Wireless Pers Commun (2018) 102:2927–2939
https://doi.org/10.1007/s11277-018-5316-2

Research and Experiment of Intelligent Natural Language Processing Algorithms

Zeliang Zhang[1] · **Xinwen Bi**[1]

Published online: 19 January 2018

Abstract Natural language processing is mainly divided into two parts: speech processing and word processing. The level of word processing is mainly studied. Natural language processing is divided into lexical analysis, syntax analysis and semantic analysis. Aiming at the scope of the language ambiguity and thesaurus in the field of smart home, the maximal matching algorithm is used to segment the natural language. Then, through the way of template matching, semantic comprehension finally forms the code form that can control the home node. In the system applied in this paper, the speech is processed into words through the existing voice input function of the mobile terminal. Then, the control instruction is obtained through the language processing method. The processed data is communicated to the server via socket. The server sends the data to the home node through the Zigbee protocol. Finally, control of home appliances is achieved.

Keywords Natural language processing · Maximum matching word segmentation algorithm · Line graph syntax analysis · Semantic understanding · Smart home

1 Introduction

Compared to Europe and America, the study of natural language started relatively late in China, and the study of Chinese understanding began in the eighties. At the same time, the characteristics of Chinese are different from the European and American languages. For some mature methods in Europe and the United States, we cannot directly copy it. Moreover, the Chinese language itself is quite complicated. All these put forward quite serious tests for understanding of Chinese. Even so, through hard work, research and

✉ Xinwen Bi
xinwenbi1289@sina.com

[1] School of Information Technology and Media, Beihua University, Jilin 132011, China

applications in this area have outstanding achievements. The typical theory has the concept of hierarchical network theory. The typical application is the cnki.

2 The Related Methods of Natural Language Processing

2.1 Natural Language Processing Techniques

Natural Language Processing technology uses computers to analyze and process natural language units at all levels, such as words, sentences, paragraphs, chapters, etc [1]. The process of Natural Language Processing is to abstract a specific problem of natural language based on the input set and output set, and design the effective algorithm process based on this model. The natural language processing model is shown in Fig. 1.

It is generally believed that the Natural Language Processing process is divided into several stages. Lexical analysis is the first step in the whole process. It is also the most basic, including the Linguistic string and Part of speech (POS). Syntactic analysis is to clarify the structure of a sentence according to the results of the previous step. Semantic analysis is to understand the intrinsic meaning of input on the basis of the above preparation. Discourse analysis and pragmatic analysis focus on the effect of paragraph or even the environment on the meaning of the sentence. In daily life, the understanding of natural language is not carried out in stages, but in parallel with several stages. Therefore, it is awkward to handle the above natural language in stages. In fact, in the absence of parallel processing, computer processing of natural language can only be carried out in such a phased form.

