On-line Learning of Wide-Domain Statistical Spoken Dialogue Systems

(Talking to Machines)

Steve Young
Objectives

To develop spoken dialogue systems which:

1. allow users to access multiple domains within a single conversation
2. support natural conversations even in rarely visited domains
3. learn automatically on-line through interaction with user

“Deploy, Collect Data, Improve”

“User in the loop” enables on-line reinforcement learning
Reinforcement Learning

Models/Parameters

ASR

User

TTS

Understand and Respond

Knowledge Base
Reinforcement Learning

- **Models/Parameters**
- **Understand and Respond**
- **ASR**
- **TTS**
- **User**
- **Optimise Reward**
- **Knowledge Base**

*Compute Reward*
I'd like a cheap Italian on the east side of town

Inform:
\[
\text{price}=\text{cheap}, \quad \text{food}=\text{Italian}, \quad \text{area}=\text{east} \quad [0.7]
\]

You'd like a cheap restaurant on the east side of town?
What kind of food would you like?

Confirm requisition:
\[
\text{confirm-request(} \quad \text{price}=\text{cheap, area}=\text{east, food}=?)\]

Confirm requisition:
\[
\text{confirm-request(food)}\]
Extending to Wide Domains

Two key problems:

1. How to expand coverage from a single limited domain to wide or even unlimited domains

2. How to measure success (and hence a reward signal)
What appointments do I have tomorrow?
You have a meeting at 10am with John and a teleconf at noon with Bill.
I need to go to London first thing, can you reschedule the meeting with John?
John is free tomorrow at 3pm, is that ok?
Yes, that's fine. I also need a taxi to the station.
What time do you need the taxi?
When does the train depart to London?
The 9.15am gets in at 10.06.

Share the Reward
Bayesian Committee Machine

• Each DM operates independently, receives speech, tracks its own beliefs and proposes system actions

• DM’s operate as a Bayesian Committee Machine, each machine’s Q-value has a confidence attached to it:

\[
\bar{Q}(b,a) = \Sigma^Q(b,a) \sum_{i=1}^{M} \Sigma^Q_i(b,a)^{-1} \bar{Q}_i(b,a)
\]

\[
\Sigma^Q(b,a)^{-1} = -(M-1) \times k((b,a),(b,a))^{-1} + \sum_{i=1}^{M} \Sigma^Q_i(b,a)^{-1}
\]

• Reinforcement learning operates on the group, distributing rewards at each turn according to previous action selection.

Modular, flexible, incremental, trainable on-line, …

Bayesian Committee Machine Training Performance

![Graph showing performance comparison]

- Laptop domain trained in parallel with Hotels and Restaurants
- Laptop domain trained in isolation

95% Confidence Interval
Training with Real Users

Most research results are obtained using paid subjects given prescribed tasks. Moving to real systems presents problems:

- Reward depends on task success which is very hard to measure
- Explicit user feedback is costly and unreliable

Solution:
- Learn an embedding function for dialogues (using a Bi-LSTM)
- Train a Gaussian Process based classifier to estimate reward success
- Use GP uncertainty estimate to limit use of explicit user feedback
- Use GP noise model to compensate for unreliable user feedback
On-line Reward Estimation

Understanding

ASR
Semantic Decoder
Turn Level
Belief Tracker
Dialogue Level

Policy Based Decision Logic

Dialog Manager

User

Dialog Manager

Database/Application

Generation

TTS
Message Generator
Turn Level
Response Planner
Dialogue Level
On-line Reward Estimation

GP-based Reward Estimator

If low confidence then
Prompt for user feedback

“good” or “bad”

LSTM Encoder

64-D embedding

Estimated Reward Signal

User

ASR

Semantic Decoder

Belief Tracker

Policy Based Decision Logic

ASR

Understanding

Turn Level

Dialog Level

Message Generator

Response Planner

TTS

Generation

Database/Application

Episodic Dialogue Features

If low confidence then
Prompt for user feedback

“good” or “bad”

Estimated Reward Signal
On-line Active Reward Learning for Policy Optimisation in Spoken Dialogue Systems."

ACL 2016, Berlin.
Conclusions

• Technology components are now in place to build large scale wide-domain spoken dialogue systems

• Capability and user acceptance of Virtual Personal Assistants (VPAs) will increase rapidly

• Key is ability to learn on-line thru interaction with users and sharing data with other VPAs

• VPAs will become autonomous entities, independent of any specific device

• This will raise many issues: ensuring veracity of VPA derived information, personal privacy, consumer protection, …