

DOES CONFIDENCE ANNOTATION MEET THE DIALOG GOAL?: A QUANTITATIVE ANALYSIS

Kadri Hacioglu and Wayne Ward

Center for Spoken Language Research
University of Colorado at Boulder
E-Mail: {hacioglu,whw}@cslr.colorado.edu

ABSTRACT

In this paper, we quantify the impact of using confidence annotation on the performance of a dialog system for a given dialog strategy. Although the ultimate goal in a dialog system is to optimize the user satisfaction, it is very difficult to quantify it. Instead, we consider the probability of success and the number of expected turns of a transaction as our dialog performance metrics. In fact, they are fairly strong predictors of user satisfaction. The desired dialog goal in terms of these metrics is mapped onto an operating region on the receiver operating characteristics (ROC) plane of the confidence annotation method. In doing so, one can not only figure out whether the confidence annotation would meet the desired goal at the dialog level but also understand how to tune the system in order to reach that goal. Such an analysis therefore allows the system designer to optimize the use of confidence annotation and tune the system without doing any time consuming online experiments.

1. INTRODUCTION

Confidence annotation, which would be unnecessary if we had perfect speech recognition/understanding performance, has become an important component in spoken dialog systems (SDSs). It allows a SDS to avoid annoying verification turns in cases where the confidence of information items is high. Similarly, it prevents a SDS from acting on an information item with low confidence. Several metrics have been used to assess the performance of confidence annotation. Almost all are related to its correct classification or detection ability. However, its actual impact on the dialog system performance (e.g. user satisfaction, successful completion rate, average number of turns etc.) is not clear without live user experiments. This is a very long process and needs to be repeated if the confidence annotation scheme has changed. So, an analytical analysis that can relate confidence annotation to dialog performance is highly desirable.

The work is supported by DARPA through SPAWAR under grant #N66001-00-2-8906.

This research has been inspired by the work reported in [1]. A finite state model was introduced for information items. Simulation-based statistical analysis was carried out, and speech understanding front-end performance was related to several dialog performance metrics.

In this paper, we extend the work in [1] to a dialog strategy that makes use of confidence annotation. In addition to speech understanding performance metrics, we relate dialog performance metrics to those of confidence annotation. We do not use any simulations to determine the finite state model transition probabilities. Our dialog strategy and assumptions on users' response patterns are simple enough to allow explicit derivation of the transition probabilities in terms of commonly used metrics to assess the speech understanding and confidence annotation components. We finally map a dialog goal, in terms of the probability of success and average number of turns, to a set of operating points on the ROC plane of the confidence annotation scheme, and compare it to the possible operating points of the confidence annotation.

The paper is organized as follows. In section 2, we introduce several finite state models for an information item considering confidence annotation and different user behaviors. A method for statistical analysis is introduced in section 3. Section 4 presents several numerical results, and conclusions are made in the final section.

2. FINITE STATE MODELS

As in [1], we assume a finite state model for an information item. An information item might be known (specified by the user) or unknown to a SDS. A known item can further be classified as "not accepted yet" and "accepted". Another classification is made regarding the value of the information item. The value can either be "correct" or "incorrect". Based on these classifications an information item can be in one of the following five states:

1. unknown; (u, ϕ)

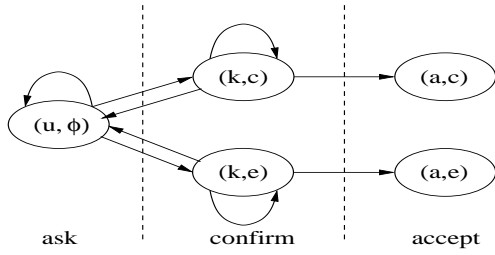


Fig. 1. Baseline finite state model (FSM0)

2. known, not confirmed, incorrect; (k,e)
3. known, not confirmed, correct; (k,c)
4. accepted, correct; (a,c)
5. accepted, incorrect; (a,e)

The possible transitions, and their probabilities, among the states is determined by the user behavior, speech understanding front-end and dialog strategy. The simplest dialog strategy, which will serve as the baseline system in this paper, asks for an information item and accepts it after positive verification. Otherwise, the information item is asked again. The user is assumed to be very cooperative and replies "yes" or "no" to any confirmation. The baseline finite state model is shown in Figure 1. Self transitions correspond to "no reply", which is a deletion error at the front-end since the user is assumed to cooperatively reply to any query.

