Large Language Models for Chinese POS Tagging and Word Segmentation: A Comparison with XLM-RoBERTa and Jieba

23/24 WS Computational Linguistics Final Project Report
Ho-Hsuan Wang
Student Number: 7038925
March 2024

***Abstract*—This report explores the efficacy of advanced Large Language Models (LLMs), specifically GPT-4, compared to neural network and statistical models in performing Chinese word segmentation and part-of-speech (POS) tagging on the Universal Dependencies Traditional Chinese dataset. Results demonstrate that while neural network models like XLM-RoBERTa significantly outperform GPT-4 in POS tagging, GPT-4 surpasses statistical models like Jieba in word segmentation accuracy with appropriate prompting. The study highlights the impact of model design and prompting methods on task-specific performance when leveraging LLMs for Chinese language processing and learning.**

***Index Terms*—XLM-RoBERTa, LLM, POS tagging, Chinese word segmentation (CWS), prompting**

I. Introduction

THE study of Chinese grammar, initiated during the Qing dynasty (1644–1911), represented a significant shift in the analysis of Chinese languages, and this shift was largely influenced by Western missionaries who adapted Western linguistic categories for teaching Chinese to Westerners [1]. In Chinese, the same word can adopt multiple syntactic functions without changing its form, diverging from conventional part-of-speech categorization standards often seen in grammar research and education. This also led to modern Chinese vocabulary not having unified standards for categorizing parts of speech for a long time [2]. Take a simple noun/verb change as an example, "研究" (yánjiū) acts as a verb in 我在研究這個問題。 (Wǒ zài yánjiū zhège wèntí, "I am studying this issue.") and as a noun in "這項研究很困難。" (Zhè xiàng yánjiū hěn kùnnán, "This study is very difficult."). This demonstrates the analytical nature of the Chinese language, where words can change function without the use of declination or conjugation affixes.

There has been a debate about whether Chinese shares all the parts of speech with Indo-European languages. Despite the differences that do exist, it is widely acknowledged that Chinese and Indo-European languages share six major categories: nouns, verbs, adjectives, adverbs, prepositions, and conjunctions. Unique to Chinese, however, are classifiers, localizers, and sentence-final particles, highlighting Sinitic language features [3].

Regarding POS tagging in Chinese, it is essential to address word segmentation first due to the structure of the Chinese language, where there is no white space between words. This process leads to the question of whether to separate word segmentation and POS tagging into two phases or combine them into a single step. Research has found that performing CWS and POS tagging simultaneously produces the best outcome [4] [5]. However, due to time limitations and the availability of more detailed labeled data that includes not only POS tags but also labels for the relative character position in a word, this final report will address these two tasks separately. CWS is a crucial initial step required to achieve high performance in various Chinese NLP tasks, such as speech recognition, machine translation, language understanding, and others [6]. This project, however, is concerned with the challenges these tasks bring to learners studying Chinese as a second language and how much can LLM be a reliable tool for them. Therefore, this project aims to evaluate the performance of advanced LLMs in CWS and POS tagging by comparing their performance with popular tools such as Jieba and a neural network model, XLM-ROBERTa. Finally, the report will evaluate whether state-of-the-art LLMs can assist learners in overcoming the challenges of learning Chinese.

II. Problem Statement

A plethora of papers exist on CWS and POS tagging. However, if not none, only a few have explored the potential of LLMs for these particular tasks. As a result, the objective of this report is to answer the following question: Can the state-of-the-art LLM compete with neural network and statistical models in performing CWS and POS tagging? Furthermore, can it be relied on as a useful tool for those learning the Chinese language?

III. Data and Preprocessing

The dataset used in this project is the Universal Dependencies (UD) Traditional Chinese dataset (`UD_Chinese-GSD`[1]), although it might not be the one that can most accurately describe Chinese part-of-speech. For example, according to

[1]https://github.com/UniversalDependencies/UD_Chinese-GSD/tree/master

the paper written by Poiret et al. [7], Chinese classifiers have a unique syntactic role that has led to specific POS tags, such as 'M' for measure words and 'q' for quantity, in the Penn Chinese Treebank and the Chinese Dependency Treebank. However, the UD framework, which avoids language-specific tags, classifies classifiers as nouns in version 2. Classifiers could also be considered PART when they play a more functional role. The UD's approach to tagging classifiers as nouns represents a compromise to facilitate comparison across languages that do not use classifiers. Nonetheless, this dataset is ideal for its comprehensiveness and standardization, useful for benchmarking, and ensuring that the findings are comparable across different studies.

