NLP Research Paper

Group: Themis

Course Name: Artificial Intelligence in the World

Course# 94010100

Abrantes

October 24, 2024

Abstract:

In this paper we are going to discuss NLP in its entirety, the techniques, the applications, and the challenges and future directions NLP will face.

Introduction to Natural Language Processing:

NLP is a field of artificial intelligence that works and focuses on how computers and humans interact with one another using natural language. It's about teaching the machines how to understand and create human language, which would make it possible for them to process and analyze large amounts of natural language using the data they created through those teachings. This has been used in a various number of applications, from chatbots to virtual assistants being used to translate.

History and Development of NLP

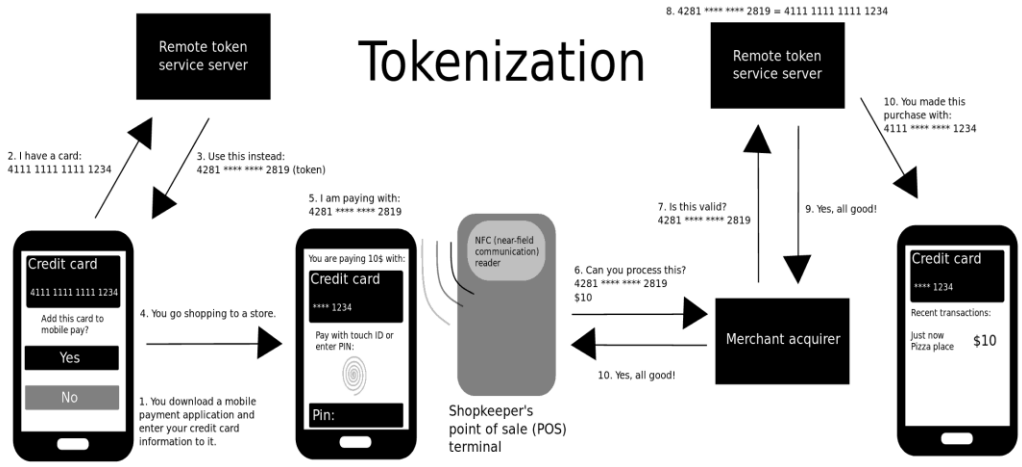
The part of (Natural language processing) was (Language as a Science) which was almost. lost to time when its creator died. In 1957 Syntactic Structures was published by Noam Chomsky. The book “revolutionized linguistic concepts and concluded that for a computer to understand a language, the sentence structure would have to be changed. With this as his goal, Chomsky created a style of grammar called Phase-Structure Grammar, which methodically translated natural language sentences into a format that is usable by computers. (The overall goal was to create a computer capable of imitating the human brain, in terms of thinking and communicating – artificial intelligence.)” Unfortunately, in 1966, the NRC and ALPAC caused the first major AI and NLPresearch tohalt by cutting off funding for studies on n atual language processing and machine translation. The research got too expensive, and their Tech cloud does not keep up. In 1966 artificial intelligence and natural language processing (NLP) research was considered a dead end by many (though not all). After 14 years AI returned. In the 1980 Natural Language Processing got a big upgrade with the introduction of probabilistic models. The focus. In this time of AI. Were statistical models. It became a huge focus. With IBM creating a few successful ones. In the 2000s, machine learning got a. larger update with Them now being able to use NLP At an advanced level. allowing systems to learn patterns from large datasets. One of the 1st the world’s first successful NLP/AI assistants. Was Siri She was created in 2011. nowadays in this new era of AI.

Key Concepts in NLP:

4.1 Tokenization:

A process in NLP that breaks down text into smaller units called tokens. This action of breaking it down into more manageable pieces helps the machines process, understand and create

human language a lot easier than before when it was in bigger pieces. A better explanation of what the tokens are generally used for is they are used to represent text document vectors, which are then used for NLP tasks such as sentimental analysis, name entity recognition, and text classifications.



