2015 International Conference on Computer Science and Applications

Chinese Natural Language Processing based on Semantic Structure Tree

Qi-jin Yin, Shao-ping Wang, Yi-nan Miao, Xin Dou
School of Automation Science and Electrical Engineering
Beihang University, BUAA
Beijing, China
e-mail: yinqijin@buaa.edu.cn, shaopingwang@vip.sina.com, miaoyinanjack@sina.com, zbzddx@163.com

***Abstract*—For the problem of limited rule bases and inaccurate matching of Chinese Natural Language Processing (NLP), this paper presents a new NLP method based on Semantic Structure Tree (SST). Through establishing SST, this paper calculates the evaluation index to find out the most suitable semantic combination from all possible SST. In order to improve the semantic recognition recall and precision, this paper carries out the semantic recognition and the word segmentation synchronously. Application indicates that the proposed method can guarantee high recall and precision in Chinese natural language semantic recognition.**

Keywords-semantic structure tree; semantic recognition algorithm; natural language processing; word segmentation

I. INTRODUCTION

Artificial intelligence [1] is a widely used technique in robots [2], machine learning [3], image recognition [4] and natural language processing [5] .etc. The concept of artificial intelligence was come up in 1950s and since then, many researchers have been working on these fields [5]. Among them, the semantic processing of NLP is the most difficult area, which not only needs to understand the meaning of NLP but also needs to distinguish the language into words and phoneme. So far, the research method of NLP could be divided by three types: semantic recognition based on rules, semantic recognition based on statistics of language models and semantic recognition based on rules and statistics [5]. Semantic recognition based on rules attempts to figure out the meaning of a sentence by analyzing the semanteme and build up rule base, while semantic recognition based on statistics of language models attempts to find the best expression by analyzing contextual links [5]. Semantic recognition based on rules and statistics combines the advantages of both methods. By integrating the existing linguistic research results and statistical models with consistency and high coverage percentages, more satisfying results can be gotten [6].

In the research of semantic recognition based on rules, Elisa Margareth Sibarani produced a simulation of Indonesian parser and realized the semantic recognition successfully. However, the above method is not suitable to Chinese because the Chinese sentences are totally different in corpus and events file in Bahasa Indonesia [7]. In semantic recognition based on statistics of language models, Peter F. Brown discussed several statistical algorithms for assigning words to classes based on the frequency of one word's co-occurrence with other words [8]. But this method cannot demonstrate the full value of the secrets [8]. In semantic recognition based on rules and statistics, Weifeng Cao found the overlapping ambiguous segmentation position. But his method is dependent on experience and cannot recognize unfamiliar words [9].

Chinese has its own characteristics, such as high frequency of meaningless words, lexical ambiguity, and dependency of word segmentation, which are different from English. As a result, the NLP of Chinese is more difficult than English [10]. In Chinese NLP, it is required to separate words. But literatures in this field are limit. Weifeng Cao obtained the segmentation through combining dictionary and statistics [9]. Muyu Zhang used the rules of association words and vocabulary to construct the basic models and figure out the semanteme of an article [11]. The methods aforementioned are lack of the logic connection of semanteme and word segmentation, so their performance on precision and recall is needed to be improved. This paper presents the method based on Semantic Structure Tree (SST) to improve the performance of NLP by considering semanteme and word segmentation synchronously.

The structure of this paper is organized as follows. Section II presents a new semantic recognition algorithm based on SST. Section III describes how to realize SST algorithm and obtain the good performance of SST in a real application. Section IV gives the conclusion.

II. SEMANTIC RECOGNITION ALGORITHM BASED ON STRUCTURE TREE

The difficulty of Chinese semantic recognition is to match the logic relation of morpheme and semantic in a sentence. Due to the various statement of a sentence, it is difficult to get absolutely correct semantic expression with intelligent robot. In order to improve the matching precision of semantic recognition, this paper presents the semantic recognition algorithm based on structure tree.

