Graduate School of Science and Technology Master's Thesis Abstract

Laboratory name (Supervisor)	Natural Language Processing (WATANABE TARO (Professor))		
Student ID	2211418	Submission date	2024 / 7 / 22
Name	SEIVERIGHT CARGILL DUJOHN		
Thesis title	Can Large Language Models Outperform Rule-based Dataset Converters for Taxonomic Entities?		

Abstract

Extracting information from a plethora of scientific works is critical for research progress and discovery. As such, experts have labeled scientific sources to create structured datasets that are used to train a variety of models. In particular, Standoff format has been widely used as the format for labeled NER datasets. However, Standoff format is not interoperable with BERT-based models and manual conversion has been shown to be inconsistent, time-consuming, laborious, and error-prone due to issues like extra spaces and entity boundary offset. Furthermore, Standoff format labelling conventions widely differ based on the tool used. Thus, we propose a pipeline that replaces rule-based Standoff format converters by incorporating large language models (LLMs). Our results show that our pipeline can be an effective rule-based converter substitute for taxon NER datasets, which can contribute to expeditious training of task-specific models.