SPOKEN LANGUAGE UNDERSTANDING IN SPOKEN DIALOG SYSTEM FOR TRAVEL INFORMATION ACCESSING

ZHANG Xin, ZONG Chengqing, HUANG Chao, ZHAO Shubin, HUANG Taiyi

National Lab of Pattern Recognition, Institute of Automation, Chinese Academy of Sciences, Beijing 100080

Email: {zcq}@nlpr.ia.ac.cn

ABSTRACT

This paper briefly introduces the VOTIRS 2.0 system — a Chinese spoken dialog system for travel information accessing. Through this system, users can query the travel information about 52 routes and finally make a transaction for a proper travel plan. The strategy and method to understand spontaneous speech in the system are discussed in detail. To understand spontaneous speech, a Semantic Constituent Spotting and Assembling (SCoSA) hierarchical model is proposed. It is a semantic-driven multi-phase parsing process similar with human's understanding process. The model is efficient in parsing spontaneous speech.

1. INTRODUCTION

The mature of speech technology has brought spoken dialog system into horizon. Spoken dialog system offers people with a very natural, friendly and convenient human-computer interaction way. It meets the increasing need for people to access information resources friendly, efficiently and timely. It's finding more and more applications in the domains such as information retrieval, transaction making and real-time spontaneous speech translation, etc. Several dialog systems have been established in different domains, such as traffic information query (ATIS, SUNDIAL), travel information query (GALAXY), advisement (WHEELS), hotel guide (DINEX), weather forecasting information (JUPITER), etc.

In a dialog system, spontaneous speech understanding is a core issue. The goal to understand spontaneous speech is to find the user's intention. It's very difficult to understand spontaneous speech because there are lots of spontaneous phenomena that make sentences ill-formed. The classic parsing methods are not good at parsing spontaneous speech. To parse spontaneous speech, some different methods are adopted. In GALAXY, TINA, the language parser, adopts an augmented CFG to describe possible utterances. It integrates ideas from CFG, ATN grammar and the unification concept. [1] In JANUS system, a two-pass strategy is adopted. A strict syntactical GLR* algorithm is applied to the whole sentence. If it's failed, the PHOENIX system spots phrases from the sentence. [2] Some statistical methods are also used for parsing. [3]

This paper introduces a Chinese spoken dialog system VOTIRS 2.0 and the parsing strategy and method adopted in the system. To parse spontaneous speech, a SCoSA model is presented. It's semantic-driven and makes full use of dialog context knowledge and background knowledge. In the following section, the system is introduced in brief. The parsing strategy and method are discussed in the third section. In the final section, the preliminary evaluation result is presented.

2. SYSTEM OVERVIEW

We developed the voice-driven tourism information retrieval (VOTIRS) 2.0 system. It's a Chinese spoken dialog system for travel information accessing. Through talking to the system naturally, users can get desirable travel information such as time, price, etc, and finally make a decision for a proper travel plan under the conduction of the system. Information about 52 routes is available in the system, including beginning time, travel duration, price, level, traffic method, etc.

The dialog process in the system is of mixed-initiative. The system gives the first prompt and waits for users' initiative. When there is no speech input for some period of time, or the system can't understand what users said or needs information from users, the system gets the initiative to give proper prompts according to the context. It's very natural and friendly for users.

The whole system consists of 5 modules: speech recognizer,

language parser, dialog manager, natural language generator, and speech synthesizer. The diagram is shown as Figure 1.

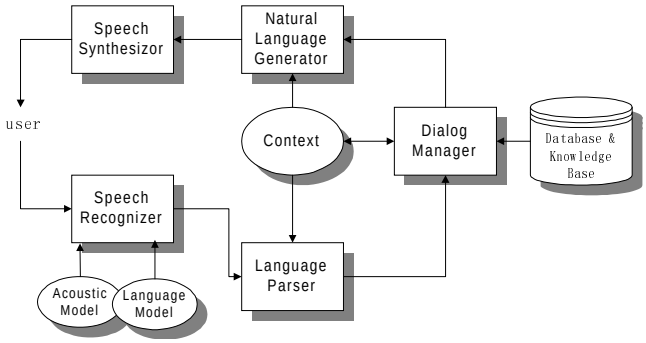

Figure 1: the system module diagram

- The speech recognizer adopts CDHMM technology. To deal with spontaneous speech, garbage models and filler models are added into the recognizer. They help to filler out speech fragments without any sense and words out of vocabulary. A mixed language model of word-based and class-based is used to solve the data sparsity problem.
- A SCoSA method is developed in the language parser. This method is efficient to deal with inputs with seriously ill-formed spontaneous speech and brings users a very flexible expression way.
- The dialog manager maintains a finite state model to manage the dialog process. For each dialog state, a transition condition and an operation are attached. The user's speech input causes the state to shift and the responding operation is done. By the transitions of the states, the manager directs the dialog to reach the final goal.
- The natural language generator adopts a template-based method and message trigging technology. Different templates are ready for different responses. When some kind of response is requested, a message trigs the proper template and fills in the blank in the template to form a natural language sentence.
- A concatenative method is used in the speech synthesizer to synthesize the final speech response according to the generated sentences.