A dialog strategy with confidence annotation may work as follows:

- If $C(I|\text{ask})$ is LOW, confirm I .
- If $C(I|\text{ask})$ is HIGH, accept I .

and

- If $C(I|\text{confirm})$ is LOW, assume "no reply".
- If $C(I|\text{confirm})$ is HIGH, accept I .

where I is the information item and $C(\cdot|\cdot)$ is the confidence score of an information item conditioned on a dialog action. The confidence score is said to be LOW if it is below a certain threshold, say θ_T . Otherwise, it is HIGH. The corresponding finite state model is exhibited in Figure 2.

Another dialog strategy might be the following:

- If $C(I|\text{ask})$ is LOW, ask again.
- If $C(I|\text{ask})$ is MEDIUM, confirm I .
- If $C(I|\text{ask})$ is HIGH, accept I .

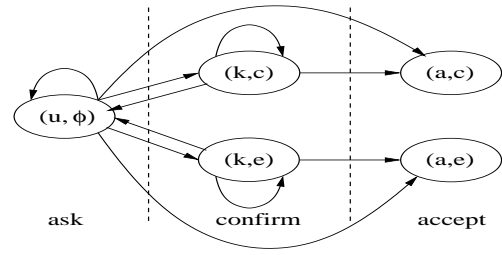


Fig. 2. Finite state model considering confidence annotation (FSM1)

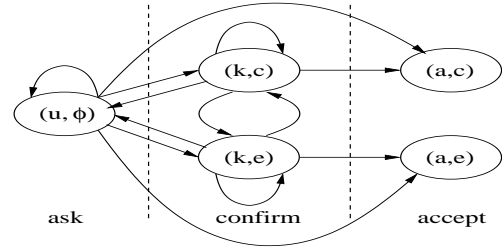


Fig. 3. Finite state model considering confidence annotation and different user behavior (FSM2)

The confidence score is said to be MEDIUM when

$$\theta_1 < C(I|action) < \theta_2$$

This strategy has the same finite state model as in Figure 2, but the transition probabilities are different.

A simple change in user behavior can result in a different model. For instance, a user might also provide the value of an information in addition to replying "no" to a confirmation. The respective model is depicted in Figure 3. Note the transitions between (k, c) and (k, e) .

3. STATISTICAL ANALYSIS METHOD

We base our analysis on the properties of Markov chains [2]. The sequence of states of a transaction can be viewed as a Markov chain. It starts at (u, ϕ) and ends at either (a, c) or (a, e) . The former state is called the initial state and the latter states are called absorption states. The transaction is considered successful if the sequence ends at (a, c) . The transaction is assumed to end when the sequence hits one of the absorption states. For the sake of analysis, self loops (not shown) with probability one are assumed for each absorption state.

We are interested in the following two metrics:

- $P_s = Pr\{state = (a, c)\}$
- $N_t = \sum_n nPr\{state \in A\}$

where P_s is the probability of success, N_t is the average number of turns and A is the set of absorption states. In order to calculate the required metrics it is sufficient to know the initial probability distribution of the states, π_0 , and the transition probability matrix, T .

The probability distribution of the states at time n is given by

$$\pi_n = \pi_0 T^n \quad (1)$$

Those probabilities at equilibrium satisfy

$$\pi = \pi T \quad (2)$$

The equilibrium probability of (a, c) gives P_s .

In the baseline model, the transition probabilities depend on the performance of the speech understanding front-end. We quantify the performance of the front-end in terms of the probability of correct understanding, P_c , and the probability of deletion, P_d . Given those probabilities, the transition matrix can be written as

$$T = \begin{bmatrix} P_d & P_e & P_c & 0 & 0 \\ P_c & P_d & 0 & 0 & P_e \\ P_e & 0 & P_d & P_c & 0 \\ 0 & 0 & 0 & 1.0 & 0 \\ 0 & 0 & 0 & 0 & 1.0 \end{bmatrix} \quad (3)$$

