Semantic MMT Model based on Hierarchical Network of Concepts in Chinese-English MT

1, 2 Wen Xiong
1 Beijing Normal University, Institute of Chinese Information Processing, 2 CPIC-BNU Joint Laboratory of Machine Translation, Beijing, China
Corresponding author
Email: stevens7979@sina.com

1, 2 Yaohong Jin
1 Beijing Normal University, Institute of Chinese Information Processing, 2 CPIC-BNU Joint Laboratory of Machine Translation, Beijing, China
Email: jinyaohong@bnu.edu.cn

Abstract—To study the generation of the semantic tree of Chinese sentence in Chinese-English Machine translation (MT), a new semantic-analysis model of Chinese multiple-branched and multiple-labeled tree (MMT) based on the hierarchical network of concepts (HNC) is proposed. Supported by word and rule knowledge-base of HNC, the model executed the semantic analysis using static and dynamic labels as a complex feature of MMT instead of a single feature of phrase structure grammar, and generated a HNC-MMT semantic tree for deep understanding of the semantic of Chinese sentence. Based on the semantic tree generated, the model can realize structure conversion of semantic chunks, and utilize a hybrid strategy of the statistical and rule-based to translate. Experiment shows one of the most important tasks of semantic analysis of Chinese sentence, the global eigen-chunk recognition, achieves accuracy above 85%, verifying the effectiveness. The model has been applied to system development of MT based on HNC.

Index Terms—machine translation (MT), semantic analysis, hierarchical network of concepts (HNC), multiple-branched and multiple-labeled tree (MMT), syntactic analysis

I. INTRODUCTION

As a sub-field of computational linguistics, machine translation (MT) is a research using computer software to translate text or speech from one natural language to another. Natural language understanding (NLU) and natural language processing (NLP) are two important means and technologies used by MT. The researches of MT can be divided into two categories [1], i.e., rule-based and corpus-based. The former includes transfer-based [2], interlingua-based [3] [4], constraint-based formalisms [5], principles-based [6], and lexicalist approaches [7], and the latter includes statistics-based [8], example-based [9], and connectionist approaches [10]. In recent years, there is a new hybrid machine translation (HMT) emerged [11], which combines the advantages of rule-based and statistics-based approaches, including deep integration and shallow integration. For example, there are methods of rule post-processing by statistics and statistics guided by rule.

The statistics-based MT has gained good effectiveness. However, there are still some cases, which do not meet the semantic constraints in the translations. Thus, it is necessary to understand source text deeply using the NLU technique to improve the quality of the translated texts, involving the semantic analysis of sentences of the source texts. In this case, the rule-based method is more conducive to merging the expert’s knowledge to analyze the semantic info of source sentence.

Hierarchical network of concepts (HNC) [12] is an important theory for semantic analysis of Chinese, which expresses the knowledge of a linguist with word knowledge-base of HNC and rule knowledge-base of HNC. The semantic analysis in HNC is to capture the deep semantic structure and semantic expression of Chinese sentence using the HNC knowledge in the process of NLU.

In this paper, we propose a novel semantic-analysis model of Chinese multiple-branched and multiple-labeled tree (MMT) based on HNC in Chinese to English machine translation, which parses the semantic of Chinese sentence and generates HNC-MMT of Chinese sentence using the semantic symbol and analysis means of HNC. The semantic HNC-MMT can be used in the generation of translated text. The rest of the paper is organized as follows. Section 2 surveys the related work on MMT, MT based on HNC, and HMT. Section 3 introduces the MMT model of Chinese information processing (CIP) and the label content of it. The proposed HNC-MMT is presented in detail in Section 4. Section 5 provides the experiment conducted. Finally, Section 6 concludes the conclusion of the paper.

