GEORGE BROWN PRESS
JOURNALS, BOOKS & PROCEEDINGS

Journal of Computer, Signal, and System Research

Article

Research on the Application of Multilingual Natural Language Processing Technology in Smart Home Systems

Xiang Chen [1,*]

[1] Microsoft, Azure, Washington, 98052, USA
* Correspondence: Xiang Chen, Microsoft, Azure, Washington, 98052, USA

Abstract: With the popularization of smart home systems, users' demand for multilingual voice interaction is increasing day by day. This paper proposes a multilingual natural language processing system architecture for the smart home environment, designs core modules such as speech recognition, semantic understanding and instruction mapping, and conducts system deployment and experimental tests in the multilingual home environment. The experimental results show that the system has a high language recognition accuracy rate and good human-computer adaptability. It can support the realization of multi-language intelligent interaction.

Keywords: smart home; multilingual processing; natural language understanding; speech recognition; human-computer interaction

Received: 01 June 2025
Revised: 06 June 2025
Accepted: 23 June 2025
Published: 26 June 2025

1. Introduction

With the development of artificial intelligence and Internet technology, the human-computer interaction mode of smart home systems is gradually evolving towards a new interaction mode with natural language as the bridge. Mainstream voice control systems only support a single language and cannot meet the actual needs of multilingual families. Therefore, this paper studies a natural language processing technology for multilingual interaction, conducts in-depth discussions on key issues such as language recognition, semantic understanding and multilingual switching, and verifies the feasibility of this technology in practical scenarios through application research.

2. The Multilingual Requirements in Human-Computer Interaction of Smart Homes

In multilingual families, cross-cultural communities and international market environments, different user groups prefer to use multiple languages for human-computer interaction. Traditional voice assistants usually only support one mainstream language. In a multilingual environment, it is difficult to achieve human-computer interaction, which may cause communication difficulties or even render the system unusable. Users generally expect the system to have the ability to automatically determine language types and switch seamlessly, while accurately understanding the semantics of multilingual inputs, thereby achieving cross-language human-machine instruction recognition and response [1]. In some areas, due to the presence of a large number of local languages, dialects or low-resource languages, the support for these languages in the existing systems is relatively limited, which further restricts the coverage and potential user scale of smart home systems. Under the background of international deployment, the multilingual adaptability of voice systems and their evaluation effects have become one of the important factors for measuring their intelligence level and market competitiveness.

3. Design of Multilingual Natural Language Processing Technology Architecture

3.1. Overall System Architecture Design

The overall framework of a multilingual natural language processing system can be divided into five main parts, including speech input, language recognition, semantic understanding, intermediate semantic layer and instruction mapping, and home control execution. Users issue voice commands through the voice terminal. Language recognition is performed by the voice recognition module, which converts the commands into standardized text information. The text is then processed by the multilingual semantic understanding module, which uses a pre-trained language model to identify user intent and extract relevant parameters. To improve language consistency, the system introduces an intermediate semantic layer, abstracts the semantic contents expressed in different languages into a unified format, converts them into control instructions, such as light switches, air conditioning adjustments, etc., and issues them to the corresponding devices through the interface module to complete the control operations [2]. This system adopts a modular design concept, supports real-time addition, deletion and on-demand loading of language packs, and is capable of adapting to users' language preferences and usage environment requirements. The application of the intermediate semantic layer has successfully decoupled language understanding from control logic. This allows semantic expression to be independent of specific languages, improves the system's maintainability and scalability, and creates favorable conditions for future expansion into multimodal interaction (such as image recognition and gesture control) and integrated applications on edge computing platforms (Figure 1).