IV. Methodology

The full implementation code is available here [2].

A. XLM-RoBERTa Training - POS Tagging

I trained an XLM-RoBERTa model on POS tagging to evaluate its performance against GPT -4's prompting accuracy for the same task.

The architecture was modified to fine-tune only the classifier layer while keeping other parameters static. The training arguments and hyperparameters that yielded the best result are as follows:

- num_train_epochs (12)
- per_device_train_batch_size (8)
- per_device_eval_batch_size (32)
- logging_steps (10)
- learning_rate (1e-3)
- max_length (512)
- padding ('max_length')

At the 12th epoch, the accuracy plateaus at 0.901, with training and validation loss being 0.376 and 0.300, respectively. The full Wandb training report [3] is available.

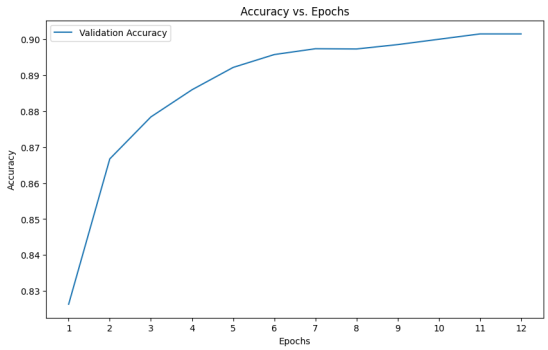

Fig. 1. Accuracy v.s. Epoch

[2]https://github.com/rozariwang/coli_final_project
[3]https://api.wandb.ai/links/hohsuann/pgy8gp3l

B. Jieba - Word Segmentation

Jieba [4] is a well-known tool for Chinese text segmentation tasks. It segments Chinese text by constructing a Directed Acyclic Graph (DAG) for each sentence to include all possible word combinations, then uses dynamic programming to select the most probable segmentation based on cumulative word frequency. For unknown words not in its dictionary, Jieba employs a Hidden Markov Model (HMM) with the Viterbi algorithm to predict segmentation, leveraging statistical algorithms rather than deep learning to adapt to new words and contexts. For the word segmentation task in this project, I used Jieba's Precise Mode, which segments sentences into words without generating all possible segments for one input.

C. GPT-4 Prompting - POS Tagging & Word Segmentation

I evaluated the proficiency of the GPT-4 Turbo (gpt-4-0125-preview) model in executing CWS and POS tagging tasks using the first 200 samples from the UD Traditional Chinese test set. The methodology involved using three different prompts for each task. The complete prompts are shown in APPENDIX A and B.

The first prompt for tokenization briefly explains the task's concept and is complemented by an example showing the tokenization result of a sentence derived from the training set. The first Prompt for POS tagging includes the full names of each POS tag, explains how some common Chinese characters are tagged in the UD dataset, and provides an example of the correct tagging in a given sentence.

The second set of prompts adopts a zero-shot approach. These prompts were minimalist, offering only a list of the 16 abbreviated names of the POS tags without further elaboration or examples.

The third set of prompts employed a Chain-of-Thought (CoT) strategy. This approach is similar to the first prompt in providing an explanatory narrative but distinguished itself by including reflective sentences such as "Let's break down the process step by step..." and "Let's analyze the sentence...". However, unlike its counterpart in the tokenization task, the CoT prompt for POS tagging did not include a fully annotated POS tagging example.

D. Prompting Accuracy Calculation

The POS tagging accuracy for each response is calculated by comparing it to the reference string. Each item in the reference and response lists is separated by a comma ", " and the accuracy for each sentence is the ratio of correct matches to the total number of items in the reference list. The overall accuracy is the average of all individual accuracies for the 200 sample sentences used in the test set. There are 500 samples in the test dataset, but I have only included the first 200 samples for both tasks for GPT-4 prompting due to resource limitations. The XLM-ROBERTa validation accuracy, however, includes the total 500 samples from the test set.

