Proceedings of the 5th International Conference on Computing and Data Science
DOI: 10.54254/2755-2721/16/20230896

Research on the applications of natural language processing

Bo Wen

Tandon School of Engineering, New York University, New York, NY 10003, USA

bw2497@nyu.edu

Abstract. In recent years, there has been a rapid advancement in natural language processing (NLP), leading to notable improvements in areas like sentiment analysis, machine translation, and text recognition. However, belonging to the same field under AI, the research materials and academic topics of NLP are not so adequate. Therefore, this paper introduces and shows the reader a general introduction to NLP for the current research environment, so that the reader can clearly understand the history of NLP, its uses, and the current and future research directions. It also analyzes the current mainstream applications of NLP, including the most basic Rule Based method and the more popular deep learning, development of pre-trained language models such as GPT. Besides, this article also analyzes the current mainstream application scenarios of NLP and its related use of technology, so that readers can have a clearer picture of the research direction of NLP. At the end, this paper also summarizes the general development of NLP, research methods, and application scenarios to provides an outlook on the future development of NLP with respect to these points.

Keywords: natural language processing, deep learning, ChatGPT.

1. Introduction

The field of Natural Language Processing (NLP) is rapidly expanding, and it mainly emphasizes the development of computational models for language comprehension and processing. NLP draws on insights and methods from linguistics, psychology, and computer science to enable computers to interact with humans using natural language. The goal of NLP is to allow for more natural and intuitive interactions between humans and machines, with applications including sentiment analysis, speech recognition, and question-answering systems. The progress in deep learning techniques and neural networks has paved the way for significant advancements in NLP, enabling computers to learn from massive amounts of language data and perform tasks that were once considered the exclusive domain of humans.

From historical review, NLP has been around since the early days of computing, with early attempts at machine translation using rule-based methods that were limited in their effectiveness. In the 1950s and 60s, statistical methods were explored, but they still had trouble with complex language phenomena. In the 1970s and 80s, expert systems were used, but they were limited by domain-specific knowledge. In the 1990s, statistical methods were improved with machine learning algorithms, which paved the way for modern deep learning techniques. Recently, the development of neural networks has revolutionized NLP, allowing computers to learn from large amounts of language data and perform tasks with

DOI: 10.54254/2755-2721/16/20230896

unprecedented accuracy. Despite challenges, NLP is a rapidly growing field with significant potential for impact in various industries [1].

NLP has various applications in different industries and domains, including healthcare, finance, customer service, and entertainment. NLP is used in sentiment analysis to determine the emotion behind a piece of text, in language translation for automatic translation of text, and in speech recognition systems for transcribing spoken language into text. NLP is also used in named entity recognition to classify and identify named entities in text, and in text summarization to summarize long pieces of text for quick understanding of the key points. These applications can be useful for information extraction, text classification, news aggregation, and document summarization.

This essay provides an overview of the recent development of NLP in the world, focusing on the current mainstream NLP methods, algorithms, and application research. It summarizes previous experiences and attempts to predict the future direction of NLP research.

2. Methods of natural language processing

To achieve the effect of identify languages and apply it, NLP techniques have evolved over time. Some of the most commonly used techniques in NLP include: Rule-Based Methods, Statistical Methods, and Neural Network Approach which also called machine learning.

2.1. Rule-based methods

Rule-based methods utilize explicit rules and patterns to process text, and are frequently applied to perform relatively straightforward tasks such as named entity recognition and part-of-speech tagging. These methods rely on a set of predetermined rules and linguistic patterns to interpret and analyze natural language data. They typically require the development of hand-crafted rules and grammars to extract meaning from text.

The rule-based approach was the earliest and most basic method used in NLP. It was first used in the 1950s and 1960s and was the dominant method until the 1990s. The algorithm used in this approach involves creating a set of rules and patterns that can be used to identify and extract specific information from text. These rules are often created by linguists and subject matter experts and are based on their knowledge of the language and the domain being analyzed. Centering Theory is one of the typical examples of rule-based method, which integrated into a formal model of inference. It first sets the Identify the different entities or centers present in a sentence or discourse segment and determine the salience score of each center. Salience is a measure of how much attention a center is currently receiving in the discourse. By identifying the transitions between different centers, a transition occurs when the speaker or writer shifts their attention from one center to another. By calculation and information, it can create a formal model of inference that can make predictions about the relationships between different parts of a text [2].

