Semantic Comprehension of Questions in Q&A System for Chinese Language Based on Semantic Element Combination

JIAXING SONG\textsuperscript{1,2}, FEILONG LIU\textsuperscript{1,3}, KAI DING\textsuperscript{2}, KAI DU\textsuperscript{4}, AND XIAONAN ZHANG\textsuperscript{1}

\textsuperscript{1}College of Field Engineering, Army Engineering University of PLA, Nanjing 210007, China
\textsuperscript{2}Science and Technology on Near-Surface Detection Laboratory, Wuxi 214000, China
\textsuperscript{3}31101 Troops, Nanjing 210016, China
\textsuperscript{4}31104 Troops, Nanjing 210016, China

Corresponding author: Kai Ding (winfast113@sina.com)

This work was supported by the Open Project of the Key Laboratory of Near-Surface Detection Technology under Grant TCGZ2015A002.

\textbf{ABSTRACT} Question understanding is an extremely important part of the question and answer (Q&A) system, which is the basis of subsequent information retrieval and answer extraction. In order to improve the semantic understanding of question, we propose a method of understanding the question based on the combination of semantic element for Chinese language. Firstly, the method uses lexical and rules recognition to extract the semantic elements of questions and recognizes the functions according to the preprocessing pattern. Secondly, combining the dependency analysis tree structure of the question and the function type, it can produce the semantic elements. Finally, a normalized semantic expression was generated. We extract the user questions in Baidu-Zhidao as the test set. The result shows that the accuracy of the proposed method is 97.8\%, so the semantics of the Chinese language questions can be effectively understood by this method.

\textbf{INDEX TERMS} Semantics, system analysis and design, logic testing.

\section{I. INTRODUCTION}
Nowadays, the focus of competition in information society is no longer how much the information we have, but how to get the needed information in the fastest speed and the most accurate way. Question and answering (Q&A) system is a kind of advanced information retrieval, which can answer user questions in natural language accurately and concisely. From the perspective of the system framework, the Q&A system is generally composed of three main parts: question comprehension, information retrieval and answer extraction \cite{1}. Question comprehension, as the primary part, is the premise and basis of the whole Q&A system, which has a crucial influence on the performance of the whole Q&A system. Namely, wrong question comprehension can lead to wrong answers certainly, and more than 78.1\% of errors in Q&A answer were caused by misunderstanding \cite{2}. Therefore, how to effectively and correctly understand the semantic meaning of questions and transform questions from fuzzy natural language to clear logical language is the key and breakthrough point of current research on Q&A system for Chinese language. Compared with Latin languages, the semantic understanding of Chinese question is more difficult. First of all, in the Latin languages, for example English, spaces are used as natural separators, while in Chinese, due to the tradition inherited from ancient Chinese, there is no separation between words. In addition, the boundary between word and phrase in Chinese is fuzzy from the sentence structure directly. For example, “机场有许多飞机” (“There are many plane in the airport”) is a typical Chinese sentence, and the phrase “飞机” (“plane”) consists of two independent characters which can combine with many other characters leading to completely different meaning. In English, it is clearly for merely using a word “plane” usually. Secondly, there is no direct relationship between the parts of speech and the structures that perform grammatical functions in Chinese. Namely, the Chinese sentences are much more various, and the analysis method for English sentence can be fail when applying directly to Chinese language. Thirdly, the corpus-related resources for
Chinese language are scarce and the construction is not good enough for Q&A system researching, and in the experiment, it is difficult to find a suitable questions set. Therefore, studying the semantic understanding of Chinese language is challenging and urgent.

Aiming at the semantic understanding of questions in Q&A system for Chinese language, this paper focuses on the transformation of natural language questions into normalized logical expressions. We start from the dependency analysis structure of questions and put forward the question understanding method based on the combination of semantic elements, as showed in Figure 1. Due to the lack of similar research reference, the input and output of this paper are natural language questions and combined semantic expressions, which are converted into semantic expressions for in-depth analysis. Therefore, firstly we need to identify the concept, entity, relationship, attribute, attribute value, and functional category corresponding to the question (pre-processing). Then, according to the dependency analysis structure and pre-defined combination rules, the identified semantic elements are combined. Finally, a structured question semantic expression is generated. Figure 1 illustrates the processing process with “波音 747 和空客 A380 相比哪个快” (“Which is faster, the Boeing 747 or the Airbus A380?”) as an example. The main work can be divided into the following three aspects: 1) collect and analyze a large number of user questions in related and relevant fields, summarize their dependent syntactic structure and language characteristics, and divide the lexical structure of the questions into eight complex functions and one basic type; 2) design a variety of complex attribute recognition rules to solve the complex attribute recognition in the user questions neglected in the current research; 3) creatively put forward the method of semantic element combination to understand the question semantics, take semantic elements as the basic unit, design basic rules and function rules to combine the basic unit, and generate logical expressions to obtain the user’s query intention.

### II. RELATED WORK

At present, the research on semantic comprehension of questions in Q&A system mainly focuses on the following four types for Chinese language: the method based on keyword matching, the method based on template, the method based on semantic relationship extraction, and the method based on semantic analysis.

Keyword matching method [3], [4] takes the set of keyword construction as the core. Firstly, the keyword phrases in the question are extracted. Then, the dictionary-based or statistical method is used to expand the keyword phrases, and a series of retrieval expressions are also generated to express the semantic meaning of the question. Namely, a series of retrieval expressions are generated to express question semantics centered on the extended keyword set. Simple keyword matching can only understand the semantics of the questions at a superficial level, which is difficult to deeply understand the user’s query intention.

