Predicting Garden Path Sentences Based on Natural Language Understanding System

DU Jia-li

School of Foreign Languages/School of Literature
Ludong University/Communication University of China
Yantai/ Beijing, China

YU Ping-fang

School of Liberal Arts/Institute of Linguistics
Ludong University/Chinese Academy of Social Sciences
Yantai/ Beijing, China

Abstract— **Natural language understanding (NLU) focusing on machine reading comprehension is a branch of natural language processing (NLP). The domain of the developing NLU system covers from sentence decoding to text understanding and the automatic decoding of GP sentence belongs to the domain of NLU system. GP sentence is a special linguistic phenomenon in which processing breakdown and backtracking are two key features. If the syntax-based system can present the special features of GP sentence and decode GP sentence completely and perfectly, NLU system can improve the effectiveness and develop the understanding skill greatly. On the one hand, by means of showing Octav Popescu's model of NLU system, we argue that the emphasis on the integration of syntactic, semantic and cognitive backgrounds in system is necessary. On the other hand, we focus on the programming skill of IF-THEN-ELSE statement used in N-S flowchart and highlight the function of context free grammar (CFG) created to decode GP sentence. On the basis of example-based analysis, we reach the conclusion that syntax-based machine comprehension is technically feasible and semantically acceptable, and that N-S flowchart and CFG can help NLU system present the decoding procedure of GP sentence successfully. In short, syntax-based NLU system can bring a deeper understanding of GP sentence and thus paves the way for further development of syntax-based natural language processing and artificial intelligence.**

Keywords- Natural language understanding; N-S flowchart; computational linguistics; context free grammar; garden path sentences.

I. INTRODUCTION

Natural language understanding (NLU), speech segmentation, text segmentation, part-of-speech tagging, word sense disambiguation, syntactic ambiguity, etc. come under the umbrella term "natural language processing (NLP)".[1] The development of NLU is briefly traced from the early years of machine translation to today's question answering and translation systems. [2-3]NLU today deals with machine reading comprehension in artificial intelligence (AI) [4]and is applied to a diverse set of computer applications.[5-6] Its subject ranges from simple tasks such as short commands issued to robots, to complex endeavors such as the full comprehension of articles or essays. A machine created to understand natural language has been one of the dreams of AI ever since computers were invented, and this encourages the systematic and rapid development of NLU. The efficient understanding of natural language requires that computer program be able to resolve ambiguities at the syntactic level

and recover that part of the meaning of its individual words taken in isolation.[7-8] The satisfaction of this requirement involves complex inference from a large database of world-knowledge, and this makes the designer of computer programs for NLU face the serious difficulty of algorithm processing.[9] The machine comprehension is embedded in the more general frame of interpersonal communication and is applied to the person-machine interaction task.[10] The further integration has proved appropriate for the design of effective and robust natural language interfaces to AI systems. Syntactic garden path phenomenon is a major source of uncertainty in NLU. Syntax-supported systems attempt to help machine to get a deeper understanding of garden path sentence and the related algorithms deserve special attention in the future of NLU developing. [11] Many hybrid researches contribute to the improvement of NLU systems. [12] For example, the fact is established that abduction rather than deduction is generally viewed as a promising way to apply reasoning in NLU. [13]The development of NLU can focus on the design of a stochastic model topology that is optimally adapted in quality and complexity to the task model and the available training data.[14] A comparative information-theoretic study is carried out to show positional letter analyses, n-gram analyses, word analyses, empirical semantic correlations between the Greek and English n-grams, and entropy calculations are useful to text processing and compression, speech synthesis and recognition as well as error detection and correction.[15]

Garden path (GP) phenomenon is a special linguistic phenomenon which comprises processing breakdown and backtracking. GP sentence (e.g. "The old dog the footsteps of the young") is an originally correct sentence which makes readers' grammatical misinterpretation linger until re-decoding has occurred. An incorrect choice in GP sentence usually is readers' most likely interpretation, leading readers initially into an improper parse which, however, finally proves to be a dead end. Thus the processing breaks down and backtrack to the original status to search the given information again for alternative route to the successful decoding. The automatic decoding of GP sentence is a challenge for NLU systems for machine has to have access to grammatical, semantic and cognitive knowledge in order to understand natural language smoothly as human brains do.

