectures, pretraining and finetuning on the perfomance of the models

4.3 Comparing results to the current achievements

Chapter 5

Conclusions

5.1 Summary

5.2 Potential for the future, current limitations

Bibliography

[1] Felix Laumann. Low resource languages, Jun 2022.

[2] Wikipedia. Named-entity recognition — Wikipedia, the free encyclopedia. http://en.wikipedia.org/w/index.php?title=Named-entity%20recognition&oldid=1136823104, 2023. [Online; accessed 01-March-2023].

[3] Wikipedia. Ukrainian language — Wikipedia, the free encyclopedia. http://en.wikipedia.org/w/index.php?title=Ukrainian%20language&oldid=1141828008, 2023. [Online; accessed 01-March-2023].