Proceedings of the 5th International Conference on Computing and Data Science
DOI: 10.54254/2755-2721/19/20231026

# Analysis on connection and translation of natural language to traditional computer language

**Hanxiang Liu**

School of Economics and Management, Xidian University, Xian, Shaanxi, China, 710126

hxliu20031204_lmx@163.com

**Abstract.** From Siri to ChatGPT, natural language processing has been applied to human-computer conversations with remarkable success. Natural language processing (NLP) is an intimate connection between humans and computers, allowing machine learning to directly translate human language into computer language. They appear to be able to communicate directly with humans. Using a literature review methodology, this paper examines the connection and translation of natural language to traditional computer language. In addition to describing the fundamental principles, primary tasks, and implementation steps of NLP, it analyzes the problems that may arise during the development process in the future. In addition to comparing, citing, and classifying human language and computer language, the paper employs appropriate computer language processing tools. Through a side-by-side comparison of natural language and traditional computer language, the fundamental principles of natural language processing are explained and analyzed concisely, and some helpful suggestions for the detailed processing of natural language are provided.

**Keywords:** natural language processing, natural language understanding, natural language generation, tokenization.

## 1. Introduction

In recent years, as a result of the advancement of science and technology, particularly in the field of computing, there are numerous inanimate companions in human life. In the process of conversing with them, it is easy to forget that they are merely programs or software, such as Siri, the first intelligent voice assistant that lacks the ability to learn, cannot constantly acquire new information from conversations with humans, and cannot engage in multiple rounds of conversation [1]. In addition, there is ChatGPT, a personalized chatbot whose introduction sparked a worldwide obsession. The algorithm behind ChatGPT is based on the Transformer architecture, a self-focusing neural network that processes incoming data. chatGPT represents a significant advancement in natural language generation technology and offers new concepts and tools for artificial intelligence research [2]. Natural language processing is essential to the development of both the emergent creative and collaborative conversation system ChatGPT and the early practical and classic Siri.

As a science that integrates computer science, mathematics, and linguistics, natural language processing (NLP) has become a hot topic in response to the current environment and people's needs [3]. In recent years, natural language processing has primarily been applied to the field of human-computer

DOI: 10.54254/2755-2721/19/20231026

dialogue, where it has produced increasingly extraordinary results. Technically, the current human-computer dialogue system has evolved from "dialogue intelligences" to "body-involved dialogue intelligences." However, current technology does not guarantee that the human-computer dialogue system can achieve its objective of achieving a natural conversation between humans and machines [4].

Natural language processing is extensively utilized in numerous fields due to its versatility and practicality. Consequently, the Internet is rife with articles discussing how natural language processing can replace or enhance the original models in other fields. However, few papers have explained how and why NLP works. This distinguishes this paper from other investigations.

Using a literature review methodology, this paper examines the relationship and transformation between natural language and traditional computer language. It also explains the fundamental principles, primary tasks, and implementation steps of NLP, as well as analyzes the issues that may arise during the development process in the future. In addition to comparing, citing, and classifying human language and computer language, the paper employs appropriate computer language processing tools. At the macro level, a horizontal comparison of natural language and traditional computer language is used to provide a more concise explanation and analysis of the fundamental principles of NLP, as well as some helpful suggestions for the detailed processing of NLP. Simultaneously, it urges more young people to engage with data science and gain a comprehensive understanding of the emerging hot field of NLP in recent years.

## 2. Principles and core tasks

### *2.1. Basic principles of computer recognition and generation of natural language*

Each creature and machine has its own language. Humans communicate using language. Dogs communicate through barking. Digital messages are also utilized by machines as a means of communication. Humans are incapable of comprehending the precise meaning of an animal's bark. Even individuals who speak different languages cannot communicate directly; translation is required. Since distinct human languages can be translated, is "translation" also possible between humans and machines? The advent of NLP has resolved this issue. To accomplish the goal of human-machine communication, it acts as a bridge between machine language and human language.

To communicate fluently with immigrants in a foreign language, you must not only understand what the other person is saying, but also respond in that language. Similarly, a computer performs natural language processing. It first converts the human language input into a computer language in its library, then processes it through an existing program, and finally converts it into sentence output that humans can directly understand, thereby enabling a conversation-like process. Using machine learning primarily, computers must construct libraries that correspond between the output and input of natural human language.