where

$$P_e = 1 - P_c - P_d \quad (4)$$

We need to quantify the performance of the confidence annotation for the system that uses it. We define P_{CR} and P_{FR} as the correct and false rejection probabilities, respectively. Similarly, one can define $P_{CA} = 1 - P_{FR}$ as the probability of correct acceptance, and $P_{FA} = 1 - P_{CR}$ as the probability of false acceptance. It is not uncommon to report the performance of confidence annotation as the plot of P_{CR} with respect to P_{FR} . This plot is known as the receiver operating characteristics (ROC). Each point on the plot corresponds to a different threshold setting. With confidence annotation as used in the model shown in Figure 2, the transition probability matrix becomes

$$T = \begin{bmatrix} P_d & P_e P_{CR} & P_c & P_c P_{CA} & P_e P_{FA} \\ P_c P_{CA} & P_d & 0 & 0 & P_e P_{FA} \\ P_e P_{FA} & 0 & P_d & P_c P_{CA} & 0 \\ 0 & 0 & 0 & 1.0 & 0 \\ 0 & 0 & 0 & 0 & 1.0 \end{bmatrix} \quad (5)$$

where

$$P_d = P_d + P_e P_{CR} + P_c P_{FR} \quad (6)$$

The initial probability distribution of the states is obvious; the probability of the initial state is one and probabilities of all other states are zero.

Finally, we define the goal G of a dialog system as follows:

$$G = \{P_s > P_s^*, N_t < N_t^*\} \quad (7)$$

where P_s^* is the minimum and N_t^* is the maximum acceptable values of the respective metrics. We are interested in the region on the plane (P_{CR}, P_{FR}) over which G is satisfied. This region can easily be obtained by numerical evaluation using equations given above.

The models assume a single information item. However, a transaction might need multiple information items. Assuming that

- SDS needs n_i information items to complete a transaction
- Information items are asked one at a time and are independent
- Speech understanding and confidence annotation performances are the same for all pairs of information items and actions

we can easily extend the analysis to multiple information items as

$$\begin{aligned} P_s &= (Pr\{state = (a, c)\})^{n_i} \\ N_t &= n_i \sum_n n Pr\{state \in A\} \end{aligned} \quad (8)$$

To be more realistic one can relax the assumptions to obtain a relatively complex FSM. For instance, one can allow the SDS to simultaneously query multiple information items.

4. NUMERICAL RESULTS

We first consider the baseline system (FSM0) shown in Figure 1. Assuming typical values of 0.88 and 0.04 for P_c and P_d , respectively, and $n_i = 5$, we numerically evaluate $N_t = 12.3$ and $P_s = 0.96$. We compare this result to the result of FSM1. We consider a typical setting of the confidence threshold that leads to $P_{CR} = 0.63$ at $P_{FR} = 0.05$. The respective results are $N_t = 6.1$ and $P_s = 0.84$. There is a significant drop in the number of average turns at the expense of the probability of success.

Figure 4 shows the dependency of the dialog metrics on P_{FR} for different values of P_{CR} . We conclude that the probability of success is very sensitive to P_{CR} . However, the average number of turns changes very slowly with P_{CR} but increases rapidly with P_{FR} .

As we mentioned earlier it is very common to use a ROC to assess a confidence annotation method. The ROC of the confidence annotation method that we have recently developed [3] within the context of CU Communicator [4] is shown in Figure 5. We wonder how good the annotation

