Available online www.jocpr.com

Journal of Chemical and Pharmaceutical Research, 2014, 6(7):630-641

Research Article

ISSN : 0975-7384
CODEN(USA) : JCPRC5

Knowledge discovery from Chinese internet public opinion short text

Shengluan Hou*[1,2], Lei Liu[1], Cungen Cao[2] and Shuying Yan[3]

[1]College of Applied Sciences, Beijing University of Technology, Beijing, China
[2]Institute of Computing Technology, Chinese Academy of Science, Beijing, China
[3]Research Development Center, VMware Information Technology(China) Co. Ltd, Beijing, China

---

ABSTRACT

Internet is becoming a spreading platform for the public opinion. Internet public opinion(IPO) discovers hot topics and spreads rapidly. It is vital to analyze the IPO and grasp their trends correctly and timely. A great deal of IPO text is short text which is oral and exits out-of-vocabulary words, such as Weibo text, products comments and so on. In this paper, we propose a novel semantic-based method of knowledge discovery from Chinese IPO short text. This method has two parts: Knowledge Discovery from Chinese IPO short text Language(KDIPOL) and parser of KDIPOL. An extensible context-free grammar(ECFG) is presented to describe the IPO short text. We adopt semantic constraints in ECFG to solve context-sensitive and ambiguous problems of Chinese. Then we design and implement a parser of KDIPOL, which provides a running platform for KDIPOL parsing and outputs frame-based knowledge representation form. Experimental results show the high performance of our method.

Key words: Internet public opinion; semantic processing; Context-Free Grammar; ontology of IPO

INTRODUCTION

Nowadays, Internet has become an important platform for information dissemination and communication. Internet users can release and get current events, policies, products and services, or other information through Weibo, BBS, blog, and many other network applications. Internet public opinion(IPO) refers to the Internet users' emotions, attitudes and perspectives about something such as hot issues, products or services through the Internet. To some extent, it reflects the whole society's attitudes and emotions [1]. However, with the rapid development of Internet and lack of effective supervision and relevant laws, IPO's impact on political and economic order and social stability is more and more serious in recent years. It is important to thoroughly analyze the emerging IPO text accurately and timely to maintain the order of society [2-5]. How to discovery knowledge [6, 7] from IPO text is the basic work of IPO analysis.

According to the China Internet Network Information Center(CNNIC)'s 33th Internet statistical report, by the end of December 2013, China has had 618 million Internet users. Interestingly, 500 million of them are mobile Internet users. Among all the Internet users, the proportion of those using mobile phones to access the Internet rose to 81.0%. Internet users can know what is happening at any moment in time, anywhere in the world. Meanwhile, most of IPO texts are short and generated constantly. Concerning the short length, weak ability to describe the characteristics of short text, traditional keywords-based language processing methods are not suitable. For example: "电信信号真的不错，但服务态度太差(China Telecom's signal is really good, but its servers' attitude is bad) ".To build classifiers for sentiment analysis, we need to collect sentiment words, like positive word"不错(good)" and negative word"差(bad)". It is not an easy task to analyze a sentence carry positive and negative feelings like this if we use traditional keywords-based methods.

We can solve this problem with semantic analysis methods, such as syntactic analysis. Syntactic analysis is one of the most important topics in the fields of NLP. We can get the following parse tree with syntactic analysis method.

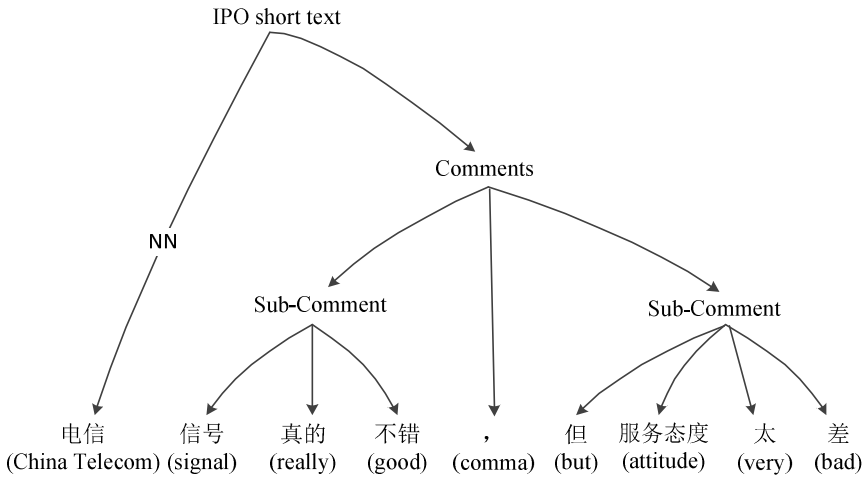

Fig. 1: Parse tree of an IPO short text

In terms of Chomsky's hierarchy of languages [8], Chinese Internet Public Opinion text is context-sensitive. However, context-sensitive grammars used in Chinese Internet Public Opinion text processing are large and complex. The difficult problems are designing, creating, testing, and maintaining them [9]. Furthermore, it is typically not efficient to parsing with a large, wide-coverage grammar due to the long running time and high space requirements.

