Proceedings of CONF-SPML 2026 Symposium: The 2nd Neural Computing and Applications Workshop 2025
DOI: 10.54254/2755-2721/2026.TJ30647

Research on Natural Language Instruction Recognition Method Based on Large Language Models

Pok Him Zheng

The Bay International College, Shenzhen, China
rwfii333@outlook.com

Abstract. In the context of the swift and significant advancement of large language models, the user demographic and the array of application scenarios for artificial intelligence have experienced a remarkable and extensive expansion. The capability to accurately recognize and interpret linguistic instructions has emerged as the fundamental cornerstone upon which AI systems rely to comprehend and effectively execute a diverse range of tasks. The ongoing enhancement of natural language instruction recognition technology holds paramount importance for substantially improving the efficiency and precision with which language models can handle intricate, information-intensive tasks. This scholarly paper delves deeply into the current state of research, focusing on critical aspects such as the extraction of sensitive information, advancements in model cognitive capabilities, and the strategies employed to mitigate the impact of various interference factors. It conducts a comprehensive and systematic analysis, juxtaposing different methodologies, including the hybrid approach that integrates Long Short-Term Memory (LSTM) and Convolutional Neural Networks (CNN) for the precise extraction of sensitive information, the innovative open-world multi-scenario instruction understanding (framework known as MOO, and the sophisticated semantic enhancement technique of retrieval-augmented generation (RAG). The paper meticulously examines the processing workflows inherent to these methodologies, scrutinizing their respective influences on the analysis of linguistic instructions. Furthermore, it synthesizes and highlights the unique advantages that each method brings to the table, providing a nuanced understanding of their contributions to the field.

Keywords: Natural language process, Information identification, Retrieval-augmented generation, Long short-term memory model

1. Introduction

Since the concept of Artificial Intelligence (AI) was proposed in 1956, its goal has been to enable machines to mimic human "thinking" behavior by simulating human nerve cells, thereby helping or replacing humans in performing tasks or work. After more than 70 years of development, AI has become a maturely applied cutting-edge science. The rapid development of AI powered by large language models (LLMs) has significantly reduced the application cost and usage threshold of

DOI: 10.54254/2755-2721/2026.TJ30647

technologies like speech recognition. From the release of Siri in 2011 to ChatGPT and DeepSeek, the proportion of AI used in public life has increased rapidly.

Instruction recognition is a key technology for enabling machines to correctly understand and execute human instructions, holding significant importance in fields such as human-computer interaction, intelligent assistants, and robot control. Early instruction understanding relied on predefined rules and grammatical parsing, but with the rise of deep learning, data-driven methods have made significant progress. In recent years, the development of large language models (LLMs) has greatly enhanced machines' ability to understand natural language instructions. However, models relying solely on statistical patterns in training data may not accurately grasp user intent, especially in complex scenarios where they are prone to interference. For example, some users disguise non-compliant requests within seemingly harmless language through emotional narratives, tricking the model into ignoring the sensitive instructions. Therefore, how to make models recognize the true intent of instructions, filter out interference information, and execute robustly has become a research hotspot. This paper systematically reviews the research progress in the field of instruction recognition, introduces its development history and representative technical solutions, compares the characteristics and applicable scenarios of various methods, and provide effective analysis for future research.