The method based on template [5]–[8] takes the pattern templates as the core, matching the questions to pattern templates in the pre-built template library at the first place. Then the semantic information of the question is extracted directly from the pattern template. This method does not involve complex semantic analysis, and can directly analyze users’ query intention to obtain semantic information. However, this method needs to construct a sufficient number of templates in advance.

The method based on semantic relationship extraction [9], [10] takes relation as the core and extracts multiple semantic relation triples. By associating the nodes in the triples that point to the same entity, the combined semantic relation triples represent the question semantics. This method does not need to construct the question template in advance and can solve a large range of questions, but its query ability for complex question is still relatively weak, and it is difficult to achieve effective reasoning. Besides, its accuracy also needs to be further improved.

The method based on semantic analysis [11]–[13] focuses on normalized semantic expressions. The semantics of the question are analyzed by using the pre-designed semantic analysis method, and the question is transformed into a normalized semantic expression. This method adopts the logical normative semantic representation as the analytical form of the question, with clear logical relationship and semantic understanding of reasoning and solving complex question. Traditional semantic analysis [11] relies on manually annotated logical word lists, which can only carry out supervised machine learning in a small range. But it is difficult to deal with samples that have not been learned in supervised learning. In order to extend the application scope, a semi-supervised model is proposed to extend semantic analyzer [12]. This method firstly trains the supervised semantic analyzer through a predefined logical vocabulary.
Then, the scope of semantic analyzer can be improved by expanding the vocabulary in the original semantic analyzer, which still relies on manually annotated logical vocabulary. Therefore, a new semantic analysis method is proposed to get rid of manual annotation [13], [14]. The multiple semantic expressions are obtained by using the DCS language. By designing the characteristic function and loss function, the existing question-answer pairs are utilized, and the training models are selected. Then, the semantic expressions are chosen to get the correct answer. Clearly, its accuracy depends on the quantity and quality of the question-answer pairs.

In this paper, the question comprehension method based on semantic element combination is proposed for Chinese language, which belongs to a method based on semantic analysis. This method can deeply understand the semantics of complex user questions by means of logical expressions, and not depend on a large number of sentence modules constructed manually. A normalized semantic expression of the question is directly generated through semantic element analysis and semantic element combination. It not only avoids the limitation of logical vocabulary on the application scope of the system, but also does not depend on the difficult collection of question-answer to the training model, so as to effectively solve the semantic understanding of complex questions in the Q&A system.

**III. OVERVIEW OF METHODS**

The set of Chinese natural language questions \( Q: \{ q_i \mid q_i \in Q \} \) is given in Q&A system. The method in this paper accepts any input \( q_i \), and then outputs the corresponding semantic expression \( s_i \) after understanding the question, where \( S: \{ s_i \mid s_i \in S \} \) is the set of semantic expression of the questions.

The question comprehension method based on semantic element combination can be divided into the following three stages, as shown in Figure 2.

1. **Pre-processing:** word segmentation, correction, part of speech tagging, stop words, dependency analysis;
2. **Semantic element recognition:** semantic elements in the question are extracted and labeled based on lexical database recognition and rules;
3. **Semantic element combination:** First the types of functions contained in the question are identified, and then according to the basic semantic element combination rules and functions, the semantic elements are recombined to generate the semantic expression of the question.

**IV. QUESTION COMPREHENSION METHOD BASED ON SEMANTIC ELEMENT COMBINATION**

**A. PRE-PROCESSING**

Before the semantic analysis of the question, pre-processing the question is necessary. The pre-processing process mainly includes the following parts: word segmentation, part of speech tagging, correction, deletion of stop words, dependency analysis.

1. **WORD SEGMENTATION, PART OF SPEECH TAGGING**

   Word segmentation and part of speech tagging are the first stage of question comprehension, which should be the basis of subsequent processing. For Chinese question, the question text consists of consecutive word sequences. Namely, there is no natural separator between words. To correctly understand the semantics of user questions, identifying the boundaries between words accurately is necessary, which is the word segmentation. Part of speech tagging assigns an appropriate part of speech to each word in the question, such as noun, verb, and adjective. In essence, it is a typical sequence labeling issue.

2. **CORRECTION**

   In the normalized form, there are not all of the questions asked by users. For example, the questions are interspersed with Pinyin, wrong characters, mixed Chinese and English, Arabic numerals in Chinese. Non-canonical forms may lead to errors in semantic element identification. So fixing the user question is an important process. It is possible to normalize the wrong way into the correct way through correction, solving the misunderstanding of the question caused by the form error.

3. **DELETION OF STOP WORDS**

   Stop words refer to auxiliary words, modal words and partial pronouns that have no practical help in understanding a sentence, which often appear in Q&A system. If these function words are not filtered out, the effect and efficiency of question comprehension will be affected negatively. There are two ways to filter stop words, part of speech approach and a lexicographical approach. In practice, most of function words can be filtered out by using the first approach. But if the first approach doesn’t work, the latter approach to filter can be used.

4. **DEPENDENCY ANALYSIS**

   The purpose of dependency analysis is to reveal the syntactic structure of a sentence by analyzing the interdependence among its components. The central element in a sentence that governs the other components is the central verb, and the
other component is in some way subordinate to the dominant. Through the dependency analysis, the semantic modification relation among sentence components could be found out, and the long-distance collocation information between sentence components could be obtained.

In this paper, LTP dependency analysis tool is adopted, and the complex attributes and semantic elements of composite questions are identified by the dependency analysis structural tree. Taking “波音747和空客A380相比哪个快” (“Which is faster, the Boeing 747 or the Airbus A380?”) as an example, LTP dependency analysis is conducted, and the results are shown in Figure 3.