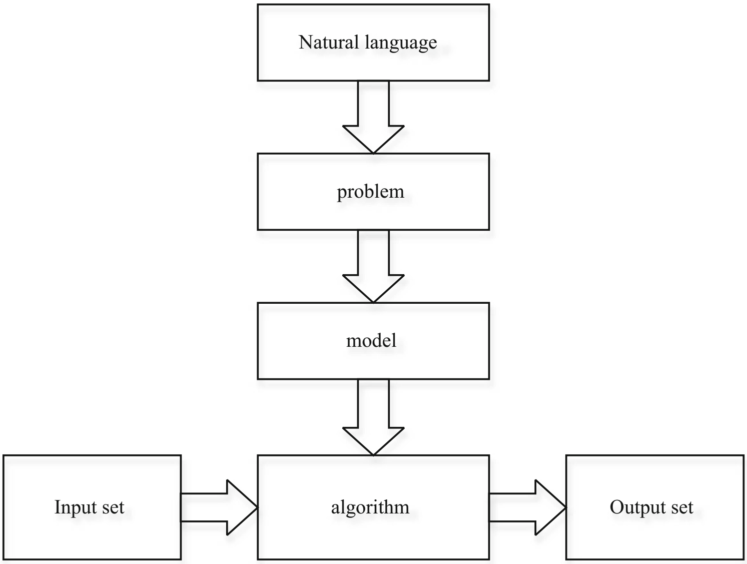


Fig. 1 The natural language processing model

2.2 Lexical Analysis

Lexical analysis is to divide a complete sentence into one word after another. Because Chinese is different from the European and American characters, there is not only the distinction between the grammatical and the number, but also no interval between words and symbols [2]. Therefore, the analysis of Chinese must be carried out first. Due to the existence of a number of words with multiple lexical [3], some people classify the part of speech tagging into the scope of syntactic analysis. However, in any case, only the division of words is not enough to make a syntactic analysis. As a result, the words that are analyzed are also tagged for the attributes of their words.

In a wide range of natural languages, a dictionary is a set of finite numerals. Therefore, it is one of the difficulties in the domain of participle to carry out the segmentation of words which are not recorded in the dictionary, and the other is to eliminate the ambiguity.

Based on the word segmentation method, the general idea is to store the word or word formation rules that may appear in the input sentence in the dictionary. When the real text is entered, the computer performs a match lookup to implement the word segmentation in a manner similar to human-query dictionaries. When this method is applied, the dictionary is first established. Several common dictionaries include dichotomous lexicon dictionaries, dichotomous lexicon dictionaries, and so on. The structure of the dictionary is similar to the dictionary in daily life. The difference is in the process of querying the dictionary. The dictionary usually includes the index as well as the dictionary text [4]. It is somewhat similar to the dictionary query process.

After the establishment of the dictionary, its specific application methods include: the maximum matching algorithm; the least word segmentation algorithm; the shortest path matching algorithm [5]. Here, the maximum matching method is taken as an example for illustration.

Maximum matching algorithm: The maximum matching algorithm is divided according to the direction is divided into positive and negative. The analysis process is shown in Table 1.

Similarly, negative is to put the starting point at the end of the sentence. For example, "Tomorrow, we go to Changchun,". The process is shown in Table 2.

In the unrestricted area, the error rate of the participle in the forward method is 1/169. The error rate of the reverse word segmentation method is 1/245. Accuracy is not enough to meet the needs of practical applications. However, it can be compared with the results of the forward reverse participle. If the result is different, the word processing is ambiguous. However, in the field of smart home control, the size of the thesaurus and grammatical libraries that lead to the size of the thesaurus and grammar libraries is significantly reduced due to the field limitation. The ambiguity is also reduced, and the accuracy rate is greatly increased.

Table 1 An example of a positive maximum matching word segmentation algorithm

Tomorrow we go, tomorrow we, tomorrow I, tomorrow	S ="Tomorrow/"
We go to Changchun, we go Chang, we go, we	S = "Tomorrow/We/"
Go to Changchun, go Chang, go	S = "Tomorrow/We/Go/"
Changchun	S = "Tomorrow/We/Go/Changchun/"

Table 2 An example of reverse maximum matching word segmentation algorithm

We go to Changchun, they go to Changchun, go to Changchun, Changchun	S = “Changchun/”
Tomorrow we go, we go, go	S = “Go/Changchun/”
Tomorrow we, we	S = “We/Go/Changchun/”
Tomorrow	S = “Tomorrow/We/Go/ Changchun/”

2.3 Syntax Analysis

Grammatical understanding can accomplish two things. First, whether the original text input is in conformity with the relevant rules of the grammar; Second, according to the grammar, the structure of the sentence can be deduced. Furthermore, it helps to understand the semantic [6].

The transfer network is a form of a graph. The graph is made up of a variety of nodes and the edges of each node, as shown in Fig. 2. These nodes include a node for the first test state. At the same time, it also includes one or more nodes that represent the final state. In these nodes, S represents the node of the initial state. The terminating state is represented by a circle with a slash string. The graph shows that when the first word of a sentence is a, the state is transferred to the node A. When the category of second words is B, then the state is transferred to node B. When the state of third words is C, it can be transferred to node A. The next word node is D, which can be transferred to the termination state to complete the grammatical representation of sentences. The transfer network syntactic representation is shown in Fig. 2.

