An Overview of Natural Language Processing

M. Danish

Computer Science, CSU Global

CSC525, Principles of Machine Learning

Dr. Joseph Issa

01/22/2023

I first began learning about Natural Language Processing technology in May of 2017. I had just graduated from Middle Tennessee State University with my B.S. in Organizational Communication—a subfield of Communication Studies—and taken a temporary job as a freelance UX Copywriter for a Denver-based agency that worked on Samsung's AI and chatbot, Bixby. At that time, I had limited experience with programming; most of what I knew came from working in MS-DOS on my family's Windows 95 and 98 systems as a young teenager, and then taking Visual Basic and HTML Web Design courses in High School. I didn't get any direct exposure to Bixby's backend during my time on that project—and couldn't talk about it if I had, thanks to the NDA I signed—but as I learned more about NLP on my own, I started to understand how the copy I was writing could fit into an NLP model and how it was probably being used by Bixby's developers.

In my current role—where I recently progressed from Technical Writer to Enterprise Services Process Analyst—I am working on chatbots and AI functions of my own that incorporate NLP. Although (almost) the entire company has been remote since the COVID-19 pandemic began—and was mostly remote even before that—my current employer's culture is still very centered around word-of-mouth interactions. Most of my colleagues don't want to be referred to pages of documentation, policies, etc. They want to talk to a person who can help them find things and guide them through unfamiliar, rarely-used processes. That's what drove me to look for ways to develop a friendly, easy-to-use, conversational chatbot assistant of my own, and to look for other ways to introduce my colleagues to the NLP features available in Microsoft Power Platform applications (such as Power BI). In the following sections, I will discuss NLP's effect on other areas of the AI field, machine learning techniques used in NLP, possible future applications of NLP, associated ethical and social concerns, and future uses for NLP technology.

NLP's Effect on the Field of AI

NLP has had effects on robotics, decision support systems, chatbots, and the software user experience in general. In all examples, it has made—and has the potential to make—AI systems more accessible to the layperson by breaking down the jargon barrier that is so often encountered with any computer technology. While there is still plenty of room for improvement, NLP has reduced the user's need to memorize specific commands and menu navigation actions by allowing them to interact with computer and robotics systems in a more casual, flexible, and conversational way.

Machine Learning Techniques in NLP

The most common machine learning techniques used in NLP are stemming and lemmatization (Srinidhi, 2020). Stemming is a technique that reduces a word to its root form by truncating it, which allows NLP models to learn vocabulary by morphology, in much the same way human children in English-speaking countries study it as part of early education (Freeman et. al, 2019). Lemmatization is a more accurate technique that uses known root words to find synonyms and antonyms (Srinidhi, 2020).

A recent development in machine learning that is also used in NLP is a technique called "attention," which weights parts of datasets according to their importance. The attention technique allows for significantly better machine translation between languages. For example, in one study, it improved English-to-French translation from a dropout rate of 0.3 to a rate of 0.1 *and* cut the training cost by 75% (Vaswani et. al, 2017).

Future Uses

Lemmatization is something I'm already using in a Python program I'm developing that will help me generate alternative phrasings for use in training chatbots that have machine

learning features. For example, Power Virtual Agents asks the developer to enter five to 10 “trigger phrases” for a given conversation flow. It can then use machine learning to extrapolate other ways users might ask the same question, as well as other questions that should lead to the same conversation topic. Using lemmatization to generate the four to nine additional trigger phrases from just one phrase entered by the developer could speed up the chatbot training process in turn and reduce the amount of human labor needed to generate the sample phrases.

I can also see potential for the attention technique in my own line of work. One Microsoft Power Platform application I work with a lot is Power BI. Power BI has an NLP feature that allows users to ask questions about the data in reports. Although I know less about the development side of it than I do with Power Platform applications, ServiceNow is another application I work in that offers an NLP feature in its report-creation tools. I think the attention technique could be applied to both Power Platform apps and ServiceNow to improve AI-generated reports and summaries, as well as to better answer users’ questions about data. Additionally, it could improve chatbot interactions with knowledge base information that are possible with the Power Virtual Agents chatbot development tool.

