SYNERGY PUBLISHING PTE. LTD.

**Modern Electronic Technology**

http://www.advmodoncolres.sg/index.php/amor/index

# Research of Paraphrasing for Chinese Complex Sentences Based on Templates

**Zhongjian Wang*** **Ling Wang**

Guangzhou College of Technology and Business, Foshan, Guangzhou, 510800, China

ARTICLE INFO

*Article history*
Received: 19 March 2022
Revised: 26 March 2022
Accepted: 9 April 2022
Published Online: 16 April 2022

*Keywords*:
Complex sentence
Associated word
Paraphrasing template

ABSTRACT

Based on the paraphrasing of Chinese simple sentences, the complex sentence paraphrasing by using templates are studied. Through the classification of complex sentences, syntactic analysis and structural analysis, the proposed methods construct complex sentence paraphrasing templates that the associated words are as the core. The part of speech tagging is used in the calculation of the similarity between the paraphrasing sentences and the paraphrasing template. The joint complex sentence can be divided into parallel relationship, sequence relationship, selection relationship, progressive relationship, and interpretive relationship's complex sentences. The subordinate complex sentence can be divided into transition relationship, conditional relationship, hypothesis relationship, causal relationship and objective relationship's complex sentences. Joint complex sentence and subordinate complex sentence are divided to associated words. By using pretreated sentences, the preliminary experiment is carried out to decide the threshold between the paraphrasing sentence and the template. A small scale paraphrase experiment shows the method is availability, acquire the coverage rate of paraphrasing template 40.20% and the paraphrase correct rate 62.61%.

## 1. Introduction

Natural language has been widely concerned by domestic and foreign scholars. Many languages, whether written or verbal language have different expressions, Chinese is no exception. With the rapid development of computer and Internet, the massive sentence needs to be processed, including a large number of complex sentences, which requires us to paraphrase the sentence of the imminent.

According to a simple classification of the complexity of paraphrasing sentences, we can paraphrase the simple sentences and sentence rewriting. The study of simple sentence paraphrasing is relatively common, and the complex sentence paraphrasing relates to a lot of lexical and syntactic parsing, it is difficult to implement because of the need for a higher level of language processing techniques.

Careful review of a large number of documents, we found that Chinese sentence research is basically at the grammar level, the operation, a formal model of the building, representation of mathematical form and algorithm procedure and practical research are less. Especially the paraphrasing of Chinese sentence, few results can be operated in the field of natural language processing.

## 2. Analysis of Complex Sentence Theory

### 2.1 The Classification of Complex Sentence

In this paper, the classification of complex sentences is basically based on the sentence grammar literature, but

**Corresponding Author:*
*Zhongjian Wang,*
*Guangzhou College of Technology and Business, Foshan, Guangzhou, 510800, China;*
*Email: zhongj_w@126.com*


Distributed under creative commons license 4.0
DOI: https://doi.org/10.26549/met.v6i1.11421

not limited to grammar rules. According to the research needs, we can take to increase, delete, and summarize the grammar of the sentence structure, in order to facilitate the implementation of the paraphrase.

Generally speaking, simple sentence contains a subject and a predicate part, complex sentence is made up of two or more than two sentences, clauses can be subject-predicate sentence, also can is a non subject-predicate sentence.

The division of the grammar studies of sentence category, there are many differences. These differences make sentence category without a clear unified standard [1]. In this paper, the classification of complex sentences is based on Jiaoyan Jia [2], which puts the sentences into the joint complex sentence, subordinate complex sentence and multiple complex sentence in three categories. The joint sentence and compound sentence contains five kinds of small class.

The joint complex sentence can be divided into parallel relationship, sequence relationship, selection relationship, progressive relationship, and interpretive relationship's complex sentences. The subordinate complex sentence can be divided into transition relationship, conditional relationship, hypothesis relationship, causal relationship and objective relationship's complex sentences. Joint complex sentence and subordinate complex sentence are divided to associated words.

Multiple complex sentences are sentences that contain two or more relations which is one of the most difficult to be rewritten and is very low in terms of overwrite coverage.

### 2.2 Complex Sentence Semantic Analysis

Complex sentence semantic analysis results containing the word segmentation, part-of-speech tagging, and the grammar of the sentence structure analysis. Word segmentation and part-of-speech tagging is the first step of rewriting, for of complex sentence word segmentation and part-of-speech tagging, this paper adopts the ICTCLAS [3] segmentation software. Part-of-speech tagging is judging each word of the sentence in given grammatical category, determining its part-of-speech and labeling process [4].

