On the utility of smoothing in spoken dialogue systems

Csaba Szepesvári*

Mindmaker Ltd.
Budapest 1121
Hungary
Konkoly Thege M. u. 29-33
szepes@mindmaker.hu

Abstract. In this article we study a stochastic model for spoken dialogue systems (SDSs). We advocate filtering methods, and in particular smoothing to be used in SDSs and argue that these methods let one design (or learn) dialogue control strategies whereas the number of unnatural (and unnecessary) clarification requests can be reduced, resulting in more natural dialogues.

1 Introduction

Spoken Dialogue Systems (SDS) must rely on the output of imperfect speech recognition systems (SRS). Most SDS use only the "best guess" of the SRS to plan the next turn, which is typically a clarification subdialogue if the score of the last best guess was not high enough.

Unfortunately, even these days the accuracy and reliability of SRSs, although both measures experienced a dramatical increase in recent years, is not high enough to be used in SDS in this very simple way: the resulting dialogues become fragmented and unnatural, leading to user frustration.

Therefore, it is an important open research problem to design and build SDSs that engage into a human-like, natural dialogue.

In this paper we study a stochastic model for dialogues and suggest the use of filtering/smoothing methods to help one to design more natural dialogues whereas the number of clarification requests is reduced without effecting the ratio of successfully completed tasks (the completion ratio).

In this approach the output of the speech recognition system (or more generally that of the spoken language understanding system) is not trimmed to the best guess, but we will work with probability distributions over the set of possible utterences. Another feature of our approach is that we propose to use the available task-related background knowledge that constraints the relationship of the possible interpretations of successive utterences. This will be shown to be very effective for combining successive utterences of the user into a single coherent hypothesis of the user's intentions. As a result, one expects the probability of successful task completion to increase, and simultaneously, the average number of interactions per task to decrease. A simplified version of the method outlined here has been in use for years with success in the commercial large-vocabulary

* This research has been supported by IKTA-3.

speech recognition system of Tisento (a subsidiary of Mindmaker). Trivially, the methods outlined here can be used in text-based or multi-model dialogue systems, as well.

Related work includes the work of Paek and Horvitz who use dynamic Bayesian networks (thus, a Markov model with structured state space and factored probabilities) and propose a dialogue strategy where the decision whether the agent should enter a grounding (actually clarification) subdialogue is computed using an expected utility framework [5]. Earlier work on modelling dialogues as Markov processes, and in particular as Markovian Decision Processes include eg. [7, 3, 8].

2 A Probabilistic Model for Spoken Dialogues

Let us call the agent participating in the dialogue Alice (A), and the user Bob (B). Assume that Alice and Bob exchange information in a dialogue with messages $Y_t^{(A)}$ and $Y_t^{(B)}$. Let us represent the mental state of the Alice and Bob at time t by the respective symbols $M_t^{(A)}$ and $M_t^{(B)}$. At this point of generality, we do not need to care about what the form of the mental state of Alice and Bob exactly are, the only restriction we make on these "mental states" is that the process

$$(M_t^{(A)}, M_t^{(B)}, Y_t^{(A)}, Y_t^{(B)})$$

is assumed to be the subject of a Markov process of the form:

$$\begin{aligned}
M_{t+1}^{(A)} &= \Pr\left(\cdot | M_t^{(A)}, Y_t^{(B)}\right), \\
Y_{t+1}^{(A)} &= \Pr\left(\cdot | M_{t+1}^{(A)}\right), \\
M_{t+1}^{(B)} &= \Pr\left(\cdot | M_t^{(B)}, Y_{t+1}^{(A)}\right), \\
Y_{t+1}^{(B)} &= \Pr\left(\cdot | M_{t+1}^{(B)}\right).
\end{aligned} \tag{1}$$

In other words, we assume that the order of events is:

1. Alice observes message $Y_t^{(B)}$ from Bob, and updates its mental state, resulting in $M_{t+1}^{(A)}$;
2. Alice computes a message and sends it to Bob;
3. Bob observes the message $Y_{t+1}^{(A)}$, and updates its mental state, resulting in $M_{t+1}^{(B)}$;
4. Bob computes a message and sends it to Alice.

