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)[3] 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[4]. 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)[3] 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

Natural Language Processing (NLP) has a section called named entity recognition (NER) that deals with extracting names, places, organizations, and dates from text. The history of NER development is extensive and spans several decades.

Rule-based techniques were employed in the 1980s to locate named items in text. These methods entailed manually establishing rules to recognize particular word patterns or word combinations that matched designated things. Unfortunately, the ability of these rule-based systems to deal with ambiguity and variety in language use was frequently constrained.

Statistical techniques started to take hold as a fresh approach to NER in the 1990s. With the use of these techniques, named entities might be automatically identified based on statistical patterns by training machine learning models on massive datasets of annotated text. The Maximum Entropy Markov Model was one of the first statistical models for NER (MEMM).

Conditional Random Fields (CRFs) sprang to prominence as a NER strategy in the early 2000s. In order to perform better on NER tasks, CRFs are a type of statistical model that can take into account the context and relationships between nearby words in a phrase.

Deep learning techniques have also recently been used to NER, notably when using recurrent neural networks (RNNs) and convolutional neural networks as neural network designs (CNNs). Results from these models have

been encouraging, especially for tasks involving big and complicated datasets.

Transformer-based models have recently revolutionized NLP and vastly enhanced NER system performance. Examples of these models include the BERT and GPT series. Transformers are a particular sort of neural network design that are very effective for NLP tasks because they can process sequences of tokens, such words or characters, in parallel.

For instance, BERT is a pre-trained transformer-based model that has been improved on a number of NER-related downstream NLP tasks. BERT has attained state-of-the-art performance on various NER benchmarks by pre-training on massive volumes of text data and then fine-tuning on certain tasks.

Transformers have also facilitated the creation of novel NER strategies, such as combining pre-trained and task-specific transformers to enhance performance on NER tasks that are specialized to a given domain.

In conclusion, transformers have become an effective tool for creating NER systems, allowing for more accurate and efficient processing of text input and fostering further advancement in the area.

2.2 Discussion about current advances for NER field

Transformer-based models have become a potent NER strategy in recent years, generating cutting-edge results on a variety of benchmark datasets. These models have been demonstrated to be very successful at modeling natural language tasks because they make use of the self-attention mechanism, which enables them to capture long-range relationships in the input sequence.

BERT is one of the most popular transformer-based NER models (Bidirectional Encoder Representations from Transformers). BERT is a pre-trained model that has been optimized for NER as well as other NER-related NLP tasks. BERT has been demonstrated to attain state-of-the-art performance on a variety of languages, including English, Chinese, and German, when fine-tuned on a particular NER dataset.

For NER, other transformer-based models like RoBERTa, ALBERT, and ELECTRA have also been created. These models expand upon the BERT architecture and enhance it in a number of ways, such as by using greater training data, more complex pre-training objectives, or more effective training techniques.

The substantial amount of computer resources needed for training and

inference is one of the difficulties of employing transformer-based models for NER. Recent studies have, however, looked into techniques to lower these models' computing costs, such as knowledge distillation, pruning, or quantization.

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

The creation of NER models for low-resource languages is a tough task, as the amount of annotated data is generally limited. Nonetheless, a number of recent studies have investigated methods to solve this issue and enhance the functionality of NER models for low-resource languages.

One strategy is to use multilingual pre-trained models that have been trained on a variety of languages, such as XLM-R or mBERT, which can be adjusted on a low-resource language with minimal annotated data. It has been demonstrated that this strategy works for a number of low-resource languages, including Vietnamese, Swahili, and Bengali.

A different strategy is to apply cross-lingual transfer learning, in which information from a language with high resources is transferred to a language with low resources. For instance, Zhang et al. (2021)[2] introduced a cross-lingual transfer learning system that converts information from English to African languages with limited resources while obtaining cutting-edge performance on numerous benchmark datasets.

Another strategy that has been investigated for NER in low-resource languages is active learning. In order to increase the performance of the model while using less annotated data, active learning entails choosing the most instructive data points to annotate. Al-Rfou et al. (2013), for instance, used active learning to create an Arabic NER system and achieved competitive performance with a minimal amount of annotated data.

Overall, there have been several promising works that have explored approaches to improve the performance of NER models for low-resource languages. However, there is still a need for further research in this area to address the challenges of limited annotated data and to develop more effective approaches for low-resource languages.

Chapter 3

Methodology

3.1 Dataset description and preprocessing

** About dataset **

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

In tasks involving named entity recognition in natural language processing (NLP), transformer models like BERT, RoBERTa, and DistilBERT have grown in popularity (NER). These models are built on a self-attention mechanism that enables them to focus on various input sequence elements and recognize intricate correlations between words.

In most cases, the transformer architecture comprises of an encoder that receives a sequence of tokens as input and generates a sequence of hidden representations that can be applied to subsequent tasks like NER. For a particular downstream job, such as NER, the encoder is typically fine-tuned on a smaller labeled dataset after being pre-trained on a huge corpus of text using unsupervised learning.

Transformer models for NER can be pre-trained and fine-tuned using a variety of techniques. Pre-training a multilingual transformer model on a sizable corpus of text from various languages, including Ukrainian, is one such method. A smaller labeled dataset of Ukrainian text can then be used to fine-tune the pre-trained model for NER.

Another strategy is to use a sizable corpus of labeled data, such as the ** Ask **

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] Peng Shi, Rui Zhang, He Bai, and Jimmy Lin. Cross-lingual training with dense retrieval for document retrieval, 2021.

[3] 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].

[4] 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].