Exploring Large Language Models for Chinese Spoken Language Understanding

Project of CS3602 Natural Language Processing, 2023 Fall, SJTU

Xiangyuan Xue (521030910387)

School of Electronic Information and Electrical Engineering

Abstract

Spoken Language Understanding (SLU) is one of the modules in spoken dialogue systems, which interprets spoken language into semantic information. With the advent of Large Language Models (LLMs), SLU meets a new era. In this project, we explore the application of LLMs to the slot filling task in SLU. We conduct adequate experiments to investigate the effects of prompt engineering, where precise description, chain-of-thought, emotional stimulation and prompt language are considered. We also explore the effects of learning from samples, where zero-shot, one-shot and few-shot paradigms are tried to form prompts. In addition, we compare the performance of different LLMs currently in the market. Besides, we give our analysis and discussion of the research status of SLU in the era of LLMs. We believe that LLMs can bring a revolution to SLU and provide more research directions, while traditional methods are still useful in some specific scenarios.

1 Introduction

In recent years, the advent of Large Language Models (LLMs) [18] has dramatically transformed the landscape of natural language processing (NLP) [3]. Their popularity and widespread use stem from their remarkable capabilities to comprehend, generate, and translate text across a variety of languages and contexts. The sheer scalability and the depth of knowledge these models demonstrate carry significant research value. Spoken language understanding (SLU) [10] is an intricate task within the field of computational linguistics, primarily due to the colloquial and often ambiguous nature of speech. LLMs hold the potential to unlock sophisticated approaches to SLU by leveraging their profound syntactic and semantic capabilities.

This project is an exploration into the apposite application of LLMs to the realm of Chinese spoken language understanding, with a focus on the analysis of semantic triples [4]. The semantic triple, consisting of an action, a slot and a value, encompasses the barebones structure necessary for encapsulating a single, coherent piece of information or fact. Successfully parsing and understanding these triples from spoken Chinese is crucial for a multitude of applications ranging from conversational agents to complex decision-support systems [15].

Our exploration covers several key areas that influence the performance and effectiveness of LLMs in the analytic task. First, we delve into the effects of prompt engineering [7], wondering how the design of prompts affects the output of LLMs, which is particularly important since subtle variations in prompts can lead to markedly different interpretations by the model. Next, we examine the input paradigm by employing zero-shot, one-shot, and few-shot settings [13] to determine the impact of example-based learning on the ability of LLMs to accurately infer semantic triples from novel spoken language inputs. We then consider the effects of the prompt language [18]. While our primary interest lies in Chinese spoken language, the choice between using Chinese and English prompts could potentially influence the performance of LLMs due to language-specific nuances

and training background. In addition, we also compare the performance of different LLMs [18] to evaluate which models excel in understanding and processing semantic triples in spoken Chinese.

2 Experiments

2.1 Prompt Engineering

Considering the fairness and effectiveness of the experiments, we use the OpenAI GPT-4 model as default except for otherwise specified, since GPT-4 is an established model and has been widely applied by researchers. Previous works [18] have shown that GPT-4 has extremely outstanding performance in multiple tasks compared with all the other models. We use English as the default prompting language, since GPT-4 has the best performance in English. With these fundamental settings, we expect to reveal the largest potential of LLMs in the SLU task.

Plain Input. Despite the strong dialogue ability of LLMs, they are in essence generative language models, which is originally trained for text completion. Hence, a natural idea is to feed the plain input to the model, expecting it to generate the correct answer, as is shown in Figure 1.

User Input

input sentence: 附近火锅店在什么位置
semantic triples:

Model Output

The input sentence "附近火锅店在什么位置" can be broken down into the following semantic triples:
(附近火锅店, 在, 什么位置)
Subject: 附近火锅店 (nearby hotpot restaurant)
Predicate: 在 (is located at)
Object: 什么位置 (what location)

Figure 1: The result of feeding plain input into the GPT-4 model in English.

Obviously, the model makes a mistake, confusing the concept of semantic triples with that in linguistics, in that we have never told the model explicitly. This is because semantic triples are more commonly used in linguistics, which indicates a higher prior probability [8]. This result shows the significance of the accurate problem description, which is the thing that good prompts should do.

Precise Description. Learning from the failure of plain input, we try to solve the problem using prompting. GPT-4 is famous for its effective finetuning to follow prompts, especially human instructions. As long as we can write a good prompt, the model is likely to give a satisfactory answer.