The model is illustrated in Figure 1. The above cycle is called a complete exchange. After a complete exchange, another exchange starts and this is repeated until the end of the dialogue. The probabilities above may encompass models of the time evoluation of dialogue states (eg. state of grounding, initiative), language models, models of the speech recognition system (for an example of this see Section 3), speech and acoustic models, or knowledge about the domain and the task. Also the mental state may include a history or even a representation of the other agent's mental state.

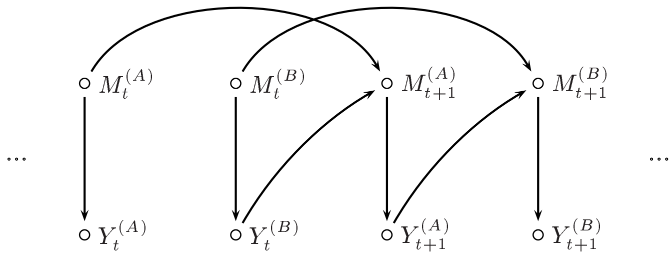

Fig. 1. A probabilistic model of dialogue

Models similar to (1) have already been proposed and used in the literature (see eg. [8]). However, in the literature, the model is typically used to estimate the most likely mental state of Bob (from Alice's perspective). Here, we propose to use the model to estimate either the full posterior:

$$\Pr\left(M_t^{(B)}, \dots, M_0^{(A)} | H_t^{(A)}\right) \tag{2}$$

or the probability of a single past mental state of the past:

$$\Pr\left(M_i^{(B)} | H_t^{(A)}\right), \quad 0 \le i \le 1. \tag{3}$$

Here $H_t^{(A)}$ is the history of the dialogue up to time t observed by Alice:

$$H_t^{(A)} = (M_t^{(A)}, Y_t^{(A)}, Y_t^{(B)}, \dots, M_0^{(A)}, Y_0^{(A)}, Y_0^{(B)}). \tag{4}$$

Estimating the probability of an earlier mental state (say at time i) can be advantegous in many ways: If the messages exchanged by Alice and Bob following stage i convey information about the mental state of Bob at time i then the entropy of the posterior $\Pr\left(M_i^{(B)} | H_t^{(A)}\right)$ can be substantially smaller than the entropy of the posterior $\Pr\left(M_i^{(B)} | H_i^{(A)}\right)$. This also allows Alice to revisit the interpretations of earlier exchanges and to build a substantially better understanding of the dialogue's state. In addition to this, by a clever organization of the dialogue Alice may delay or completely eliminate clarification subdialogues (which would otherwise be necessary when the entropy of $\Pr\left(M_i^{(B)} | H_i^{(A)}\right)$ is large) in the hope of being able to get a better estimate of the most likely mental state of Bob at time i after making some new observations. In the next section, we give an explicit example when Alice can follow such a strategy when task knowledge is used for disambiguating the meaning of past utterences.

In general, the probabilies above would be reduced to some model probabilities using Bayes' theorem and eg. sequential algorithms could be used to obtain an approximation of the respective posterior distributions (eg. particle filtering, see [1, 4]).

3 Example: Information Seeking Dialogues

Assume that Alice's goal is to obtain the values of n variables, $x_1, \ldots, x_n$ taking on values in the spaces $\mathbb{X}_1, \ldots, \mathbb{X}_n$ by means of a dialogue with Bob using some spoken language understanding (SLU) system. In general, SLU systems work on the output of SRSs, which typically produce a series of hypothesis-score pairs of the last utterance of Bob. The scores typically represent the log-likelihood of the associated utterence. Therefore, the output of SRSs can be modelled as a (random) probability distribution over the set of possible utterences. Actually, the output of SRSs is an approximation of the *a posteriori* probability distribution conditioned on the observed speech. Here, we assume that this approach is followed by the SLU system, i.e., the output of the SLU system approximates the *a posteriori* probability distribution of the speech-acts of Bob given the observed speech.

In the scenario considered here we restrict the set of possible speech-acts of Bob to those of the form $Inform(i, x_i)$, i.e., to speech-acts when Bob informs Alice that the value of the ith variable is x_i. Further, we assume that Alice blindly performs the speech acts $Ask(i), i = 1, 2, \ldots, n$ one after another, no matter what response it receives from Bob. This example is used for showing the advantages of our approach in the simplest possible set-up. Extensions that include feedback will be considered later.