The research target of this paper is mainly tag complex sentence that compared with tag complex sentence, no-marked complex sentence's paraphrasing is difficult, so we only extract the main part of the sentence to paraphrase such as object, predicate and subject.

The first category is the joint complex sentence of complex sentences. The joint complex sentence includes parallel, sequence, selection, progressive and interpretive complex sentences. Parallel complex sentence is composed of several clauses, each clause shows one thing, a kind of situation, a phenomenon or a particular aspect of a thing.

## 3. Complex Sentence Paraphrasing Strategy

On the basis of simple sentence paraphrasing, we try to paraphrase the complex sentences that use template method. Through construct corpus as the resources necessary to paraphrase complex sentences and through the simple sentences template combined to achieve complex sentences paraphrasing, and then expand the corpus size to further paraphrase complex sentences to lay the foundation for further study.

There are a lot of theoretical research on complex sentences, as mentioned in the literature [5] proposed three methods for long sentences into short sentences which are dispersion method, iterative method and segmentation method.

The basic principle of paraphrasing is the same for the simple sentence and complex sentence which is to paraphrase the sentence structure without changing the meaning of the sentence, we will use the following 4 kinds of sentence paraphrasing strategy:

1) Extract the sentence trunk, extraction of the main components for no-marked complex sentence and complex sentence having many clauses.

2) The sentence in a complex sentence merged into an attributive clause, other clauses remain unchanged. The sentence is a set of clauses in the same or similar structures, the scattered sentence is a set of sentence structure irregular.

3) On the basis of simple sentences, we add the two clause positional inverted which exchange the front and rear position between the two clauses. Simple sentence paraphrasing strategy includes the replacement, deletion, addition, repetition and locomotion of words.

4) For a sentence with metaphor, human and other rhetorical methods, we change the non obvious, ambiguous words to the obvious modification.

### 3.1 Template Extraction

In the process of rewriting template extraction, we use the above methods, or combination of several methods. The following template "[]" has two kinds of the contents, one is part of speech, the other is the associated word and its part of speech, there is a comma in the "< >", there is a replaceable associated word in the "{}". This thesis selects P and Q as the variable, the variable P and Q are characterized as follows:

1) P and Q are just symbols, representing different sentence elements.

Distributed under creative commons license 4.0 DOI: https://doi.org/10.26549/met.v6i1.11421 

2) The contents of P and Q can be a sentence, phrase, word, punctuation or the combination of the above.

3) In the same sentence template for the original sentence and the paraphrasing sentence, P in the original sentence template and correspondingly in the paraphrasing template is the same sentence components. Similarly, Q in the original sentence template and correspondingly in the paraphrasing template is the same sentence components.

Paraphrase the complex sentence template extraction method:

A. Word segmentation and part of speech tagging on sentence segmentation system using ICTCLAS.

B. Each word and its part of speech in the complex sentence respectively compares with each word and its part of speech in a template of in the template library, if there is the same word and the same part of speech components between sentence and template. We use position label to replace the extracted words, at the same time, the position of the label and the extracted words are in the same position in the original sentence. If there is not the same word and the same part of speech components between sentence and template, then the loop terminates.

C. In a complex sentence, the non extracted parts are bundled into a whole between the two positions. That is bundled into a block. The word is a block that had been extracted and it is the same key word with the template.

Case 1 Original sentence: 如果讳疾忌医 , 就可能小病拖成大病。

Word segmentation, part of speech tagging:

如果 /c 讳疾忌医 /i , /w 就 /d 可能 /v 小 /a 病 /n 拖 /v 成 /v 大 /a 病 /n 。/w

The template matching with the original sentence is:

[ 如果 /c]+[/i]+{, /w}+[ 就 /d]+[/v]+[/a]+[Q]

The ingredients contained in Q are: { /n, /v, /v, /a, /n }

As shown in Figure 1 and Figure 2, complex sentence is divided into 7 blocks, ingredients 1 to 7. Figure 1 and Figure 2 in the contents of the corresponding relationship, Figure 2 is a diagram of sentence components. Among them, the composition of one to six, and the template in the same key words. Elements 7 is the uncertain variables in the template, the component 7 contains the part of speech bundled into a whole as a variable Q, the contents of ingredients 1 to 7 are arranged in the order of the original sentence.

| Componen1 | Componen2 | Componen3 | Componen4 | Componen5 | Componen6 | Componen7 |
|---|---|---|---|---|---|---|

**Figure 1.** Sentence composition block diagram

| 如果/c | /i | ，/w | 就/d | /v | /a | Q |
|---|---|---|---|---|---|---|

**Figure 2.** Sentence component diagram

The template has the following four categories:

1) It doesn't contain variable template, template does not contain P or Q.