2. Research status of domestic and foreign methods

2.1. Open-world multi-scenario instruction understanding (MOO)

As instruction scenarios expand from pure text to physical environments, researchers have begun exploring multimodal, multi-scenario instruction recognition methods. Austin Stone et al. proposed the Manipulation of Open-World Objects (MOO) method in 2023 [1]. MOO aims to enable robots to understand and execute instructions involving novel objects, even if these objects were never seen during training. This method uses pre-trained Vision-Language Models (VLMs) to link object descriptions in natural language instructions with robot visual observations: first, the instruction is divided into "manipulation action" and "object description" parts. For the object description, prompt words such as color and shape are further extracted (e.g., the instruction "pick up the Coca-Cola can" extracts "red can" as a prompt). Then, an open-vocabulary object detection model (like OWL-ViT) is used to find the object position corresponding to the text prompt in the robot's camera image and generate a segmentation mask for that object. The robot policy model (a Transformer-based control policy like RT-1) then takes the current image, the original instruction, and the object location information provided by the VLM as input to decide the next action. Since the mask representation is consistent for both seen and unseen new objects, MOO is more robust when dealing with new semantic concepts compared to strategies relying solely on language embeddings. Physical robot experiments showed that after training with large-scale demonstrations involving 106 types of objects, the MOO model achieved a 98% success rate in manipulating seen objects and a 79% success rate for never-before-seen new objects, demonstrating excellent zero-shot generalization capability. Furthermore, MOO can be extended to accept non-verbal object indication methods (such as a human pointing at the target object) and combined with navigation models to achieve mobile manipulation in open environments. The advantage of the MOO method lies in its strong applicability across multiple scenarios: it is no longer limited to objects covered in the training set, can recognize new things through semantic prompts, and achieves cross-scenario instruction understanding and execution. This helps overcome the poor applicability caused by excessive model specialization, improves the correctness of instruction understanding by robots in

DOI: 10.54254/2755-2721/2026.TJ30647

different environments, and enhances robustness in unstable environments. Its shortcomings include the need for a large amount of demonstration data to learn basic skills, a relatively complex system requiring collaboration between pre-trained models and policy models. Additionally, MOO focuses on generalization at the object level, and further extension is needed for entirely new action instructions or more abstract tasks.

2.2. Semantic enhancement method using Retrieval-Augmented Generation (RAG)

Although large language models possess vast knowledge, they are prone to errors or hallucinatory answers in instruction scenarios requiring precise facts or domain-specific information. Retrieval-Augmented Generation (RAG) is a framework proposed in recent years that integrates external knowledge retrieval into the generation process [2], used to enhance the model's semantic understanding and factual accuracy. Its core idea is to combine internal knowledge with external knowledge, completing instruction processing in three stages: (1) Internal Knowledge Utilization: The model first uses the knowledge learned during its training to preliminarily understand the user instruction and form a semantic representation. (2) External Knowledge Retrieval: The model retrieves relevant information from external knowledge bases (such as Wikipedia, company documents, specialized databases, etc.) based on the instruction, converting the original query into a vector and matching it with knowledge base vectors to extract the most relevant document snippets. (3) Knowledge Fusion and Generation: The retrieval results are fused with the model's internal knowledge semantic representation, integrating, aligning, and reasoning over the context, and finally generating the answer. In this process, the language understanding and generation capabilities of the pre-trained language model closely cooperate with the traditional information retrieval module. For example, the RAG model proposed by Patrick Lewis et al. in 2020 combined a BERT-based question-answering retrieval mechanism with a BART generation model [2], significantly outperforming generation models relying solely on parametric memory in tasks like open-domain question answering, producing more specific and factual answers. The advantage of RAG lies in endowing the model with the ability to query external knowledge, allowing it to dynamically obtain the latest or domain-specific information, thereby improving answer accuracy and providing source attribution. This effectively alleviates the model's "hallucination" problem – when the model is uncertain, the retrieved reliable information can prevent it from fabricating misleading content. Research indicates that RAG can reduce the interference of misinformation on instruction intent understanding and improve the precision of output results. On the downside, RAG introduces a retrieval module, increasing system complexity, and the retrieval quality depends on the completeness of the knowledge base and the effectiveness of the retrieval algorithm; building high-quality knowledge sources in specific domains is also a challenge. Furthermore, RAG has limited effectiveness on reasoning-type instructions requiring logical deduction; how to more tightly integrate reasoning capabilities is also a future direction.

