Research on Knowledge Representation and Automatic Recognition of Dynamic Words for Chinese Automatic Syntactic Analysis

DONGDONG GUO, WEIMING PENG, AND JIHUA SONG
School of Artificial Intelligence, Beijing Normal University, Beijing 100875, China
Corresponding author: Jihua Song (songjh@bnu.edu.cn)

This work was supported by the National Natural Science Foundation of China under Grant 61877004.

ABSTRACT There are many temporarily constructed dynamic words in Chinese sentences. Dynamic words are sentence building units that are not included in the general lexicon and are not suitable for further syntactic analysis. Automatic recognition and analysis of dynamic words in sentences play an important role in improving the efficiency and accuracy of Chinese automatic syntactic analysis. The existing researches on dynamic words mainly focus on the qualitative description of concepts and categories. There is no overall algorithm design and experimental exploration on automatic recognition of dynamic words. In the practice of automatic syntactic analysis, dynamic words are generally segmented, and the components are analyzed according to syntax, while the automatic recognition and analysis of dynamic words as a whole are ignored. In this study, the dynamic word is separated from syntactic analysis as the content of lexical analysis and recognized and analyzed as a whole. This paper uses the method of knowledge engineering to research and analyze dynamic words for Chinese automatic syntactic analysis based on sentence pattern structure, initially designs a knowledge representation method of dynamic words, secondly constructs the dynamic word structural mode knowledge base by annotating the dynamic words in the corpus of a certain scale of international Chinese textbooks, and finally explores the automatic recognition methods of dynamic words based on regular expressions, semantic category combinations and machine learning classification algorithms. The experimental results show that the three algorithms can cover the recognition of all types of dynamic words, and achieve relatively ideal accuracy and recall rate.

INDEX TERMS Automatic syntactic analysis, dynamic word, structural mode, knowledge base, automatic recognition.

I. INTRODUCTION
There are many temporarily constructed words and intermediate state combinations between words and phrases in Chinese sentences [1], [2]. In the field of linguistics and Chinese information processing, theoretical research and application practice are generally focused on the definite words included in the normative lexica, and less attention is paid to the dynamic words not included in the lexicon, and even less systematic research on dynamic words. Dynamic words are sentence building units that are not included in the general lexicon and are not suitable for further syntactic analysis [3]–[5]. The typical types of dynamic words include temporary nouns (“阅览室 [yuelan shi](reading room)”, “铁路工人 [tielu gongren](locoman)”), resultative/directional structures (“拿走 [na zou](take away)”, “走进 [zou jin](walk into)”)), reduplication (“看看 [kan kan](have a look)”, “听一听 [ting yi ting](give an ear to)”)), quantitative structures (“一个 [one]”, “五米二 [wu mi er](five meters two)”)), etc [6]–[9]. The dynamic word is an objective language unit, which is an important object to be defined and dealt with in Chinese lexical analysis or syntactic analysis. A comprehensive and in-depth study of dynamic words is of great significance not only to the theoretical research of Chinese morphology and Chinese teaching, but also to the research and practice of Chinese information processing.

Dong [10] divides vocabulary knowledge into two parts: the lexicon and morphology, and strictly distinguishes the two concepts. Based on the modern Chinese word segmentation
vocabulary for information processing, the research focuses on the lexical patterns with high productivity and the strong structural types and main semantic patterns of compound words. Under the guidance of Hierarchical Network of Concepts, Tang [11] adopts the method of combining corpus extraction and deduction in material acquisition, and classifies and describes dynamic words according to the actual situation. He focuses on the rule research of dynamic word combination patterns and the role of the single character knowledge base for information processing in dynamic word recognition. Song [12] pointed out that the boundary between words and phrases in Chinese is fuzzy, and it is hoped that computers can expand the granularity of language units from words to phrases when processing Chinese in many cases. He further proposed the concept of simple phrases. From the practical point of view, some combination types which are easily recognized by computers automatically, have the same grammatical function and have high frequency are defined as simple phrases artificially. The recognition of several types of simple phrases is carried out by using the method of human-computer combination (first design rough extraction rules composed of parts of speech or word forms, and then manually check the candidates one by one, eliminate errors and correct rules). The above researches mainly describe the concept and category of dynamic words qualitatively. For the computer automatic recognition of dynamic words in real language text, the above researches do not involve, or only describe a kind of knowledge base construction scheme for dynamic word recognition, or use the way of human-computer combination to identify several combination types by examples. There is no overall algorithm design and experimental exploration on automatic recognition of dynamic words.

Automatic syntactic analysis is one of the key links and difficulties in the basic research of Chinese information processing, and dynamic word recognition has a very important impact on the efficiency and accuracy of automatic syntactic analysis. In the current mainstream syntactic analysis of Chinese information processing (phrase structure syntactic parsing and dependency syntactic parsing), word segmentation is generally based on the lexicon, and the result of word segmentation is directly used as the leaf node of the syntactic tree. The recognition and lexical analysis of dynamic words outside the lexicon are ignored, and the problems that should be solved at the lexical level are pushed to the syntactic level. Dynamic words are segmented, resulting in the complexity of syntactic tree structure, which undoubtedly increases the burden of syntactic analysis and seriously affects the efficiency and accuracy of syntactic analysis. This paper focuses on the knowledge representation and automatic recognition of dynamic words in Chinese text for automatic syntactic analysis based on sentence pattern structure [13]. The main contents are as follows: initially designs a knowledge representation method of dynamic words, secondly constructs the dynamic word structural mode knowledge base by annotating the dynamic words in the corpus of a certain scale of international Chinese textbooks, and finally on the basis of the structural mode knowledge base explores the automatic recognition methods of dynamic words. The innovative work of this paper is as follows: (1) designing a symbolic knowledge representation scheme of dynamic words; (2) constructing an effective and systematic knowledge base of dynamic word structural modes for Chinese information processing; (3) designing and realizing three automatic recognition algorithms of dynamic words based on regular expressions, semantic category combinations and machine learning classification algorithms.

The organizational structure of this paper is as follows: Chapter 1: introduction, which introduces the research background and significance, research status, research content and innovation, as well as the organizational structure of the article; Chapter 2: sentence pattern structure syntactic parsing, which introduces the syntactic analysis in Chinese information processing and the formalization, syntax and morphology of sentence pattern structure; Chapter 3: knowledge representation of dynamic words, which introduces the knowledge representation method of the dynamic word structural mode, corpus annotation and construction of the structural mode knowledge base; Chapter 4: automatic recognition of dynamic words, which introduces the automatic recognition algorithms and related experiments based on regular expressions, semantic category combinations and machine learning classification algorithms; Chapter 5: conclusion, which summarizes the research and looks forward to the future work.

