International Journal of Education and Humanities
ISSN: 2770-6702 | Vol. 13, No. 2, 2024

Challenges of Natural Language Processing from a Linguistic Perspective

Dingli Chen[1, a]

[1]School of Foreign Languages, Sichuan Normal University, Chengdu 610000, China
[a]775743977@qq.com

Abstract: As the foundation of artificial intelligence, natural language processing is a research field with great development prospects, closely related to linguistics. Linguistic research has realized the initial natural language processing, and the development of natural language processing has raised higher requirements for linguistic research. Although natural language processing technology has achieved some results, there is still a distance from the most natural human-machine interaction. Language is the expression form of human thinking, and the fundamental purpose of artificial intelligence is to simulate human thinking with computers. To achieve this goal, it is necessary to start with simulating human natural language. This paper mainly analyzes the current challenges of natural language processing from a linguistic perspective, aiming to further improve the level of natural language processing.

Keywords: Natural Language Processing; Linguistics; Artificial Intelligence; Machine Translation.

1. Introduction

With the development of the times, the demand for language services is increasing day by day, and convenient and efficient electronic devices have become the preferred means of language services. Such language services mainly rely on natural language processing (NLP) technology to achieve. At the same time, the development of NLP has brought about significant changes in the working mode of translators in modern times. Nowadays, translation work relies more on the cooperation between translators and machines, and machines play an increasingly important role in the translation process. Mr. Feng Zhiwei, a pioneer in Chinese NLP research, once said, “In the field of artificial intelligence, those who control language control the world” [1]. Despite the efforts of technicians and linguists, NLP technology has reached a certain level, but there are still many linguistic challenges that are difficult to overcome. Contemporary linguists need to adapt to the background of the era of artificial intelligence, apply linguistic knowledge to technology, study machine language issues, and contribute to the development of technology.

2. What is Natural Language Processing

To understand NLP, it’s essential to grasp the concept of natural language first. Natural language refers to the language that emerges with the development of human culture, used for communication between people, and evolves over time. Languages like Chinese, English, and Japanese are examples of natural languages. NLP involves inputting natural language into machines and then having the machines output information to humans, enabling people to interact with computers using natural language. NLP consists of two main components: Natural Language Understanding (NLU) and Natural Language Generation (NLG). NLU focuses on enabling computers to understand the natural language we use, while NLG involves computers generating data in natural language format. Technologies such as voice assistants like Siri, speech-to-text features in chat applications, machine translation, and text generation techniques are all based on NLP.

In the early days, machine control relied on high and low voltage levels, where high voltage represented "1" and low voltage represented “0.” This binary system of “0” and “1” was initially used as the machine language. With the advent of computers, this control mechanism was also applied to them. To facilitate human-computer interaction, a universal encoding system called ASCII was developed. This encoding system assigned different decimal numbers to common symbols, digits, and uppercase and lowercase letters. For example, in computers, “Y” typically represents “yes,” and “N” represents “no,” with ASCII defining the uppercase letter “Y” as number 89, which translates to binary as 1011001. When we press the “Y” key, the binary number 1011001 is input into the computer, and the computer executes the “yes” command. Over time, even Chinese characters could be input into computers using this method. From initially using voltage levels to interact with machines to now controlling computers using familiar natural language, it’s evident that computers are gradually adapting to our language. This progression represents the development process of natural language processing.

3. Technical Challenges in Natural Language Processing

Firstly, the current challenges in NLP can be broadly categorized into two aspects: technical and linguistic. From a technical perspective, NLP relies on machine’s statistical and learning abilities, making corpus data the most direct material for machine learning language. However, constructing a good corpus requires manual efforts and is extremely time-consuming, making many corpora inaccessible, which severely hampers machine learning progress in language acquisition [2]. Secondly, many conversations and texts in daily life are initiated based on shared knowledge between the initiator and the receiver. However, current NLP models, trained on large corpora, often lack knowledge of historical or

cultural contexts, making them unaware of many issues related to history, culture, or common sense. This lack of background knowledge makes it difficult for machines to capture key factors when processing conversations and texts.

