Capabilities and Advancements of Natural Language Processing

Oceana Li '27

Introduction

As the use of artificial intelligence (AI) technology rises in the 21st century, so does the need for a machine's ability to comprehend human language. NLP (Natural Language Processing), where linguistics and AI intersect, incorporates deep learning techniques to teach machines to interpret and generate human language (Stryker et al, 2024). State-of-the-art technologies and big-tech companies have utilized NLP for years, developing LLMs (large language models) and popular tools like Chat-GPT, Google Translate, and Grammarly. Recent advancements include Google's BERT model (a transformer-based model) and OpenAI's GPT-4o, which have revolutionized the applications of artificial intelligence in a multitude of fields. This paper will discuss and explore the inner processes of NLP and its recent breakthroughs and applications.

How Natural Language Processing Works

For a machine to interpret human-inputted text, it must undergo a series of word-normalization tasks and word-embedding processes that rearrange and translate the text into an input that a machine can understand. After the machine receives a text, the section of the text is broken down into separate words, undergoing a task known as "tokenization." "Lemmatization," on the other hand, identifies similar words by removing suffixes (Jurafsky et al, 2024). For example, *created*, *creating*, and *creates* are lemmatized into the base word *create*. These words are then converted into numerical data that can be processed through an algorithm, allowing the machine to make predictions. This process, OHE (one-hot encoding), pairs an index

value with a word and creates a binary vector. The vector, or sequence of numerical values that represent categorical data, of zeros and ones is unique to each word (Imran, 2023). To capture the frequency of a word in a sentence, BoW (Bag of Words) ignores syntax and grammar and sums up the binary vectors to produce a final vector. This vector represents the number of times each word in the vocabulary occurs, and the machine may make interpretations based on this result. OHE and BoW are limited in their restrictive use. They are only able to indicate the presence and frequency of a word rather than its relationship with other words. Furthermore, treating each word as an independent element and ignoring its syntax and context can result in an inaccurate analysis by the machine. For example, the clauses “The service was good but the food was not” and “The food was good but the service was not” may lead the machine to label the clauses as equal in meaning due to the identical word frequencies, yet the clauses are completely opposite in meaning.

Fortunately, word embedding techniques like Word2Vec can overcome the challenge of capturing word order. Word2Vec compares words based on the contexts in which they appear; similar words appear in similar contexts. By this premise, the word “dog” when paired with “animal” has a higher weight value or strength of connection than when paired with “man.” Unlike BoW, this word-embedding technique evaluates word meaning by considering surrounding words instead of analyzing the individual word itself. For Word2Vec to produce accurate results, it needs a large amount of training data to learn the common patterns and contexts of a word. Sources often include Wikipedia articles, social media posts, and any forms of human-produced text.

Applications of NLP

NLP is used in multiple fields ranging from biomedical to legal due to its proficiency in completing tasks such as sentiment analysis, language translation, and other AI-powered functions. In finance, NLP is used to make stock predictions through sentiment analysis of social media posts. Similarly, enterprises use these tools to streamline daily processes by taking advantage of text summarization and installing AI assistants. Even more popular products include those released by big-tech companies like Google Translate and Chat-GPT, used by millions of people worldwide. These developments efficiently convert or generate text with the combination of large datasets and advanced NLP techniques.

Neural Networks and Transformers

Initially, machines used deep-learning models like CNNs (convolutional neural networks) and RNNs (recurrent neural networks) that simulate the neurons of a human brain to process natural language. These networks consist of an input layer, intermediate or hidden layers, and an output layer. Inputs are assigned a weight, which is a value that measures its importance to the output. In addition to these layers are functions that optimize the prediction-making process. A few examples include an activation function, often applied to determine what causes a certain output; a loss function, which measures a model's performance in classifying data; and the pooling function, which simplifies data to prevent *overfitting*—when large, complex amounts of data actually harm a model's accuracy. While these neural networks excel at tasks like image recognition, they are limited in data collection due to the amount of time taken to train and process text.

Instead, NLP has shifted towards transformer-based models, neural networks that use "self-attention" and an encoder-decoder structure. While encoding layers convert text into their embeddings, decoding layers convert embeddings from the encoding layers back into a sequence of words. However, both layers each consist of a fully-connected FNN (feed-forward neural network), neural networks that are fast in training and processing data and "self-attention" (Ankit, 2024). *Self-attention* is a mechanism that focuses on a single word and assigns weight values to its surrounding words that represent how important the word is in relation to them. This process is repeated for each word in the clause. Then, the sum of these weights produces a final "context-aware" vector that allows the transformer to better prioritize which pieces to focus on by understanding the reliance and relationship between each word (Verma, 2023). In order for this vector to be processed by a machine, an activation function is applied in the output layer. The softmax function, for instance, is a mathematical equation that turns these values into probabilities that add up to one.