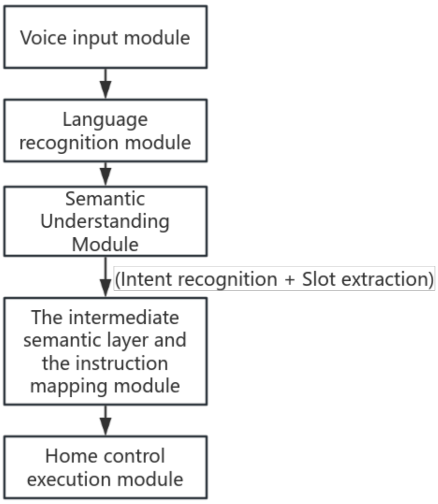
Voice input module
Language recognition module
Semantic Understanding Module
(Intent recognition + Slot extraction)
The intermediate semantic layer and the instruction mapping module
Home control execution module

Figure 1. Overall Architecture Diagram of the Multilingual Natural Language Processing System.

3.2. Design of Multilingual Natural Language Understanding Module

The multilingual Natural Language Understanding (M-NLU) module is the core foundation of the intelligent voice interaction system. Its main functions are to achieve intention recognition and slot filling in different language environments, accurately understand the input content in different languages, and map it into structured home appliance control commands. Therefore, the multilingual natural language understanding module needs to address issues such as disordered word order and ambiguous semantic expression caused by differences in language structure, as well as the difficulty of model transfer in low-resource languages [3].

The system uses multilingual pre-trained language models with shared parameters (e.g., mBERT or XLM-R) as its base encoder. The input of the model is the multilingual text sequence transcribed by the user through speech recognition, and the output is the corresponding intention classification result and the semantic vector of slot labeling. The input sequence is denoted as $X = \{x_1, x_2, \dots, x_\mathrm{n}\}$. After being encoded by the encoder f_θ, the context representation is computed as follows:

$$H = f_\theta(X) = \{h_1, h_2, \dots, h_n\}, h_i \in R^d \tag{1}$$

Here, h_i represents the context semantic representation of the *i*-th word, and *d* is the dimension of the hidden layer. The semantic summary vector h_{CLS} of the entire sequence is used for sentence-level intention recognition, and its intention category prediction formula is:

$$\hat{y}_{int\,ent} = soft\,max(W_{int\,ent} \cdot h_{CLS} + b_{int\,ent}) \tag{2}$$

Slot filling uses a sequence labeling method to classify each word vector h_i:

$$\hat{y}_i^{slot} = soft\,max(W_{slot} \cdot h_i + b_{slot}) \tag{3}$$

The model training adopts the joint optimization strategy of intent classification loss and slot filling loss, and the total loss function is defined as follows:

$$L(1-\lambda)_{slot} joint_{joint} \tag{4}$$

Among them, $\lambda \in [0,1]$ is a hyperparameter, which is used to adjust the training weights of the two types of tasks. Through end-to-end joint modeling, the system can better capture semantic relations in a multilingual context and has good cross-language generalization ability.

To adapt to low-resource languages, this system uses a cross-language transfer strategy during the training process. By sharing the representation space, the language knowledge in high-resource languages is effectively transferred to low-resource languages. This cross-language migration strategy effectively alleviates the data imbalance problem among multiple languages by sharing the word embedding layer and encoding layer between the source language and the target language.

This module supports language-independent semantic encoding and includes a joint modeling mechanism for intent recognition and slot parsing. It also enables cross-language transfer learning and can achieve high-precision semantic and command parsing in a multilingual smart home environment [4].

3.3. Multilingual Speech Recognition and Synthesis Interface

In the voice interaction system of smart home, the speech recognition (ASR) and speech synthesis (TTS) modules are the core parts of the human-computer natural language closed-loop interaction. For a multilingual environment, ASR should have the functions of language recognition and dynamic model selection. During voice input, it should automatically identify the language and invoke the corresponding recognition model to ensure accurate multilingual transcription. To improve the matching degree and robustness, this paper adopts a unified multilingual recognition model and combines it with a lightweight language detector, effectively simplifying the processing steps during the operation. The system performs semantic analysis on the recognized text, extracts instruction information, and achieves accurate intent understanding. The TTS module generates speech responses. It must support multilingual synthesis and provide features like personalized speaker settings and pitch control to produce speech that closely resembles natural human conversation. Meanwhile, in order to improve the response efficiency and scalability, the architecture adopts the decoupling of ASR and TTS modules and the modular design of interfaces, enabling ASR and TTS to operate independently. This not only supports local deployment but also allows for flexible networked installation. At the same time, it is convenient to integrate with mainstream voice apis or self-developed models. It can be adapted to the computing power and performance requirements of different household devices to ensure that the smart home system achieves a smooth and natural voice interaction experience in a multi-language environment.

