

# Toy spoken dialogue system POMDP: Problem definition

Jason D. Williams

Machine Intelligence Lab  
Department of Engineering  
University of Cambridge

[jdw30@cam.ac.uk](mailto:jdw30@cam.ac.uk)

# Overview of the problem

A user is interacting with a machine. There are three places called  $a$ ,  $b$ , and  $c$ . The user's goal is to travel from one of these places to another – for example, if the user's goal is to travel from  $a$  to  $b$ , we write this as  $(a,b)$ . However, the machine cannot observe the user's goals, or *even the user's action* (because of speech recognition errors) directly.

The machine's goal is to determine this pair as best it can through interacting with the user, and *submit* it – for example, to a ticket printing machine.

This document frames this problem as a POMDP.

More detail, and motivation, in Ph D first year report:

<http://mi.eng.cam.ac.uk/~jdw30/>

<http://mi.eng.cam.ac.uk/~jdw30/willams2003PhDFirstYearReport.pdf>

# (Machine) actions

When interacting with the user, the machine can ***greet*** the user\*, ***ask*** questions (e.g., “Where do you want to go from?”) and make ***confirming*** statements (e.g., “So you want to go to ***a***, is that right?”).

At the end of the dialogue (*which is at a time of the machine’s choosing*), the machine can either *submit* a pair of locations (e.g., ***submit-a-b***), or ***fail***.

(\*) The machine automatically greets the user during the first turn and after that, the “greet” action is no longer available.

# State space - overview

There are three components to the state space:

- The user's goal
- The user's action
- The state of the dialogue.

# State space – user's goal & action

The user's goal is simply a pair of places, indicating (*from*, *to*). For example, (*a*,*b*).

The user's action may indicate:

- Just a place (e.g., ***a***)
- To or from a place, (e.g., ***from-a***)
- To one place from another (e.g., ***from-a-to-b***)
- ***Yes*** or ***No***
- ***Null*** – which means the user didn't say or do anything

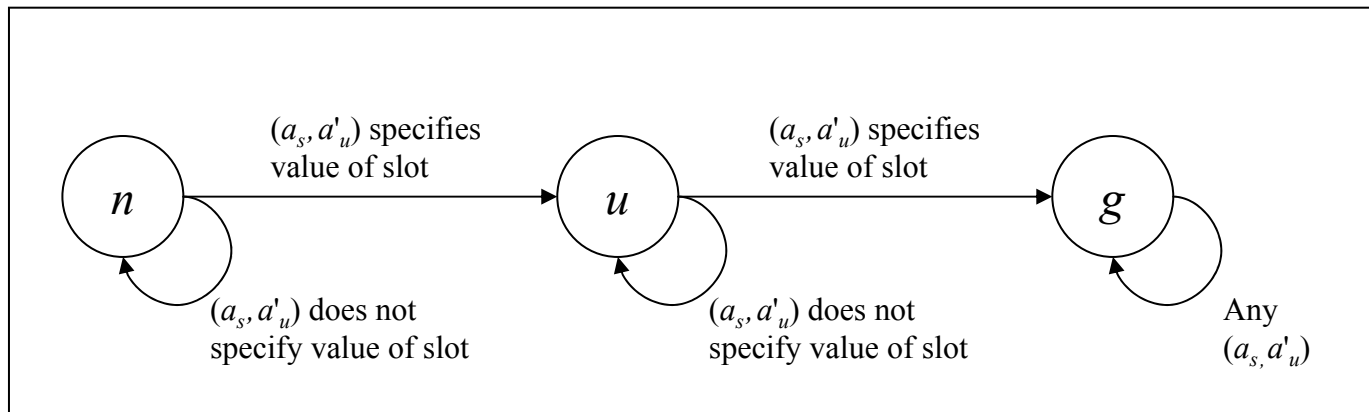
# State space – state of dialogue

We also have a state variable for the state of the *dialogue*. This variable indicates whether the *from* and *to* components have been *grounded* or not. ***n*** means “not yet mentioned”. ***u*** means “mentioned but not grounded”. ***g*** means “grounded”.

Here we take “grounded” to mean “mentioned or confirmed by the user more than once.”