A. Basic Definition of SST

SST is a kind of tree, in which the leaves express the basic morpheme and the nodes indicate the semantic of a sentence. SST can reflect the logic and semantics relation inside a sentence directly.

Definition 1: Node of SST

$Node = (N, C)$ is defined as the nodal point of SST. There are two kinds of nodes: leaf node and non-leaf node. When $Node$ is not a leaf node, $N = \{nature_1, nature_2, \ldots\}$ represents the semantic attribute set of the node. $C = \{Node_1, Node_2, \ldots\}$ is an attribute ordered set and is the set of the node's child nodes; When $Node$ is a leaf node, $N = \{morpheme\}, C = \varnothing$. In this article, $Node.N$ stands for the attribute set N in Node and $Node.C$ stands for the

978-1-4799-9961-3/15 $31.00 © 2015 IEEE
DOI 10.1109/CSA.2015.65

attribute ordered set C in Node.

Definition 2：Semantic structure tree

Assume that $Tree = (Node_{root}, F)$ is defined as SST, in which $F = (T_0, T_1, T_2, T_3, \ldots)$ indicates the logic relation following the node root. $T_i = (Node_i, F_i), (i = 0,1,2...)$ expresses the logic relation of the child nodes shown in Fig.1. Hence, the relationship between the node root and child nodes can be described as

$RF = \{\langle Node_{root}, Node_i \rangle | i = 1,2,...,m; m > 0\}$.

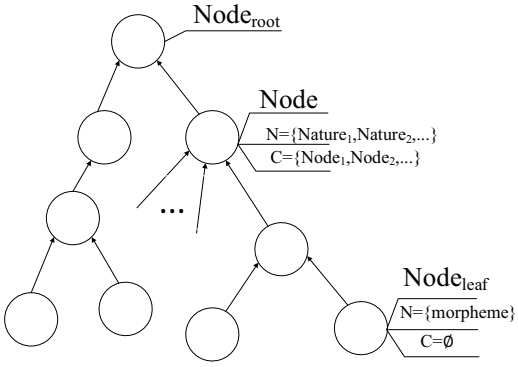

Figure. 1 The Structure of a SST

Example A sentence "明天我要开会" can be described as a SST shown in Fig.2.

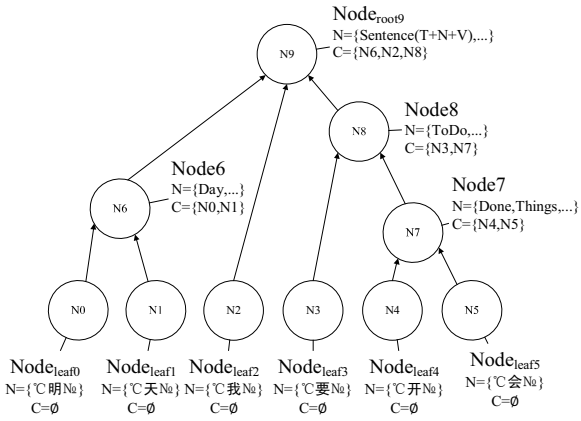

Figure 2. The SST of a sentence "明天我要开会"

B. Construction of the SST

1) Construction of Leaf Nodes

When decomposing a Chinese sentence into basic morphemes(e.g. in character), these morphemes form set $S = \{w_0, w_1, \ldots, w_n\}$ and generate n leaf nodes $Node_{leaf_i}, (i = 0,1,\ldots,n)$. The corresponding relationship can be described as $Node_{Leaf_i} = (N_i, C_i), N_i = \{w_i\}, C_i = \varnothing$.

2) Construction of Non-leaf Nodes

Definition 3: Semantic rule

$Rule_i(NR_i, NN_i)$ is defined as semantic rule, in which $NR_i = \{nature_1, nature_2, \ldots\}$ represents semantic attribute set of current semantic rule and $NN_i = \{nature_1, nature_2, \ldots\}$ indicates the semantic attribute ordered set of its child nodes. In this article, $Rule_i.NR$ stands for the attribute set NR in $Rule_i$ and $Rule_i.NN$ stands for the semantic attribute ordered set NN in $Rule_i$.