II. RELATED WORK

The analytical capacity of phrase structure grammar is limited, and the generative capacity of it is too strong, resulting in an ungrammatical content generated. Feng et al. presented MMT model of Chinese information [13] to overcome the defects of the phrase structure grammar and
to meet the requirements of the computer processing of Chinese in 1983. Linear and hierarchical structures are shown among linguistic symbols in the same time, which cannot be presented by a single feature. Thus, functional unification grammar [14] was proposed, which parses the sentence for the multiple characteristics of the linguistic symbols. The translation of Chinese sub-sentence is a difficulty in the MT of Chinese-English patent, thus, the model of degraded sentence was introduced, and related rules were proposed [15] to form the translation algorithm for the Chinese sub-sentence. For the solid expressions in MT of Chinese-English patent, especially for the expression of the claim sentence of Chinese, the rule-based method [16] can improve the quality of translations effectively.

The selection of the main verb in the case of multiple verbs used together, and the decision of the boundary of the long noun-phrase (NP) are still two difficulties of the syntactic analysis of Chinese. However, the semantic analysis based on HNC theory and the use of principle of language-logic dynamic-representation (lv) [17] can handle the two situations effectively. Integrated with the system of the syntax trees, the research formed a hybrid-strategy method, which can help to improve the performance of the MT of Chinese-English patent. Passive voice is often appeared in patent document. Thus, it is an important step of MT to recognize the corresponding Chinese sentence having a passive expression, and then to handle the generation of translated text using a rule-based method [18]. Element sub-sentence is widely existed in Chinese patent documents. Most of the problem of Chinese-English MT about the element sub-sentence can be perfectly resolved on an online system of patent MT in SIPO by the analysis of semantic structure of element sub-sentence in three types, and the proposed rules of Chinese-English MT [19].

The analysis of long Chinese sentence is the fourth difficulty. Thus, the method can improve the performance of MT of Chinese-English patent by the segmentation of long Chinese sentence using features coming from HNC theory [20]. The annotated method of corpus of sentence-level semantic [21] annotates the semantic info in three categories: sentence category, semantic chunk, and sub-sentence included in the semantic chunk, which is a degradation of sentence or a chunk extension. The corpus annotated by the method has been an important knowledge resource for Chinese information processing and foreign language study.

There are different directions in hybrid methods. For example, a hybrid approach is to integrate information from a rule-based machine translation system into a statistical machine translation framework [22]. The techniques of hybridization are grouped into three parts: the morphological, lexical, and system level. As an opposite direction, another method is to add statistical bilingual components to a rule-based system [23], which has a higher degree of grammaticality than a phrase-based statistical MT system, where grammaticality is calculated according to correct verb-argument implemented and translation of long-distance dependency.

III. MMT Model of CIP

One of the most important features of the MMT model is the adoption of multiple-labels to describe the Chinese sentence, which can be seen as a complex-feature to describe the Chinese sentences. In the MMT model, the concept of "multi-value label-function" was proposed. Many MT systems have adopted the Chomsky's phrase structure grammar as a theoretical basis of system design. According to the phrase structure grammar, each node of the syntax-tree obtained by parsing only has one corresponding label, which can be expressed by single-value label-function L as: L (node) = label, where the node expresses a node of the syntax-tree, and the label represents a corresponding label of the node. The relation between the node and the label is a many-to-one relationship. Because the language feature represented by the single-value label-function is quite limited, a large number of ambiguous structures are generated for long sentence. Thus, there will be many different syntax-trees for the same long sentence, which has brought great difficulty to disambiguation and a significant increase in the overhead of parsing. Unlike the phrase structure grammar, MMT model uses a multi-value label-function to replace the single-value label-function, which can be expressed as: M (node) = \{label_1, label_2, ..., label_n\}, where the label of a node is no longer one, but corresponds to multiple labels \{label_1, label_2, ..., label_n\}. Thus, the method of multi-value label-function improves the ability in knowledge representation and knowledge description. Each node can record as much as possible grammatical and semantic features. These features compose a complex feature to overcome the weaknesses using phrase structure grammar to describe Chinese sentence fundamentally and to become a new kind of syntactic analysis model of Chinese.