II. SENTENCE PATTERN STRUCTURE SYNTACTIC PARSING

Sentence level analysis technology in natural language processing can be roughly divided into three levels: lexical analysis, syntactic analysis and semantic analysis. Syntactic analysis is the process of analyzing the input text sentence to get the syntactic structure of the sentence. The analysis of syntactic structure, on the one hand, is the need of language understanding. On the other hand, it also provides support for other natural language processing tasks. For example, syntax driven statistical machine translation requires syntactic analysis of the source language or the target language.

A. SYNTACTIC ANALYSIS IN CHINESE INFORMATION PROCESSING

In the field of Chinese grammar research and teaching, there have been two prevalent sentence structure analysis methods, one of which is Sentence Component Analysis advocated by the traditional grammar school, and the other is Analytical Hierarchy Process advocated by the structuralism grammar school. At present, the mainstream syntactic analysis in Chinese information processing mainly uses Analytical Hierarchy Process or other analysis methods based on structuralism grammar [14], [15]. The phrase structure syntactic parsing is to recognize the phrase structure in sentences and the hierarchical syntactic relationship between phrases [16];
FIGURE 1. (a) The phrase structure: the functional markers “sj”, “dj”, “np” and “vp” respectively indicate the whole sentence, single sentence, noun phrase and verb phrase; the relational markers “ZW”, “DZ”, “PO”, “Z2”, “XX” respectively indicate subject-predicate, attributive-headword, predicate-object, adverbial-headword and default; the part of speech markers “n”, “a”, “v” and “w” respectively represent the noun, adjective, verb and punctuation. (b) The dependency structure: the “HED”, “ATT”, “SBV”, “ADV”, “VOB” and “WP” respectively refer to the head relationship, attributive-headword, subject-verb, adverbial-headword, verb-object and punctuation; the part of speech markers “n”, “a”, “v” and “w” respectively refer to the noun, adjective, verb and punctuation.

D. Guo et al.: Research on Knowledge Representation and Automatic Recognition of Dynamic Words

In order to make the sentence pattern structure syntactic parsing accepted by computers, it is necessary to formalize the sentence pattern structure, that is, to design a set of standard symbol system and data structure, so as to establish a sentence pattern structure grammar system for Chinese information processing. In the history of Chinese grammar teaching and research, Sentence Component Analysis has a corresponding formal description scheme, that is, various types of “diagrams”. The most famous is Li Jinxi’s diagrammatic method [19], followed by the widely used “adding line marking method”. The adding line method is a simple diagrammatic method, which cannot describe the complex structure level. Therefore, the formalization of sentence pattern structure is designed systematically based on Li Jinxi’s diagrammatic method [20]–[24].

The design of the diagrammatic method includes two aspects: one is the diagrammatic style; the other is the corresponding data structure. Considering the hierarchy and extensibility of syntactic structure, XML (Extensible Markup Language) is used as the data storage format. The two-way transformation of lossless information can be carried out between the diagrammatic style and XML structure, that is, the diagrammatic style can be encoded and saved as a certain level of XML structure; otherwise, the XML structure can be decoded into the original diagrammatic style renewedly.

C. SYNTAX OF SENTENCE PATTERN STRUCTURE

There are seven types of sentence components: subject, predicate, object, attribute, adverbial, complement and independent. Function words do not directly act as sentence components, but occupy a specific “function word position”. In XML structure, the sentence structure continuously embeds component nodes and function word position nodes according to certain rules until the part of speech node as the leaf node. The XML element tag set is shown in Table 1.

It is the most important thought of syntactic analysis reflected in the diagrammatic method to separate the main components and additional components of sentences by a long horizontal line, that is, the so-called “grasping the main components”. Among the six components of a sentence, the subject, predicate and object are defined as the main
TABLE 1. The XML element tag set. XML elements mainly include sentences, components, function word positions and parts of speech.

| Sentence/Component/Function word position | Tag   | POS  | Tag   |
|-----------------------------------------|-------|------|-------|
| whole sentence                          | ju    | noun | n     |
| clause                                  | xj    | time word | t   |
| subject                                 | sbj   | localizer | f   |
| predicate                               | prd   | numeral | m   |
| object                                  | obj   | quantifier | q   |
| attribute                               | att   | pronoun | r   |
| adverbial                               | adv   | verb  | v     |
| complement                              | cmp   | adjective | a   |
| independent                             | ind   | adverb | d     |
| preposition position                    | pp    | preposition | p   |
| conjunction position                    | cc    | conjunction | c   |
| structural auxiliary position           | uu    | auxiliary word | u   |
| auxiliary position (attached to nominal phrases) | un | interjection | e   |
| auxiliary position (attached to verbal phrases) | uv | onomatopoeic word | o   |
| localizer position                      | ll    | punctuation | w   |

components, which are above the main horizontal line when they are diagrammatized; the attribute, adverbial and complement are defined as the additional components, which are below the main horizontal line when they are diagrammatized. The basic pattern of Chinese sentence trunk is the “subject-predicate-object” structure. The sentence structure that does not break the basic pattern in the trunk and does not contain any additional components is called “basic sentence pattern”, and its diagrammatic style and XML structure are shown in Fig. 2. The sentence structure that does not break the basic pattern in the trunk, but has additional components, or has structural factors such as double objects, juxtaposition, apposition and independent is called “extended sentence pattern”. Its diagrammatic style and XML structure are shown in Fig. 3. The sentence structure that breaks the basic pattern in the trunk is called “complex sentence pattern”. Within the scope of a single sentence, there are five categories: compound predicate sentence, parataxis-structure predicate sentence, subject-predicate predicate sentence, pivotal sentence and serial verb sentence. Their diagrammatic style and XML structure are shown in Fig. 4. The above is the normal sentence pattern of Chinese sentences. In addition, there are three kinds of special sentence patterns: one-word sentence, no-subject sentence and inverted sentence. Their diagrammatic style and XML structure are shown in Fig. 5.

D. MORPHOLOGY OF SENTENCE PATTERN STRUCTURE

The final state of syntactic analysis is to analyze the sentence layer by layer to “word” units. In the practice of Chinese information processing, syntactic analysis usually uses the lexicon to realize lexical analysis (word segmentation and part of speech tagging). In the construction of Chinese Treebank based on sentence pattern structure, Modern Chinese Dictionary (6th Edition) [25] is selected as the basic lexicon, and the part of speech and sense code of the word are marked.
In the process of parsing sentences into words, it is not easy to master the granularity of analysis, which can be regarded as the “word” [26]–[28]. At present, the mainstream syntactic parsing basically accepts the standard of the lexicon of word segmentation in terms of word granularity, that is, taking the result of word segmentation as the leaf node of the syntactic tree. Theoretically speaking, the internal structure analysis of words should not be reflected in the syntactic tree, and the leaf node of the syntactic tree should be the word. However, there are a large number of dynamic words in Chinese sentences that are generally segmented in the process of word segmentation and are finally syntactically analyzed. Dong [10] pointed out that the lexicon and morphology are two parts of vocabulary knowledge, and they have both differences and connections. The definition of words cannot only be from the perspective of the lexicon. All forms that conform to the lexical pattern, even if they are not in the lexicon, should be considered as words. If dynamic words outside the lexicon are syntactically segmented, the syntactic tree structure will expand exponentially. This undoubtedly increases the burden of syntactic analysis, makes the realization of automatic syntactic analysis more complex, and affects the accuracy of the analysis results. In addition, neglecting dynamic words will cause the leaf nodes of the syntactic tree to be too fragmentary, which will further affect the semantic understanding of sentences. Based on the reality of Chinese information processing, the sentence pattern structure syntactic parsing considers dynamic words besides the basic lexicon in terms of word granularity division. The dynamic word is separated from syntactic analysis as the content of lexical analysis and recognized and analyzed as a whole.