Dependency grammar is a kind of grammatical expression based on rewriting rules and transfer networks, which focuses on revealing the relationship between words and words. The verb is regarded as the center of the meaning of the sentence. All the remaining components are attached to it and are limited by the meaning it expresses. In the example of “we go to Changchun tomorrow”, “we” and “Changchun” are governed by the “go” and are dependent on the verbs. Therefore, as the root part of the tree, the verb forms the structure as shown in Fig. 3.

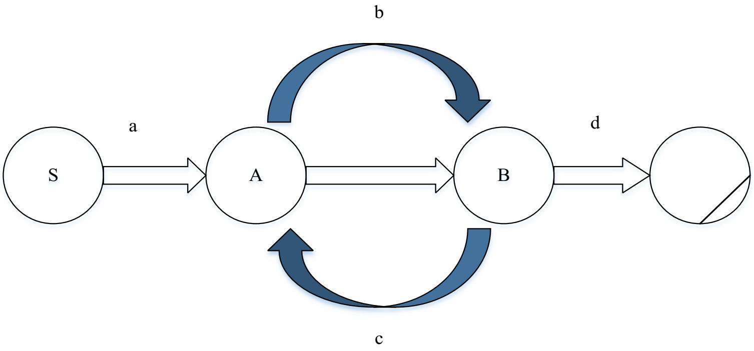


Fig. 2 The transfer network syntactic representation

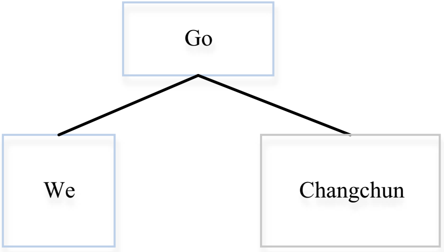


Fig. 3 Dependency grammar tree

2.4 Semantic Analysis

People understand natural language, not through the language itself, not even through words and grammar, but through the understanding of the concept. When people listen to the last sentence of a sentence, they recognize the meaning of the sentence. The reason is that people have clearly understood the meaning of the concept and the relationship between the concept and the concept. Therefore, the method of understanding semantics is mainly focused on the two aspects of the relationship between the meaning of the concept and the relationship between the concepts. The concept itself includes two parts: Connotation and extension. The connotation refers to the essential attribute of the concept, and the denotation refers to the concept of things. Specifically, the connotation describes the characteristics of the concept itself. Epitaxy is the set of things that the concept contains [7]. For example, the concept of "animal" includes concepts such as "man" and "tiger." Therefore, "human" and "tiger" have inherited from the "animal" content. At the same time, "people" and "tigers" are also extensions of the concept of "animals."

Due to its characteristics in smart home control, the information is also relatively limited. Therefore, the understanding of language meaning by template matching is suitable for use in the understanding of this article. Each template is a semantic block that expresses a certain semantic meaning. In combination with the understanding of the concept subordination theory, there is a connection between each semantic block. Therefore, in the process of semantic analysis, semantic templates are filled in according to the connections between semantic blocks. Then, the concept of the semantic block itself is understood, and the semantic information of the statement can be obtained.

3 Design of Natural Language Processing Algorithm in the Smart Home

3.1 The Overall Process of the Algorithm

The whole process is divided into several stages. The first step is to prepare for the next step, namely lexical analysis, syntactic analysis, semantic analysis, pragmatic analysis and discourse analysis. Among them, pragmatic analysis and text analysis are based on a relatively long original text, and the context is related to the analysis. The semantic analysis and understanding of a single sentence are considered only. As a result, the total includes the first three aspects of [8]: lexical analysis, syntactic analysis, and semantic analysis. Among them, the grammatical analysis divides the original sentence into word strings, and annotates each of them. According to the word part of speech, syntactic analysis is to clarify the relationship between words. A set of semantic blocks that can express semantics relatively independently is formed. The meaning of the sentence is understood after the above steps are completed, based on the above results.