For example, if, so far in the conversation, *a* has been mentioned as the *from* location by the user, and *b* has been mentioned by the user *and confirmed* by the user, then the state of the dialog is ***(u,g)***.

The diagram below shows a state transition diagram for the grounding model.



# Observations

Observations are simply the user's actions combined with a measure of their recognition confidence, which may be ***h*** (high) or ***l*** (low).

Example observations include:

- **from-a-to-b-h**
- **to-a-l**
- **a-l**
- **yes-l**
- **no-h**
- **null-h**

# Summary of variables & values

## State variables & values

---

$$S = (s_u, s_d, a_u)$$

where:

$$a_u \in \begin{cases} x, & x \in \{a, b, c\} \\ \text{from-}x, & x \in \{a, b, c\} \\ \text{to-}x, & x \in \{a, b, c\} \\ \text{from-}x_1\text{-to-}x_2, & x_i \in \{a, b, c\}, x_1 \neq x_2 \\ \text{yes} \\ \text{no} \\ \text{null} \end{cases}$$

$$s_d \in (d_1, d_2), d_i \in \{n, u, g\}$$

$$s_u \in (x_1, x_2), x_i \in \{a, b, c\}, x_1 \neq x_2$$

## (System) action values

---

$$a_s \in \begin{cases} \text{greet} \\ \text{ask-to} \\ \text{ask-from} \\ \text{conf-from-}x, & x \in \{a, b, c\} \\ \text{conf-to-}x, & x \in \{a, b, c\} \\ \text{submit-}x_1\text{-}x_2, & x_i \in \{a, b, c\}, x_1 \neq x_2 \\ \text{fail} \end{cases}$$

## Observation values

---

$$o \in \{x-y\}, x \in a_s, y \in \{h, l\}$$

# Reward function - definition

The reward function is defined as:

$$r(S, a_s) = r(s_u, s_d, a_u, a_s) \Rightarrow \mathfrak{R}$$

The goal of the values of the reward function is both to incent the machine to achieve the goal as well as conform to conversational norms.

The system is rewarded for submitting the correct pair.

The system is penalized for submitting the wrong pair, and for taking the *fail* action.

In addition, the system is also penalized for asking for a field which has already been mentioned, confirming a field which has already been confirmed. There is also a small per-step penalty for any other action.

# Reward function - values

$$r(s_u = (x, y), s_d, a_u, a_s = \text{submit-}x\text{-}y) = +10$$

$$r(s_u = (x, y), s_d, a_u, a_s \neq \text{submit-}x\text{-}y) = -10$$

$$r(s_u, s_d, a_u, a_s = \text{fail}) = -5$$

$$r(s_u, s_d = ([u, c], *), a_u, a_s = \text{ask-from}) = -3$$

$$r(s_u, s_d = (*, [u, c]), a_u, a_s = \text{ask-to}) = -3$$

$$r(s_u, s_d = (*, c), a_u, a_s = \text{conf-to-}x) = -2$$

$$r(s_u, s_d = (c, *), a_u, a_s = \text{conf-from-}x) = -2$$

And if none of the rules above apply, then:

$$r(s_u, s_d, a_u, a_s) = -1$$

# Observation function – definition & values

The observation function is defined as:

$$\begin{aligned} p(o' | s', a) &= p(o' | s'_u, s'_d, a'_u, a_s) \\ &= p(o' | a'_u) \\ &= p(o | a_u) \end{aligned}$$

Where

$$\begin{aligned} p(o = (a-l) | a_u) &= 0.35, & a = a_u \\ p(o = (a-l) | a_u) &\approx 0.015, & a \neq a_u \\ p(o = (a-h) | a_u) &= 0.425, & a = a_u \\ p(o = (a-h) | a_u) &\approx 0.004, & a \neq a_u \end{aligned}$$

# Transition function - definition

First we define three models:

- **The user belief model**  $p(s'_u | s_u)$
- **The user action model**  $p(a'_u | s'_u, a_s)$
- **The dialogue model**  $p(s'_d | s_d, a_s, a'_u)$