Any natural language is very ambiguous [10], and it is well known that humans use world knowledge and contextual knowledge to do disambiguation which cannot be solved by computers now. So we propose an extensible context-free grammar(ECFG) to describe Chinese Internet Public Opinion text. ECFG has the semantic constraints to solve the context-sensitive problems. we present a semantic-based Chinese IPO short text knowledge discovery method. This method has two parts:

1) Knowledge Discovery from Chinese Internet Public Opinion Language(KDIPOL): KDIPOL is a particular program language for Chinese IPO short text. KDIPOL consists of several intelligent interacting knowledge discovery agents. Every agent is partially independent, self-aware and autonomous. An extensible context-free grammar(ECFG) is bind to each agent. ECFG works as a mapping from formal grammar symbols to Chinese IPO semantics.

2) High efficiency parser of KDIPOL: Parsing the Chinese IPO short text with ECFG efficiently. The parser can parse the IPO short text into structured and computer-readable form. The running time and space requirements are the key factor of parser. We present a heuristic algorithm to lower the running time and reduce space requirements.

RELATED WORK

Due to the booming and widespread of public opinion on the Internet, it is difficult to analyze IPO text not only accurately but also timely. Many domestic and foreign research institutes and enterprises have been carried out relevant research work. Some theoretical achievements [2-5] and products of IPO text analysis systems such as Buzzlogic and TRS OM have been worked out. Most of these achievements and products are based on keywords statistics which are inaccurate due to lack of necessary semantic processing [6]. Recently, scholars have been studying more effective ways to solve this problem. Content identification methods of semantic-based [7] are the focus of current research. Relevant research includes KAT, ontology of IPO, formal grammar-based NLP.

Knowledge Discovery from Text(KDT) is an important research area in artificial intelligence [11]. KDT can turn unstructured text into computer-readable form using artificial intelligence methods, such as machine learning. KDT methods include concepts learning, concept semantic relations learning, concept attributes learning and some other aspects of knowledge acquisition. There are two main approaches for KDT. One is pattern-based (also called rule-based), and the other is statistics-based. The former approach uses the linguistics and natural language processing techniques (such as lexical and parsing analysis) to obtain patterns, and then makes use of pattern matching algorithm to discover knowledge. The latter approach is based on corpus and statistical language model, and uses text mining algorithm to discovery knowledge.

Wang Haitao proposed a semi-structured text processing method based on knowledge ontology (Ontology-Mediated Knowledge Processing for Semi-structured Text, OMKP) [12]. OMKP is mainly designed to process semi-structured text automatically (but not only) by adopting technologies and theories of multi-agents, pattern matching and automata theory. OMKP is a kind of general-purpose method, which isolates knowledge processing from factual fields but bridging them with the help of ontologies. OMKP is both domain-independent and language-independent.

After over 15 years of research, the method of constructing domain ontology has become a matured method [13]. Ontology of one domain reflects an abstraction between a set of concepts and their relationships. At present, research areas on ontology of domestic and foreign scholars include ontology description languages, ontology construction methods and so on. As for the ontology description language, there have been some artificial intelligence-based ontology languages in the early 1990s, such as KIF, Ontolingua, OKBC and so on. With the development of Web technologies, some Web-based ontology languages, such as RDF, RDFS and OWL, are worked out in turns. OWL is becoming an important knowledge representation language. In the aspects of ontology building tools, so far, there have been many ontology building tools such as OntoSaurus, Protégé, WebODE, OntoEdit and so on. Ontology of IPO can be seen as a domain ontology. It includes all the concepts of IPO and their concept relations which can help computers to understand the IPO text.

Formal grammar-based method is an effective method of sentence structure analysis [14, 15]. It is often used in natural language processing(NLP). Unstructured natural language based on formal grammar can be transformed into a structured parse tree, then generating natural language semantic representation. The Chomsky hierarchy consists of the following levels: Type-0 grammars (unrestricted grammars) include all formal grammars. Type-1 grammars (context-sensitive grammars) generate the context-sensitive languages. Type-2 grammars (context-free grammars) generate the context-free languages. Type-3 grammars (regular grammars) generate the regular languages [8]. The context-free grammar (Context-Free Grammar, CFG) is often used to represent a natural language. Partly because the form of context-free grammar is simple and the summary of grammar natural language processing required is convenient. On the other hand, context-free grammars have good algebra. It is easy to calculate and implement on computers.