III. KNOWLEDGE REPRESENTATION OF DYNAMIC WORDS

The automatic analysis and recognition of dynamic words is the key link and basic work of Chinese automatic syntactic analysis based on sentence pattern structure. The formal knowledge representation of dynamic words is the first important work to be completed. This paper designs the dynamic word structural mode to realize the knowledge representation of dynamic words.
A. THE DYNAMIC WORD STRUCTURAL MODE

Dynamic words are characterized by cohesive meaning, suitable syllables and relatively stable structure, which cannot be expanded freely. Four kinds of information, including the whole part of speech of the dynamic word, the part of speech of the internal component, the syllable number of the internal component and the structural relationship between internal components, can better reflect the lexical characteristics of dynamic words and distinguish various types of dynamic words. Therefore, they are selected to describe the dynamic word structural mode. The structural relationship between internal components of dynamic words is shown in Table 2.

The specific design scheme of the dynamic word structural mode is as follows:

- `<structural relationship symbol> ::= ... | / | -> | ← | | | | | -`

The “‘n”, “t”, “f”, “m”, “q”, “v”, “a”, “d”, “p”, “c”, “u”, “e”, “o”, “Ng”, “Vg”, “Ag”, “Dg” and “Ug” denote nouns, time words, localizers, numerals, quantifiers, pronouns, verbs, adjectives, adverbs, prepositions, conjunctions, auxiliary words, interjections, onomatopoeic words, nominal morphemes, verbal morphemes, adjective morphemes, adverbial morphemes and affix morphemes (such as the category of the suffix “儿[er]” in “玩儿[wán er][play]”) respectively. Punctuation symbols are denoted by the “w”. The syllable number is the default value 1 when it is NULL.

Examples of the dynamic word structural mode are shown in Table 3. The structural mode of “阅读室[yuelan shi](reading room)” is “n: v2 / n”. In “n: v2 / n”, the “n” before the colon indicates the whole part of speech is a noun; the “v2” indicates the part of speech of “读[du](read)” is a verb and its syllable number is 2; the final “n” represents “室[shi](room)” is a noun and its syllable number is 1 (The syllable number is the default value 1 when it is NULL); the “/” represents attribute-headword relationship.

B. ANNOTATING OF CORPUS

In order to obtain the actual various sentence pattern structures (to provide recognition rules for the realization of automatic sentence pattern structure syntactic parsing) and dynamic word structural modes in Chinese text, the undergraduates or postgraduates with the background of linguistics and language teaching are organized to annotate syntactic information as well as all the dynamic words in sentences. The corpus is mainly international Chinese textbooks, including New Practical Chinese Textbook, Happy Chinese, Great Wall Chinese, Chinese with Me, Mandarin Teaching Toolbox, Contemporary Chinese, Chinese Paradise and other international Chinese textbooks. Modern Chinese Dictionary (6th Edition), as the basic lexicon, is a systematic, comprehensive and relatively stable word system. The internal components of dynamic words are annotated with the words and their corresponding part of speech.

### Table 2. The symbol set of dynamic word structural relationships.

| Relationship | Symbol | Example | Note |
|--------------|--------|---------|------|
| 并列 (juxtaposition) | ... | 花(花) | including reduced juxtaposition |
| 定中 (attributive-headword) | / | 鸡(chicken) | including "name-appellation" structures |
| 状中 | (extremely)→ | 鸡(鸡) | including auxiliary verb and serial verb structures |
| 述补 (predicate-complement) | ← | 营(drive) | resultative structures; directional structures |

### Table 3. Examples of the dynamic word structural mode.

| Dynamic word | Mode | Dynamic word | Mode |
|--------------|------|--------------|------|
| 木桥 (wooden bridge) | n: n / n | 一个 (one) | m: m-q |
| 阅览室 (reading room) | n: v2 / n | 看清楚 (see clearly) | v: v→a |
| 严谨性 (rigor) | n: a2-Ug | 举起来 (raise up) | v: v→v2 |
| 制药厂 (pharmaceutical factory) | n: v / n / n | 看一看 (take a look at) | v: v→v+v |
| 铁路工人 (locoman) | n: n2 / n2 | 休息休息 (have a rest) | v: v2 / v2 |

### Table 3. Examples of the dynamic word structural mode.
TABLE 4. Examples of the annotating result of dynamic words in the corpus. The outermost part of speech node is the whole part of speech of the dynamic word, and the attribute "mod" contains other structural mode information except the whole part of speech, that is, the part of speech and syllable number of each internal component and the structural relationship between internal components.

| ID | Example |
|----|---------|
| 1  | 就是这个，我猜<n mod="v-v">你</n sen="001">喝</n sen="003">不</n sen="005">一滴。
(That is it, and I just finished a bottle.) |
| 2  | 我可以<n mod="v-v">试</n sen="001">试</n sen="005">吗？
(May I have a try?) |
| 3  | 有人来给娘送信，当妈的先问，那家有几口<n mod="n2\n">人</n sen="001">？
(When someone comes to serve as a matchmaker for the girl, the mother first asks how many pickled vegetable jars the family has.) |
| 4  | 小区南面有一个<n mod="n2\n">n</n sen="001">地铁<n sen="102">站</n sen="105">，到城里要半个小时。
(There is a subway station in the south of the community. It takes half an hour to get to the city.) |
| 5  | 我今天有<n mod="v2\n">n</n sen="001">课</n sen="005">。
(I will have a writing class tomorrow. The teacher asked us to hand in the composition. I have not finished it yet!) |

TABLE 5. Parts of dynamic word structural modes. The structural modes of No. 1–18 are partial structural modes of trisyllabic nounal dynamic words, and that of No. 19–30 are structural modes of verb reduplication. The "frequency" field indicates the number of the dynamic words corresponding to the structural mode in the annotated corpus.