We consider the following problem: infer the most likely value of

$$x = (x_1, \ldots, x_n)$$

on the basis of Bob's utterences observed by Alice. We will take a Bayesian perspective and compute the posterior (or rather an approximation of it) over the set of possible values of x.

Let us thus assume that we are given an *a priori* probability distribution p over $\mathbb{X}_1 \times \ldots \times \mathbb{X}_n$. This probability distribution can be obtained (approximately) by some learning mechanism. In another approach, p will model that we have no a priori preferences for values of x, but we treat every single value of x as being equally probable. Often, this will combined with some background knowledge about the set of values of x that are admissible. In this case

$$p(x) = \begin{cases} \frac{1}{N}, & \text{if } x \in A; \\ 0, & \text{otherwise.} \end{cases} \tag{5}$$

Here $A \subset \mathbb{X}_1 \times \ldots \times \mathbb{X}_n$ represents the set of admissible value combinations and N is the cardinality of A. Typically, N is of orders of magnitudes smaller than the cardinality of $\mathbb{X}_1 \times \ldots \times \mathbb{X}_n$, in which case the above prior will strongly confine the posterior. As an example, consider a telephone book, where $\mathbb{X}_1$ is the set of surnames, $\mathbb{X}_2$ is the set of family names, $\mathbb{X}_3$ is the set of city names, etc. Clearly, the number of records in a telephone book is much smaller than the number that we would get by multiplying the cardinalities of the sets $\mathbb{X}_1, \mathbb{X}_2, \ldots$.

As it was already suggested, observations $P_1, \ldots, P_n$ are (random) probability distributions over the respective domains $\mathbb{X}_1, \ldots, \mathbb{X}_n$:

$$P_i : \mathbb{X}_i \to [0, 1], \quad i = 1, 2, \ldots, n.$$

Our goal is to compute various functions of the posterior probability distribution $\Pr(x|P_1,\ldots,P_n)$ over $\mathbb{X}$. Our primary target is to compute

$$x^* = \operatorname*{argmax}_{x\in\mathbb{X}} \Pr(x|P_1,\ldots,P_n),$$

i.e., the most likely value of x given our observations: x^* represents our best possible guess of what Bob might have said, given our observations $P_1,\ldots,P_n$. Another quantity of interest is the entropy $H(\Pr(X|P_1,\ldots,P_n))$, representing the uncertainty of X remaining after the observations $P_1,\ldots,P_n$. Here

$$X = (X_1,\ldots,X_n) \propto p$$

is a random variable representing the choice of Bob.

Using Bayes' theorem we obtain

$$\Pr(X|P_1,\ldots,P_n) \propto \Pr(P_1,\ldots,P_n|X)\Pr(X), \tag{6}$$

where we have dropped the normalizing constant $\Pr(P_1,\ldots,P_n)$. Now, we shall assume that the observations $P_1,\ldots,P_n$ are independent given X:

$$\Pr(P_1,\ldots,P_n|X) = \prod_{i=1}^{n} \Pr(P_i|X).$$

This is quite a reasonable assumption, as normally, the value of P_i will depend on X_i through the SRS and the SLU systems. Therefore, we make the assumption that P_i depends on X only through X_i:

$$\Pr(P_i|X) = \Pr(P_i|X_i), \quad i = 1,2,\ldots,n.$$

We will further assume that $\Pr(P_i|X_i)$ is given in the form

$$\Pr(P_i = p|X_i = x) \propto \exp\left(-d(p,p_x^{(i)})\right), \quad x \in \mathbb{X}_i,\ i = 1,2,\ldots,n, \tag{7}$$

where $d(\cdot,\cdot)$ is a distance-like measure for probability distributions, and $p_x^{(i)}$ is a probability distribution over $\mathbb{X}_i$ representing the expected response of the SR/SLU system given that the input is x. In practice, d will be chosen by the experimenter and $p_x^{(i)}$ will be learnt from experimental data , either directly, or e.g. using information in x actively. For example, if xs are represented by words then their phonetic representations could be used to guide the learning process (ie. one may learn what the probabilities of mistaking one phoneme by another one). The assumption about the form of $\Pr(P_i|X_i)$ is admittedly ad-hoc and could be replaced by a more generic one that models this distribution by some more generic parametric or non-parametric form. The important point to note here is that we assume that $\Pr(P_i|X_i)$ is "known" and can be computed. Note that $\Pr(P_i|X_i)$ models the errors introduced by the SR/SLU systems.