In view of the characteristics of the above steps, the relevant information needed for each step is stored in the memory. In the stage of lexical analysis, a dictionary needs to be designed to make the computer understand the words in the text. In the phase of syntactic analysis, the grammatical format of the natural language needs to be stored in the computer. The meaning of a sentence generates the corresponding code. In some ways, the corresponding code is queried.

CD theory holds that concept is the smallest unit of human cognition of the world, and the process of semantic analysis is to understand the meaning of concepts and the relationship between concepts [9]. Therefore, the analysis of each word and the analysis of the structure of the sentence are dealt with in the two aspects.

The overall structure of the approach to the processing of the input Chinese is shown in Fig. 4.

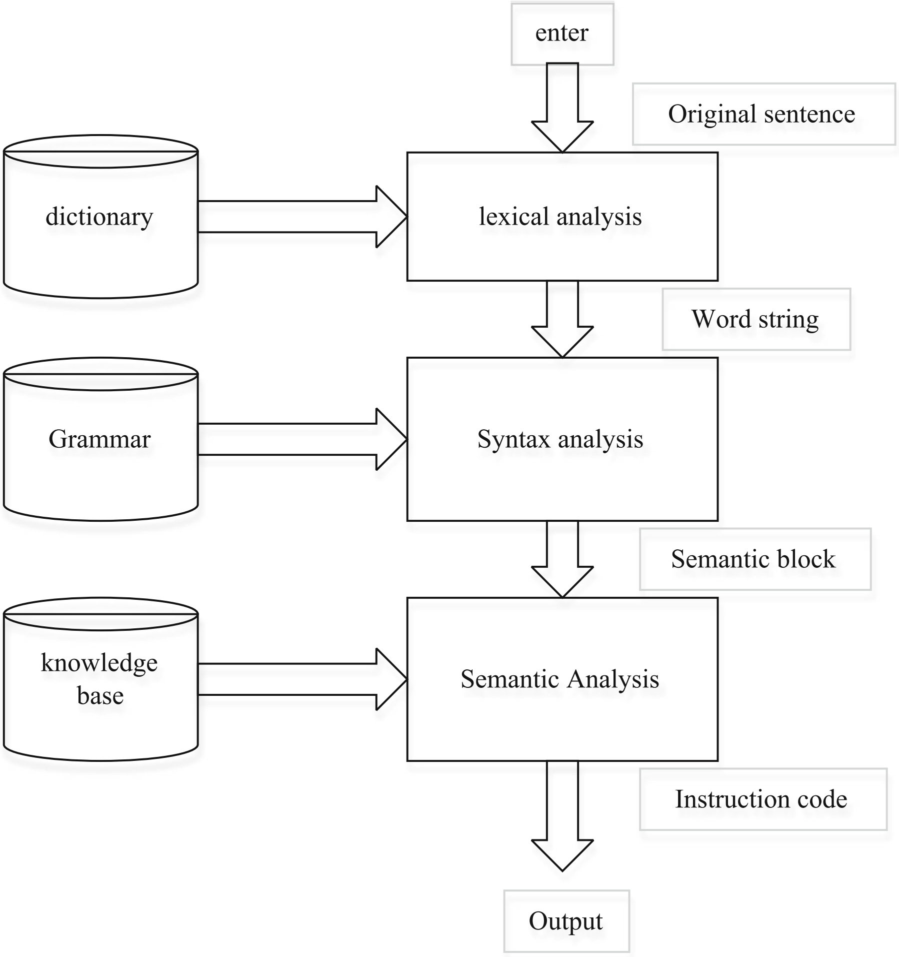


Fig. 4 The overall structure of the approach to the processing of the input Chinese

3.2 Semantic Representation

Because the ultimate aim of this article is to understand and feedback the meaning of Chinese characters, all the steps are carried out around this purpose. Therefore, the first definition of semantic representation can play a guiding role in a series of preprocessing.