In this paper, a syntax-based and algorithm-originated approach to understanding GP sentence in natural language domain is presented. The discussion consists of four sections. Firstly, a logic-based model of NLU system is shown to

provide an overview of NLU. Secondly, N-S flowchart is used to present the special feature of processing breakdown of GP sentence. Thirdly, context-free grammar (CFG) is introduced to analyze the backtracking of GP sentence in detail and the machine's automatic decoding procedure of GP sentence is analyzed. The last section brings the conclusion.

II. A MODEL OF NATURAL LANGUAGE UNDERSTANDING SYSTEM

In the domain of Intelligent Tutoring Systems, the high-precision NLU system is helpful to accurately determine the semantic content of learners' explanations. For example, the NLU system developed in the context of the Geometry Cognitive Tutor combines unification-based syntactic processing with Description Logic based semantics to achieve the necessary accuracy level. [16] The syntactic and semantic processing of natural language is useful for specific decoding problems, like metonymy resolution and reference resolution. On the basis of linguistic theories and computational technologies, NLU system architecture building the syntactic structure and offering the semantic interpretation of learner' explanations is structured by Octav Popescu in 2005. (Fig. 1)
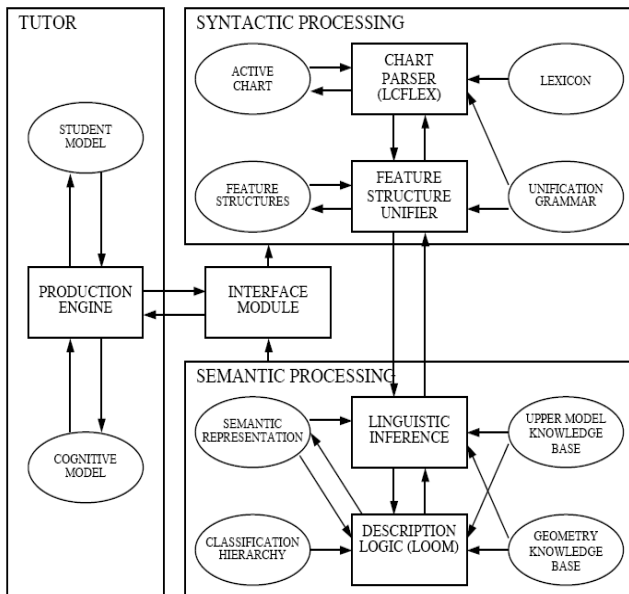

Figure. 1. Octav Popescu's NLU system architecture

The research is related to practical application of NLU system. Popescu's architecture comprises two sections: the syntactic and the semantic processing subsystems, which are interactive during the NLU processing. The syntactic processing subsystem uses an active chart parser while the semantic one bases its action on a Description Logic system. The key part is interface module connecting the NLU subsystem to the tutor itself, functioning asynchronously to the NLU system, taking the input sentence from the tutor, and passing the checked words to the chart parser. System functions in real time until the parser finishes and passes the resulting classifications back to the tutor. The chart parser, adopting linguistic knowledge about the target natural language from the unification grammar and the lexicon, is the center of NLU system. Words of a sentence are taken one by

one by the parser according to grammar rules, and feature structures, which can store lexical, syntactic, and semantic properties of corresponding words and phrases, are built simultaneously.

Popescu's system highlights the integration of syntactic and semantic information during the machine decoding, showing that hybrid knowledge is necessary for machine to understand natural language. GP sentence is a more complex linguistic phenomenon than the common sentence since GP sentence possesses special features of processing breakdown and backtracking. Therefore, the emphasis of integration of linguistics and computational science is helpful for system to automatically decode GP sentence. The computational skill (e.g. N-S flowchart) can present in detail the syntactic processing procedure, making learners have access to the machine understanding of GP sentence and thus enhancing the effectiveness of system.

