Introduction to statistical spoken dialogue systems

Milica Gašić

Dialogue Systems Group, Cambridge University Engineering Department

October 28, 2016
In this lecture...

Architecture of a spoken dialogue system

Turn-taking in dialogue

Dialogue acts

Modular architecture of a dialogue system
What is a spoken dialogue system?

- A spoken dialogue system is a computer system that enables human computer interaction where primary input is speech.
- Speech does not need to be the only input. We can interact with machines also using touch, gesture or facial expressions and these are multi-modal dialogue systems.
Examples from popular culture
Dialogue and AI

- Turing test: Are we talking to a machine or a human?
- Chat-bots ▶ Eliza
- Goal-oriented dialogue - there is a goal or several goals that must be fulfilled during conversation ▶ Medical Bayesian Kiosk
Personal assistants

- Most commonly used dialogue systems are personal assistants such as Siri, Cortana, Google Now and Alexa.
- These are server-based accessed via a range of devices: smart-phones, tablets, laptops, watches and specialist devices such as Amazon Echo (Alexa).
Properties

What constitutes a spoken dialogue system?

- Being able to understand the user
- Being able to decide what to say back
- Being able to conduct a conversation beyond simple voice commands or question answering
ontology a database that defines properties of entities that a dialogue system can talk about

system-initiative vs user-initiative who takes the initiative in the dialogue:

▶ System: Hello. Please tell me your date of birth using the six digit format.
▶ System: Hello, how may I help you?
Turn-taking in dialogue – Who speaks when?

Dialogue can be described in terms of system and user turns

- System: *How may I help you?*
- User: *I’m looking for a restaurant*
- System: *What kind of food would you like?*
- ...

Turn taking can be more complex and characterised by

- **barge-ins** System: *How may I... User: I’m looking for a restaurant*

- **back channels** User: *I’m looking for a restaurant* [System: *uhuh*] *in the centre of town*
Turn-taking in dialogue – Multi-party dialogue

- Dialogue system can be built to operate with multiple users and also be situated in space.
- Example: Robot giving directions
- In this case a complex attention mechanism is needed to determine who speaks when. For that both spoken and visual input can be used.
Dialogue acts

One simple dialogue act formalism would consist of

dialogue act type - encodes the system or the user intention in a (part of) dialogue turn

semantic slots and values - further describe entities from the ontology that a dialogue turn refers to

Is there um maybe a cheap place in the centre of town please?

inform (price = cheap, area = centre)
Dialogue acts

Dialogue act formalism describes meaning encoded in each dialogue turn [Traum, 2000].

- Relation to ontology
- Encode intention of the speaker
- Relation to logic
- Context
- Partial information from ASR (primitive dialogue acts)
Traditional approach to dialogue systems – Call flow

What are you looking for a restaurant or a hotel?

restaurant

hotel

What kind of food?

Chinese

three

How many stars?

What area?

center
Node processing in a call flow

Speech generation

Ask question

Speech understanding

Load Grammar

Conf. score

no

med

yes

low

high

14 / 32
Part of a deployed call-flow [Paek and Pieraccini, 2008]
Problems

What breaks dialogue systems?

- Speech recognition errors
- Not keeping track of what happened previously
- Need to hand-craft a large number of rules
- Poor decisions
- User’s request is not supported
- ...
Modular architecture of a dialogue system

I'm looking for a restaurant
inform(type=restaurant)

Speech recognition → Semantic decoding → Dialogue management → Ontology

Speech synthesis ← Natural language generation

What kind of food do you have in mind?
request(food)
Machine learning in spoken dialogue systems

- Dialogues
- Labelled user intents
- Transcribed speech

- Regression
- Classification
- Markov decision process
- Neural networks

- What the user wants
- What is the best response
- How to formulate the response
Architecture of a statistical dialogue system

- Speech recognition
- Semantic decoding
- Dialogue management
- Natural language generation
- Speech synthesis
- Ontology

Waveform distribution over text hypotheses
Distribution over dialogue acts
Automatic speech recognition for dialogue

Provide alternative recognition result

- N-best list
- Confusion network
- Lattice

**Figure 1:** Confusion network
Automatic speech recognition for dialogue systems

Recognise when the user has started speaking
- Key-word spotter running on a smartphone - always listening [Chen et al., 2015]
- Requirements: low memory footprint, low computational cost and high precision

Recognise when the user has stopped speaking
- This is studied in the broad context of voice activity detection
Acoustic modelling for dialogue systems

- Spoken dialogue systems are meant to be used everywhere: busy street, noisy car
- Advantage: the conversation spans over several turns so it is possible to perform adaptation in the first turn to improve future interactions
- Advantage: the same speaker through-out the dialogue
Language modelling for dialogue systems

- The vocabulary in limited domain dialogue systems is small so the language model can be trained with in-domain data.
- A general purpose language model can be combined with in-domain language model to provide better recognition results and also deal with out-of-domain requests.
Semantic decoding for dialogue systems

- Extract meaning from user utterance

- *Do they serve Korean food*
- *Can you repeat that please*
- *Hi I want to find an Italian restaurant*
- *I want a different restaurant*

- confirm(food=Korean)
- repeat()
- hello(type=restaurant, food=Italian)
- reqalts()
Statistical semantic decoding

Italian restaurant please

Delexicalise

<tagged-food-value> restaurant please

Count N-grams

please 1
restaurant please 1
restaurant 0
<tagged-food-value> restaurant 1

Query SVMs

Ontology

classifier for dialog act
inform() 0.9
request() 0.1

classifier for food=<tagged-food-value>
0.9

classifier for area=<tagged-area-value>
0.1

classifier for price=<tagged-price-value>
0.2

inform(food=Italian) 0.85
request() 0.15

Produce valid dialogue acts and renormalise distributions
Dialogue management

- Maintain belief about what user said: dialogue states/user intent
- Choose the best answer

I’m looking for a Thai restaurant.

Semantic input: inform(type=restaurant, food=Thai)

States/user intent:
- type
- food

Actions/system response:
- request(area)
Belief tracking and policy optimisation

- Speech recognition
- Semantic decoding
- Natural language generation
- Policy optimisation
- Supervised learning
- Reinforcement learning

Relations:
- Waveform
- Distribution over text hypotheses
- Distribution over dialogue acts

Outputs:
- Ontology

Inputs:
- Speech synthesis
Reinforcement learning for dialogue

Q-function: \( Q^\pi(b, a) = E_\pi \left( \sum_{t=k}^{T} r_k | b_t = b, a_t = a \right) \)

Policy: \( \pi(b) = a \)
Natural language generation for dialogue systems

- Generate semantic representation of system action into natural text

- request(area)
- select(pricerange=expensive, pricerange=cheap)
- confirm(food=Korean)
- inform(name="Little Seoul", food=Korean, area=centre)

- What part of town are you looking for?
- Are you looking for something cheap or expensive?
- Did you say Korean food?
- Little Seoul is a nice Korean restaurant in the centre.
Speech synthesis for dialogue systems

- In a dialogue system the context is available from the dialogue manager.
- Text-to-speech system can make use of the context to produce more natural and expressive speech [Yu et al., 2010].
Goal directed limited domain dialogue systems
Turn taking mechanism between system and user
Dialogue act formalism for conveying meaning
Limitations of traditional approach
Architecture of statistical dialogue systems
Role of each component in a spoken dialogue system
References


