



INFORMATICS INSTITUTE OF TECHNOLOGY

In Collaboration with
UNIVERSITY OF WESTMINSTER

Scholar Scope : Extracting Entities on CS related Research Papers using Named Entity Recognition

Thesis by

M.N. Jahzan Ahamed
2019611 | w1790143

Supervised By

Ms. Krishnakripa Jayakumar

Submitted in partial fulfilment of the requirements for the BSc in Computer Science
degree at the University of Westminster.

02nd May 2023

Abstract

Natural Language Processing has conducted a great deal of research and development on Named Entity Recognition. The development of domain-general NER models has advanced significantly, but there is still a lot of room for research in domain-specific NER models for several reasons. The lack of annotated datasets for domains, which are difficult and time-consuming to produce, is one of the main reasons of this. To address this problem, many researchers are working on numerous projects to produce datasets that are domain-specific so that NER systems can be constructed for that specific domain. Every time a dataset is developed, tests should be conducted using all available approaches to determine which one is best for that dataset. Because not all methods will perform equally well across all datasets

Computer Science is such a domain where NER is less focused because of the non-availability of standardized datasets. As it has not been tested on more recent transformer-based models, the primary goal of this research is to conduct an experiment using BERT and its appropriate variants to a novel dataset that has been created to perform NER on Computer Science related Research Papers.

Although it has been determined by this research that BERT and its variants do not provide a good performance for this dataset, pathways have been found on how state-of-the-art methods can be found for this particular dataset in the future using the results of the author's critical literature review.

Keywords: Named Entity Recognition, Natural Language Processing, Computer Science, Research Papers, BERT, SciBERT.