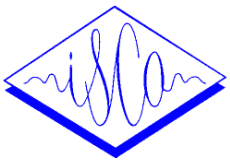

6th International Conference on Spoken Language Processing (ICSLP 2000)
Beijing, China
October 16-20, 2000

Chinese Spoken Language Understanding Across Domain

Yunbin Deng, Bo Xu, Taiyi Huang
National Lab of Pattern Recognition, Institute of Automation
Chinese Academy of Sciences, Beijing 100080
Email: (ybdeng, xubo, huang)@nlpr.ia.ac.cn

ABSTRACT

A robust parsing model for spontaneous Chinese based on semantic constituent spotting and concept assembling model (SCAM) had been successfully developed in our "LOADSTAR "dialog system[1]. It is a travel information accessing system and the SCAM is rule based. Considering the domain portability, a statistical model for spoken language understanding is adopted. The statistical spoken language understanding model is developed in the domain of hotel reservation. Then the statistical model was ported to the domain of travel information accessing within four weeks.

1. INTRODUCTION

Under the SCAM model (figure 1), when a new domain is introduced, most of the work done in the previous understanding component is not transferable. The work includes definition of vocabulary, semantic, concept, and description of the rules used in the combination of semantics and concepts. Further more, the rules are written and amended through analysis of the corpus of a specific domain, such work is time consuming and the rules of a domain can not describe the character and concept of another domain.

Language understanding involves syntactic analysis, to determine how the words group together, and semantic analysis, to determine the meanings of the constituents. These two processes may be kept separate at the representational level in order to maintain generality to other domains, but they tend to be combined during processing for reasons of efficiency.

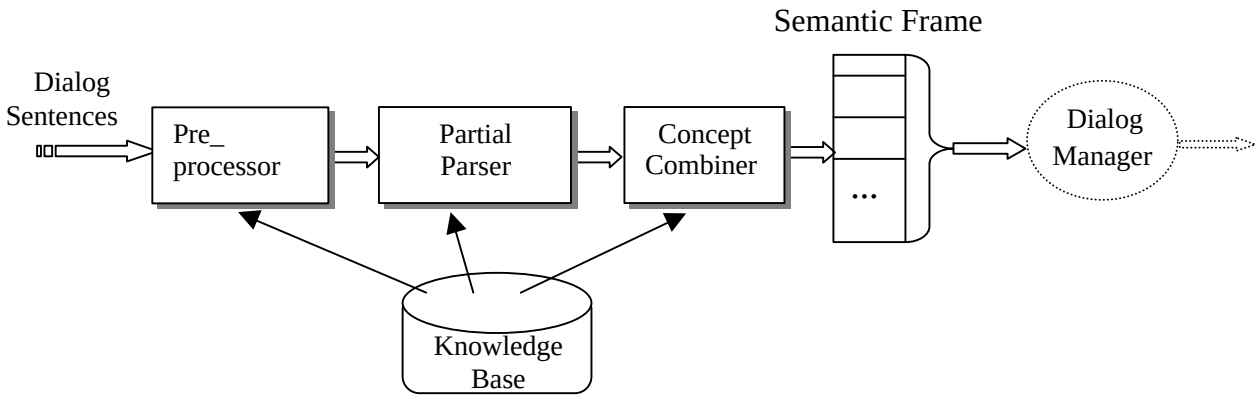

Figure 1: The Workflow Diagram of SCAM

2.KNOWLEDGE REPRESENTATION

Statistical methods have been used in parsing. For example, PCFG is used to analysis the structure of a sentence. In spoken language, the structure is usually incomplete, and the semantic analysis is more efficient. We adapt a statistical model, semantic case is used to annotate the training corpus, which acts as the semantic analysis. Through training, we get the parameter of HMM model, which acts as the syntactic analysis. By this way the semantic analysis and semantic analysis is Separated at representation level and combined by the HMM when parsing a sentence.

Considering the compatibility with other components of our dialog system, we define the semantic frame as follows:

HEAD：the information about sentence type
TOPIC：the main idea of a sentence
CASE： sub-topic of a concept

CASE MARKER：case identifier

Garbage：the component which has no semantic function

For example "*多少 钱 一 间 啊 ?*" is annotated as follows

{h:whq} {t:cost} {c:roomnum} {m:roomnum} {g:}

This semantic sequence can be easily mapped to semantic frame. The form of semantic frame is composed of three sections.

SemanticFrame ::= Head + TopicItem + $SubTopicItem_1$ {…… +$SubTopicItem_n$ }

Head ::= WHQ| YNQ|…… WHQ ： interrogative, YNQ：yes or no question……

TopicItem ::= (Topic, Value)

$SubTopicItem_i$ ::= ($Topic_i$, $Value_i$)

Value ::= Vi {op Vi }， op ::= &/ |

3. STATISTICAL UNDERSTANDING MODEL

To parse without rules, we adapt a statistical language understanding model [2].