Under these assumptions the r.h.s. of (6) can be computed. We prefer to compute the log-likelihood and thus obtain

$$\log\Pr(X = x|P_1,\ldots,P_n) \propto \log p(x) - \sum_{i=1}^{n} d(P_i,p_x^{(i)}).$$

Therefore, computing x^* amounts to maximizing

$$\Lambda(x) = -\log p(x) + \sum_{i=1}^{n} d(P_i, p_x^{(i)}).$$

Note that if (5) is used to model p then $\Lambda(x) = -\infty$ whenever $x \notin A$ and $\Lambda(x)$ is proportional to

$$\Lambda'(x) = \sum_{i=1}^{n} d(P_i, p_x^{(i)})$$

otherwise. In this case, finding x^* is equivalent to maximizing $\Lambda'(x)$ subject to $x \in A$. This is a widely studied problem of computer science and special purpose, efficient (tree-based) algorithms are available for solving it (see eg. [**?**] in the context of classification). Also, the entropy of $\Pr(X = \cdot | P_1, \ldots, P_n)$ can be computed in an efficient way by eg. computing the most significant peaks of the posterior.

Now, let us consider the case when the SRS returns an N-best list of the form

$$\{(X_i^{(1)}, S_i^{(1)}), \ldots, (X_i^{(N)}, S_i^{(N)})\},$$

where $X_i^{(j)} \in \mathbb{X}_i$ is the jth hypothesis with score $S_i^{(j)}$ returned by the SRS on the ith query. Then, assuming the SRS returns log-probabilities, we have

$$P_i(z) \propto \begin{cases} \frac{1}{Z}\exp(S_i^{(j)}), & \text{if } z = X_i^{(j)}; \\ 0, & \text{otherwise.} \end{cases}$$

Now, choose

$$d(P_1, P_2) = \sum_{z \in \mathbb{X}_i, P_1(z) \neq 0} P_2(z) \log(P_1(z))$$

and for simplicity assume that

$$p_x^{(i)}(z) = \begin{cases} 1, & \text{if } x = z; \\ 0, & \text{otherwise.} \end{cases}$$

Then

$$d(P_i, p_x^{(i)}) = \begin{cases} S_i^{(j)}, & \text{if } x = X_i^{(j)}; \\ 0, & \text{otherwise} \end{cases}$$

and

$$\operatorname{argmax}\left\{ \sum_{i=1}^{n} d(P_i, p_{x_i}^{(i)}) \,|\, x = (x_1, \ldots, x_n) \in A \right\} =$$
$$\operatorname{argmax}\left\{ \sum_{i=1}^{n} S_i^{(j_i)} \,|\, x = (X_i^{(j_1)}, \ldots, X_i^{(j_n)}) \in A,\, (j_1, \ldots, j_n) \in \{1, \ldots, N\}^n \right\}.$$

Now, the advantage of the method should be evident: without any clarifications, Alice may recover the correct message even when the best candidate returned by the SRS in the individual turns is not the correct one. A detailed analysis of the probability of this happening is out of the scope of the present paper and will be presented elsewhere.

4 Conclusions

In this paper we have proposed a method to increase the naturalness of dialogues by reducing the number of clarification subdialogues. In particular, we proposed that the posterior on the meaning of past utterences should be recomputed in the face of new observations if there is a restrictive relationship between past and new interpretations. An example was given for information seeking dialogues where the advantage of the method was made explicit. As another example we mention that often anaphore resolution should be delayed in the hope that new observations will reduce the number of probable candidates (ie. the uncertainty).

An important open research issue is to explore the ways how to structure dialogue control strategies so that the dialogue can be progressed without clarifications. This is a non-trivial problem because the cost of progressing the dialogue in the face of misunderstandings may be high for two reasons: the cost associated with the wasted time (if the exchanges following the point of misunderstanding turn out to be useless) can be high and the cost associated with the clarification itself can be high as well.

