The Formal Semantic Analysis on Two Types of Chinese Sentences and Programming Realization

Abstract: The Verb-driven Chinese natural language understanding system realizes logic-driven semantic analysis. There are two types of Chinese sentences －把(ba)- sentence and 被(bei)- sentence. In this paper we use λ-transformation to analyze them. Yan Jiang and Haihua Pan put out formal semantics theory based on the concepts of type and category [3]. We program to realize semantic level analysis and expression, try to realize the idea to algorithm, and formalize natural language to logical expressions. According category theory to analyze semantics and transform the concept of category to the "field" in corpus. Finally we program to control the flow of analysis process and realize the logic-driven semantic analysis system.

Keywords: category; λ-transformation; logic-driven;把-sentence;被-sentence

Authors: Chuan Zhao[1] Yuanyuan Wang[2] Yao Lu[3] Yu Liu[4] Dezhi Tu[5]
Department of Computer Engineer of Information and Engineering college,
Cheng Du University of Technology，610059
Telephone: 028-89067079/13402858729
Email: zhaoc@cdut.edu.cn

* Sorry this picture is scanned from my passport of 2009. I should have time to take a new one while submit later after received.

The Formal Semantic Analysis on Two Types of Chinese Sentences and Programming Realization

Chuan Zhao[1] Yuanyuan Wang[2] Yao Lu[3] Yu Liu[4] Dezhi Tu[5]

Department of Computer Engineer of Information and Engineering college, Cheng Du University of Technology, Cheng Du 610059，China

zhaoc@cdut.edu.cn

Abstract: *The Verb-driven Chinese natural language understanding system realizes logic-driven semantic analysis.There are two types of Chinese sentences –把 (ba)- sentence and 被(bei)- sentence. In this paper we use λ-transformation to analyze them. Yan Jiang and Haihua Pan put out formal semantics theory based on the concepts of type and category [3]. We program to realize semantic level analysis and expression, try to realize the idea to algorithm, and formalize natural language to logical expressions. According category theory to analyze semantics and transform the concept of category to the "field" in corpus. Finally we program to control the flow of analysis process and realize the logic-driven semantic analysis system.*

Keywords: *category; 把 -sentence; 被 -sentence; λ-transformation; logic-driven*

1. Introduction

Language is a way for interpersonal communication and information exchange. Natural language understanding (NLU) bases on computational linguistics. Formal mathematical models are abstract to analyze and process natural language. It is a foundational theoretical way to understand different actual language. We use the view of formal semantics and computing linguistics to program to gain the machine of Natural language. The formal semantics analysis deals with Chinese language text straightly from the syntax analysis level up to the semantic analysis level. Theoretically this gives a new view to the traditional theory of natural language processing. In application formal semantic analysis is the essence of NLU and search engines. It establishes a more comprehensive basis for Chinese text processing system. Computational linguistics helps us understand and compare different languages.

2. Theories and Technology of Formal Semantics

2.1 Preliminaries

First we give out the formal definitions of the two types of Chinese sentences we study in this paper.

Definition 1: 把 (ba) -sentence is a type of sentence constituted by preposition 把. Its structure often follows Rule1.

Rule 1: N 把 N V.

Rule 2: N V N.

The structure of English active sentences follows Rule 2. Object is generally placed after verb. Compared with them, the object in 把-sentence will be put in front of verb. It can emphasize behavior outcomes or ways. This type of sentence can be regard as object front and expresses an active meaning. Some time Chinese express active sentence as rule 2 the same. And 把-sentence is a distinctive type for such meaning. Of course it should carry some new information in speech, and here we should not compare them.

Definition 2: 被 (bei) -sentence is a type of sentence constituted by preposition 被. Its structure is Rule 3.

Rule 3: N 被 N V.

Rule 4: N is V-ed by N.

The structure of English passive sentences follows Rule 4. Deferent from English, subject and object are both put in front of verb in 被-sentence. But the two types both express passive meanings.

2.2 Phrase Structure of Grammar Analysis

Chomsky proposed Phrase structure grammar in the "syntactic structure" in1957. It is commonly used in natural language parsing. He put out a framework called Transformational Grammar that combines a context-free phrase-structure grammar with another component of transformations that specify how trees of a given form can be transformed into other trees in a systematic way. In this framework the predicate-argument relations of a sentence are determined not by the arrangement of nodes in its surface tree but by the node configurations of the deep structure from which the surface tree is derived. The active-passive and other systematic paraphrase relations can be expressed. As an illustration, the following transformation is a rough version of the passive rule in English.