By reading the smart home control system, this paper holds that the state of intelligent home control mainly includes temperature, brightness, humidity and air quality. For smart home control, the object will eventually be positioned as a home node, namely, the form of a noun. For different nodes, there are a variety of control methods [10], which eventually forms the code in the form of selecting the file. In other words, in semantic analysis, the expected control object is analyzed from the statement. Then, according to the specific target, in the control node code set, the code is looked up. It can analyze the original statement of the input into the form of the corresponding code. The focus of this article is to determine the home node and the related operations to the home node. Therefore, the framework of the semantic template is defined as:

(framework semantic template (slot object) (slot direction) (slot direction) (slot syntax))

The semantic templates are shown in Table 3.

The “control object” slot is usually filled by the NP part of the core nouns in the syntax tree. The “location information” represents the description information for the controlled object. Normally, it is filled by the modifier in the NP phrase, for example, the object “lamp” has many kinds. After the restriction, the “kitchen lamp” and “the bathroom lamp” separate the same types of objects. The “state” slot represents the environmental attributes of control, such as “illumination”, “brightness,” and so on, which are usually filled by some AP core adjectives. “Direction” indicates that the direction of changing the environment attribute is “up” or “down”. It can also indicate operation ways such as “opening” and “closing”. It is usually filled by the core verbs of VP in syntax analysis tree.

“Degree” indicates how much of an environmental attribute has changed. Normally it is populated by the AVP part of the parse tree. The combination method is the relationship between several slots obtained through syntactic analysis, which represents the semantic information contained in the syntax analysis to combine the semantics of the sentences. Through the understanding of the statement, the five grooves are filled. It can be further generated in the form of code. There are certain connections between the first four grooves, which can be conjectured or supplemented by each other in certain circumstances. If the state is filled with “Brightness”, the controlled object is likely to be a “light”. Some specific “lights” can be changed by two kinds of “temperature” or “Brightness”. The node address of the controlled object is analyzed by first, second, third slot values, and the

Table 3 Semantic template

Slot name	Explanation
Object	Node
Location	The difference between the same type of object
State	Expected change status
Direction	Increase or decrease status value
Degree	The magnitude of the status adjustment
Formula	Sentence structure brings the meaning of information

desired selected gear is analyzed by fourth, fifth grooves. For example, the input sentence "gives me a little bit of light in the kitchen", and its semantic analysis results in (Table 4).

3.3 Lexical Analysis Module

Because there are no gaps between Chinese words, each sentence appears in the form of strings. Therefore, lexical analysis is the first step to prepare for subsequent processing.

The purpose of lexical analysis is to make the computer understand the words that appear in the sentence and the simple surface meaning of the word. In order to make the computer have this ability, the dictionary is first needed, and then the lexical analysis is carried out on the input sentences based on the dictionary computer. Therefore, this section follows this order. First, the establishment of the dictionary is introduced. Then, the algorithm of lexical analysis based on a dictionary is introduced.

The process of setting up a dictionary: first, the words related to the smart home products are extracted. According to the basic standard of processing modern Chinese corpus, the words are tagged. The most commonly used types of speech involved in this article are summarized in Table 5.

Based on these parts of speech, the following describes the process of modeling various parts of speech.

In the non-restrictive field, the result of the positive maximum matching method has reached nearly 1/200 error rate. In the field of smart home control, the accuracy of the maximum forward matching principle will be greatly improved, which is sufficient to meet the requirements of the application. Therefore, the general idea of the forward maximum matching principle is used to design the algorithm.

According to the established dictionary, the segmentation algorithm applies the principle of positive maximum matching, divides the longest word in the dictionary, and annotate the word character. The part without storage is filtered as a stop word.

Based on the above idea, in the process of removing the stop word, the maximum matching segmentation algorithm is adopted. If the length of the segmentation word is 1 or the word does not exist, the algorithm needs to carry out the process of L query word library, and the time cost of the process is large. In order to solve this problem, an improved forward maximum matching algorithm is proposed. By using hash lookup, the algorithm can quickly locate the characteristics of the key words. The entire matching process requires only once word library search, which reduces the number of search and increases the speed of the algorithm. The following is an introduction to the preprocessing method of the dictionary and the process of the specific segmentation of the algorithm.

The flow chart of the algorithm is shown in Fig. 5.

