A Study on Domain Recognition of Spoken Dialogue Systems

T. Isobe\textsuperscript{1,2}, S. Hayakawa\textsuperscript{1,3}, H. Murao\textsuperscript{1,4}, T. Mizutani\textsuperscript{1}, K. Takeda\textsuperscript{1}, F. Itakura\textsuperscript{1}

\textsuperscript{1}Center for Integrated Acoustic Information Research, Nagoya University, Japan
\textsuperscript{2}NTT Data Corp., \textsuperscript{3}Fujitsu Lab. Ltd., \textsuperscript{4}SANYO Electric Co., Ltd.

\texttt{isobet@nttdata.co.jp}

Abstract

In this paper, we present a multi-domain spoken dialogue system equipped with the capability of parallel computation of speech-recognition engines that are assigned to each domain. The experimental system is set up to handle three different domains (restaurant information, weather report, and news query) in an in-car usage. All of these tasks are of information retrieval nature. The domain of a particular utterance is determined based on the likelihood of each speech recognizer. In addition to the human-machine interaction, synthesized voice of the route sub-system interrupts the dialogue frequently. Experimental evaluation has yielded 95 percent recognition accuracy in selecting the task domain based on a specially designed scoring method.

1. Introduction

In spoken dialogue systems, it is easier to build a statistical language model rather than hand-written grammars, typically of the finite-state-automata (FSA) class. Hence, the use of N-gram language models has become common for developing spoken dialogue systems. However, unlike the FSA grammar, the modularity of N-gram language models is not high. For example, no systematic way to combine two or more N-gram language models of different tasks is known, especially when the size of training corpora vary. This is a serious problem in building a dialogue system that can deal with multiple topics. Simply switching language models according to a predetermined scenario is an effective strategy for the multi-topic dialogues, however, that strategy is difficult to apply to a dialogue where topic changes frequently. Question-and-answer sessions for traffic, navigational information retrieval, or driver assistance in a moving vehicle setting is the best example for this case.

In the Center for Integrated Acoustic Information Research (CIAIR) at Nagoya University, Japan, we have an on-going comprehensive study on the construction and experimental use of the dialogue system, human-machine interface model aiming at realization of information retrieval and operation support of a driver in a moving vehicle, and the behavioral characteristics of drivers [1, 2, 5]. As an important part of this comprehensive effort, we approach the multi-domain dialogue system by means of selective use of parallel decoding results (SUPDR).

Towards that end, we present in this paper a pilot system with three independent dialogue subsystems. These subsystems have their own specialized task domains, namely, “weather report”, “restaurant information” and “news query.” Since all of them are running simultaneously, only one response is found for every input utterance from three possible responses and the winner response is presented to the user.

We describe the architecture of the proposed system in Section 2. In Section 3, scoring method for topic selection is discussed with experimental findings. Section 4 summarizes the paper.

2. Multi-Domain Spoken Dialogue Systems

2.1. System Configuration

The configuration of the proposed multi-domain spoken dialogue system is shown in Figure 1. It consists of three speech-recognition engines, three task processors, a speech-synthesis engine, a dialogue controller, a response store buffer, and a route navigation event generator. A speech recognizer and a task processor are assigned each of three domains ("restaurant information", "weather report", and "news query").

Figure 1: Configuration of multi-domain spoken dialogue system

Utterances from a given user are recognized by each of these domain-specific speech-recognition engines. And then the task
processor is initiated by the hypothesis of selected domain and the system prompt is generated. In addition to the information retrieval mode, the vehicular route navigation guidance is provided by an interrupt and response procedure.

2.2. Speech Recognition Engines

When an utterance is detected from a given speaker, the speech input event, the speech recognition results for each domain, and an evaluation score (likelihood) are written on a “Black Board”. In particular, we have used “Large Vocabulary Japanese Continuous Speech Recognition Engine Julius v3.1” [3] for speech recognition. As an acoustic model in each of the three domains, however, the gender-independent PTM model [4] is employed. Finally, language models are trained and learned for each domain with sentences of the in-the-car spoken dialogue collected in CIAIR [5] and the sentences generated synthetically from the text corpora.

2.3. Speech Synthesis Engine

A speech-synthesis engine based on a waveform concatenation method is used. As an ear pleasant factor, the sex of the synthetic speech and the height of pitch are modified for each domain.

2.4. Dialogue Controller

The dialogue controller consists of a Black Board, a domain controller, and an event controller. The track record of both speech recognition results and the scores for each domain are kept in the Black Board.