To form an effective prompt, we first describe the problem in a precise way, as is shown in Figure 2, but it is not concrete enough for our task. Hence, we add the range of action and slot to limit the answer space. The model also tends to extract the semantic triples into a single action. Therefore, we add a requirement that different semantic triples should be extracted separately.

User Input

Interpret the given sentence into semantic triples, namely the form of "<act>(<slot>=<value>)".
Note that <act> should be either "inform" or "deny", <slot> can be "操作", "对象", "起点", "终点", "途径", “位置", "方式", "偏好".
If there are multiple semantic triples, output them separately instead of including multiple slots in one action.
The given sentence is: 附近火锅店在什么位置

Model Output

inform(操作=查询)
inform(对象=火锅店)
inform(位置=附近)

Figure 2: The result of feeding precise description into the GPT-4 model in English.

We can see that the model gives a correct answer, which indicates the effectiveness of our prompt and the strong ability of GPT-4 to follow human instructions. Actually, other models cannot generate such satisfactory results, which will be discussed in the following experiments.

Chain-of-Thought. Proposed in [14], Chain-of-Thought (CoT) is a prompting trick to improve the performance of LLMs in the dialogue task. The idea is to let the model decompose the complex problem into several simple ones and then solve them sequentially. Previous works have shown that CoT can effectively improve the inference ability of LLMs.

There are two major categories of CoT tricks, namely few-shot CoT [14] and zero-shot CoT [5]. Few-shot CoT provides the model with a few examples of thinking steps, while zero-shot CoT adds a single requirement to the prompt. It is hard to form a valid few-shot CoT prompt in this task, so we utilize zero-shot CoT here to test its effectiveness, as is shown in Figure 3.

User Input

Interpret the given sentence into semantic triples, namely the form of "<act>(<slot>=<value>)". Note that <act> should be either "inform" or "deny", <slot> can be "操作", "对象", "起点", "终点", "途径", "位置", "方式", "偏好". If there are multiple semantic triples, output them separately instead of including multiple slots in one action.
The given sentence is: 附近火锅店在什么位置
Let's think step by step.

Model Output

The given sentence "附近火锅店在什么位置" translates to a request for information about the location of a nearby hotpot restaurant. Here are the semantic triples:
inform(对象=附近火锅店)
inform(位置=什么位置)

Figure 3: The result of utilizing chain-of-thought in the GPT-4 model in English.

Unluckily, the model fails to provide a good answer as expected. Compared to the prompt without CoT, the result is much more inaccurate, indicating that it fails to fully understand the sentence, which shows that CoT is not useful for all the tasks. This is not surprising because understanding spoken language does not require complex reasoning, not to mention analyzing the problem step by step. Therefore, using CoT is likely to make the model confused and thus leads to a worse result.

Emotional Stimulation. Latest research finds that emotional stimulation can improve the quality of the generated results by LLMs [6]. By letting the model believe they are human beings, the model can fit in the identity and show empathy, thus generating better results. Here we give a specified identity of servant to the model as an emotional stimulation, as is shown in Figure 4.

User Input

You are an experienced servant and never misunderstand the intention of guests.
Your task is to interpret the given sentence into semantic triples, namely the form of "<act>(<slot>=<value>)". Note that <act> should be either "inform" or "deny", <slot> can be "操作", "对象", "起点", "终点", "途径", "位置", "方式", "偏好". If there are multiple semantic triples, output them separately instead of including multiple slots in one action.
The given sentence is: 附近火锅店在什么位置

Model Output

inform(对象=火锅店)
inform(位置=附近)

Figure 4: The result of utilizing emotional stimulation in the GPT-4 model in English.

It is interesting that the model outputs a relatively accurate answer, much better than that of CoT. Such a result indicates that emotional stimulation is useful as a prompting trick. This may derive from the training data of LLMs, where the context is usually aligned with a specific identity.

Chinese Prompt. Another interesting problem is whether the model can perform well with Chinese prompts. We translate the prompt into Chinese and feed it into the model, as is shown in Figure 5.

User Input

将指定句子解析为语义三元组，即“<act>(<slot>=<value>)”的形式，其中<act>只能为“inform”或“deny”，<slot>可以是“操作”、“对象”、“起点”、“终点”、“途径”、“位置”、“方式”、“偏好”。如果句子包含多个语义三元组，请将它们分别输出，不要在一个<act>中包含多个<slot>。
句子：附近火锅店在什么位置

Model Output

inform(对象=火锅店)
inform(位置=附近)

Figure 5: The result of feeding Chinese prompt into the GPT-4 model in Chinese.

