Natural Language Processing and Deep Learning Project Proposal

Christian S. Fleischer
chfl@itu.dk

Hugo L. Hoydal
huly@itu.dk

Sander E. Thilo
saet@itu.dk

1 Topic Introduction

We aim to identify named entities in football transfer articles from two different online football news websites: 90min.com, which is in English (90m), and Bold.dk (bol), in Danish. The objective is recognising common named entity tags, such as person, location, organisation, etc., with the idea that this in many cases would correspond to football specific terms, like player and manager names or club or national team locations or names. Using multiple different tagger models, pre-trained on non-sports specific domains, allows us to employ domain adaptation techniques. We can then compare their abilities in tagging entities in new domains: football data in two different languages.

The premise of the project is based on one dataset extracted from kaggle.com (kag, 2023), containing the title, date, and content of 6722 articles from 90min.com. Data from Bold.dk will be extracted manually. Thus, none of our data will have tags included already. We plan to combat this by creating gold standard data sets by manually tagging the articles ourselves. This hands-on approach will ensure the creation of high-quality evaluation data sets, contributing to the reliability and authenticity of our results.

The amount of existing research relating to domain adaptation in named entity recognition (NER) is already considerably substantial. The task in various ways, with one example being Zhai & Jiang's research upon the usage of Instance Weighting (Jiang and Zhai, 2007). It is here concluded that exploiting more information through Instance Weighting in the target domain, provided more effective adaptation. Moreover, Jiang et al. addressed the challenge of cross-domain NER with limited labeled resources (Jia et al., 2019). Their study introduced an approach leveraging cross-domain language models (LM) for domain adaptation in NER tasks. By bridging domains through LM, they effectively extracted domain-specific features, facilitating knowledge transfer across diverse domains. Their method demonstrated promising results, showcasing its potential for enhancing NER performance across diverse domains. Smădu et al. have explored NER within the legal domain for Romanian and German languages (Smădu et al., 2022). The authors propose a domain adaptation method using multi-task learning, which shows slight performance gains, but significant improvements specifically in the recall metric, emphasizing its efficacy in legal entity recognition. Despite this, there seems to be a scarcity of work specifically concerning sports data. Thus, exploring the possibility of expanding the field to football transfer articles becomes compelling.

2 Project Novelty

This project seeks to provide a new scope upon domain adaptation to what has previously been worked on. Furthermore, the inclusion of danish football articles will furnish originality and unconventionality to our project. We expect that this added multilingual aspect can cause the taggers to perform in a couple of different ways: the danish data could prove an challenge to models that are not pre-trained in on data in this language, or it might make no difference at all. However, we believe that the most likely result turns to be a smoother adaptation, as the entities will be even more distinguished from the Danish language than the English.

3 Research Question

Therefore, the main objective of this project is to address the following research question:

How do named entity recognition models trained on non-sports-specific domains perform when applied to football transfer articles in two different languages?

References

90min.com. https://www.90min.com. Accessed: April 2, 2024.

bold.dk. https://bold.dk. Accessed: April 2, 2024.

2023. Football Transfer News Articles for NLP. https://www.kaggle.com/datasets/crxxom/football-transfer-news-for-nlp. Accessed: March 16, 2024.

Chen Jia, Xiaobo Liang, and Yue Zhang. 2019. Cross-domain NER using cross-domain language modeling. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 2464–2474, Florence, Italy. Association for Computational Linguistics.

Jing Jiang and ChengXiang Zhai. 2007. Instance weighting for domain adaptation in NLP. In *Proceedings of the 45th Annual Meeting of the Association of Computational Linguistics*, pages 264–271, Prague, Czech Republic. Association for Computational Linguistics.

Răzvan-Alexandru Smădu, Ion-Robert Dinică, Andrei-Marius Avram, Dumitru-Clementin Cercel, Florin Pop, and Mihaela-Claudia Cercel. 2022. Legal named entity recognition with multi-task domain adaptation. In *Proceedings of the Natural Legal Language Processing Workshop 2022*, pages 305–321, Abu Dhabi, United Arab Emirates (Hybrid). Association for Computational Linguistics.