THE FRAMEWORK OF OUR METHOD

Our method consists of four phases. In Phase 1, we randomly select about 25% of total corpus as training corpus. The remaining corpus is test corpus. In Phase 2, we cluster the training corpus $D = \{d_1, d_2, \cdots, d_n\}$ into $C = \{c_1, c_2, \cdots, c_m\}$ using text clustering technologies. The purpose is to understand the whole hierarchy of Chinese IPO short text. In Phase 3, we design the KDIPOL. This phase has two steps: (1) Design syntax rules of every sentence, (2) Induce them together and program the semantic actions. Last but not least, we test KDIPOL with KDIPOL parser in an iterative way, including add new grammars, correct the wrong grammars according to parse trees. The framework of our method is presented in Fig.2.

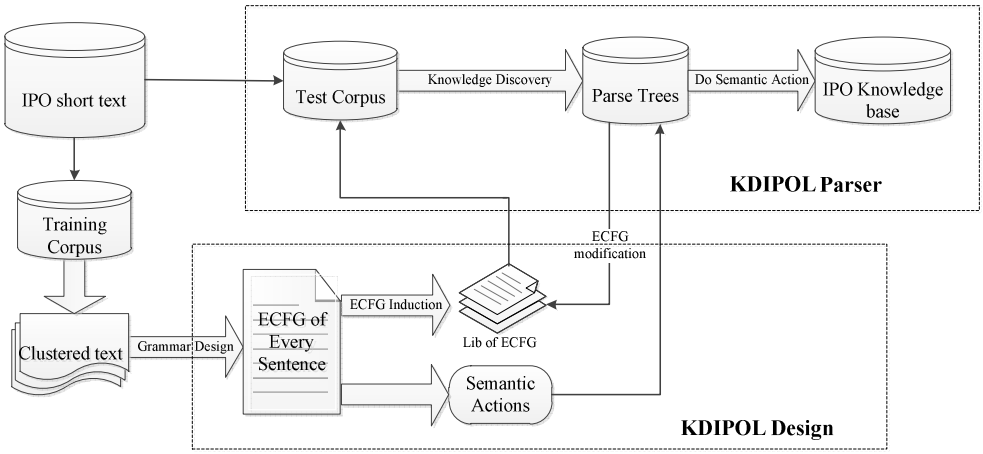

Fig. 2: The framework of our method

THE DESIGN OF KDIPOL

According to the characteristics of Chinese IPO short text, we propose a common program language to meet different demands. Users can program this language according to their processing granularity. In accordance with this idea, we design a Knowledge Discovery from Chinese Internet Public Opinion Language, KDIPOL for short. The basic structure of KDIPOL is as follows:

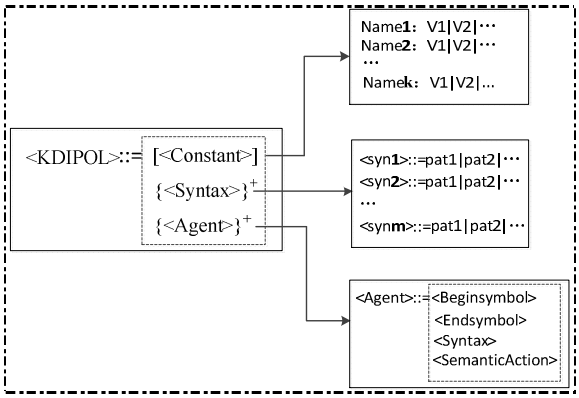

Fig. 3: The basic structure of KDIPOL

In Fig.3, KDIPOL has three parts.

“Constant” are constant words sets. Symbol “|” means logical “or”. We define high frequency words, words appear in similar position and some punctuation as “Constant”. This will be convenient for syntax rules to call them. Take “*程度副词：很|非常|极其(Adverbs of degree: very | quite | extremely)*” for an example.

“Syntax” is a set of syntax rules. Every syntax rule has one or more patterns. They are the key factors to parse IPO short text accurately. The pattern is described by ECFG which is mentioned previously.

“Agent” is a set of intelligent parse unit. In KDIPOL, one or more agent exists. Each agent has triggering conditions(string’s begin symbol and end symbol), begin symbol of syntax rules and semantic actions. Agent will be active when environment meet triggering conditions. Then text will be parsed by the binding syntax and do semantic action to acquire structured knowledge. An example as follows:

```
Constant
{
  标点：，|。|？、(punctuations)
程度副词：太|很|非常(Adverbs of degree: too | very)
差评形容词：冷漠|差(Adjectives of negative feedback: indifferent | unconcerned)
}
Syntax
{
  <消费者评价>::=<服务方><评价内容>(<Consumers’ comments>::=<Server><Comments>)
  <服务方>::=<#服务方>的[服务]态度(<Serving attitude>::=<#server>’s[serve]attitude)
   <评价内容>::=[也][<!程度副词>]<!差评形容词>[了][吧]
  (<Evaluation contents>::=[AD][< Adverbs of degree!>[<! Adjectives of negative
    feedback >[AS][AS])
}

Agent
{
  <Beginsymbol>::= “*” or <! Beginsymbol >
  <Endsymbol >::= “。” or <! Endsymbol >
  <Syntax>::=<消费者评价>
   SemanticAction(¬streq(<#服务方>,“”))--> Creatframe(<#服务方>,“评价”)∧Insertslot(“差评”,<
    差评形容词>)
  SemanticAction--> Closeframe(<#服务方>)
}
```