| ID | Mode | Frequency | ID | Mode | Frequency |
|----|------|-----------|----|------|-----------|
| 1  | n: n2/N | 1548      | 16 | n: v | 26 |
| 2  | n: a/n2 | 810       | 17 | n: Ng...n2 | 24 |
| 3  | n: n/n2 | 462       | 18 | n: a/n-Lg | 23 |
| 4  | n: v2/n | 339       | 19 | v: v | 548 |
| 5  | n: n2-Lg | 318      | 20 | v: v-m-v | 96 |
| 6  | n: n2/Ng | 228      | 21 | v: v2-v2 | 65 |
| 7  | n: v2/Ng | 174      | 22 | v: v-v2 | 47 |
| 8  | n: a2/n | 126      | 23 | v: u-v | 27 |
| 9  | n: a2/Ng | 58       | 24 | v: u-v-u | 7 |
| 10 | n: m/n2 | 54       | 25 | v: v-v3 | 5 |
| 11 | n: v2-u | 51       | 26 | v: u-m-v-v | 4 |
| 12 | n: n/Ng | 49       | 27 | v: u-v2 | 3 |
| 13 | n: v | 42       | 28 | v: m-v2 | 2 |
| 14 | n: m-q/n | 35     | 29 | v: v-v...v-v | 1 |
| 15 | n: Ag/n2 | 34       | 30 | v: u-d-v | 1 |

information in Modern Chinese Dictionary. In order to annotate dynamic words, the internal components should be separated first, and the segmentation standard is based on the unity of structure and meaning until each internal component can find the corresponding sense and part of speech information in Modern Chinese Dictionary. For example, for the dynamic word “足球迷[zuqiu mi](football fan)”, because “足球[zuqiu](football)” and “迷[mi](fan)” are included in Modern Chinese Dictionary, the correct segmentation result should be “足球[zuqiu](football)” and “迷[mi](fan)”. Although “足[zu](foot)” and “球[qi][mi](ball game fan)” are also included in Modern Chinese Dictionary, the segmentation of “足[zu](foot)” and “球[qi](ball game fan)” is obviously not correct, and this combination structure is not consistent with the meaning of the word.

The annotating work is carried out on the semi-automatic corpus annotating platform which integrates the corpus of international Chinese textbooks, Modern Chinese Dictionary and visual annotating tools [29], [30]. The dynamic word annotating content includes the whole part of speech, the corresponding part of speech of each internal component (the syllable number of the internal component is automatically obtained by the semi-automatic corpus annotating platform), the structural relationship between internal components and other structural mode information, and the specific annotation results are shown in Table 5.

In order to ensure the accuracy and consistency of the annotating results, the same corpus text was annotated by two different annotators, and the annotating results were reviewed by professional auditors. The data with consistent annotating results and approved shall be regarded as valid data. If the annotating results are inconsistent or not approved, the final results need to be determined through discussion and analysis between the annotator and the auditor.

C. CONSTRUCTION OF THE STRUCTURAL MODE KNOWLEDGE BASE

In this paper, the corpus data of 29465 sentences (498965 words) with dynamic word structural mode information are collected. Regular expressions are used to match and extract the dynamic words and their structural mode information in the annotated corpus. The regular expression is a formula that can match a class of strings with a certain pattern, which is composed of several common characters and special characters (metacharacters). Common characters include uppercase and lowercase letters, numbers, Chinese characters, punctuation marks and other characters. Metacharacters refer to special characters with special meanings. For example, the character “\w” can match the letter, digital or underscore; the character “\.” can match any single character except “\n” and “\r”; the character “\+” can match the preceding subexpression one or more times; the character “?” is used after “\+” to indicate that the matching pattern is non-greedy, and the non-greedy pattern matches...
the searched string as few as possible. The rules of dynamic words and their structural mode information in the annotated corpus are clear. All the information to be extracted can be accurately matched by the regular expression "<\w mod="n.".+?"</\w>".

After statistical analysis of the extracted information, a knowledge base with 672 kinds of dynamic word structural modes is preliminarily established. Parts of dynamic word structural modes are shown in Table 5.

The structure of the structural mode knowledge base is shown in Table 6. Table 7 shows the specific contents of the structural mode "n: v\|Ng \nearrow n" in the knowledge base. The "detail" field lists all the dynamic words in the annotated corpus corresponding to the structural mode and their frequency of occurrence, as well as the sense code of each internal component.

### IV. AUTOMATIC RECOGNITION OF DYNAMIC WORDS

The realization of automatic recognition of dynamic words is the preparation of automatic syntactic analysis based on sentence pattern structure. On the basis of the previous work, this chapter explores automatic recognition methods of dynamic words in the raw corpus of international Chinese textbooks.

#### A. RECOGNITION BASED ON REGULAR EXPRESSIONS

The composition of reduplication and quantitative structures is very regular, the formal features are obvious, and the internal components are clear or relatively closed. It is suitable to construct regular expression rules reflecting their distinct features, and use regular expressions to identify the corresponding kinds of dynamic words. In this section, the regular expression method is used to explore the automatic recognition of verb reduplication in corpus.

Before using regular expressions to recognize verb reduplication, it is necessary to summarize and determine reasonable regular expressions for verb reduplication according to its own language characteristics and the structural mode knowledge base. The regular expressions corresponding to various structural modes of verb reduplication are shown in Table 8. The regular expression defines the information of syllable number, the position of the reduplication character, and the fixed component of verb reduplication corresponding to the specific structural mode. A combination can generally be judged to be verb reduplication if it satisfies the regular expression rule and part of speech information, and the specific type of its structural mode can also be determined.

#### 1) RECOGNITION ALGORITHM

The automatic recognition algorithm of the verb reduplication in the raw corpus is as follows. In the process of automatic recognition, the sequence of verb reduplication in different structural modes is different.

1) First, the regular expression "(\·)\{1}\cdot(\·)\{2" of the structural mode "v\cdot v\ldots v\cdot v" is used to match the combinations in the raw corpus, and obtain the verb reduplication candidates with the structural mode "v\cdot v\ldots v\cdot v". Then use Modern Chinese Dictionary to filter the candidates and delete the vocabularies that have been included in Modern Chinese Dictionary. Finally, extract the first character (use "A" to represent)
TABLE 8. The regular expressions corresponding to various structural modes. The fields “type” and “frequency” respectively indicate the number of the types and the frequency of the dynamic words corresponding to the structural mode in the annotated corpus.