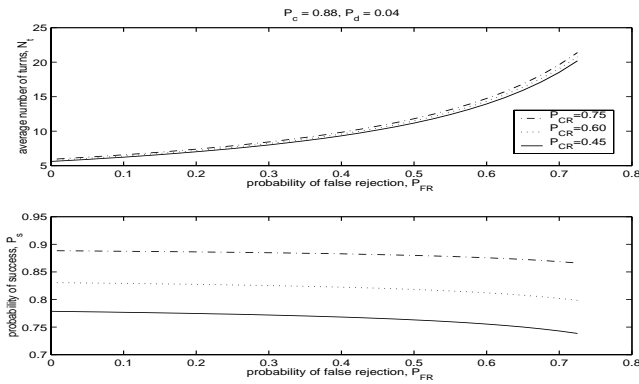


Fig. 4. The dependency of dialog metrics to confidence annotation metrics

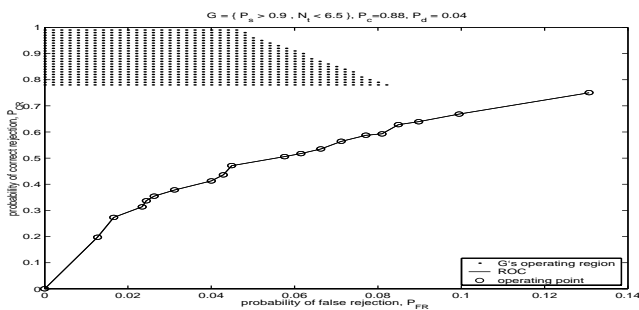


Fig. 5. The ROC of the confidence annotation and the operating region of the dialog goal $G = \{P_s > 0.9, N_t < 6.5\}$. The front-end performance is ($P_c = 0.88, P_d = 0.04$).

scheme is as compared to a given dialog goal G . The set of points on (P_{FR}, P_{CR}) plane that satisfy the goal constitutes the so called operating region. We say that the goal is achieved if there is at least one operating point of the ROC that falls within that region. The operating region of $G = \{P_s > 0.9, N_t < 6.5\}$ is also shown in Figure 5. It clearly shows that our confidence annotation scheme does not meet the goal and must be improved if we are not willing to lower the goal. Another possibility is to improve the performance of the speech understanding front-end. The impact of lowering the goal and improving speech understanding are shown in Figure 6 and 7, respectively.

5. CONCLUSIONS

We have presented a quantitative method that can be used to relate confidence annotation metrics to dialog metrics. This method avoids time consuming online experimentation and allows the SDS designer to optimize the use of confidence annotation by either selecting an appropriate dialog strategy

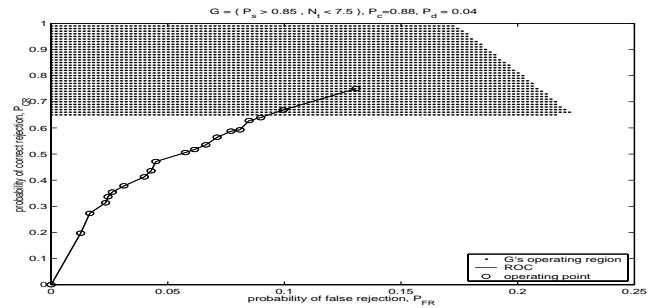


Fig. 6. The ROC of the confidence annotation and the operating region of the dialog goal $G = \{P_s > 0.85, N_t < 7.5\}$. The front-end performance is ($P_c = 0.88, P_d = 0.04$).

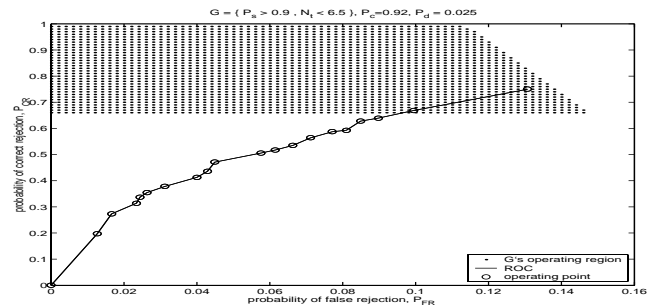


Fig. 7. The ROC of the confidence annotation and the operating region of the dialog goal $G = \{P_s > 0.9, N_t < 6.5\}$. The front-end performance is ($P_c = 0.92, P_d = 0.025$).

or tuning the speech understanding front-end. The framework is so general that it can be extended to account for different user behaviors and more complex dialog strategies. Our research is continuing in those directions.

6. REFERENCES

- [1] B. Lin and L. Lee, "Computer-aided analysis and design for spoken dialogue systems based on quantitative simulations," *IEEE Transactions on Speech and Audio Processing*, vol. 9, no. 5, pp. 534–548, July 2001.
- [2] James R. Norris, *Markov Chains*, Cambridge University Press, 1998.
- [3] K. Hacioglu and W. Ward, "A concept graph based confidence measure," to appear in *International Conference on Acoustics, Speech, and Signal Processing*, Orlando, Florida, May 2002.
- [4] W. Ward and B. Pellom, "The CU communicator system," in *IEEE Workshop on Automatic Speech Recognition and Understanding*, Keystone, Colorado, 1999.