The rule-based approach works well for simple and well-defined tasks, such as information extraction, where the input language and expected output are clearly defined. However, it has difficulty handling the complexity and ambiguity of natural language, making it less suitable for tasks such as sentiment analysis or language translation. The rule-based approach provides a high level of accuracy for specific tasks when domain experts can identify a finite set of rules to capture the relevant language patterns. It is also transparent and interpretable since the rules are explicitly defined. However, it is also a time-consuming and expensive to develop since it requires human experts to define the rules and update them constantly. It also has limited flexibility since it cannot handle variations in language use that were not foreseen in the rule set.

2.2. Statistical methods

Statistical methods involve the use of probabilistic models to analyze and generate text. These methods have been widely used for tasks such as machine translation, language modeling, and sentiment analysis. It uses mathematical models to analyze and understand natural language data. It involves the use of machine learning algorithms to train models that can automatically identify patterns and relationships

DOI: 10.54254/2755-2721/16/20230896

in language data. The statistical approach emerged in the 1990s as a more advanced alternative to the rule-based approach. It has become the dominant approach in NLP in recent years.

The algorithm used in this approach involves training statistical models using large amounts of annotated language data. These models use probabilistic methods to predict the most likely interpretation or translation of a given text. The most typical example must be IBM alignment models. The basic idea behind IBM alignment models is to learn a set of parameters that can be used to probabilistically align words in a source language with their translations in a target language. These parameters are learned using a training corpus that contains aligned pairs of sentences in both languages. There are several different versions of IBM alignment models, each of which makes different assumptions about the alignment process (Figure 1). But both are much stronger than those rule-based methods model in NLP [3].

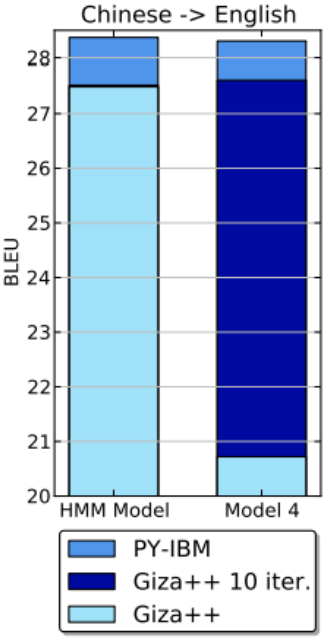

Figure 1. BLEU scores of Giza++'s and PY-IBM's HMM model and model 4 translating from Chinese into English [3].

The statistical approach can handle a wide range of tasks and language variations without requiring a fixed set of rules. It is also more flexible and less expensive to develop than the rule-based approach, as it requires less manual intervention in creating models. However, it also requires enormous labeled data to train, which could be difficult and costly to collect, especially in domains with specialized language use. It can also be difficult to interpret and explain the results of statistical models, as they do not provide any explicit rules for decision making.

2.3. Neural network approach

The neural network approach involves the use of artificial neural networks to analyze and understand natural language data which can also be called Artificial Neural Network. It involves the use of deep learning algorithms to automatically extract features from language data and make predictions based on those features. The neural network approach emerged in the early 2010s and has become cumulatively popular in recent years due to the availability of large amounts of labelled data and improvements in computing power. The algorithm used in this approach involves training deep neural networks using large amounts of annotated language data [4]. These networks can automatically learn to extract useful features from text data and use them to make accurate predictions about the meaning or translation of a given text. For example, Neural Machine Translation (NMT) is a popular machine translation approach

DOI: 10.54254/2755-2721/16/20230896

that uses a neural network consisting of an encoder and decoder to learn the mapping between a source language and a target language. It is typically implemented using recurrent neural networks or their variants, allowing the model to capture complex dependencies between the input and output sequences. The NMT model is instructed to reduce the dissimilarity between its projected translation and the accurate translation by means of a loss function while undergoing training.

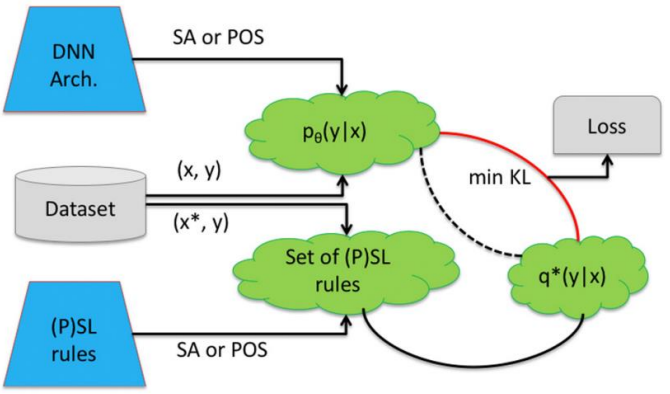

Figure 2. Model show soft logic and neural network for NLP [5].