B. SEMANTIC ELEMENT RECOGNITION

The purpose of semantic element recognition is to extract the semantic elements in the question and mark their types, which can provide support for the subsequent combination of semantic elements. There are two usual methods for identification. The first method is to construct relevant thesaurus for recognition. The advantage of this method is that as long as the information of thesaurus is comprehensive enough, its recognition accuracy is very high, while the drawback is that it is pretty difficult to expand thesaurus indefinitely. The other is to make rules and algorithms for identification. The advantage of this method lies in its generality, which can identify complex semantic elements with high coverage rate and wide identification types while the drawback is the limited accuracy. In this paper, semantic elements are identified by the combination of lexicon and rule. Firstly, semantic elements are identified by thesaurus, while words that fail to be identified by thesaurus are identified by rules. This method effectively guarantees the accuracy and coverage of the recognition.

1) SEMANTIC ELEMENT DEFINITION

In this paper, the data types of the knowledge base include entity (E), concept (C), attribute (A), relation (R) and attribute value (V) [15], [16]. Among them, the entity (E) is the most basic element in the knowledge base, such as “波音747” (“Boeing 747”). Concept (C) are the collections of the same class of entities, such as “飞机” (“plane”). Attributes (A) refer to the attributes that the entity or concept may have, while the attribute value (V) represents the value of the specific attribute. For example, in the Chinese sentence “波音747的巡航速度是900千/小时” (“The Boeing 747 has a cruising speed of 900 km/h”), the attribute of “波音747” (“Boeing 747”) is “巡航速度” (“Cruising speed”) and the attribute value is “900 千米/小时” (“900km/h”). Relationship (R) refers to the association between the entity and concept. For example, in Chinese sentence “波音747装备GenX涡轮发动机” (“Boeing 747 equipped with GenX turbine engine”), the word in Chinese sentence “装备” (“equipped”) shows the relationship between the two entities, “波音747” (“Boeing 747”) and “GenX涡轮发动机” (“GenX turbine engine”).

Based on the data types of the above knowledge base, depending on the type of attribute value, additional $V_d$ and $V_o$ represent numeric attribute value and object attribute value ($\{V_d, V_o\} = V$). Then, we defines the following two categories of semantic element forms: \{e, p, e, C, A, R, V, V_d, V_o\} are used in total.

Unary semantic element: $e$ and $V$, and $e \in \{E, C\}$

Binary semantic element: $p, p \in \{A, R\}$. The main difference between $A$ and $R$ is: $A = \{A\} < e, A, V >$, $R = \{R\} < e, R, e >$

If the knowledge base triples are represented as < subject, 'relation', object >, the unary semantic elements are set: \{ subject, object \} while the binary semantic elements are set: \{ 'relation' \} ('Relation' is a generalized concept, including attributes and relationships)

2) THESAURUS RECOGNIZES SEMANTIC ELEMENT

Lexical recognition is the first step of semantic element recognition. According to the types of semantic elements to be recognized, each type is established its own lexicon. The thesaurus used in this paper includes the following parts:

a: DOMAIN THESAURUS

This is the most important part of the lexicon, which is extracted from the elements in the knowledge base, including entity library, concept library, relation library, attribute library and attribute value library, and the above libraries respectively store the entities, concepts, relations, attributes and attribute values. To be sure, the library of attribute value mainly refers to attribute value library for the object type.

b: GENERAL WORD THESAURUS

To solve the difference between user vocabulary and knowledge base vocabulary, general word thesaurus is mainly used to improve the accuracy of the stage of high score word, and it can also ensure the feasibility of subsequent word recognition. This paper adopts the general word thesaurus in the LTP word segmentation program of Harbin Institute of Technology.

c: SYNONYM THESAURUS

This part represents a synonymy extension of the domain thesaurus. Synonyms are often used in the question expression by users in practice, so it is necessary to expand the synonyms of the domain thesaurus to improve the coverage and accuracy of recognition.

FIGURE 3. The results of LTP dependency analysis.
Some users ask questions with partial acronyms, abbreviations or aliases, for example, the nickname of “波音” (“Boeing”) in Chinese is “珍宝客机” (“Jumbo jet airplane”). Therefore, it is necessary to collect relevant acronyms, abbreviations and aliases in advance, and establish the corresponding user thesaurus of the domain thesaurus for expansion.

The main steps to identify semantic elements in the lexicon can be summarized as followed: receive the preprocessed question, identify and marks \{E, p, C, A, R, V, V_d, V_o\} based on established lexicon, where the subcategory and the parent category can be marked a same semantic element simultaneously, for example, “波音747” (“Boeing 747”) is marked as <e, E >.

3) RULES IDENTIFICATION FOR SEMANTIC ELEMENT

Although thesaurus recognition can cover many semantic elements, some complex semantic elements are difficult to be identified, such as “有多远” (“How far”), whose corresponding attribute A is “distance” (“Distance”). Due to the complexity of the structure, these semantic elements are difficult to be recognized by thesaurus, and their recognition requires the help of part of speech, predefined text structure, text pattern. In this paper, the following rules are formulated to assist semantic elements of thesaurus recognition:

Recognition of Entity E, Concept C, Relation R:

Entity identification and relationship extraction are the key technologies in information extraction. Named entity recognition based on conditional random field (CRF) and relationship extraction based on rule and heuristic algorithm were selected for E, C, and R identification.

