In this paper, we present LOADSTAR, a Mandarin Spoken Dialogue system developed for travel information retrieval, which could provide clients tourism information and plan proper itinerary according to clients’ favor. In addition to description of the system architecture and individual modules, several significant strategies such as mixed class-based and word-based language model, robust semantic-based hierarchical parsing paradigm, dynamically refreshed tracing of dialog, unified scoring schema integrating multiple information sources, and mixed-initiatives control mechanism are proposed and applied to realize efficient interaction. Evaluation of system performance on 13 dialog include 120 utterances, which contains ample spontaneous speech phenomena is reported. The response accuracy of system achieved 90.9%.

1. INTRODUCTION

With widely spread and highly increased on-line information and services, the research of constructing spoken dialogue system attract more and more researchers’ interests, it could help user attain the information efficiently through voice interaction with machine. The perceptible progress in the speech recognition technology and unbelievable computational ability of computer provide such possibilities. However, we have to face several great challenges to built such system and make it work in real world. The first of all, spontaneous speech is notoriously problematic, full of “improper” language usage like repetitions, disorder, repairs, fragments and no-sense insertion which interrupt words and grammatical constructions. Secondly, there exists plenty of phenomena such as exhalation, laughter, cough, hesitation, and non-voice utterance and no-meaning filled pause such as /ah/, /wol/, /um/ in conversational talk, which lay great burden to speech recognition. Additionally, recognition errors make later processing more problematic. Moreover, the mechanism of human dialog is not very clear and therefore any model applied to such process is short of the theoretical foundation. As the results, the dialog system developed now are restricted to specified domain, such as railway timetable inquiry [1][7], used automobile advertisements [4], automated appointment scheduling [2], directory information switchboard [5]. With the help of domain knowledge, the perplexity and ambiguity could be decreased at certain degree. Here we present our mandarin dialogue system, LODESTAR, which could provide clients tourism information about route, departure date, traffic mode, duration, cost, sights and so on. Additionally, it could help user plan the proper routine according to user’s favor. Our aim is to realize efficient and graceful human-machine interaction. To achieve such aim, we developed many strategies: mixed class-based and word-based language model, robust hierarchical parsing paradigm, mixed-initiatives control mechanism and unified scoring schema. Evaluation test for system performance on 13 dialogs shows they are feasible and practical.

The paper is organized as follows: First we introduce the corpus and database, then we give the framework of the whole system, describe each module in detail and emphasize on the method we proposed. Next we present the evaluation; finally we make a conclusion.

2. CORPUS AND DATABASE

Corpus includes two parts, raw speech data collecting and corresponding text transcription. It plays a great role in analysis of the phenomena in different levels such acoustics, language and dialogues. Technologies based on statistics methodologies almost rely heavily on the corpus, for example, the training of acoustic model and language model. Even the parsing algorithm and dialog strategy is built up according to the observation, extraction and abstraction of corpus. Evaluation and refine of prototype system also need the corpus set. Therefore, we paid much attention to data collecting. The work was taken on two modes, first we predefined some scenarios and organize persons to call agency for requesting travel information. However such mode represent the transaction between the human beings. Some researchers [7] argues a human behaves much differently when he or she is talking to a machine rather
than to another human. As the second mode, we adopted a simulated mode in which a human expert emulates the machine and the user is led to believe that she or he is actually talking to a computer. Table 1 concludes the materials we collect through these two modes.

<table>
<thead>
<tr>
<th></th>
<th>dialogs</th>
<th>utterance</th>
<th>words</th>
<th>lexicon</th>
</tr>
</thead>
<tbody>
<tr>
<td>Method 1</td>
<td>60</td>
<td>2716</td>
<td>21228</td>
<td>1381</td>
</tr>
<tr>
<td>Method 2</td>
<td>13</td>
<td>530</td>
<td>3280</td>
<td>-</td>
</tr>
</tbody>
</table>

Table1: Corpus collection

From the corpus, we learned that commonly used words appear very frequently, for example, the top ten exceed 400 times on averages, occupied 19.4% of the total number of the words, and the top 105 words occupied 67.3%. On the other hand, some words appear only once or twice. We should notice such case when training language models.

With the corpus, we build up our lexicon, whose size is 1687 words.

### 3. SYSTEM ARCHITECTURES

Figure 1 sketches architecture of LODESTAR and the relation of individual modules.

The system comprises five core modules: speech recognizer (SR), natural language understanding (NLU), dialog manager (DM), language generator (NLG), text-to-speech synthesizer (TTS), and supporting module including database, inquiry generator and history tracker. They cooperate with each other and work as follows: SR transcribes the acoustic events into word sequences, in our system the output is N-Best sentences. NLU parses the input and maps the results into semantic frames. DM accepts the semantic frame output and interprets it in the context of large dialogue structure that mainly coordinates interaction history, browses the database if necessary, and decides the proper response template through NLG. After TTS, clients could get voice feedback.