4.2 Part-of-Speech Tagging:

Part-Of-Speech Tagging, (POS) tagging is a method that an NLP uses to give a response from a prompt with high accuracy. POS tagging works by having the A.I scan though each word in the prompt and give a part of speech tag with the context that is given and then give a response to the prompt. For example, the sentence," I like learning French" is given to the A.I and the A.I will scan the words and find the tag for" I" a pronoun, the word" like" a verb, the word" learning" a noun, and the word" French" a Noun. After tagging each word, the A.I will process a prompt to respond to this statement such as" That is nice, would you like to learn more about French?" The reason an NLP uses POS tagging is so the A.I cannot get confused with words with two different parts of speeches and give the wrong answer. As used in the example, the word" French" can be used to describe a language of France or a person from France.

4.3 Named Entity Recognition:

Named Entity Recognition (NER) is a technique in natural language processing that focuses on identifying and classifying information in unstructured text. The reason we use NER is to automatically take structured information from unstructured text. This helps machines understand and categorize entities in a helpful manner for many different applications like text summarization, building knowledge graphs, and Q&A's. The way NER works is first the NER system analyses the whole input text to look for and find named entities. The system then identifies the sentence structure by taking into account capitalization rules. This means that it will know when a sentence ends when the word after it starts with a capital letter. Knowing sentence structure helps the NER contextualize entities within the text which helps the model understand relationships and meanings. NER then deploys machine learning algorithms to analyze the labeled datasets. The datasets contain examples of annotated entities which help the model recognize similar entities in new data.

4.4 Sentiment Analysis:

Sentiment analysis is the process of classifying if a string of text is positive, negative, or neutral. The goal is to analyze people's opinions in a way that helps businesses expand. It focuses on polarity (positive, negative, and neutral) and emotions (happy, disappointed, mad). This uses NPL algorithms like rule-based, automatic and hybrid. Places where sentiment analysis could help businesses is Customer Feedback Analysis. Businesses can analyze customer reviews, comments, and concerns to understand how the customer thinks so they can improve and address the customer's needs. It can also help with brand reputation because sentiment analysis allows

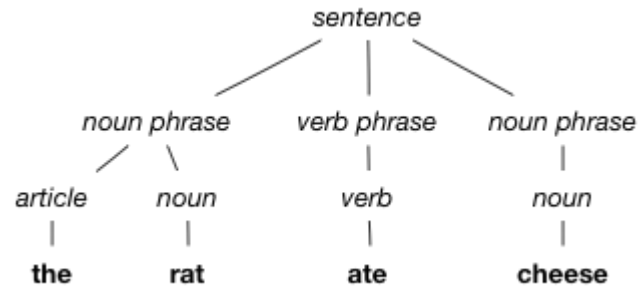
businesses to track their mentions and comments on social media in real time so they can respond quickly to positive negative comments. Sentiment analysis also helps with product development and research because understanding how a customer feels about a certain product helps the business to know what works and what needs improving.

4.5 Stop Words:

Frequently used words that are removed from (NLP) because they have little meaning and do not contribute too much to the context of a text. It was added in to prove the efficiency and accuracy of (NLP).

4.6 Parsing

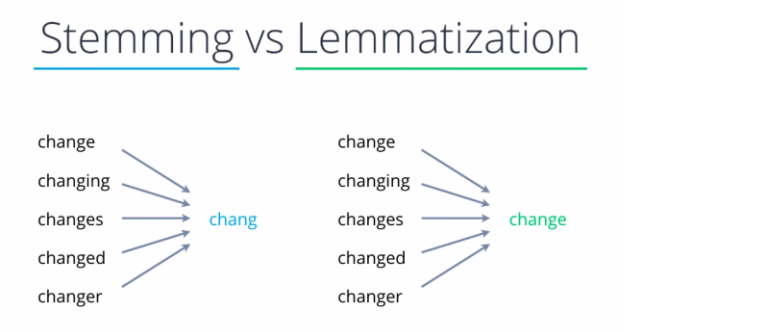
This is the grammatical analysis of a sentence structure. For example, a (NLP) algorithm is fed the sentence. “The cat meow” parsing in volves breaking the sentence into parts of speech ~ i.e., cat is an noun and meow is a verb.