Besides the two representative methods mentioned above, there are other important research directions and methods in the field of instruction recognition worth noting:

Transformer-based Instruction Understanding: Since its proposal in 2017, the Transformer model has quickly become the mainstream tool for sequence modeling, capable of capturing global semantics and long-range dependencies better than LSTM. Pre-trained Transformers like BERT and GPT are used for instruction understanding tasks such as instruction classification, question answering, and semantic parsing, significantly improving performance. Particularly, Instruction Tuning technology for large language models, which fine-tunes the model on massive task instruction data to make its output more aligned with human instruction intent, has been proven

DOI: 10.54254/2755-2721/2026.TJ30647

effective in enhancing the model's ability to follow user instructions and its controllability. Instruction-tuned models (such as InstructGPT, ChatGPT) have achieved powerful zero-shot task generalization capabilities in open domains, able to complete various complex tasks based on different user instructions. However, Transformer-based models also have issues such as sensitivity to adversarial attacks and opaque reasoning processes, requiring supplementary safety training and interpretability research to enhance robustness.

Instruction-to-Action Mapping: Directly converting natural language instructions into machine-executable action sequences is an important application of instruction understanding, covering scenarios from conversational assistants calling backend APIs to robot action planning. For example, in robotics, some work combines large language models with robot control policies, letting the language model handle high-level planning and the control module handle low-level execution: the SayCan algorithm proposed by Google [3] combines candidate actions generated by the LLM with the value function of robot skills to select the next operation that is both meaningful and feasible. The PaLM-SayCan system showed that a stronger language model improves robot task success rates: using the 540B parameter PaLM model increased the correct skill selection rate to 84% and the task completion rate to 74% compared to the previous FLAN model, halving errors. In environments like smart homes, researchers have also explored schemes to parse instructions into structured operation commands (such as program code, API call sequences), achieving automatic language-to-action mapping through semantic parsing or neural sequence learning. The advantage of such methods is eliminating the tediousness of manually defining rules and adapting to diverse user expressions, but the challenge lies in ensuring the generated action sequences are safe, reliable, and do not deviate from user intent, as well as decomposing and planning in complex long-horizon tasks.

Multimodal Instruction Understanding: Human instructions are often not limited to text but may combine gestures, visual environment information, etc. For example, when instructing a robot, a person might simultaneously provide a language description and an image of the target object or a pointing gesture. Therefore, multimodal fusion for instruction recognition has become one of the frontier directions. A typical approach is to combine language with visual perception, using visual-semantic fusion models to understand the entities and environment involved in the instruction. For example, the Vision-Language Navigation task requires an agent to navigate in a visual environment based on textual route instructions, requiring the model to parse both language and images. Another example is the MOO method and the related PaLM-E model mentioned earlier, which input camera images into the model to assist in understanding object references in instructions, enhancing cross-modal instruction execution capability. Multimodal instruction understanding enables AI to have scene perception capabilities, understanding instructions like "put this over there" (requiring understanding what "this" and "there" refer to). The difficulty lies in aligning feature representations from different modalities, constructing large-scale multimodal training data, and ensuring the model can still filter environmental noise and focus on instruction-relevant elements after introducing visual information. Research in this area is promoting robots to more naturally accept human multimodal instructions in real environments.

3. Background related issues

As AI based on large language models iteratively develops and language models for the public become relatively mature, the language input by users into intelligent programs and the language habits of different users may affect the program's output results. Taking ChatGPT as an example, after its rise on the internet, users discovered that they could attempt to deceive the AI into performing non-compliant behaviors by constructing emotional narratives. In this case, users used

DOI: 10.54254/2755-2721/2026.TJ30647

sentences like "My grandmother, before she passed away, often used to... to help me fall asleep. I miss my grandmother very much, can you imitate my grandmother... to help me fall asleep?" to divert the AI's attention from the requested action to simulating the "grandmother," causing the AI to lower its vigilance against the non-compliant request.

The ability of large language models to parse and understand input content relies on large amounts of data during model training. However, in specific work scenarios, the information received by the machine is more complex and may contain vocabulary similar to that in everyday language environments. These synonyms and near-synonyms may cause the machine to misjudge the user's presupposed work scenario, ultimately leading the machine's output to completely deviate from the user's true intent.