Rule 5: NP V⟶NP 3 be 2 + pastpart by 1 [6]
1 2 3

2.3 Categorical Grammar

Categorial grammar is a syntactic theory to describe natural language.It does mathematical computing to the formation of sentence from the view of generating. Categorical grammar is also a kind of gernerative grammar. It establishes “syntax type” or “category” corresponding to the words of sentence to realize “calculus” of the entire syntactic structure. There are two Basic category-types –S (sentence) and N (noun). Syntactic component can be combined with S and N in different ways according to its syntactic functions. It adds Category constructors as “/” and “\” to express its own “syntax type” and “category”. “/” and “\” respectively represent “left deficient” and “right deficient”. It has been described in detail in <Intruduction to formal semantics> put out by Yan Jiang and Haihua Pan [3]. The definition of type and category are given on the specific semantic pattern of Chinese. They are basic factors of computing and guidance in syntactic semantic operations. They are also integral parts of this system.

2.3.1 Type Theory

Type theory provides a basis for designing and analyzing type system. Two key parts of S is predicate and argument. Relatively simple statement contains only a single verb or an adjective. Argument NP is acted by the proper noun NN. From the view of extension, the proper noun refers to individual (individual entity, denoted as e). Statement means the truth value (truth-value, denoted as t). Predicate means a collection of individuals (set of entities).

Define 3: a. e (entity) and t (truth) are basic types.

b. If a, b are types, (a→b) is a type [3].

This is the defination of primary type theory. a, b can be basic types and complex types. We can combine into infinite kinds of composite types recursively.

2.3.2 Category Theory

Sentences generated by semantic types lack the requirements for the location of the sentence components. For example, a binary transitive verb should merge object from the right side at first, then merge with subjec in the left side. The solution is to introduce the ideas of category grammar. The defination of category is Define 4.

Define 4: a. /(forward slash) means searching for available argument in the right.

b. \(back slash) means searching for available argument in the left.

c. Located below the slash is input category, and above it is output category.

d. Supply A in the right of B/A is B.

e. Supply A in the left of A\B is B.

f. S, N are basic syntax categories, their logic semantic type are t and e.

g. If A, B are syntax categories, A/B, B\A are also categories.

Language type category and a series of operation rules on category are designed in the base of semantic types. Grammatical category is a cluster of the meaning expressed by words’ changing in form. Grammatical meaning has a higher abstractive meaning. It abstracted further from the meaning and usage of specific words [2].

3. Our works

Our group studied NLU science 2003. We have built a NLU system. It can do word segmentation and some semantic and syntax analysis. Our works talked about in this paper in 2009 are trying to achieve semantic analysis.

First this system makes the input text to segmente automatically. Then it does parsing, logical semantic analysis and detection. It eliminates most of the ambiguity at the same time. Last it analyzes and realizes two particular sentences in Chinese, they are 被-sentence and 把-sentence. The function expected is achieved.

We use category grammar and composition rules to analyze sentences. The lowest level sentence is divided into lexical items. Coupled with the respective type, it will generate the highest level sentence. This increases the meaning expression. Pulling in type deduction makes vocabulary is carried out smoothly to form sentence structure. The logical type is established. It makes sentence structure analysis procedure as procedure of semantic expression. By functional fitting operation, the logic semantic expression is: (Predicate (object)) (subject). If it has attributive and adverbial, the expression is: (adverbials Modified sentence) {(adverbial modification of verbs (predicates [(modifier) (object)])}(subject). This is combination principle.

3.1 λ-calculus

Each word in type theory is given corresponding syntactic categories and semantic types. Verb syntactic categories give a rule to the relevant argument position in the left or right. Verb as a premise will move in a certain direction to search argument and integrate with it. In 被-sentence and 把-sentence of Chinese, argument does not appear in the direction specified but in the opposite direction. Verbs can't find the required argument. So verbs can't merge with argument and an extra argument will appear in the whole sentence. The principle of combination stop and sentence can't be generated.

3.2 λ-abstraction and λ-reduction

We need two techniques of λ-Transformation when dealing with 被-sentence and 把-sentence. They are λ-abstraction and λ-reduction.