Recognition of Attribute Value V:

The attribute value V to be recognized consists of object type attribute value V_o and numeric type attribute value V_d. The method of lexicon recognition can effectively identify so many object type attribute values V_o. The partially uncovered values need to define rules for identification. Therefore, the following heuristic rules are defined to supplement the attribute value V of thesaurus recognition:

a: RECOGNITION BY PART OF SPEECH

LTP part-of-speech tagging will mark specific parts of speech of nouns, for example, the part of speech for “现在” (“At present/ now”) is “nt” (noun of time), and the part of speech for “城郊” (“outskirts of a town”) is “nl” (noun of location). Most words of this kind correspond to the attribute values V_o. Therefore, it can be marked as V_o according to the part of speech, mainly including “nl, nd (noun of direction) and ns (noun of speed)”.

b: RECOGNITION BY PATTERN

A word marked as a numeric attribute value V_d must contain some numeric data. But these words are sometimes not so obvious or direct. The value V_d has its special mode in the question, which combines the keyword mode with the dependency analysis structure. So, the following mode is defined for V_d recognition:

Keyword pattern: Numerical data + quantifier set {‘千米’, ‘米’, ‘公里’, ‘吨’, ‘斤’} \{‘km’, ‘m’, ‘km’, ‘ton’, ‘a unit of weight about half of kilogram’\}, for example, “车长5米的汽车有哪些?” (“What kinds of cars are there with a length of 5 meters?”). Dependency tree structure pattern: Attribute A - word_prep – numeric, for example, “速度_A 为 100 的汽车有哪些?” (“What kinds of cars are there with a speed_A of 100?”), where the word_prep is the preposition set {‘是’, ‘有’, ‘为’, ‘属于’} (\{‘is’, ‘have/has’, ‘be’, ‘belongs’\}).

Recognition of Attribute Type A:

Complex attribute recognition is one of the most difficult parts of semantic element recognition. From the current research, most of the papers to deal with the process are generally simple word recognition. The experimental results of this paper show that 77.08% of the errors in semantic expressions caused by complex attributes fail to recognize the combinations in question comprehension, which indicates the importance of complex attribute recognition. A large number of questions are collected from Baidu, through statistical analysis the following complex attribute recognition rules are formulated:

c: IDENTIFY ATTRIBUTES BY E/C AND PART OF SPEECH RULES

Most attributes are closely related to entity E and concept C, and there are rarely separate attribute units. Therefore, the following recognition patterns can be defined by combining part of speech and recognized E/C:

\(E/C/\text{word}_{(a)} + (\text{len(word}_{(a)}) > 1)\)

where “+” represents a combination of conditions, and “()” means it is required, “f” represents the mean of “or”. word_{(a)} represents the noun word. The meaning of the above model is that it needs to meet two conditions: 1) Specific lexical structures; 2) Chinese character length (word_{(a)}) shall be at least two. For example, “客机_E 的速度(n)是多少” (“What is the velocity/speed of the plane_E (n)”), where “的 速度” (“the velocity/speed of...”) could satisfy the above pattern, so “速度” (“velocity/speed”) can be recognized attribute A.

d: ADD OMITTED ATTRIBUTES IN CONJUNCTION WITH V_o

The attribute A corresponding to object type attribute V_o is generally omitted by the user. For example, the complete sentence of “城郊_V_o 的 机场_E 名字_A 是 什么” (“What is the name_A of airport_E in the suburb_V_o?”) should be “位置_A 在 城郊_V_o 的 机场_E 名字_A 是 什么” (“What is the name_A of the airport_E at the location_A of the city suburb_V_o?”). Omitting the attribute “位置_A” (“location_A”) will result in the semantic combination of “城郊_V_o” (“the city suburb_V_o”) and “机场_E” (“airport_E”) not being combined. Therefore, when identifying V_o, its corresponding attribute A should be attached at the same time. This paper summarizes the correspondence of additional V_o attribute A based on part of
TABLE 1. Part of speech - attribute correspondence table.

| Part of speech | Corresponding attribute A | Examples |
|----------------|---------------------------|----------|
| nd             | 方向 Direction            | 右侧(nd)的汽车是谁的？（“Who’s the car on the right (nd)?”） |
| nl             | 位置 Location             | 西郊(nl)的机场叫什么？（“What is the airport in the western suburbs (nl)?”） |
| ns             | 国籍 Nationality          | 中国的客机有哪些？（“What are the passenger planes of China (ns)?”） |
| ns             | 位置 Location             | 北京(ns)的天气如何？（“What’s the weather like in Beijing (ns)?”） |

TABLE 2. Query block-attribute corresponding dictionaries.

| Attribute A | Query block |
|-------------|-------------|
| {有多久，多久、什么时间、什么时候，何时…} | Time |
| {哪里、哪儿、何处、什么地方、什么地方…} | Address |
| {第几次，排第几，排在前、排在后…} | Ranking |
| {飞多远，能飞多远…} | Speed |
| {飞多远，能飞多远…} | Mileage |
| {有多长、多长…} | Length |
| {怎么样，多大，多大…} | Size |

speech, as shown in Table.1. The range of parts of speech is {nd, nl, ns}. For the attribute values that do not satisfy the situation in Table.1, the new attribute value - corresponding attribute \( V_o \), \( A \) need be established.

e: IDENTIFY COMPLEX ATTRIBUTES BY QUERY BLOCKS

The query block in the question implies an attribute, such as “客机有多长?” (“How long is the passenger plane?”), where the query block “有多长” (“How long”) implies the attribute “长度” (Length). Sometimes, however, the implicit attribute in the query block could be repeated with the previous, for example, “客机的长度 \( E_A \) 多长” (“What is the length \( A \) of the passenger plane?”). Therefore, if it is repeated, recognized attribute should be ignored. The complex attribute recognition of query block consists of three parts:

Part 1: query Block Confirmation:

The simple definition of the query block is the semantic block containing the query word in the question. In order to extract query block and determine the coverage of query block, this paper proposes a retrospective method based on dependency analysis tree, which mainly includes two steps:

The first step is to locate query words in the dependency analysis tree and obtain the following query words by statistics:

- Query words without real meaning: {‘怎么样’, ‘如何’, ‘怎么’, ‘怎样’, ‘那么’, ‘这么’} ({‘How’, ‘What if’}, {‘How’, ‘How’, ‘so’, ‘so’});
- Query words with practical meaning: {‘多少’, ‘多’, ‘几’, ‘什么’, ‘啥’, ‘那’, ‘哪’} ({‘How much’, ‘so’, ‘some’, ‘what’, ‘what’, ‘who’, ‘there’, ‘where’});
- Causative query word: {‘为什么’, ‘为何’} ({‘why’, ‘why’});

The second step is starting with the query word, and going back up or down to the first \( E, C, R \) or \( A \), and then stopping backtracking. The node passed in the forward output backtracking process is the query point block.