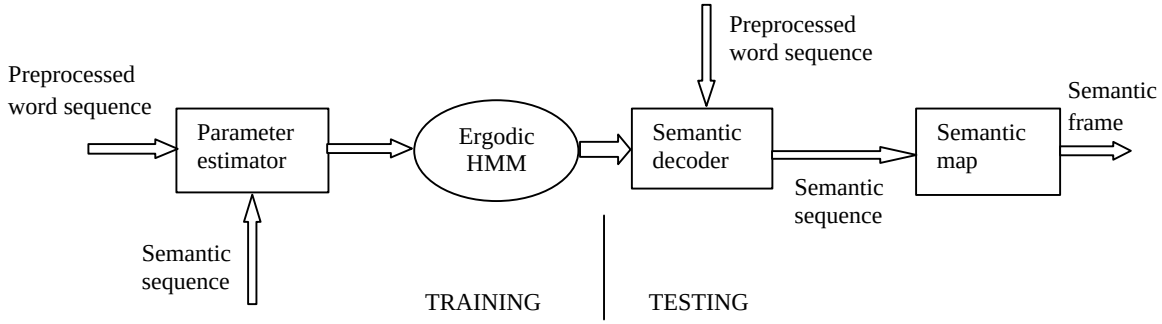

Figure 2: The Workflow Diagram of Statistical Spoken Language Understanding Model

The statistical model is mainly composed of flowing four components.

Preprocessor: With the support of semantic-class dictionary, words that have the same semantic function are clustered as a class. Considering the flexibility of Chinese spoken language, We use domain independent rule to preprocess the phrase of time and number. These tow approaches can reduce the size of statistical model considerably.

For example: "*到 九寨沟 一千 三百 多*" is preprocessed as:

"*到 N_O_PLACE num2 garbage*"

Paremeter estimator: We use the semantic case to annotate the corpus of specific domain. The preprocessed word sequence is the observation of the HMM, and the manually annotated semantic sequence is the state of the HMM. Because of the flexibility of word sequence of Chinese spoken sentence, we adapt an Ergodic HMM[3]. Discounting method is adapted to coupe with the sparse data problem.

Semantic decoder: Given a sentence (a word sequence), we get its preprocessed form automatically. With the the pre-processed word sequence (observation sequence) and the parameter of HMM, the semantic decoder is to find the "optimal" semantic sequence (state sequence). This is realized through the Viterbi algorithm.[4]

Semantic map: Semantic sequence is mapped to the semantic frame with domain independent rules. Topic is the most important thing the speaker cares. In most cases, the concept that is nearest to the "Head" is regard as topic. In some cases, it is hard to find the topic of a sentence. Fortunately, in most cases, the selection of topic will not affect the component of dialog management. For example:

"*暑期 有没有 去 敦煌 的 旅行团 啊*" **can be parsed as** "*REQ :(ROUTE,敦煌) :(TIME,暑期)|)&&*" **or** "*REQ：(TIME，暑期)：(ROUTE，敦煌)|)&&*"

Under the statistical understanding model, when a new domain is introduced, we need to add new words to the semantic entry to get preprocessed form of these new words. The drawback is that corpus should be annotated manually or semi-automatically.

4. EXPERIMENT

We build up the statistical understanding model at domain of hotel reservation, then we port the understanding model to the domain of travel information accessing within four weeks. The work of domain portability including: adding new entry of a new domain to the semantic lexicon and annotation of the training corpus of the new domain. The statistical model can be ported in a short period of time compared with the rule model.

Table 1 describes the model size of the two domains.

As we can see, the two domains have similar entry number and model size. And they share nearly half entry and state.

Table 2 shows the training corpus and test corpus and the corresponding result of the two domains. The input is a word sequence, and the output is semantic frame. A parsing result is regard as correct only if the sentence type, topic, subtopic and corresponding value are correct. The performance of the statistical model is pretty good. In the domain of Travel Information Accessing, we firstly annotated 903 sentences. With those training sentences, we got the parameter of HMM. Using this model, We annotated 894 sentences automatically. Then manually correct these sentences and added them to the training corpus. With these 1797 sentences, we got a new HMM. As it shows, the error rate decreased when the training corpus increases.

Table 3 illustrates the robustness of statistical understanding model. A Chinese spoken sentence**"*有 没有 坐 火车 去 上海 呢*"** is recognized as "*有 没有 走 火车 七 上海 能*" by error. The word "***走***" and "***七***" is parsed as meaningless words under such context. The meaning of the whole sentence is well understood.

Table 4 demonstates how we annotate a sentence when there is ellipsis. A Chinese spoken sentence**"*要 一 个 好/便宜 一点儿 的 房间 吧*"** in the domain of hotel reservation is usually represented as "*要 一 个 好/便宜 一点儿 的 吧*". If those sentences with ellipsis are well trained, they can be well understood during test.