Since the model construction of the syntactic analysis is an important composition of knowledge representation from the viewpoint of artificial intelligence, the resolving of this problem can serve knowledge reasoning, and to form finally intelligent decisions.

A. MMT Definition

The MMT [24] is a knowledge-representation model of syntactic analysis defined as follows:

1) An MMT has and only has one root;
2) If the root has a child node, then, every child node is an MMT;
3) Any node in the MMT has child nodes with the number from 0 to n. If a node has not a child node, the node is a terminal node (a leaf); otherwise it is a non-terminal node;
4) The label in each of the nodes of MMT is a set of multiple labels.

B. Label Content of MMT of CIP

There are two kinds of label in MMT. The label kind is static if it can be given in a lexical independently, or if it is inherent in the word itself, and the label kind is dynamic if it is generated when the relation between words is occurred. Therefore, the dynamic label is added.
into the MMT along with the procedure of syntactic analysis gradually.

For automatic analysis of Chinese sentences and automatic generation of syntax tree, MMT adopts a combination method using the static label of word knowledge and the dynamic label of syntax, semantic, and logical function. The content of static labels of MMT is described as follows.

- **The part of speech (POS) tag**: nouns, premises words, position words, time words, difference words, numerals, quantifiers, body sexual pronouns, predicate pronouns, verbs, adjectives, adverbs, prepositions, conjunctions, particles, modal particles, onomatopoeia, and interjection.

- **The phrase type tag**: the verb phrase, noun phrase, adjective phrase, and the number and quantity phrases.

- **The inherent semantic label of words**: images, materials, phenomena, time and space, measure, abstract, attributes, and actions.

- **The inherent grammar label of words**: a different noun required a different quantifier, the different valence of verb, and substance properties.

The content of dynamic labels of MMT is described as follows.

- **The labels of syntactic function**: subject, predicate, object, attribute, adverbial, complement, adnex, the center language.

- **The semantic relationship tags**: agent, patient, dative, involved, moment, period, start time, spatial points, space segment, starting point of space, end point of space, initial state, final state, causes, results, tools, way, purpose, conditions, roles, concept, scope, modification, comparison, accompanying, judgments, statements, attached and so on.

- **Logical relationship labels**: argument 0 (deep subject of sentence), argument 1 (deep direct object of sentence), and argument 2 (deep indirect object of sentence).

**IV. HNC-MMT**

The HNC theory is founded by Z.Y. Huang [12], which is a theoretic framework of NLU oriented. Based on the researches of traditional sinology and modern linguistics, HNC starts with the semantics of the language, and uses conceptualization, hierarchy, and network as a basic means, getting rid of the shackles of the syntactic analysis using formal language theory. With foundation of language concept space and methods of the formal, HNC links the surface structure and deep semantics of natural language using the sentence category (SC), and is a unified theory of the syntactic, the semantic, and the pragmatic.

HNC theory simulates the cognitive mechanisms of human brain, and divides the cognitive structure of the human brain into local and global types of the associative skeleton. It thinks of the expression of the associative skeleton is the deep fundamental problem of the language.

The local association is of a lexical level, and the global association is of a sentence level and a chapter level. The concept representation-system of HNC theory focuses on the expression of abstract concepts, which uses a five-tuple for the diversity of the abstract concept, and uses a network-hierarchical symbol for the connotation of it. Three semantic networks were proposed, such as semantic network of primitive concept, that of basic concepts, and that of logic concept.

HNC thinks of that there is a conceptual space existed in the human brain, which is a base for human to recognize the world and to think. The existence of the conceptual space is a basic assumption of the HNC, and is one of the axioms of the HNC. The language concept-space is a subspace of the whole concept-space, which is used for human to understand and to apply the natural language, and corresponds to the natural language space, so human beings have a common language concept-space and a variety of natural language spaces. Therefore, there is a one-to-many mapping between the language concept-space and the natural language space.