4.7 Stemming and Lemmatization

Stemming and lemmatization are both used to reduce words to their base in order to simplify text analysis. They are commonly used in NLP tasks such as mining, information retrieval and sentiment analysis. They reduce the variability of words in a dataset which can improve the performance of machine learning algorithms by reducing the number of tokens. While both are used in natural language processing to reduce words to their base form they operate differently.

(geeksforgeeks) Stemming involves cutting down words to their base or root form, which may not be a valid word. For example, "running" might be stemmed to "run," but "better" would be stemmed to "better," as it doesn't have a simpler root. Some benefits in stemming include its speed which is typically faster because it uses simple rule-based approaches and requires less computational resources making it easier to implement. On the other hand, the stemmed form of the word may not be a valid word which can lead to confusion, and it does not consider context which can oversimplify the meaning of words.



(Toporkov & Agerri 2023) Lemmatization reduces words to the dictionary form or their lemma. Like in the earlier example "running" becomes "run," and "better" becomes "good." In lemmatization it considers the context and part of speech, ensuring that the root form is a valid word. This allows for more accurate applications where understanding is important, like in-search engines such as Google and Bing because the dictionary forms ensure accurate representations of the words. While this is true, lemmatization is generally slower because it uses more resources due to its linguistic analysis and may require complex algorithms and linguistic resources such as dictionaries or databases. Additionally, in some cases lemmatization can still struggle with ambiguous words where context is crucial.

Applications of NLP:

5.1 Machine Translation:

(Hu, J. 2022) defines machine translation as the task of translating text from a source language to another target language. NLP facilitates this translation through statistical models and neural networks. Essentially, the statistical model works by the identification of patterns from large database sets that can allow for more accurate translations between languages. While this is still in effect today, more modernized translations use deep learning (Neural Machine Translation) to translate more effectively. The NMT neural network is inspired by the human

brain, and it consists of multiple layers that are interconnected like how a neural network in the human brain would be. There are different types of neural networks systems used to increase the accuracy of machine translation: Recurrent Neural Networks (RNNs), Long Short-Term Memory (LSTM) Networks, Transformers. Each of these are designed with specific traits that try to help improve the understanding of the model to interpret the meaning of the words. RNN's are suitable for text processing and maintain the context of sequences of words. LSTMs are designed to overcome the limitations of traditional RNNs, essentially, they can remember information for longer periods.

Ethical Implications of NMT:

NMT systems often require vast amounts of data for training, which can include sensitive information from users. If this data is not handled correctly, it can lead to breaches of privacy. Additionally, it is crucial to ensure that personal identifiers are removed from datasets. However, complete anonymization can be challenging, and there is a risk that user data could be re-identified. This is why developers must implement stringent data protection measures, such as encryption and secure storage, to safeguard user information. They should also be transparent about data usage and obtain informed consent from users. Developers should not be the only ones with responsibility as users should be aware of what data is being collected and how it will be used, advocating for their privacy rights when using NMT services.

5.2 Chatbots and Virtual Assistants

Chatbots are automated programs that can engage in text or voice conversations with users. They can be integrated into websites, messaging platforms, or apps. Virtual assistants are more advanced AI systems that can perform a wider range of tasks beyond simple conversations. They often integrate with various services and devices to help users manage tasks. Rule-Based

Chatbots Operate based on predefined rules and scripts. They respond to specific commands and keywords but may struggle with complex queries. AI-Powered Chatbots utilize machine learning and NLP to understand and generate responses. They can learn from interactions, improving their performance over time. They can be used in customer support to answer frequently asked questions, e-commerce to assist customers with things like order tracking or recommendations, and entertainment by engaging users in storytelling within games.