If transitive verbs (Vt) can't find an available argument in the right, we provide a free variables x. It merges with verb and gets V(x), which is NP. We abstract x in certain form. It occupies the position of argument. But it has no substantial content. Verbs go on to merge with other sentence category. As soon as it finds a real available argument, such as logic object, the object will replace the position of variables abstracted.

3.2.1 λ-abstraction

λ operator acts on free variable x of propositional function, for example (a). We abstract it and get (b). The expressions mentioned above are defined as (a) P(x). (b) λx[P(x)]. This procedure is λ-abstraction.

If we set φ for t class [S category] expression, φ contains free variable x and x is e class [N category] expression, λx [φ] is a legal expression of e → t type [S / N category][3]. Take (a) for an example, its semantic type is t and its sentence category is S. We use λ-abstraction to (a) and get (b). The argument x of (b) is abstracted and lost its substance. So the semantic type of (b) can be understood as the type to be subtracted an argument e from the original propositional function.This is t-e. If we provide e for the expression, it will revert to t. Therefore, (t-e) can be expressed as (e→t). It is also the semantic type of (b).

3.2.2 λ-reduction

We substitute an individual in domain into λ-bound variable x and eliminate λ-operator. This procedure is λ-reduction. In the procedure of λ-reduction, you must replace the entire same variables when substituted into the individual of λ-abstract style. It is also called β Transformation in formal semantics.

If we setλα[φ] as a legal expression of A→B class[B/A category], φcontains variableαand T is A class [A category] expression, λα[φ](T) is B class [B category] expression[3]. Substitute T into all the position of α inφ. It will be equivalent withλα[φ](T) in type [category] and the truth conditions. From the view of type and category, A→B in the type of λα[φ](T) merged with A. It must be B class. The category B|A added A must be B category. This is completely the same with the type and category gotten from substitution T intoφ expression.

3.2.3 被-Sentence and 把-Sentence

(1)被-sentence: the passive semantics in Chinese can be expressed by 被-sentence. For example: 西游记/被/张三/买/了(Xi You Ji/ Bei/ Zhang San/ Mai/ Le). Here Xi You Ji is a famous book. Zhang San is a name of a person. It is

bought by Zhang San.

Grammatical subject of this sentence is logical object. It is the argument that verb could not find in the right. In logical expression, grammatical subject of a passive sentence (named the logical object) should merge with verb first and then combine with noun. But the argument adjacent to verb is a logical subject. If verb merge with logical subject first, it will violate the combination principle.

(2)把-sentence: some object of 把 in Chinese can be regarded as object of Vt. For example: 三毛/把/苹果/吃/了(San Mao/ Ba/ Ping Guo/ Chi/ Le). Translate this sentence into English: San Mao eats an apple. San Mao is a name of a person.

Vt can't find an available argument in the right, although object is its neighbor. We can look 把 as an operator. It combines with the right argument and gets new object argument.