Fig. 4: An example of KDIPOL

IPO short text like “中国银行态度太差了(Bank of China has a bad attitude to depositors and lenders)”, “麦当劳服

务态度很差(McDonald's service attitude is very poor)" can be parsed by the KDIPOL described above.

The Design of ECFG

1) Definition of ECFG

The Extensible Context-Free Grammar(ECFG) is defined by the 6-tuple:

$$ECFG =< V_N, V_T, T_K, T_C, T_V, \wp >, \text{where} \quad (1)$$

(1) V_N is a finite set of non-terminal symbols. Each element represents a different type of phrase or clause of IPO short text. Every agent in the parser will bind a V_N. V_N used to represent a whole sentence of IPO short text.

(2) V_T is a finite set of terminal symbols. V_T has two types of elements: keywords and variables. Disjoint from V_N, V_T makes up the actual content of IPO short text. The set of V_T is the alphabet and variables of IPO short text language defined by the grammar $ECFG$.

(3) T_K is a finite set of keywords. Each shows as a keyword or a string in the grammar. T_K can be defined as an optional item. Its definition format is: "Keyword" or "[String]". "Keyword" is a required item and "[String]" is an optional one.

(4) T_C is a collection of constant items. Constant item is composed of one or more Chinese words or punctuations. It will be more convenient to define higher frequencies, similar meaning words into a set. The constant should be defined in "Constant".

(5) T_V is a collection of text variables. It is a non-empty finite set, too. Variable item is a string variable. Its definition format is: " <#VariableName>". Variable item can be defined with some semantic restrictions, such as non-empty. "<#非空字符串>(<#non-empty string>)" means the variable item must match a non-empty string. "<#非空字符串$<!标点>>(<#non-empty string$<! punctuations >>)" means the variable item must match a string which is not empty and not contain keyword "<!标点>(<punctuations>)" defined.

(6) $\wp$ is a finite relation from V to $(V \cup T)^*$, where the asterisk represents the Kleene star operation. The members of $\wp$ are called the (rewrite) rules or productions of the grammar. Its definition format is: "<Syntaxname>".

Furthermore, the items of $ECFG$ follow the following rules:

(1) $V_N \cap T_K \cap T_C \cap T_V = \phi$;

(2) $T_K \cup T_C \cup T_V = V_T$.

2) The design principles of ECFG

The design of ECFG is the key to parse IPO short text accurately. The design principles of ECGF are as follows:

(1) Prepare comprehensive corpus.

Before designing ECFG, the comprehensive corpus of one area is necessary. The syntax rules are designed according to the IPO text. Only if you design different structure and different semantic characteristics as much as possible, can you generalize the high performance ECFG.

(2) Naming should be normal.

The name of non-terminals and terminals must be able to explain the meaning of their definitions effectively. For example, it is irregular if we define "发生于(happen)" as the happen time symbol name. Because "发生于(happen)" means both happen time and happen place. We should use "发生时间(happen time)" . Naming will be unambiguous and comply with the general the language principle.

(3) ECFG should be extensible.

It means adding new syntax rules without modifying the original and existing grammar. In other words, ECFG should

be in a clear hierarchy and has a distinctive structure. We can add and modify the ECFG according to the different processing demand.

3) The design method of ECFG

Following the design principles of ECFG mentioned above, we summarized an iterative method of designing ECFG.

(1) Cluster the IPO short text $D = \{d_1, d_2, \cdots, d_n\}$ into $C = \{c_1, c_2, \cdots, c_m\}$ using text clustering technologies. The purpose of clustering is grasp the whole hierarchy of Chinese IPO short text.

(2) For every cluster $c_i (1 \le i \le m)$ of clustering result $C = \{c_1, c_2, \cdots, c_m\}$, we design syntax rules of every sentence under the design principles of ECFG. We mark these as $G = \{G_1, G_2, \cdots, G_m\}$. The design process of an example sentence is shown in Fig.5.

电信信号真的不错，但服务态度太差
(China Telecom's signal is really good, but its servers' attitude is bad.)