3.4. Intermediate Semantic Layer and Instruction Mapping Mechanism

In multilingual intelligent speech systems, different languages have significant differences in grammatical structure and expressed meaning. Directly mapping the NLU results of different languages to operation commands can lead to semantic inconsistency due to structural and syntactic differences among languages, making it difficult to maintain uniform intent interpretation and hindering future system expansion. To solve these problems, this paper introduces an intermediate semantic layer as an intermediary mechanism between language understanding and device control to unify the expression forms of different language intentions. This process will uniformly transform the input intent and its parameters of all languages. For example, for the command "Turn on the bedroom light", regardless of the language used for expression, the system can standardize it into a unified format (such as intent: "TURN_ON", parameters: "LOCATION: BEDROOM", "DEVICE: LIGHT"), achieve language-independent processing at the semantic level. Based on this unified semantic representation, the system can apply rule mapping or template matching to translate user intents into executable commands, which in turn enables control over various types of devices such as lights, air conditioners, and audio equipment. The instruction format is designed hierarchically and modularly to ensure compatibility with different manufacturers' protocols, thereby offering excellent scalability and integration flexibility. The intermediate semantic layer not only enhances the interpretability and stability of semantic information processing, but also improves the adaptability of the system in complex interaction scenarios such as multilingual mixed input and user-defined expressions, providing a unified semantic structure basis for subsequent language expansion and functional upgrades.

4. Applied Research and Result Analysis

4.1. Multilingual Home Environment Deployment

To verify the practicality and language adaptability of multilingual natural language processing technology, this paper selected six families with multilingual usage experience, deployed and tested the proposed system in the actual living environment, including six language environments covering Chinese, English, Spanish, French, Arabic and Russian. Each family includes 2 to 5 individuals from different countries and regions. Each member has their own native language and there are frequent language switching behaviors in daily life [5]. The experimental period lasted for one month, and the research contents included functions such as voice control, device linkage, automatic language recognition and multi-round command response.

The system deployment involves voice interaction terminals (smart speakers or voice central control screens), cloud-based multilingual NLP processing modules, local semantic parsing and control interfaces, as well as smart lighting, curtains, thermostats and other devices that support multi-brand protocols. Family members do not need to manually switch language modes. The system recognizes the language used by the user through the preset language detection mechanism and matches the corresponding language processing model, allowing family members to freely communicate with the system in any language. The language composition, the number of devices and the average daily voice interaction frequency of each family are shown in Table 1 as follows:

Table 1. Statistics Table of Basic Information for Multilingual Home Deployment.

Family number	Main language used	Number of members	The number of connected devices	The average daily number of voice interactions
F1	Chinese + English	4	12	52
F2	Spanish + English	3	9	35
F3	Chinese + French	5	15	58
F4	Arabic + English	4	10	40

F5	Chinese + Russian	2	7	26
F6	English + French + Arabic	5	13	47

It can be seen from the table that in a multilingual environment, the average daily number of voice interactions of users is approximately 26 to 58 times, and the language switching rate is positively correlated with the diversity of native languages spoken by family members. Especially in the case of frequent switching among multiple languages (such as F3 and F6), the overall operation of the system is stable without any delay, indicating that the language recognition and model switching mechanism demonstrating strong real-time responsiveness and robustness against interference.

4.2. Evaluation Indicators

To objectively measure the performance of multilingual natural language processing systems in the smart home environment, this paper designs a multi-level evaluation index system, covering multiple aspects such as the accuracy of speech recognition, the effect of semantic understanding, the system response efficiency and the subjective satisfaction of users.