III. N-S FLOWCHART BASED NATURAL LANGUAGE UNDERSTANDING OF GP SENTENCE

GP sentence seems to be a grammatically correct sentence at the original processing stage and is usually used in syntax, linguistics, psycholinguistics, and computational linguistics. According to syntactic theory, while a person reads a GP sentence, he builds up an original meaning structure and processes natural language one word at a time. With the advancement of processing, the person finds that he has been constructing an incorrect structure and that the next word or phrase cannot be incorporated into the structure originally created thus far. The "garden path" is a reference to the saying "to be led down the garden path", meaning "to be misled". Examples are as follows: "The horse raced past the barn fell", "The man who hunts ducks out on weekends", "The cotton clothing is made of grows in Mississippi", "The prime number few", "Fat people eat accumulates", "The old man the boat", "The tycoon sold the offshore oil tracts for a lot of money wanted to kill JR", etc (Some examples cited are from http://home.tiac.net/~cri/1998/garpath1.html). The verbs (e.g., "raced", "ducks", "grows", "number", "accumulates") in the above examples are not the words which a reader is assuming to be the verbs. Some words that look as though they were modifying nouns are actually being used as nouns and some words which look as though they were nouns are verbs. The revised non-GP sentences are as follows: "The horse that was raced past the barn fell down", "The man, who hunts, ducks out on weekends.", "The cotton, of which clothing is made, grows in Mississippi", "The prime (group) number few", "Fat that people eat accumulates", "The old (people) man the boat", "The tycoon, who was sold the offshore oil tracts for a lot of money, wanted to kill JR", etc. Please see the explanation in Fig. 2.

Fig. 2 shows us that GP sentence which has special features of processing breakdown and backtracking is exclusive of multi-meanings, namely, only one meaning involved in the final result of decoding even though the deviation of the original meaning and the final meaning is obvious. The special features of GP sentence can be presented with the help of computational linguistics skills, e.g. N-S flowchart which was raised by Nassi and Shneiderman [17] to

analyze structured programming [18-19]. This kind of analysis is helpful for the effectiveness of discussion boards. [20]

Sentence	Initial likely partial parse	Final parse	Alternative form of original sentence
The horse raced past the barn fell.	The horse was racing past the barn....	The horse that was raced past the barn fell down.	The horse was raced past the barn and fell down.
The man who hunts ducks out on weekends.	The man who hunts ducks....	The man, who hunts, ducks out on weekends.	The man hunts and ducks out on weekends.
The cotton clothing is made of grows in Mississippi.	The clothing, which is made of cotton, is made of....	The cotton, of which clothing is made, grows in Mississippi.	The cotton that clothing is made of grows in Mississippi.
The prime number few.	The prime number....	The prime (group) number few.	The major group count few.
Fat people eat accumulates.	Fat people eat....	Fat that people eat accumulates.	An oily substance (fat) that people eat accumulates
The old man the boat.	The old man...	The old (people) man the boat.	The old people sail the boat.
The tycoon sold the offshore oil tracts for a lot of money wanted to kill JR.	The tycoon sold the offshore oil tracts for a lot of money....	The tycoon, who was sold the offshore oil tracts for a lot of money, wanted to kill JR.	The tycoon was sold the offshore oil tracts for a lot of money and wanted to kill JR

Figure. 2. Explanation for garden path sentence

Flowchart's IF-THEN-ELSE statement in computational processing can be used to show the procedure of decoding GP sentence now that this technique adopts a dichotomy between Yes and No. For example, Fig.3 consists of four pairs of dichotomies, namely, A and not A, B and not B, C and not C, D and not D. Nassi and Shneiderman introduce seven possible processing results, i.e. (1) not A; (2) not A and not B; (3) not A and B; (4) A and not C; (5) A and not C and not D; (6) A and not C and D; (7) A and C.
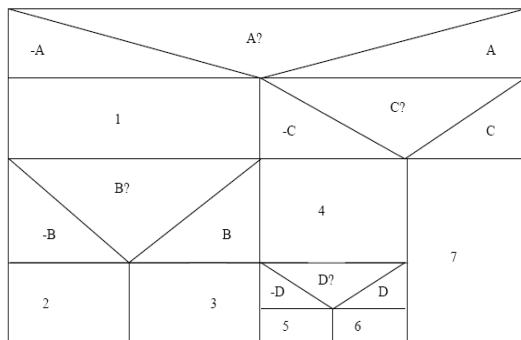

Figure. 3 Nassi and Shneiderman's flowchart of IF-THEN-ELSE statement

The dichotomy of IF-THEN-ELSE statement can be applied to analysis of GP sentence which possesses processing breakdown and backtracking.