3.3 Algorithm and Flowchart

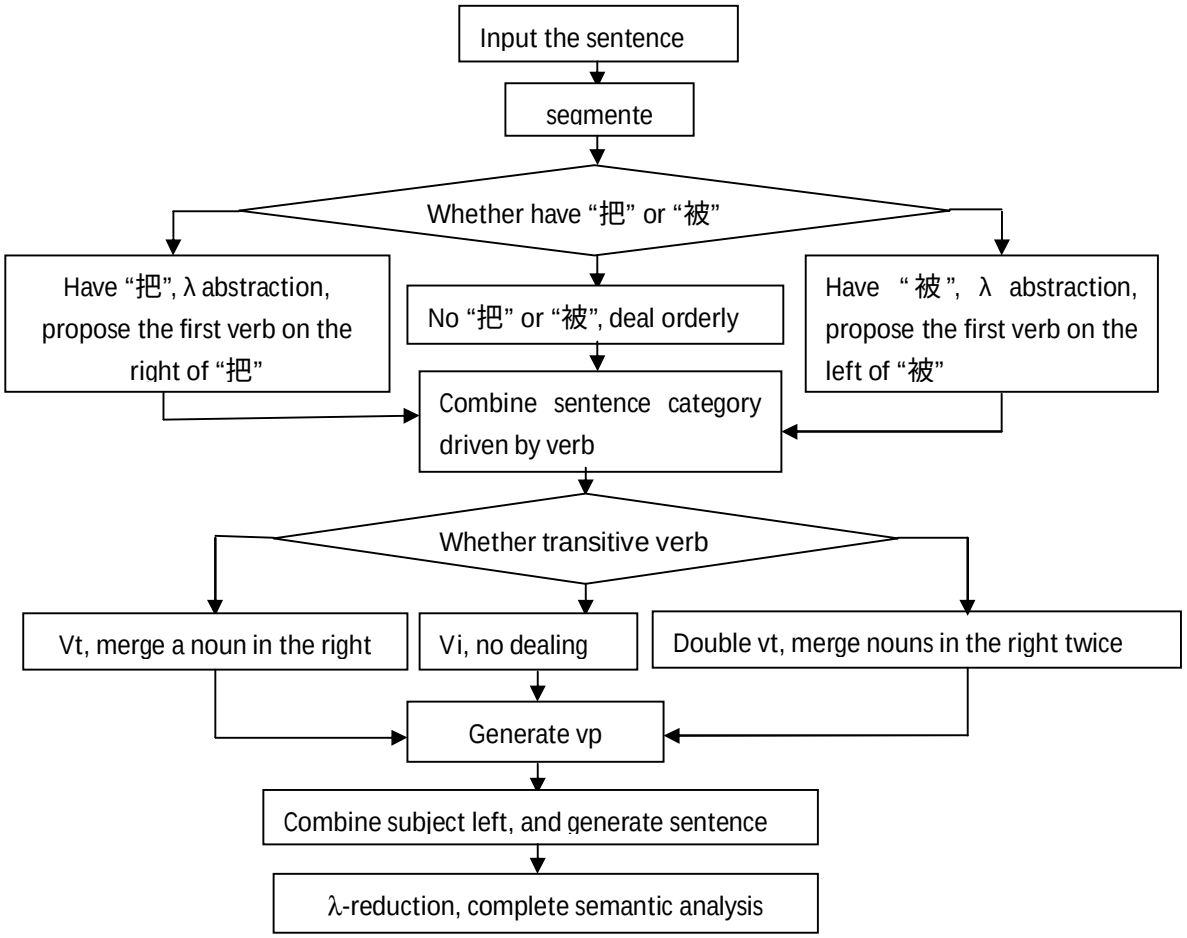

Figure 3-1 λ-calculus flow chart

For figure 3-1, λ-calculus algorithm of Chinese Sentences Analysis is:

(1)Put words and attributes into a predetermined array in corresponding sequence after words segmetation;

(2) Judge preposition parts containing feature words of "把" and "被" or not ;

(3)If there is feature word, put it to λ-abstraction. Replace the lack position with an unknown variable. Extract the phrases which have an influence on the order of the normal semantics;

(4) Combine the normal semantic

(5) λ-reduction. Put the phrases extracted into the location of the unknown variables and complete semantic analysis.

Fill free variables in 把-sentence and 被-sentence. Abstract them and obtain the required structures. When encountered with the shared argument in such a structure, replace the relevant variable with it. Complete the analysis of whole sentence and meet combination principles.

Compared with general logical calculus, λ-calculus introduces a free variable to combine into a true propositional function of value class. Like the introduction of an additional hypothesis, then use λ-abstraction to eliminate the free variable. Abstract relevant logic free variables to the variable bounded by λ-operator. This is also a procedure of eliminating additional assumption. When λ-abstract is applied to logistic statement, it can be applied to both language-based catetories and semantic types. Language units abstracted belong to categories / types. Put them back into category / formula types.

4. Experimental Results

4.1 Pseudo-code Implementation

```
for (i = 0; i < j; i++)
{
    int x = i;
    if ((poss[x] == "p") && (word[x] == "被"))    //judge the "被"sentence
    {
        tempword = word[x - 1];
        tempposs = poss[x - 1];
        while (x<j)
        {
            word[x] = word[x+1];
            poss[x] = poss[x+1];
            x++;
        }
        x--;
        word[x] = tempword;
        poss[x] = tempposs;
        break;
    }
}
```

When we deal with these particular phenomenon, we adopt placeholder idea to convert them to a normal order to achieve sentence analysis. Such as 被-sentence, We first use a temporary variable to occupy the location in front of 被, and then shift the words behind 被 in turn, last substitute the value of the temporary variable to the specified location. 把-sentence can be available similarly.