In the speech recognition stage, the Word Error Rate (WER) is adopted as the core evaluation index and is defined as follows:

$$WER = \frac{S+D+I}{N} \tag{5}$$

Among them, *S* represents the number of replacement errors, *D* represents the number of deletion errors, *I* represents the number of insertion errors, and *N* is the total number of reference words. Lower WER indicates higher speech recognition accuracy. The language identification accuracy is also used as a supplementary indicator to assess the system's ability to correctly detect the language of input speech.

In the semantic understanding part, the system adopts the Accuracy evaluation for the intention recognition task:

$$Accuracy = \frac{\text{Number of Correct Predictions}}{\text{Total Samples}} \tag{6}$$

Slot filling is framed as a sequence labeling task, requiring token-level classification to extract structured information. the F1 score is used as the primary metric, integrating both Precision and Recall to provide a balanced evaluation:

$$F1 = \frac{2 \cdot P \cdot R}{P+R} \tag{7}$$

Among them,

$$P = \frac{TP}{TP+FR} \tag{8}$$

$$R = \frac{TP}{TP+FN} \tag{9}$$

The system interaction efficiency is measured by the average Response Time and instruction execution success rate. Response time measures the delay between voice input and device action, while success rate represents the ratio of executed instructions to total issued commands.

Based on the above evaluation criteria, the system can quantitatively measure the interaction effects in multiple languages and scenarios, providing theoretical basis and practical reference for subsequent algorithm optimization and user experience improvement.

4.3. Experimental Comparison

To comprehensively evaluate the applicability of multilingual natural language processing systems in different language environments, this paper conducts comparative tests in six language environments: Chinese, English, Spanish, Arabic, French and Russian. And eight performance indicators are set as the evaluation basis, including word error rate (WER), intent recognition accuracy rate, slot filling F1 score, multi-round dialogue understanding accuracy rate, average response time, execution success rate, first recognition success rate and system stable operation time, etc. (Table 2).

Table 2. Experimental Performance Expansion Comparison Table of Multilingual Smart Home System (n = 300 Instructions/Language).

Language	WER	Intention recognition	F1 score	Multi-round accuracy rate	Response time (seconds)
Chinese	0.062	0.960	0.93	0.952	1.21
English	0.087	0.940	0.91	0.928	1.38
Spanish	0.093	0.913	0.89	0.895	1.52
Arabic	0.108	0.890	0.86	0.872	1.67
French	0.089	0.925	0.90	0.914	1.45
Russian	0.095	0.910	0.88	0.887	1.58

It can be seen from the table data that Chinese performs the best among various indicators. Its word error rate (WER) is the lowest, only 6.2%, the intent recognition accuracy rate is as high as 96.0%, the F1 value is the highest (0.93), and with a response time of only 1.21 seconds, it outperforms all other languages in speed. English and French also perform well in terms of language understanding and interaction effects, while Spanish and Russian exhibit slightly lower performance in recognition accuracy and response time. Although its performance is slightly lower, the overall system stability still meets the practical requirements of smart home usage. However, due to the significant phonetic changes in Arabic, it brings certain difficulties to the system recognition. As a result, it lags behind other languages in terms of WER, initial recognition rate, and response latency. But it achieves more than 89% intention understanding and more than 90% instruction success rate.

4.4. User Feedback Data Analysis

To further assess user interaction experiences with multilingual NLP systems in real-world home environments, this paper conducts a subjective satisfaction survey on multilingual user groups (including Chinese, English, Spanish, French, Arabic and Russian). The survey content involves the accuracy of speech recognition, the naturalness of speech synthesis, the satisfaction of response speed, the smoothness of language switching, the coherence of multi-round dialogues and the overall satisfaction. A five-point Likert scale (ranging from 1: very dissatisfied to 5: very satisfied) was adopted for quantitative analysis. The data organization results are shown in Table 3.

Table 3. Summary of Multilingual User Feedback Ratings (Full Score: 5 Points).

