Proceedings of the 5th International Conference on Signal Processing and Machine Learning
DOI: 10.54254/2755-2721/99/20251770

Review of Language Structures and NLP Techniques for Chinese, Japanese, and English

Jingxuan Du[1,a,*]

[1]*Tianjin University, Tianjin, 300072, China*
a. jingxuandu@126.com
**corresponding author*

Abstract: The increasing demand for cross-lingual communication in a globally interconnected world has spurred significant advancements in Natural Language Processing (NLP), a branch of artificial intelligence aimed at enabling computers to comprehend, generate, and process human language. Over the past few years, NLP has transitioned from traditional rule-based and statistical approaches to deep learning-based methods, with pre-trained models like BERT and GPT demonstrating remarkable performance across a variety of tasks, including text classification and machine translation. However, the linguistic diversity of the world's languages presents ongoing challenges, as many NLP models are primarily developed for high-resource languages like English, leaving other languages underserved. This paper examines the linguistic features of three representative languages—Chinese, Japanese, and English—and explores NLP methodologies tailored to their distinct grammatical, lexical, and structural properties. Through comparative linguistic analysis, this study discusses the challenges of multilingual NLP and the potential for cross-lingual applications. The findings underscore the importance of adapting NLP models to different language characteristics to enhance the effectiveness and inclusivity of cross-lingual processing.

Keywords: Natural Language Processing (NLP), Vocabulary, Cross-lingual, Grammatical Structure.

1. Introduction

Language is the core medium for human communication and information exchange. With increasing globalization, the demand for cross-lingual communication is on the rise, making language processing a crucial topic across various fields.

Natural Language Processing (NLP), a significant branch of artificial intelligence, focuses on enabling computers to understand, generate, and process human language[1]. It encompasses multiple core tasks, such as text classification, sentiment analysis, machine translation, and named entity recognition[2]. Hence, NLP as a key technology enabling computers to comprehend and generate human language, has made significant progress. Despite these advancements, NLP models face unique challenges due to the diversity of the world's languages and significant imbalances in data availability.

Inspired by the above, this paper selects Chinese, Japanese, and English as representative languages to explore NLP methodologies and applications, focusing on their distinct yet overlapping

DOI: 10.54254/2755-2721/99/20251770

linguistic characteristics. These languages present significant challenges due to their differences in grammatical structures, vocabularies, and character systems, necessitating specialized strategies for effective multilingual NLP processing. Traditional approaches often fail to capture the complexities of each language, making it essential to develop tailored models that address these unique features. The study focuses on analyzing the grammatical, lexical, and structural differences between Chinese, Japanese, and English to propose effective NLP strategies for each language. Using a comparative analysis of these three languages, this study evaluates existing methodologies and identifies areas for improvement. By doing so, the paper proposes strategies that enhance individual language processing while improving cross-lingual capabilities. The goal is to offer both theoretical insights and practical solutions for developing more adaptable and inclusive multilingual NLP models. The findings of this study will contribute to the ongoing development of NLP models that can better handle diverse linguistic systems, addressing key challenges in multilingual contexts and offering potential improvements for future cross-lingual applications.

2. Language features

An examination of the grammatical structures, vocabularies, and character systems of Chinese, Japanese, and English is essential for understanding the relationship between these languages. By analyzing the similarities and differences, it is possible to gain deeper insights into their cross-language relationships, which in turn facilitates the development of NLP models. Such an understanding not only aids in tackling language-specific challenges but also enhances the ability to adapt NLP applications to diverse linguistic contexts, thereby improving multilingual and cross-lingual language processing.

2.1. Chinese

2.1.1. Language features

The Chinese language is comprised of characters that were originally pictographs but have evolved into complex logographic symbols. In contrast to alphabetic systems, where the meaning is derived from the phonetic combinations, each Chinese character possesses an independent meaning. requiring NLP models to focus on the holistic interpretation of words rather than individual characters.

The grammar of the Chinese language is relatively simple, with no complex tense, gender, number, or case inflections. In Chinese, the sentence structure is primarily determined by the order of words, with a typical subject-verb-object (SVO) pattern. In addition, Chinese employs a range of particles, including “的”, “地”, “得” and “了” to indicate completion, possession, or modification.