2) It has a variable template, the template is only one Q or a P.

3) It has two variable templates has P and Q, or P1 and P2, or Q1 and Q2.

4) It has three variable templates contains P1, P2 and Q, or P and Q1, Q2.

The template with several uncertain variables is more complex, template extraction in the process must be refined template, this reduces the template coverage rate.

The following is a template for the paraphrasing of different associated words:

Example 1 contains the word " 如 果 " complex sentence paraphrase.

Original sentence: *如果我听妈妈的话，我就不会拉肚子了。*

*Original sentence template:*

*[ 如果 /c]+[/r]+[/v]+[P]+<, />+[/r]+{[ 就 /d], [ 就要 /d], [ 就是 /d]}+[Q]*

*Paraphrasing sentence template:*

*[/r]+[/v]+[P]+<, />+[/r]+{[ 就 /d], [ 就要 /d], [ 就是 /d]}+[Q]*

*[ 假如 /c]+[/r]+[/v]+[P]+<, />+[/r]+{[ 就 /d], [ 就要 /d], [ 就是 /d]}+[Q]*

[/r]+[Q]+<, />+[ 如果 /c]+[/r]+[/v]+[P]

Paraphrase the template corresponding to paraphrase the sentences as follows:

我听妈妈的话 , 我就不会拉肚子了。

假如我听妈妈的话 , 我就不会拉肚子了。

Example 2 contains the word " 只有……才 " complex sentence paraphrase

Original sentence: 只有国家强盛了 , 才不会受欺负。

Original sentence template:

[ 只有 /c]+[/n]+[P]+<, />+[ 才 /d]+[/d]+[/v]+[Q]

Paraphrasing sentence template:

[ 唯有 /c]+[/n]+[P]+<, />+[ 才 /d]+[/d]+[/v]+[Q]

[ 只有 /c]+[ 在 /c]+[/n]+[P]+[ 的 /u]+[ 条件 /n]+[ 下 /f]+<, />+[ 才 /d]+[/d]+[/v]+[Q]

Paraphrase the template corresponding to paraphrase the sentences as follows:

唯有国家强盛了 , 才不会受欺负。

只有在国家强盛了的条件下 , 才不会受欺负。

## 3.2 Paraphrasong Process

In order to improve the success paraphrasing rate, input of complex sentences need to match template in the templates library, by sentence similarity calculation to find the appropriate paraphrase template. We need to set a similar


Distributed under creative commons license 4.0
DOI: https://doi.org/10.26549/met.v6i1.11421

level, the similarity threshold which determine by preliminary test.

We put forward an improved algorithm based on similarity calculation and paraphrase the flow chart as shown below:

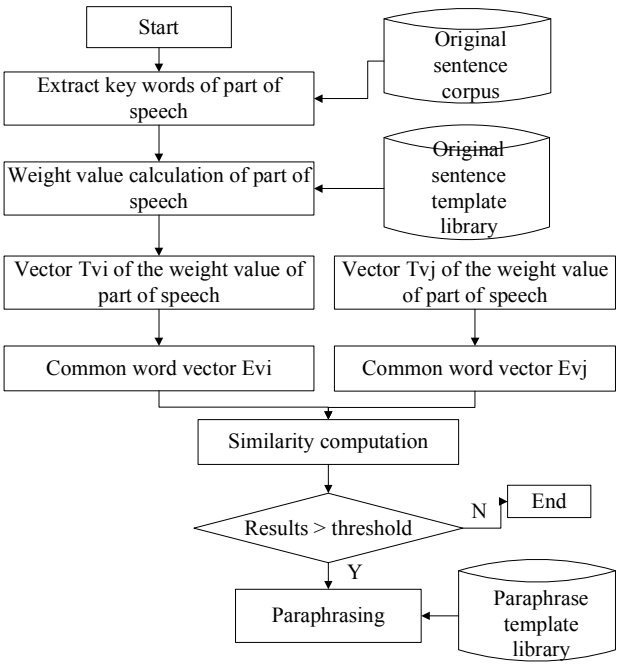



**Figure 3.** Paraphrase the flow chart

As shown in Figure 3, we calculate the similarity:

