INFO 521 Final Project Proposal v2

Matthew Hernandez

September 29th 2024

1 Motivation

Deep learning methods dominate many modern natural language processing (NLP) tasks, however there are legitimate reasons to use classic machine learning approaches besides pedagogical reasons. This project proposal aims to implement the perceptron to solve the named entity recognition (NER) task. Not only does implementing an older algorithm offer comparative insights to cutting-edge techniques, but advancements in computation allow us to carry out previous research.

2 Project Plan

2.1 Data Acquisition

To begin, we must source the English and German dataset from CoNNL-2003.

2.2 Current Practices & Limitations

Previous research employed various supervised techniques ranging from the voting perceptron to neural-based approaches that have state-of-the-art performance; however, this comes at a cost of interpretability and computational overhead [Liu and Ritter, 2023].

We will instead focus a more interpretable method that will be augmented by feature selection and augmented by the 'kernel trick' to represent nonlinearities in the data[1]. The project is largely influenced by the work carried out by Carreras et al. [2003] during which their algorithm did not

[1] Ideally this will offer alternate and fair evaluation to early neural networks.

converge but achieved competitive results. We hope to implement a version of their algorithm that reaches convergence.

Furthermore, this application both stays true to the course objectives and fulfills personal goals, namely:

1. Algorithm can be written from scratch.
2. NER is a well-defined task and state-of-the-art performance can provide relevant results for comparison without the computational overhead needed to test them.
3. Evaluation can be plainly reported with precision, recall, and F1-score.
4. The NER task is new to me.
5. Ensemble methods and the kernel trick are new to me.
6. More experience with NLP methods on non-English data.
7. I have a soft spot for the Perceptron.

References

Xavier Carreras, Lluís Màrquez, and Lluís Padró. Learning a perceptron-based named entity chunker via online recognition feedback. In *Proceedings of the Seventh Conference on Natural Language Learning at HLT-NAACL 2003*, pages 156–159, 2003. URL `https://aclanthology.org/W03-0422`.

Shuheng Liu and Alan Ritter. Do CoNLL-2003 named entity taggers still work well in 2023? In Anna Rogers, Jordan Boyd-Graber, and Naoaki Okazaki, editors, *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 8254–8271, Toronto, Canada, July 2023. Association for Computational Linguistics. doi: 10.18653/v1/2023.acl-long.459. URL `https://aclanthology.org/2023.acl-long.459`.