3. SPONTANEOUS SPEECH UNDERSTANDING

3.1 Characteristics of Spontaneous Speech in Dialogs

In spoken dialogs, the sentences are different from those in written language. When designing the parsing algorithm, the characteristic of spontaneous speech should be considered. From the corpus collected, two main characteristics can be observed:

1) From the view of meaning, in dialog sentences, the meaning is usually simple and relates just one topic the user cares about. There is no complex rhetorical relation among the constituents of the sentence. A complex meaning relating several topics is expressed in several turns. Therefore, the meaning of a sentence can be represented in a simple form.

2) From the view of structure, the sentences are ill-formed as a whole. There are lots of spontaneous phenomena such as hesitations, stutters, insertions, corrections, etc, which destroy the normal structure. However, the structures of phrases in the sentences are strict. The word orders of the phrases have to conform to some strict linguistic rules. Furthermore, as a constituent, the meaning of the phrase is complete. In all, a sentence in dialogs is consisted of some constituents which have strict inner structure accompanied by some spontaneous speech phenomena.

The following sentence is an example to demonstrate spontaneous phenomena in dialogs.

Example: *我想问一下有没有去去就是说泰山曲阜的最近*

3.2 Basic Strategy

Based on the above characteristics in spoken dialog, some basic ideas should be taken into consideration to parse spontaneous speech:

1. The final goal to understand spontaneous speech input is to find out the user's intention and extract goal-related information. In a special domain, a dialog process is goal-oriented. As they talk, users express their inclination or supply some information about several finite topics. In the travel information domain, topics about time, travel routes, traffic vehicles, prices, landscapes are the things the user cares about mostly. To parse a user's input is to find the main topic and restrictive information.
2. The parsing process should be semantic-driven, aided with syntactical parsing. Since the whole sentence structure is ill-formed, there is no grammar that can describe the syntactical structure perfectly. It's difficult to parse spontaneous speech with syntactical-driven algorithms. On the other hand, the user's intention is ultimately embodied in the semantic information. Therefore, it's natural to adopt semantic-driven methods. Because the structures in phrases are strict and the structure is the media to carry semantic information, it's

efficient and helpful to parse the phrases syntactically.

3. The meaning of the whole sentence can be obtained by merging the meanings of the constituents that consist of the whole sentence. The meanings of these constituents are relative to each other by some latent semantic relationship. These relations are free from syntactical structure in spontaneous speech.

4. The words describing action, behavior, property and state, etc play a key role in describing the latent semantic relations. They act as the semantic connector of other concepts in sentences. It's an important cue for parsing with semantic-driven methods.

5. In a dialog, users talk as they think. The thinking process is usually a disfluent one. Thus there are lots of spontaneous phenomena in spontaneous speech. But when they finish talking, the whole meaning is clear. Therefore, the latter part of the sentence often reflects what the user really wants to express. When there are some conflict between the former part and the latter part, the meaning of the latter is correct and complete.

Considering these basic ideas, a Semantic Constituent Spotting and Assembling (SCoSA) model is proposed to parse spontaneous speech in the dialog system. The model can be defined formally as the following sextuple.

(*S, f, CS, CA, SF*)

S denotes the input sentence.

f represents the pre-processor or filler. The function is to eliminate useless words and get basic information for each word.

CS represents the Constituent Spotter. The spotter spots semantic constituents from the sentence and forms parsing sub-trees.

CA represents the Constituent Assembler. The assembler assembles the parsing sub-trees into the parsing forest for the whole sentence to get the meaning.

SF denotes semantic frames. The parsing forest is mapped to the slots of the frames. The frames are used to represent the meaning of the sentence.

3.3 Parsing Algorithm

In the SCoSA model, the semantic constituent is an important concept. It means the basic semantic element of a sentence. It's strict and regular in structure and self-complete in meaning. It may be a phrase, a word or a simple sentence. So, to understand a sentence is to find these constituents out and get the whole meaning by combining them.

