Natural Language Processing

By TOMMY DOBSON III, MARVENS LONGCHAMP, NICHOLAS NEGLIA, ETHAN-JAMES STIDHUM, NATHANIEL THOMAS, BREYNER ORDONEZ.

Abstract

Natural Language Processing (NLP) is a rapidly evolving field of Artificial Intelligence (AI), focused on enabling machines to interpret, understand, and respond to human language. NLP is the foundation behind virtual assistants like Apple's Siri and OpenAI's ChatGPT, which emulate human conversation by analyzing and generating natural language. This research paper provides an overview of NLP's evolution, tracing its origins from early models like ELIZA and SHRDLU to advanced neural network models that have revolutionized the field. Key NLP concepts, such as tokenization, part-of-speech tagging, named entity recognition, sentiment analysis, and semantics, are discussed as fundamental building blocks that enable machines to parse and interpret complex linguistic patterns. Additionally, it examines NLP applications, including machine translation, chatbots, and text summarization, which have permeated industries and daily life. However, challenges such as language ambiguity, high computational power demands, and lengthy training times present hurdles for further advancement. Future directions of NLP research aim to deepen contextual understanding, improve response accuracy, and optimize resource usage. This paper concludes by highlighting NLP's impact on human-machine interactions and potential for broader societal applications.

Introduction to Natural Language Processing

Natural Language Processing (NLP) is a continuously evolving branch of Artificial Intelligence that focuses on enabling machines to read, understand, and emulate human language. Chatbots and virtual assistants like Open AIs' ChatGPT and Apple's Siri are some examples of NLP, as they use complex NLP processes to aid in generating human-like responses. The goal of natural language processing is to enable machines to look beyond individual words or phrases and understand the context that it is being delivered with. For instance, if someone told you “Lets hit the books” you would automatically understand what they meant by hitting the books, on the other hand, machines without NLPs would process that as physically hitting a book.

Teaching a machine to adopt human language is exceedingly difficult to achieve. Not only is the human language complex, but you are trying to instill AI over 6500 languages each with its own languifical rules. Additionally, teaching AI to pick up social cues such as sarcasm or irony, is significantly more difficult than teaching AI the language itself.

History and Devlopment of NLP

In 1966, Joseph Weizenbaum created ELIZA at MIT, which was one of the earliest chatbots. ELIZA mimicked a therapist by using pattern matching and pre-written replies, which amazed users with its almost human-like conversations. Some important milestones in the early development of natural language processing (NLP) include:

- 1970: Terry Winograd developed SHRDLU, a system that could manipulate virtual blocks using natural language commands, highlighting an early grasp of context and instructions.

- 1980s-1990s: Statistical models gained popularity in NLP, leading to improved data analysis and language processing methods.

- 2001: Yoshio Bengio and his team introduced the first neural language model, which used a feed-forward neural network that processes information in one direction, marking a major move towards deep learning in NLP.

- 2011: The introduction of Apple's Siri was a major advancement in AI assistants, using automated speech recognition to understand user commands and respond correctly. For example, if someone asked about their account balance, Siri could interpret "Yes" or "No" and act accordingly.

Key Concepts in NLP

- 4.1- Tokenization- This concept of NLP is the breaking down of a larger text into an organized piece by breaking down each word, this also allows us to exclude punctuation and make segmentation easier.

- 4.2- P.o.S Tagging- Part of speech tagging is an NLP technique that tags each individual word in a body of text as a noun, adjective, verb, or pronoun. This is a fundamental task

that in NLPs used in a variety of applications, such as machine translation, text summarization, and search engine optimization. However, tagging can be difficult due to some words having distinct functions depending on the context they are used in. For example, the word “novel” can mean a fictional book, or it can mean an original idea, making classification often difficult for NLPs.

- 4.3- Named Entity Recognition- Named entity Recognition, this process identifies unique names for people, places, events, companies, and more. NLP software uses named-entity recognition to determine the relationship between different entities in a sentence.

- 4.4- Sentiment Analysis- Sentiment Analysis helps to understand the emotion meaning behind a of text, whether it be positive or negative, it can also determine the type of emotion or feeling, like happy, sad, angry. For example, sentiment analysis can be performed on Twitter to determine overall opinion on a particular trending topic. Companies and brands often use sentiment analysis to monitor brand reputation across social media platforms or the web.

- 4.5- Semantics- Semantics refers to the direct meaning of the text, understanding individual word meanings, including synonyms, and antonyms. For example, two sentences can both use the same word, but they can have different meanings.
- A bat is a nocturnal creature.
- Baseball players use a bat to hit the ball.