5.3 Text Summarization

Text summarization is the process in which the essential points of a text are pulled from the text. In this process irrelevant text is removed from the prompt for better analysis through condensing meaning. There are two types of summarizations: extractive and abstractive.

Extractive summarization involves selecting key sentences directly from the source text to create a summary, ensuring that the most important points are retained in their original wording. On the other hand, abstractive summarization involves generating new sentences that capture the essence of the source text, similar to how humans summarize by rephrasing and condensing information. While extractive summarization is straightforward and preserves the source text's original language, abstractive summarization is more complex and aims to provide a more coherent and human-like summary by interpreting and rewording the content.

The limitations with text summarization are the information loss and the variety of text structure. Information loss can occur during summarization process which can lead to critical details being omitted and the determination of which parts of the text are most relevant for inclusion in the summary can be context dependent subjective. As for variety of text structures, there are different formats the prompt texts can come in and the differences in structure and style can

make Standarization difficult. Additionally, the length and complexity of the prompt may impact summarization as well.

While these limitations are true, text summarization is a valuable tool in today's information-driven world, offering significant benefits in efficiency, comprehension, and accessibility. Often seen in News articles like Google News that provide summaries of articles and social media platforms that summarize lengthy online discussion.

Challenges and Future Directions in NLP:

NLP (Natural Language Processing) faces significant challenges and offers exciting future directions. One major challenge is accurately interpreting the nuances and complexities of human language, including slang, sarcasm, and context. Another is the ethical use of NLP, ensuring biases in data do not lead to unfair or harmful outcomes. The future looks promising, with ongoing advancements in deep learning and neural networks likely to improve language understanding and generation. Researchers are also focusing on making NLP more accessible and useful across different languages and dialects. Balancing these advancements with ethical considerations will be crucial for the future of NLP.

Conclusion:

In conclusion, as shown NLP has many key points that contribute to it. Techniques such as Tokenization and Named Entity Recognition help NLP compute and understand the languages it is given, by shortening the amount of information or connecting name to face. Applications of NLP are things it itself actually does. For example, Machine Translation and Chatbots/Virtual Assistants both use NLP to communicate with humans. Machine Translation specifically in

translating speech and Chatbots/Virtual Assistants speaking with the person through voice or text. Using both the techniques and the applications NLP has and will become very popular.

References:

Hu, J. (2022). *Leveraging Word and Phrase Alignments for Multilingual Learning*. (Thesis). C arnegie Mellon University. Retrieved from http://hdl.handle.net/10.1184/r1/21707984.v1

Natural language processing (NLP) - what is it and how is it used? Hyperscience. (2024, February 19). https://www.hyperscience.com/knowledge-base/natural-language-processing/#:~:text=NLP%20algorithms%20use%20statistical%20models%20to%20identify%20patterns,improve%20the%20quality%20of%20machine%20translation%20even%20further.

Otten, N. V. (2023, October 31). *What is neural machine translation? & 4 easy python tools*. Spot Intelligence. https://spotintelligence.com/2023/01/04/neural-machine-translation/#What_is_neural_machine_translation

History and Development of NLP A Brief History of Natural Language Processing - DATAVERSITY

Advancing AI with integrity: Ethical challenges and ... (n.d.). https://arxiv.org/pdf/2404.01070

Named Entity Recognition https://www.geeksforgeeks.org/named-entity-recognition/

What is Sentiment Analysis https://www.geeksforgeeks.org/what-is-sentiment-analysis/

What is POS (Parts-Of-Speech) Tagging - https://www.geeksforgeeks.org/nlp-part-of-speech-default-tagging/

Lemmatization: On the Role of Morphological Information for Contextual Lemmatization.: EBSCOhost

Stemming: Introduction to Stemming - GeeksforGeeks