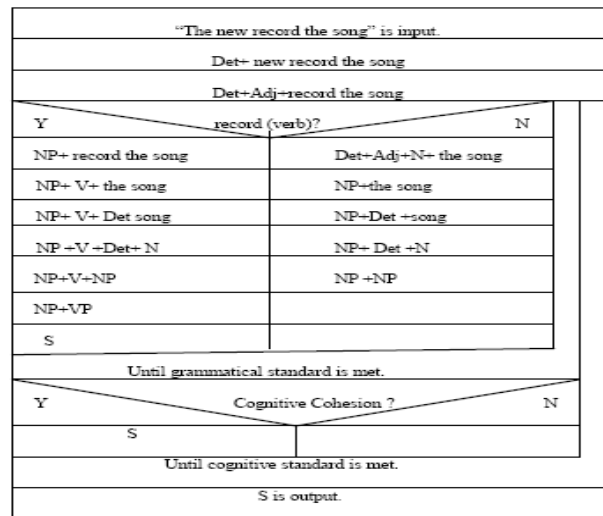Example 1. The new record the song.

Figure. 4 The example of two pairs of dichotomies of N-S flowchart

There are two pairs of dichotomies involved in Fig. 4. The first pair is used to distinguish the verb definition and the noun definition of "record". If the verb definition is chosen, the machine decoding will run along the left column which is the grammatically correct presentation.

Otherwise, the noun definition of "record" is the alternative choice; then the right column is the preferable choice. According to the cognitive rules, "Det+Adj+Noun" is the prototype model in which "Adj" is used to modify the "Noun". Therefore, the right column is the original decoding procedure until system fails to parse more. The decoding has to return to the triangle to choose the verb definition of "record" till every string involved is processed completely.

The second pair of dichotomy appears after the grammatical requirements are met successfully. That is to say, system usually abides by the grammatical rules firstly for the syntactic cohesion is the basic condition for decoding GP sentence.

After satisfying the grammatical requirements, system then considers cognitive cohesion. This is the reason why cognitive rules are put after the grammatical decoding. Example 1 is accepted by system according to cognitive rules and therefore the result is output finally. If the sentence "The dead record the song" which has the similar decoding procedure like Example 1 is submitted to system, the cognitive disagreement makes system reject the decoding and no final result will be output.

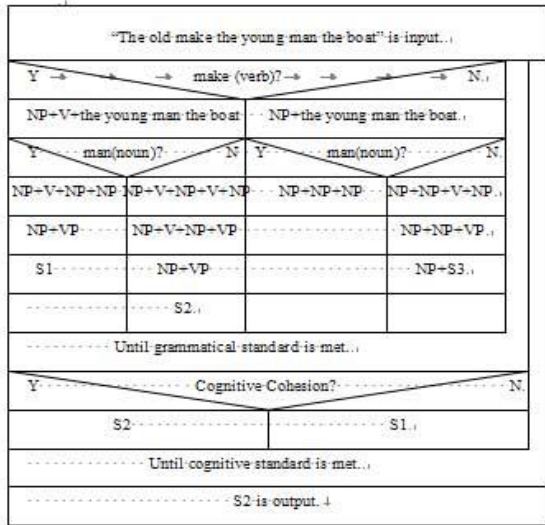Example 2. The old make the young man the boat.

Figure.5 The example of four pairs of dichotomies of N-S flowchart

Fig. 5 consists of four pairs of dichotomies of N-S flowchart. The first pair is used to differentiate the verb definition from the noun definition of "make". If the verb definition is chosen, the left column is system's decoding path. Otherwise, the noun definition in the right column is the choice. The second central triangle in left column is applied to distinguish the verb definition from the noun definition of "man". The result is that the noun one is in left part and the verb one is in right part.

The third pair in right column is similar decoding procedure as the second pair. The difference between the second and the third lies in "make" in the former is a verb while "make" in the latter is a noun. The last pair of dichotomy is used to test the output from the cognitive level. If the result is for the cognitive rules, the left column is activated and it will be output successfully. Otherwise, the right column starts and the failed system has to return to the first pair of dichotomy to parse again.