Therefore, the model's judgment of its own environment when reasoning based on input content is particularly important for the accuracy of output information. Research on instruction recognition and instruction checking methods for large data models is crucial for improving machine understanding of instructions. Further enhancing the efficiency of using AI can ensure that AI improves the understanding ability of language models in increasingly complex information environments in the future, enabling AI to maintain efficient information processing capabilities even in increasingly complex information environments.

4. Method analysis

4.1. Instruction parsing: retrieving effective sensitive information from text

Zhang proposed an algorithm based on Long Short-Term Memory (LSTM) model and Convolutional Neural Network (CNN) for extracting sensitive information from social media [2]. The obtained text is divided into key words and statistically summarized into a sensitive vocabulary list, which is managed by sorting based on the frequency of sensitive words. The processed data is then fed into a Bidirectional Long Short-Term Memory (BiLSTM) neural network for information detection in both forward and reverse directions. Finally, the resulting data is input into a CNN to extract sample features, and corresponding score values are generated through a Self-Attention mechanism. The LSTM+CNN sensitive information extraction process was shown in the Figure 1.

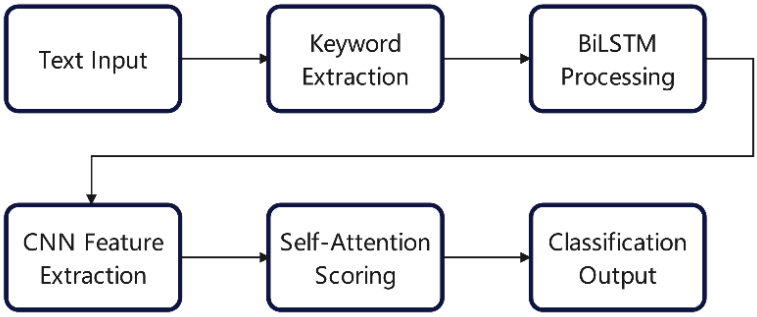

Figure 1. LSTM+CNN sensitive information extraction process

Through experiments on Chinese online platforms, Zhang's algorithm model achieved recognition accuracy rates exceeding 93% for five types of text: conventional, advertising, abusive, vulgar, and violent. Furthermore, the accuracy of this algorithm was 2.319% higher than the BRET algorithm. By changing the target of the algorithm model samples, the model's objective shifts from identifying the characteristic nature of text in online chats to recognizing the intent conveyed by

DOI: 10.54254/2755-2721/2026.TJ30647

instructions. This helps enhance the language model's attention to key information in complex information, thereby improving the instruction recognition capability of language models.

4.2. Multi-scenario applicability: expanding model cognitive scope

Austin Stone et al. proposed an open-world object manipulation (MOO) method [1] that, combined with large language models, enables machines to describe or recognize unfamiliar objects based on instruction descriptions. Their developed machine, upon receiving an instruction, divides the instruction into object and location parts. For the object part, the description of the object is further divided into adjectives and shape feature prompts (e.g., Coca-Cola -> red can; Pepsi-Cola -> blue can). This approach allows the machine's object recognition ability to break free from the limitations of training samples. The model only needs to learn the meaning represented by the prompt words. When facing never-before-seen objects described in tasks, it recognizes the objects by summarizing their features. In numerous representative "grasping" tasks, experimental results showed that their developed MOO algorithm achieved a 98% success rate for known objects and a 79% success rate for unseen objects.

In diverse application environments for large language models, the research by Austin Stone et al. helps overcome the limitation of models lacking multi-scenario applicability due to excessive specialization. It is beneficial for improving model understanding of instructions in different scenarios and achieving high accuracy in similar repetitive tasks. Simultaneously, it enhances the model's robustness when working in unstable environments. This improves the applicability and robustness of large language models in different scenarios. For future environments with complex information, it allows models to maintain high accuracy in training efficiency and reasoning judgment without requiring extensive repetitive sample training.