Definition 4: NodePool

$NodePool = \{Node_0, Node_1, \ldots\}$ is defined as an attribute set of all nodes.

Definition 5: Judge condition

$fitNN(Node_x, Node_y)$ is defined as a judge condition whether $Node_y$ satisfies a child of $Node_x$ or not. Its expression can be described as:

$\exists Rule_i, st.Rule_i.N = Node_x.N, Rule.NN \bigcap Node_y.N \neq \varnothing$

Definition 6: Node connection

$FitNN(Node_x, Node_y)$ expresses that $Node_x$ can connect $Node_y$ to generate the new node $Node_{x-y}$ according to $Rules$.

$$FitNN(Node_x, Node_y).N = Node_x.N,$$
$$Node_y \subset FitNN(Node_x, Node_y).C$$

Definition 7: New node generation judgment

$fitNS(Node_x, Rules)$ is an expression used to estimate whether a new node can be created by $Node_x$ according to $Rules$. Its representation is as follows:

$$\exists Rule, st.Rule.NN \bigcap Node_x.N \neq \varnothing.$$

Definition 8: New node generation based on rule

$FitNS(Node, Rules)$ expresses that $Node$ can generate the new node $Node_{New}$ according to $Rules$. Its representation is as follows:

$$FitNS(Node, Rules).N = Rules.NR,$$
$$Node \subset FitNS(Node, Rules).C$$

The non-leaf node can be obtained through iteration in the $NodePool$ according to $Rules$. Its iteration process can be described as:

$do\{$

$$NodePool \leftarrow NodePool \bigcup \{FitNN(Node_i, Node_j) | Node_i, Node_j \subset NodePool, fitNN(Node_i, Node_j) = true\} \quad (1)$$

$$NodePool \leftarrow NodePool \bigcup \{FitNS(Nodes_i, Rules) | Node_i \subset NodePool, fitNS(Nodes_i, Rules) = true\} \quad (2)$$

$\}While(new\ node\ is\ created)$

Above iteration can generate multiple topology semantic structure.

Equation (1) indicates that the node to join into the $NodePool$ with possible $Node_i$ and $Node_j$.

Equation (2) shows the possible new nodes $Node_{New}$ into $NodePool$ according to the $Rules$.

When the *NodePool* expands to the limitation with the iteration aforementioned, the iteration stop. At that time, the nodes in $NodePool$ are the possible root nodes, nodes

in various level and the leaf nodes. The process of SST generation can be described in Fig.3.