The concepts of HNC are divided into abstract concept and concrete concept. The former refers to the concepts expressed by object do not have physical attributes, and the latter refers to the concepts expressed by the object have a physical property. Compared with the concrete concepts, abstract concepts have primitive and systematic characteristics, so HNC semantic network is designed for abstract concepts, but it expresses the concrete concepts by linking it to the abstract concept. The HNC semantic network is a hierarchical structure; each layer has some nodes, called conceptual nodes, which are labeled with the numbers 0-13, wherein the numbers 10-13 expressed as hexadecimal lowercase letters: a, b, c, and d.

The semantic network of the basic concepts has a total of nine one-level nodes; that is, \( j_0 \) expresses sequence and generalized space; \( j_1 \) is for time; \( j_2 \) is for space; \( j_3 \) is for the number; \( j_4 \) is for quantity and range; \( j_5 \) is for quality and class; \( j_6 \) is for the degree; \( j_7 \) is for the basic properties of objective; \( j_8 \) is for the basic properties of subjective.

The semantic network of primitive concepts has 14 one-level nodes, divided into two categories. The numbers of 0-5 belong to one category, called principally primitive concept, and the numbers of 6-9 belong to another category, called primitive concept of extension. The principally primitive concepts consist of six one-level nodes; that is, \( 0 \) is for action; \( 1 \) is for process; \( 2 \) is for transfer; \( 3 \) is for effect; \( 4 \) is for relation; \( 5 \) is for state. The six nodes are the six basic viewpoints for natural language to represent everything as a whole, and the six basic chains for everything happening, developing, and disappearing, called action-effect chain in HNC.

The primitive concepts of extension describe the content including activity of physiological instinct, the mental, thinking, the intellectual, the professional, the social, pursuit, and stipulation. The primitive concepts of extension are designed for the natural language to describe human activity, which are a compound of principally primitive concept and basic concept.
Logic concepts are divided into two categories, such as language-logic concept and basic-logic concept. The former has 12 one-level nodes, including single identifiers of main semantic block, whose purpose is to build a variety of semantic signs and symbols that serve the analysis of SC and apperceiving of semantic-chunk. The basic-logic concept has two one-level nodes: comparison, and basic judgment.

The semantic network of HNC has the basic characteristics of conceptualization, the primitive, hierarchy, and network. Any node in the semantic network corresponds to a concept, but one natural language has not necessarily the word having the meaning corresponding to the concept. The nodes of semantic network must be no ambiguity, but the word may be ambiguous, since a word can have several different meanings.

The concepts represented by the nodes of semantic network are primitive concepts, whose number is finite, but the combinations of them are infinite. The structure of concepts is hierarchical with high level and low level. For example, "deserved thing" is high-level concept, and wealth, experience, lesson, and knowledge are underlying concepts.

The hierarchy of semantic network embodies the gradual expression approach to the concept in HNC, from high-level to low-level, which helps the establishment of the associative skeleton of concept, and expresses the natural connection between the high-level concept and underlying concept. Since the characteristic of network is inherent in the concept system, there is a longitudinal and transverse correlation between the nodes of concept, so the characteristic of network is an important part of semantic network.

HNC expresses the global associative skeleton using semantic chunk and SC, and proposes the axiom of the action-effect chain, which reflects the largest common of everything and describes the basic rule of existence and development of everything in the universe by using six chains: action, process, transfer, effect, relation, and state.

To illustrate the basic content of the human mind, HNC theory introduces the judgment as a supplement of the action-effect chain, which is a response of subjective against objectivity and emotion, therefore, the action-effect chain and judgment fully express the relationship between subjective and objective, the rational and the emotional, and their relations. They constitute the generalized action-effect chain, which are the basis for the classification of principally primitive concept and the basic for the partition of semantic category of sentence (SC).