| ID | Mode       | Example (Chinese) | Regular expression | Type | Frequency |
|----|------------|-------------------|--------------------|------|-----------|
| 1  | v·v·v      | 看看 (have a look) | (.j)l              | 70   | 548       |
| 2  | v·m·v·v    | 看一看 (take a look at) | (.j)一l | 38   | 96        |
| 3  | v·v2·v2   | 休息休息 (have a rest) | (.j)l1            | 32   | 65        |
| 4  | v·v2      | 散散步 (take a walk) | (.j)l             | 17   | 47        |
| 5  | v·u·v      | 看了看 (have a look) | (.j)了l           | 10   | 27        |
| 6  | v·u·u·v-u | 玩儿玩儿 (just play) | (.j)儿l1 儿      | 2    | 7         |
| 7  | v·v3      | 亲亲头头 (chat)   | (.j)l1           | 2    | 5         |
| 8  | v·u·m·u·v-u| 玩儿玩儿 (just play) | (.j)儿一儿 儿    | 1    | 4         |
| 9  | v·v2·v2   | 吼了号脉 (feel the pulse) | (.j)了l1       | 3    | 3         |
| 10 | v·m·v2   | 拍一拍手 (shake hands) | (.j)一l           | 2    | 2         |
| 11 | v·v·v·v   | 进进出出 (shuttle in and out) | (.j)l(.j)l2   | 1    | 1         |
| 12 | v·u·d·v | 想了又想 (think again and again) | (.j)了又l1 | 1    | 1         |

2) RECOGNITION EXPERIMENT

In order to objectively reflect the automatic recognition effect of verb reduplication, 3 groups of raw corpus of international Chinese textbooks (hundreds of thousands of Chinese characters in each group) were selected randomly for the experiment. According to the above recognition algorithm, the automatic recognition experiment is carried out on 3 groups of randomly selected raw corpus, and the result is shown in Table 9.

TABLE 9. Recognition results of verb reduplication.

| Raw Corpus | Accuracy Rate | Recall Rate | F-Measure |
|------------|---------------|-------------|-----------|
| Group 1    | 86.42%        | 97.15%      | 91.47%    |
| Group 2    | 84.78%        | 94.96%      | 89.58%    |
| Group 3    | 82.35%        | 95.06%      | 88.25%    |

and the third character (use “B” to represent) in the candidate. By looking up Modern Chinese Dictionary to determine whether “AB” exists and whether it is a verb, if it exists and the part of speech is the verb, the candidate would be defined as verb reduplication of the structural mode “v·v·v”. If it exists and is the multi-category word, the method of determining the part of speech is the same as step 1 above. Once again according to this method recognize the verb reduplication of the structural modes “v·v·v”, “v·v-u·v·v” and “v·v·m·v·v” in turn.

3) First, the regular expression of the structural mode “v·v·v·v·v·v” is used to match the combinations in the raw corpus, and obtain the verb reduplication candidates. Then use Modern Chinese Dictionary to filter the candidates and delete the vocabularies that have been included in Modern Chinese Dictionary. Finally, extract the first two characters in the candidate. By looking up Modern Chinese Dictionary to determine whether it exists and whether it is a verb, if it exists and the part of speech is the verb, the candidate would be defined as verb reduplication of the structural mode “v·v·v·v·v·v”. If it exists and is the multi-category word, the method of determining the part of speech is the same as step 1 above. Once again according to this method recognize the verb reduplication of the structural mode “v·v·v·v·v·v”. For the rest of structural modes, separately match the combinations in the raw corpus, and obtain the verb reduplication candidates with various types of structural modes. Then use Modern Chinese Dictionary to filter the candidates and delete the vocabularies that have been included in Modern Chinese Dictionary. The remaining candidates can be defined as the verb reduplication corresponding to various modes.

2) RECOGNITION EXPERIMENT

In order to objectively reflect the automatic recognition effect of verb reduplication, 3 groups of raw corpus of international Chinese textbooks (hundreds of thousands of Chinese characters in each group) were selected randomly for the experiment. According to the above recognition algorithm, the automatic recognition experiment is carried out on 3 groups of randomly selected raw corpus, and the result is shown in Table 9.
According to the experimental result, it can be concluded that the regular expressions of structural modes of verb reduplication have a significant effect on automatic recognition of verb reduplication. Both the recognition accuracy rate and recall rate of verb reduplication are high, and the recognition results are relatively satisfactory.

3) EXPERIMENTAL ANALYSIS
Through analysis of the experimental data, it is found that the reasons for reducing the accuracy rate and recall rate of automatic recognition are mainly the following aspects:
1) For the multi-category words which have verb and adjective two parts of speech at the same time, it is not rigorous enough to determine the specific part of speech only based on the frequency of the two parts of speech. It is necessary to improve the recognition effect by studying the reduplication phenomenon of this kind of words.
2) It is difficult to avoid some ambiguities in the process of recognition. For example, for the “好好说说[haohao shuo shuo](tell earnestly)” from the sentence “你和他好好说说这件事(You talk to him about this matter earnestly)”, because there is the verb “好 说[haoshuo](no problem)” in Modern Chinese Dictionary, so it is easy to be defined as verb reduplication of the structural mode “v: v...v...v”. The ambiguities need to be further eliminated by improving the recognition algorithm.
3) The current regular expressions can only reflect the partial external characteristics of verb reduplication. It is necessary to dig out more effective and quantifiable feature information with the in-depth study of verb reduplication.
4) In the algorithm of recognizing verb reduplication, restrictions and screening conditions are limited. The research resources and achievements in the field of linguistics, language teaching and Chinese information processing are needed to further enrich the recognition conditions of verb reduplication.

B. RECOGNITION BASED ON SEMANTIC CATEGORY COMBINATIONS
The formal features of temporary nouns and resultative/directional structures are not as strong as those of reduplication and quantitative structures, and their internal components are also rich and relatively open, so they are not suitable for the recognition algorithm based on regular expressions. However, the combination of their internal components is obviously restricted by semantics, so the combination of semantic categories corresponding to the internal components can be explored to identify the corresponding type of dynamic words. This section attempts to identify the corresponding dynamic words by using the semantic category combination features of the internal components of dynamic words, taking the recognition of resultative/directional structures as an example. This paper uses Modern Chinese Semantic Dictionary [31] to annotate the semantic categories corresponding to the internal components of resultative/directional structures in the structural mode knowledge base, and obtains the semantic category combinations corresponding to various structural modes. The semantic category combinations corresponding to the mode “v: v2—a2” are shown in Table 10.

1) RECOGNITION ALGORITHM
The automatic recognition of resultative structures and directional structures in the raw corpus is carried out separately. The automatic recognition process is as follows.
1) Using the automatic word segmentation tool (with Modern Chinese Dictionary as the basic lexicon) [32], the original corpus is automatically segmented and tagged with parts of speech to obtain the processed corpus which has been segmented and carries part of speech information. In order to avoid the negative effect of segmentation and part of speech tagging errors on the automatic recognition of resultative structures (or directional structures), the results of segmentation and part of speech tagging are manually proofread.
2) Sort resultative structure (or directional structure) structural modes according to the number of internal components from more to less, the number of syllables of internal components from large to small, and the frequency of corresponding resultative structures (or directional structures) from high to low. For example, there are at most three internal components in the resultative structure structural modes, so the structural mode “v: v-d-a2” with three internal components, four
syrables (the highest) and the highest frequency of corresponding resultative structures ranks first in the queue of structural modes.