Future Applications of Natural Language Processing

NLP encompasses a variety of specific language-related tasks. These range from basic tasks such as identification and prediction, grouping and categorization, stemming and lemmatization, and word embedding to part-of-speech recognition, chunking, named entity recognition, voice processing, language generation, and other complex categorization and prediction tasks (Meyer, 2021).

Future applications seem infinite. Many NLP tasks are already featured in our favorite word processors, grammar correctors, plagiarism checkers, and even our smartphones’ predictive

text modules. Machine learning will only make AI-generated language better, more coherent, and more natural sounding. Voice recognition and simulation in particular are already becoming ubiquitous—Computer generated voiceover is now a standard feature on TikTok, with other internet content creation platforms undoubtedly soon to follow.

Ethical and Social Concerns

As we've seen with the recent trend in AI-generated visual art (Rogers, 2022), there are several ethical and social concerns that can be associated with NLP. First, AI can be used to plagiarize human artists and writers. Similarly, it's all too easy to plagiarize from AI itself—in fact, Business Insider (Syme, 2023) recently reported on a new app that's been developed to detect plagiarism from the ChatGPT AI in college essays.

Another ethical concern is that of bias; e.g., the variety of grammatical frameworks, vocabularies, and other mannerisms related to speech and writing that occur just within the English language depending on *who* is writing or speaking and what their racial, ethnic, and socioeconomic backgrounds are. We have seen an example of that come up already in facial recognition as it fails to correctly identify dark-skinned people due to hard-coded biases in the algorithms that power it (Crockford, 2020). Considering the well-documented phenomenon of bias and discrimination against both people of color and people from disadvantaged socioeconomic backgrounds based on their natural speech and writing patterns, it's not hard to see how NLP could be applied to existing discriminatory practices.

Finally, there's the matter of informed consent. Do users understand what is being collected for NLP uses? Has the way it will be used been explained to them in a way that maximizes comprehension for the layperson? Will any of their data be sold or otherwise released to third parties, with or without their knowledge? What happens to their data after it's been

processed by the NLP model? Are there any ways their data could be used against them or put them at risk?

Conclusion

NLP has affected other areas of AI by making it easier for—and more accessible to—laypersons. The most common machine learning techniques used with NLP are stemming and lemmatization, which allow NLP models to get better at understanding root words, synonyms, and antonyms. Attention is a more recently-developed machine learning technique that is used in NLP, and it allows parts of datasets that include—but are not limited to—language samples to be weighted according to importance, which helps NLP get better at understanding context and relevance.

Lemmatization—which is more accurate than stemming—and attention have the most potential for future use in NLP. NLP itself has seemingly limitless potential for future use, as it can be used to improve any type of software interface that handles language. Ethical and social concerns for the future of NLP include plagiarism and intellectual property rights, bias and discrimination, and informed consent with regard to data protection.

References

Crockford, K. (2020, June 16). *How is face recognition surveillance technology racist?* ACLU.org. Retrieved January 21, 2023, from https://www.aclu.org/news/privacy-technology/how-is-face-recognition-surveillance-technology-racist

Freeman, N. D., Townsend, D., & Templeton, S. (2019). *Thinking about words: First graders' response to morphological instruction.* The Reading Teacher, 72(4), 463–473. https://doi.org/10.1002/trtr.1749

Meyer, P. (2021, October 19). *Natural Language Processing tasks*. Towards Data Science. https://towardsdatascience.com/natural-language-processing-tasks-3278907702f3

Rogers, R. (2022, December 9). *What you should know before using the Lensa AI app.* Wired. https://www.wired.com/story/lensa-ai-magic-avatars-security-tips/

Srinidhi, S. (2020, February 26). *Lemmatization in Natural Language Processing (NLP) and Machine Learning*. Towards Data Science. https://towardsdatascience.com/lemmatization-in-natural-language-processing-nlp-and-machine-learning-a4416f69a7b6

Syme, P. (2023, January 4). *A Princeton student built an app which can detect if ChatGPT wrote an essay to combat AI-based plagiarism.* Business Insider. https://www.businessinsider.com/app-detects-if-chatgpt-wrote-essay-ai-plagiarism-2023-1

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, L., & Polosukhin, I. (2017). *Attention is all you need.* In *arXiv [cs.CL]*. arXiv. http://arxiv.org/abs/1706.03762