FUNDAMENTAL PERFORMANCE ANALYSIS FOR SPOKEN DIALOGUE SYSTEMS BASED ON A QUANTITATIVE SIMULATION APPROACH

Bor-shen Lin, Lin-shan Lee

Department of Electrical Engineering, National Taiwan University
Taipei, Taiwan, Republic of China

ABSTRACT

Currently the performance of dialogue system is mostly measured based on the analysis of a large dialogue corpus. In this way, the dialogue performance can not be obtained before the system is on line, and the dialogue corpus should be recollected if the system is modified. Also, the effect of different factors, including system's dialogue strategy, recognition and understanding accuracy or user's response pattern, etc., on the dialogue performance can not be quantitatively identified because they can not be precisely controlled in different corpora.

In this paper, a fundamental performance analysis scheme for dialogue systems based on a quantitative simulation approach is proposed. With this scheme the fundamental performance of a dialogue system can be predicted and analyzed efficiently without having any real spoken dialogue system implemented or having any dialogue corpus actually collected. How the dialogue performance varies with respect to each factor, from recognition accuracy to dialogue strategy, can be individually identified, because all such factors can be precisely controlled in the simulation. The quality of service for the spoken dialogue system can also be flexibly defined and the design parameters easily determined. This approach is therefore very useful for the design, development and improvement of spoken dialogue systems, although the on-line system and real corpus will eventually be needed in the final evaluation and analysis of the system performance in any case.

1. INTRODUCTION

The performance analysis of recently developed spoken dialogue systems is mostly based on the analysis of a large dialogue corpus using some metrics such as the number of dialogue turns or transaction success rate [1-4]. One of the major difficulties for such scheme is that it can not predict the dialogue performance before the system is on line. The developers, after hard working in writing the dialogue control flows, can know very little about the dialogue performance until many people are involved in the on-line test. Some further problems also exist even if the system can be on line. For example, when the system dialogue strategy is modified, the performance analysis based on the old strategy is no longer valid, and all the data collection and analysis should be repeated once more. What is even worse is that the analysis for the modified system should be performed based on the new data collected with the new dialogue strategy, while the old corpus becomes useless for the analysis of new strategy. This makes the cycle for testing and modifying spoken dialogue systems relatively long. Other serious problems on the methodology of performance analysis also occur. For example, it is actually not rigorous nor reliable enough to compare the dialogue

performance for different dialogue strategies using different corpora, because many factors in the corpora, such as the user's response pattern, the slot accuracy or word error rate, can not be precisely controlled. All these difficulties and problems are intrinsic in the analysis methodology and not solved yet for such corpus-based analysis schemes.

On the other hand, dialogue performance studies based on analytical or mathematical model instead of on-line corpus analysis is always highly desired [5]. With these studies, such parameters as dialogue turns for different dialogue strategies can even be expressed in closed forms. However, in order to achieve closed forms, usually some impractical assumptions have to be made implicitly for simplification, for example on system initiative or deterministic user's response pattern [5]. It is thus difficult to extend those closed forms to other cases when some of the constraints are relaxed. In other words, such methods may not be flexible and efficient enough for the developers of spoken dialogue systems, even if with closed form solutions.

In this paper, a fundamental performance analysis scheme for dialogue systems based on a quantitative simulation approach is proposed. With this scheme the fundamental performance of a dialogue system can be predicted and analyzed efficiently without having any real spoken dialogue system implemented or having any dialogue corpus actually collected. How the dialogue performance varies with respect to each factor, from recognition accuracy to dialogue strategy, can be individually identified, because all such factors can be precisely controlled in the simulation. The quality of service for the spoken dialogue system can also be flexibly defined and the design parameters easily determined. This approach is therefore very useful for the design, development and improvement of spoken dialogue systems, although the on-line system and real corpus will eventually be needed in the final evaluation and analysis of the system performance in any case.

2. QUANTITATIVE SIMULATION APPROACH

2.1 State Representation

The finite state machine for each semantic slot, s_i , is first represented in a two-tuple expression, as shown in Figure 1. The first argument denotes the dialogue state of the slot as unknown (u), known but not yet verified (k), or verified (v), while the second argument denotes the correctness of the slot value as correct (c), error (e) or meaningless (x). When a slot is unknown, the correctness of its value is meaningless. Assuming that there are a total of n semantic slots necessary for a transaction, the overall state S can therefore be represented as n finite state machines, that is, $S = (s_1, s_2, \dots, s_n)$.