Table 4 An example of a semantic template

Slot name	Semantic value
Object	Light
Location	Kitchen
State	Brightness
Direction	+
Degree	A little
Formula	S- > VPNP

Table 5 The part of speech and examples in this article

Mark	Name	Example
n	Noun	Light
v	Verb	Turn on
a	Adjective	Bright
d	Adverb	Slightly
m	Numeral	Half
s	Place words	Kitchen
u	Particle	Grasp

3.4 Syntax Analysis Module

Similarly, it is similar to the process of human language learning, which requires some knowledge of relevant grammar. Therefore, the related syntax needs to be stored in the computer. Therefore, first, the syntax representation method and the creation of the grammar library are introduced. Then the algorithm of syntactic analysis for applied Grammar Library is introduced.

In the overall design of the algorithm, the input of the grammar analysis is a word string that is labeled part of speech. This article does not consider sentences and sentences in the law. For ease of expression, it is described in natural language during the example. In fact, in the process of the algorithm, the natural language has not appeared in parsing, but the separated word strings. After the above steps, the algorithm has processed the input original sentence into a semantic block expressing a certain state. At the same time, the relationship between concepts and concepts is analyzed, and the semantic blocks are filled in the corresponding semantic slots. For example, the semantic processing template for the sentence "slightly lit up the kitchen light" is shown in Table 6.

The intelligent home control instruction can obtain the corresponding code only by analyzing the address (id) of the target node and the selection stall. Therefore, according to the semantic expression of the syntax template, the corresponding instruction code can be obtained by combining the knowledge base. The node that matches the input text is stored in the candidate node set Node Aggregate, and the set Node Aggregate is increased or decreased through matching of the slots. Then, it matches the operations that conform to the input text. From the node mapping set code Set, the semantic - compliant operation instruction is selected, and the candidate instruction set Instructions Aggregate is filled. After the matching operation of each slot, the semantic instruction is selected as the answer.

4 Algorithm Test

In terms of computation time, the length of the statements in the collection varies greatly, it inevitably leads to a great difference in the running time of the word segmentation algorithm, which is not convenient for observing the experimental results. Therefore, the statements in the test set are divided into 12 groups. Each group includes 10 sentences. The total number of Chinese characters contained is approximately the same. For each of these 12 sets of data, two methods are used to perform word segmentation, respectively. To obtain the time required by each of the two algorithms for each sentence, an average of the

