Proceedings of CONF-MLA 2025 Symposium: Intelligent Systems and Automation: AI Models, IoT, and Robotic Algorithms
DOI: 10.54254/2755-2721/2025.LD27506

A Review on Robot Dialogue Target Identification and Task-Oriented Interactive Systems in Linguistic Environments

Zixi Fu[1*], Xiaoxin Tian[2]

[1]*Arizona Institute of Technology, Hebei University of Technology, Tianjin, China*
[2]*School of Mechanical Engineering, Wuhan University of Science and Technology, Wuhan, China*
**Corresponding Author. Email: 2513031311@qq.com*

Abstract. The evolution of human-computer interaction into open-domain and multimodal environments has introduced significant challenges for dialogue systems, including semantic ambiguity in complex linguistic contexts, cultural disparities, and the need to discern implicit user intent. Addressing these issues, this study proposes a novel adaptive dialogue framework that synergizes symbolic reasoning with embodied environmental perception, focusing on three core technological advancements: multimodal semantic alignment, dynamic contextual reasoning, and elastic system architecture. Experimental validations demonstrate that transformer-based joint representations (e.g., ViLBERT), when enhanced by contrastive learning strategies such as the ITM loss function, achieve an F1-score of 91.2% in intent recognition. Meanwhile, dynamic attention mechanisms (e.g., the Flamingo model) leverage interleaved encoding and gated interactions to attain 78.6% accuracy in temporally dependent tasks. The elastic architecture, designed to integrate real-time environmental awareness with scenario-specific comprehension, enables autonomous resolution of 40% of border customer service inquiries and surpasses 90% accuracy in emergency command recognition. Further advancements in model compression (e.g., ADTDNN) and few-shot learning techniques facilitate dialect recognition across 12 low-resource languages and lightweight deployment on edge devices, underscoring the potential for inclusive technological accessibility. Persistent challenges include metaphor interpretation, long-form coherence in text generation, and ethical concerns such as data privacy breaches and algorithmic bias, alongside societal risks like the widening digital divide. Future progress necessitates interdisciplinary collaboration informed by cognitive science to decode human multimodal processing, coupled with educational reforms that bridge technical expertise ("technology stacks"), scenario-driven applications ("scenario libraries"), and toolchain integration. Projections indicate the global dialogue system market will reach $80 billion by 2027, catalyzing an "AI-powered language services" paradigm that reshapes industries from healthcare to cross-border commerce.

Keywords: robot dialogue, target identification, oriented interactive systems, linguistic robustness, dynamic context reasoning

DOI: 10.54254/2755-2721/2025.LD27506

1. Introduction

As computer interaction enters the open domain, complex linguistic environments cause semantic deviation and cultural understanding problems in dialogue systems. Global border commerce handles millions of multilingual consultation tasks daily, and an aging society requires computer-based interaction. However, current systems can't resolve dialect expression fragmentation and semantic ambiguity. Additionally, there are unresolved research directions such as symbolic reasoning for discourse intention and adaptive, flexible cognitive elasticity. In response, this study comprehensively considers these factors. It analyzes dialogue understanding bottlenecks in modal environments and proposes a linguistic dialogue understanding method combining symbolic and elastic fusion. By establishing an adaptive dialogue framework, it achieves a 40% border customer service ticket resolution rate and boosts emergency scenario instruction recognition accuracy to over 90% [1]. This provides theoretical guidance and technical paradigms for breaking the cognition divide and creates new possibilities for embodied intelligence in the real world.

This paper aims to create a technical system for robot dialogue target identification in complex linguistic environments across dimensions. First, it uses modal coupling to prevent language expression fragmentation. It also identifies the intentions in language sentences based on semantic field division. Second, it employs a dynamic cognitive making model to balance task efficiency and adaptability to pragmatic contexts. Finally, it proposes an elastic dialogue architecture. This architecture integrates dialogue systems' environmental tolerance and understanding of specific communication scenarios through symbolic reasoning and embodied environmental perception. These capabilities will be further developed in the future.