For example, "S1" is a grammatical correct output of Example 2, and it means that "The old people make the young man be the boat" which is against the cognitive rules. Therefore, "S1" is rejected by system and has to return to the first central triangle until the correct "S2" is chosen. "S2" means "The old people make the young people sail the boat".

From the discussion above, we can find that N-S flowchart is useful to analyze GP sentence since this technique has the obvious function by which the special features of processing breakdown and backtracking can be presented in detail.

Besides N-S flowchart, N. Chomsky's context-free grammar (CFG) is often applied to natural language understanding and the formalized processing is helpful in decoding GP sentence.

IV. CFG-BASED NATURAL LANGUAGE UNDERSTANDING OF GP SENTENCE

A context-free grammar (CFG) is also called a phrase structure grammar in formal language theory. According to the grammar, it can naturally generate a formal language and clauses can be embedded inside clauses arbitrarily deeply. In terms of production rules, every production of a CFG is of the form: "V→w". Generally speaking, "V" is a single nonterminal symbol, and "w" is a string of terminals and/or nonterminals. CFG used to analyze the syntax of natural languages plays an important role in the description and design of programming languages in NLU system.

Noam Chomsky has argued that natural languages are based on CFG and the processing procedure can be presented by formal languages. In Chomsky's opinion, if "A→BC"or "A→a" can be accepted by system in which "A (B.C)" represents the nonterminal symbol and "a" is regarded as terminal symbol, the CFG (usually called Chomsky Normal Form) is a dichotomy possessing the binary tree form. Thus the programming languages are designed to decode GP sentence.

According to formal languages, quaternion parameters are involved in the programming, namely, "Vn" (the collection of nonterminal symbols), "Vt" (the collection of terminal symbols), "S" (the original start), and "P" (rewriting programs). Thus, the natural language can be understood by a machine if the programming language is created. The analysis of GP sentence of Example 2 is as follows in Fig.6.

$G=\{Vn, Vt, S, P\}$

$Vn=\{S, NP, VP, Det, Adj, V, N\}$

$Vt=\{the, old, make, young, man, boat\}$

$S=S$

P:

S → NP VP	(a)
NP → Det Adj	(b)
NP → Det N	(c)
NP → Det Adj N	(d)
VP → V NP NP	(e)
VP → V NP VP	(f)
VP → V NP	(g)
Det → {the}	(h)
Adj → {old, young}	(i)
N → {make, man, boat}	(j)
V → {make, man}	(k)

Processing Procedure (Part I: NP+NP+NP)

1. The old make the young man the boat	
2. Det old make the young man the boat	(h)
3. Det Adj make the young man the boat	(i)
4. Det Adj N the young man the boat	(j)
5. NP the young man the boat	(d)
6. NP Det young man the boat	(h)
7. NP Det Adj man the boat	(i)
8. NP Det Adj N the boat	(j)
9. NP NP the boat	(d)
10. NP NP Det boat	(h)
11. NP NP Det N	(j)
12. NP NP NP	(c)

Processing Procedure (Part II: NP+S3)

13(7). NP Det Adj man the boat	
14. NP NP man the boat	(b)
15. NP NP V the boat	(k)
16. NP NP V Det boat	(h)
17. NP NP V Det N	(j)
18. NP NP V NP	(c)
19. NP NP VP	(g)
20. NP S3	(a)

```
Processing Procedure (Part III: S1)
21(3). Det Adj make the young man the boat
22. NP make the young man the boat          (b)
23. NP V the young man the boat             (k)
24. NP V Det young man the boat             (h)
25. NP V Det Adj man the boat               (i)
26. NP V Det Adj N the boat                 (j)
27. NP V NP the boat                        (d)
28. NP V NP Det boat                        (h)
29. NP V NP Det N                           (j)
30. NP V NP NP                              (c)
31. NP VP                                   (e)
32. S1                                      (a)
Processing Procedure (Part IV: S2)
32(25). NP V Det Adj man the boat
33. NP V NP man the boat                    (b)
34. NP V NP V the boat                      (k)
35. NP V NP V Det boat                      (h)
36. NP V NP V Det N                         (j)
37. NP V NP V NP                            (c)
38. NP V NP VP                              (g)
39. NP VP                                   (f)
40. S2                                      (a)
```

Figure 6 CFG-based understanding of Example 2

The whole processing procedure of Example 2 is shown above and comprises 40 steps for a machine to understand the GP sentence. The flowchart of the decoding consists of four parts. The first part is the production of "NP+NP+NP"; the second, "NP+S3"; the third, "S1"; the fourth, "S2".