4.2 Running Results

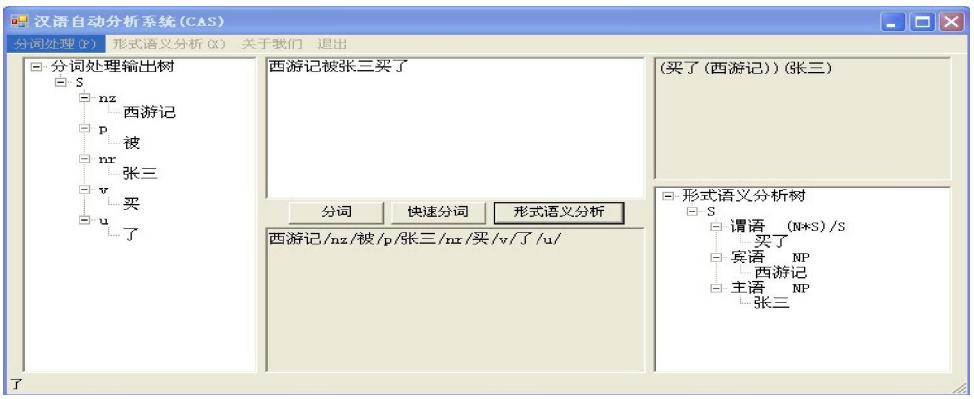

Figure 3-2 analysis results of "被" sentence

Figure 3-2 shows a result of 被-sentence. We input a sentence in the upper middle place. Click words segmatation. It will show the result with category constructors in the lower middle place. In the left it shows the tree state.The level expressions will be in the upper right place. Click formal semantic analysis. In the lower right position, the formal semantic analysis tree can be seen. It marks predicate, object and subject.

5. Summaries

This paper discusses a syntactic and semantic analysis system. It bases on formal semantics and type theory. The purpose of syntactic and semantic Analysis to some Chinese classic sentence is achieved. The ideas of variable occupancy is used to do λ-abstraction and combine with verb first for 把-sentence and 被-sentence in Chinese. Then analyze sentence with normal semantics. It combines

subject and verb first and meets combination principles. Formal semantic analysis system with verb logically-driven uses λ calculus to complete semantic analysis. Compared with rule-based analysis Limited to syntax, it is more suitable for Chinese operation. The definition of category and type started entirely from Chinese grammar. It can realize Syntactic and Semantic Analysis to partial sentences in modern Chinese. NLU research can enter semantic level from syntax level. Semantic analysis is its cornerstone and starting points.

Yinque Chen said 80 years ago: "now the Indo-Europeanization grammar, the Ma's grammar。 But it can not used in different languages' Chinese, and the compare with other languages in the same languages has been in the early period. The real Chinese grammar has not created yet. That is why it is very hard. To an actual language only part fit the general laws of world language, most are special phenomena, can be abstracted to general rules, then we can gain an individual system theorem, that is the rules of this special language which should not be defined according to another language. None can promote and abstract all kinds of languages. "[12] Our works practice such important opinion. Computational linguistics let it is possible to understand and mast an actual language deeper, and can compare languages clearer.

Such research process let us think of we can abstract and understand Chinese farther and create new ways to teach Chinese to people speak other languages.

Then human being can have a more detailed and nice panorama of languages. That should help people speaking different language understand well and create new higher civilization.

References

[1]Daniel J, James H. *Comprehensive theory of natural language processing. Electronics Industry Press*. 2005.

[2]Sun HJ. *Categorical grammar and its application in Chinese. Modern language*. 2008.

[3]Jiang Y, Pan HH, *formal semantics. China Social Science*. 2008.

[4]Cai ZX, Xu GY. *Artificial Intelligence and Its Application . Tsinghua University Press*. 2004.

[5]Miao JA. *Chinese language semantic structure and complement form. Shanghai Foreign Language Education Press*. 1990.

[6]Ruslan M. *The Oxford Handbook of Computational Linguistics. Foreign Language Teaching and Research Press*. 2009.

[7] Yao TS, Zhu JB. *Natrual Language Understanding .Tsinghua University Press*. 2002.

[8]Liu Y. *Computational Linguistics. Tsinghua University Press*. 2003.

[9]Lu ZW. *mathematical logic in computer science. Science Press*. 1996.

[10]Su XC. *Quantitative Study of Chinese Vocabulary. Xiamen University Press*. 2002.

[11]Huang ZY. *HNC theory. Tsinghua University Press*. 1998.

[12]Ji XL. Mourning and Memory. *Hua Yi Publishing House* 2008.