3) Recognize resultative structures (or directional structures) corresponding to the current resultative structure (or directional structure) structural mode in order. The specific method is to extract the combinations consistent with parts of speech and syllables of the internal components of the current mode from the processed corpus in step 1 as the candidate set of the structure to be identified. Then, through Modern Chinese Semantic Dictionary, semantic category information is automatically added to all the internal components of each combination in the candidate set. The semantic category combination of each candidate will be compared with the “combination” field of the current mode in the queue of structural modes. If it exists in the “combination” field, the candidate is qualified and will be included in the resultative structure (or directional structure) set; if it does not exist in the “combination” field, the candidate is not qualified. Once identified as resultative structures (or directional structures), the contents are no longer involved in the following recognition operations.

2) RECOGNITION EXPERIMENT
In order to objectively reflect the automatic recognition effect of resultative/directional structures, 3 groups of raw corpus of international Chinese textbooks (tens of thousands to hundreds of thousands of Chinese characters in each group) were selected randomly for the experiment. According to the above recognition algorithm, the automatic recognition experiment is carried out on 3 groups of randomly selected raw corpus, and the recognition results of resultative structures and directional structures are shown in Table 11 and Table 12 respectively.

According to the experimental result, it can be concluded that the semantic category combinations corresponding to the resultative/directional structure structural modes in the knowledge base have a significant effect on automatic recognition of resultative structures and directional structures. The accuracy and recall rate of automatic recognition are both more than 80% or even more than 90%. The directional complements of directional structures are very clear, and they are monosyllabic or multisyllabic directional verbs, which are both fixed and few in number in Chinese, so the accuracy and recall rate of automatic recognition of directional structures are better than resultative structures.

3) CONTRAST EXPERIMENT
The resultative structures and directional structures in the corpus are identified only according to the parts of speech of the internal components, the number of syllables of the internal components of the dynamic words and monosyllabic or multisyllabic directional verbs in the contrast experiment. The experimental results are as follows in Table 13. It can be seen from Table 13 that the accuracy rate and F-measure of automatic recognition of resultative structures are relatively low; the accuracy rate of automatic recognition of directional structures is not far from the above recognition experiment results, mainly because the limitation of directional verbs is very important for the recognition of directional structures, so as to ensure the validity of recognition.

4) EXPERIMENTAL ANALYSIS
Through the comparison of the above recognition experiment results and the contrast experiment results, it is found that the recognition algorithm based on semantic category combinations can significantly improve the automatic recognition effect of resultative/directional structures. Through analysis of the recognition experiment data, the recognition algorithm based on semantic category combinations has some shortcomings and can be improved in the following aspects:

1) Limited by the scale of the annotated corpus, the semantic category combinations obtained cannot cover all the semantic category combinations corresponding to the resultative/directional structures actually appearing in the texts, which limits the recall rate of resultative/directional structure recognition. In the later stage, it is necessary to expand the scale of annotated corpus to further collect and improve the semantic category combinations corresponding to various structural modes.
2) It is difficult to avoid some ambiguities in the process of recognition. The recognition error occurs during the specific recognition process. For example, in the sentence “格林先生不但汉语说得好，而且对中国文化也很有研究(Mr. Green not only speaks Chinese well, but also studies Chinese culture very well)”, the “好(hao)(well)” in “说得好(shuo de hao)(speaks well)” should be used as the modal complement of “说(shuo)(speak)”, but because its formal structure is identical with the possible form “说得好(shuo de hao)(can come to an agreement)” of the resultative structure “说好(shuo hao)(come to an agreement)”, it is easy to be confused into the resultative structure set of the structural mode “v: v-u-a”. The ambiguities need to be further eliminated by improving and perfecting the recognition conditions and process.

3) The information of structural modes and the semantic category combination can reflect the structure and semantic features of resultative/directional structures to a certain extent, and cannot fully reflect the characteristics of resultative/directional structures. It is necessary to dig out more effective and quantifiable feature information with the in-depth study of resultative/directional structures.

C. RECOGNITION BASED ON MACHINE LEARNING CLASSIFICATION ALGORITHMS

The problem of automatic recognition of dynamic words is essentially a binary classification problem. This section explores the specific performance of various machine learning classification algorithms in the automatic recognition of dynamic words, taking the recognition of a typical nounal dynamic word – trisyllabic nounal dynamic word as an example.

1) RECOGNITION ALGORITHM

The automatic recognition algorithm of trisyllabic nounal dynamic words is as follows.

1) Sort dynamic word structural modes according to the frequency of corresponding dynamic words from high to low. The dynamic words corresponding to each structural mode are recognized in order. For example, there is the highest frequency of corresponding trisyllabic nounal dynamic words, so the structural mode “n: n2/’n” ranks first in the queue of structural modes.

2) Extract the combinations consistent with parts of speech and syllables of the internal components of the current structural mode from the annotated corpus, and whether each combination is a dynamic word or not is marked (the corpus has carried relevant information, with “yes” for dynamic words and “no” for non-dynamic words).

3) Identify the corresponding semantic categories of internal components of the combinations in Modern Chinese Semantic Dictionary automatically (stipulating and restricting the internal combination of trisyllabic nounal dynamic words from the semantic perspective). For the case that the internal component does not exist in Modern Chinese Semantic Dictionary, the internal component itself is temporarily used as its semantic category.

4) Build the dataset for machine learning. The dataset includes the semantic categories corresponding to the internal components (For trisyllabic nounal dynamic words with the same structural mode, the number of internal components is fixed, which is 2 or 3) and the field whether it is a dynamic word (“yes” or “no”). Different semantic categories are mapped to continuous different integers. In the dataset, semantic categories are represented by corresponding numbers.

5) Evaluate different machine learning classification algorithms. The classification algorithms include: Logical Regression (LR), Linear Discriminant Analysis (LDA), K Nearest Neighbor (KNN), Classification and Regression Tree (CART), Naive Bayesian Classifier (NB) and Support Vector Machine (SVM). The 10 fold cross validation is used to separate data, and the same random number allocation method is used to ensure that all algorithms use the same data. The parameters of different classification algorithms are the default parameters provided by scikit-learn.

6) Select the classification algorithm with the highest accuracy to create a classification model for the prediction of trisyllabic nounal dynamic words.