Machine learning systems store words and their combinations as they would any other type of data. Everything from words and phrases to sentences, paragraphs, and sometimes even entire books is fed into machine learning algorithms, where they are processed using either grammar rules, people's actual language habits, or both. Both natural language processing and machine learning are operational techniques that arose in the context of the development of computer technology as one of the manifestations of artificial intelligence techniques. At its foundation, the system is programmed to function automatically by simulating human intelligence [5].

In natural language processing, when a computer receives a piece of text, it executes the procedures outlined below. First, the text in the corpus is preprocessed (Preprocess), then each sentence is tokenized (Tokenize), then the corresponding feature vector is generated (Make Feature), and lastly it is fed to a learner (Machine Learning) for learning.

### *2.2. Core task: NLP=NLU+NLG*

By describing the fundamental principles and steps of NLP, it is evident that the central mission of NLP is the same as translation between two human languages, i.e., understanding and expressing the language.

DOI: 10.54254/2755-2721/19/20231026

NLU and NLG are the accepted acronyms for "natural language understanding" and "natural language generation," respectively.

*2.2.1. The principle of NLU and the existing problems.* The objective of natural language understanding is to have machines behave like humans and understand language normally. However, how is it possible for a computer to comprehend natural human language? The principle has undergone three iterations as the field of artificial intelligence has undergone three iterations. Initially, individuals determined the purpose of natural language by summarizing principles. Later, as the use of mathematics and statistics in computing became widespread, the NLU founded on statistics emerged. NLU's foundational principles were established when deep learning was applied to natural language processing. The most cutting-edge of them is Transformer, which has been implemented in ChatGPT in addition to BERT and GPT-2, both of which have gained significant traction recently. It is qualitatively superior, more parallel, and more rigorous than previous models [6].

NLU has a wide spectrum of applications and is used in nearly all applications involving written language and speech. Examples include machine customer service, which is utilized in a number of industries, as well as smart furniture and smart speakers, which have gained popularity in recent years.

Despite the rapid development of natural language understanding technology over the past few years, it still confronts a number of obstacles. Because it is, after all, a machine, it does not always correctly interpret the meaning of each sentence in its various settings. Natural language can be combined in an extremely versatile manner. Various permutations can communicate multiple meanings. Furthermore, it is challenging to identify universal principles for natural language. Always, there are numerous exceptions. In addition, different words convey distinct meanings based on context, which is difficult for a computer to distinguish like a human. Future development shouldn't ignore these concerns.

*2.2.2. Steps and applications of NLG.* The generation of natural language is a crucial component of natural language processing. NLG is the process by which machines convert data from a non-linguistic format into a human-understandable linguistic format. Early NLG could merely consolidate data or generate output using a predetermined template. The state-of-the-art NLG used in today's NLP sector brings the field closer to mimicking human thought. It comprehends intent, adds intelligence, considers context, and presents the results in an easy-to-read and comprehend narrative.

The process of natural language generation can be broken down into five steps.

Step 1: Content determination. The NLG system must choose which pieces of information it receives in text or data format should be incorporated into the text it is creating, and which should be discarded. Like a funnel, it sorts through the incoming data to find the best possible course of action.

Step 2: text sorting. Syntactic awareness is a skill that is frequently put to use in casual conversation. The sequence of words is also important during natural language production. Location nouns, for instance, might refer to anything from a single building to an entire country. Dates should be written out using the standard format of day, month, and year. (Of course, the order of things may be different in various languages. In Chinese, for instance, the date is written from the big to the little to indicate the year, the month, and the day. The NLG system must be able to differentiate between these.)

Step 3: Sentence aggregation. Not all information has to be expressed in its own sentence, just as people don't always write in simple sentences. It can be more readable and natural if you combine several pieces of information into one statement. The NLG system must then assemble the individual pieces of information into sentences that adhere more closely to the rules of human language.

Step 4: grammaticalization. Once the meaning of each sentence has been established, the data must be structured to mimic normal speech. In this stage, we add conjunctions between the various sections of text to help it read more smoothly.