The parsing process is a semantic-driven multi-phase process. It can be divided into two main phases: spotting and assembling. In the spotting phase, semantic constituents are spotted from the sentence based on structural knowledge. In the assembling phase, these constituents are assembled to form the whole semantic hierarchy of the sentence. The workflow diagram is shown as figure 2.

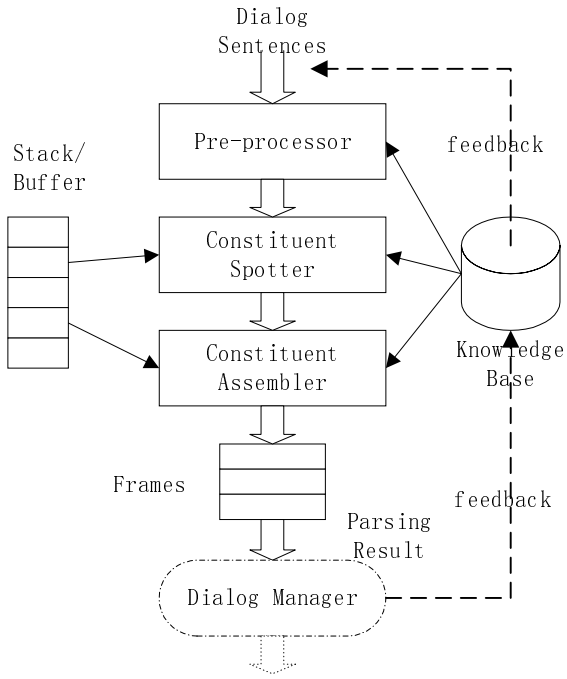

Figure 2: the workflow diagram of language parser

The workflow is explained in the following subsections.

3.3.1. Pre-processor

After the input speech is recognized, the N-best sentences are generated. In the N-best sentences, there are many inessential words as to the meaning. They destroy the sentence structure and increase the complexity of later parsing. It's necessary to eliminate them before further processing. The task for pre-processing is to filler out these inessential parts.

The lexical knowledge and some formal rules are used to find out the inessential words. In a special domain, some words are sure to be inessential to the sentence meaning, such as "一下". When these words are omitted, the meaning of the sentence

keeps the same. There are several obvious principles lying in the sentence form to identify the inessential words. For example, when two identical words appear adjacently, a repetition appears. One of them can be eliminated. By preprocessing, part of spontaneous speech phenomena can be eliminated efficiently.

For later parsing, the basic information for each word is needed. The information is defined by part of speech (POS). In the system, the vocabulary contains total 1416 words. All the words are divided into 17 classes according to syntactical information. The classes are also divided into subclasses according to semantic information. Some subclasses are divided into more detailed subclasses again. About 110 subclasses are got after final dividing. Each class or subclass has some common property. The individual information is attached to each word.

3.3.2 Constituent Spotter

After preprocessing, the sentence is simplified and the basic information for each word is formed. To understand the sentence meaning, it's necessary to find out all the semantic constituents. The constituent spotter spots the constituents from the sentence to generate parsing sub-trees.

Lexical and structural knowledge is made full use of in the spotting phase. Rules are established to describe the structural information of the semantic constituents. The rules are augmented with semantic operations attached to them. When a rule is matched in a sentence, the corresponding operation is done to form a new parsing node with compositive syntactical and semantic information. Two kinds of constituents are described, domain independent and dependent. The former includes usual utterances appearing in all kinds of dialogs, such as utterance to express users' inclination. The latter includes special phrases in travel information domain, such as phrases for place, time, price, etc.

The spotting phase can be divided into two stages: locating possible constituents and spotting them out. Keywords trigger technology is applied to locate possible constituents according to the lexical knowledge. The sequence of words is scanned from the left to the right. When some word matches some rule, there is a possible constituent and the rule is added into a rule buffer. After all possible constituents are located, a matching algorithm is applied to the buffer to find the rules which can be reduced successfully. Then, constituents are spotted out and the parsing sub-t0rees are generated by the operations attached to each rule.

3.3.3 Constituent Assembler

After the spotting phase, semantic constituents with full information are extracted from the sentence. To find out the whole meaning or the user's intention, the semantic relations of these constituents must be determined. This task is implemented by the constituent assembler. The assembler assembles the constituent to form the whole semantic hierarchy and parsing tree of the sentence.