2) RECOGNITION EXPERIMENT

In this experiment, the trisyllabic nounal dynamic words corresponding to the two typical structural modes “n: n2/’n” and “n: n2/’Ng” are recognized. The preparation of the dataset is as follows: (1) Dynamic words with structural modes “n: n2/’n” and “n: n2/’Ng” are extracted from the annotated corpus by regular expressions and marked as “yes”; then, the combination of “the disyllabic noun + the monosyllabic noun or nominal morpheme” is extracted from the remaining corpus with the extracted dynamic words removed and marked as “no”. (2) The dynamic words or non-dynamic words extracted above all contain two internal components. The corresponding semantic category of each internal component is extracted from Modern Chinese Semantic Dictionary automatically. For example, for the dynamic word “地铁(ditie zhan)(subway station)”, the internal components “地铁(ditie)(subway)” and “站(zhan)(station)” respectively correspond to the semantic categories “交通工具(jiaotonggongju)(vehicle)” and “空间(kongjian)(space)”. For the case that the internal component does not exist in Modern Chinese Semantic Dictionary, the internal component itself is temporarily used as its semantic category. (3) In
order to train the machine learning model, the semantic categories corresponding to the internal components of the combinations are mapped into continuous integers. For example, “交通工具[jiaotonggongju](vehicle)” and “空间[kongjian](space)” are mapped to “23” and “18” respectively. (4) The semantic category data of each internal component and the classification result whether the combination is a dynamic word are selected to form the dataset (including three data feature attributes). In this experiment, a total of 2226 data records were extracted, of which 1872 were classified as “yes” and 354 were classified as “no”.

The classification effect of different classification algorithms is shown in Fig. 7. The box-plot shows the accuracy of various algorithms and the result distribution in 10 fold cross validation. It can be seen from Fig. 7 that CART algorithm has the best recognition effect. The average accuracy and standard deviation of each classification algorithm are shown in Table 14. The average accuracy of CART algorithm is up to 86.51%.

### Table 14. Recognition results of trisyllabic nounal dynamic words.

| Algorithm | Accuracy Rate | Standard Deviation |
|-----------|---------------|--------------------|
| LR        | 83.10%        | 0.0563             |
| LDA       | 82.97%        | 0.0506             |
| KNN       | 84.09%        | 0.0517             |
| CART      | 86.51%        | 0.0483             |
| NB        | 79.86%        | 0.0758             |
| SVM       | 83.78%        | 0.0500             |

3) CONTRAST EXPERIMENT

The experimental corpus data is divided into 3 parts. The trisyllabic nounal dynamic words in the corpus are identified only according to the parts of speech or morpheme categories of the internal components and the number of syllables of the internal components of dynamic words. The experimental results are as follows in Table 15. Although the contrast experiment guarantees the recall rate of automatic recognition, the accuracy rate of automatic recognition is not ideal compared with the above recognition experiment, and the accuracy rate is only between 70% and 80%.

Use the above recognition algorithm based on semantic category combinations to recognize the trisyllabic nounal dynamic words in the three groups of data, and the recognition results are shown in Table 16. It can be seen from Table 16 that the recognition algorithm based on semantic category combinations has also achieved good results, and the accuracy rate is higher than most of the above-mentioned machine learning classification algorithms, which shows that the recognition algorithm based on machine learning classification algorithms has a large space for improvement.

### Table 15. Recognition results of the contrast experiment.

| Raw Corpus | Accuracy Rate | Recall Rate | F-Measure |
|------------|---------------|-------------|-----------|
| Group 1    | 78.03%        | 100%        | 87.66%    |
| Group 2    | 76.15%        | 100%        | 86.46%    |
| Group 3    | 72.24%        | 100%        | 83.88%    |

### Table 16. Experiment results of recognition based on semantic category combinations.

| Raw Corpus | Accuracy Rate | Recall Rate | F-Measure |
|------------|---------------|-------------|-----------|
| Group 1    | 90.33%        | 92.49%      | 91.40%    |
| Group 2    | 93.23%        | 87.33%      | 90.18%    |
| Group 3    | 91.16%        | 89.48%      | 90.31%    |

4) EXPERIMENTAL ANALYSIS

By comparing the above recognition experiment results and the contrast experiment results, it can be found that the recognition algorithm proposed in this section is effective. At the same time, it also embodies the scientificity, validity and completeness of the structural mode knowledge base. In the recognition algorithm of trisyllabic nounal dynamic words, the automatic acquisition of semantic categories relies too much on the completeness, accuracy and strong practicability of Modern Chinese Semantic Dictionary, and machine learning classification algorithms use fewer types of features, which limits the accuracy of automatic recognition. In the future work, it is necessary to improve the automatic recognition of trisyllabic nounal dynamic words by improving
Modern Chinese Semantic Dictionary and extracting more effective features.

The main positive impacts of the proposed algorithms in this study are as follows: (1) They provide a systematic solution for the problem of automatic recognition of dynamic words in Chinese text. The above three algorithms can cover the recognition of all types of dynamic words. (2) They have achieved good results in recognition accuracy and recall rate. Although the recognition accuracy rates and recall rates of different types of dynamic words are different, the recognition accuracy rates are basically above 80%, most of the recall rates are above 90%, and the recognition accuracy rates of some types of dynamic words can even reach more than 90%. (3) They provide solving ideas or reference value for other natural language processing tasks. For a specific natural language processing task, this study extracts relevant knowledge and rules by annotating large-scale corpus, constructs a task-oriented knowledge base, and then designs algorithms based on the constructed knowledge base to complete the specific task.

V. CONCLUSION

In this paper, based on the dynamic word structural mode knowledge base for Chinese information processing, we propose dynamic word recognition algorithms based on regular expressions, semantic category combinations and machine learning classification algorithms respectively. This study provides a systematic solution for the task of automatic recognition of dynamic words in Chinese text. The proposed three recognition algorithms can cover the recognition of all types of dynamic words, and achieve good recognition accuracy and recall rate. The recognition algorithms use the structural form or semantic pattern of dynamic words to realize the integration of theory research and automatic recognition practice of dynamic words, which makes them verify each other and develop together. In addition, the recognition algorithms can be applied to the process of Chinese automatic syntactic analysis based on sentence pattern structure, which can effectively improve the efficiency and accuracy of syntactic analysis. However, the information of language rules of dynamic words is not fully mined in this study, especially the context information of dynamic words in sentences, which affects the accuracy and recall rate of automatic recognition to a certain extent. In future work, we hope to further analyze the theoretical research results of dynamic words and large-scale annotated corpus to mine the feature information of dynamic words, so as to continue to improve the effect of the recognition algorithms, and improve the recognition algorithm based on machine learning classification algorithms by adding feature attributes, integrating various classification algorithms and adjusting parameters.

REFERENCES

[1] W. Peng, J. Song, and N. Wang, “Issues in formalization of Chinese syntactic analysis,” J. Chin. Inf. Process., vol. 30, no. 3, pp. 30–35 and 67, 2016.

