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

Social-technical development of NLP AI

A Thesis Prospectus
In STS 4500
Presented to
The Faculty of the
School of Engineering and Applied Science
University of Virginia
In Partial Fulfillment of the Requirements for the Degree
Bachelor of Science in Computer Science

By
Kevin Liu

October 27, 2022

On my honor as a University student, I have neither given nor received unauthorized aid on this assignment as defined by the Honor Guidelines for Thesis-Related Assignments.

ADVISORS

Kent Wayland, Department of Engineering and Society

Briana Morrison, Computer Science

General Research Problem

How can natural language processing artificial intelligence benefit society?

Who do you think will be replaced by artificial intelligence last? I believe most people will answer programmers. Interestingly a recently released AI program will probably change your answer to that question. The program GitHub Copilot is an artificial intelligence plugin designed to help programmers write code. Copilot interprets and predicts what the programmer wants to write and provides code suggestions that the programmer can implement. Every friend of mine that has used Copilot is shocked by its capabilities and accuracy. If something like Copilot already exists in 2021, then it would not be long until even programmers are threatened with unemployment, or will they?

Programming jobs might not be threatened until the future, but translation jobs have already been impacted by Machine Translation AI (MT). MT is a sub-field of computational linguistics and Natural Language Processing (NLP) that deals with the usage of software to translate text or speech from one language to another. The development of MT in history has been significantly influenced by society, specifically the Cold War, globalization trends, and computer manufacturing. The STS research topic of this thesis examines the impact of society on the development of MT and the effect of MT on society.

The technical topic of this thesis is closely linked to the STS topic. MT as a sub-field of NLP handles translations between one human language and another. Meanwhile, the technical research project attempts to improve computers' understanding of human language. The high-level goal of the technical project is to improve NLP AI models' understanding of corpus in the business and healthcare sectors. More specifically, the technical project aims to improve Named Entity Recognition scores by combining training datasets and rule-based matching

lexicon pipes. In a business context, this allows for a more accurate understanding of customer feedback. Therefore, reducing the amount of human capital that is invested in customer experience management.

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

What is customer experience management (CXM)? Customer experience management (CXM) is the discipline of understanding customers and deploying strategic plans to improve customer satisfaction. Are customers satisfied with the product or service? Or are they complaining about a specific issue such as battery life or customer service? Discovering and resolving customer issues is CXM. Companies on the decline can be saved from bankruptcy by performing proper CXM. On the other hand, CXM can also assist stagnant businesses to grow and beat their competitors by large margins. Most major companies include some form of CXM in their operation. Some can gather detailed data and deeply understand their customers while others can only grasp a shallow knowledge of their clients. Every year, $40 billion dollars are left on the table by companies that fail to provide “simple” experiences to their consumers (Aussant, 2022). In addition, 52% of US consumers sometimes or always walk away from purchases due to a bad customer experience. These are just a few examples that showcase the importance of CXM.

Collecting customer experience data is not a newly adopted practice in the business world. These data have been collected in the past through multiple channels. From paper and online surveys to upvoting, downvoting, and chatting with customer support. These are all channels through which customer experience data are being gathered. The recent advancement in

CXM lies in the ability to process the data gathered. Previously, someone had to manually read hundreds of reviews and possibly listen to hundreds of hours of customer support conversations. It took an immense amount of effort to correctly identify customer issues manually. Recent advancements in natural language processing have made this process much easier. Artificial intelligence can process large amounts of customer experience data that were once impossible to do manually. Hundreds of hours of call conversations can be listened to and analyzed in an instant. It can read and understand thousands of reviews and comments from multiple platforms at incredible speeds.

Qualtrics, a software company that provides customer experience management solutions, wishes to increase its presence in the healthcare and business sectors. For Qualtrics, providing valuable customer experience insights requires understanding customers first, so Qualtrics needs to improve its natural language processing (NLP) capabilities—specifically, named entity recognition (NER). NER is a subtask of information extraction that identifies and classifies entities mentioned in text into pre-defined categories such as names, organizations, locations, etc. The goal of my project at Qualtrics was to improve NER precision, recall, and f-1 scores in the business and healthcare sectors. Different sectors have different vocabulary, punctuation, and sentence structure, so another part of the project was also to find out if models can be created to function across both sectors. 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. 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 is determined for both the healthcare and business sectors. For the

future of the project, the model should be further refined to improve the precision and recall scores, so it can be deployed in the production environment.

STS Question: The Mutual Shaping of NLP Machine Translation and Society

I am certain almost every person reading this Prospectus has heard of Google Translate and most of us have probably tried it at least once in our lives. If you have used Google Translate in recent years after 2020, you might have been surprised by its accuracy and convenience. With the recently updated Google Translate, we can take pictures of a menu in a foreign language such as Japanese, and see the translation in English. I rely on Google Translate extensively during my trips to foreign countries, from reading road signs to interacting with strangers. Machine translation services such as Google Translate were never this accurate until recently. In fact, just ten years ago when I moved to the US and used Google Translate for homework readings, I could barely understand the translation from English to Chinese. Interpreting the translation was a learning process on its own. Machine translation has evolved dramatically since its conception in the 1950s, with significant progress made in the past decade.

How Society Shaped the Development of Machine Translation?