The neural network approach is known for its ability to handle complex language use cases and achieve higher accuracy than statistical and rule-based methods in many NLP tasks (Figure 2). This is because neural networks allow for end-to-end learning of a higher-level task, as opposed to relying on a pipeline of separate intermediate tasks. This approach has proven to be highly effective, as the neural network can learn to represent the relationships between words and phrases in a text and make predictions based on the learned patterns. It can also learn from unstructured data, making it well-suited for tasks where there is no explicit labeling of the data. However, this approach requires large amounts of labeled data and computing resources to train, making it expensive and time-consuming to develop. It is also less transparent and interpretable than the rule-based approach since it relies on a large number of hidden parameters and layers.

2.3.1. Deep learning. Deep learning which includes using neural networks to learn representations of text data, as a standout in the field of artificial intelligence, of course, also plays a great role in the neural network approach, further improving the efficiency of NLP processing. Its techniques have been particularly successful in NLP, achieving state-of-the-art results on many tasks such as language modeling, machine translation, and sentiment analysis.

2.3.2. Transformer architecture. Transformer is a neural network architecture based on the idea of an attention mechanism. The attention mechanism allows the model to focus on some parts of the input sequence and ignore others, allowing the model to learn to make more informed decisions. In traditional neural network architectures, information flows through a fixed set of layers, with each layer processing the entire input sequence. However, in the Transformer architecture, the input sequence is processed in parallel, allowing the model to exploit dependencies between different parts of the input sequence. The self-attentive mechanism of the Transformer architecture enables the model to capture the relationships between various components of the input sequence. By allowing the model to attend to different parts of the input sequence at different times based on the input's context, the self-attention mechanism enhances the model's ability to process the input sequence effectively.

ChatGPT is an example of an application of NLP and transformer architecture. It is a large language model (LLM) trained on a massive amount of text data, using sophisticated machine learning algorithms and techniques to understand the meaning and structure behind human’s language. ChatGPT employs a type of neural network known as a transformer network, which is highly effective in processing

DOI: 10.54254/2755-2721/16/20230896

sequential data such as text. The transformer architecture allows ChatGPT to model complex relationships between words and phrases in a given text, enabling it to generate coherent and contextually appropriate responses to user inputs.

In terms of its relationship to NLP, ChatGPT represents a cutting-edge example of how NLP can be applied to enable more natural and effective communication between humans and machines. By leveraging the power of NLP, ChatGPT can understand and generate text-based conversations in a way that closely mimics human-to-human communication, allowing it to provide a wide range of information and services to users in a seamless and intuitive way [6]. Table 1 shows the comparison of three types of Natural Language Processing.

Table 1. Comparison of three methods of natural language processing.

Types	Representative Models	Advantages	Disadvantages	Performance	History
Rule-Based Methods	Centering Theory	Transparent, interpretable	Time-consuming, expensive	Perform well when the amount of training data	1950-1990
Statistical Methods	IBM alignment models	Flexible, less expensive	Need large data, hard to interpret	Perform well when statistical interpretability and transparency is required	1990-2010
Neural Network Approach	Large Language Model (GPT-4)	highly effective, most accurate	Need large number of data, time-consuming	The most efficient and mainstream method	2010-Now

3. Application

Natural Language Processing finds extensive utilization in diverse industries and domains, such as finance, customer service, entertainment. NLP helps to make people better communicate with the machine through its own ways. Thus, NLP is used in different scenes to help the human, which can also be called NLP techniques [7].

3.1. Sentiment analysis

In the realm of natural language processing, sentiment analysis is a technique that utilizes computational algorithms to detect the prevailing sentiment or emotion expressed within a text, such as a product review or social media post. The primary goal of Sentiment Analysis is to classify a piece of text as emotional response such as positive, negative. NLP has played a crucial role in the development and success of Sentiment Analysis. With the aid of machine learning algorithms, NLP models are capable of scrutinizing vast amounts of text and sorting them into either positive, negative, or neutral categories, based on the linguistic attributes displayed in the text. LP techniques used in Sentiment Analysis include natural language understanding, feature extraction, and machine learning algorithms. Natural language understanding allows NLP models to interpret and analyze human language, while feature extraction involves identifying relevant features from the text that can be used to classify sentiment. Machine learning algorithms are used to train NLP models on large volumes of annotated data, enabling them to learn how to classify text accurately.