[4]https://github.com/fxsjy/jieba

For tokenization accuracy, the predicted tokens (also split by ", ") are converted into a set. The accuracy for each response is calculated by finding the intersection of predicted tokens and reference tokens (correct tokens) and dividing the size of this intersection by the total number of reference tokens. I also used the same 200 samples and accuracy calculation method to assess the tokenization accuracy of Jieba.

In other words, POS tagging accuracy is based on a direct comparison of linear sequences (lists) and is calculated by comparing each element positionally in the sequence, suitable for POS tagging where the order is crucial. For tokenization accuracy, it is calculated based on the presence of tokens using sets, which is more appropriate for tokenization where the order might not be as critical.

In the beginning, I made the mistake of using the same accuracy calculation method for both tasks; namely, I was using the way to calculate POS tagging accuracy to calculate tokenization accuracy as well, but that led to tokenization accuracy for GPT-4 with Prompt 1 being as low as 34%. For example, suppose a sentence "只是二選一做決擇" is tokenized into "只是", "二", "選", "一", "做", "決擇" while the correct tokenization result in the UD dataset is "只", "是", "二", "選", "一", "做", "決擇". In this case, the entire tokenization accuracy will become 0% even though the rest of the tokenization result except the first token is correct since the first token messes up the order for the rest of the sentence. The tokenization accuracy increased drastically after applying a more suitable way to calculate tokenization accuracy.

V. Results and Discussion

The result shows that the neural networks-based XLM-ROBERTa model performs the best for POS tagging, with about a 15% improvement in accuracy compared to the GPT-4 CoT (Prompt 3) prompting result. However, neural models are capable of achieving ever higher accuracies. In the paper by Tian et al. [4], they used a character-based neural model with multi-channel attention of n-grams for the joint task of CWS and POS-Tagging. Their model achieved over 95% accuracy on the UD Chinese dataset. This result has confirmed that the SOTA LLMs still cannot be compared with neural models in these tasks. Nevertheless, regarding statistical word segmentation tools like Jieba, the LLMs could easily outperform them with the right prompting. As shown in Table 1, Prompt 1 achieves about 5% higher accuracy in word segmentation than Jieba.

For the zero-shot prompting approach (Prompt 2), it has also confirmed the findings by Moghaddam et al. [8] that GPT-4 performs better when some in-context learning is provided. This is even more prominent in the word segmentation task where the zero-shot approach performs almost 50% worse than the prompts that provide task explanations and tokenization examples. It can also be inferred from the table that GPT-4 is inherently much better at POS tagging than tokenization under the zero-shot condition. This might be because of its extensive pre-training on diverse linguistic data, enabling it to understand underlying grammatical rules and part-of-speech relationships without explicit examples. Conversely, Chinese tokenization requires a deeper understanding of Chinese characters and phrases' specific context and nuances. Therefore, it may benefit more from task-specific examples, which are unavailable in a zero-shot scenario.

TABLE I
Comparison of Word Segmentation and POS-Tagging Accuracies

Task/Model	XLM-R	Jieba	GPT-4 Turbo (gpt-4-0125-preview)		
			P1	P2	P3
Word Seg.	-	0.748	**0.796**	0.388	0.747
POS-Tag.	**0.901**	-	0.747	0.71	0.75

Note: The best results are highlighted in bold. XLM-R refers to XLM-RoBERTa. P refers to Prompt. P2 is zero-shot. P3 is Chain-of-thought (CoT). P1 is close to P3, and the difference is that it provides a full example of each task.

VI. Conclusion

The investigation into the capabilities of state-of-the-art LLMs for CWS and POS tagging confirms the superiority of neural network models for these tasks. Conversely, GPT-4, through strategically crafted prompts, demonstrates a competitive edge over statistical models like Jieba in word segmentation, suggesting that LLMs can serve as a valuable tool for Chinese language learners. This is especially true since NN models dedicated to these tasks are often not readily available to the general public compared to GPT-4.