In the next part of this section, we will discuss each module in more detail.

#### 3.1 SPEECH RECOGNIZER

The framework and search strategy for speech recognition are based on a large vocabulary continuous speech recognition system we developed [8], the main characteristics are:

- **Features**: MFCC and energy and their 1-order and 2-order difference, add up to 59 dimensions;
- **Acoustic models**: 138 HMM models with mixture continuous Gaussian density function. They are based on context dependent initials and finals.

Compared with LVCSR, dialog system has several distinct features,

- Spontaneous speech is more flexible and ungrammatical. Besides, it contains many sound like /en/, /a/, /um/, which fill the pause between contents words.
- The vocabulary is often comparatively small, the impact of Out-of-Vocabulary (OOV) becomes acute.
- For a certain domain, it is very difficult to obtain sufficient data to train language model.

Considering above, we improve our recognizer in following aspects:

1. We use the LM-directed search mechanism instead of the Finite State Network (FSN) which was adopted by many systems. It completely liberates clients from the constrains of the uttering.
2. Mixed class-based and word-based LM is built up. We classify the lexicon into 733 catalogs according to their semantics, 45 contain more than one word, like the time representations, numbers, interesting sites, greetings and so on. The rest are single-word class. As the result, such mixed LM relieves data sparseness of proper noun, while keeps the predictability of common words.
3. Syllable filler are introduced to model the OOV.
4. The garbage models are established for sound like /en/.

There are two output forms from recognizer: word lattice and N-best sentences traditionally. Considering the complexity of the processing for the word lattice, we utilize N-best candidates as the output.

#### 3.2 NATURAL LANGUAGE UNDERSTANDING

Because of extreme flexibility of expressions like words in disorder, repetitions, ellipsis, and self-repair, even fragments, it is very difficult to seek solution based on any regular grammatical analysis. Recognition error makes such problems acute. Since achieving perfect parsing is very difficult, we think of picking up the meaning of the utterance as possible as could. Semantic-driven parsing will be the top principle to design our parser. According to the principle, a robust semantic-based hierarchical parsing paradigm has been proposed. We call it SCAM, a semantic constituents spotting and concept assembling model [3]. The workflow could be illustrated with figure 2.
• Filtering out the simple surface repetition and no meaning words from the input by preprocessing.
• Detecting the significant semantic constituents individually from the sentence with help of predefined patterns by semantic detector.
• Merging the small individual constituents into a bigger concept unit according to rules by concept assembler.
• Catching and generating a meaning representation of the sentence.

In the parsing process, several guidelines are obeyed.
• Concentrated on figuring out the dialog topic and their attributes;
• Directed by semantic parsing, assisted with syntactic analysis;
• Parsing process are hierarchically

After parsing, the parsing tree is mapped into a semantic frame that was predefined.

3.3 DIALOG MANAGER

As one of the most important module, dialog manager performs two main functions: interpretation and action. The former resolves the ambiguities from ellipsis, anaphora and that could not be resolved by NLU. The roles of later include defining the response templates, updating the history, and generating databases inquiry and so on. To achieve efficient and fluent transaction, we incorporate the following mechanism into our system.

Mixed Initiatives:
The first and unavoidable question is who take control of the conversation. In other word, who has the initiative, client or computer? There are two typical modes: user-initial, in this case, user asks and system answers, so-called Q-A mode, and system-initial where system prompts user to fill quasi-form in a certain routine, also called directive mode. Later improves the system robustness and simplifies the control flow while sacrificing freedom of clients’ input. The former is on the contrary. Both are not efficient enough since directive mode may need several system prompts to accomplish a simple client’s request. And Q-A mode may need rectifying their misunderstandings once and again when some errors taken place from SR and NLU. It also led to clients to plough around helplessly on availability of certain service because of system’s inactivity to supply information actively. Accordingly, we implemented a mixed-initiative strategy, which both parts could direct the dialog depending on their familiarity about the topics, and on the dialog states. Each one could possess the initiative when found himself was misconceived or thought he know more about the topic. In our system we assumed that user has the initiative in most case except that:

1. Long time waiting input from user
2. Detecting conflicts between current utterance and common beliefs (history);
3. No reasonable parsing result from input;
4. Too much records returned from database inquiry;
5. Null records returned from database inquiry;
6. Too low confidence measure from NLP.

At these cases, system dominates the transaction by asking for confirmation (case 2, case 6), more constrains (case 4), releasing constrains (case 5) or reissue some topic according to context (case 1 and case 3).