Figure 1. Finite state machine for each semantic slot

With this definition, the initial state S_i of the overall system is then $(s_1=(u,x), s_2=(u,x), \dots, s_n=(u,x))$, while the goal state S_g in which all slots are verified is $(s_1=(v,y), s_2=(v,y), \dots, s_n=(v,y))$, where the symbol y means that the slot value may be either correct or error. The purpose of the dialogue between the system and the user is therefore to make each of the finite state machine to transit from the state (u,x) to the state (v,y) as shown in Figure 1, such that the overall state S may transit from the initial state S_i to the goal state S_g . How these states transit is determined by the simulation scheme given below.

2.2 Simulation Scheme

The simulation scheme can be represented with the pseudo codes as shown in Figure 2. The cycle of a dialogue turn is divided into four segments: system prompt, user response, channel effect and system update, as shown within the for-loop in Figure 2. In the 'system prompt' segment, how the system decides which slots should be queried and which slots should be confirmed is simulated. In the 'user response' segment, how the user decides to respond to the system's prompt is simulated. The schemes simulated in these two segments are referred to as 'system's prompt strategy' and 'user's response pattern' respectively.

In the 'channel effect' segment, the slots can be considered as transmitted by the user through a channel, and the effect of recognition and understanding errors is simulated so as to decide the actually received slots. Conventionally, the speech understanding errors are measured with slot error rate, which includes the rates for inserted, deleted, and substituted slot errors, R_{ins} , R_{del} and R_{sub} respectively. The inserted slots are those slots leading to misunderstanding and therefore regarded as 'misunderstanding slots' here, while the deleted slots are those slots 'lost' in slot transmission and regarded as 'lost slots' here. On the other hand, each substituted slot is effectively equivalent to an inserted slot plus a deleted slot, or a misunderstanding slot plus a lost slot. The understanding error can therefore be represented by the following two metrics:

$$R_m = R_{ins} + R_{sub}, \quad R_l = R_{del} + R_{sub}$$

Here R_m is the slot misunderstanding rate and R_l is the slot lost rate. As a result, for each slot transmitted by the user two degradation effects may occur. One is that the transmitted slot

```

for( $N_t=0$ ; !Goal();  $N_t++$ ) {
    SystemPrompt(); // system's prompt strategy
    UserResponse(); // user's response pattern
    ChannelEffect(); // understanding performance
    SystemUpdate(); // state transition
}
if(AllSlotCorrect())  $T_s = 1$ ; // successful transaction
else  $T_s = 0$  // error transaction
 $E = n/n'$ ; // slot transmission efficiency

```

Figure 2. Pseudo codes for simulation of a dialogue

Figure 3. Simulation of channel effect for each transmitted slot

may be lost with probability R_l , and the other is that some other undesired misunderstanding slot may be generated with probability R_m . The channel effect can therefore be simulated using the random tests as shown in Figure 3.

In the 'system update' segment, how the system controls the state transition is simulated. The scheme simulated in this segment is referred to as 'system's update strategy'. For example, the system may decide that all slots being confirmed are verified (from (k,y) to (v,y)) based on the condition that a 'Yes' is detected or the condition that these slots are consistent with the previously received slots.

2.3 Characteristic Parameters of the Dialogue

For the simulation of each dialogue, after the four segments are iterated for enough number of times, the goal state is achieved and the dialogue terminated, as shown in Figure 2. Some characteristic parameters of the dialogue can be extracted. First, the number of dialogue turns, N_t , can be obtained in the for-loop in Figure 2. The transaction success flag, T_s , which denotes whether a successful transaction is achieved, can be determined by checking the second argument of the finite state machines. If all of them are correct, or equivalently the goal state equals $(s_1=(v,c), s_2=(v,c), \dots, s_n=(v,c))$, this dialogue leads to a successful transaction, otherwise an error transaction. The transmission efficiency for the semantic slots, E_s , can be obtained by dividing the total number of slots, n , by the total number of transmitted slots, n' , where n' can be observed in the 'user response' segment. The transmission efficiency E_s here indicates whether the user can transmit the slots efficiently, whose value ranges from 0 to 1. For example, if E_s is 50%, this means $n' = 2n$, or each slot is transmitted twice in average and the user may feel bored by speaking the slots again and again. All these characteristic parameters of the dialogue, N_t , T_s and E_s derived in the simulation will be further applied to the statistical analysis in the next section.

3. STATISTICAL ANALYSIS

In the analysis here the symbols T_s , N_t and E_s , are random variables for the transaction success flag, the number of dialogue turns and the slot transmission efficiency as defined perviously, while samples of these random variables are extracted for each dialogue iteration. After the simulation is performed for a large number of iterations, the mean values of these random variables, $\bar{T}_s$, $\bar{N}_t$ and $\bar{E}_s$, can be estimated. They are respectively the transaction success rate, the average dialogue turn, and the average slot efficiency, to be used to measure the dialogue performance. The variances for these random variables, $\sigma_{T_s}^2$, $\sigma_{N_t}^2$ and $\sigma_{E_s}^2$ respectively, can also be estimated for further analysis.

