DESIGNING A MULTIMODAL DIALOGUE SYSTEM FOR INFORMATION RETRIEVAL

Sadaoki Furui and Koh'ichiro Yamaguchi

Tokyo Institute of Technology
2-12-1, Ookayama, Meguro-ku, Tokyo, 152 Japan
furui@cs.titech.ac.jp

ABSTRACT

This paper introduces a paradigm for designing multimodal dialogue systems. An example system task of the system is to retrieve particular information about different shops in the Tokyo Metropolitan area, such as their names, addresses and phone numbers. The system accepts speech and screen touching as input, and presents retrieved information on a screen display. The speech recognition part is modeled by the FSN (finite state network) consisting of keywords and fillers, both of which are implemented by the DAWG (directed acyclic word-graph) structure. The number of keywords is 306, consisting of district names and business names. The fillers accept roughly 100,000 non-keywords/phrases occuring in spontaneous speech. A variety of dialogue strategies are designed and evaluated based on an objective cost function having a set of actions and states as parameters. Expected dialogue cost is calculated for each strategy, and the best strategy is selected according to the keyword recognition accuracy.

1. INTRODUCTION

Recent progress in the field of spoken natural language understanding has expanded the scope of spoken language systems to include a number of dialogue strategies. Currently, there are no agreed-upon theoretical foundations for the design of such systems. In the present work, we therefore define a dialogue system as a system that attempts to achieve an application goal in an efficient way through a series of interactions with the user. We show that by quantifying the cost for achieving the application goal in terms of an objective function, the dialogue strategy for a given application can be optimized as a function of the keyword recognition accuracy.

We have recently proposed an efficient method for large-vocabulary continuous-speech recognition, using a compact data structure and an efficient search algorithm [1]. The data structure was modeled by DAWG (directed acyclic word-graph) [2] for reducing the search space. The search algorithm composed of forward search and traceback can obtain N-best hypotheses using the DAWG structure. This method is used in this work to build the speech recognition part of a multimodal dialogue system for information retrieval.

2. DIALOGUE SYSTEM SPECIFICATIONS

2.1 Overall Structure

An example system task of the dialogue system is the retrieval of particular information about different shops in the Tokyo Metropolitan area, such as their names, addresses and phone numbers. The system accepts speech and screen touching as input, and presents retrieved information on the screen display or by synthesized speech as shown in Fig. 1.

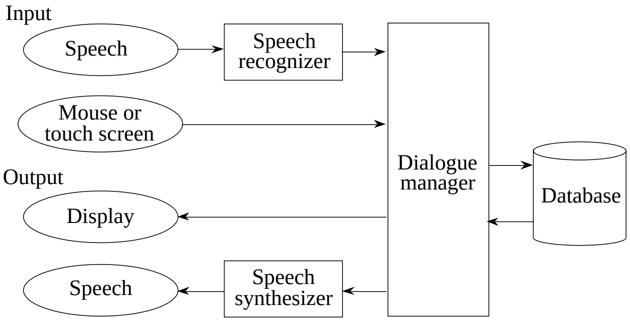


Fig. 1. Dialogue system structure

2.2 Acoustic Modeling

Task-independent triphone HMMs were constructed as acoustic models by using phonetically-balanced sentence utterances and dialogue utterances spoken by 53 male speakers. The total length of the training utterances was roughly 20 hours.

2.3 Langulage Modeling

The decoder is modeled by the FSN (finite state network) consisting of keywords and fillers, both of which are implemented by the DAWG structure. The filler networks are made to accept various non-keywords and phrases. Figure 2 outlines the basic network structure. The number of keywords is 306, consisting of district names and business names. In order to accept utterances having no district names or having repetitions of district and/or business names, the actual FSN is constructed as shown in Fig. 3.