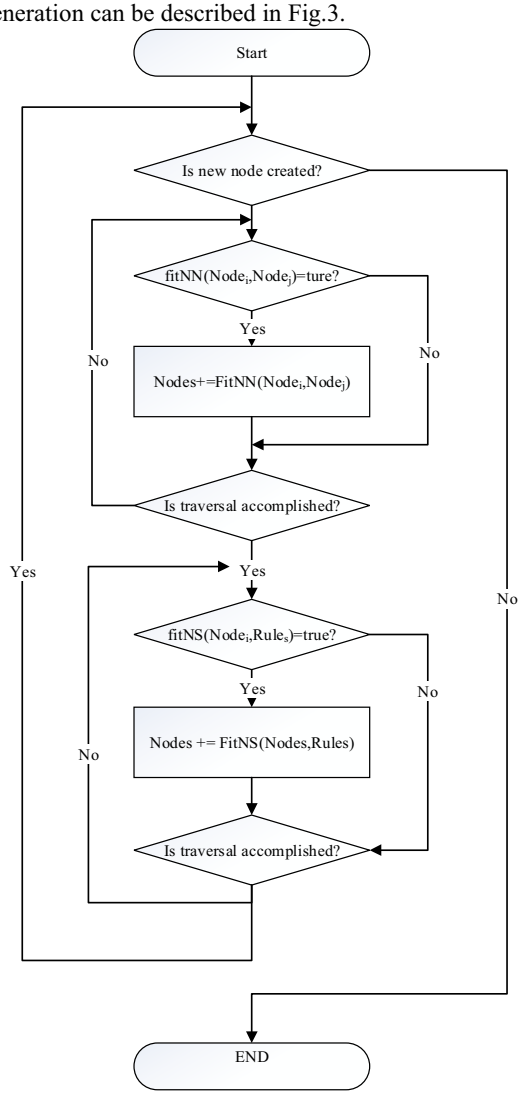

Figure 3. The process of SST generation

C. Evaluation of SST

In order to screen the optimal SST, it is necessary to evaluate its semantic recognition result. Suppose *Node.N* contains n nodes in SST. α_{Node} indicates its evaluation index. When $n=0$, the *Node* is a leaf node and $\alpha_{Node}=c$ (c is a constant). When $n>0$ and the evaluation index of child *Node* are $\alpha_1,\alpha_2,...,\alpha_n$, the evaluation index of node can be described as

$$\alpha_{Node}=\delta\sum_{i=0}^{n}\alpha_i,(0<\delta<1) \quad (3)$$

The evaluation index of a SST can be described by the root nodes. The higher the evaluation index, the better the SST.

III. Application And Discussion

A. SST Algorithm

SST algorithm is composed of two parts, SST construction algorithm and rule base. SST construction algorithm is in Section II. The rule base can be added continuously by inputting by hand, in order to apply to various fields.

The rule base is made of the following rules:

(1) Transform morpheme into semantic terms

(2) Aggregate semantic terms to meaningful expressions. This kind of rule combines nodes according to their nature to a more completeness semantic set.

(3) Combine one or many kinds of nodes according to their nature into a complete sentence according to natural language sequence.

In the program design process of SST algorithm, theoretical analysis and engineering application are both considered. The SST algorithm has the following characteristics:

- Arborescent semantic logic
  SST algorithm can generate a tree with the best semanteme, in which the nodes indicate multilayer semanteme. So the tree reflects the layer of the sentence meaning intuitively and can be extended easily. More information can be included based on the existing tree.
- Removal of the meaningless word
  In the SST algorithm, meaningless morpheme will be filtered out automatically. Because leaf node produced by invalid input can't combine with other node to create a new semantic node and can't connect to the SST. As a result, SST algorithm can apply to Chinese with the characteristic of many meaningless words, and improve the precision of the system.
- Self-adaptation of unfamiliar words
  Morphemes that haven't been included by the rule base, won't be connected to the tree during the construction of the tree. They will be backfilled in the best semantic tree when recalling the leaf nodes. Such characteristic makes words that haven't been included in the dictionary fit well in the SST algorithm.

B. Experiment Statistics and Evaluation Method

To prove the algorithm's feasibility and high efficiency, a rule base is constructed according to Chinese syntax rule, including rules of "making phone calls", "sending messages", "setting alarms" and "question answering". Taking the above four tasks as the research objects, the usability of SST algorithm can be tested and analyzed.

The corpus of the experiment is from Google's open source base. By keyword searching and manual tagging, 2000 task sentences for each object are chosen, totally 8000 sentences. Half of them are used to extract rules, and another half are used to test the effect. That's the corpus base of our experiment.

To judge the performance of SST algorithm, two measurements are used: recall and precision. Here, we use positive example and negative example to analyze. The positive example refers to corpus whose semantic intend matches the testing task. The negative example refers to corpus whose semantic intend does not match

the testing task.

