Thesis Project Portfolio

Machine Learning: Improving Named Entity Recognition in Healthcare and Business Sector

(Technical Report)

The Mutual Shaping of NLP Machine Translation and Society

(STS Research Paper)

An Undergraduate Thesis

Presented to the Faculty of the School of Engineering and Applied Science

University of Virginia • Charlottesville, Virginia

In Fulfillment of the Requirements for the Degree

Bachelor of Science, School of Engineering

Kevin Liu

Spring, 2023

Department of Computer Science

Contents of Portfolio

Executive Summary

The topic of Natural Language Processing (NLP) is fascinating and rapidly evolving. NLP is a subfield of artificial intelligence and computational linguistics that focuses on enabling computers to understand, interpret, and generate human language in a way that is both meaningful and contextually relevant. NLP seeks to bridge the gap between human communication and machine understanding, thereby revolutionizing the ways we interact with technology and the world around us. In this context, I have recently conducted research on the following subjects. The technical portion of the thesis aims to improve Named Entity Recognition (NER) in the healthcare and business sectors. NER is a subtask of NLP that focuses on identifying and categorizing specific entities, such as people, organizations, and locations, within a given text. By extracting these named entities and assigning them to predefined categories, NER helps convert unstructured text into structured data, enabling more efficient information retrieval and analysis in various applications. The STS portion focuses on the development and impact of Machine Translation (MT). MT is a subfield of NLP that involves automatically translating text from one language to another with a computer and without human intervention. Both research topics are connected under the broader subject of NLP.

The technical paper is derived from my internship experience at Qaultrics. Qaultrics is a software company that provides customer experience management solutions and wishes to increase its presence in the healthcare and business sectors. To upgrade their NLP engine's performance in these particular areas, I was tasked with improving the NER aspect of the pipeline, increasing the precision, recall, and F1 scores of models. I combined machine learning models with rule-based matching to improve NER scores. Since ML models are unpredictable, I used the trial and error method to determine which ML and rule-based matching combination achieves the highest scores. The models are evaluated and ranked based on their precision, recall,

and F1 scores. Additionally, I used the spaCy library to train and test NLP models, and Wikidata Query Service to obtain larger lexicons. By the end of the internship, the optimal and highest scoring combination of ML model and rule-based matching was determined for both the healthcare and business sectors.

The STS research paper answers the question, how did global trends and events shape the development of MT, and what effect does MT have on commerce, communication, and culture exchange? Through reviewing multiple scholarly articles, the three major factors that affect the direction and development of MT are political environment, commercial incentives, and cultural communication needs. On the other hand, MT's effect on society can be summarized as simply expediting communication and exchange of information, especially in the cultural and commercial worlds. Political pressure in the US focuses the attention of MT development on certain languages such as Russian, Chinese, and Arabic. Economic development and cultural diffusion require mass, speedy, and accessible translation which increases the demand for MT. In reverse, MT mutually shapes society in key aspects such as commerce and culture, allowing people to access parts of the world outside of their domain.

Together, these two research papers provide an exploration of these critical NLP tasks, showcasing their significance, current state-of-the-art, and potential future developments in both Named Entity Recognition and Machine Translation. Both research projects were successful and fruitful. They offered valuable insights into the direction and potential future of both projects. For the technical project, further work can be done to refine the most optimal performing model and prepare it for production. To reach the production standard, precision and recall scores will need to be further optimized by methods such as hyperparameter tuning. After creating a production-level model, the model should be integrated with the existing NLP engine at

Qualtrics. For the STS project, further research can be done to evaluate the side effect of the transformer model, that is Large Language Models (LLM) like ChatGPT. It is no doubt that LLM is a revolutionary technology, so future work can be done exploring this area of NLP. Both topics offer exciting and interesting possibilities for future research.