Step 5: Language realization. The next step is to establish all the pertinent words and phrases, the basic units of information, and then integrate them into a well-structured complete sentence that may be output into paragraphs or articles one at a time.

DOI: 10.54254/2755-2721/19/20231026

Natural language generation has shown renewed life in recent years, thanks to the fast progress of NLP and deep learning. NLG's widespread application in areas such as intelligent assistants, machine translation, human-machine communication, intelligent writing, etc., is a reflection of AI's growing intelligence [7].

## 3. Tokenization (Chinese&English)

### *3.1. Tokenization and its importance*

Words are the fundamental unit for sophisticated syntactic and semantic analysis in current NLP algorithms that are widely utilized around the world. As a result, word segmentation is typically NLP's first order of business [8]. Tokenization is a common term for disambiguation in the field of natural language processing. In order to facilitate further processing and analysis, lengthy texts like sentences, paragraphs, and articles are tokenized into word-based data structures. This procedure is reminiscent of the syllable-by-syllable breakdown of English that we learned in kindergarten. For example:

"NLP is a technology that makes a computer to understand human language."

This sentence can be broken down into:

"NLP/ is /a technology /that makes a computer /to understand /human language."

The fundamental nature of words is the starting point for comprehending the necessity of Tokenization. A word contains the maximum amount of information possible. Because of the white space between English words, they often serve as the building blocks of sentences. Smaller characters combine to form Chinese and Japanese words. Using letters as the fundamental building block of these languages could lead to uncertainty. However, if the computer directly processes the textual information in the sentences, it would require too much time to do the task. Therefore, NLP's first step is segmentation of the received textual information, which kicks off the process of processing and transformation that follows.

Overall, word segmentation is the most fundamental and crucial component of NLP technology. The accuracy of the segmentation has a direct bearing on the processing impact, which in turn influences the quality of the subsequent syntactic and semantic analysis [9].

### *3.2. Tokenization in Chinese & English*

Languages have different approaches to word segmentation. We use Chinese and English as case studies in this research. Both the Chinese and English tokenization processes have their own unique challenges and quirks. Words can be retrieved simply and reliably from Latin languages like English because of the use of spaces between words to signify word boundaries. In contrast, there are no clear dividing lines between words in Chinese, Japanese, and other languages, making it more challenging to extract words [8]. Furthermore, many terms in Chinese have numerous meanings, which might lead to confusion. The same word might mean something entirely different depending on the context in which it is used.

In contrast, English words undergo numerous changes that provide depth and nuance to the language. For instance, nouns can be either singular or plural, countable or uncountable, and verbs can take on a variety of tense and person forms. Word form reduction and stem extraction are special processing processes in English NLP that aren't present in Chinese; they involve reducing changed words to prototypes.

### *3.3. Common methods and comparison*

There are three common word segmentation methods: dictionary segmentation, Understanding segmentation and Statistical segmentation.

First, dictionary segmentation, whose foundation is in the concept of matching words from dictionaries. Words in the dictionary are matched with segments of text that have been modified according to predetermined guidelines. If a match is found, the words will be separated into dictionaries according to their meanings; if not, the system will make necessary changes. The downside to this fast and cheap strategy is its lack of flexibility.

DOI: 10.54254/2755-2721/19/20231026

Second, understanding segmentation; recognizing that syntactic and semantic examination of segmentation words is central to the concept of segmentation. Participles are avoided, and ambiguity is resolved with the help of syntactic and semantic information [10]. Participle analysis, however, calls for an extensive vocabulary and familiarity with grammar rules. This is because it mainly utilizes AI, which is both expensive and time-consuming.

Third, statistical segmentation, whose central tenet is the labeling and education of words. It's able to learn from both the context and the frequency with which words appear. As a result, it is excellent at picking up on obscure and nebulous vocabulary.