Part 2: Complex Attributes Recognition:

The identification of complex attribute is based on the obtained query block. First, the implicit attribute is extracted from the query block. Then, whether the verb is included or not in the query block need to be judged and at the same time, the extracted properties should be extended, for example, “波音 747 在哪里制造(v)” (“Where is the Boeing 747 made (v)?”), where (v) means the part of speed is a verb and “哪里制造” (“Where made”) is the obtained query block. As for query block, “哪里” (“Where”) corresponds to the attribute “地址” (address), combined with the verb “制造” (manufacture), the extended attribute is “制造 地址” (“manufacture address”), and the final returned attribute is “制造地址” (“manufacture address”). The complex attribute recognition process is as follows:

As the first step, based on the corresponding dictionary of query block-attribute, the attribute of query block is extended. In this paper, some query block-attribute corresponding dictionaries are collected, as shown in table.2.

The second step is to determine whether the question block contains a verb. If a verb exists, the query block is extended according to the attributes obtained above. The extension means is that the verb comes first and the attribute comes last. For example, the query block “多久完成(V)” (“How long to complete (v)”) → the attribute “完成时间” (“the finish time”), the query block “哪里 制造(V)” (“Where to make (v)”) → the attribute “制造 地址” (“manufacture address”).

Part 3: Attribute Repeat Judgment:

The attribute \( A' \) prime obtained by the above complex attribute recognition may duplicate the previous attribute \( A \) in the question. So it is necessary to judge the attribute repeat. The judgment process is mainly based on the dependency analysis tree, which allows us to determine whether the identified attribute \( A' \) and its most recently connected attribute \( A \) in the dependency tree structure are repeated. If repeated, the attribute \( A' \) will be ignored. If not, attribute \( A' \) will be returned and the node part of query block in the dependency analysis tree will be replaced with attribute \( A' \).
TABLE 3. Combination rules for connection relationships.

| Combination units | Dependency tree structure | Results of combination |
|-------------------|---------------------------|------------------------|
| e/V, p            | V—A / A—V                | A.V                    |
| e/V, p            | C—A / A—C                | C(e?) / A              |
| e/V, p            | E—A / A—E                | E.A                    |
| e/V, p            | C—R, R—A.C              | C(R)e?, R. R.C(e?)    |
| e, V              | e—V—A                    | e.AV                   |
| e, V              | e—V / V—A                | e.AV                   |
| e, V              | e—e / e.R—e              | e, e.R                 |
| e, V              | e.A—V / A.V—e            | e.AV                   |
| e, V              | e—A / A—V                | A.V                    |
| e, V              | e—A / A—V                | A.V                    |
| e, V              | e—A / A—V                | A.V                    |

Where “—” represents a bidirectional interdependent connection, and “→” represents a one-way dependency connection. “?” is same to the meaning of “or”.

C. SEMANTIC ELEMENT COMBINATION

The purpose of semantic element combination is to combine the identified semantic elements through certain rules to form a standardized semantic expression to represent the question semantics. In order to ensure the accuracy of semantic element combination, a combination method based on dependency analysis tree is proposed in this paper. This method uses the dependency relationship between semantic elements and the structure of dependency analysis tree to identify and replace function units. The dependency analysis tree is then simplified to eliminate the nodes of function units and semantic elements. Finally, according to the depth of the tree, this method can combine layer by layer from leaf nodes according to the rules (if leaf nodes are attributes or relationships, this method allows to skip some nodes and connect to entities or concepts) to get the semantic expression of the question. The combinations of semantic elements fall into two categories: basic combinations and functional combinations. The basic combination is executed firstly, and the function combination is executed later. Through the implementation of two types of combination, the semantic expression of the question can be accurately generated and effectively expressed.

1) BASIC COMBINATION

Basic combination is the basis of semantic element combination. It mainly performs simple concatenation and combination of unary semantic elements and binary semantic elements. The specific rules are shown in table 3:

a: CONNECTION

Connection relation is the most commonly used combination rule, which mainly involves the connection between unary semantic elements and binary semantic elements. The specific rules are shown in table 3:

b: INTERSECTION

"Intersection" involves two conditions, and the symbol is "∩", meaning that both conditions need to be satisfied at the same time. For example, "客机_C[e?]∩大飞机_C[e?]" (客机: passenger plane; 大飞机: giant aircraft) means the category of entity "e?" has both attribute "客机" ("passenger plane") and attribute "大飞机" ("giant aircraft"). The activation conditions of "∩" for the combination rule are as follows:

1. The entities "e1 e2 e3" are connected to the same entity "e?" in the dependency analysis tree, such as e1—e2—e3, and the semantic expression is e1∩e2∩e3.