In HNC, the SC includes 57 basic SCs, 3192 mixed SCs, and more than 10 million compound SCs. 57 basic SCs can be divided into seven kinds, such as the action sentence, the effect sentence, the process sentence, the transfer sentence, the relation sentence, the state sentence, and the judgment sentence, which describes one of the seven parts of the generalized action-effect chain.

The SC mixed by basic SC and describing two or more parts of the chain of the generalized action-effect is called hybrid SC, and the SC mixed by basic SC or hybrid SC is called compound SC in the HNC. The sentence of basic SC or mixed SC has at most one eigen semantic-chunk (that is similar to the role of the predicate of a sentence), and the sentence of compound SC has two or even more eigen chunks.

HNC theory proposed a complete theoretical framework for the engineering realization, including the sentence processing, the sentence-group processing, and the chapter processing. According to the viewpoint of HNC, language translation is a mapping from one natural language space to another, including the analysis and understanding of source language, the generation process of target language, as well as the intermediate process between the analysis and generation. The analysis of source language in HNC analyzes the constitution of the semantic chunks of sentence using the techniques of SC analysis.

The generation of the target language relies on the mapping word knowledge-base from HNC concept symbols to the target language words, and the rule knowledge-base from words to semantic chunk and from semantic chunk to sentence. The intermediate processing includes six sub-processes, such as SC conversion and sentence format conversion, primary and auxiliary transformation of semantic chunk and constitution transformation of semantic chunk, and the reorder of auxiliary chunk and clause, which is a necessary condition for producing high-quality translation.

The machine translation based on HNC theory uses the techniques of SC analysis to activate, to extend, to enrich, to convert, and to store the associative skeleton of concepts in the space of language concept, and completes the mapping from source language to the space of language concept to achieve the understanding of the source language. According to the analysis result of source-language SC, the MT can determine the semantic-chunk type, semantic-chunk number, and semantic-chunk order of target language, and analyzes the internal composition of semantic-chunk according to the expected knowledge and constitution information of semantic-chunk to be translated.

The sentence handled by HNC-MMT is terminated according to Chinese punctuation, such as full stop, question label, and exclamatory. Since there are different situations of single sentence and complex sentence in the unit of sentence, HNC-MMT model is different from MMT of CIP on the content of label and its value owing to HNC-MMT is semantic analysis oriented. Therefore, it adopts more semantic-labels, such as conceptual categories and sentence categories for the sentence structure, which serving for automatic knowledge reasoning, named concept associative skeleton.

A. Label Content of HNC-MMT

There are also two kinds of label in HNC-MMT like in MMT. There are about 261 static labels of HNC, distributed in six categories, which are listed as follows.

- **Class of generalized concept**: dynamic concept, concept of abstract noun, concept of specific objects, people, property, and logical concept.
• Conceptual categories: specific concept (associated with persons or things), abstract concept, and ambident concept.

• Lv attribute: flag of main chunk, flag of auxiliary chunk, the back flag of auxiliary chunk, connection specifier within chunk, coreference specifier, and logic specifier within sentence.

• Morpheme: adjective prefix, person prefix, verb prefix, noun suffix, verb suffix, people suffix, and substance suffix.

• Pure dynamic-representation verb.

• Sentence categories: generalized role sentence, the number of main chunk, chunk-extension sentence, prototype sentence-degradation, passive voice, lead of sentence-category conversion, concise state sentence, the first main chunk, sentence category applicable to language-logic-0 (l0), symbol of HNC, effect category, basic concept category, composite structure of eigen chunk (EK), SC code, type of concept category, and word form of translation.

There are about 487 dynamic labels of HNC, distributed in four categories.

• Features of word form: Chinese character, string, starting Chinese characters of string, end Chinese characters of string, English string.

• Chunk features: concept category of chunk, semantic interpretation of chunk, sentence category of chunk, attribute of sentence category, weighted value of EK, semantic expression, result of semantic analysis, the number of child node, level symbol, semantic relation between sentences, share of clause, grade of adjective, tense, style, voice, and deformation of verb.