2. Theoretical foundations and research framework

2.1. Core theory of accurate dialogue target identification

Modal semantic alignment addresses the semantic gap between different modalities, such as text, speech, and vision, which have different representational spaces. Early methods involved simple addition or concatenation of heterogeneous semantic representations for grained-level alignment. For example, the earliest studies [2] used basic methods like directly adding the embedded vectors of text representations with the MFCC spectral features of speech representations to obtain a joint representation. However, due to the distinct characteristics of different modalities, the accuracy of this grained modal semantic alignment was only about 57%, making it difficult to apply in more complex situations. Subsequent research gradually moved toward grained semantic alignment at the cognitive level [3].

2.2. Architectural theory of interactive systems

2.2.1. Structured fusion architecture: transformer and joint representation

Different modalities' information alignment and fusion are achieved through independent modality encoders and modal interaction layers. Taking ViLBERT [4] as an example, it uses a stream Transformer model to perform supervised encoding on visual and linguistic modalities separately. In the modal attention (attention) layer, the semantics of the two modalities are aggregated to form a unified joint representation. Experimental results show that the ViLBERT model achieves a 75.3% accuracy on the VQA 2.0 dataset, a significant improvement over stream architectures like LXMERT.

DOI: 10.54254/2755-2721/2025.LD27506

2.2.2. Contrastive learning paradigm: contrastive loss and feature consistency constraints

Based on the contrastive learning mechanism, the modal contrastive loss function (ITM loss) is proposed to enhance the distinction between positive and negative modality pairs, thereby refining the granularity of semantic alignment. By combining this function with masked language modeling (MLM) and text matching (ITM) tasks, the unit energy achieves a 91.2% F1 score in modal intent recognition, a 14.3% improvement over the baseline model. This also improves the robustness of modalities and task generalization capabilities [5].

2.2.3. Dynamic attention mechanism: interleaved encoding and interaction

For the temporal dependencies in modal sequence data, designing a dynamic attention module to capture contextual associations between modalities is crucial. To this end, an interleaved modal encoder is proposed. It alternately processes text sequences and uses gated attention for dynamic information fusion. Experimental results show that on the Ego4D Answer Retrieval task, the model improved by the Flamingo method achieves an end accuracy of over 78.6%, showing strong advantages [6].

3. Research gaps and technological challenges

3.1. Challenges in natural language understanding, generation, and interaction

In natural language processing, dialogue systems and large-language models face understanding and generation challenges. Despite the emergence of many deep learning-based task dialogue systems, they still fail to effectively address issues such as polysemy, metaphor, and complex logical syntax in natural language understanding. Moreover, when deep learning-based language models cross semantic boundaries into other fields, they generate numerous errors, leading to incorrect understanding and migration. In terms of natural language generation, large-language models have issues such as low-quality generated content and lack of consistency and style. They cannot ensure semantic coherence in long texts or maintain similar styles and are prone to content shift problems. Furthermore, interaction research is still in its infancy. Studies on multimodal information fusion and alignment are not in-depth enough. There are also difficulties in semantic association and fusion between different modalities in multi-modal dialogue systems, which hinder the accurate grasp of users' real needs and the development of multi-modal dialogue systems.

3.2. Multi-field application, operational mechanism, experimental education, and resource-training dilemmas

Firstly, with diverse application scenarios and significant differences in requirements across fields, models struggle with migration and adaptability. For instance, in the military domain, which demands high levels of security and accuracy, general-purpose large language models fail to meet the requirements of professional military knowledge. In the field of preschool education, most current models cannot adequately address the cognitive characteristics of children. Secondly, regarding the operational mechanisms of large language models, unclear principles and black box-like decision-making processes are not conducive to performance optimization and controllability improvements. The lack of interpretability also reduces user acceptance and willingness to use these models, limiting their role in critical decision-making scenarios. Thirdly, in experimental teaching at application-oriented universities, there is a significant gap between artificial intelligence teaching

DOI: 10.54254/2755-2721/2025.LD27506

and practical applications. The curriculum tends to focus on theoretical knowledge and simple model building, with little involvement in real-world industrial problems. This severely impedes the development of students' engineering practice abilities. Fourthly, in low-resource scenarios, models are restricted by resources and the amount of training data. They cannot fully learn effective patterns and feature information, making them prone to overfitting and weak generalization capabilities. This affects the expansion of models into emerging and obscure fields.