Knowledge about the relations of the semantic constituents is extracted and refined from the application domain. Constituents are also divided into several classes according to semantic information. IF-THEN rules are used to describe the knowledge. The condition part includes positional and semantic match. The action part is to assemble semantic information. A simple searching algorithm is used to find the proper constituents and assemble them.

Because some pairs of constituents have closer semantic relations while others have looser ones, priorities are assigned to each kind of constituents to indicate the difference. Closer the relation is, higher the priority is. Constituents with higher priorities are assembled firstly. Therefore, the assembling process is a multi-stage process. In each stage, constituents with the equal priority are assembled from the right to the left. It is based on that the latter part of a sentence supplies more correct and complete information, as stated in the section 3.2.

3.4 Frame Representation

A whole parsing tree or forest is generated after the assembling phase. It's finally mapped to frames for representation. Because the rhetorical relations in dialog sentences are simple, the form of frame with three sections is chosen for semantic representation. A frame is consisted of the type section, the topic section and the sub-topic section. The frames can be

defined formally as the following.

SemanticFrame ::= SentenceType + TopicItem + $SubTopicItem_1$ {+ $SubTopicItem_n$ }

TopicItem ::= Topic + Relation + Value

Relation ::= EQUAL | NEQUAL | LARGER | SMALLER | … | ε

Value ::= Vi {op Vi }

op ::= $ | ||

$SubTopicItem_i$ ::= TopicItem

In the above representation, the type section indicates the type and mood of sentences, such as questions, answers, confirms, etc. The topic section indicates the main topic the user is talking about, such as routes, prices, etc. The sub-topic sections give the constraints for the main topic. The Relation section represents a relation between the topic and the concrete value. The op section represents logic AND and OR for logic operation, and the Vi section represents the concrete value for a certain topic.

4. PRELIMINARY EVALUATION AND CONCLUSTION

A test speech corpus with 12 dialogs is designed to evaluate the system preliminarily. The 12 dialogs contain 112 sentences and 927 words. For the validity of evaluation, when designing the corpus, all kinds of real phenomena in dialogs are included as many as possible in the acoustic, linguistic and dialog level. In the acoustic level, some spontaneous speech, such as “en”, and some words out of vocabulary are recorded in the corpus to test the garbage model and filler model. In the linguistic level, there are many kinds of spontaneous phenomena, including repetition, omission, insertion, disorder of words, etc. In the dialog level, a cooperative attitude from users and an uncooperative one are taken into consideration.

The performances of the speech recognizer, the language parser and the system response are evaluated respectfully on the above corpus. The speech data are recorded with 14.4kb/s sampling. The results are listed in the table 1 as following.

Correct Rate of Speech Recognizer	Correct Rate of Language Parser	Correct Rate of System Response
92.3%	90.7%	85.7%

Table 1: evaluation result for each stage in the system

From the test result on the language parser, we found that the parser has a good performance in parsing spontaneous speech in the dialog system. The parser is efficient to deal with most kinds of spontaneous speech in the application domain. The parsing errors in the test partly come from the recognizing errors on some key words. These key words are necessary to understand the meaning. Another source of errors comes from the restriction of this method. It’s difficult to draw all the knowledge from the application domain to describe the complex language.

In all, we put up the basic strategy to parse spontaneous speech and develop the SCoSA model for the VOTIRS 2.0 spoken dialog system. The preliminary evaluation has argued that the SCoSA model is efficient to understand spontaneous speech in the dialog system. This method is similar to the process people understand. It doesn’t depend on the high-level syntactical information, but focuses on finding out users’ intention by spotting and assembling semantic constituents. This method also has considerable advantages in flexibility and easy to implementation.

5. REFERENCE

[1]. S. Seneff, TINA: A Natural Language System for Spoken Language Applications, Computational Linguistics, 1992.

[2] W. Ward, Understanding Spontaneous Speech: The PHOENIX System, ICASSP'91, VOL 1, pp. 365-367.

[3] R. Schwartz, S. Miller, D. Stallard, J. Makhoul, Language Understanding using Hidden Understanding Models, ICSLP'96, VOL 2.

[4] James Glass, Multilingual spoken-language understanding in the MIT Voyager system, Speech communication, 1995.

[5] A. Thanopoulos, N. Fakotakis, G. Kokkinakis, “Linguistic Processor for a Spoken Dialogue System Based on Island Parsing Technoniques”, in Proceedings of EUROSPEECH97’, Patras, Sep. 22-25, 1997

[6] 殷峰，何克抗，语句级拼音-汉字转换系统的设计与实现，计算机研究与发展，第 34 卷，第 5 期，1997，5.