Evaluation dimension	Chinese	English	Spanish	French	Arabic	Russian
Accuracy of speech recognition	4.8	4.6	4.4	4.5	4.2	4.3
Naturalness of speech synthesis	4.6	4.5	4.3	4.4	4.1	4.2
Satisfaction with response speed	4.7	4.4	4.2	4.3	4.0	4.1
Fluency in language switching	4.6	4.5	4.1	4.2	3.9	4.0
Coherence of multiple rounds of dialogue	4.7	4.5	4.3	4.4	4.0	4.1
Overall satisfaction	4.7	4.5	4.3	4.4	4.1	4.2

Among all groups, Chinese users reported the highest satisfaction, with an average score of more than 4.6 points, indicating that the system has good recognition accuracy and response speed in the Chinese environment. Secondly, there are English and French users, indicating the system's strong performance in Western European language environments as well. The ratings of Spanish and Russian users were slightly lower, mainly reflected in the slightly inferior evaluation of the diversity processing of speech formats and the response speed. Arabic users reported the lowest satisfaction, primarily due to the system's limited ability to handle diverse morphological structures and maintain contextual continuity in dialogue, suggesting that existing models require further optimization to better accommodate morphologically rich and context-dependent languages.

Despite variations in language-specific performance, all user groups gave average satisfaction scores above 4.1, indicating that the system has reached a usable level in terms of the naturalness of human-computer interaction, cross-language adaptability, and overall operational stability, demonstrating its practical viability for deployment in multilingual, multi-user smart home environments.

5. Conclusion

This paper comprehensively studies the application of multilingual natural language processing technology in the intelligent home system, proposes a natural language interaction architecture that integrates multilingual recognition, semantic understanding and instruction mapping, and conducts systematic testing and user evaluation of its performance in multiple language environments such as Chinese, English and Spanish. The experimental results show that the average word error rate (WER) of the system is less than 9.5%, the accuracy rate of intention recognition is not less than 92%, the F1 value of slot filling reaches above 0.90, the response speed is between 1.2 and 1.7 seconds, and the average user satisfaction score is 4.5 points. These results demonstrate the system's strong adaptability and stability across both high-resource and low-resource languages. This study verified the practical application value of the proposed system architecture in the multilingual home environment. Furthermore, it provides a scalable technical framework and empirical foundation for future advancements in multilingual human-computer interaction (HMI) system design and deployment.

References

1. Z. Wei, Y. Liu, M. Zhao, J. Chen, F. Wang, H. Li, et al., "Optimized attention enhanced temporal graph convolutional network espoused research of intelligent customer service system based on natural language processing technology," *Appl. Artif. Intell.*, vol. 38, no. 1, p. 2327867, 2024, doi: 10.1080/08839514.2024.2327867.
2. Y. Chen, H. Zhang, and S. Zhong, "Design and implementation of smart home system based on IoT," *Results Eng.*, vol. 24, p. 103410, 2024, doi: 10.1016/j.rineng.2024.103410.
3. Z. Li, "Research and implementation of low resource voice awakening technology in smart home scene," *Acad. J. Comput. Inf. Sci.*, vol. 7, no. 11, pp. 96–101, 2024, doi: 10.25236/AJCIS.2024.071113.
4. X. Zhang, D. Luo, and X. Wu, "Based on Raspberry Pi voice controlled smart home system," *Front. Comput. Intell. Syst.*, vol. 8, no. 2, pp. 38–42, 2024.
5. M. Orosoo, D. Bat-Erdene, B. Ganbat, B. Naranchimeg, B. Tuvshintugs, B. Khurelbaatar, et al., "Enhancing natural language processing in multilingual chatbots for cross-cultural communication," in *Proc. 5th Int. Conf. Intell. Commun. Technol. Virtual Mobile Netw. (ICICV)*, 2024, doi: 10.1109/ICICV62344.2024.00027.

Disclaimer/Publisher's Note: The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of GBP and/or the editor(s). GBP and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.