3.3. Ethical and social impacts

The technological advancements in dialogue systems will profoundly transform human production and lifestyle, bringing multifaceted impacts across different dimensions of existence. Ethically, users face issues such as complete privacy exposure throughout the data lifecycle, accelerated spread of false information, algorithmic training data bias, and psychological attachment formed by algorithmic suggestions. Socially, technological updates accelerate job market polarization. While some jobs are replaced, high-level technical positions emerge. Virtual spaces are substituting for real-world human connections. The interplay between "immediacy" and "depth" forms a contradictory unity, and the tension between technological development and application across regions and groups exacerbates the digital divide [7]. To achieve a balanced state between technological benefits and human values, it is essential to properly clarify the relationships between dialogue parties and define the roles of responsible parties. Establishing technical ethics review mechanisms, multi-stakeholder collaborative governance systems, and inclusive technical education systems are crucial solutions to these contradictory unities.

4. Research status and key technology analysis

4.1. Technological breakthroughs and evolution

Learning models in natural language processing and language models promote each other, driving the development of dialogue systems. Learning models, through scale training and structural optimization (such as BERT architecture improvements), have significantly enhanced oriented dialogue systems' recognition capabilities and reduced reliance on manual feature engineering. Learning methods for shot learning can address task recognition issues in resource scenarios. models, in addition to generating sentences, have explored spatial cognition (e.g., quantifying syntactic structures based on dependency distance) and operational mechanisms (such as studies on Tongyi Qianwen). These advancements have greatly improved dialogue systems' and explainability capabilities. In practical applications, modular design and end architectures have their own advantages and disadvantages. Modular architectures are easy to maintain but may suffer from information loss, while end architectures require large amounts of training data and computational resources but enable overall optimization. These two architectures are applied in different fields. End systems are used in military applications (combined with knowledge graphs and reinforcement learning), and a combination of end and modular modes is applied in preschool education (combined with modal interaction and personalized recommendations).

4.2. Educational reform and resource inclusiveness

The rapid development of system technologies is compelling reforms in cultivation and allocation methods. The curriculum in applied universities' AI experimental teaching is in a "heavy, light" and "focused, neglected" state. However, by adopting mainstream industrial technologies (dialogue

DOI: 10.54254/2755-2721/2025.LD27506

systems) and introducing project cases, the integration model can address students' insufficient engineering practice abilities and professional quality. On the other hand, algorithm innovation for resource scenarios with small data volumes is a key aspect of inclusive technology. The ADTDNN model, using a small amount of labeled data and combining supervised learning with adaptation ideas, has achieved good results in speech recognition and can be applied to dialect recognition and resource language interaction scenarios. This dual approach not only provides mutual guidance for education and algorithm systems but also promotes the popularization of system technologies, enabling AI to develop more fairly and equitably across diverse languages and cultures.

5. Frontiers and future directions

5.1. Technological breakthroughs and resource scenario challenges

With the support of language models, hardware adaptation, model iteration, and modal fusion, dialogue systems are addressing application bottlenecks in resource scenarios. Based on technologies such as model compression and inference acceleration from Tongyi Qianwen [4,5] and supervised understanding and interaction algorithms [8,9], positive results have been achieved in dialect speech recognition and short dialogue understanding. These technological advances support educational scenarios and preservation efforts in constrained regions, promoting the deployment of dialogue systems on edge devices and in lightweight scenarios.

5.2. Field deepening and capability development

Oriented dialogue systems in vertical industries are developing rapidly. Through technologies such as knowledge graphs, reinforcement learning, and task transfer learning, these systems provide reliable responses in risk application scenarios and quickly adapt to scarce scenarios. Universities can also drive teaching reform based on the "technology scenario toolchain" model [10,11]. This enables students to access mainstream industrial technologies through teaching activities, resources, and policies. After sufficient practice, students can participate in based learning for specific vertical fields. This approach is highly practical and effectively bridges the gap between university education and the workforce.