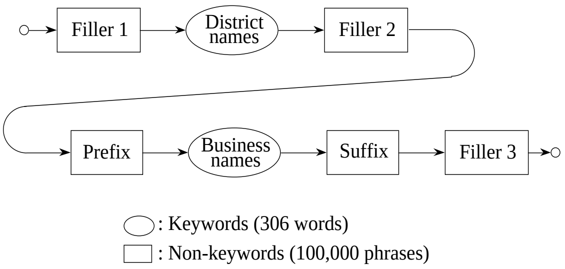


Fig. 2. Basic network for recognizing inquiry utterances

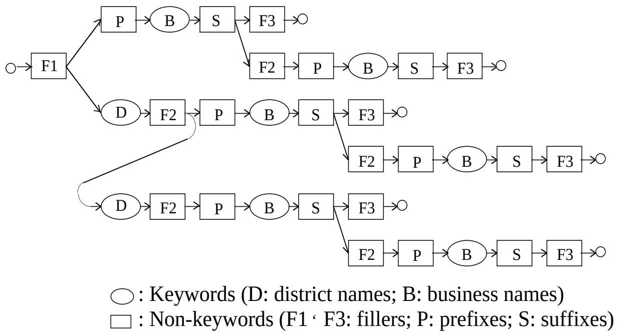

F1
P
B
S
F3
F2
D
: Keywords (D: district names; B: business names)
: Non-keywords (F1 ‘ F3: fillers; P: prefixes; S: suffixes)

Fig. 3. Actual network for recognizing inquiry utterances

Figure 4 presents an example of the DAWG structure, which represents a word/phrase set K. The DAWG structure features two advantages. First, it significantly reduces the search space compared with the tree structure. Second, words and phrases can be easily added to and removed from the graph [3]. Using the latter feature, various non-keywords and phrases observed in training utterances were easily included in fillers. The fillers accept roughly 100,000 non-keywords/phrases occuring in spontaneous speech.

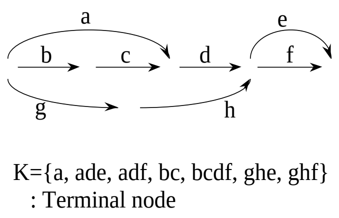


Fig. 4. DAWG (directed acyclic word-graph) structure for a word/phrase set K

3. DIALOGUE SYSTEM FORMATION

We formalize a dialogue system by describing it in terms of a state space, an action set, and a strategy. The basic idea is similar to the approach proposed by Levin et al. [4] The state of a dialogue system represents all of the knowledge the system has about internal and external resources with which it interacts. For our simple task, the state of the system includes only two entries: the district and business names, whose values can be either empty or filled through interaction with the user. The action set of the dialogue system includes all possible actions it can perform, inclusive of interactions with the user. For our task, the action set consists of 15 actions:

0. An open-ended question to the user for the value of the district and business names

For the district name:

1. Displaying the top hypothesis
2. Displaying the next hypothesis
3. Displaying the 2nd - 10th hypotheses
4. Displaying the top 10 hypotheses
5. A question to the user asking for the value of the district name
6. A request to the user to choose Yes/No
7. A request to the user to choose the correct hypothesis or "None"

For the business name:

8 - 15. Same as above

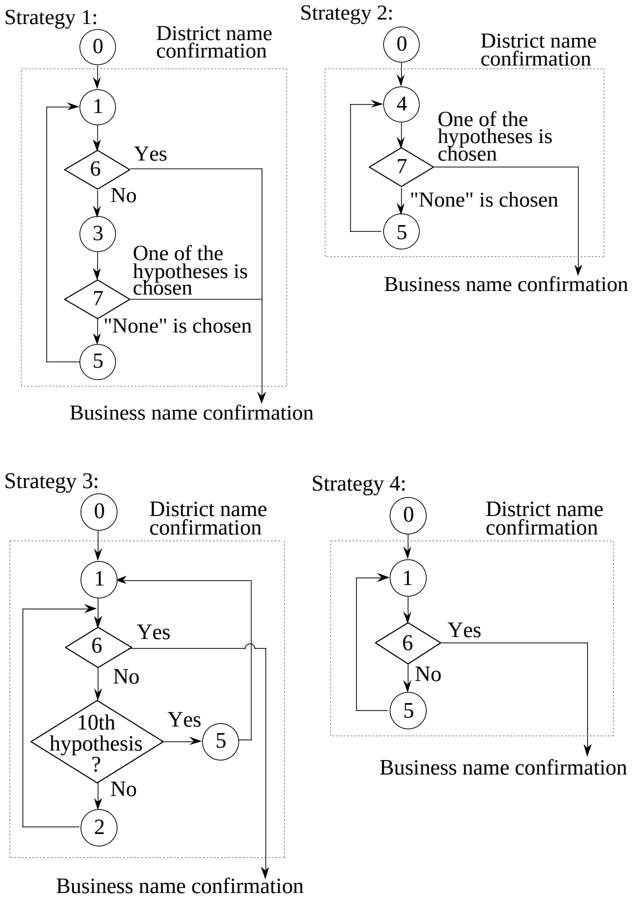
A dialogue strategy specifies a series of actions to be invoked. For our task, we propose four strategies constructed by combining various actions as presented in Fig. 5. In this work, the district and business names are always confirmed in this order following the first utterance of each speaker. For example, Strategy 1 is made up of a series of the following actions as noted in Fig. 5;