The domain controller regulates the switching process between domains using scores of speech recognizers recorded in the Black Board and the duration time of the user’s non-utterance. This controller can be visualized as a four-state machine with state assignments: “restaurant information”, “weather report”, “news query” and “neutral,” respectively. Neutral state corresponds to the case of none of the above three domains. The controller functional flow diagram is shown in Figure 2.

2.5. Task Processor

Since the system has three domains, it has three domain dependent task processors. The specific processor of the domain selected by the dialogue controller makes a query for information retrieval by extracting keywords from its speech recognition result and fills them into a table slot. After getting requested information, an appropriate answer sentence is generated and sent to a Response Store Buffer.

2.6. Response Store Buffer

During a human-machine dialogue —when the navigation event occurs—the dialogue system is interrupted and a system guidance message is generated and announced repeatedly. This function is realized by buffering the response sentence temporally in the response store buffer.

3. Domain Recognition Experiments

3.1. Domain Recognition from Speech Recognition Scores

In the proposed system, domain recognition is achieved by selecting the output of speech recognizer with the highest score. The evaluation score (speech recognition score) of a hypothetical $h$ consisting of $n$ words can be expressed in terms of:

$$f(h) = AC(h) + LM(h) \cdot WLM + n \cdot PLM$$

(1)

where $AC(h)$: Logarithmic output probability of acoustic model for hypothetical $h$
\[ LM(h) \text{: Logarithmic output probability of language model for hypothetical } h \]

\[ WLM \text{: Weight of language model} \]

\[ PLM \text{: Word insertion penalty} \]

The weight of a given language model \( WLM \) is used in order to rectify the difference in contributions of the acoustic model and the language model. The word insertion penalty \( PLM \) is used in order to adjust the influence of the language model by the number of words uttered. These parameters are experimentally determined most of the time.

During the domain recognition from the evaluation scores of multiple speech-recognition engines, a comparison is made among the evaluation scores from different language models. Furthermore, it became necessary to take into consideration that the influences of the language model since they differ from domain to domain.

In our experiments, we have compared the domain recognition results between two different conditions: (i) weight of language model \( WLM \) and word insertion penalty \( PLM \) are optimized for speech recognition rate and (ii) both of these parameters are optimized for domain recognition rate. In addition, we have evaluated a score adjusting method based on the entropy of language model [6].

### 3.2. Experimental Conditions

The domain recognition is made by comparing the evaluation scores from multiple speech-recognition engines which correspond to the three specific domains: “restaurant information”, “weather report” and “news query”. 540 utterances by 12 male speakers are used for performance evaluation and the specifics of the experimental conditions are illustrated in Table 1.

<table>
<thead>
<tr>
<th>SR Engine</th>
<th>Acoustic Model</th>
<th>Language Model (tri-gram)</th>
<th>Domain Dependent Model</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Julius v3.3 (Japanese LVCSR)</td>
<td>Domain Independent Model</td>
<td>Gender Independent PTM Model (3000 states)</td>
</tr>
</tbody>
</table>

Table 1: Experimental Conditions


### 3.3. Experimental Results

#### Speech Recognition:

Speech recognition rate and the optimal weight of the language model for each domain are shown in Table 2. ACC corresponds to the word accuracy rate and the COR stands for the word correct rate. The values in parentheses represent the language model weights. The word insertion penalty is fixed to -2.0 in all cases. It is also worth noting that the weight of language model for “restaurant information” is higher than others.

#### Domain recognition:

We have considered five different methods for domain recognition as discussed below and have investigated the optimal settings for the language model weight and the insertion penalty. We have measured the domain recognition rates for optimized \( WLM \) and \( PLM \) for each method. The results are given in Table 3. The methods under study were:

(a) Use the optimal \( WLM \) for the rate of speech recognition for each domain.

(b) Re-optimize \( WLM \) of the hypothesis in method (a) for the domain recognition performance. \( AC(h) \) and \( LM(h) \) are the same in method (a).

(c) Re-optimize both \( WLM \) and \( PLM \) of the hypothesis in method (a) for the domain recognition performance. \( AC(h) \) and \( LM(h) \) are the same as in method (a).

(d) Re-optimize \( WLM \) from all hypotheses for the domain recognition performance. \( AC(h) \) and \( LM(h) \) can be different from method (a)).