Domain	Entry	Observation	State
Hotel Reservation	951	100	89
Travel Information Accessing	986	120	92
Common parts	533	65	45

Table 1: The statistical model of Hotel Reservation and Travel Information Accessing

Domain	Training Corpus (Sentence number)	Test Corpus (Sentence number)	Error Rate
Hotel reservation	1037	230	28.0
Travel Information Accessing	903	216	26.4
	1797(903+894)		22.7

Table 2: The training and test corpus of Hotel Reservation and Travel Information Accessing

Word Sequence	***有 没有 走 火车 七 上海 能***					
Preprocessed Form	*V_Q_AVAILABILITY*	*V_S_GOOUT*	*N_O_TRAFFIC*	*num1*	*N_O_PLACE*	*V_S_CAN*
Value Sequence	{}	{}	{***火车***}	{7}	{***上海***}	{}
Semantic Sequence	*{h:req}*	*{m:traffic}*	*{c:traffic}*	*{g:}*	*{t:route}*	*{g:}*
Semantic Frame	*(req:(traffic,**火车**):(route,**上海**):)&&*					

Table 3: The Robustness of the statistical understanding model

Word Sequence	***要 一 个 好 一点儿 的 吧***
Preprocessed Form	**要** num1 uPerson **好** advDegree C_FUNC_DE C_FUNC_YEH
Value Sequence	{} {1} {} {} {} {} {}
Semantic Sequence	*{t:bookroom} {c:roomnum} {m:roomnum} {c:roomlevel} {m:roomlevel} {m:roomlevel} {g:}*
Semantic Frame	*((bookroom,):(roomnum,1):(roomlevel,better):)&&*

Table 4: The annotating tactic for ellipsis.

As we can see from table 2, the error rate is as high as 22.7%. Through analysis, there are several reasons for these errors.

1.Ellipsis

Under the statistical understanding model, the observations and states are one-to-one. If there is an ellipsis in a spoken sentence and this kind of ellipsis is not well trained, the corresponding semantic sequence will have an ellipsis. For example, there is an ellipsis of "***钱***" in the sentence " ***标准间 是 多少 啊*** *?* ". The understanding result is as follows:

{t：roomlevel} {g：} {h：whq} {g：}
*(whq：(roomlevel，**标准间**)：)&&*

Though the correct understanding result should be

*(whq：(cost，)：(roomlevel，**标准间**)：)&&*

2. Bigram

We adapted bigram for the parameter estimator of the ergodic HMM. This model can not reflect the structural information of two words that are widely separated if such structure is not well trained. For example, the phrase "***一 个 人***" and "***一 个 单间***" can be parsed correctly. But the understanding result for the sentence "***呵 就 一 个 普通 的 单间 吧***" is as follows.

(：(personnum，1)：(roomlevel，***单间***)：)&&

Though the correct understanding result should be (：(roomnum，1)：(roomlevel，***单间***)：)&&

3. Annotation

Some times the meaning of a spoken sentence is connotative. For example: "***就 这 一 种 啊*** ? ". The mean of is sentence is "***有没 有 其他 类型 的 房间***" In such case, the meaning is determined by the whole sentence, but the word and semantic annotation is one-to-one. The annotation for such sentence is very difficult. The understanding result for this sentence is that this sentence has no meaning.

就 这 一 种 啊 *?*

{g：} {g：} {g：} {g：} {g：}

4. OOV

Because the entry and topic of the statistical model is got from the training corpus, new entry and topic of the test corpus can not be well understood. For example:"***黄庄路 位置 是 在 哪儿 林业大学 附近 吧*** *？*" The word "**黄庄路**", "**位置**" "**林业大学**"and the topic "***位置*** " have not been defined during training. The understanding result of this sentence is that this is a interrogative but have no topic.

{g：} {g：} {g：} {g：} {h：whq} {g：} {g：} {g：}
(whq：(，)：)&&

As we have discussed, the performance of the statistical model will improve as we increase training corpus. With more training data, We can use tri-gram to take the place of Bi-gram to improve the performance of the statistical model.

5.ACKNOWLEDGMENTS

We would like to thank our sponsor: Intel China Research Center, Intel China Ltd, in particular Prof. Yonghong Yan and Zhiwei Lin.

6.REFERENCES

1. Huang C., Xu P., Zhang X., Zhao S.B., Huang T.Y., Xu B.(1999), "Lodestar: A Mandarin Spoken Dialogue System For Travel Information Retrieval", *Proc. Eurospeech*, September, pp.1159-1162.
2. W.Minker(1997),"Stochastically-based Natural Language Understanding Across Tasks and Languages," *Proc. Eurospeech*, September, pp.1423-1426.
3. Lawrence R.Rabiner ,"A Tutorial on Hidden Markov Model and Selected Application in Speech Recognition" (1993) Fundamentals of Speech Recognition. Prentice Hall
4. Weng Fu-Liang, Wang Ye-Yi, "Introduction to Computer Linguistics" Chinese Social Science Publishing Company.