In this conversation I had with ChatGPT4 [5], besides completing the tokenization and POS tagging tasks, it repeatedly reminded me that the results it has provided were very basic and could not be the most accurate. In this conversation, I tried both tasks in a zero-shot manner using the first sentence in the UD development dataset [6]. After its failed attempt to access Jieba, it did not provide any tokenization result. Afterward, I removed the keyword "Chinese" and simply asked it to perform "word segmentation", it gave a result that matches the UD tokenization very well. To the point that I, as a native Traditional Chinese speaker, would say that I agree with ChatGPT4's tokenization output more than the UD way of tokenizing the sentence, most likely due to the language-specific compromises that UD has to make for it to be universal. While the results from GPT-4 need to be taken with a grain of salt, just like in most other scenarios, and GPT-4 is clearly aware of this too, LLMs can indeed serve as a quick and convenient way to segment Chinese sentences and offer POS tags for Chinese as second language learners.

References

[1] M. Gianninoto, "The development of chinese grammars and the classification of the parts of speech," *Language & History*, vol. 57, no. 2, pp. 137–148, 2014.

[2] Z. Ping, "On part of speech of modern chinese vocabulary," *Yinshan Academic Journal*, 2007.

[5] https://chat.openai.com/share/d944e477-fc84-4a80-8271-bbea1d58227b

[6] https://github.com/UniversalDependencies/UD_Chinese-GSD/blob/master/zh_gsd-ud-dev.conllu

[3] C. C.-H. Cheung, *Cheung, Chi-Hang Candice. (2016) Chinese: Parts of speech. In Sin-Wai Chan, ed., The Routledge Encyclopedia of the Chinese Language. New York: Routledge.*, 02 2016, pp. 242–294.
[4] Y. Tian, Y. Song, and F. Xia, "Joint Chinese word segmentation and part-of-speech tagging via multi-channel attention of character n-grams," in *Proceedings of the 28th International Conference on Computational Linguistics*, D. Scott, N. Bel, and C. Zong, Eds. Barcelona, Spain (Online): International Committee on Computational Linguistics, Dec. 2020, pp. 2073–2084. [Online]. Available: https://aclanthology.org/2020.coling-main.187
[5] H. Ng and J. Low, "Chinese part-of-speech tagging: One-at-a-time or all-at-once? word-based or character-based?" 01 2004, pp. 277–284.
[6] H. Li and B. Yuan, "Chinese word segmentation," in *Proceedings of the 12th Pacific Asia Conference on Language, Information and Computation*, J. Guo, K. T. Lua, and J. Xu, Eds. Singapore: Chinese and Oriental Languages Information Processing Society, Feb. 1998, pp. 212–217. [Online]. Available: https://aclanthology.org/Y98-1020
[7] R. Poiret, T.-S. Wong, J. Lee, K. Gerdes, and H. Leung, "Universal dependencies for mandarin chinese," *Language Resources and Evaluation*, vol. 57, no. 2, pp. 673–710, 2023. [Online]. Available: https://doi.org/10.1007/s10579-021-09564-2
[8] S. R. Moghaddam and C. J. Honey, "Boosting theory-of-mind performance in large language models via prompting," 2023.

APPENDIX A
TOKENIZATION PROMPT

A. Prompt 1

Tokenization splits words and punctuation in a sentence, keeping meaningful words intact. For example, if we tokenize the following Chinese sentence: 看似簡單，只是二選一做決擇，但其實他們代表的是你周遭的親朋好友，試著給你不同的意見， the tokenized result is: 看似, 簡單, ，, 只, 是, 二, 選, 一, 做, 決擇, ，, 但, 其實, 他們, 代表, 的, 是, 你, 周遭, 的, 親朋, 好友, ，, 試, 著, 給, 你, 不同, 的, 意見, 。 Now, please tokenize the following Chinese sentence in the same way: 他花費了許多時間來比較加拿大地質調查局博物館中的恐龍化石。 Output only the tokenized result separated by commas.

B. Prompt 2 - Zero-Shot

Tokenize the following Chinese sentence: 他花費了許多時間來比較加拿大地質調查局博物館中的恐龍化石。 Output the tokenized result separated by commas.