Apparantly, such mixed-initiative mode combines the advantages of Q-A and directive mode. It ensures that system could process transaction smoothly and avoid the breakdown. Both efficiency and transaction success rate have been improved remarkably.

Unified Scoring schema:

Since recognizer output is N-best sentences, there is null or more than one result after the parsing. How to choose an optimal result that can catch the intention of user is an important issue. We developed a unified scoring schema, which take account of processing results from several stages such as recognition and parsing stage. Because our current parsing algorithm are not based on probability and it is not achievable to get parsing score directly. As an alternative, we use principle of rewards and punishment according to the parsing results. As a whole, in addition to the recognition scores, other scores are weighted according to:

• The more completed of parsing , in other word, more semantic blocks contains in the semantic frames, higher reward score is given;
• According to the topic, different weight is given, for example, topic ROUTE is rewarded a comparative higher weighted score than topic COST;
• In case inconsistency was detected in the current parsing or with the context, punishment scores are given.

Dialog state stack and Topic Stack

We build two structures to modeling the interaction history. One is dialogue state stack (DSS), which records agreements that have been reached. Factualy, the agreements are intersection between clients’ requirement and availability that the database could provide. The other is topic stack (TS), which tracks the topic shift. The operation of the DSS is a dynamic process. The DSS
Organizing Response Information
To speed the process of information interaction between machine and clients, it is proper to provide some additional information beyond active topic. In our system, we accomplished such mechanism through a data structure called prompt topic selector (PTS), which was decided by the DSS and topic stack. Experiments show that it is a very efficient strategy to make both sides exchange their ideas and knowledge. Another tactics has been implemented is summarizing user’s request (according to system’s own belief) to the user whenever new constraints are spotted. Factually, this provides implicit confirmation. As result, misunderstandings could be found and corrected as soon as possible. Furthermore, it clearly states where the transaction is.

3.4 RESPONSE GENERATOR

In language generation, a template-based method is applied. Dialog manager decides which template should be chosen. The template defines the structure of the sentence and the content was filled out by dialog state stack. Such mode is easy to realize the separation of domain-dependent part from the core module. At the same time, the structural information of template may provide the information needed for high-quality speech synthesis. A disadvantage is that the responses generated are tedious. One solution is to get ready multiple templates for each case and pick randomly among them.

4. EXPERIMENTS AND EVALUATION

An extensive evaluation experiments was carried out, the test data include two sets, set one is made up of 139 single sentences and the other one made up of 13 dialogues, which contain 120 utterances, as shown in Table2. Table 3 summarizes all kinds of spoken phenomena in test set 2 by the times of appearance in the set.

<table>
<thead>
<tr>
<th></th>
<th>Utterance</th>
<th>Words</th>
<th>Word/Utterance</th>
</tr>
</thead>
<tbody>
<tr>
<td>Set 1</td>
<td>139</td>
<td>1455</td>
<td>10.5</td>
</tr>
<tr>
<td>Set 2</td>
<td>120</td>
<td>1025</td>
<td>8.5</td>
</tr>
</tbody>
</table>

Table 2: The test sets for evaluation

<table>
<thead>
<tr>
<th>Repetition</th>
<th>Ellipsis</th>
<th>Repair</th>
<th>Multi-clause</th>
<th>Disorder</th>
<th>Insertion</th>
</tr>
</thead>
<tbody>
<tr>
<td>App.</td>
<td>6</td>
<td>56</td>
<td>5</td>
<td>7</td>
<td>9</td>
</tr>
</tbody>
</table>

Table 3: Summarization of test set 2

Table 4 and table 5 is the test results of the performance SR, NLU and system.

<table>
<thead>
<tr>
<th></th>
<th>Word Accuracy</th>
</tr>
</thead>
<tbody>
<tr>
<td>Top=1,5,10</td>
<td></td>
</tr>
<tr>
<td>Set 1</td>
<td>90.38</td>
</tr>
<tr>
<td>Set 2</td>
<td>91.76</td>
</tr>
</tbody>
</table>

Table 4: Recognizer performance

<table>
<thead>
<tr>
<th>Module</th>
<th>Parser Accuracy</th>
<th>System Correct</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Original Text</td>
<td>After SR (top=1)</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Set 2</td>
<td>96.67%</td>
<td>76.3%</td>
</tr>
</tbody>
</table>

Table 5: NLU and System performance

5. CONCLUSIONS

It can be observed from the evaluation of system performances that through all the strategies we developed, satisfied results have been reached under natural spontaneous speech input. Also the graceful strategies of dialog control and responses organization we proposed improve the performance effectively under comparative high sentence error. The future work will focus on improving the portability of the system.

REFERENCE