In the first part, not all strings are decoded completely according to the rewriting rules and correspondingly system fails to go ahead. The decoding returns to the central triangle where "man" is considered a verb or a noun. The verb definition of "man" is chosen since the noun definition proves to be a dead end. Thus, system comes to the second part of processing procedure in which "NP+S3" is the conclusion. The decoding numbers of first part is "1-2-3-4-5-6-7-8-9-10-11-12-12-11-10-9-8-7".

The second part is the result of backtracking of "NP+NP+NP". "Man" is alternatively regarded as a verb and system reads the sentence of "The old make the young man the boat" into "NP+S3" in which "make" is considered a noun. Since no more rewriting rules support the systematic decoding of "NP+S3", the syntactic output is against the grammatical requirements and as a result, system rejects the final output in which both the noun definition and the verb definition of "man" are chosen as the syntactic choices when "make" is used as a noun. Therefore, system has to backtrack to another central triangle where "make" is considered to be a verb, which brings the appearance of the third part. The decoding numbers of second part is "13(7)-14-15-16-17-18-19-20-20-19-18-17-16-15-14-13-12-11-10-9-8-7-6-5-4-3".

The third part is the result of backtracking of "NP+NP+NP" and "NP+S3" where noun definition of "make" is preferable choice now that, according to cognitive and psychological prototype theory, "Det+Adj+Noun" is the default setting. Since the noun definition of "make" results in a dead end, system becomes to consider "make" a verb. Thus the left central triangle is activated. If "man" is chosen as a noun, S1 in which Example 2 means "The old people make the young man (be) the boat" is the temporary decoding result. There is no cognitive cohesion involved in S1 even though it

is grammatical correct. System rejects S1 as a successful output (see Fig. 5) and backtracks to the central triangle where "man" can be chosen as a verb, which results in the decoding of S2. The decoding numbers of third part is "21(3)-22-23-24-25-26-27-28-29-30-31-32-32-31-30-29-28-27-26-25".

The last part is the result of backtracking of "NP+NP+NP", "NP+S3" and "S1". The only choice for system is that both "make" and "man" should be verbs. The meaning of Example 2 is "The old people make the young people sail the boat". This explanation satisfies the grammatical, cognitive and semantic requirements and is accepted by system perfectly. The automatic understanding of GP sentence in NLU system is completely concluded. The decoding numbers of last part is "32(25)-33-34-35-36-37-38-39-40".

CFG-based natural language understanding is helpful to analyze the special features of GP sentence. Both processing breakdown and backtracking can be presented on the basis of CFG grammar. For example, in the decoding of Example 2, all the figures in parentheses refer to the stage in which processing breaks down and the converse decoding numbers represent the stage in which backtracking occurs. The whole decoding numbers of NLU system is as follows: "1-2-3-4-5-6-7-8-9-10-11-12-12-11-10-9-8-7-13(7)-14-15-16-17-18-19-20-20-19-18-17-16-15-14-13-12-11-10-9-8-7-6-5-4-3-21(3)-22-23-24-25-26-27-28-29-30-31-32-32-31-30-29-28-27-26-25-32(25)-33-34-35-36-37-38-39-40".

V. CONCLUSION

Natural language understanding (NLU) dealing with machine reading comprehension is the umbrella term of "natural language processing (NLP)". The advancement of machine technologies and computational skills develops NLU system. Its subject ranges from sentence decoding to text understanding. The decoding of GP sentence, a special linguistic phenomenon which possesses processing breakdown and backtracking, belongs to the domain of NLU system. GP sentence is a grammatical correct sentence in which original misinterpretation lingers until re-decoding has occurred. An incorrect choice in GP sentence usually is readers' most likely interpretation and lures readers into an default parse which, however, finally proves to be a dead end. If the system can present the special features of GP sentence, namely, processing breakdown and backtracking, the automatic decoding can be successful and the effectiveness of system can be improved. After showing Octav Popescu's model of NLU system, we emphasize the importance of integration of syntactic, semantic and cognitive backgrounds in system, focus on the programming skill of IF-THEN-ELSE statement used in N-S flowchart, and highlight the function of context free grammar (CFG) created to decode GP sentence. On the basis of GP sentences-supported analyses, we come to the conclusions that the machine comprehension can be embedded in the general frame of communication via programming languages, that interaction between the person and machine can improve NLU system, that programming technology (e.g. N-S flowchart) can help NLU system present the decoding procedure of GP sentence, and that syntax-supported linguistic skill (e.g. CFG) can bring a deeper understanding of GP sentence.