• Position Features Used by HNC: Such as relative position, parent node, child nodes, sibling nodes, a comma at the beginning, a period at the end, relative position, parent node, child nodes, sibling nodes, a comma at the beginning, a period at the end, relative position, parent node, child nodes, sibling nodes, a comma at the beginning, a period at the end, relative position, parent node, child nodes, sibling nodes.

• Functions Used by HNC: Such as function of node Generation, assignment function, function of value transfer, selection function, condition function, equivalent judgment function, consistent judgment function of high-level semantic, function of node deletion, function of eigen-value deletion, function of adding nodes, function of replacing node, and function of moving node.

B. Implementation

Directed by HNC theory, the linguist constructed word knowledge-base and rule knowledge-base to express the static word knowledge and dynamic rule knowledge. We used a production rule to express our HNC rules; that is, to express rules using HNC functions and static labels attained by the analysis of rules into the HNC-MMT, wherein the new static labels are not given by word knowledge-base initially.

Rules of semantic analysis of HNC-MMT are divided into two categories, which are listed as:

• Analysis rules: verb-object (vo) and object-verb (ov) processing, auxiliary chunk generation, recognition of global eigen, generation and weighting of upside load and down load, rule of compound constitutes of eigen, rule of eigen excluding, rule of eigen queuing, rule of eigen generation, rule of format conversion, rule of clause analysis, rule of clause segmentation, and generation of sentence structure.

• Transformation rules: conversion of compound constitute of eigen, format conversion, basic format conversion, clause transformation, adjustment of chunk sequence, long distance collocation of chunks, apposition of chunks, and processing of repeat structure.

The flow of processing is always to judge whether the word pointed currently is matched to the content of position 0 in the rule. When matched, the processing matches the left nodes of the position 0 one by one, if succeed, it matches the right nodes of the position 0 one by one. We use nine representative rules to describe the characteristics of the rule knowledge-base as follows.

\[
(0)CHN[Ke]+(1)LC_CC[l01, l02, l1]==> \text{LC_TREE(QE,0,0)/$}.
\]

Where \((0)CHN[Ke]\) represents the word 0 pointed by processing currently is a special Chinese character \(ke\); \((1)LC_CC[l01, l02, l1]\) expresses the word 1 after the word 0 has a special semantic category, such as \(l01, l02,\) and \(l1\); \(LC_TREE(QE, 0, 0)\) indicates that the processing will add a label of chunk before the eigen chunk \((QE)\) into the position 0, which is the position of word \(ke\) in this case. For example, this rule will be activated by the following three sentences:

1) Chuliqi ke genju celiang jieguo lai xuanze.
2) Beiyoung dianci ke jiang dianyuan wending di tigong gei bianxieshi zhongduan.
3) QiZhong renhe huan ke bei renyi qudai.

When the position pointer points the Chinese word \(Ke\), after it is a Chinese word "genju" ("jiang" in 2) and "bei" in 3)) which has a word concept-category \(I (l02\) in 2 and \(l01\) in 3)), \((1)\) is activated by the match between the three sentences and the left of it, and the position of Chinese word "Ke" will be added a \(QE\) label.

\[
(-1)LC_CC[uv, uu]+(0)LC_CC[v]==> \text{LC_TREE(EU,-1,-1)/$}.
\]

Where \((0)LC_CC[v]\) represents the concept-category of word 0 pointed by processing currently is a dynamic-representation concept \(v\); \((-1)LC_CC[uv, uu]\) expresses the word -1 before the word 0 is an adverb \((uv)\) or pure adverb \((uu)\) concept; \(LC_TREE(EU, -1, -1)\) indicates that the processing will add a modification of E chunk \((EU)\).
into the position -I, which is the position of word -I in this case. For example, this rule will be activated by the following sentence:
4) Fenbie shezhi cheng sudong he lengcang liangzhong zhenghang fanshishi.