[2] W. Peng, J. Song, and S. Yu, “Lexical issues in Chinese information processing: In the background of sentence-based diagram treebank construction,” J. Chin. Inf. Process., vol. 28, no. 2, pp. 1–7 and 36, 2014.

[3] D. Guo, “Analyzing on dynamic words and their structural modes in building the sentence-based treebank,” M.S. thesis, Dept. Softw. Eng., Beijing Normal Univ., Beijing, China, 2016.

[4] D. Guo, S. Zhu, W. Peng, J. Song, and Y. Zhang, “Construction of the dynamic word structural mode knowledge base for the international Chinese teaching,” in Proc. Chin. Lexical Semantics Workshop (CLSW), Singapore, 2016, pp. 251–260.

[5] D. Guo, J. Song, and W. Peng, “Research on dynamic words and their automatic recognition in Chinese information processing,” in Proc. Chin. Lexical Semantics Workshop (CLSW), Leshan, China, 2017, p. 479–488.

[6] D. Guo, J. Song, W. Peng, and Y. Zhang, “Analysis of three syllable noun dynamic words based on the corpus of international Chinese textbooks,” J. Chin. Inf. Process., vol. 32, no. 6, pp. 12–18, 2018.

[7] D. Guo, J. Song, W. Peng, and Y. Zhang, “Research on verb reduplication based on the corpus of international Chinese textbooks,” in Proc. Chin. Lexical Semantics Workshop (CLSW), Chia-yi, China, 2018, pp. 707–719.

[8] D. Guo, J. Song, W. Peng, and Y. Zhang, “Research on resultive/directional structures based on the corpus of international Chinese textbooks,” in Proc. Int. Conf. Knowl. Sci., Eng. Manage. (KSEM), Athens, Greece, 2019, pp. 184–200.

[9] D. Guo, J. Song, W. Peng, and Y. Zhang, “Research on quantifier phrases based on the corpus of international Chinese textbooks,” in Proc. Chin. Lexical Semantics Workshop (CLSW), Beijing, China, 2019, pp. 840–852.

[10] X. Dong, Chinese Lexicon and Morphology. Beijing, China: Peking Univ. Press, 2004.

[11] X. Tang, The Study of Dynamic Words in Chinese Understanding Processing. Beijing, China: Science Press, 2012.

[12] R. Song, Song Rou’s Collected Papers on Language Engineering. Beijing, China: Beijing Language and Culture Univ. Press, 2012.

[13] W. Peng, J. Song, and N. Wang, “Design of diagrammatic parsing method of Chinese based on sentence pattern structure,” Comput. Eng. Appl., vol. 50, no. 6, pp. 11–18, 2014.

[14] J. Song, E. Yang, and Q. Wang, Chinese Information Processing Tutorial. Beijing, China: Higher Education Press, 2011.

[15] S. Yu, An Introduction to Computational Linguistics. Beijing, China: Commercial Press, 2003.

[16] Q. Zhou, “Annotation scheme for Chinese treebank,” J. Chin. Inf. Process., vol. 18, no. 4, pp. 1–8, 2004.

[17] T. Liu, J. Ma, and S. Li, “Building a dependency treebank for improving Chinese parser,” J. Chin. Lang. Comput., vol. 16, no. 4, pp. 207–224, 2006.

[18] W. Peng, J. Song, N. Wang, and M. Kang, “Traditional Chinese grammar and its potential application in Chinese information process,” J. Chin. Inf. Process., vol. 26, no. 4, pp. 50–60, 2012.

[19] J. Li, The New Chinese Grammar. Beijing, China: Commercial Press, 2001.

[20] W. Peng, J. Song, Z. Sui, and D. Guo, “Formal schema of diagrammatic Chinese syntactic analysis,” in Proc. Chin. Lexical Semantics Workshop (CLSW), Beijing, China, 2015, pp. 701–710.

[21] W. Peng, “Digital platform construction of sentence-based grammar and its application study,” Ph.D. dissertation, Dept. Chin. Philol., Beijing Normal Univ., Beijing, China, 2012.

[22] J. He, W. Peng, J. Song, and H. Liu, “Annotation schema for contemporary Chinese based on JinXi Li’s grammar system,” in Proc. Chin. Lexical Semantics Workshop (CLSW), Zhengzhou, China, 2013, pp. 668–681.

[23] J. He, “The construction of contemporary Chinese treebank based on the ‘sentence-based’ grammatical system,” M.S. thesis, Dept. Comput. Appl. Technol., Beijing Normal Univ., Beijing, China, 2012.

[24] T. Song, W. Peng, J. Song, D. Guo, and J. He, “The construction of sentence-based diagrammatic treebank,” in Proc. Chin. Lexical Semantics Workshop (CLSW), Singapore, 2016, pp. 306–314.

[25] Modern Chinese Dictionary, Dictionary Editing Room Linguistics, Inst. Chin. Acad. Social Sci., Beijing, China, 1979.

[26] S. Lv, Problems in Analyzing Chinese Grammar. Beijing, China: Commercial Press, 1979.

[27] B. Zhang, Descriptive Grammar of Modern Chinese. Beijing, China: Commercial Press, 2010.

[28] J. Li, Theory and Practice of Distinguishing Words From Phrases. Beijing, China: China Social Sciences Press, 2013.

[29] T. Yang, W. Peng, and J. Song, “High efficiency syntax tagging system based on the sentence pattern structure,” J. Chin. Inf. Process., vol. 28, no. 4, pp. 43–49 and 67, 2014.
[30] M. Zhao, W. Peng, J. Song, and T. Yang, “Development and optimization of syntax tagging tool on diagrammatic treebank,” J. Chin. Inf. Process., vol. 28, no. 6, pp. 26–33, 2014.
[31] H. Wang, W. Zhan, and S. Yu, “Structure and application of the semantic knowledge-base of modern Chinese,” Appl. Linguistics, no. 1, pp. 134–141, 2006.
[32] M. Kang, “Study of Chinese word segmentation system for international Chinese textbook analysis,” M.S. thesis, Dept. Comput. Softw. Theory, Beijing Normal Univ., Beijing, China, 2014.

**DONGDONG GUO** received the B.E. degree from the North China University of Science and Technology, China, in 2013, and the M.Eng. degree from Beijing Normal University, China, in 2016, where he is currently pursuing the Ph.D. degree in computer application technology. His research interests include natural language processing and knowledge engineering.

**WEIMING PENG** received the M.Sc. and Ph.D. degrees from Beijing Normal University, China, in 2009 and 2012, respectively. He is currently a Lecturer with Beijing Normal University. His research interests include natural language processing, computational linguistics, and Chinese syntactic parsing.

**JIHUA SONG** received the M.Sc. and Ph.D. degrees from Beijing Normal University, China, in 1995 and 2000, respectively. He is currently a Professor with Beijing Normal University. His research interests include language information processing and computer applications in education.