REFERENCES

[1] W. L. Wu, R. Z. Lu, J. Y. Duan, H. Liu, F. Gao, and Y. Q. Chen, "Spoken language understanding using weakly supervised learning," Computer Speech & Language, vol. 24, April 2010, pp. 358-382.

[2] Y. Wilks, "A position note on natural language understanding and artificial intelligence," Cognition, vol. 10, July-August 1981, pp. 337-340.

[3] R. C. Schank, "Conceptual dependency: A theory of natural language understanding," Cognitive Psychology, vol. 3, October 1972, pp. 552-631.

[4] C. Mellish, and J. Z. Pan, "Natural language directed inference from ontologies," Artificial Intelligence, vol. 172, June 2008, pp. 1285-1315.

[5] M. Jeong, and G. G. Lee, "Practical use of non-local features for statistical spoken language understanding.," Computer Speech & Language, vol. 22, April 2008, pp. 148-170.

[6] M. Jeong and G. G. Lee, "Machine learning approaches to spoken language understanding for ambient intelligence," Human-Centric Interfaces for Ambient Intelligence, 2010, pp. 185-224.

[7] K. Hird and K. Kirsner, "Objective measurement of fluency in natural language production: A dynamic systems approach," Journal of Neurolinguistics, March 2010.

[8] A. Lockman, and D. Klappholz, "The control of inferencing in natural language understanding," Computers & Mathematics with Applications, vol. 9, 1983, pp. 59-70.

[9] R. López-Cózar, Z. Callejas, and D. Griol, "Using knowledge of misunderstandings to increase the robustness of spoken dialogue systems," Knowledge-Based Systems, vol. 23, July 2010, pp. 471-485.

[10] C. Madden, M. Hoen, and P. F. Dominey, "A cognitive neuroscience perspective on embodied language for human–robot cooperation," Brain and Language, vol. 112, March 2010, pp.180-188.

[11] K. J. Lee, Y. S. Choi, and J. E. Kim, "Building an automated English sentence evaluation system for students learning English as a second language," Computer Speech & Language, May 2010.

[12] D. Kayser, and F. Nouioua, "From the textual description of an accident to its causes," Artificial Intelligence, vol. 173, August 2009, pp. 1154-1193.

[13] J. Bos, "Applying automated deduction to natural language understanding," Journal of Applied Logic, vol. 7, March 2009, pp. 100-112.

[14] W. Minker, "Design considerations for knowledge source representations of a stochastically-based natural language understanding component," Speech Communication, vol. 28, June 1999, pp. 141-154.

[15] E. J. Yannakoudakis, I. Tsomokos, and P.J. Hutton, "N-grams and their implication to natural language understanding," Pattern Recognition, vol. 23, 1990, pp. 509-528.

[16] O. Popescu, Logic-Based Natural Language Understanding in Intelligent Tutoring Systems (PhD Thesis), Carnegie Mellon University, 2005.

[17] I. Nassi and B. Shneiderman, "Flowchart techniques for structured programming," ACM SIGPLAN Notices, vol. 8, 1973, pp. 12-26.

[18] J. L. Du and P F Yu, "A computational linguistic approach to natural language processing with applications to garden path sentences analysis," International Journal of Advanced Computer Science and Applications. Vol. 3, September 2012, pp. 61-75.

[19] J. L. Du, P. F. Yu, "Syntax-directed machine translation of natural language: Effect of garden path phenomenon on sentence structure," International Conference on Intelligent Systems Design and Engineering Applications, 2010, pp. 535－539.

[20] A. A. Badawy, "Students' perceptions of the effectiveness of discussion boards," International Journal of Advanced Computer Science and Applications. Vol. 3, September 2012, pp. 136-144.