First of all, the sentence and the template, we extracted keywords and take vector representation. A given complex sentence $T_i$ vector representation of $T_i$={$m_1$, $m_2$, $m_3$,..., $m_n$}, the number of Ti words called vector of length Ti, $m_1$ to $m_n$ is $T_i$ keyword words.

Secondly, we will introduce the calculation method of the keyword weight value. The initial weight value of each word is 1/n, weights constitute the vector called the weight value vector. Keywords vector's length is Len($T_i$), key words in this method are the sentence elements contained in the template which also include punctuation marks.

Next, we will introduce the method of calculating the common word vector. Given two sentences $T_i$ and $T_j$, k and n are the length of the vector, respectively, in the $T_i$ and $T_j$, among them k<=n. Every word of the mi for $T_i$={$m_1$, $m_2$, $m_3$,..., $m_k$}, If mi is also present in vector $T_j$={$m_1$, $m_2$, $m_3$,..., $m_K$}, the vector of the same words in $T_i$ and $T_j$ is called public word vector. This public word vector and keyword vector are the same, they are expressed as $E_{i,j}$={$e_1$,$e_2$,...,$e_p$}.

Finally, the similarity between the sentence and the paraphrasing template is calculated, similarity degree formula is shown below:

$$Sim(T_i, T_j) = \frac{\sum_{k=1}^{p} v_k + \sum_{l=1}^{p} v_l}{\sum_{i=1}^{n_1} v_i + \sum_{j=1}^{n_2} v_j} \times \frac{2Len(T_i)}{Len(T_i) + Len(T_j)} \quad (1)$$

In (1), $v_k$ represents the value of item K in the common word vector Evi.

In this formula, the calculation method of the weight value is as follows:

If any one of the key words $w_i$ in $T_i$ or the synonym of the keyword appears in $T_j$, and in $T_j$ and $T_i$, $w_i$ and $w_{i-1}$ are equal or are synonymous with each other, and the corresponding weights of Tbi value bi to increase the α times, in the same way, in $T_j$ and $T_i$, $w_i$ and $w_{i+1}$ are equal or are synonymous with each other, the corresponding weights of Tbi value bi also increase α times, after several tests to determine the alpha =1.3. If the $w_i$ not in $T_j$, the Tbi corresponding weight value remains the same.

After a lot of preliminary experiments we got a paraphrasing threshold of 0.7598, the similarity of input complex sentence template and template library of up to 75.98%, we can paraphrase the sentence according to the template.

## 4. Paraphrasing Experiment and Results Analysis

### 4.1 Experiment Procedure

We randomly selected 1500 sentences with associated words from the joint and compound complex sentence, corresponding template sentence is 603. We use the word segmentation software to carry out word segmentation and part of speech tagging, the original sentence corpus is a sentence that has been marked by word segmentation and part of speech tagging.

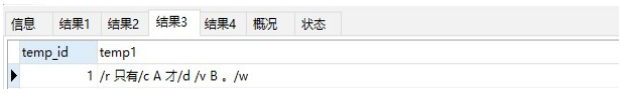
The experimental process is divided into two steps, one is needed to create a database, two is to write programs.

### 4.2 Experimental Results Analysis

In the process of manual checking paraphrasing results, we found that the small errors in the template have a great impact on the paraphrasing results. The absence of spaces will not only make a serious error in paraphrasing the results, but also the lack of spaces of different locations in the same template can lead to a lot of different errors in the result. The absence of a comma and period has a negligible effect on the correct rate of paraphrasing. Error types are the following, respectively, give examples:

(1) The original sentence missing comma in the template, such errors account for 77% of the total errors,such as Figure 4.

