A MULTILINGUAL SPOKEN DIALOG SYSTEM

Yunbiao XU*, Masahiro ARAKI**, Yasuhisa NIIMI**
Department of electronics & information science,
Kyoto Institute of Technology
Matsugasaki, Sakyo-ku, Kyoto, 606-8585
yunbiao@vox.dj.kit.ac.jp*, araki@dj.kit.ac.jp, niimi@dj.kit.ac.jp**
Tel: +81-75-724-7477(NIIMI) Fax: +81-75-724-7400

ABSTRACT
This paper will briefly introduce MSDSKIT-1 (Multilingual Spoken Dialogue System Version 1.0 developed by Kyoto Institute of Technology) which integrates Japanese and Chinese now. It is a promotion vision of the SDSKIT-3 (Spoken Dialogue System in Japanese). This system can provide services such as sight-seeing introduction, traffic guidance, hotel reservation. A user can also plan his itinerary under the conduction of the system. We regard a spoken dialogue system as an integrated system with a language-dependent speech interface and a language-independent dialogue controller. We must carefully consider the linguistic characteristics of the particular language for the language-dependent interface during designing a multilingual spoken dialogue system, for example, the syntactical structural features for the language parser. In order to promote SDSKIT-3 into a multilingual system (called as MSDSKIT-1), a great effort has been taken. This paper will present such effort on two aspects: (1) Chinese speech recognizer (2) Chinese language parser.

1. INTRODUCTION
With the quickly increase of computer system performance, a spoken dialogue system, a large source-consuming human-machine dialogue application, which integrates several speech technologies such as speech recognition, nature language understanding, speech synthesis technologies, becomes more and more mature. It is generally realized today that the vocal communication technology between human and computer has a huge applicable prospect in the coming 21‘th century. Especially for Chinese spoken dialogue systems, because the population of speaking Chinese is so large, the market will someday be huge, and the potential impact on related areas is almost unlimited[1] if Chinese language can be conveniently, efficiently and friendly processed by computer. So more and more researcher are working hard to develop such spoken dialogue systems to benefit world’s people. Now, many such applications have been developed in different domains for different languages[1,2,3,6], such as traffic information query (ATIS) for English, travel information accessing (VOTIRS 2.0) for Chinese, tourist information service (SDSKIT-3) for Japanese.

In the following, we will briefly introduce MSDSKIT-1 in section 2, then present the Chinese speech recognizer in section 3, the Chinese language parser in section 4, and finally, the conclusion and the future work will be presented in section 5.

2. SYSTEM OVERVIEW
MSDSKIT-1, a promotion vision of SDSKIT-3, is a mixed-initiative system, which integrates Japanese and Chinese now. The system diagram was shown in figure 1. Viewing this diagram, we can find that it can be divided into two parts: one is language-dependent which runs in the front end and another is language-independent which runs in the back ground, in which message is conveyed from one to another through case frame form. MSDSKIT-1 operates in a language such as Chinese or Japanese in the front end, while another language such as Japanese is adopted in the back ground. We call the former language as target language, the latter language as standard language.

2.1 The language-dependent part
The input-side of the language-dependent part includes a speech recognizer and a language parser, while the output-side includes a sentence generator and a speech synthesizer.
- The speech recognizer converts user’s utterances into strings of words as the input of the succeeding language analyzer.
- The language analyzer converts such string of words into the corresponding case frame form. This case frame form of user’s utterance is based on the framework of a case grammar and described by language-independent terms for which a set of words of the standard language is used in the current implementation.
- The sentence generator firstly translates a language-independent case frame form which
was described by a set of words of the standard language into the case frame form of the target language, then creates a sentence in the target language based on some templates.

- The speech synthesizer presents the preceding generated sentence to the user in synthesized speech in the target language.

The speech synthesizer presents the preceding generated sentence to the user in synthesized speech in the target language.

2.2 The language-independent part

The language-independent part is composed of a discourse analyzer, a dialogue controller, a system response generator, some databases and task-dependent knowledge bases. We represent such task-dependent knowledge bases here by a frame-system which is a set of frames to specify all the topics which might appear in dialogs on a particular task.

- The discourse analyzer extract the topic and focus, and determines the act type of the user’s utterance from the input case frame form by applying a set of predefined rules and the discourse history which was managed and updated by the dialogue controller. The references [2,4,5] described more details on the discourse analyzer and the dialogue controller.

- The dialogue controller based on a frame-driven control scheme decides system’s actions (for example, to retrieve necessary information from the database or generate a response to the user) according to the preceding discourse analysis.

The dialogue controller is driven by a task-dependent frame-system. It works as follows.

As a dialogue proceeds, the dialogue controller constructs a tree (called a dynamic topic tree hereafter) by a topic frames appearing in the dialogue, and marks a slot of a topic frame currently focused as a focus node. When it takes the initiative, it searches for an unfilled slot of the dynamic topic tree in the depth-first way from the focus node and tries to fill that slot, for example, by asking the user. If the response from the user is a direct answer, it fills the slot. On the other hand, if the response is over-informative, or instead of answering the question, the user asks a question, it tries to locate the user’s topic on the appropriate slot of the dynamic topic tree by searching the tree equidistantly from the focus node.