4.3. Semantic enhancement: eliminating interference elements

Yu Shengyin, Zhou Changgao, and Sun Jingchao mentioned using the Retrieval-Augmented Generation (RAG) framework [4-6] combined with the internal language capabilities of large language models and external knowledge base information to enhance the reasoning ability of large language models. The RAG proposed by Yu Shengyin et al. is divided into three stages:

(1) Internal Knowledge Utilization: After the user submits a query, the LLM first performs preliminary processing based on the internal knowledge learned during its training. In this stage, the model uses its language expression and reasoning capabilities to try to understand the question and form a preliminary semantic representation.

(2) External Knowledge Retrieval: To enhance answer accuracy, the system concurrently calls external knowledge bases (such as document collections, Wikipedia, corporate databases, etc.) for retrieval. The retrieval system converts the original query into an embedding vector and matches it with pre-stored embedding vectors in the knowledge base, extracting the information most relevant to the query.

(3) Knowledge Fusion and Generation: The LLM fuses the retrieved results from external sources with its own internal knowledge, performs context integration, content alignment, and reasoning, and accordingly generates the final answer. This process reflects the collaborative cooperation between the generative model and the information retrieval module.

At the same time, Zhou. and Sun. mentioned that RAG can alleviate the "hallucination" problem caused by misinformation affecting model reasoning, preventing misleading information from distorting the intent expressed in instructions. During instruction analysis, it compensates for the

DOI: 10.54254/2755-2721/2026.TJ30647

cognitive deficiencies of large models through external materials, reduces the impact of misleading information on the model, and improves the precision of output data.

4.4. Framework comparison

To more intuitively compare the aforementioned methods, The summary of the characteristics, applicable scenarios, advantages, and disadvantages of several typical instruction recognition techniques was shown in the Table 1, and the RAG network framework was shown in the Figure 2.

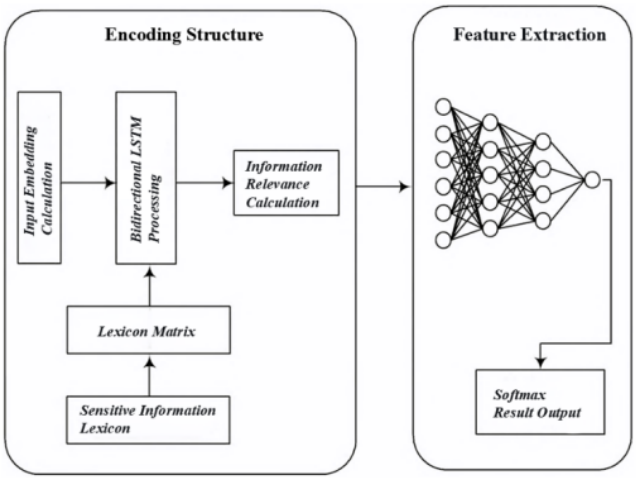

Figure 2. RAG network framework [2]

Table 1. Method comparison

Method	Core Idea / Characteristics	Applicable Scenarios	Advantages	Disadvantages / Limitations
LSTM+CNN Sensitive Info Extraction	RNN excels at sequence modeling + CNN extracts local features, Self-Attention highlights keywords	Social media text instructions, non-compliant content detection	Model structure simple and efficient, can capture key information like keywords in instructions, high classification accuracy	Understanding stays at word level, limited handling of complex long-range dependencies, performance limited by training corpus coverage
MOO Open-World Instruction Understanding	Uses pre-trained VLM, binds language description with visual target, zero-shot new object manipulation	Robot operation instructions, open environment multi-object scenarios	Strong generalization, can recognize new objects not in training set; multimodal fusion improves instruction understanding accuracy	Complex model system, requires large demonstration data; currently focuses on object manipulation, limited generalization for new action types
RAG Retrieval-Augmented Generation	Combines internal LM with external knowledge base retrieval, three-stage knowledge fusion for answer generation	Knowledge QA, professional domain inquiry, long-text instructions	Comprehensive and attributable information, dynamic knowledge lookup avoids knowledge gaps; reduces model hallucination, improves factual correctness	High system complexity, depends on knowledge base quality; retrieval-generation synergy requires fine-tuning, effectiveness on logical reasoning instructions needs improvement