The principal parts of speech mainly include nouns, verbs, adjectives, adverbs, pronouns, conjunctions, prepositions, particles, and measure words. In contrast to some languages, Chinese does not differentiate parts of speech through morphological changes. In contrast, Chinese grammar relies more heavily on contextual cues and the structure of sentences. Furthermore, Chinese words do not undergo morphological changes in accordance with the number, tense, or person of the sentence.

Word segmentation is a pivotal aspect of Chinese linguistics, particularly in the context of NLP applications. In Chinese script, words are not visually separated. Sentences are written as continuous strings of characters, necessitating the segmentation of text into meaningful units for analysis.

Overall, the language system of Chinese conveys meaning through flexible structures and contextual relationships. Compared to other languages, Chinese grammar relies more on context and logical relationships rather than on morphological changes in words. For instance, the meaning of a phrase can vary significantly based on the context in which it is used.

DOI: 10.54254/2755-2721/99/20251770

2.1.2. NLP methods in Chinese

Yan Shao et al. explored the joint modeling of word segmentation and part-of-speech (PoS) tagging in Chinese, emphasizing that the identification of word boundaries facilitates more accurate PoS tagging[3]. Conversely, accurate PoS tags can facilitate the process of word segmentation. By jointly modelling these tasks, their approach achieves better overall performance in language processing for Chinese. Building upon this concept, Canasai Kruengkrai et al. proposed an error-driven word-character hybrid model for joint word segmentation and PoS tagging[4]. The model is trained on erroneous examples from a training corpus, thereby enhancing the understanding of Chinese vocabulary and addressing the inherent challenges of its lack of spaces.

2.2. Japanese

2.2.1. Language features

The Japanese language has been significantly influenced by Chinese, with the introduction of kanji (Chinese characters) influencing vocabulary and writing. This led to a dual script system combining logographic symbols (kanji) and syllabic scripts (hiragana and katakana), which provides flexibility in word formation and also adds complexity to sentence structure analysis. Following the Meiji Restoration of 1868, Japan rapidly adopted Western loanwords, especially from English, commonly found in fields like science, technology, and culture. These words are often written in katakana, such as “ﾃﾚﾋﾞ” (terebi, television) and “ﾊﾞｽ” (basu, bus), reflecting the integration of Western concepts into the language.

The Japanese sentence structure adheres to a subject-object-verb (SOV) order, with the roles of nouns indicated by particles such as “が” (ga), “は” (wa), and “を” (wo). The accurate handling of these particles is of great consequence for NLP tasks such as syntactic and semantic analysis, as they define grammatical relationships.

The honorific system, which encompasses forms of respectful and humble language, is a key characteristic of Japanese. It reflects social hierarchy and cultural norms, presenting challenges for NLP due to the way honorifics can alter sentence meaning and tone significantly. Additionally, the diverse forms of honorifics necessitate extensive annotated data to train effective NLP models, thereby rendering data collection and model evaluation more complex.

2.2.2. NLP methods in Japanese

Due to the unique emotional characteristics of the Japanese language, M. Fahim Ferdous Khan et al. addressed the challenges of sentiment analysis on social media posts, which often contain informal language not found in standard dictionaries[5]. They improved the accuracy of sentiment recognition by enhancing sentiment polarity dictionaries and utilizing advanced word embeddings. Similarly, Hirokuni Maeta et al. proposed a framework for understanding procedural texts by tokenizing input, identifying key concepts, and linking them comprehensively[6]. This approach effectively dealt with the ambiguity and complex syntax of Japanese text, enhancing NLP capabilities for varied applications. Additionally, Siqi Peng et al. implemented a clustering strategy to extract and summarize temporal expressions from Japanese news archives, leveraging grammatical features for the efficient identification of relevant terms[7]. Each of these works contributes to enhancing the processing of Japanese language complexities within NLP.

DOI: 10.54254/2755-2721/99/20251770

2.3. English

2.3.1. Language features

The English language originated in the 5th century as Old English, which was influenced by the Germanic tribes. Over time, it absorbed a substantial amount of vocabulary from French and Latin, gradually evolving into a globally dominant language. Today, English is characterized by its simplified grammatical structure, rich vocabulary, and flexible word formation, which have contributed to its status as a primary tool for international communication. The Latin alphabet, comprising 26 letters, constitutes the foundation of its vocabulary. The relatively straightforward script facilitates ease of learning and widespread dissemination.