**Table 1.** Comparison of three methods of word segmentation.

| Comparative aspects | Dictionary segmentation | Understanding segmentation | Statistical segmentation |
|---|---|---|---|
| Ambiguity recognition | √ | √√√ | √√ |
| Neologism Finding | √ | √√√ | √√ |
| Algorithm complexity | √ | √√√ | √√ |
| Technological maturity | √√ | √ | √√ |
| Difficulty of implementation | √ | √√√ | √√ |
| Segmentation accuracy | √ | √√√ | √√ |
| Word segmentation speed | √√√ | √ | √√ |
| Wide application | √√ | √ | √√ |

(The more "√" it has, the more it fits this dimension)

The advantages and disadvantages of the three approaches are laid out in the table 1. To get the most out of each, multiple strategies are required. When multiple techniques are used together, segmentation can be completed more quickly and accurately.

The dictionary-and-statistics based approach to word segmentation is one such example. In order to solve the issue, this technique first employs the dictionary word segmentation approach [11], and then employs the statistical method to resolve the ambiguity and unknown word problem caused by the first step. This composite approach combines the dictionary storage structure with the robust screening power of the statistical method to maintain the speed of dictionary matching while also enhancing the accuracy of intersection ambiguity segmentation.

## 4. Conclusion

The paper examines the role of Dialog Systems in modern society, beginning with Siri and ChatGPT, and presents the widespread application of natural language processing in the era of big data. The second section compares and contrasts human natural language with computer information language to expound on the fundamental principle and process of natural language processing. The two main subfields of natural language processing that are presented are natural language understanding and natural language generation. Tokenization, the first and most important stage in natural language processing, is discussed in the third section. It explains how to segment words, how the Chinese language differs from the English language, and how to segment words using various approaches.

The fundamental ideas and techniques of NLP research are elucidated and described in this publication. Thus, limitations exist. It does not make use of the computer's massive storage capacity or its underlying working principle to perform calculations or statistical analysis. Additional procedure-based research on additional elements will be conducted in the follow-up study.

DOI: 10.54254/2755-2721/19/20231026

The study of NLP has expanded fast during the 1990s and is now widely applied. NLP is currently the most important basic technology in the study of artificial intelligence. NLP is a hot topic in AI research since it is one of the most fundamental challenges in the field. It has been called the "jewel in the crown" of AI systems.

The future of NLP growth is not without its share of obstacles. To begin, there is ambiguity on several levels, including the morphological, syntactic, semantic, pragmatic, and phonetic. Second, uncharted language phenomena are made possible by the introduction of novel lexicons, terminologies, semantics, and syntax. Finally, extensive nonlinear computation of parameters is necessary for semantic computing [12], and a fuzzy and intricate correlation of semantic information is hard to represent using a simple mathematical model. However, these challenges will eventually be resolved thanks to the progress of human research and technology.

## References

[1] Zhihui Yang, Language ability assessment of voice assistant Siri from the perspective of Linguistics[J], Sinogram Culture, vol.6, 2022, pp.156-157.
[2] ChatGPT Revolution[J], Insight China, vol.9, 2023, pp.28-29.
[3] Mengnan Wang, Research on Text Classification Method Based on NLP[J], Advances in Computer, Signals and Systems, vol.7, no.2, 2023, pp.93-100.
[4] ZHANG Fan, The Dilemma and Solution of Man-Machine Dialogue System[J], Philosophical Analysis, vol.6, no.13, 2022, pp.124-134.
[5] Ling Wei, Text Classification Based on Natural Language Processing and Machine Learning and its Application[J], Science & Technology Vision, vol.27, 2019, pp.88-89.
[6] Ashish Vaswani, Attention Is All You Need, arXiv preprint arXiv:1706.03762, 2017.
[7] Tong Li, Research on Natural Language Generation Algorithm Based on Transformer[D], Xidian University, 2022.
[8] Yuhan Xie, Application Research of Chinese Word Segmentation Model Based on Deep Learning[D], Chongqing University, 2017.
[9] Kaichang Chen, Chinese Word Segmentation in Natural Language Processing[J], Information & Computer, vol.19, 2016.
[10] Guohe Feng, Wei Zheng, A Review of Chinese automatic word segmentation[J], Library and Information Service, vol.2, no.55, 2011.
[11] Fengwen Di, Fengling Hao, Wanli Zuo, A Chinese Word Segmentation Method Combining Dictionary and Statistics[J], Journal of Chinese Computer Systems,vol.9, 2006.
[12] Ping Jiang, Research on Natural Language Processing Technology Based on Deep Learning[J], Digital Communication World, vol.1, 2021.