Cross-Domain Named Entity Recognition

Mathis Gravil mvgr@itu.dk
Katarina Kraljevic katkr@itu.dk
Maria Momanu mmom@itu.dk
Sneha Shrestha snsh@itu.dk

1 Introduction

In today's world, natural language processing (NLP) has become an essential part of our everyday lives. Even with a simple use of our mobile phones, many NLP processes are triggered. One of the main tasks of NLP is named entity recognition (NER) which performs information extraction and identifies entities within a given text. In 2001, research conducted by T. Poibeau and L. Kosseim (Thierry Poibeau, 2001) proved that NER models developed on a specific domain typically don't perform well when used on a different domain. Therefore, our project aims to approach this topic by focusing on transfer learning between high and low-resource datasets. To accomplish this goal, we will utilize the CrossNER dataset, which includes the CoNLL-2003 dataset and five smaller domain-specific datasets in AI, science, politics, literature, and music.

2 Topic and Current state

The lack of annotated data for training a robust model for easy domain adaptation is a weakness of cross-domain NER that has been addressed by a few researches. One research paper from 2023 (Xiang Chen, 2023) shows that transferring knowledge from multiple source domains instead of only one improves the accuracy of a model tested on the target dataset. Another study (Nitisha Jain, 2022) focuses on the cultural heritage domain and gives a semi-structured approach to generate annotations for identifying artwork mentions from art collections.

Although current research in cross-domain NER provides valuable approaches for improving the adaptability of NER models, it is still difficult to find an ideal model that performs well in any low-resource domain.

In this project we examine different approaches to transfer knowledge from highly-sourced dataset and different other domains for training an NER model to test its performance on low-resource domain-specific target data. If successful, such approach can be used as a guide for future work with low-resource domains.

3 Research Question

In this project, we aim to answer the following research question: How can we improve the performance of cross-dataset NER models on a low-resource target dataset by using highly-resourced domain adaptation?

4 Experiment Setup

For this project, the CoNLL-2003 dataset is considered as our source dataset with high resources, while the remaining datasets as low-resource datasets. In the baseline experiments, we will test how the RNN model, pre-trained on source data, performs on each low-resource dataset and the one with the worst performance will be our target domain. The goal is to fine tune our model to work better for the target domain. The first approach will be to run an uncertainty sampling algorithm based on entropy score to find uncertain words in the target data and give human input to re-annotate them. The second approach is to add additional randomly selected training data from other low-resource datasets and then selecting data based on dissimilarity criteria measured using text similarity between two datasets. The uncertainty sampling algorithm will be used again with re-annotation of the data to increase the performance. At each step, the model's performance will be evaluated and the best model will be compared to domain-specific model that has only seen domain-specific data for training, development, and testing. This is to check if knowledge from a high-resource dataset and unrelated low resource domain-specific data can improve the performance of the state-of-the-art models in cross-domain NER scenarios.

References

Jan Ehmueller Ralf Krestel Nitisha Jain, Alejandro Sierra-Munera. 2022. Generation of training data for named entity recognition of artworks.

Leila Kosseim Thierry Poibeau. 2001. Proper name extraction from non-journalistic texts.

Shuofei Qiao Ningyu Zhang Chuanqi Tan Yong Jiang Fei Huang Huajun Chen Xiang Chen, Lei Li. 2023. One model for all domains: Collaborative domain-prefix tuning for cross-domain ner.