➔ 电信<信号质量短语>，但<服务态度短语>
<信号质量短语>::=信号[真的]不错
<服务态度短语>::=[服务]态度[太]差

➔ <#非空服务商><信号质量短语><!逗号>[但]<服务态度短语>
<信号质量短语>::=信号[<!程度副词>]<!积极评价词>
<服务态度短语>::=[服务]态度[<!程度副词>]<!消极评价词>
逗号：,
程度副词：真的|太
积极评价词：不错
消极评价词：差

➔ <#非空服务商><评价短语> //syntax rule
<评价短语>::=<信号质量短语>[<!逗号>][<评价短语>]|[但]<服务态度短语>[<!逗号>][<评价短语>] //syntax rule
逗号：, |,| //Constant
程度副词：真的|太|非常|很|特别|… //Constant
积极评价词：不错|好|… //Constant
消极评价词：差|恶劣|不好|… //Constant

Fig. 5: Design process of an example sentence

First, we analyze the main sentence structure and split every part. If a part is phrase or clause, we name a non-terminal symbol to represent it. Then generalize non-terminal symbols. We extract constants and variables. At last, we add keywords and syntax rules to strengthen the matching capabilities. If a syntax rule has several parallel components, we deal with it in a recursive way. Only right recursion is allowed in ECFG. In order to simplify the form of syntax rule, we can merge the syntax symbols together. We adopt symbol "|" which means logical "or" in this place and symbol "[]" means optional. Syntax symbol in the middle of "[]" means it's an optional symbol.

(3) Generalize the $G = \{G_1, G_2, \cdots, G_m\}$. After generalizing the G, we can obtain the lib of ECFG.

Step 4: Test the ECFG in an iterative way. According to the parse tree obtained by the parser, we add new grammars, correct the wrong grammars.

The Design of Parse Agent

After finish designing the KDIPOL, we can design one or more agents to parse KDIPOL. Each agent in the parser parses different syntax rules of ECFG and does different semantic actions.

We adopt frame-structured knowledge representation language to describe our result. The terminologies are as follows:

Frame----A frame is the basic semantic unit. It represents a set of entities and their relations in Chinese IPO short text. A frame is made of a set of slot and slot value pairs. A frame can be either intentional or extensional on the basis of whether its values are instances or types.

Slot-----A slot is a property or a binary relation extracted from the parse tree. A slot corresponds to one or more slot values.

Slot value-----A slot value is either the lemma form of a term from the sentence being extracted or a concept. A slot value corresponds to a slot.

A complete agent has three parts: begin and end symbols of Chinese IPO short text, syntax rules and semantic actions. Fig.4. shows an example of a complete agent.

The main purpose of semantic actions is turn parse tree to structured knowledge representations. Table 1 shows part of the semantic actions.

Table. 1: Semantic Action Functions(Part)

Type	Name	Function
Logical Symbol	∧	Logical “and”
	∨	Logical “or”
	¬	Logical “not”
Bool Function	Streq(p1,p2)	Parameter “p1” and “p2” are both string. Return TRUE if “p1” is equal with “p2”, else return FALSE.
	Contain(p1,p2)	Return TRUE if “p1” contains “p2”, else return FALSE.
	Beginwith(p1,p2)	Return TRUE if “p1” is begin with “p2”, else return FALSE.
	Endwith(p1,p2)	Return TRUE if “p1” is end with “p2”, else return FALSE.
String Operation	Strcat(p1,p2[,pn])	Concatenate string “p1” and “p2” and “pn”(if exist) together, return the concatenating result.
	Prefix(seq, prefix)	Add string “prefix” to each element of “seq” sequence.
	Suffix(seq, suffix)	Add each element of “seq” sequence with string “suffix”.
Frame Operation	Createframe(framename,classname)	Create a frame, the frame type is “classname”, frame name is “framename”.
	Closeframe(framename)	Close the frame whose name is “framename”.
	Insertframe (framename,slot,value[,slot1,value1])	Insert a slot to frame whose name is “framename”, the slotname is “slot” and slotvalue is “value”, if exist “slot1”, insert it all the same.
	Insertslot(slot,value[,slot,value])	Insert a slot to all the opening frame, the slotname is “slot” and slotvalue is “value”, if exist “slot1”, insert it all the same.

After finishing the above phases, we can get a complete KDIPOL of Chinese IPO text.

THE PARSER OF KDIPOL

We design the parser of KDIPOL by adopting theories of multi-agents, heuristic pattern matching and automata theory. This is a multi-agent system which is a computerized system composed of multiple interacting intelligent agents. Each agent is partially independent, self-aware and autonomous, but they collaborate together to accomplish parsing the Chinese IPO short text.

In order to enhance the matching efficiency, we present a heuristic search algorithm in the process of syntax analysis. For every pattern in ECFG syntax rules, we extract the keywords characters which work as the restricted string set in preprocess step. In this way, unnecessary backtrack will be avoided in process of syntax pattern matching. The execute efficiency of parser is in a high performance. We can get the parse tree in reasonable time. Concrete model of KDIPOL parser is shown in Fig. 6.