Fig. 5 Flow chart of segmentation algorithm

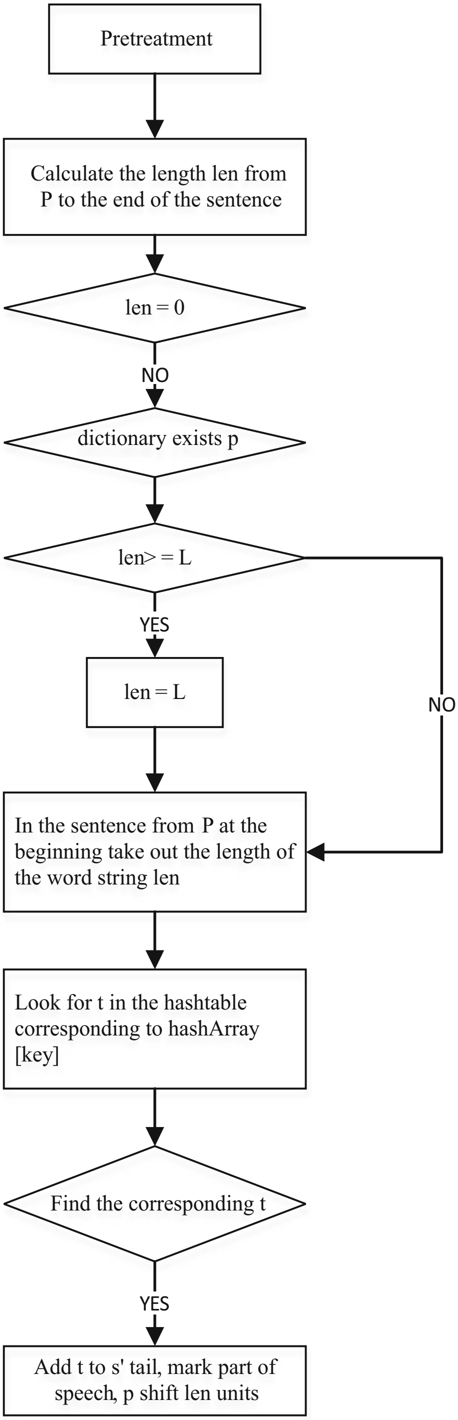

Pretreatment
Calculate the length len from P to the end of the sentence
len = 0
NO
dictionary exists p
len> = L
YES
NO
len = L
In the sentence from P at the beginning take out the length of the word string len
Look for t in the hashtable corresponding to hashArray [key]
Find the corresponding t
YES
Add t to s' tail, mark part of speech, p shift len units

Table 6 Semantic processing stage semantic template

Slot name	Semantic value
Object	Light
Location	Kitchen
State	Brightness
Direction	–
Degree	A little
Formula	S- > VPNP

time spent on the two algorithm words in each group of data is calculated. For the convenience of observation, the data is plotted with a line chart by Matlab, and the result is shown in Fig. 6.

As can be seen from Fig. 6, the time required for the improved algorithm is significantly less than before. After analyzing the experimental data of each group, it is found that there are fewer stop words in the experimental data in the eleventh group, so the running time of the algorithm is relatively small. It is also found that stop words have a great influence on the running time of traditional word segmentation algorithms. Because every time a stop word appears, the traditional algorithm needs to traverse the thesaurus, which consumes a lot of time. However, the improved method based on hash list can remove the stop word after one operation, so the stop word has a relatively small influence on the operation time of the algorithm. This is evident from the other group data. At the same time, the improved word segmentation algorithm outperforms the traditional algorithm before improvement, even in the 11th group with fewer stop words. The improved algorithm can effectively improve the efficiency of the algorithm while ensuring the accuracy of the original algorithm. To a certain extent, it proves the feasibility of the improved algorithm.

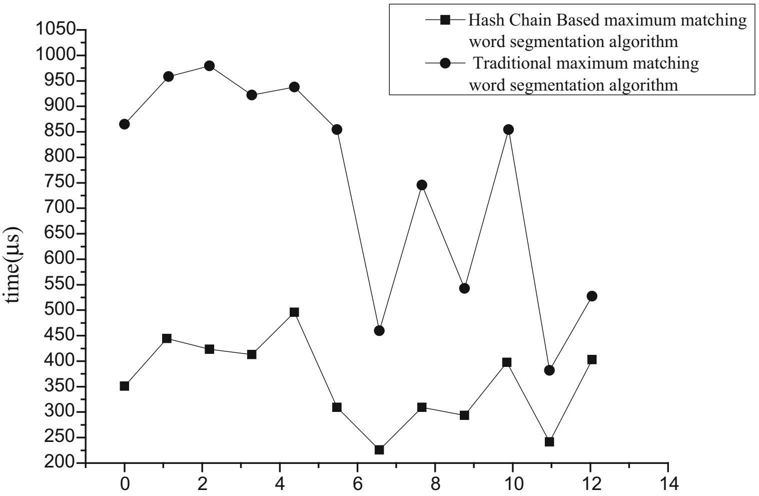


Fig. 6 Experimental results: The horizontal axis indicates that the test set is divided into 12 groups. The vertical axis represents the average time spent on each group of words

5 Conclusion

Control instructions in the form of smart home text are translated into computer-readable code. After reading natural language processing, the algorithm is divided into three stages: lexical analysis, syntax analysis and semantic understanding. By reading and studying the algorithms which may be involved, the corresponding algorithm idea is finally chosen. First, the implementation of these algorithm ideas is completed. In this process, the semantic knowledge base, such as the word library, the Grammar Library and the mapping set, is set up in accordance with the needs of the actual application.