Therefore, the technique is most useful when the dialogue can progress the same way independently of the observed uncertain value. This is the case eg. in information seeking dialogues when "independent" variables have to be queried considered in Section 3.

The same idea (of using smoothing) can also be used to delay (self-)repairs for another reason: when Alice observes a discrepancy in the dialogue she may decide not to enter a repair dialogue but to continue, letting the dialogue "unfold" again in the hope that the probability of discrepancy might reduce with time anyway without any explicit actions.

This leads to the next very interesting topic for further research: to build (domain independent) dialogue control strategies that decide about the progress of dialogues in the face of (substantial) uncertainty. Alice's options could eg. be

1. Avoid clarification (eg. continue with a query on an "independent" aspect of the task);
2. Implicit clarification (continue with a query on a "dependent" aspect of the task);
3. Use an explicit clarification subdialogue.

In fact, one may extend model (1) with actions and rewards in which case the model becomes a two-player, sequential Markov game with incomplete observations:

$$
\begin{aligned}
M_{t+1}^{(A)} &= \Pr\left(\cdot | M_t^{(A)}, Y_t^{(B)}, A_{t+1}^{(A)}\right), \\
Y_{t+1}^{(A)} &= \Pr\left(\cdot | M_{t+1}^{(A)}\right), \\
M_{t+1}^{(B)} &= \Pr\left(\cdot | M_t^{(B)}, Y_{t+1}^{(A)}, A_{t+1}^{(B)}\right), \\
Y_{t+1}^{(B)} &= \Pr\left(\cdot | M_{t+1}^{(B)}\right), \\
R_{t+1}^{(A)} &= \Pr\left(\cdot | M_{t+1}^{(A)}, A_{t+1}^{(A)}, M_{t+1}^{(B)}, A_{t+1}^{(B)}\right).
\end{aligned}
$$

Here $A_{t+1}^{(A)}$ is the action of Alice at time $t+1$ (after observing $Y_t^{(B)}$) and symmetrically, $A_{t+1}^{(B)}$ is the action of Bob at time $t+1$ (after observing $Y_{t+1}^{(A)}$), and $R_{t+1}^{(A)}$ (a real valued random variable) is the payoff of Alice at time $t+1$ that depends on the mental states of the agents (actually, the progress of the dialogue – here we assume that the mental states of the agents is sufficient for measuring progress).

One then hopes to be able to use learning techniques (eg. reinforcement-learning, see eg. [2] or [6] for learning in sequential Markov-games) to learn domain independent strategies that maximize some performance index (e.g. maximize task completetion ratio subject to a bound on the task completion time). This could be done eg. along the lines of [3] or [7], but keeping track of the posterior state distribution, allowing one to exploit the known uncertainty of the state in governing the dialogue.

References

1. Arnaud Doucet. On sequential simulation based methods for Bayesian filtering. *Statistics and Computing*, 10(3):197–208, 1998.
2. L.P. Kaebling, M.L. Littman, and A.W. Moore. Reinforcement learning: A survey. *Journal of Artificial Intelligence Research*, 4:237–285, 1996.
3. D. Litman, M. Kearns, S. Singh, and M. Walker. Automatic optimization of dialogue management. In *COLING*, 2000.
4. Pierre Del Moral and Salut G. Non-linear filtering using monte carlo particle methods. *C.R. Acad. Sci. Paris*, pages 1147–1152, 1995.
5. Tim Paek and Eric Horvitz. Uncertainty, utility, and misunderstanding: A decision-theoretic perspective on grounding in conversational systems. In Susan E. Brennan, Alain Giboin, and David Traum, editors, *Working Papers of the AAAI Fall Symposium on Psychological Models of Communication in Collaborative Systems*, pages 85–92, Menlo Park, California, 1999. American Association for Artificial Intelligence.
6. Cs. Szepesvári and M.L. Littman. A unified analysis of value-function-based reinforcement-learning algorithms. *Neural Computation*, 11:2017–2059, 1999.
7. Marilyn A. Walker, Jeanne Frommer, and Shrikanth Narayanan. Learning optimal dialogue strategies: A case study of a spoken dialogue agent for email. In *COLING-ACL*, pages 1345–1351, 1998.
8. S. Young. Probabilistic methods in spoken dialogue systems. In *Proceedings of the Royal Society, London*, 1999.