Where "fenbie" is the word belonging to uv concept-category, and "shezhi" is the word belonging to v concept-category. Equation (2) is activated by the match between the sentence 4) and the left of (2), and the position of word "fenbie" will be added an EU label.

(0)LC_CC[\nu]+(1)LC_QH[hSG]\Rightarrow LC_TREE
(E,0,0)& PUT(fp, LC_E_SCORE,VOOV)$.

Where (1)LC_QH[hSG] expresses the word before the word 0 is a suffix of noun (hSG); E is an E chunk; PUT (fp, LC_E_SCORE, VOOV) represents a father-parent node (fp) will be assigned an E score (LC_E_SCORE) with verb-object (vo) and subject-predicate (ov) appeared sequentially (VOOV). For example, this rule will be activated by the following sentence:
5) Zai zhaoshe dian shi gai gongjian de bei zhaoshe qu ronghua.

Where "zhaohe" is the word belonging to v concept-category, and "dian" is the word belonging to hSG concept-category, (3) is activated by the match between the sentence 5) and the left of (3), and the position of word "zhaohe" will be added an E-score VOOV, showing the verb "zhaohe" is dissimilated into a noun.

(0)CHN[bei]+(1)(1)LC_CC[u] & END%\Rightarrow LC_TREE(E,1,1)$.

Where (1)LC_CC[u] & END% expresses the processing searches for the concept-category adjective (u) from the current word rightward, and the word searched must be an end of sub-sentence (END%). For example, this rule will be activated by the following two sentences:
6) Juhewu de hanliang bi toushe quyu nei di.
7) Lingyi ceng de zheshelv bi jidi di.

Where "di" is the word belonging to u concept-category, and the position of "di" is END%, (4) is activated, and the position of the word with u concept-category will be added an E label.

(0)CHN[bei] & BEGIN%\Rightarrow LC_TREE(QE,0,0)& PUT(fp, VOI, P)$.

Where (0)CHN[bei] & BEGIN% expresses the word 0 is a Chinese character "bei", which does not locate at the start of sub-sentence; LC_TREE (QE, 0, 0)& PUT (fp, VOI, P) represents QE label will be added to the position of "bei", and the fp of it will have a passive voice (VOI, P). For example, this rule will be activated by the following sentence:
8) Zhege qduan bei xiugai.

Where "xiugai" has a concept-category v and word "bei" is within the sub-sentence.

(0)CHN[yongyu]+(1)LC_CC[l01,l02,l11] \Rightarrow LC_TREE(EQ,0,0)& PUT(fp, LC_EXP,CENTER_EQ)+
(1,LC_E_SCORE,V_COMP)$.

Where PUT (fp, LC_EXP, CENTER_EQ) expresses the fp will have a semantic label (LC_EXP) with a value center of EQ (CENTER_EQ); PUT(1, LC_E_SCORE, V_COMP) represents the word f will have a score; that is, complex structure of E chunk (V_COMP). For example, this rule will be activated by the following sentence:
9) Tiaoziqiu yongyu dui yinshua de tuxiang shixian hang zhihe.

"Dui" is a Chinese character with concept-category l02 in the above sentence.

(0)CHN[Yizhong]+(1)(1)END%&
CHN[fangfa, zhuangzhi, shebei]$\Rightarrow LC_TREE(BK,0,1)$&
PUT(fp, LC_EXP, AN+N)$.

Where (fp, LC_EXP, AN+N) represents the processing finds a Chinese word "fangfa" or "zhuangzhi" or "shebei" at the end of sub-sentence, and appoints the position of the word with 1; LC_TREE (BK, 0, 1)& PUT(fp, LC_EXP, AN+N) expresses the processing will add a block or chunk (BK) label into the position of "Yizhong" and the position of "fangfa" (or "zhuangzhi", or "shebei"), and put fp a noun-phrase (AN+N) label. For example, this rule will be activated by the following sentence:
10) Yizhong jiguangshu jiaong fangfa.

