# CHINESE SPOKEN LANGUAGE ANALYZING BASED ON COMBINATION OF STATISTICAL AND RULE METHODS

*Guodong Xie, Chengqing Zong, Bo Xu*

National Laboratory of Pattern Recognition, Institute of Automation
Chinese Academy of Sciences, Beijing, 100080
E-mail:{gdxie,cqzong,xubo}@nlpr.ia.ac.cn tel:(010)82614468

## ABSTRACT

A combination of statistical and rule methods has been developed for Chinese spoken language analyzing. The analyzing result is a middle semantic frame, which can be converted to different language according people's needs. We adopt the statistical method in the stage of extracting semantic meaning and the rule method in the stage of mapping the semantic units to middle semantic frame. Experiment shows this method has high robustness and can analyzing Chinese spoken language effectively.

## 1. INTRODUCTION

In this paper we present a Chinese spoken language analyzing method based on the combination of statistical and rule methods. The analyzing result is a middle semantic frame[7]---IF(Interchange Format),which is adopted by C-STAR (Consortium for Speech Translation Advanced Research)[8], This consortium aims to build a speech to speech translation system

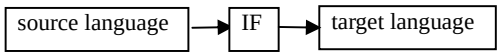


*Figure 1:* Translation between different languages

among multi-language in travel planning domain. Now the system includes Chinese, English, Japan and other languages. IF expressions can be converted to different language according people's needs, so to accomplish the translation between different languages. Figure 1 shows this process.

The task of the spoken language analyzing is to extract the semantic meaning from the sentence. In spoken language, the sentence seldom up to the grammar, there are full of repetitions, omitting, reversal, etc[6], so it is difficult to analyze the sentence base on the pure rule method. Some systems improve the rule method through adopting some technique and get a better result, for example Alon Lavie's [10] and YanPengju's [11], but all these systems aim at a certain domain, when being transplanted to other domain, people have to spend long time and do a great deal work[9].

Recently, statistical approach has shown its advantage in natural language processing. [2] and [5] presented a statistical method to analyze the natural language, the semantic analyzer is a ergodic HMM(Hidden Markov Model) [1] indeed. One characteristic of the HMM is that it need to be trained with enough annotated corpus. If only we have enough corpus and enough time to annotate this corpus, we will get the statistical language analyzing model---HMM through training with the corpus. If this model need to be transplanted to a new domain, we only need to annotate the corpus of the new domain, then use this corpus to train the model, and will get model that can meet the need of the new domain.

[2] and [5] were designed for the spoken dialog system. They adopt the case grammar frame[3] as the analyzing result which are not fine enough to meet the need of translation. We hope analyzing result is IF expression which can represent the meaning of the sentence honestly. So we put forward the method based on the combination of the statistical and rule approaches.

This paper is organized as: section 2 introduce the middle semantic frame---IF; section 3 introduce the analyzing method which include preprocessing, design of the semantic units and statistical analyzing model, etc; section 4 introduce the experiment; section 5 is conclusion; section 6 is acknowledge and section 7 is references.

## 2. MIDDLE SEMANTIC FRAME---IF

IF is a man-made system based on middle semantic relation[7] . It can represent the meaning of the dialog in travel planning domain, but it doesn't include peculiarities of any natural language[4].

IF is composed with four parts.

- Speaker tag. It indicates who is speaking. There are two speaker tags, either 'c' indicating the client or 'a' indicating the agent.
- Speech Act. It indicates the speaker's intent such as *giving information, requesting information, making suggestions or beg pardon,* etc.
- Concept. It indicates the topic of the sentence. such as *reservation, room*. Different concepts can be combined together according the IF rules to express more complex topics. Speech Act and concepts can form simple IF expression. For example: *c:give-information+reservation+room* is a IF expression, Its meaning can be following the sentence: *I want reserve room*.
- Argument. It indicate the specific detail of the sentence, such as *room price, room level,* etc. An Argument consists of an argument name and a value

separated by an equal sign[4]. For example, *room-spec=single* which means a *single room*.

A IF expression represent a sentence in dialog. Normally, each IF expression has a speaker tag and at least one speech act optionally followed by string of concepts and optionally, a string of arguments[4] as follows:

speaker: speech act+concept*(argument*)

the '*'indicate the concept and argument can appear in a IF expression repeatedly. Here is an example of IF expression:

*c:give-information+reservation+room(room-spec=single)*, its meaning is *I want reserve a single room*.