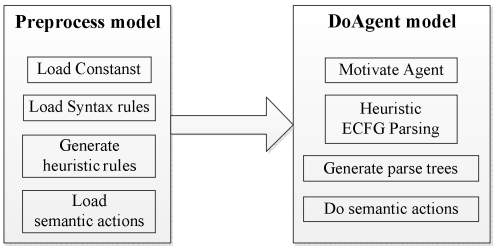

Fig. 6: Concrete model of KDIPOL parser

1) Preprocess model

Preprocess model is to build KDIPOL and load string data to memory, including four steps. In addition to generate heuristic rules, the remaining three steps are required for lexical checking and grammar checking, which is to ensure the correct format of KDIPOL program.

Load Constants. Constants used to be called by system in the process of syntax analyzing. The main purpose of this step is to construct corresponding constant dictionary, which serves for system to access to the contents of dictionary efficiently in syntax parsing process.

Load Syntax rules. This step is to construct syntax rule structure and write it to memory. We use linkList structure to save syntax rule structure. First, we check the writing legality of ECFG. Because non-terminals exist in syntax rules, this step will map all the non-terminals to their corresponding syntax rule.

Generate heuristic rules. This step is the key step of heuristic search. The heuristic rules in this system are structures of restricted symbols. Such as syntax rule must begin with what keywords or contain what strings. The details of generate heuristic rules are presented in algorithm 1.

Algorithm 1. Generate heuristic rules

Input: Constant dictionary, syntax rule linkList.

Let R be structures of restricted symbols.

We denote Constant dictionary as D, syntax rule linkList as L and the node pointer of L as L_{IDC}.

Begin

While L has next node

Denote U_{IDC} as the unit pointer of L_{IDC}, initially the first unit of L_{IDC}. $U_{IDC}++$ means U_{IDC} point to the next unit.

IF U_{IDC} is keyword or constant calling, add keyword or the corresponding words in D to R. $U_{IDC}++$. Continue.

IF U_{IDC} is optional or variables, jump to the next unit. $U_{IDC}++$. Continue.

IF U_{IDC} is non-terminal, get restricted symbols of its corresponding syntax rule. $U_{IDC}++$. Continue.

End While

End

Output: Structures of restricted symbols.

Load semantic actions. The function of semantic actions is turn parse tree to the frame-based knowledge representation form. This step is to find functions from system pre-defined functions and write them to memory.

2) DoAgent model

After ECFG preprocessing, the following step is corresponding agent executing. Once an agent is motivated, ECFG syntax analysis and semantic action will be executed.

Execute process of doAgent model is shown in Fig.7.

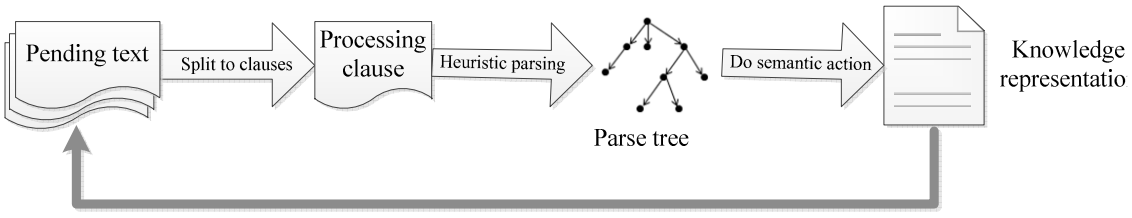

Fig. 7: Execute process of doAgent model

Motivate Agent. If the cut text which is cut according beginsymbols and endsymbols meets the demanding definition of the main conditions, the corresponding agent will be motivated.

Heuristic ECFG Parsing. EGFG parsing will be conducted by the above agent. This is a process of pattern matching.

We introduce heuristic rules and backtracking mechanisms to do ECFG pattern matching. We use the depth-first search algorithm to match the correct symbols. The details of heuristic ECFG parsing algorithm are presented in algorithm 2.

Algorithm 2. Heuristic ECFG parsing

Input: Chinese Internet Public Opinion text string S, syntax rules R.
Let syntax rule stack L empty.

Step 1: Mark the begin unit of R as N_0. Push N_0 to stack L.

Step 2: If S is null and L is empty, the parsing is success, turn to step3. Else if S is null but L is not empty or S is not null but L is empty, R parse S failed, turn to end. Pop the top element of L, denote the element as n, and delete n from L.

Case 1, n is a non-terminal. If n meets heuristic rules, push units of n's syntax rule to L. If not, if n is optional, n matches an empty string, else backtracking. Turn to Step 2.

Case 2, n is a keyword. If n is begin string of S, n's value is this keyword and intercept it from S. If not, if n is optional, n matches an empty string, else backtracking. Turn to Step 2.