English adheres to a subject-verb-object (SVO) order, which creates a clear sentence structure and enhances readability. English grammar is notable for its complex tense system, which indicates actions in the past, present, and future. Additionally, parts of speech, including nouns, verbs, and pronouns, undergo inflections to express tense and number. These grammatical characteristics present challenges for NLP, as accurately identifying word forms and context is crucial for syntactic and semantic understanding, particularly in tasks like pronoun resolution.

Word segmentation in the English language is relatively straightforward, with words typically separated by spaces, in contrast to languages like Chinese or Japanese. However, challenges remain in handling punctuation and compound words. English has been central to NLP research, partly because it dominated the early development of computer science and artificial intelligence, providing a rich base of computational resources, datasets, and research literature.

2.3.2. NLP methods in English

Natural language processing initially focused primarily on English, which led to many NLP methods being developed and applied specifically for English. This is due to the fact that English dominated the early development of computer science and artificial intelligence, and a considerable number of computational resources, datasets, and research literature were based on English.

The history of English natural language processing can be divided into several significant phases. The initial stage (1940s-1950s) commenced with the proposal of American mathematician Warren Weaver, who was the first to suggest using computers for language translation. In the 1950s and 1960s, two major trends emerged in NLP: the symbolists and the frequentists[8]. The developmental stage (1960s-1970s) saw machine translation divided into three principal phases: lexical and syntactic analysis, transformation of lexical and grammatical structures, and generation of lexical and syntactic outputs. With the advent of more powerful computational resources, NLP entered a period of rapid growth, characterized by the processing of large-scale data sets and a notable enhancement in usability. In the 21st century, the introduction of deep learning, particularly with the advent of large pre-trained models like BERT and GPT, has led to tremendous breakthroughs in NLP research and applications, driving advancements in text generation, information extraction, and dialogue systems. This evolution demonstrates the rapid transformation of NLP from early theoretical explorations to its current practical applications today.

3. Cross-language NLP

Cross-language NLP is a subfield of NLP that focuses on the processing, understanding, and generating natural language across multiple languages. There are many challenges in Cross-language NLP, such as the limited digital resources, grammatical and structural differences, and code-switching.

For the cross-application of NLP across multiple languages, there are many challenges to consider. Firstly, the issue of insufficient multilingual corpora and the presence of mixed languages within the

DOI: 10.54254/2755-2721/99/20251770

same text must be addressed. Given the considerable diversity of different languages and their various writing systems, multilingual NLP models require strong adaptability. Secondly, it is crucial that trained models are capable of adaptive recognition and generation across different languages. Additionally, in cross-lingual information retrieval, there may be various expressions for certain idiomatic phrases, which necessitates improving the training accuracy.

Due to the limited availability of resources for multilingual training, Lample, G. et al. introduced a new unsupervised and supervised learning method for in their work on Cross-lingual Language Model Pretraining[9]. This method has the potential to enhance cross-lingual pretraining and mitigate the challenges encountered by low-resource models. Based on Wikipedia, Zhao, X. et al. categorized three patterns of acquiring and representing facts as follows: Language-independent knowledge refers to the ability of the model to understand and express certain knowledge independently across different languages, without relying on information from other languages[10]. Cross-lingual shared knowledge involves leveraging information from the same knowledge source, allowing for similar knowledge representations to be shared across multiple languages. Cross-lingual transferred knowledge pertains to applying knowledge learned in one language to others through transfer mechanisms, enhancing the model's performance across languages. Additionally, the multilingual Text-to-Text Transfer Transformer (mT5) model developed by Xue, L. et al., facilitates the transfer of knowledge between languages[11].

These studies demonstrate the potential of multilingual models in multilingual situations, particularly in tasks such as summarization and translation. As cross-lingual NLP techniques continue to develop, future research will focus on optimizing optimizing model performance and applicability across different language structures. Additionally, advancements in cross-lingual technologies contribute to the creation of multilingual corpora and facilitate the widespread deployment of language models in real-world applications, enhancing multilingual interaction and information exchange. Ran Z. et al. summarized and abstracted historical documents in different languages, thereby facilitating the exchange and understanding of information and contributing to the accessibility of culture[12]. Puduppully, R et al. generated fluent and accurate translations by evaluating relevant languages from different language families[13]. This approach allowed for the transfer of knowledge between languages, facilitating the accessibility of cultural information. Etxaniz, J. et al. leveraged the translation capabilities of multilingual language models in a few-shot setting through self-translate, eliminating the need to rely on external translation systems[14].