In the process of using the segmentation algorithm, the algorithm of the lexical analysis is improved. Compared with the original algorithm, the accuracy of the algorithm is guaranteed, and the efficiency of the algorithm is improved to some extent. In the process of syntactic analysis, according to the characteristics of smart home, several rewriting rules are selected in the 33 rules of grammatical rewriting commonly used in natural language. In the semantic analysis stage, a matching template suitable for this paper is established by combining the idea of template matching semantic analysis. At the end of the experiment, the overall verification of the design algorithm is completed.

In the process of testing, it is found that the natural language processing method is used to deal with the spoken language. In some few cases, there will be an ambiguous problem that cannot be eliminated. In the future study, combined with machine learning knowledge, this problem can be solved. At the same time, with the further research on smart home, thesaurus and semantic templates also need to be reasonably expanded and adjusted in order to achieve the requirements of language control for smart home nodes.

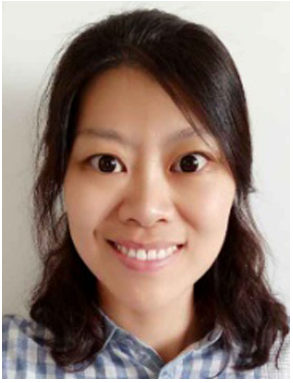
Acknowledgements The authors acknowledge the follow projects:(1) Science and Technology Department of Jilin Province, countermeasure research on current situation and promoting transformation of scientific and technological achievements of the provincial universities Changjitu area (20170418035FG); (2) Beihua University, the professional application of the “C programming language” case teaching practice (XJQN2016035).

References

1. Allen, L. K., Snow, E. L., & McNamara, D. S. (2015, March). Are you reading my mind?: modeling students’ reading comprehension skills with natural language processing techniques. In *Proceedings of the fifth international conference on learning analytics and knowledge* (pp. 246–254). ACM.
2. LeCun, Y., Bengio, Y., & Hinton, G. (2015). Deep learning. *Nature, 521*(7553), 436–444.
3. Burger, G., Abu-Hanna, A., de Keizer, N., & Cornet, R. (2016). Natural language processing in pathology: A scoping review. *Journal of Clinical Pathology, 69*(11), 949–955.
4. Poria, S., Cambria, E., Hussain, A., & Huang, G. B. (2015). Towards an intelligent framework for multimodal affective data analysis. *Neural Networks, 63,* 104–116.
5. Agarwal, B., Mittal, N., Bansal, P., & Garg, S. (2015). Sentiment analysis using common-sense and context information. *Computational Intelligence and Neuroscience, 2015,* 30.
6. Pons, E., Braun, L. M., Hunink, M. M., & Kors, J. A. (2016). Natural language processing in radiology: A systematic review. *Radiology, 279*(2), 329–343.
7. Fisher, I. E., Garnsey, M. R., & Hughes, M. E. (2016). Natural language processing in accounting, auditing and finance: A synthesis of the literature with a roadmap for future research. *Intelligent Systems in Accounting, Finance and Management, 23*(3), 157–214.
8. Lin, J. R., Hu, Z. Z., Zhang, J. P., & Yu, F. Q. (2016). A natural-language-based approach to intelligent data retrieval and representation for cloud BIM. *Computer-Aided Civil and Infrastructure Engineering, 31*(1), 18–33.

9. Heift, T. (2017). History and key developments in intelligent computer-assisted language learning (ICALL). In S. Thorne & S. May (Eds.), *Language, education and technology. encyclopedia of language and education* (3rd ed.). Cham: Springer.
10. Hearst, M. A. (2015). Can natural language processing become natural language coaching? *ACL, 1,* 1245–1252.
11. Lee, L. H., Yu, L. C., & Chang, L. P. (2015). Guest editorial: special issue on chinese as a foreign language. *Computational Linguistics & Chinese Language Processing, 20*(1), i–v.

Zeliang Zhang Doctor of Engineering, Associate Professor. Graduated from Jilin University in 2012. Worked in Beihua University. His research interests include pattern recognition and natural language processing.



Xinwen Bi Master of Engineering, Lecturer. Graduated from Beihua University in 2010. Worked in Beihua University. Her research interests include pattern recognition and intelligent system.