2. Triples are connected to the same node. For example, e1—R—e2—A—V, the semantic expression is e1.R.e2∩e2.A.V.

3. The structure in the dependency tree is C1—C2. The semantic expression is C1[e?]∩C2[e?].

c: NEGATION

"Negation" is mainly a negation of the result of subsequent conditions or semantic combinations, which is activated by the negative keyword set with the symbol "!". For example, the semantic expression of "不配备" R 波音飞机_E 的 航司_C" (“The Airline_C not equipped_R with Boeing plane_E") is “航司 [e?] 配备! 波音飞机” (“The airline[e?] equipped! Boeing plane”). Typical negative keywords are listed as {'不是', '非', '没有', '不包括', '不包含'} ({'not', 'rather', 'yet', 'not including', 'excluded'}).

d: CLASS INCLUSION

The class inclusion contains a combination of rules between C and E, which has the highest priority in the base combination. However, only C and E that meet the activation conditions can execute the combination, and the combination result is C[E], which represents E of category C, whose activation rules are as follows:

1. The dependency tree structure is E → C and the dependency relationship is ATT;

2. E and C are adjacent to each other in the user question;

3. The dependency tree structure is E—[有], [是] —C— question word, where “[ ]” means that the contents in parentheses are optional. (有: have/has; 是: be)

e: OTHER SPECIAL STRUCTURES

Some special dependency tree structures need to be modified and/or perform special semantic combinations, mainly including the following three types:

1. E—A1—A2 is modified to A1—E—A2.

2. As for the structure of e—R1—R2, it executes special rule of semantic combination as e.R1.R2.

3. As for the structure of e—R—A, it executes special rule of semantic combination as e.R.A.

Function Combination:

Function combination is a high order operation of semantic element combination. Based on the basic combination and predefined function combination rules, it uses dependency analysis tree structure to generate semantic expression with function body to understand complex question semantics. For example, the semantic expression of “速度最快的飞机是什么” (“What is the fastest speed
plane?”) is “MEntity(飞机[e?] 速度, MAX)” (“飞机; plane; 速度; speed”), where MAX parameter represents the entity for which the function MEntity() takes the maximum, namely, the maximum attribute value. Function combination is mainly divided into the following aspects:

**Function Category:**

In this paper, the combination of functions includes the following 8 categories of functions, and there are 17 of them in total. Detailed function types and functions are shown in Table 4:

**Type Confirmation of Query Block:**

The query block implies the user’s focus, for example, “波音 747 和空客 A380 比速度哪个快?”. (“Which is faster, the Boeing 747 or the Airbus A380?”) and “波音747和空客A380相比速度快多少?” (“How much faster is the Boeing 747 compared to the Airbus A380?”). The former focuses on fast entities, corresponding to the function CompareSelect() while the latter is concerned on how much faster the velocity is, corresponding to the function CompareValue(). Therefore, it is necessary to confirm the query block type before function recognition. In combination with the keyword set, this paper defines the query block type as follows:

1. Omissible query block. Recognized condition: only the question word + [word_prep].

2. ?Value. Recognized keyword: {‘多少’, ‘几’} (‘how much’, ‘how many’), ?notValue represents it is not the query block of ?Value.

3. ?Num. Recognized condition: {多少’, ‘几’}(‘how much’, ‘how many’) + {个,架,辆,发,艘,...}(the unit in Chinese language to express the number of countable thing).

4. ?ListNum. Recognized condition: {‘多少’, ‘几’}(‘how much’, ‘how many’) + {种,类,种类,类型}(‘type’, ‘form’, ‘kind, ‘type’).

5. Others.

3). Function recognition and semantic element combination

The difficulty of function combination lies in function recognition. In order to accurately identify the functions in the questions, this paper proposes a pattern-based function recognition, which is characterized by the types and numbers of semantic elements in the questions, such as the types of query blocks, keyword set, structure of dependency tree, and the dependency relationship, and formulates the rules of semantic element combination according to different functions.

In order to obtain the ideal function combination effect, this paper analyzes and summarizes a large number of user questions. Based on the features of query block type and dependency relationship, a set of accurate recognition patterns covering most users’ question expressions is established for each function. And the following symbols are set to accurately describe the pattern structure: "[ ]" means optional, "()" means required, and "" is the same to the meaning of "or". "+" represents a combination of multiple conditions. "-" represents a bidirectional dependency connection while "→" "←" represents a one-way dependency connection. "NumE" represents the number of entities E in the question. Taking comparative functions (belongs to comparison class) as an example, the function recognition pattern and semantic element combination rules established in this paper are as follows:

**CompareBase(E.A/A):**

Function performance: Basic comparison class function outputs multiple E’s respective attribute comparisons. If there is an attribute input, only that attribute is compared, otherwise all attributes of E are compared. For example, the semantic expression of “波音 747 和空客 A380 相比速度快多少?” (“How much faster is the Boeing 747 compared to the Airbus A380?”) is CompareBase(波音 747, 空客 A380) (Boeing 747, Airbus A380);

Recognized rules: there are four following conditions: NumE >= 2, key_cmp, [C/A] and not satisfy CompareSelect() and CompareValue() functions.

Positioning to replace: key_cmp is replaced by CompareBase()

Combination rules: The combination rules are shown in Figure 4 based on the dependency tree structure. Attributes are generally associated to entity E by a basic combination after they are joined to the entity. Therefore, the position of attribute A is not marked in the Figure 4, only the entity structure associated with the function is marked.

**CompareSelect(E.A,0/1):**

Function performance: Output E after attribute comparison if there are attributes, otherwise attribute ‘rank’ is used by default. 0/1 means that the output is the larger E or the smaller E. Words_high / words_cmp corresponds to 1 while words_low corresponds to 0.