## 3. ANALYZING METHOD

The process of the analyzing method is: Firstly, collecting the corpus. Our analyzing system aims at the hotel reservation domain. We collected 2000 sentences of this domain, of which 1500 sentences is the training corpus and 500 sentences is the testing corpus. Secondly, Annotate the 1500 sentences manually and train the HMM with these sentences. Thirdly, Use the HMM to analyzing the sentence and the result is the semantic units sequence. Fourthly, Map the semantic units sequence to IF expression use the rule method. Figure 2 shows this process.

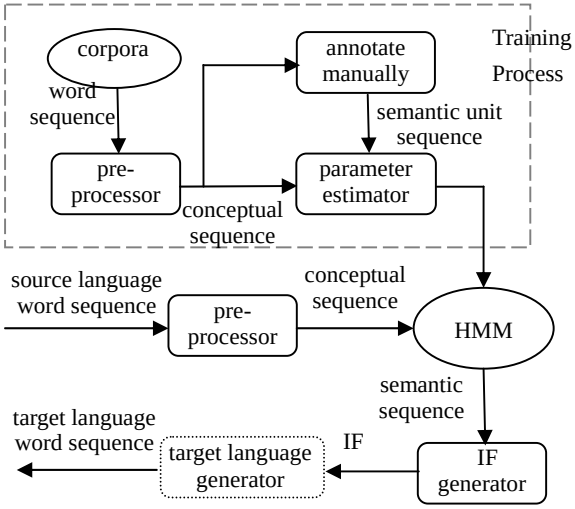


*Figure 2:* Overview of Chinese spoken language analyzing system

### 3.1 Preprocessing

The purpose of the preprocessing is to map the words into conceptual category. After the preprocessing, all the words of the sentence will be mapped into corresponding concepts, the output of the preprocessing is conceptual sequence.

For natural language understanding, the analyzing result we need is the semantic representation of the sentence. Different sentences have different semantic representation, but the similar sentence structure will have the similar semantic representation. The following two sentences have same middle semantic frame but different values of the argument "room-spec". (the latter parts is their IF representation).

我(I)想(want)预定(reserve)一个(one)单人间(single room)
----give-information+reservation+room(room-spec(single, quanlity=1))
我(I)想(want)预定(reserve)一个(one)双人间(double room)
----give-information+reservation+room(room-spec(double, quanlity=1))

We adopt the statistical model to analyzing the sentence, but the statistical model is obtained through training. Let's imagine when the word "单人间(single room)" was trained in the training process but the "双人间(double room)"wasn't, in this case if the statistical model meet the "单人间(single room)" and "双人间(double room)", it can analyze the first one properly but the second one. There are many examples like this in natural language. The corpus we can collect is limited and it is impossible that every word we meet will be trained. The solution to this problem is to classify the word into different category. There is no standard to classify the words, common principle is to classify according the semantic function of the word in the sentence.

For example: "单人间(single room)", "双人间(double room)", "豪华间(senior suite)"and "总统套间(royal suite)" and other words with the meaning of the room class are belonged to the conceptual category "N_O_ROOMLEVEL".

But some special process is necessary to the words that express the number. In natural language, the manners to express number are very flexible. A number always needs a phrase composed of several words. For example, "one thousand one hundred and twenty" is composed of five words, but as a phrase its semantic function in sentence is the same as "five thousands and three hundred" and "twenty". So it is necessary to classify the number phrase as one category. We adopt some simple rules to merge the words expressing the number to one category, so no matter "one thousand one hundred and twenty" or "twenty", they are treat as number phrases, we classify all them to one conceptual category.

In our system, we classify all the words to 230 categories, we call this category as conceptual category. After preprocessing, the word sequence become the conceptual name sequence. For example, word sequence *I want to reserve room*, its corresponding conceptual name sequence is *P_FIRST_PERSON V_STATE_DESIRE V_STATE_RESERVE N_C_ROOM*.

In training and analyzing process all words are treated as their corresponding conceptual category name.

### 3.2 Design of the Semantic Units

To annotate the corpus is the first step to build HMM. After preprocessing, the word sequence become the conceptual category name sequence, For every conceptual category name, we annotate its semantic unit, at last every conceptual name sequence corresponding to a semantic unit sequence. Then we use this conceptual sequence and semantic sequence pairs to train the model. In [2] and [5], the case grammar frame was adopted as the result of analyzing, the semantic units are some Cases extracted from the case grammar frame. In our analyzing system, the target of the analyzing is the IF, we hope the result of semantic analysis could be mapped into IF easily. So we extracted part of semantic units from the Speech Act, Concepts, Arguments of IF, another part are designed according the characteristic of the Chinese spoken language and rules of IF. All the semantic units can be classed to three categories:

- Independent semantic units. These semantic units only represent a single semantic meaning. For example,

*who, room-spec, currency* etc.