We can see that the model correctly understands the prompt and gives a satisfactory answer. Although the answer is not as perfect as that using English prompts, it is quite acceptable, which demonstrates the strong multilingual ability of LLMs. With tremendous training data, LLMs can learn and understand multiple languages without any complicated design.

2.2 Learning from Samples

User Input

Interpret the given sentence into semantic triples, namely the form of "<act>(<slot>=<value>)". Note that <act> should be either "inform" or "deny", <slot> can be "操作", "对象", "起点", "终点", "途径", “位置", "方式", "偏好". If there are multiple semantic triples, output them separately instead of including multiple slots in one action.
The given sentence is: 我要去娄桥高速收费站

Model Output

inform(操作=去)
inform(对象=娄桥高速收费站)

User Input

Interpret the given sentence into semantic triples, namely the form of "<act>(<slot>=<value>)". Note that <act> should be either "inform" or "deny", <slot> can be "操作", "对象", "起点", "终点", "途径", “位置", "方式", "偏好". If there are multiple semantic triples, output them separately instead of including multiple slots in one action.
The given sentence is: 导航到安德第六地税所拍照
The semantic triples are: inform(操作=导航), inform(终点=安德第六地税所)
The given sentence is: 我要去娄桥高速收费站

Model Output

The semantic triples are: inform(操作=去), inform(终点=娄桥高速收费站)

User Input

Interpret the given sentence into semantic triples, namely the form of "<act>(<slot>=<value>)". Note that <act> should be either "inform" or "deny", <slot> can be "操作", "对象", "起点", "终点", "途径", “位置", "方式", "偏好". If there are multiple semantic triples, output them separately instead of including multiple slots in one action.
The given sentence is: 导航到安德第六地税所拍照
The semantic triples are: inform(操作=导航), inform(终点=安德第六地税所)
The given sentence is: 我要走宁夏回族自治区党委大院
The semantic triples are: inform(操作=导航), inform(途径=宁夏回族自治区党委大院)
The given sentence is: 附近火锅店在什么位置
The semantic triples are: inform(对象=火锅店), inform(位置=附近)
The given sentence is: 我要去娄桥高速收费站

Model Output

The semantic triples are: inform(操作=导航), inform(终点=娄桥高速收费站)

Figure 6: From top to bottom are respectively the results of utilizing zero-shot, one-shot and few-shot paradigm to form prompts in the GPT-4 model in English.

Learning from samples is an important ability of deep learning models [13], especially for LLMs. Previous works have proved that providing the model with a few samples can bring much higher generation quality. Therefore, we form three prompts in the zero-shot, one-shot and few-shot paradigms respectively to explore whether LLMs can learn from samples in slot filling tasks.

We use a different sentence to raise the difficulty. The results are shown in Figure 6. We can see that all the prompts lead to the correct targets, which demonstrates the deep understanding with named entities of LLMs. However, more samples promise more accurate operations, since the samples help the model realize that it is probability a navigation related task. Although one-shot prompt does not help generate the correct answer, it successfully guides the model to output in the given format. Such a result proves the strong ability of LLMs to follow human instructions and learn from samples.

2.3 Cross-model Comparison

User Input

Interpret the given sentence into semantic triples, namely the form of "<act>(<slot>=<value>)". Note that <act> should be either "inform" or "deny", <slot> can be "操作", "对象", "起点", "终点", "途径", "位置", "方式", "偏好". If there are multiple semantic triples, output them separately instead of including multiple slots in one action.
The given sentence is: 附近火锅店在什么位置

ChatGPT Output

The given sentence "附近火锅店在什么位置" can be interpreted into the following semantic triples:
inform(对象=火锅店, 位置=附近)
inform(对象=火锅店, 位置=什么位置)

Claude Output

inform(位置=附近)

LLaMA Output

Sure, I'd be happy to help! Here are the semantic triples for the given sentence:
inform(位置="附近")
inform(对象="火锅店")
So, the sentence can be interpreted as "Inform me about the location of a hot pot restaurant nearby."

ChatGLM Output

The sentence "附近火锅店在什么位置" can be interpreted as a query asking for the location of a nearby hotpot restaurant. In the form of semantic triples, this would be:
inform(位置=附近火锅店的位置)

Ernie Output

inform(位置=附近火锅店的位置)

Qwen Output

inform(位置=附近)
inform(对象=火锅店)

Figure 7: The result of feeding the same prompt into different models in English. ChatGPT and Claude completely fail to solve the problem. ChatGLM and Ernie provide similar but inaccurate results. LLaMA and Qwen give best results close to that of GPT-4.