The benefits of Sentiment Analysis are vast, particularly for businesses looking to understand customer feedback and monitor brand reputation. By using NLP-based Sentiment Analysis, businesses can quickly analyze large volumes of customer feedback, identify areas of improvement, and make informed decisions to improve customer satisfaction.

DOI: 10.54254/2755-2721/16/20230896

3.2. Language translation

Language Translation is the process of translating the given text from one language to another language. This is an important task in today's increasingly globalized world where communication across different languages is becoming more and more important. Natural Language Processing (NLP) has played a significant role in making language translation more efficient and accurate. NLP techniques used in Language Translation involve the use of statistical and neural machine translation models. Statistical models use statistical algorithms to determine the most likely translation of a given text based on large corpora of bilingual texts. Neural machine translation models, on the other hand, use deep learning neural networks to learn how to translate text by being trained on large datasets of translated texts.

The use of NLP in Language Translation has led to significant improvements in translation accuracy, speed, and cost. This technology has made it possible for individuals and businesses to communicate with people from different cultures and regions without the need for human translators. In recent years, NLP-based Language Translation has also led to the development of advanced translation tools such as translation apps, real-time translation systems, and automatic subtitle generators for videos. These tools have made it possible for people to communicate effectively across different languages in real-time.

3.3. Speech recognition

Speech recognition refers to the process of using technique to convert the spoken sentences or words into text. This technology can be applied in various domains such as voice assistants, automated transcription, and language learning tools, among others. NLP techniques such as acoustic modeling, language modeling, and deep learning algorithms have been used to develop sophisticated speech recognition systems. Acoustic modeling involves the analysis of audio signals to identify and extract features that can be used to recognize speech. Language modeling involves the use of statistical models to predict the probability of a particular sequence of words based on previous words in the sentence. Deep learning algorithms, such as Convolutional Neural Networks (CNNs) and Recurrent Neural Networks (RNNs), have been used to improve the accuracy and speed of speech recognition systems.

3.4. Named entity recognition

Named Entity Recognition (NER) is a specialized area that focuses on detecting and categorizing named entities like people, organizations, locations, and dates in textual data. It is widely used in different applications including extracting information, classifying the text, and question answering systems. NLP's usage in NER entails utilizing machine learning algorithms, such as Hidden Markov Models (HMMs), Conditional Random Fields (CRFs), and deep learning models like Recurrent Neural Networks (RNNs). These algorithms are fed annotated datasets and trained to recognize the characteristics and patterns of named entities in textual data.

To identify named entities, the NER system initially tokenizes the input text into separate words or phrases. Subsequently, the system employs different linguistic attributes like part-of-speech tags, dependency parsing, and word embeddings to detect the named entities within the text. Finally, each named entity is classified into predetermined categories such as person, organization, location, or date.

3.5. Text summarization

Text Summarization is a crucial process in natural language processing that involves condensing a longer document while preserving its essential information. NLP techniques have significantly improved the accuracy and efficacy of text summarization systems. The two primary methods of text summarization are extractive and abstractive. In extractive summarization, significant sentences or phrases from the source text are identified and arranged in a structured summary. Conversely, in abstractive summarization, new sentences are generated that encapsulate the meaning of the original document.

NLP techniques applied in text summarization exploit machine learning algorithms, particularly deep learning models, to recognize the most significant sentences or phrases in a document. To achieve this,

DOI: 10.54254/2755-2721/16/20230896

these models are trained on massive datasets of text documents and utilize features like word frequency, sentence structure, and semantic similarity to identify the most relevant information.

4. Application scenarios

With the knowledge of the technology and applications implemented by NLP, the market quickly responded to these technologies, giving birth to a number of common NLP usage scenarios in the market. Sometimes, different techniques of NLP will also be used together to achieve the assigned application scenario. Here are explicit common application scenarios.