The transition function is defined as:

$$p(s' | s, a) = p(s'_u, s'_d, a'_u | s_u, s_d, a_u, a_s)$$

Using the terms above, and conditional independence, we can re-write the transition function as:

$$\begin{aligned} &= p(s'_u | s_u, s_d, a_u, a_s) p(a'_u | s'_u, s_u, s_d, a_u, a_s) p(s'_d | a'_u, s'_u, s_u, s_d, a_u, a_s) \\ &= p(s'_u | s_u) p(a'_u | s'_u, a_s) p(s'_d | a'_u, s_d, a_s) \end{aligned}$$

# Transition function - values

The **user belief model** randomly assigns a (*from,to*) pair at the beginning of the task; after that, the user's goal does not vary

$$p(s'_u = s_u | s_u) = 1$$

$$p(s'_u \neq s_u | s_u) = 0$$

# Transition function - values

The **dialogue model** is a deterministic probability distribution which gives the current state of a dialogue, given the user's action, the system's action, and the prior state of the dialogue

$$p(s'_d = (d_1, d_2) | a'_u = \text{null}, s_d = (d_1, d_2), a_s) = 1$$

$$p(s'_d = (u, d_2) | a'_u = x, s_d = (n, d_2), a_s = \text{ask-from}) = 1$$

$$p(s'_d = (u, d_2) | a'_u = \text{from-}x, s_d = (n, d_2), a_s) = 1$$

$$p(s'_d = (d_1, u) | a'_u = x, s_d = (d_1, n), a_s = \text{ask-to}) = 1$$

$$p(s'_d = (d_1, u) | a'_u = \text{to-}x, s_d = (d_1, n), a_s) = 1$$

$$p(s'_d = (c, d_2) | a'_u = x, s_d = (u, d_2), a_s = \text{ask-from}) = 1$$

$$p(s'_d = (c, d_2) | a'_u = \text{from-}x, s_d = (u, d_2), a_s) = 1$$

$$p(s'_d = (d_1, c) | a'_u = x, s_d = (d_1, u), a_s = \text{ask-to}) = 1$$

$$p(s'_d = (d_1, c) | a'_u = \text{to-}x, s_d = (d_1, u), a_s) = 1$$

$$p(s'_d = (c, d_2) | a'_u = x, s_d = (c, d_2), a_s = \text{ask-from}) = 1$$

$$p(s'_d = (c, d_2) | a'_u = \text{from-}x, s_d = (c, d_2), a_s) = 1$$

$$p(s'_d = (d_1, c) | a'_u = x, s_d = (d_1, c), a_s = \text{ask-to}) = 1$$

$$p(s'_d = (d_1, c) | a'_u = \text{to-}x, s_d = (d_1, c), a_s) = 1$$

$$p(s'_d = (d_1, u) | a'_u = \text{yes}, s_d = (d_1, n), a_s = \text{conf-from-}x) = 1$$

$$p(s'_d = (u, d_2) | a'_u = \text{yes}, s_d = (n, d_2), a_s = \text{conf-to-}x) = 1$$

$$p(s'_d = (d_1, c) | a'_u = \text{yes}, s_d = (d_1, u), a_s = \text{conf-from-}x) = 1$$

$$p(s'_d = (c, d_2) | a'_u = \text{yes}, s_d = (u, d_2), a_s = \text{conf-to-}x) = 1$$

$$p(s'_d = (d_1, d_2) | a'_u = \text{no}, s_d = (d_1, d_2), a_s = \text{conf-from-}x) = 1$$

$$p(s'_d = (d_1, d_2) | a'_u = \text{no}, s_d = (d_1, d_2), a_s = \text{conf-to-}x) = 1$$

$$p(s'_d = (u, u) | a'_u = \text{from-}x_1\text{-to-}x_2, s_d = (n, n), a_s) = 1$$

$$p(s'_d = (c, u) | a'_u = \text{from-}x_1\text{-to-}x_2, s_d = (u, n), a_s) = 1$$

$$p(s'_d = (c, u) | a'_u = \text{from-}x_1\text{-to-}x_2, s_d = (c, n), a_s) = 1$$