As OpenAI embarks on the wave of LLMs, more companies invest in the research of LLMs and more models emerge in the market. It is meaningful to compare the performance of different models for the slot filling task. We compare multiple LLMs with the same prompt in English, including OpenAI ChatGPT [9], Anthropic Claude [2], Meta LLaMA [12], Tsinghua ChatGLM [17], Baidu Ernie [11] and Alibaba Qwen [1]. The results given by different models are shown in Figure 7.

We can see that ChatGPT fails to follow the instructions and outputs multiple slots in a single action. Claude also fails to capture the key information in the given sentence. ChatGLM and Ernie present the key information, but the results are not accurate enough. LLaMA and Qwen show the best performance, which provides the correct answer close to that given by GPT-4. This is an exciting

result because LLaMA and Qwen are both open-source models that are available to the public. Their outstanding performance provides an enormous value for the research community. Companies and institutes can apply them in their own downstream tasks, while researchers can use them as reliable baselines to compare with their own models.

3 Discussion

The advent of LLMs has revolutionized the field of SLU by providing systems with a profound capability to interpret human language with greater context sensitivity and accuracy. Our exploration reveals that the dominant tasks within SLU, namely intent detection and slot filling, can now be addressed with relative ease through LLM inference mechanisms. These tasks, essential for finding the underlying meaning and requisite actions from spoken input, are handled more effectively by LLMs due to their extensive linguistic databases and sophisticated predictive algorithms. As a result, LLM-based methods not only simplify the process of understanding spoken language but also outperform traditional paradigms that have been employed in the past.

Despite these advancements, traditional methods continue to maintain their significance by offering advantages in terms of speed and computational costs. They are found to be particularly suitable for scenarios where resources are constrained, and the requirement for computation efficiency and economic feasibility is of utmost importance. These lightweight scenarios speak to a significant subset of real-world applications where the balance between performance and resource consumption must be carefully managed.

Looking towards the future, there are several promising directions for SLU research. One of these angles involves enhancing the performance of SLU systems on specific tasks through supervised finetuning (SFT) [9] on high-quality, domain-specific datasets. SFT offers a pathway to giving LLMs with an even greater degree of precision and adaptability for specialized applications.

Another emergent research vector is the concept of reconstructing spoken dialogue systems by integrating speech directly as a modality within Multi-modal Large Language Models (MLLMs) [16]. This approach could potentially bypass the need for a separate SLU module entirely, leading to more seamless and integrated processing of spoken dialogue.

Additionally, the application of knowledge distillation presents an opportunity to distill the capabilities of LLMs into more compact and efficient models suitable for deployment on terminal units with limited resources. Through this technique, the wealth of knowledge encapsulated in LLMs can be transferred into lightweight SLU models that can operate on edge devices without the need for constant connectivity to powerful cloud-based systems.

To conclude, the impact of LLMs on the landscape of SLU is profound and far-reaching, reshaping the way spoken language is processed and understood. Amid this transformative period, it is paramount for researchers to engage with these challenges and opportunities, striving to optimize performance and democratize the benefits of SLU technologies across diverse applications.

4 Conclusion

In this project, we explore the application of LLMs to the slot filling task in SLU, which interprets spoken language into semantic triples. We conduct adequate experiments to investigate the effects of prompt engineering, where precise description, chain-of-thought, emotional stimulation and prompt language are considered. We find that precise description and emotional stimulation can significantly improve the performance of LLMs in the slot filling task, while chain-of-thought is not that useful. We also explore the effects of learning from samples, where zero-shot, one-shot and few-shot paradigms are tried to form prompts, where we find that LLMs can benefit from learning from samples, especially in the few-shot paradigm. In addition, we compare the performance of different LLMs currently in the market. We find that GPT-4 has the leading performance in the slot filling task, while LLaMA and Qwen can also achieve satisfactory results as open-source models, which is promising for future research. Besides the above experiments, we also give our analysis and discussion of the research status of SLU in the era of LLMs. We believe that LLMs can bring a revolution to SLU and provide more research directions, while traditional methods are still useful in some specific scenarios.

References

[1] Jinze Bai, Shuai Bai, Yunfei Chu, Zeyu Cui, Kai Dang, Xiaodong Deng, Yang Fan, Wenbin Ge, Yu Han, Fei Huang, et al. Qwen technical report. *arXiv preprint arXiv:2309.16609*, 2023.