(e) Modifying evaluation scores (speech recognition score) using entropies of language model.

In the method (e), the evaluation score \( f(h) \) is calculated from the expression given in Equation (2), where \( LM(h) \) is compensated by the average entropies of all of the models. Entropies of language models are given in Table 4. We like to point out that \( WLM \) is optimized for the domain recognition performance, and \( AC(h) \) and \( LM(h) \) are the same as in method (a).

\[
f(h) = AC(h) + WLM \cdot (LM(h) - m(H_i - \bar{H})) \quad (2)
\]

where

\[
\bar{H} = \frac{1}{m} \sum_{i=1}^{m} H_i \quad m: \text{number of domains}
\]

\[
H_i = -\frac{1}{3} \sum_{w_{t-2},w_{t-3},\Omega} P(w_{t} \mid w_{t-2},w_{t-3}) \log_{10} P_i(w_{t} \mid w_{t-2},w_{t-3})
\]

\( P_i() \): \( tri \)-gram probability

\( \Omega_i \): the entire set of training data for \( i \)-th domain

<table>
<thead>
<tr>
<th>Domain recognition rate</th>
<th>WLM</th>
<th>PLM</th>
</tr>
</thead>
<tbody>
<tr>
<td>RS / WT / NW</td>
<td></td>
<td></td>
</tr>
<tr>
<td>a</td>
<td>91 %</td>
<td>7.0 / 2.0 / 2.0</td>
</tr>
<tr>
<td>b</td>
<td>94 %</td>
<td>8.0 / 3.0 / 3.0</td>
</tr>
<tr>
<td>c</td>
<td>94 %</td>
<td>7.0 / 2.0 / 2.0</td>
</tr>
<tr>
<td>d</td>
<td>95 %</td>
<td>9.0 / 2.0 / 3.0</td>
</tr>
<tr>
<td>e</td>
<td>94 %</td>
<td>7.0 / 3.0 / 3.0</td>
</tr>
</tbody>
</table>

Table 3: Domain recognition rate
Table 4: Entropies of each language model

<table>
<thead>
<tr>
<th></th>
<th>$H_i$</th>
<th>$\bar{H}$</th>
<th>$H_i - \bar{H}$</th>
</tr>
</thead>
<tbody>
<tr>
<td>RS</td>
<td>1.576</td>
<td>1.236</td>
<td>0.34</td>
</tr>
<tr>
<td>WT</td>
<td>1.076</td>
<td>0.283</td>
<td>-0.79</td>
</tr>
<tr>
<td>NW</td>
<td>1.056</td>
<td>1.236</td>
<td>-0.18</td>
</tr>
</tbody>
</table>

In the case of using the optimal weight for the language model marking the best COR, the domain recognition rate was 91 percent, which was the lowest. Improvement on the domain recognition rate can be achieved by making the setup of the weight of language model and the word insertion penalty more flexible, or by re-casting the hypothesis. The score-compensating method based on the entropies of language models has resulted at a domain recognition rate of 94 percent.

3.4. Discussion

Since the acoustic model is independent of each other for domains under study, it is presumed that the evaluation score of the speech-recognition engine for each domain contains only the difference in the impact of the language model used. The improvement of the ability to recognize domain is observed by and optimizing the weight of language model and the word insertion penalty, and these parameters optimized for speech recognition are different from those optimized for the domain recognition mode.

It can be said that the weight of language model and the word insertion penalty adjusted for speech recognition process are critical for the compensation to the difference between the hypothetical evaluation score of the acoustic model and the language model. Furthermore, these parameters optimized for domain recognition are effective for compensation of the difference between language models.

The performance of the score-compensating method using entropies of language models was competitive with the result using optimal parameters.

4. Conclusions

A spoken dialogue system that can deal with multiple task domains, namely, "restaurant information", "weather report", and "news query", is introduced. In this system, domain recognition is performed comparing the recognition scores of three speech-recognition engines. From the domain recognition experiments, the necessity of controlling the language model weights and the word insertion penalties for optimal selection was clarified. The optimized values of the parameters were different from those optimized independently for individual speech recognition task. Finally, it was confirmed that the system can discriminate the task domain of the input utterance by about 95 percent accuracy.

In the future, we are going to examine the method of controlling evaluation score for taking the complexity of a language model into account.

5. Acknowledgements

This research was performed in response to assistance of a Ministry of Education and science fellowship grant COE formation basic research cost (subject number 11CE2005).

6. References