(1) An open-ended question to the user (Action 0).
(2) Displaying the top hypothesis of the district name (Action 1).
(3) A request to the user to choose Yes/No (Action 6).
(4) If the choice is "Yes" (The top hypothesis is correct.),the strategy proceeds to the business name confirmation.
(5) If it is "No" (The top hypothesis is uncorect.), the 2nd - 10th hypotheses are displayed (Action 3).
(6) A request to the user to choose the correct hypothesis or "None"

(Action 7).

(7) If the correct hypothesis is chosen, the strategy proceeds to the business name confirmation.

(8) If "None" is chosen (No hypothesis is correct.), the user is asked to say the district name (Action 5), and the strategy proceeds to (2).



Strategy 1:
0
District name confirmation
1
6
Yes
No
3
One of the hypotheses is chosen
7
"None" is chosen
5
Business name confirmation
Strategy 2:
0
District name confirmation
4
One of the hypotheses is chosen
7
"None" is chosen
5
Business name confirmation
Strategy 3:
0
District name confirmation
1
6
Yes
No
10th hypothesis ?
Yes
5
No
2
Business name confirmation
Strategy 4:
0
District name confirmation
1
6
Yes
No
5
Business name confirmation

Fig. 5. Dialogue strategies

4. DIALOGUE SYSTEM EVALUATION

We assume that the goal of a dialogue system is to achieve an application goal in an efficient way through a series of interactions with the user. We propose evaluating the system utilizing an objective cost funtion having a set of actions and states as parameters:

$$Cost = \mathrm{E}\ \{ \sum_{S} C\ (a, s\ )\} \qquad (1)$$

$C\ (a, s)$: expected cost for action a in state s

$C\ (a, s\ )$ consists of three cost terms:

W_i : hearing/reading the instructions and questions

$W_s(S_p + C_p)$: speaking and waiting for recognition results

S_p, C_p: duration of speaking and waiting

W_c <#choices>: selecting one from the choices and touching

Therefore, the expected dialogue cost for achieving an application goal is written by

$$Cost = k_i W_i + k_s W_s + k_c W_c \qquad (2)$$

where k_i, k_s and k_c are determined according to the strategy. This paradigm allows us to objectively evaluate and compare different strategies and different systems for a specific application.

5. EVALUATION EXPERIMENTS

Speech recognition experiments were performed using 104 utterances spoken by each of eight male speakers. The recognition rates are given in Table 1. The ratio of correct keywords being included in the top 10 candidates for each utterance was 95.2% on average.

Table 1. Recognition rates

	Top	Top 5	Top 10
District names	87.5%	91.3%	92.3%
Business names	80.8%	94.2%	98.1%
Average	84.2%	92.8%	95.2%

(104 utterances by 8 male speakers)

Based on this recognition performance, we evaluated the four strategies using the evaluation method described above. The expected dialogue cost calculated for each strategy is as follows:

Strategy 1:

$$(2 + p_2 + p_4)W_i + (p_3 + p_4)W_s + (4.1 + p_1 + 2.7p_2 - 2p_3)W_c \qquad (3)$$

Strategy 2:

$$(2 + p_4)W_i + (p_3 + p_4)W_s + (5.5 - 2p_3)W_c \qquad (4)$$

Strategy 3:

$$(2 + p_1 + p_2 + p_4 + \alpha)W_i + (p_3 + p_4)W_s + (4 + 2p_1 + 2p_2 - 2p_3 + \alpha)W_c \qquad (5)$$

Strategy 4:

$$(2 + p_1 + p_2)W_i + (p_1 + p_2 + p_3)W_s + (4 - 2p_3)W_c \qquad (6)$$

where p_1 / p_2 is the error rate for the top hypothesis for the district/ business names, p_3 / p_4 is the error rate for the top 10 hypotheses for the district/business names, and α is the accumulation of the probabilities that the correct value does not occur within the 2nd - 10th hypotheses (accumulated over district and business names together).