$$p(s'_d = (u, c) | a'_u = \text{from-}x_1\text{-to-}x_2, s_d = (n, u), a_s) = 1$$

$$p(s'_d = (c, c) | a'_u = \text{from-}x_1\text{-to-}x_2, s_d = (u, u), a_s) = 1$$

$$p(s'_d = (c, c) | a'_u = \text{from-}x_1\text{-to-}x_2, s_d = (c, u), a_s) = 1$$

$$p(s'_d = (u, c) | a'_u = \text{from-}x_1\text{-to-}x_2, s_d = (n, c), a_s) = 1$$

$$p(s'_d = (c, c) | a'_u = \text{from-}x_1\text{-to-}x_2, s_d = (u, c), a_s) = 1$$

$$p(s'_d = (c, c) | a'_u = \text{from-}x_1\text{-to-}x_2, s_d = (c, c), a_s) = 1$$

and if not specified above,

$$p(s'_d = (d_1, d_2) | a'_u, s_d = (d_1, d_2), a_s) = 1$$

where

$$x, x_1, x_2 \in \{a, b, c\}, x_1 \neq x_2$$

$$d_1, d_2 \in \{u, c, g\}$$

# Transition function - values

The **user action model** specifies how a user responds to the system given the system's action and the user's goal.

$$p(a'_u = \text{null} \mid s'_u = (x_1, x_2), a_s) = 0.1$$

$$p(a'_u = \text{from-}x_1\text{-to-}x_2 \mid s'_u = (x_1, x_2), a_s = \text{greet}) = 0.54$$

$$p(a'_u = \text{from-}x_1 \mid s'_u = (x_1, x_2), a_s = \text{greet}) = 0.18$$

$$p(a'_u = \text{to-}x_2 \mid s'_u = (x_1, x_2), a_s = \text{greet}) = 0.18$$

$$p(a'_u = \text{from-}x_1 \mid s'_u = (x_1, x_2), a_s = \text{ask-from}) = 0.225$$

$$p(a'_u = x_1 \mid s'_u = (x_1, x_2), a_s = \text{ask-from}) = 0.585$$

$$p(a'_u = \text{from-}x_1\text{-to-}x_2 \mid s'_u = (x_1, x_2), a_s = \text{ask-from}) = 0.09$$

$$p(a'_u = \text{to-}x_2 \mid s'_u = (x_1, x_2), a_s = \text{ask-to}) = 0.225$$

$$p(a'_u = x_2 \mid s'_u = (x_1, x_2), a_s = \text{ask-to}) = 0.585$$

$$p(a'_u = \text{from-}x_1\text{-to-}x_2 \mid s'_u = (x_1, x_2), a_s = \text{ask-to}) = 0.09$$

$$p(a'_u = \text{from-}x_1 \mid s'_u = (x_1, x_2), a_s = \text{conf-from-}x_3) = 0.03375$$

$$p(a'_u = x_1 \mid s'_u = (x_1, x_2), a_s = \text{conf-from-}x_3) = 0.10125$$

$$p(a'_u = \text{yes} \mid s'_u = (x_1, x_2), a_s = \text{conf-from-}x_3) = 0.765, \quad x_1 = x_3$$

$$p(a'_u = \text{no} \mid s'_u = (x_1, x_2), a_s = \text{conf-from-}x_3) = 0.765, \quad x_1 \neq x_3$$

$$p(a'_u = \text{to-}x_2 \mid s'_u = (x_1, x_2), a_s = \text{conf-to-}x_3) = 0.03375$$

$$p(a'_u = x_2 \mid s'_u = (x_1, x_2), a_s = \text{conf-to-}x_3) = 0.10125$$

$$p(a'_u = \text{yes} \mid s'_u = (x_1, x_2), a_s = \text{conf-to-}x_3) = 0.765, \quad x_2 = x_3$$