4. Conclusion

This paper provides an in-depth analysis of the challenges and methodologies associated with NLP for Chinese, Japanese, and English, emphasizing the need for language-specific approaches due to their distinct grammatical, lexical, and structural features. Key challenges include the logographic writing system of Chinese, which requires specialized word segmentation techniques, the dual-script system of Japanese, combining kanji with syllabic kana, and the complex tense and grammatical structures of English. These linguistic complexities highlight the importance of developing tailored NLP strategies for each language. Additionally, cross-lingual techniques such as transfer learning, multilingual transformers, and unsupervised learning methods have demonstrated significant potential in bridging linguistic gaps and improving multilingual model performance. Future research should focus on incorporating finer language-specific nuances, such as particles in Chinese, honorifics in Japanese, and complex verb tenses in English, into multilingual models. Moreover, addressing resource disparities for low-resource languages is critical to ensure the broader applicability of NLP models. Ultimately, this research contributes to advancing more adaptable and inclusive NLP models, capable of handling a wider range of languages and improving global communication.

DOI: 10.54254/2755-2721/99/20251770

References

[1] Das, S., & Das, D. (2024) Natural Language Processing (NLP) Techniques: Usability in Human-Computer Interactions. Proceedings of the 2024 6th International Conference on Natural Language Processing (ICNLP): 783-787.

[2] Qin, L., Chen, Q., Zhou, Y., Chen, Z., Li, Y., Liao, L., Li, M., Che, W., & Yu, P.S. (2024) Multilingual Large Language Model: A Survey of Resources, Taxonomy and Frontiers. arXiv preprint arXiv:2404.04925.

[3] Ng, H.T., & Low, J.K. (2004). Chinese Part-of-Speech Tagging: One-at-a-Time or All-at-Once? Word-Based or Character-Based? Proceedings of the 2004 Conference on Empirical Methods in Natural Language Processing (EMNLP) : 277-284.

[4] Kruengkrai, C., Uchimoto, K., Kazama, J., Wang, Y., Torisawa, K., & Isahara, H. (2009). An Error-Driven Word-Character Hybrid Model for Joint Chinese Word Segmentation and POS Tagging. Proceedings of the Joint Conference of the 47th Annual Meeting of the ACL and the 4th International Joint Conference on Natural Language Processing of the AFNLP: 513-521.

[5] Khan, M.F.F., Kanemaru, A., & Sakamura, K. (2022). Sentiment Analysis of Japanese Tweets Using Auto-Augmented Sentiment Polarity Dictionaries and Advanced Word Embedding. 2022 IEEE 11th Global Conference on Consumer Electronics (GCCE): 462–466.

[6] Maeta, H., Sasada, T., & Mori, S. (2015). A Framework for Procedural Text Understanding. Proceedings of the 14th International Conference on Parsing Technologies, 50–60.

[7] Peng, S., Yamamoto, A., Mori, S., & Sekino, T. (2022). Event Time Extraction from Japanese News Archives. 2022 IEEE International Conference on Big Data (Big Data): 2556–2564.

[8] Jiang, K., & Lu, X. (2020). Natural Language Processing and Its Applications in Machine Translation: A Diachronic Review. In 2020 IEEE 3rd International Conference of Safe Production and Informatization (IICSPI) :210–214.

[9] Lample, G., & Conneau, A. (2019). Cross-lingual language model pretraining. arXiv preprint arXiv:1901.07291.

[10] Zhao, X., Yoshinaga, N., & Oba, D. (2024). Tracing the Roots of Facts in Multilingual Language Models: Independent, Shared, and Transferred Knowledge. Proceedings of the 18th Conference of the European Chapter of the Association for Computational Linguistics, 2088–2102.

[11] Xue, L. et al. (2021). mT5: A Massively Multilingual Pre-trained Text-to-Text Transformer. In Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, 483–498.

[12] Zhang, R., Ouni, J., & Eger, S. (2024). Cross-lingual Cross-temporal Summarization: Dataset, Models, Evaluation. Computational Linguistics, 50(3), 1001–1047.

[13] Puduppully, R., Kunchukuttan, A., Dabre, R., Aw, A. T., & Chen, N. (2023). DecoMT: Decomposed Prompting for Machine Translation Between Related Languages Using Large Language Models. Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing, Singapore, December 2023. 3735–3747.

[14] Etxaniz, J., Azkune, G., Soroa, A., Lopez de Lacalle, O., & Artetxe, M. (2024). Do Multilingual Language Models Think Better in English? Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, 550–564.