(0)LC_CHK[L1]&CHN[yi]$\Rightarrow (f){(1)LC CHK[L1]+CHN[fangfa]}\Rightarrow
LC_TREE(ABK,0,1)$.

Where (0)LC_CHK[L1]&CHN[yi]$ expresses the word 0 is "yi" having a chunk label L1; ABK represents an auxiliary chunk label. For example, this rule will be activated by the following sentence:
11) Jiguangshu yi bi suoshu yamno de guang tousheqi da de fanwei zhaoshe dao suoshu yamno shang.

"Yi" has a chunk label L1 in the above sentence.

(0)LC_CHK[L1]&CHN[Genju]$\Rightarrow (f){(1)CHN[fangfa]}\Rightarrow
LC_TREE(ABK,0,1)$.

Where (0)LC_CHK[L1]&CHN[Genju]$ expresses the word 0 is "genju" having a chunk label L1. For example, this rule will be activated by the following sentence:
12) Genju qianshu renyi quanli yaoqiu suoshu de fangfa.

The total processing flow in HNC-MMT for semantic analysis of Chinese is illustrated as follows.
In the above Figure 1, after pre-processing, the Chinese sentence is sent to the module of semantic analysis and conversion. Finally, there is a post-processing, and the results of analysis and translation are given, wherein the pre-processing is illustrated as follows.

![Figure 2. Pre-processing of semantic analysis of HNC-MMT.](image)

In the above Figure 2, the module of pre-processing loads the word lexicon and word knowledge-base, and then segments the Chinese sentence using the forward maximum segmentation to acquire the sequence of HNC words, which includes word form and knowledge-feature of words, that is, static label of HNC. The semantic analysis and conversion for Chinese in the Figure 1 are illustrated in Figure 3.

![Figure 3. Semantic analysis and conversion of HNC-MMT for Chinese.](image)

In the Figure 3, the module of semantic analysis and conversion loads the semantic rule-base, and executes the rule analysis on the Chinese sentence to generate the HNC-MMT. Finally, a rule conversion for Chinese is performed, which realizes a reorder on the HNC-MMT.

After that, a post-processing is performed, illustrated as follows.

In the Figure 4, the module of post-processing executes the English word selection to generate E chunk of English corresponding to the E chunk of Chinese, and performs the transformation of English verb to generate the final English sentence.

![Figure 4. Semantic analysis of post-processing of HNC-MMT.](image)

V. EXPERIMENTS

The experimental corpus was 10 bilingual documents of Chinese-English from China Patent Information Center (CPIC). We extracted 1000 Chinese sentences from these documents sequentially as the data set of experiment. The main content of the experiment was the recognition of the global E chunk of Chinese sentence since global E chunk plays an important role in the semantic analysis of Chinese, and we took the accuracy as our measure for the evaluation of experiment. The accuracy is defined as follows:

\[
\text{Accuracy} = \frac{\# \text{recognized correctly}}{\# \text{needed to be recognized}}
\]  

The experimental result of open test showed the accuracy of global E chunk was 85.1%, and the throughput of the processing was about 455 Chinese characters per second.

VI. CONCLUSION

In this paper, we proposed a semantic analysis model of Chinese multiple-branched and multiple-labeled tree (MMT) based on HNC in Chinese-English machine translation, achieved a HNC-MMT of semantic analysis for Chinese using static labels and dynamic labels of HNC, utilized word knowledge-base and rule-base. The experimental result showed the accuracy of E chunk of Chinese sentence was above 85%. The results generated by the HNC-MMT can be used in the generation of translation, and hybrid methods of MT can be used based on the results. For example, we can send the Chinese chunk into a statistical MT engine, and then integrate the output of the engine, and execute a special processing on the predicate verb to generate the translation. Therefore, the method of multiple strategies hybrid using HNC-MMT will be our future works.

ACKNOWLEDGMENT

The paper was supported by "the National High Technology Research and Development Program of
China (No. 2012AA011104)” and "the Fundamental Research Funds for the Center Universities.”

REFERENCES