So define recall as

$$\mathrm{Re}\,call = \frac{TruePositive}{TruePositive + FalseNegative} \quad (4)$$

In (4), TruePositive indicates the number of target morphemes that have been identified as target morphemes correctly. FalseNegative indicates the number of target morphemes that have been identified as nontarget morphemes incorrectly.

$$\mathrm{Pr}\,ecision = \frac{TruePositive}{TruePositive + FalsePositive} \quad (5)$$

In (5), FalsePositive indicates the number of nontarget morphemes that have been identified as target morphemes incorrectly.

$$F1_{Measure} = \frac{2 \times \mathrm{Pr}\,esiction \times \mathrm{Re}\,call}{\mathrm{Pr}\,esicion + \mathrm{Re}\,call} \quad (6)$$

In (6) $F1_{Measure}$ is a measurement to measure recall and precision, which shows the performance of SST algorithm synthetically [6].

C. Experiment and Analysis

Extract 1000 positive examples and 1000 negative examples from the corpus base, and test each task separately. The testing results are shown in Table 1.

Table 1. Statistics and Evaluation

$\delta = 0.97$	RECALL	PRECISION	F1-MEASURE
MAKING PHONE CALLS	98.3%	95.3%	96.78%
SENDING MESSAGES	91.43%	90.4%	90.91%
SETTING ALARMS	94.29%	93.75 %	94.02%
QUESTION ANSWERING	98.2%	92.11 %	95.06%

The results can be shown more intuitively in Fig.4.

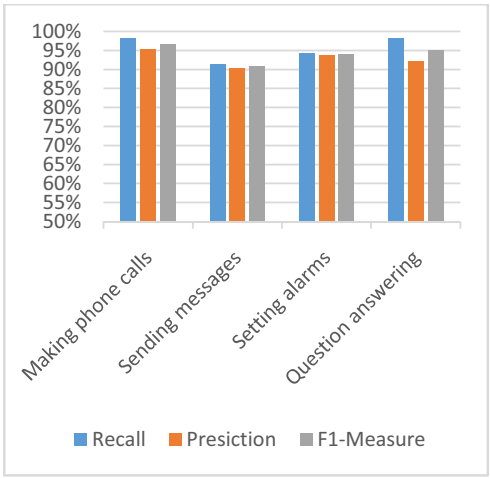

Figure 4. Bar chart under four tasks in application

In the above task experiments, the recall and precision of SST algorithm can both reach to a high level. It results from the filtration of meaningless words. In different tasks, the recall and precision are different. For the making phone call task and question answering task, recall and precision percentage are relatively high. It's because that corpus is relatively simple and the sentence structure is simple in these two tasks. In addition, the recall percentage is a bit higher than precision, indicating that SST algorithm can identify positive examples more accurately and can suit the tasks better.

IV. CONCLUSIONS

This paper presents a novel algorithm for Chinese NLP based on SST. In this SST algorithm, semantic recognition and word segmentation can be carried out at the same time. Comparing with the traditional algorithm which separate semantic processing and word segmentation, SST algorithm can guarantee the completeness of semantics by constructing all the semantic structure trees and evaluate all the value of the nodes to choose the best SST. SST can get very good solution for Chinese sentences with high frequency of meaningless words, lexical ambiguity, and dependency of word segmentation. Application indicates that SST can improve the recall and precision percentage effectively.

REFERENCES

[1] An He, Kyung Kyoon Bae, Timothy R.Newman, Joseph Gaeddert, Kyouwoong Kim and Rekha, et al.. A survey of artificial intelligence for cognitive radios. In proceedings of IEEE Transactions on Vehicular Technology, vol. 59, No. 4, pp. 1578~1592.

[2] Kortenkamp.David, R. Peter Bonasso and Robin Murphy, Artificial intelligence and mobile robots: case studies of successful robot systems, MIT Press, 1998.

[3] Christopher Bishop, Pattern Recognition and Machine Learning, Springer Press, 2011, pp. 1~126.