Case 3, n is a constant calling. If the corresponding constant dictionary has begin string of S, n's value is this string and intercept it from S. If not, if n is optional, n matches an empty string, else backtracking. Turn to Step 2.
Case 4, n is a variable. n's value is decided by its next unit. Pop the top element of L, denote the element as n', and delete n' from L.

① n' is a non-terminal. Treatment approach likes Case 1. If n' meets heuristic rules, push n' and units of n''s syntax rule to L. If not, if n' is optional, n' matches an empty string, else backtracking. Turn to Step 2.
② n' is a keyword. If n' is substring of S, n''s value is this keyword and n's value the string before n''s value. Intercept them from S. If not, if n' is optional, n' matches an empty string, else backtracking. Turn to Step 2.
③ n' is a constant calling. If n''s corresponding constant dictionary is substring of S, n''s value is this keyword and n's value the string before n''s value. Intercept them from S. If not, if n' is optional, n' matches an empty string, else backtracking. Turn to Step 2.
④ n' is a variable. It is invalid of two adjacent variables. Backtracking and turn to Step 2.

Step 3: Save the ancestor and all its sons and grandsons of parsing result.

Output: The ancestor node of parsing result.

In algorithm 2, if a node n is not a valid solution, backtracking will be applied. The whole sub-tree rooted at n will be pruned. If n has another pattern, pop this pattern to stack L and continue parsing. If success, continue; else backtracking recursively.

Generate parse trees. Parse-tree will be generated in this step according to the result of ECFG parsing. System will traversal from the root node to its sons recursively, output the matched node and its value.

Do semantic actions. After the above steps, system finished parsing. The agent will do semantic actions defined in agent according to parse trees. The purpose of semantic actions is turn parse trees to frame-based knowledge representation form. It is a structured and computer-readable form.

RESULTS AND DISCUSSION

In order to validate the validity of the method we proposed above, we apply it in the analysis of IPO short text about corruption and product(service) comments. They are both hot topics nowadays in China nowadays.

We extract Chinese IPO short texts about corruption and product or service industry from Weibo, blog, BBS and other popular network applications. They are a single sentence, a phrase or several phrases. This corpus worked as the experimental corpus.

First, we cluster the corpus. After clustering, the corpus is clustered into 4 categories. The texts in one cluster have same theme or have same structure.

Then we split the corpus of every cluster into sentences or clauses with Chinese punctuations. An example of splitting corpus is shown in Fig.8.

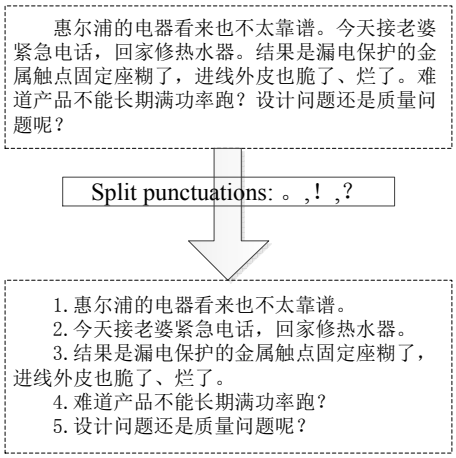

Fig. 8: An example of splitting the corpus

After splitting, the total number of sentences of this corpus is 2087, including 54,923 Chinese characters.

Table. 2: Clustering and Splitting Results Statistics

Cluster	Number of Sentences or Clauses	Topic	
		Corruption	Comments
Cluster 1	682	298	384
Cluster 1	497	177	320
Cluster 3	332	154	178
Cluster 4	576	246	330

Next, we design KDIPOL about this corpus. We randomly select about 20% of above sentences. According to the design method described in Section IV, we summarize every sentence or clause's syntax rule and generalize them together. After that, we design agents in the light of characters of every cluster, including beginsymbol and endsymbol, head of syntax rules and write semantic actions. After that, we get the KDIPOL and let the remaining sentences as the text corpus.

We employ the standard measures to evaluate the performance of our method, i.e. precision, recall and F1 measure. Precision(P) is the proportion of actual correct slots. In this experiment, a slot is correct means slot name and slot value are both correct. Recall(R) is the proportion of actual obtained and correct slots in the data.
F1 is the harmonic average of precision and recall as shown below:

$$F_1 = 2 * P * R / (P + R) \quad (2)$$

We do this experiment in our parser of KDIPOL. The total number of sentences of test corpus is 1670. The execute result is shown in Table 3.

Table. 3: Execute Result

Total number of test sentences	System executing time	Number of parse trees	Number of frames
1670	340s	879	330

An example frame is presented in Fig.9.

defframe 电器：产品质量问题
{
产品品牌：惠而浦
产品类型：热水器
电器质量：不太靠谱
电器设计：有问题
电器质量：有问题
}

Fig. 9: An example frame

In order to evaluate the correctness, we selected 417 sentences from the test corpus randomly and added semantic tagging to them manually. Then compared the experimental results and marked results.