4.1. Explainable and interpretable NLP

As NLP models become more complex, it is increasingly important to ensure that they can be understood and interpreted by humans. Explainable and interpretable NLP models can help build trust in the technology and enable users to better understand how the models make decisions. By using NLP models that can explain how they arrive at a particular diagnosis, doctors and patients can have greater confidence in the accuracy of the diagnosis and a better understanding of the reasoning behind it. This can also help to identify potential biases or errors in the model and improve its overall performance [8].

4.2. Context-aware NLP

The accuracy of several NLP tasks, including entity recognition and sentiment analysis, is reliant on the availability of contextual information. Context-aware NLP models can take into account the broader context of a piece of text, such as the speaker's tone or the social and cultural context, to produce more accurate and nuanced results. One example of context-aware NLP is in the field of recommender systems. By using NLP models that can understand the user's context, such as their search history or location, these systems can provide more personalized and relevant recommendations for products or services [9].

4.3. Multilingual and cross-lingual NLP

NLP has primarily focused on English-language text, but there is a growing need for NLP models that can handle multiple languages. Multilingual and cross-lingual NLP models can help bridge language barriers and enable communication and information exchange across diverse populations. One example of multilingual NLP is in the field of machine translation, where NLP models are trained to translate between different language pairs. Multilingual NLP can also be applied to sentiment analysis, named entity recognition, and other NLP tasks that involve multiple languages. Cross-lingual NLP is important for applications such as information retrieval, where NLP models are used to search for information across multiple languages, or cross-lingual document classification, where NLP models are used to classify documents written in different languages [10].

5. Conclusion

In conclusion, NLP has come a long way since its inception, and it continues to be an exciting and rapidly evolving field of research. In this review paper, we have explored some of the main methods and applications of NLP, including machine translation, sentiment analysis, and speech recognition. We have also discussed some of the application scenarios in NLP research today, such as Explainable and Interpretable NLP, Context-Aware NLP, and Multilingual and Cross-Lingual NLP. Looking ahead, there are several promising directions for future development in NLP. One area of focus is on developing more sophisticated models that can better capture the nuances of human language and improve performance on tasks such as question answering and dialogue systems. There is also a focus on merging NLP with other areas, including computer vision and robotics, to develop advanced systems that can communicate with humans in a more natural manner. Thus, NLP has already had a significant impact on many industries and domains, and its potential for further innovation is vast. With researchers continuously striving to explore the full potential of NLP, we can anticipate witnessing even more thrilling advancements in the field in the upcoming years.

DOI: 10.54254/2755-2721/16/20230896

References

[1] D. Ganguly, et al., "Legal IR and NLP: The History, Challenges, and State-of-the-Art," in Lecture Notes in Computer Science, 2023, pp. 1-10.

[2] A. K. Joshi and S. Weinstein, "Control of Inference: Role of Some Aspects of Discourse Structure-centering," in Proceedings of the International Joint Conference on Artificial Intelligence, 1981, pp. 385-387.

[3] Y. Gal and P. Blunsom, "A Systematic Bayesian Treatment of the IBM Alignment Models," in Proceedings of the North American Chapter of the Association for Computational Linguistics, 2013, pp. 969-977.

[4] M. Gridach, "A Framework Based on (Probabilistic) Soft Logic and Neural Network for NLP," in Applied Soft Computing, vol. 93, 2020.

[5] X. Lyu, et al., "Refining History for Future-Aware Neural Machine Translation," in IEEE Transactions on Audio, Speech, and Language Processing, vol. 31, no. 2, 2023, pp. 500-512.

[6] M. Perkins, "Academic Integrity Considerations of AI Large Language Models in the Post-Pandemic Era: ChatGPT and Beyond," in Journal of University Teaching and Learning Practice, vol. 20, no. 2, 2023, pp. 1-24.

[7] R. Varaprasad and G. Mahalaxmi, "Applications and Techniques of Natural Language Processing: An Overview," in Journal of Computer Science, vol. 16, no. 3, 2022, pp. 7-21.

[8] Q. Wang, et al., "Explainable APT Attribution for Malware Using NLP Techniques," in Proceedings of the International Conference on Software Quality, Reliability and Security, 2021, pp. 70-80.

[9] "NLP-Based Context-Aware Log Mining for Troubleshooting," IBM Corporation, Armonk, NY, 2022, pp. 1-11.

[10] M. M. Agüero-Torales, et al., "Deep Learning and Multilingual Sentiment Analysis on Social Media Data: An Overview," in Applied Soft Computing, vol. 107, 2021, article no. 107373.