The concept of Machine Translation first debuted to the public in 1954 at IBM's New York office (Gordin, 2014). The debut, later called Georgetown–IBM experiment, caught the world's attention and impressed many experts in linguistic and computer science. Many other countries began tackling the challenges of Machine Translation and developing their own systems. In the beginning, US and USSR dominated the MT field. The world stage at that time was still in the middle of the Cold War. This meant the US and USSR competed in every aspect,

including MT. The competition drove both sides to devote more budget and human resources to this subsection of NLP. The Georgetown MT project received the most funding out of any NLP projects at its time. It was funded by the National Science Foundation (NSF) as well as the Central Intelligence Agency (CIA). The US intelligence service wanted to understand Russian communication and more importantly Russian scientific advancements. During the Cold War period, the USSR made it illegal to publish scientific papers in English in an attempt to prevent Western nations from stealing information. This resulted in the US strongly promoting the education and translation of "scientific Russian". Scientific Russian is significantly different from typical conversational Russian. With simpler grammatical structures and rules, it should be easier for a computer to translate. Near all early MT projects in the US focused only on scientific Russian to English translation.

After more than 10 years of research and $20 million dollars, in 1964, the US Automatic Language Processing Advisory Committee (ALPAC) determined that MT was impractical and not worth the trouble or expense. This essentially halted the US's development of MT, but other countries carried on where the US stopped (Pestov, 2022). Post WW2, the wave of globalization swept over every country; however, the language barrier hindered many's progress, especially in places where English was not spoken. Nations like Japan were extremely interested in MT since it provided a solution to the language barrier. The US used MT exclusively for translating scientific Russian into English, whereas other countries focused on translation between their language and English.

The Cold War sparked MT development, while globalization fueled the fire. In the past 70 years, MT technology has evolved significantly, from dictionary translation to statistical rule-based translation to deep neural network translation (Gordin 2014). Neural network

translation was previously discussed, but only recently achieved. This is in large part due to the limitations of computers. Machine Learning capability is always capped by computing power. The growth trend of computing power in the world roughly follows Moore's law. This means that the computing power of the world grows at an exponential rate (Tardi, C). The increase in computing power and the drop in computing power per dollar have promoted significant advancements in ML. With cheaper and more powerful computers than ever in history, Google attempted to utilize Deep Learning (DL) for MT and achieve phenomenal results. In 2014, Google began the development of DL MT, and in two years, by 2016, the DL MT model surpassed every MT project (Jones 2021). From incomprehensible gibberish to coherent language translation, MT has made incredible advancements in the past 70 years.

From war to global trends to computer manufacturing, the progress and development of MT are closely tied to society. In my research, I plan to focus on specific advancements and time periods of MT. I will examine articles about the Georgetown MT project and its relation to the Cold War environment. In addition, I will read articles about the development of MT in other countries and analyze the effect of globalization on MT; specifically, I will focus on the time period after ALPAC terminated MT development in the US. I will also research the impact of accurate and convenient MT on society, specifically, how the translation market has changed and how the role of human translators has changed. I will also read articles about how MT has assisted and increased international interactions.

Conclusion

Both technical and STS topics are centered on NLP. Through the technical internship project, I gained experience working with NLP technology and without that opportunity, I probably would never have been exposed to it. I hope to learn more about the mutual shaping of MT and society through STS research. The accessibility and quality of MT can have a profound impact on society. The STS research hopes to shed light on those who have been positively and negatively affected by this technology. Artificial intelligence's effect on society is tremendous, to say the least. AI has created new jobs, new sectors, new research, and even new ways of making decisions. Some are optimistic and see a future where humanity is freed from demeaning work and world productivity skyrockets. In contrast, others envision a dystopian society where our lives are jeopardized by the technology we created. No one can see the distant future, but we can try our best to point it in the right direction. It is quite astounding that we have achieved such dramatic and incredible progress in AI technology in roughly 70 years. We have come a long way in the development of NLP artificial intelligence, and we still have a long way to go.

References

Aussant , P. (2022, April 3). Top 35+ customer experience statistics to know in 2022. Emplifi. Retrieved October 26, 2022, from https://emplifi.io/resources/blog/customer-experience-statistics

Brooks, R. (2021, July 15). *A brief history of machine translation Technlogy*. K International. Retrieved November 28, 2022, from https://www.k-international.com/blog/history-machine-translation/

Foerst, A. (2022, July 1). The ethics of AI and robotics. *Media Development* (3), 15 - 20.

Gordin, M. D. (2014). The Dostoevsky Machine in Georgetown: Scientific Translation in the Cold War. *Annals of Science*, *73*(2), 208–223. https://doi.org/10.1080/00033790.2014.917437

Jones, G. (2021, September 15). *History of machine translation*. The Translator's Studio. Retrieved November 28, 2022, from https://translatorstudio.co.uk/machine-translation-history/

Kolmar, C. (2021, October 12). 23+ artificial intelligence and job loss statistics. Zippia. Retrieved October 26, 2022, from https://www.zippia.com/advice/ai-job-loss-statistics/

Nicol, V. (2022, August 12). Translation industry trends and statistics. My Language Connection. Retrieved October 26, 2022, from https://www.mylanguageconnection.com/translation-industry-trends-and-statistics/#:~:text=Translation%20Industry%20Statistics,America%20follows%20this%20at%2039.41%25.

Pestov, I. (n.d.). *A history of machine translation from the Cold War to deep learning*. freecodecamp. Retrieved November 28, 2022, from https://www.freecodecamp.org/news/a-history-of-machine-translation-from-the-cold-war-to-deep-learning-f1d335ce8b5/

Semuels, A., De La Garza, A., & Zorthian, J. (2020, August 17). Fewer Jobs, More Machines. *TIME Magazine*, 196(7/8), 64 - 71.

Tardi, C. (2022, October 8). What is Moore's law and is it still true? Investopedia. Retrieved October 21, 2022, from https://www.investopedia.com/terms/m/mooreslaw.asp#:~:text=In%201965%2C%20Gordon%20Moore%20posited,and%20more%20efficient%20over%20time.