Thus we can control naturally both the shift of topics and the exchange of initiatives. In table 1, we can find such initiative exchange in User/4.

<table>
<thead>
<tr>
<th>Table 1: an example of initiative exchange</th>
</tr>
</thead>
<tbody>
<tr>
<td>System/1</td>
</tr>
<tr>
<td>User/1</td>
</tr>
<tr>
<td>System/2</td>
</tr>
<tr>
<td>User/2</td>
</tr>
<tr>
<td>System/3</td>
</tr>
<tr>
<td>User/3</td>
</tr>
<tr>
<td>System/4</td>
</tr>
<tr>
<td>User/4</td>
</tr>
<tr>
<td>System/5</td>
</tr>
<tr>
<td>User/5</td>
</tr>
</tbody>
</table>

After introducing MSDSKIT-1, now, we will concentrate on the language-dependent part, strictly speaking, the input-side language-dependent part.

3. CHINESE SPEECH RECOGNIZER

Mandarin Chinese is a monosyllabic and tonal language. There are about 1300 syllables in total number with 5 lexicon tones. If we ignore the tone information, the syllables will be reduced to only 408 base syllables. Each syllable is composed of one INITIAL and one FINAL in pronunciation. The total number of the INITIALs and the FINALS are 22 and
38 respectively. We use such *INITIALs* and *FINALs* as base unit to train mono-phone HMM models for the speech recognizer of MSDSKIT-1.

HTK was used to compute MFCC and delta MFCC parameters and to train HMM models. We constructed prototype HMM model with 5 left-to-right states and 5 mixture Gaussian distributions for each state. 4680 sentences uttered by 9 male speakers were adopted during training HMM models. The detail experiment conditions was shown in table 2.

**Table 2: experiment condition to train HMM models**

<table>
<thead>
<tr>
<th>Sampling frequency</th>
<th>16 KHz</th>
</tr>
</thead>
<tbody>
<tr>
<td>Encoded precision</td>
<td>16 bit</td>
</tr>
<tr>
<td>Encode method</td>
<td>Linear</td>
</tr>
<tr>
<td>Window type</td>
<td>Hamming</td>
</tr>
<tr>
<td>Window Size</td>
<td>25.6 ms</td>
</tr>
<tr>
<td>Window shift rate</td>
<td>5.0 ms</td>
</tr>
<tr>
<td>Parameter type</td>
<td>MFCC + Power + ΔMFCC + ΔPower</td>
</tr>
<tr>
<td>Vector component size</td>
<td>12 + 1 + 12 + 1 = 26</td>
</tr>
<tr>
<td>States per model</td>
<td>5</td>
</tr>
<tr>
<td>Mixtures per state</td>
<td>5</td>
</tr>
</tbody>
</table>

In MSDSKIT-1, we adopted speech recognizer-JULIAN (developed by Kyoto University) as a baseline speech input interface, in which Context-Free grammar(CFG) was used for such interface.

4. CHINESE LANGUAGE_PARSER

The language parser in MSDSKIT-1 first performs the syntactic analysis for input word strings. Next, interprets the meanings of the pronoun words, and determines the semantic concepts( called cases) of constituents and word senses of content words. Finally, the results are represented with a frame-based data structure called a case frame form which is described by a set of words of the standard language.

In other words, the semantic interpretation of an utterance is performed based on the case grammar in which the meaning of a sentence is represented by a case frame associated with a main verb of that sentence. A case frame is described by a set of slots, each indicating one of such relations between a verb and a noun phrases, like an agent, object and instrument. Noun phrases included in an utterance are assigned to some slot of the case frame based on semantic markers of the noun phrases. Thus, the semantic interpretation of an utterance is represented by a list of three terms, a main verb, case information with slots filled, modality information including the style of an utterance.

**Table 3: an example of case frame form:**

<table>
<thead>
<tr>
<th>Verb</th>
<th>介绍(introduce)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Modality</td>
<td>依赖(Request)</td>
</tr>
</tbody>
</table>

Here, we show the case frame form of the utterance “请介绍京都市的景点(Please introduce the landscapes of Kyoto City)” in table 3. In such a Chinese language parser, we used Definite Clause Grammar (DCG), which is the most popular approach to parsing in Prolog, to construct the semantic constraints. Before designing such a parser, we find Chinese language is different from Japanese in several aspects.

4.1 Different in syntax structure --- word ordering.

Both in Chinese spoken and written sentences, verbs almost always appear in the middle, but not in the tail such as in Japanese sentences. We can use following two formulations to describe the syntax structure of Chinese sentences and Japanese sentences respectively.