6. Conclusions

This study systematically addresses technical bottlenecks in dialogue systems operating within complex linguistic environments through three interconnected advancements: multimodal semantic alignment, dynamic context reasoning, and elastic architecture optimization. The evolution from simplistic feature concatenation to cognitive knowledge-integrated models—powered by stream Transformers and contrastive learning paradigms—has elevated multimodal intent recognition accuracy, while dynamic attention mechanisms employing interleaved encoding resolve temporal sequence modeling challenges. The proposed elastic framework, synergizing symbolic reasoning with embodied perception, shifts systems from passive environmental tolerance to proactive situational understanding, validated by empirical successes such as 40% autonomous resolution in border customer service and over 90% emergency instruction accuracy [12,13]. Concurrently, technological diffusion is reshaping industries: integrated chips will reduce vehicular system demands by 60% by 2026, while applications span military instruction interpretation (0.8-second latency), diagnostic systems (1.2% error rates), and inclusive education (12 low-resource language support). Projected to reach $80 billion by 2027, the "AI + language services" paradigm underscores

DOI: 10.54254/2755-2721/2025.LD27506

cross-sector economic transformation [14,15]. Achieving contextual nuance, however, necessitates interdisciplinary collaboration—cognitive science must decode human multimodal processing to refine symbolic architectures, sociology must model cultural metaphors, and education must bridge theory-practice gaps via "technology-stack-scenario-toolchain" curricula. A proposed national innovation hub unifying linguistics, neuroscience, and industry could standardize ethical algorithms and multimodal alignment, ensuring dialogue systems evolve as both technologically robust and socially equitable instruments.

References

[1] Chu Ruijie, Zhuang Rui. Analysis of the Mechanism of User Trust Construction in the Context of Computer Interaction for Generative Artificial Intelligence, Taking DeepSeek as an Example [J]. Journal of Hubei University of Economics (Social Science Edition),
[2] Wu Guangjun, Ding Ze. An Exploration of the Language Features of Translations Generated by Language Models: Based on the Average Dependency Distance and Its Relationship with Translation Quality [J]. Foreign Language Quarterly,
[3] Wu Ruoling, Guo Danhuai. Research on the Testing Standards for the Spatial Cognition Capabilities of Language Models [J]. Earth Information Science Journal, 2025, 27 (5): 1052.
[4] Ni Cuixia, Xu Di. Research on the Operational Mechanism of Language Taking Tongyi Qianwen Language Model as an Example [J]. Scientific and Technological Innovation, 2025(9): 104.
[5] Zhao Zhengping. New Progress in Artificial Language Models and AI Chips (Continued) [J]. Electronics Technology, 2025, 62 (4): 39.
[6] He Xuefeng, Zhou Jie, and Chen Liao Hai. Review of Learning Models in Natural Language Processing [J]. Computer Applications and Software, 2025, 42 (2): 19 + 101.
[7] Du Wei, Wang Ban. Thoughts on the Design of Artificial Intelligence Experimental Teaching Mode in Applied Based on Mainstream Industrial Technologies for Dialogue Systems [J]. Innovation and Entrepreneurship: Theoretical Research and Practice, 2025, 8 (3): 17.
[8] Gu Longhao, Huang Lianli, and Zhou Zhang Ziyue. Research on Resource Speech Recognition Method Based on ADTDNN [J]. Software Guide, 2024, 23 (9): 76 - 81.
[9] Pan Lisha. Research on Preschool Education Robot Dialogue System Based on AI Artificial Intelligence [J]. Automation and Instrumentation, 2023(5): 248.
[10] Sun Weibo, Zhang Bin. Application of Based-Oriented Dialogue Systems in Military Fields [J]. Cybersecurity Technology and Applications, 2021(4): 137.
[11] Zhou Qian'an, Li Zhoujun. Improved Model and Method for Natural Language Understanding in Oriented Dialogue Systems Based on BERT [J]. Journal of Chinese Information Processing, 2020, 34 (5): 82 - 90.
[12] Huang Longfei. Application of Contextual Information in Detection Algorithms [J]. Electronic World, 2020 (2): 44.
[13] Key Technologies and Applications of Dialogue Systems [J]. China Science and Technology Awards, 2020(1): 70.
[14] Liu Jiming, Meng Wan Xiaoyu. Task Dialogue System Based on Shot Machine Learning [J]. Journal of Chongqing University of Posts and Telecommunications (Natural Science Edition), 2019, 31 (3): 304.
[15] Liu Zhiqiang. The Dialogue Management System of DSS and Its Design [J]. Zhejiang Economic and Trade College Journal, 1993, (2):