Recent Models

A popular transformer model recently released by Google is BERT (Bidirectional Encoder Representations from Transformers), a pre-trained LLM that specializes in understanding context and classifying text. Since it is bidirectional, it is able to read text both left and right, meaning it does not have to take the time to process each word sequentially. In addition to transformer encoder layers, BERT consists of additional layers like pooling and position embedding which translates the data into a sequence that represents each element's specific location (Maverick, 2023). Ultimately, the combination of all these layers makes this LLM capable of advanced processing and comprehension. BERT is also preferred for its high

accuracy and efficient training time. Thus, common applications range from optimizing search engines to identifying emotional tone of sentiments. BERT is also being utilized in other fields including science and medicine, with variants such as SciBERT and BioBERT (Casey, 2023).

Although BERT is exceptional in understanding existing text, it is inferior to OpenAI's most recent release, GPT-4o, when generating new information. These generative models are trained through "autoregressive language modeling," an approach that uses past data to make predictions based on probability formulas (Mehra, 2023). There are many supposed improvements, such as more realistic human interactions and complex text generation in more than 50 different languages. In addition to GPT-4o's improvements in response time and cost-effectiveness, what distinguishes the model from GPT-4 is its "multimodality," which allows users to enter many types of inputs ranging from audio to images (Craig, 2024). ChatGPT is also now available in other languages besides English after OpenAI implemented new NLP techniques in the latest model. To accommodate multiple languages that use different alphabets, developers improved and created new tokenization functions for these languages. In hopes of making NLP products like ChatGPT more globally accessible, it is important for these machines to support a variety of languages.

Conclusion

The relevance and importance of NLP across various fields increases as long as the age of artificial intelligence progresses. Language translation, chatbots and AI-assistants, sentiment analysis, and other products that utilize NLP techniques are utilized daily by individuals and larger enterprises. Through word-normalization processes to layers of neural networks, machines are able to replicate and understand human language. Advanced LLMs continue to emerge, and

existing transformer models like BERT and GPT-4o further develop. As a result, machines carry out tasks that involve high-levels of rapid computation, conquering the long-standing “language barrier” between humans and machines.

References

Alam, M. F. (2022, September 8). Applications of Natural Language Processing. *Datascience Dojo*. https://datasciencedojo.com/blog/natural-language-processing-applications/

Alaoui, S. (2023, July 21). Advancements in Natural Language Processing (NLP) and Future Expectations. *Medium*. https://medium.com/@soukaina./advancements-in-natural-language-processing-nlp-and-future-expectations-33bec2a42d14

Ali, M. (2024, September 12). *Introduction to Activation Functions in Neural Networks*. Datacamp. Retrieved October 6, 2024, from https://www.datacamp.com/tutorial/introduction-to-activation-functions-in-neural-networks

Ankit, U. (2024, May 24). Transformer Neural Networks: A Step-by-Step Breakdown. *Builtin*. https://builtin.com/artificial-intelligence/transformer-neural-network

Casey, M. (2023, December 27). BERT models: Google's NLP for the enterprise. *Snorkel AI*. https://snorkel.ai/bert-models/#:~:text=BERT%20stands%20for%20Bidirectional%20Encoder,used%20for%20various%20NLP%20tasks.

Craig, L. (2024, July 26). GPT-4o vs. GPT-4: How do they compare? *TechTarget Enterprise AI*. https://www.techtarget.com/searchenterpriseai/feature/GPT-4o-vs-GPT-4-How-do-they-compare#:~:text=OpenAI's%20testing%20indicates%20that%20GPT,idioms%2C%20metaphors%20and%20cultural%20references.

Imran, R. (2023, June 11). Comparing Text Preprocessing Techniques: One-Hot Encoding, Bag of Words, TF-IDF, and Word2Vec for Sentiment Analysis. *Medium*. https://medium.com/@rayanimran307/comparing-text-preprocessing-techniques-one-hot-encoding-bag-of-words-tf-idf-and-word2vec-for-5850c0c117f1

Jeet. (2020, August 14). One Hot encoding of text data in Natural Language Processing. *Medium: Analytics Vidhya*. https://medium.com/analytics-vidhya/one-hot-encoding-of-text-data-in-natural-language-processing-2242fefb2148

Jurafsky, D., & Martin, J. H. (2024). *Speech and Language Processing An Introduction to Natural Language Processing, Computational Linguistics, and Speech Recognition with Language Models* [PDF]. https://web.stanford.edu/~jurafsky/slp3/ed3book.pdf

Maverick, A. (2023, February 24). BERT — Bidirectional Encoder Representations. *Medium*. https://samanemami.medium.com/bert-bidirectional-encoder-representations-e98833f9dfcd#:~:text=BERT%20uses%20a%20series%20of,meaning%20of%20words%20and%20sentences.

Mehra, A. (2023, May 25). A Deep Dive into GPT Models: Evolution & Performance Comparison. *KDnuggets*. https://www.kdnuggets.com/2023/05/deep-dive-gpt-models.html

Stryker, C., & Holdsworth, J. (2024, August 11). *What is NLP (natural language processing)?* IBM. Retrieved October 6, 2024, from https://www.ibm.com/topics/natural-language-processing

Verma, A. (2023, June 14). Self-Attention Mechanism Transformers. *Medium*. https://medium.com/@averma9838/self-attention-mechanism-transformers-41d1afea46cf