Table. 4: Experimental Result About P, R and F1

Text Category	P	R	F_1
Quality of services	90.74	90.74	90.74
Service attitude	89.66	91.95	90.79
Bribery	90.74	93.52	92.11
Embezzlement	88.89	88.89	88.89

From the experiment results shows above, we can conclude that our method has a good performance both in accuracy and recall. This sufficiently illustrates the effectiveness of the method mentioned above. Although the accuracy and recall have both reached a high level, errors come from two main aspects:

1)The mismatch of ECFG. This is mainly because matching symbols semantic constraints on grammar are not strong, resulting in a false match. The solution is in the semantics of the original without compromising the ability of generalization grammar patterns, causing mismatched generative grammar strengthen semantic constraints or add new grammar production rules.

2)The absence of ECFG. This is mainly because there is no description of the corpus of existing grammar grammar or semantic constraints too. The solution to the situation in general is adding new grammar rules in the original production rules on grammar.

CONCLUSION

With the rapid development of Internet technologies, more and more IPO short text is appearing on the Internet. It is important for governments, business and even our own to grasp the IPO trends and control them. In order to overcome the problem that traditional method of IPO is low performance because of lack of semantic processing which is necessary. We propose a novel method base on semantic. We skillfully apply an extensible context-free grammar which is a formal grammar to this area. This method shows high performance through experiments in different types of corpus.

It is the first time to apply pattern-based combine with keywords-based methods to Chinese IPO short text. There are still some more works to be done in the future. First of all, the structure of Chinese IPO short text is so complicated that current patterns can’t parse it. Moreover, the knowledge we have acquired is not accurate at all.

Above all, it is necessary for us to solve some problems in the future. In the future, our work includes automatic syntax learning and add knowledge verification model. We will combine some methods (such as ontology knowledge base, concept space etc.) to verify the acquired knowledge.

Acknowledgments
This work is supported by the National Natural Science Foundation of China (Grant No. 61105040, 61203284, 61272361), the Beijing Natural Science Foundation (Grant No 4133085), the Beijing municipal commission of education young top-notch personnel plan (006000543114509) and the Beijing University of Technology Science Foundation (Grant No. 00600051 4311002).

REFERENCES

[1] N. Thanthry, M. S. Ali, R. Pendse. Security, Internet connectivity and aircraft data networks. *International Carnahan Conference on Security Technology*. pp.251-255, **2005**.
[2] Xin M, Wu H, Niu Z. A Quick Emergency Response Model for Microblog Public Opinion Crisis Based on Text Sentiment Intensity. *Journal of Software*, **2012**.
[3] Xiaohu Y L T W H, Lan J. Analysis and Early Warning to Government's Network Public Opinion. *Journal of Modern Information*. **2011**.
[4] Guan, Q., Ye, S., Yao, G., Zhang, H., Wei, L., Song, G., & He, K. Research and design of internet public opinion analysis system. In Services Science, Management and Engineering. *IITA International Conference*. pp.173-177, **2009**.
[5] Wang G. A Novel Algorithm of Internet Public Opinion Evaluation. TELKOMNIKA *Indonesian Journal of Electrical Engineering*, pp.3990-3996, **2013**.
[6] Poon H, Domingos P. Unsupervised semantic parsing. //Proceedings of the **2009** Conference on Empirical Methods in Natural Language Processing: Volume 1-Volume 1. *Association for Computational Linguistics*, pp. 1–10, **2009**.
[7] De Silva D, Alahakoon D. Incremental knowledge acquisition and self-learning from text.//Neural Networks (IJCNN), *The 2010 International Joint Conference*. pp.1-8, **2010**.
[8] Hopcroft J E. Introduction to Automata Theory, Languages, and Computation, 3/E[M]. *Pearson Education India*, **2008**.
[9] Kešelj V, Cercone N. A formal approach to subgrammar extraction for NLP. *Mathematical and computer modeling*. pp. 394-403, **2007**.
[10] Fan J, Kalyanpur A, Gondek D C. Automatic knowledge extraction from documents. *IBM Journal of Research and Development*. pp. 1-10. **2012**.
[11] Gama J, Rodrigues P P, Spinosa E J. Knowledge discovery from data streams. Boca Raton: *Chapman & Hall/CRC*, **2010**.
[12] Wang Haitao. Research on Text Knowledge Processing and Its Application in Intelligent Narrative Generation. *Chinese Academy of Science*, **2010**.
[13] Zhao L, Ichise R. Mid-ontology learning from linked data.//The Semantic Web. *Springer Berlin Heidelberg*. pp.112-127, **2012**.
[14] Klein D. The unsupervised learning of natural language structure. *Stanford University*, **2005**.
[15] Clark A. A learnable representation for syntax using residuated lattices.//Formal Grammar. *Springer Berlin Heidelberg*.pp.183-198, **2013**.