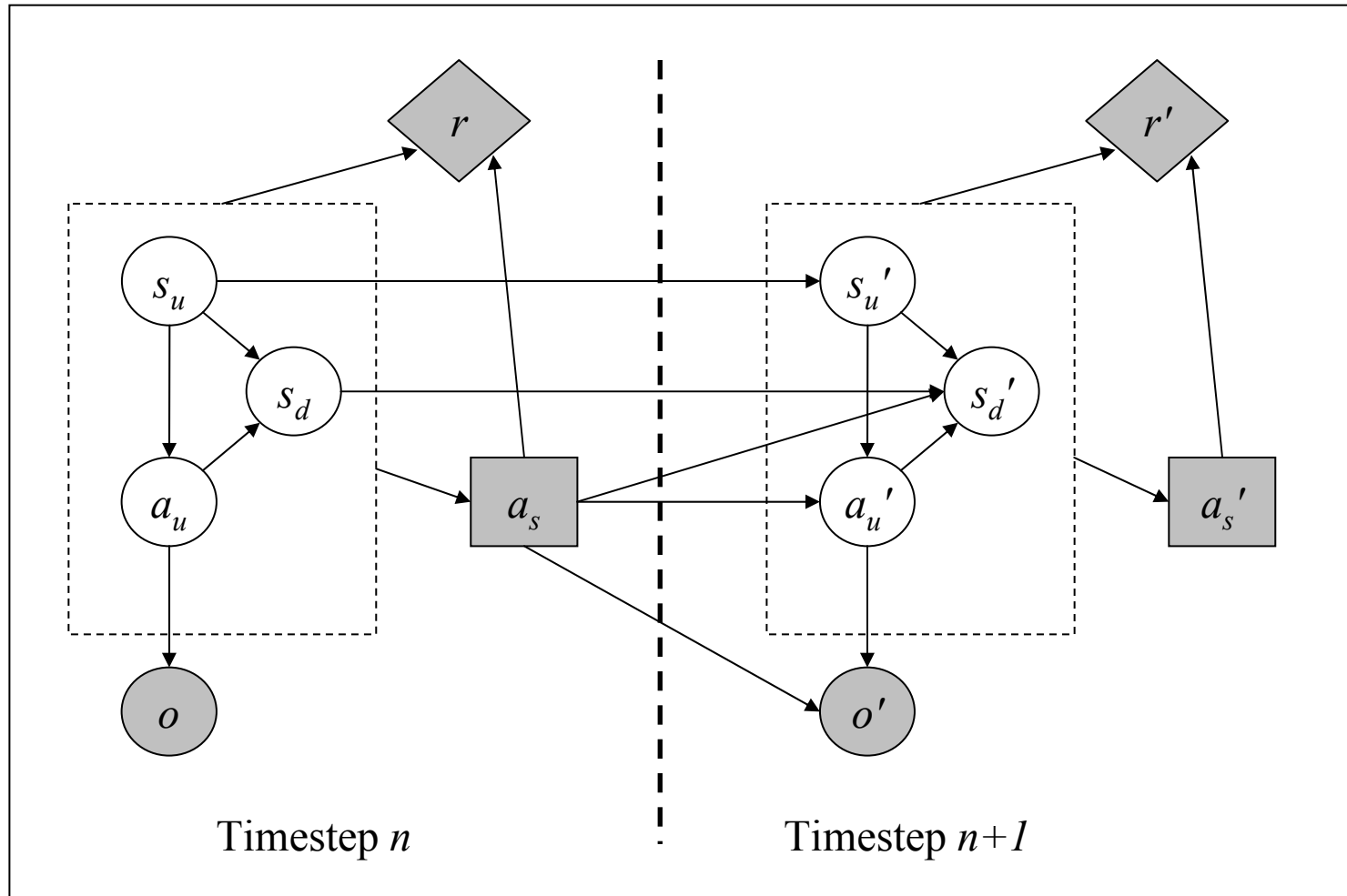
$$p(a'_u = \text{no} \mid s'_u = (x_1, x_2), a_s = \text{conf-to-}x_3) = 0.765, \quad x_2 \neq x_3$$

where

$$x_1, x_2, x_3 \in \{a, b, c\}, x_1 \neq x_2$$

# Influence diagram

This POMDP can be depicted in a customary influence diagram:



# Solutions...

Exact solutions intractable

Witness runs for days

Easy to compute belief state; hard to compute value function

Techniques to approximate value function most appropriate

ANN?