Distributed under creative commons license 4.0
DOI: https://doi.org/10.26549/met.v6i1.11421



| temp_id | temp1 |
|---|---|
| 1 | /r 只有/c A 才/d /v B 。/w |

**Figure 4.** Original sentence template

| sum | re_id | or_id | temp_id | or_temp_id | result |
|---|---|---|---|---|---|
| 0.8 | 1 | 1 | 1 | 1 | 我们唯有在假期里，，才可以出去旅游。 |
| 0.8 | 2 | 1 | 1 | 2 | 我们唯独在假期里，，我们才可以出去旅游。 |

**Figure 5.** Error type 1

Paraphrasing results with two comma is because the original sentence and paraphrasing template each have a comma, there is no period in the original sentence template, adds a full stop to the variable A in the process of program processing, a comma in the paraphrasing template is also added to the paraphrasing result, so there are two comma in the result, as Figure 5.

(2) Phrase collocation error, this error accounted for 18% of the total error, as Figure 6.

| sum | re_id | or_id | temp_id | or_temp_id | result |
|---|---|---|---|---|---|
| 0.77 | 22 | 7 | 7 | 16 | 我们可以邀约出去春游，只要明天天气晴朗。 |
| 0.77 | 23 | 7 | 7 | 17 | 我们就可以邀，如果明天天气晴朗约出去春游。 |

**Figure 6.** Error type 2

Without considering the clause phrase collocation, the sentence did not exchange the position and previous clauses together, programming is not reasonable.

(3) Long sentence similarity is low, this error is 5% of the total error, as Figure 7.

| sum | re_id | or_id | temp_id | or_temp_id | result |
|---|---|---|---|---|---|
| 0.47 | 31 | 8 | 6 | 12 | 发芽可以证明它还活着，只要这颗树的枝条还在。 |
| 0.47 | 32 | 8 | 6 | 13 | 发芽可以证明它还活着，如果这颗树的枝条还在。 |

**Figure 7.** Error type 3

The experimental data included rewriting correct rate, the template coverage, and the rate of not being rewritten, the following specifically introduces the calculation method of all kinds of data.

We define the total number of sentences as Psum, in the result, the total number of sentences to be rewritten is Pasum, paraphrase the correct number of sentences is Rres, one of the original sentences only corresponds to a correct paraphrasing sentence. The total number of templates is Tsum.

The proportion of the sentence that has not been paraphrased is shown in the (2):

$$NPrate = \frac{Psum - pasum}{Psum} \times 100\% \quad (2)$$

Paraphrase correct rate calculation as shown in the (3):

$$PRrate = \frac{Rres}{Psum - pasum} \times 100\% \quad (3)$$

The formula for calculating the template coverage is shown in (4):

$$Trate = \frac{Tsum}{Psum - Pasum} \times 100\% \quad (4)$$

According to paraphrase the correct sentence and the total sentence compared, the proportion of sentences that have not been rewritten by the (2) is 7%. The correct rate of rewriting is 62.61%, which is obtained by the (3). The template coverage rate was 40.2%, which was obtained by the (4).

## 5. Conclusions

This paper presents the method of paraphrasing Chinese sentence based on template, by building to associated words as the core of the corpus, provides the basis for sentence paraphrasing. The experimental results show the effectiveness of the method and its deficiency.

The template coverage rate and correct rate is the key to paraphrase the sentences based on template. In the process of rewriting the sentences, we need a further deeper level of syntax and semantic analysis of sentences, get a more efficient paraphrasing template, raise paraphrasing accuracy and template coverage.

## References

[1] Rinaldi, F., Dowdall, J., Moll, D., et al., 2003. Exploiting Paraphrases in a Question Answering System. Proceedings of Workshop in Paraphrasing at ACL2003, Sapporo, Japan.

[2] Li, W.G., Liu, T., Zhang, Y., et al., 2005. Automated Generalization of Phrasal Paraphrases from the Web. The 3rd International Workshop on Paraphrasing. Jeju Island, South Korea. pp. 49-57.

[3] ICTCLAS (Institute of Computing Technology, Chinese Lexical Analysis System): http://www.ictclas.org/index.html.

[4] Zhao, Sh.Q., Liu, T., Yuan, X.Ch., et al., 2007. Automatic Acquisition of Context-Specific Lexical Paraphrases. Proceedings of IJCAI, Hyderabad, India. pp. 1789-1794.

[5] Wang, Z., Wang, L., 2010. Paraphrase of Chinese Sentences Based on Associated Word. ASIA-ICIM 2010, Wuhan, China.


Distributed under creative commons license 4.0
DOI: https://doi.org/10.26549/met.v6i1.11421