[4] Md. Iqbal Quraishi, J Pal Choudhury and Mallika De, Image Recognition and Processing Using Artificial Neural Network. In proceeding of Recent Advances in Information Technology, 2012.

[5] Erik Cambria and Bebo White, Jumping NLP Curves: A Review of Natural Language Processing Research. In proceedings of Research Review Article, IEEE Computational intelligence magazine, 2014, Vol. 9, pp.48~57.

[6] Bing Guo, Intelligent Speech dialog system for the semantic recognition based on rules and statistics, Master Dissertation, Beijing University of Posts and Telecommunications, 2011. (In Chinese)

[7] Elisa Margareth Sibarani, Mhd. Nadial, Evy Panggabean and Meryana S, A Study of Parsing Process on Natural Language. In processing in Bahasa Indonesiam. In proceedings of International Conference on Computational Science and Engineering, 2013, pp.309~316.

[8] Peter F.Brown, Vincent J.Della Pietra, Peter V.deSouza, Jenifer C.Lai and Robert L.Mercer, Class-Based n-gram Models of Natural Language. In proceedings of Association for Computational Linguistics. 1992, Vol. 18, No. 4, pp.467~479.

[9] Weifeng Cao, research of key technology in Chinese word segmentation, Master Dissertation, Nanjing University of Science and Technology, 2009. (In Chinese)

[10] Hailiang Yin, Study On The Quasi-affix Of Modern Chinese, Doctoral Dissertation, Shandong University, 2004. (In Chinese)

[11] Muyu Zhang, Yuan Song, Bing Qin and Ting Liu, Chinese Discourse Relation Recognition. In proceeding of Journal of Chinese Information Processing, Vol. 27, No. 6, pp.51~57. (In Chinese)

[12] Hasi and Enbo Tang, The Frame Design of Mongolian Noun Semantic Network. In proceedings of International Conference on

Asian Language Processing, 2013, pp.201~204.

[13] Er-Qing Xu, Fusion approach of formal semantics for the parsing of UDC sentences. In proceedings of the Fourth International Conference on Machine Learning and Cybernetics, 2015, pp.3734~3743.

[14] Chung-Chi Huang, Maxine Eskenazi, Jaime Carbonell, Lun-Wei Ku and Ping-Che Yang, Cross-Lingual Information to the Rescue in Keyword Extraction. In proceedings of 52nd Annual Meeting of the Association for Computational Linguistics, 2014, pp.1~6.

[15] Yingxu Wang, A Semantic Algebra for Cognitive Linguistics and Cognitive Computing. In proceedings of Cognitive Informatics & Cognitive Computing (ICCI*CC13), 2013, pp.17~25.

[16] Mustafa Al Emran and Khaled Shaalan, A Survey of Intelligent Language Tutoring Systems. In Proceedings of International Conference on Advances in Computing, Communications and Informatics (ICACCI), 2014, pp.393~399.

[17] HUANG Chang-ning and ZHAO Hai, Chinese Word Segmentation: A Decade Review. In proceeding of journal of Chinese Information Processing, Vol. 21, No. 3, pp. 8~18. (In Chinese)

[18] LONG Shu-quan, ZHAO Zheng-wen and TANG Hua, Overview on Chinese Segmentation Algorithm. In Proceedings of Computer Knowledge and Technology, Vol. 5, No. 10, pp. 2605~2607. (In Chinese)

[19] CHEN Ya-dong, HONG Yu, YANG Xue-rong, WANG Xiao-bin, YAO Jian-min and ZHU Qiao-ming, Automatic target identification in frame semantic parsing. In Proceedings of Journal of Shandong University, Vol. 50, No. 7, pp.45~65. (In Chinese)

[20] Luciano Del Corro, Rainer Gemulla and Gerhard Weikum, Werdy: Recognition and Disambiguation of Verbs and Verb Phrases with Syntactic and Semantic Pruning. In Proceedings of Conference on Empirical Methods in Natural Language Processing (EMNLP), 2014, pages 374~385.