- Flag of other semantic units. These semantic units are indicators of other semantic units. For example, *family-name=, for-whom=* etc.
- Composite semantic units. These semantic units represent two or more semantic meaning. In Chinese, some words' meaning can't be represented with one independent semantic unit, so we design this composite semantic units to represent these words. For example, *trip:destination=here, price:quantity,* etc.

### 3.3 Statistical Analyzing Model---HMM

HMM can be named as <S,O,A,B, $\pi$ >, it is compose with five parts[1]:

- S, states of the model. The number of states is N
- O, output observation of the model. The number of observation is M.
- $A = a_{ij}$ , state transition matrix of the model.

  $a_{ij} = P(q_t = S_j \mid q_{t-1} = S_i)$ ,

  $1 \le i, j \le N$ , $a_{ij} \ge 0$ , $\sum_{j=1}^{n} a_{ij} = 1$
- B= $b_j(k)$, the probability matrix which the state $s_j$ output the observation $v_k$ according to.

  $b_j(k) = P(o_t = v_k \mid q_t = S_j)$ ,

  $1 \le j \le N$ , $1 \le k \le M$

  $b_j(k) \ge 0$ , $\sum_{k=1}^{m} b_j(k) = 1$
- $\pi = \pi_i$ ,the initiation probability vector.

$$\pi_i = P(q_1 = S_i) \quad , \quad 1 \le i \le N$$

$$\pi_i \ge 0, \qquad \sum_{i=1}^{n} \pi_i = 1$$

Every sentence represents a certain semantic meanings, Every semantic meaning need sentences to express them. The relation between sentence the sentence's semantic meaning can be embodied by the HMM. Words in the sentence can be the output observations of HMM, Semantic meaning of the sentence that we want to get can be the inside states of HMM. The parameter of the HMM can be estimated by the training with annotated corpus. In order to get the HMM, We first annotated some corpus manually, then use this corpus to train the HMM and estimate the parameters. Then we use this HMM to analyzing some other corpus, and collate the analyzing result manually. After collating, the result can be add to the training corpus to train the HMM again, continue like this, after several repetitions, we can get a better HMM.

In our system the HMM's states transition probabilities model is bigrams, which can model only the adjacent semantic relation but cannot cope with the longer distance semantic relation[2], so we adopt a ergodic HMM, that means all semantic units can follow each other. We annotated 1500 sentences altogether to train the model, there are 171 semantic units and 169 conceptual categories involved, which corresponding to the *states* and *observations* of the HMM.

### 3.4 Map Semantic Units Sequence to IF Expression

The output of the HMM is the semantic units sequences. The semantic units are discrete, the restriction between semantic units still not show in form.

| semantic Units | parts of IF expressionIF |
|---|---|
| quantity=1, time-unit=day | duration=(quantity=1,time-unit = day) |
| who=I, disposition=disire | disposition=(desire,who=I) |
| Room-spec= double, or, room-spec= single | room-spec= (operator=disjunct,[double,single]) |
| Price:quantity=100, currency=us_dollar | price=(quantity=100),currency =us_dollar |

*Table 1:* Semantic units and corresponding parts of IF expression mapped from them.

Our target of analyzing is IF expression, which must obey strict IF rules. CFG(Context Free Grammar)[1] has been used in the natural language parsing successfully, in order to map the semantic units sequence into IF expression, we use the method based on the CFG. The word sequences of the spoken language sentence become the semantic unit sequences after the analyzing of HMM, the grammar relation in origin sentence become the semantic relation which is very simple compare to the relation between words in natural language. We use some context free rules and can map the semantic units sequence to IF expression. Through analyzing the IF document and the results of the HMM, we generalized 60 rules, then we use this rules and a simple LR parsing method[1] to analyze the semantic sequences and map them to IF expression. Experiments prove this method is feasible and error rate is lower. Here we give a simple example of mapping process.

Chinese sentence:*您好(Hello), 还(still)有没有(available or not) 房间(room)?* corresponding English sentence is: *Hello! Is single room available?* The analyzing result of HMM is the semantic unit sequences: *greeting=begin, availability=question, room.* When the system map the semantic unit sequence to IF expression, it use three rules as follows:

*greet=?→greeting(greeting=?)* (1)

*?=question→request-information+?* (2)

*availability, room→availability+room* (3)

The sign '?' indicate that position can be fill with any semantic unit.When the system scan the semantic unit sequence, it first find *greeting=begin*, which match the rule (1), so the system replace the it with *greeting(greeting=begin)*, then the system find the *availability=question* which match the rule (2), the system replace it with the *request-information +availability*, last the system find the *availability,room* which match the rule(3), the system replace it with the availability+room. At last the semantic

become the IF expression:
greeting(greeting=begin), request-information+availability+room.