Recognized rules:

(1) NumE >= 2 + [key_cmp] + [‘哪个’,‘哪’,‘那一个’,‘谁’,‘哪那种’] + [A] + [‘更’] + words_cmp/words_high/words_low + with query block;
A. EXPERIMENTAL SETUP

Since there is no suitable test set similar to QALD in Chinese, in order to test the question comprehension method, this paper collects question sets by means of Baidu Zhidao. Firstly, the entity set and the entity-relation set were constructed by combining the knowledge base, and the question set was retrieved in Baidu Zhidao. Then, based on this question set, it was searched again in Baidu Zhidao to collect synonymous questions. Finally, the extraneous and repetitive questions were eliminated. A total of 8,732 questions were collected, and 6,215 questions were obtained after removing questions with similar structures. 1000 questions were randomly selected as the final test set, and 5,215 questions were selected as the auxiliary analysis set.

In terms of experiment, the language technology platform (LTP) [17] tools of Harbin Institute of Technology does not support additional thesaurus to assist word segmentation, so this paper uses the Jieba word segmentation tool is adopted, and semantic elements are added to identify all thesaurus. Jieba word segmentation uses dynamic programming to query the maximum probability path to find the maximum short combination based on word frequency. For unregistered words, it uses HMM model based on Chinese character word-forming ability to identify them. The language technology platform (LTP)-3.4.0 is adopted to analyze the part-of-speech tagging and dependency, whose support vector machine (SVM) is selected as the basic classifier to tag part-of-speech, with relatively high labeling accuracy. Additionally, for the problem of data sparsity, especially the unregistered words identified in the word segmentation stage, it introduces the characteristic of Chinese characters' partial radicals which has good generalization ability. The dependency analysis adopts the decoding algorithm based on columnar search and the two-stage syntactic analysis method based on punctuation. The efficiency of syntactic analysis is greatly improved without loss of accuracy. The thesaurus is an extension of the thesaurus forest of the social computing and information retrieval research center of Harbin Institute of Technology.

B. EVALUATION STANDARD

First, the semantic expression $F_m$ of each test question is given manually. Then the semantic expression $F_c$ is obtained by using the method in this paper. Then we compare the string similarity between $F_m$ and $F_c$. If it is exactly the same, $F_c$ is considered correct; otherwise, it will be judged manually by 7 people. When more than half of the people agree with $F_c$, $F_c$ is correct; otherwise, it is wrong. The experimental effect was evaluated by the accuracy rate $P$. The calculation formula is as follows:

$$P = \frac{\text{The number of correct semantic express}}{\text{Total number of test questions}}$$  (1)

C. RESULTS AND ANALYSIS

Experimental analysis was conducted on 1000 test questions, and 4 groups of experiments were set for comparison according to the different methods of semantic element recognition. The first group (1#) only used the thesaurus to identify semantic elements. The second group (2#) used thesaurus...
TABLE 5. Experimental results of different semantic element recognition methods.

| Experimental number | Correct number | Accuracy % |
|---------------------|---------------|------------|
| 1#                  | 774           | 77.4       |
| 2#                  | 865           | 86.5       |
| 3#                  | 904           | 90.4       |
| 4#                  | 978           | 97.8       |

FIGURE 5. 1000 test questions category statistics.

FIGURE 6. The accuracy of understanding different types of function questions.

+ synonym extension to identify semantic elements. The third group (3#) used thesaurus + synonym extension + named entity recognition to identify semantic elements, and the fourth group (4#) used thesaurus + synonym extension + named entity recognition + complex attribute recognition to identify semantic elements. The experimental results are shown in table 5.

The experimental results show that the highest accuracy of the question comprehension based on semantic element combination in this paper is about 97.8%. Even using a simple thesaurus to identify semantic elements, the accuracy rate can reach 77.4%. It can be considered that this method can effectively generate normalized semantic expressions for users’ natural language questions.

To further analyze the cause of the error, 1000 test questions were counted by function category, as shown in Figure 5. (The base combination class does not contain any function types). The effect of fourth group of experimental methods is shown in Figure 6.

Through experiments, it can be found that the forth method can completely and accurately understand the questions of concept class, basic combination class and calculation class. For comparison class, reason class and choice class, the effect is slightly worse. The reasons are as follows:

1) Dependency analysis errors cause semantic composition errors. The questions of concept class, basic combination class and calculation class do not involve complex dependency analysis structure, so they are not affected by it. However, other functions are more dependent on the dependency analysis structure and are more affected by it.

2) The complex functions are recognized wrongly, for example, the comparison functions, which have complex identification rules.

3) The user’s questions are too complicated. Some user questions may contain multiple sub-questions and complex referential relationships, leading to errors in semantic element recognition.

VI. CONCLUSION

This paper proposes a question understanding method based on semantic element combination. Through semantic element analysis and semantic element combination, the semantic semantics of the questions are understood and the normalized semantic expressions of the questions are directly generated. It not only avoids the limitation of the logical vocabulary on the scope of application of the system, but also does not depend on the difficult collection of question-answer training model. The user questions extracted from Baidu Zhidao are used as test sets for experiments. The results showed that this method had high accuracy in semantic understanding. It can solve the difficulty of question semantic analysis in Q&A system, and can also understand and reason complex questions effectively. The accuracy rate of all types of questions on the test set reached 97.8%, among which the accuracy rate of non-function types reached 100% and that of function types reached 96.8%. This paper tries to give the relevant researchers a potential method which can be effective to promote the in-depth development of question semantic understanding and Q&A system, and lays the foundation for the next step to convert the logical expression into a machine language (such as SPARQL), and further execute the corresponding handler to achieve a more intelligent knowledge data.

Although the method in this paper can understand the vast majority of questions, some complex questions are still misunderstood. Through careful analysis of the reasons affecting the experimental results, it is found that most of these errors are related to the function pattern and referential relationship.