4. Linguistic Challenges in Natural Language Processing

From a linguistic perspective, machines face several challenges in understanding natural language. These include:

- Diversity of languages: Different languages spoken in various countries and regions make it challenging for machines to adopt a standard approach to process them.
- Ambiguity of language: The same sentence or word in a conversation or text may have different meanings, leading to ambiguity.
- Robustness of language: Machines need to demonstrate consistent performance when processing different accents and conversation styles.
- Dependency on knowledge: Language comprehension often relies on practical knowledge, posing a common challenge for both technology and language.
- Contextual understanding: Machines need to understand language based on its environment and context.

Compared to technical challenges, linguistic challenges appear to be more critical. Deep analysis of linguistic difficulties and further optimization of language processing methods are essential for enhancing machine performance. This paper focuses on linguistic challenges, particularly from the perspective of linguistics, analyzing language processing in terms of phonetics, semantics, and pragmatics.

4.1. Challenges in Phonetics

As mentioned earlier, NLP involves both input and output processes. Therefore, the first challenge in processing speech input into machines is the diversity of languages. For instance, can machines distinguish between British and American English accents? How much difficulty does a Japanese-accented English speaker, who cannot pronounce "r" and "f," pose for speech recognition? Furthermore, people from different regions within the United States and the United Kingdom may have different accents. Considering that each person's spoken language varies in terms of usage habits, fluency, and standardization, and the various noise, signal interference, mispronunciations, and grammatical errors present in real-life conversation scenarios, these undoubtedly test the robustness of machines in handling speech. Additionally, when processing speech, machines need to consider intonation. The same sentence spoken with different intonations can convey completely different meanings. For example, the sentence "你忘了" ("You forgot") can be punctuated differently depending on the intonation. If the final character "了" is spoken with a rising tone, it becomes "你忘了？" ("Did you forget?"), expressing the speaker's question. If "了" is spoken with a falling tone and increased volume, it becomes "你忘了！" ("You forgot!"), expressing the speaker's dissatisfaction.

After speech input into machines, we also hope that machines can output speech. However, various speech robots on the market today are perceived as unnatural because they overlook issues of intonation and tone. Currently, the speech output of natural language processing lacks any emotional expression. It merely objectively recounts a particular event. These utterances lack expression in tone and mood, making the language expression rigid and monotonous [3]. This also points out the future research direction for phoneticians, and it is believed that phonetics research on intonation and tone will make significant contributions to the further development of artificial intelligence.

4.2. Semantic Challenges

Semantic analysis in NLP has always been a headache for researchers, mainly due to the ambiguity of language. Some philosophers argue that semantic issues are a major obstacle to the development of artificial intelligence. John Searle, an American philosopher, once argued that although artificial intelligence can achieve syntactic operations based on symbolic language, it cannot achieve understanding and therefore lacks semantic capabilities [4]. Early NLP models, whether rule-based or statistical, could not solve the problem of different meanings of the same word in different contexts. Such models could only rigidly process language rules and the frequency of word usage for each meaning. Later, neural network-based NLP models could select expressions closer to the correct one by referring to a large amount of parallel text to analogize similar contexts, but still could not completely solve the semantic problem. When parallel texts cannot be found or there are no similar corpora for reference, machines resort to probabilistic methods to choose word meanings [5]. Moreover, ambiguity is most common in the Chinese language, making semantic analysis in Chinese much more difficult than in other languages. For example, "我的抽屉没有锁" ("My drawer is not locked") can mean "My drawer is not locked (at all)" or "My drawer is locked, but not locked securely"; "那位先生走了" ("That gentleman left") can mean "That gentleman left the place" or "That gentleman passed away"; "那家店关门了" ("That shop is closed") can mean "That shop closed for the day" or "That shop went out of business." English also has such ambiguities, such as "Are you engaged?" which can mean "Are you busy?" or "Are you engaged to be married?" These lexical ambiguities often require the consideration of context to choose the most accurate meaning, which machines cannot autonomously consider when processing language. In addition to ambiguity, some expressions also pose problems for semantic analysis. For example, "冬天能穿多少穿多少，夏天能穿多少穿多少" ("Wear as much as you want in winter, and wear as much as you want in summer"), "骑车的时候突然打滑了，还好我一把把把把住了" ("When riding a bike, I suddenly slipped, but fortunately I grabbed it"), "小龙女说：我也想过过过过过的生活" ("Xiaolongnü said: I also want to experience a life that has passed by"). As native speakers, we often need to read these sentences several times to discover their meanings. It can be imagined how difficult it is for machines to process such sentences, let alone translate them into another language.