Table 1 shows some examples of semantic units and corresponding parts of IF expression mapped from them.

## 4. EXPERIMENT

The HMM is trained by 1500 annotated sentences. We test this system using 500 sentences of this domain. The results show in table 2.

| different method | single semantic unit | last analyzing result |
|---|---|---|
| 1 | 91.4% | 52.8% |
| 2 | not given | 72.0% |
| 3 | 91.2% | 79.2% |

*Table 2:* Compare of accuracy of different methods in different analyzing stage

method 1: stochastic case frame approach in [2]
method 2: statistical understanding model in [5]
method 3: methods presented in this paper

| different method | training corpus (sentence number) | state number | observation number |
|---|---|---|---|
| 1 | 6439 | 330 | 737 |
| 2 | 1037 | 89 | 100 |
| 3 | 1500 | 171 | 169 |

*Table 3:* Compare of HMM in Different methods

When analyzing the sentences to semantic units, the accuracy for single semantic unit is 91.2% which is a bit lower than the method 1, this is because our training corpus is not enough. The accuracy for last result of analyzing is 79.2%, which is higher than the method 1 and 2. Table 3 shows the compare of HMM adopted in different methods.

On the other hand, both aimed at the hotel reservation domain, our system has 171 semantic units, the system in [2] has only 89 semantic units. This means that our system can extract more semantic information from the sentence, which will lay a good foundation for high quality translation. This can be explained by the following:

- Define more conceptual categories. We defined more conceptual categories than [5], so our system can capture more difference in grammar and meaning between words
- Define more semantic units. We defined more semantic units considering the characteristic of the Chinese. These semantic units can represent the meaning of the sentence more clearly, this mean more fine analyzing.
- Use the rules. We adopted plenty rules to map the semantic units sequence to IF expression, These rules ensure we can get better IF expression.

But we still face the 20.8% error rate. Through analyzing the result, we can find several reasons for the errors.

- Dialog history. In our system, we didn't consider the dialog history when analyzing, but sometimes the meaning of a sentence is implied by the dialog history but not show literally.
- Bigrams as HMM's states transition probabilities model. Bigrams can't cope with the longer distance semantic relation of words.
- Peculiarities of IF. IF expression has strict form. Also we adopted 60 rules to map the semantic unit sequences to IF expressions, but there are still some exceptions.

Next step we hope improve the performance of our system by increasing the training corpus, adopting the trigrams model and considering the dialog history, etc.

## 5. CONCLUSION

We adopted the combination of statistical and rule methods to analyze the Chinese spoken language, the analyzing result is IF. This method adopt the statistical approach and rule approach in different analyzing stage, it make use of statistical model's high robustness characteristic and can cope the Chinese spoken language effectively.

## 6. ACKNOWLEDGE

The research work described in this paper is supported by the National Natural Science Foundation of China under grant number 69835003 and 60175012, and the National Key Fundamental Research Program (the 973 Program) of China under the grant G1998030504.

## 7. REFERENCES

[1] Weng Fu-Liang, Wang Ye-Yi. Introduction to Computational Linguistics, *China Social Science Publisher.1996*
[2] W.Minker, S.Bennacef. A stochastic Case Approach for Natural Language Understanding. *Proc. ICSLP, 1996*
[3] B.Bruce. Case Systems for Natural Language. *Artificial Intelligence. 1975. 6:327-360.*
[4] Lori Levin, Donna Gates. An Interlingua Based on Domain Actions for Machine Translation of Task-Oriented Dialogues. *Proc. ICSLP, 1998*
[5] Yunbin.Deng, Bo Xu. Chinese Spoken Language Understanding Across Domain. *Proc. ICSLP, 2000*
[6] Chengqing Zong, Hua Wu. Analysis of Spoken Dialog Corpus in Restricted Domain. *Proc. JSCL-99*
[7] Hua Wu, Taiyi Huang. Interlingua-Based Response Generation. *Proc. JSCL-99*
[8] Jun Park, Jae-Woo Yang. ETRI Speech Translation System. *C-STAR Workshop. Schwetzingen, 1999*
[9] Yunbin Deng. statistical Language Understanding---Transplanting Cross Domains in Spoken Language Dialog System. *Master thesis. Institute of Automation, CAS. Beijing, Jun, 2000*
[10] Alon Lavie. GLR*: A Robust Grammar-Focused parser for Spontaneously Spoken Language. *PhD. Thesis. Carnegie Mellon University. Pittsburge, PA, May 1996*
[11] Yan P.J., Zheng F.. Robust Parsing in Spoken Dialogue Systems. *7th European Conference on Speech Communication and Technology. Aalborg, Denmark, 2001*