REFERENCES

[1] X. L. Mao and X. M. Li, “A survey on question and answering system,” J. Frontiers Comput. Sci. Technol., vol. 6, pp. 193–207, 2012.
[2] D. J. Ye, “Question analysis in automatic question and answer system,” Ph.D. dissertation, Tianjin Univ., Tianjin, China, 2010.
[3] Y. Chen, “Research and realize of Chinese question answering system based on fact questions,” Ph.D. dissertation, Shanxi Univ., Taiyuan, China, 2006.

[4] K. Jia, X. Pang, Z. Li, and X. Fan, “Query expansion based on semantics and statistics in Chinese question answering system,”, Wuhan Univ. J. Natural Sci., vol. 13, no. 4, pp. 505–508, Aug. 2008.

[5] D. S. Wang, “A domain-specific question answering system based on ontology and question templates,” in Proc. 11th ACIS Int. Conf. Softw. Eng., Artif. Intell., Netw. Parallel/Distrib. Comput., Jun. 2010, pp. 151–156.

[6] X.-H. Zhu, Q.-H. Cao, and F. F. Su, “A Chinese intelligent question answering system based on domain ontology and sentence templates,” Int. J. Digit. Content Technol. Appl., vol. 5, no. 11, pp. 158–165, Nov. 2011.

[7] X. H. Zhu, C. Gong, and H. C. Chen, “Construction and match of question templates based on domain ontology and semantic block,” Appl. Mech. Mater., vols. 373–375, pp. 1776–1779, Aug. 2013.

[8] R. Wilensky, D. N. Chin, M. Luria, J. Martin, J. Mayfield, and D. Wu, “The Berkeley UNIX consultant project,” Artif. Intell. Rev., vol. 14, pp. 43–88, 2000.

[9] X. Yao and B. Van Durme, “Information extraction over structured data: Question answering with freebase,” in Proc. 52nd Annu. Meeting Assoc. Comput. Linguistics, vol. 1, 2014, pp. 956–966.

[10] A. Bordes, J. Weston, and N. Usunier, “Open question answering with weakly supervised embedding models,” in Proc. Eur. Conf. Mach. Learn. Knowl. Discovery Databases. New York, NY, USA: Springer-Verlag, 2014, pp. 165–180.

[11] C. Unger and P. Cimiano, “Pythia: Compositional meaning construction for ontology-based question answering on the semantic Web,” in Proc. Int. Conf. Appl. Natural Lang. Inf. Syst., in Lecture Notes in Computer Science, vol. 6716, 2011, pp. 153–160.

[12] Q. Cai and A. Yates, “Large-scale semantic parsing via schema matching and lexicon extension,” in Proc. 51st Annu. Meeting Assoc. Comput. Linguistics, vol. 1, 2013, pp. 423–433.

[13] Y. E. Adlouni, H. Rodríguez, M. Meknassi, S. O. El Alaoui, and N. En-Nahnahi, “A multi-approach to community question answering,” Expert Syst. Appl., vol. 137, pp. 432–442, Dec. 2019.

[14] J. Berant and P. Liang, “Semantic parsing via paraphrasing,” Meeting Assoc. Comput. Linguistics, vol. 1, pp. 1415–1425, Jun. 2014.

[15] Z. Xu, Y. P. Sheng, L. R. He, and Y. F. Wang, “Review on knowledge graph techniques,” J. Univ. Electron. Sci. Technol. China, vol. 45, no. 4, pp. 589–606, 2016.

[16] M. Banko, M. J. Cafarella, S. Soderland, M. Broadhead, and O. Etzioni, “Open information extraction for the Web,” in Proc. 20th Int. Joint Conf. Artif. Intell. New York, NY, USA: ACM, vol. 1, 2007, pp. 2670–2676.

[17] W. Che, Z. Li, and T. Liu, “LTP: A Chinese language technology platform,” in Proc. Int. Conf. Comput. Linguistics, Demonstrations Assoc. Comput. Linguistics, vol. 1, 2010, pp. 13–16.

JIAXING SONG was born in Baoji, China, in 1993. He received the B.S. and M.S. degrees from the College of Field Engineering, University of Science and Technology, Nanjing, in 2015, and the Ph.D. degree from the Army Engineering University of PLA, in 2019. Since 2019, he has been a Research Assistant with the Army Engineering University of PLA. His research interests include software engineering, mathematical modeling, nanomaterials science, and signal processing. Dr. Song was awarded the First Prize in Mathematical Modeling for Chinese Graduate Students twice, in 2015 and 2016.

FEILONG LIU was born in Fuyang, China, in 1994. He received the B.S. degree from the University of Science and Technology, Nanjing, in 2017, and the M.S. degree from the Army Engineering University of PLA, in 2017. Since 2017, he has been a Research Assistant with the Army Engineering University of PLA and 31101 Troops. His research interests include software engineering, mathematical modeling, and intelligence algorithms.

Mr. Liu has received the First Prize from the Advanced Mathematics Competition, in 2013, and was awarded a prize from the Underwater Robot Competition, in 2014.

KAI DING was born in Nanjing, China, in 1988. He received the B.S. and Ph.D. degrees from the University of Science and Technology, in 2006 and 2013, respectively. Since 2013, he has been with the Science and Technology on Near-Surface Detection Laboratory, Wuxi. His research interests include signal processing, signal detection technology, and intelligence algorithms. Dr. Ding is an Editor of the Journal of Ordnance Equipment Engineering.

KAI DU, photograph and biography not available at the time of publication.

XIAONAN ZHANG, photograph and biography not available at the time of publication.

**VOLUME 8, 2020**