3.1 Quality of Service

In principle, for spoken dialogue systems higher transaction success rate, smaller average dialogue turn and higher average slot efficiency are desired. However, due to the uncertainty in speech recognition and understanding, the dialogue behavior is stochastic instead of deterministic. It is therefore appropriate to describe the 'quality of service' (QOS) for the spoken dialogue systems as the probability that a user may acquire the service above some minimum acceptable criteria. Here in this paper three metrics of QOS for spoken dialogue systems are defined as follows.

1. $R_{ts} \equiv \bar{T}_s$
2. $P_{nt} \equiv \Pr(N_i < N^*_i)$
3. $P_{es} \equiv \Pr(E_s > E^*_s)$

The first metric R_{ts} is the probability that a user may achieve successful transaction in the dialogue. The second metric P_{nt} is the probability that a user may achieve the goal within N^*_i dialogue turns. The third metric P_{es} is the probability that a user may achieve the goal with slot efficiency above E^*_s . Here N^*_i or E^*_s are some minimum acceptable criteria.

3.2 Operating Region

Because the metrics R_{ts} , P_{nt} and P_{es} are all functions of the channel effect simulated by the two parameters R_m and R_l , it is therefore easy to obtain R_{ts} , P_{nt} and P_{es} for a given (R_m, R_l) pair. If some criteria for R_{ts} , P_{nt} or P_{es} is set, say $P_{es} > P^*_{es}$, the region on (R_m, R_l) plane within which all (R_m, R_l) pairs satisfy such criteria can be directly obtained by numerical evaluation. This region for the given criteria is the operating region for the channel effect. In other words, in order for the spoken dialogue system to meet the given criteria of QOS, the channel effect has to be within this region. In this way, it is possible to determine not only whether a speech recognition front end is capable to achieve the criteria of QOS, but also to what extent this front end should be tuned or improved so as to meet the criteria. Furthermore, the criteria of QOS for a spoken dialogue system can be defined flexibly by giving any combinations of the conditions for R_{ts} , P_{nt} , P_{es} , and so on, and the corresponding operating region can be derived accordingly.

Such analysis can in fact be performed for other segments of the dialogue such as the system's prompt strategy, the user's response pattern, and the system's update strategy as well, although in those cases it may not be easy to define numerically characterizing parameters such as R_m or R_l in the case of channel effect. Furthermore, though the analysis proposed here is limited to the case of parameters T_s , N_i and E_s , it is in fact very flexible and extensible. For example, the users may be impatient about transmitting slots repeatedly or conversing for too many turns, and thus give up during the dialogue. All these events can be easily observed in simulation, and the statistical analysis for 'user's patience' can be performed as well.

4. SIMULATION RESULT

Based on the simulation scheme described in Section 2.2, the dialogue performance is a function of system's prompt strategy, user's response pattern, channel effect and system's update strategy, which can be expressed as $P_D = F(S_p, U, C, S_U)$, where P_D may be any set of metrics for dialogue performance, and S_p , U ,

C and S_U represent the simulation schemes for the four segments respectively. The scheme of the 'channel effect' segment, C , can be described by two parameters R_l and R_m , as shown in Figure 3 and discussed above. However, the schemes of the other segments are in fact difficult to parameterize. A set of simple base-line schemes is therefore defined descriptively below for illustration purpose.

$S_p : (AQ, AC)$

All slots in unknown state are queried (AQ), while all slots in known state are confirmed (AC).

$U : (AQR, YC, NT)$

All queried slots are replied (AQR). For those slots to be confirmed, a 'Yes' is transmitted if correct for all (YC), otherwise incorrect slots are retransmitted with 'No'. Those slots neither queried nor confirmed are not transmitted (NT).

$C : (R_m, R_l)$

The transmitted slots are passed through a random test with parameters R_m and R_l to determine the finally received slots.

$S_U : (QUK, VSC, SI)$

Queried slots in unknown state will enter known state if they are received (QUK). Those slots in known state being confirmed will enter the verified state if they are received consistently, or the slots are verified by slot consistency (VSC). Those slots neither queried nor confirmed are ignored completely, which means system initiative (SI).

For the base-line schemes assumed above how the average dialogue turn increases with the degradation of understanding performance is shown in Figure 4, where the number of simulated dialogues for each (R_m, R_l) pair is 100,000 and the number of semantic slots, n , is 5.