[2] Yuntao Bai, Andy Jones, Kamal Ndousse, Amanda Askell, Anna Chen, Nova DasSarma, Dawn Drain, Stanislav Fort, Deep Ganguli, Tom Henighan, et al. Training a helpful and harmless assistant with reinforcement learning from human feedback. *arXiv preprint arXiv:2204.05862*, 2022.

[3] KR1442 Chowdhary and KR Chowdhary. Natural language processing. *Fundamentals of artificial intelligence*, pages 603–649, 2020.

[4] Kunho Kim, Rahul Jha, Kyle Williams, Alex Marin, and Imed Zitouni. Slot tagging for task oriented spoken language understanding in human-to-human conversation scenarios. In *Proceedings of the 23rd Conference on Computational Natural Language Learning (CoNLL)*, pages 757–767, 2019.

[5] Takeshi Kojima, Shixiang Shane Gu, Machel Reid, Yutaka Matsuo, and Yusuke Iwasawa. Large language models are zero-shot reasoners. *Advances in neural information processing systems*, 35:22199–22213, 2022.

[6] Cheng Li, Jindong Wang, Yixuan Zhang, Kaijie Zhu, Wenxin Hou, Jianxun Lian, Fang Luo, Qiang Yang, and Xing Xie. Large language models understand and can be enhanced by emotional stimuli. *arXiv preprint arXiv:2307.11760*, 2023.

[7] Pengfei Liu, Weizhe Yuan, Jinlan Fu, Zhengbao Jiang, Hiroaki Hayashi, and Graham Neubig. Pre-train, prompt, and predict: A systematic survey of prompting methods in natural language processing. *ACM Computing Surveys*, 55(9):1–35, 2023.

[8] Carl N Morris. Parametric empirical bayes inference: theory and applications. *Journal of the American statistical Association*, 78(381):47–55, 1983.

[9] Long Ouyang, Jeffrey Wu, Xu Jiang, Diogo Almeida, Carroll Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, et al. Training language models to follow instructions with human feedback. *Advances in Neural Information Processing Systems*, 35:27730–27744, 2022.

[10] Libo Qin, Tianbao Xie, Wanxiang Che, and Ting Liu. A survey on spoken language understanding: Recent advances and new frontiers. *arXiv preprint arXiv:2103.03095*, 2021.

[11] Yu Sun, Shuohuan Wang, Shikun Feng, Siyu Ding, Chao Pang, Junyuan Shang, Jiaxiang Liu, Xuyi Chen, Yanbin Zhao, Yuxiang Lu, et al. Ernie 3.0: Large-scale knowledge enhanced pre-training for language understanding and generation. *arXiv preprint arXiv:2107.02137*, 2021.

[12] Hugo Touvron, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Timothée Lacroix, Baptiste Rozière, Naman Goyal, Eric Hambro, Faisal Azhar, et al. Llama: Open and efficient foundation language models. *arXiv preprint arXiv:2302.13971*, 2023.

[13] Yaqing Wang, Quanming Yao, James T Kwok, and Lionel M Ni. Generalizing from a few examples: A survey on few-shot learning. *ACM computing surveys (csur)*, 53(3):1–34, 2020.

[14] Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Fei Xia, Ed Chi, Quoc V Le, Denny Zhou, et al. Chain-of-thought prompting elicits reasoning in large language models. *Advances in Neural Information Processing Systems*, 35:24824–24837, 2022.

[15] Tsung-Hsien Wen, David Vandyke, Nikola Mrksic, Milica Gasic, Lina M Rojas-Barahona, Pei-Hao Su, Stefan Ultes, and Steve Young. A network-based end-to-end trainable task-oriented dialogue system. *arXiv preprint arXiv:1604.04562*, 2016.

[16] Shukang Yin, Chaoyou Fu, Sirui Zhao, Ke Li, Xing Sun, Tong Xu, and Enhong Chen. A survey on multimodal large language models. *arXiv preprint arXiv:2306.13549*, 2023.

[17] Aohan Zeng, Xiao Liu, Zhengxiao Du, Zihan Wang, Hanyu Lai, Ming Ding, Zhuoyi Yang, Yifan Xu, Wendi Zheng, Xiao Xia, et al. Glm-130b: An open bilingual pre-trained model. *arXiv preprint arXiv:2210.02414*, 2022.

[18] Wayne Xin Zhao, Kun Zhou, Junyi Li, Tianyi Tang, Xiaolei Wang, Yupeng Hou, Yingqian Min, Beichen Zhang, Junjie Zhang, Zican Dong, et al. A survey of large language models. *arXiv preprint arXiv:2303.18223*, 2023.