- 4.6- Lemmatization- Lemmatization groups different forms of a word by reducing them to their root form. It changes tenses and unifies synonyms, considering context to distinguish identical words. Some examples of this can be “thought” being changed to “think,” “building” to “build.”

- 4.7- Stemming- Stemming removes affixes from words by slicing prefixes or suffixes. This often at times can be mistaken for lemmatization, but stemming primarily slices the beginning or end of words to remove affixes.

- 4.8- Stop Word Removal - Stop Word Removal is used to remove common words such as “a,” “the,” “to,” “an,” and “is.” Such words do not add significant meaning to the text and are often removed from the text to improve the efficiency and accuracy of various language processing tasks. By eliminating these words, the focus is shifted to more significant terms, allowing algorithms to better understand the context and meaning of a given text.
- **Examples**
- “I am going to the store to buy some groceries."
- After stop word removal is applied**: "Going store buy groceries"**
- "The quick brown fox jumps over the lazy dog."
- After stop word removal is applied**: "Quick brown fox jumps lazy dog."**
- Common stop words like "the," "am," "to," and "some" are removed to emphasize the important parts of the text

- 4.9- Bag of Words- Refers to using a model that creates a matrix of all the words given in text based on an unordered collection of words. It is used in NLPs and information retrieval, disregarding word order and showing multiplicity through its simple and effective representation of text data.
- **Example**: Consider the following three short sentences:

  "The dog ran."

  "The dog ran in the house."

  "The dog with the house."

First, identify all unique words in the sentences: "the," "dog," "ran," "in," "house," "with." Next, count how many times each word appears in each sentence:

Sentence	the	dog	ran	in	house	with
1	**1**	**1**	**1**	**0**	**0**	**0**
2	**2**	**1**	**1**	**1**	**1**	**0**
3	**2**	**1**	**0**	**0**	**1**	**1**

Finally, represent each sentence as a vector:

[1, 1, 1, 0, 0, 0]

[2, 1, 1, 1, 1, 0]

[2, 1, 0, 0, 1, 1]

Applications of NLP

- 5.1- Machine Translation- Machine translation relies on AI and NLP to automatically convert text or speech from one language to another. An example is Google translate.

- 5.2 - Chatbots and Virtual Assistants- AI programs are created to simulate human conversation and help with different tasks using natural language, either through text or voice. These systems use NLP to understand what users input and give appropriate replies. An example is Alexa and ChatGPT.

- 5.3- Text Summarization- This NLP technique involves automatically creating a shorter version of a longer text, capturing the most important information while maintaining coherence and readability. An example of this is "Rewriter AI."

Extras

Market Intelligence

- NLP helps businesses understand the market by examining unstructured data to spot trends and feelings of consumers. Companies use this insight to make smart choices in marketing and developing their products.

Urgency detection

- Urgency detection identifies expressions indicating immediate concern or need within text communications. This is useful in customer service to prioritize responses effectively based on the urgency of requests.

Challenges of NLP

Despite its rapid advancements over the years, NLP still faces several challenges. Some reoccurring issues include language ambiguity, learning time, and computer power.

- Language Ambiguity: English is the most spoken language around the world with over 1.5 billion speakers, yet in the English language different meanings can be shared among the same word. Due to these confusions in language NLP can get confused.

- Learning Time: NLPs take time to remember and show reliable information on their own.

- Computing Power- Chatbot NLPs (one of the most used NLPs of today) require immense amounts of computer power. For example, while asking for a chatbot's assistance one liter of water is used.

Future Directs of NLP- Future directions may involve improving models to better understand context, relay information better, and cost less resources.

- Understand Context: In the future NLPs may be able to understand the meaning of various phrases and allow more casual talk leading to less robotic conversations.

- Relaying Information: When NLPs develop in understanding so many more possibilities can be opened. One example being the use of NLPs in advertising businesses.

- Resource Saving: NLP models can be programmed to amplify resource saving by focusing on data-efficient techniques, allowing for high performance with little data requirements.

Conclusion

In conclusion, Natural Language Processing represents the field of artificial intelligence, enabling machines to engage with human language in increasingly sophisticated ways. From its early beginnings with programs like Eliza and SHRDLU to creative applications such as chatbots and machine translation, NLP has undergone remarkable evolution. Key concepts like tokenization, part-of-speech tagging, and named entity recognition have laid the groundwork for more advanced functionalities, including sentiment analysis and text summarization.

References

Challenges of NLPs- https://i2group.com/articles/the-10-biggest-issues-facing-natural-language-processing

ChatGPT: How does this NLP algorithm work?

Natural language processing - Wikipedia

An In-depth Handbook: What is Sentiment Analysis in NLP?

What is NLP? - Natural Language Processing Explained - AWS

History of natural language processing - Wikipedia

Visual Aids:

Presentation9.pptx