As we can see above, many factors in the four segments may affect the dialogue performance simultaneously. For the base-line schemes, the first factor of user's response is AQR, which means all queried slots are replied. Two other schemes for this factor were simulated. One is OQR, or only one of the queried slots is replied. The other is PQR, or only a part of the queried slots are replied, in which the number of the replied slots is determined by a random test. During the simulation, all other factors are the same as the previous base-line simulation. Figure 5 shows the average dialogue turn for different user's response patterns. It can be seen in Figure 5 that, for the given system's strategy, different behavior of the user may lead to quite different performance, approximately ranging from the curve for AQR to the curve for OQR. This verifies that it is not necessarily reliable for conventional corpus-based analysis to compare different system strategies using different corpora, because the user's response is an important factor but can not be controlled.

Figure 4. Average dialogue turn for base-line schemes ($R_l = 0.1, 0.2, 0.3$)

Figure 5. Average dialogue turn for different user's response patterns ($R_l = 0.1$)

On the other hand, the second factor of system's update strategy is also of interest. In the base-line simulation, the strategy 'verified by slot consistency' (VSC) is used. Another possible strategy 'verified by yes detection' (VYD), i.e., the verification is completed if a 'Yes' is detected, elsewhere the slot consistency is checked. In some languages such as Mandarin Chinese, the words meaning 'No' may be easily confused with those meaning 'Yes', which may lead to incorrect verification and therefore incorrect transaction, while in other languages such problem doesn't exist at all. It is therefore interesting to see how the rate for such misrecognition of taking 'No' as 'Yes' (R_{ny}) affects the transaction success rate, which is shown in Figure 6. According to Figure 6, the choice of the strategy VSC or VYD depends on the value of R_{ny} . If 'No' is easily confused with 'Yes' (e.g. $R_{ny} \geq 0.18$), it is proper to choose VSC. Otherwise, the best strategy should be determined based on the operating point of R_m and R_{ny} . In other words, for different systems or different languages with different values of R_{ny} , the best choice of system's confirmation strategy can in fact be decided quantitatively based on the simulation. Due to the limited space of this paper, not all simulations for all possible factors can be presented here. However, as can be seen, such simulation approach is a very useful basis for further research on these factors in dialogue systems.

Finally, the operating regions for the criteria R_{ts} , P_{nt} , $P_{es} > 0.9$ and R_{ts} , P_{nt} , $P_{es} > 0.99$ with $N^* = 5$ and $E_s^* = 0.5$ for the base-line schemes are shown as the left-bottom area of the respective curves in Figure 7. Therefore, in order to maintain the desired

Figure 6. Transaction success rate for different system's update strategies ($R_{ny} = 0.1, 0.14, 0.18, R_l = 0.2$)

Figure 7. Operating Regions for different criteria and the operating curve of speech understanding front end ($N^* = 5$ and $E_s^* = 0.5$)

QOS, the speech understanding front end must be able to operate within the operating region. For example, for a speech understanding front end in Mandarin Chinese with a task of train ticket reservation based on key-phrase spotting and a hierarchical tag-graph search [6], the operating curve can be derived by updating the spotting rate, also shown in Figure 7. It is therefore possible for this front end to be tuned such that it can operate within the desired criteria of R_{ts} , P_{nt} , $P_{es} > 0.9$. However, it is actually impossible for this front end to satisfy the criteria of R_{ts} , P_{nt} , $P_{es} > 0.99$ in any case by simply tuning the spotting rate. Such analysis gives a very good direction for designing spoken dialogue systems based on QOS criteria.

5. CONCLUSION

We have presented in this paper an analysis scheme for dialogue system performance based on simulation. Such a scheme is very efficient without using dialogue corpora. How dialogue performance varies with respect to different factors and parameters can be analyzed individually because all the factors can be precisely controlled in the simulation, while the quality of service for spoken dialogue systems can be well observed. It can be a powerful tool for developing spoken dialogue systems, and an important basis for further research on spoken dialogue systems.

6. REFERENCES

- [1] Walker Marilyn A., etc., "PARADISE: A Framework for Evaluating Spoken Dialogue Agents", *ICACL*, 1997.
- [2] Aust Harald and Ney Hermann, "Evaluating Dialog Systems Used in the Real World", *Proc. ICASSP*, 1998.
- [3] Hanrieder Gerhard, Heisterkamp Paul and Brey Thomas, "Fly with The EAGLES: Evaluation of the "ACCeSS" Spoken Language Dialogue System", *Proc. ICSLP*, 1998.
- [4] Devillers L. and Bonneau-Maynard H., "Evaluation of Dialog Strategies for a Tourist Information Retrieval System", *Proc. ICSLP*, 1998.
- [5] Niimi Yasuhisa and Nishimoto Takuya, "Mathematical Analysis of Dialogue Control Strategies", *Proc. EUROSPEECH*, 1999.
- [6] Lin Bor-shen, etc., "Hierarchical Tag-Graph Search for Spontaneous Speech Understanding in Spoken Dialogue Systems", *Proc. ICSLP*, 1998.