**Chinese:** Sentence → Subject + Verb + Object

**Japanese:** Sentence → Subject + Object + Verb

And furthermore, for the same verb word, it changes regularly in different tenses and different aspects of the sentence in Japanese. It is quite different with Chinese sentence.

4.2 Different in the delimiter boundary between adjacent syntax constituents --- the auxiliary word.

Between the adjacent syntax constituents, there is almost not any auxiliary word as the delimiter boundary in Chinese sentences. In contrast, in Japanese, the auxiliary words are essential components to construct a complete sentence as the delimiter boundary. And, the usage of such auxiliary words have some obvious rules. For example, the auxiliary word “は” is always followed with the topic word, “を” is always followed with the object. Because of the existence of the auxiliary words in Japanese sentences, it makes easier to segment and determine the proper syntax constituents. In other words, the structure constituents in the case grammar is relative fixed with the proper auxiliary word in Japanese. Such properties reduce the complexity to design language parse grammar.

4.3 The parser’s grammar rules

We adopted some enhancement items in DCG grammar rules as complementary constraints to construct Chinese sentences. Such constraints come partially from the predefined case frames which was used to specify all the case frames might appear in user’s utterances. And the rest of the constraints come from the predefined case constraints which was
used mainly to deal with preposition words in Chinese sentences.

4.3.1 Case Frames

We predefined a set of case frames to describe all possible user’s utterances. Here is the model of the case frames: \( (V, M, C) \) where

\( V \) denotes the value of verb, for example “introduce” corresponding to table 3.

\( M \) denotes the modality information, it compose one to several modality information, such as Request, Negative, Hope and etc. \( M = "Request" \) for table 3.

\( C \) denotes the case information with one or several slots, in which each slot indicates the relation between the verb \( V \) and the noun phrase filling this slot, i.e.,

\[
C = case\_id,(case\_i),...,case\_id\_n,(case\_n)
\]

\( case\_id \) is ACCusative for table 3.

\( case \) indicates the noun phrase or the macro concept’s name of the case constituent, such as “landscape” to “ACCusative” for table 3.

4.3.2 Case Constraints

In order to enable to parse some special sentence, such as

<table>
<thead>
<tr>
<th>Chinese phrase: noun/NP + preposition_word</th>
</tr>
</thead>
<tbody>
<tr>
<td>“it建造于什么年代? Which era was it build at?”</td>
</tr>
<tr>
<td>“它坐落在什么地方? Where does it locate in?”</td>
</tr>
<tr>
<td>“请把**列入行程. Please add ** into my itinerary.”</td>
</tr>
</tbody>
</table>

we find that all such sentences have a preposition word to reflect the direction of the action. We must use some additional case constraints to match such styles.

We defined two types of case constraints in MSDSKIT-1, one is for GENetive for the word “of”, another is for SPC_AT and TIM_AT, for the words “in” and “at” respectively.

\[
\begin{align*}
\text{case_def1}( & case\_id, \text{preposition\_word}, \text{noun/NP}) \\
\Rightarrow \text{Chinese phrase: noun/NP+preposition\_word}
\end{align*}
\]

\[
\begin{align*}
\text{case_def2}( & case\_id, \text{preposition\_word}, \text{noun/NP}) \\
\Rightarrow \text{Chinese phrase: preposition\_word+noun/NP}
\end{align*}
\]

4.3.3 Grammar

After defining case frames and case constraints, now we can create some DCG grammar rules with above complementary condition as follows.

\[
S( V, \{ M0,..,Mn \}, \{ C0,..,Ck \}) \Rightarrow
\]

\[
\begin{align*}
& \text{case\_constituent}( C0),\\
& \{ \text{case\_frame}( V, \{ M0,..,Mn \}, \{ C0,..,Ck \}) \},\\
& \text{Verb\_word}( V),\\
& ...\\
& \text{case\_constituent}( Ck).
\end{align*}
\]

Here, \( \text{case\_constituent}( C_) \) will connect to the case constraints in order to check whether belong to the predefined case constraints or not, and which it belongs to.

4.4 Transform of the case frame forms

In order to use the whole source of the language-independent part of MSDSKIT-1 without any modification for it, we must replace the words of the target language with the corresponding words of the standard language included in the case frame forms as the output of the language parser. The inverse translation was performed as the preprocess for the sentence generator. Such translation is simple to realize by adopting a multilingual words table.

Based on above efforts, combining with a framework system created for the succeeding discourse analyzer and the dialogue controller which can be run language-independently, the whole spoken dialogue system is implemented now by our research group.

5. CONCLUSION AND FUTURE WORK

MSDSKIT-1, Multilingual Spoken Dialogue System has been developed. Currently, it can operate in two languages, Chinese and Japanese. It mainly consists of two parts, a language-dependent speech interface and a language-independent dialogue controller. This paper was devoted to description of Chinese speech input part.

In future, We will try to improve the acoustic models and the language models for the speech recognizer, the grammar rules for the language parser, the templates for the sentence generator, and a set of rules to the language-independent part in order to increase performance of MSDSKIT-1.

REFERENCES