4.3. Pragmatic Challenges

Pragmatics provides three basic principles to standardize conversations. Under such principles, what people say in conversations reflects what they want to express. If people's conversations were all standard like this, then machines would handle them very easily and accurately. However, in reality, conversations in people's daily lives often violate the three major principles, which creates trouble for machines in processing conversations. People violate pragmatic principles not unintentionally; the study of pragmatics explores the

motives behind people's violation of these principles. From the perspective of pragmatics, when people violate pragmatic principles, they often intend to express implicit meanings, known as implicatures. For example, when a mother asks her child, "Have you finished your homework?" and the child responds, "Tomorrow is Saturday," or when a leader asks an employee, "Can you come to the company tomorrow?" and the employee responds, "Tomorrow is Saturday," these answers that deviate from the questions clearly violate the relevance principle, which states that the response of the hearer should be related to the speaker's utterance. In the first dialogue, the child indirectly expresses that they won't go to school tomorrow, the homework is not completed yet, and it can be finished tomorrow. The employee indirectly expresses that they won't work tomorrow and don't want to come to the company for overtime. This also reflects the problem of linguistic knowledge dependency. As mentioned earlier, machines do not have common sense; they do not know what weekends are, so they cannot understand the actual meaning of "Tomorrow is Saturday." Additionally, in everyday conversations, people often use humor and irony, which are implemented based on violations of pragmatic principles. For example, when a passenger asks a ticket seller, "What should I do if I miss my train?" and the ticket seller responds, "Try to see if you can run over and catch it," the ticket seller clearly says something untrue, violating the quality principle. In fact, the ticket seller means there is no way to catch it anymore. For example, in an argument between A and B, when A says, "I could not bear such a fool," and B responds, "Your mother could," B deviates from the original context and interprets "bear" as "give birth to," responding with a sarcastic remark, "Your mother can give birth to fools," which is highly ironic. Examples like these require machines not only to distinguish between the different meanings of "bear" but also to analyze conversations by stepping out of the original context.

In conclusion, NLP has made significant advancements, yet machines still have a way to go to truly comprehend human natural language. However, considering the development trajectory of technology, from entirely manual to fully intelligent, and the booming demand in the language service market, it's evident that we need more advanced and efficient technologies to meet market demands. Language research is a crucial foundation for the development of NLP technology, which necessitates contemporary linguists to adapt to the demands of the times by integrating linguistic research with machinery. It is hoped that linguists can conduct deeper analyses and research on the three linguistic aspects mentioned above, proposing effective solutions to elevate NLP technology to new heights.

References

[1] Feng Zhiwei. History and Current State of Natural Language Processing. Chinese Foreign Language, 2008, (01), 15.

[2] Information on: blog.csdn.net/heyc861221/article/details/80130981

[3] Zhang Le, Tang Liang. Opportunities and Challenges for Linguists in the Age of Artificial Intelligence. Computer Knowledge and Technology, 2020, 16(24), 197.

[4] Wu Ge. Research on Semantic Problems from the Perspective of Artificial Intelligence (Academic Degree, Jilin University, China, 2021). 26.

[5] Feng Zhiwei. History and Current Situation of Natural Language Processing. Chinese Foreign Language, 2008, (01), 37.