Figure 6 shows k_i, k_s and k_c in Eq.(2) as a function of the error rate p_2 for each strategy, assuming that p_1 is in proportion to p_2. $\tilde{p}_2$ indicates the present system performance. These results demonstrate that the best strategy differs according to the performance of the recognition system and the importance of individual cost weight, W_i, W_s and W_c. For example, with the present system performance, if we want to reduce the W_i-related cost, we should choose Strategy 2; if we want to reduce the W_s-related cost, we should select Strategy 1, 2 or 3; if we want to reduce the W_c-related cost, we should pick Strategy 4; and if we want to reduce the cost on average, we should apply Strategy 1.

For Strategy 1, which is generally the best, the average number of instructions and questions together is 2.2, the average number of utterances is 1.1, and the average number of screen-touchings is 2.1.

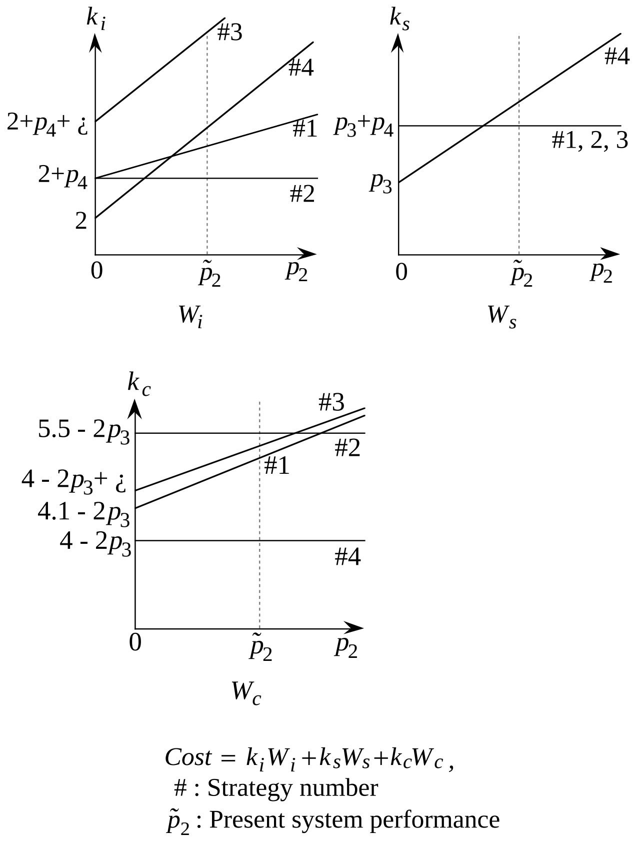


Fig. 6. Multiplier for each cost term as a function of p_2 in each strategy

From Fig. 6, it can also be observed that, if the system performance improves, Strategy 4 becomes the best choice.

6. CONCLUSION

This paper proposed a paradigm for designing multimodal dialogue systems. A variety of dialogue strategies were designed and evaluated based on an objective cost function which consists of three terms and has a set of actions and states as parameters. As an example, a system was built for retrieving particular information about different shops in the Tokyo Metropolitan area, such as their names, addresses and phone numbers. The system accepted as input a mixture of speech and screen touching, and presents the retrieved information on a screen display. The speech recognition part was modeled by the FSN consisting of keywords and fillers, both of which were implemented by the DAWG structure. Expected dialogue cost was calculated for each strategy, and the best strategy was selected according to the importance of the individual cost term and the recognition performance. The system and the approach investigated in this paper are applicable to general tasks using large-vocabulary spoken-dialogue systems.

ACKNOWLEDGMENTS

The authors wish to express their thanks to Dr. Yasuhiro Minami of NTT Human Interface Labs and Mr. Ken Hanazawa of NEC for their contribution in building the basic framework of the recognition system used in this research. The authors would also like to thank NTT Business Information Service, Inc. for providing the yellow-page database used in the evaluation experiments.

REFERENCES

[1] K. Hanazawa, Y. Minami and S. Furui: “An efficient search method for large-vocabulary continuous-speech recognition,” Proc. ICASSP97, Munich, pp. 1787-1790 (1997)

[2] E. Fredkin: “Trie memory,” Commun. ACM, 3, 9, pp. 490-550 (1960)

[3] J. Aoe, K. Morimoto and M. Hase: “An algorithm of compressing common suffixes for trie structures,” Trans. IEICE, J75-D-II, 4, pp. 770-799 (1992)

[4] E. Levin, R. Pieraccini and W. Eckert: “Learning dialogue strategies within the Markov decision process framework,” Proc. 1997 IEEE Workshop on Automatic Speech Recognition and Understanding, pp. 72-79 (1997)