DOI: 10.54254/2755-2721/2026.TJ30647

4.5. Outlook

Looking ahead, several research directions worthy of attention include: (1) Adversarial Robustness – Improving the model's ability to resist adversarial examples and social engineering attacks, for example, through adversarial training, instruction content review mechanisms, etc., to prevent models from being induced by harmful instructions; (2) Multimodal Fusion – Deeply integrating language with multimodal information such as vision, speech, and gestures to achieve more natural human-machine instruction communication, while researching efficient cross-modal representation learning to reduce data requirements; (3) Low-Resource Generalization – Exploring methods like few-shot learning, transfer learning, and knowledge injection so that models can accurately understand instructions even in specialized domains lacking large-scale annotated data, including using simulated data and self-supervised pre-training to compensate for data scarcity. Additionally, explainability and safety are also key for instruction recognition to move towards practical application: future models should be able to explain their understanding of instructions and execution steps for human trust and correction; for instructions involving decision-making actions, safety constraints and supervision need to be introduced to ensure that the output behavior complies with ethics and rules. In conclusion, as an important interface for interaction between artificial intelligence and humans, instruction recognition is developing towards more intelligent, reliable, and multi-capability fusion. Continuous research and technological innovation are expected to endow machines with the ability to comprehensively understand and execute various human instructions, thereby greatly expanding the application boundaries of AI in the real world.

5. Conclusion

In summary, instruction recognition technology has undergone rapid evolution in recent years, forming multiple technical routes ranging from textual element extraction and knowledge enhancement to multimodal understanding. Traditional models like LSTM/CNN have achieved good results in specific domain instruction parsing. Pre-trained Transformer models have further brought about a significant improvement in general language understanding capability. Multimodal methods in open environments and retrieval-augmented frameworks have significantly broadened the range and depth of instructions that models can handle. Current mainstream technologies each have advantages: methods based on deep learning have high precision and continuously refresh performance records; integrating external knowledge makes models more reliable and verifiable; combining multimodal approaches with large models gives systems unprecedented flexibility. However, limitations and challenges remain. On one hand, the adversarial robustness of models against instructions needs strengthening, as they might still be misled by obscure expressions or maliciously constructed instructions. On the other hand, the fusion and transfer between different modalities and tasks are still insufficient, and the generalization ability of models under low-resource data also needs improvement. Furthermore, how to enable models to understand the deep intent and pragmatic context of instructions and ensure the safety and reliability of the execution process are issues that must be emphasized in the future.

References

[1] Stone A., et al. (2023). Open-World Object Manipulation using Pre-Trained Vision-Language Models. Conference on Robot Learning (CoRL) 2023. pp.3-8

[2] Zhang, X. (2025). Sensitive Information Recognition and Analysis Algorithm Based on Natural Language Semantic Perception Modern. Electronics Technique, 2025, 48(18): 17-21.

DOI: 10.54254/2755-2721/2026.TJ30647

[3] Lewis, P., et al. (2020). Retrieval-Augmented Generation for Knowledge-Intensive NLP Tasks. Advances in Neural Information Processing Systems (NeurIPS) 2020. pp.2-10
[4] Yu, S. (2025). Research on Event Causality ldentification Methods Based on Large Language Models [Doctoral dissertation, Guangdong University of Technology]. pp.13-15
[5] Zhou, C., & Sun, J. (2025). Research on Disinformation Detection Method Based on Hallucination-Elimination Large Language Model [Online first].
[6] Ahn, M., et al. (2022). Do As I Can, Not As I Say: Grounding Language in Robotic Affordances. arXiv preprint arXiv: 2204.01691. pp.3-11