C. Prompt 3 - CoT

Tokenization is the process of splitting a sentence into meaningful units such as words and punctuation. Let's break down the process step by step using an example sentence in Chinese. Consider the sentence: 看似簡單，只是二選一做決擇。 First, identify independent words and symbols. For instance, 看似 is a word that means 'seems' and is kept together because it forms a single concept. Similarly, 簡單 means 'simple' and is another unit. Punctuation like ， is separated as it marks the sentence boundary. Following this logic, the tokenized result is: 看似, 簡單, ，, 只, 是, 二, 選, 一, 做, 決擇, 。 Now, using the same thought process, please tokenize the following Chinese sentence: 他花費了許多時間來比較加拿大地質調查局博物館中的恐龍化石。 Output only the tokenized result separated by commas.

APPENDIX B
POS TAGGING PROMPT

A. Prompt 1

Note: '不, 了, 事' are often tagged as PART, '之, 他, 她' as PRON, and '一切, 全, 每' as DET. Some lemmas like '了' and '是' can be AUX or VERB based on context.

For example, the sentence '然而，這樣的處理也衍生了一些問題。' would be tokenized and tagged as follows:

然而: Subordinating Conjunction (SCONJ), ，: Punctuation (PUNCT), 這樣: Pronoun (PRON), 的: Particle (PART), 處理: Noun (NOUN), 也: Subordinating Conjunction (SCONJ), 衍生: Verb (VERB), 了: Auxiliary (AUX), 一些: Adjective (ADJ), 問題: Noun (NOUN), 。: Punctuation (PUNCT)

Now, for the following tokenized Chinese sentence, provide the POS tags. The possible tags with their full names are: SYM (Symbol), ADP (Adposition), ADV (Adverb), AUX (Auxiliary), PRON (Pronoun), PUNCT (Punctuation), PART (Particle), SCONJ (Subordinating Conjunction), DET (Determiner), X (Other), NOUN (Noun), NUM (Numeral), PROPN (Proper Noun), ADJ (Adjective), CCONJ (Coordinating Conjunction), VERB (Verb).

Here is the tokenized sentence:

他, 花費, 了, 許多, 時間, 來, 比較, 加拿大, 地質, 調查, 局, 博物, 館, 中, 的, 恐龍, 化石, 。

Output only the POS tags (short forms: "SYM", "ADP", "ADV", "AUX", "PRON", "PUNCT", "PART", "SCONJ", "DET", "X", "NOUN", "NUM", "PROPN", "ADJ", "CCONJ", "VERB") separated by commas and in the same order as the words.

B. Prompt 2 - Zero-Shot

Here is the tokenized Chinese sentence: 他, 花費, 了, 許多, 時間, 來, 比較, 加拿大, 地質, 調查, 局, 博物, 館, 中, 的, 恐龍, 化石, 。 Tag each token with the correct part of speech from the following list: SYM, ADP, ADV, AUX, PRON, PUNCT, PART, SCONJ, DET, X, NOUN, NUM, PROPN, ADJ, CCONJ, VERB Output only the POS tags separated by commas and in the same order as the tokens.

C. Prompt 3 - CoT

POS (Part of Speech) tagging is the process of labeling tokens with their corresponding part of speech. This helps in understanding the role of each word in a sentence. For example, nouns are subjects or objects, verbs express actions, and adjectives describe nouns. Let's analyze the sentence 然而，這樣的處理也衍生了一些問題。 step by step. 然而 is a subordinating conjunction (SCONJ) as it connects clauses. 處理 is a noun (NOUN) because it represents an entity or concept. The process continues for each token to assign the correct tag. Now, apply the same method to tag the following tokenized sentence. The possible tags with their full names are: SYM (Symbol), ADP (Adposition), ADV (Adverb), AUX (Auxiliary), PRON (Pronoun), PUNCT (Punctuation), PART (Particle), SCONJ (Subordinating Conjunction), DET (Determiner), X (Other), NOUN (Noun), NUM (Numeral), PROPN

(Proper Noun), ADJ (Adjective), CCONJ (Coordinating Conjunction), VERB (Verb). Here is the tokenized sentence: 他, 花費, 了, 許多, 時間, 來, 比較, 加拿大, 地質, 調查, 局, 博物, 館, 中, 的, 恐龍, 化石, 。