

NPFL099 Statistical Dialogue Systems

1. Introduction

<https://ufal.cz/npfl099>

Ondřej Dušek & Vojtěch Hudeček

29. 9. 2020



Charles University
Faculty of Mathematics and Physics
Institute of Formal and Applied Linguistics



unless otherwise stated

Organizational: 2/1 Z+ZK – 4 Credits

- Lecture (Tue 10:40am) + labs (Tue 9:50am)
 - on this Zoom link + <https://ufaldsg.slack.com/>
 - labs approx. bi-weekly (announced in advance), starting next week
 - Lecture: theory
 - Labs: practical projects & how-tos for experiments
- To pass the course:
 - written/spoken exam – freeform questions (covered by the lectures)
 - main ideas, not specifics of a random system we show for 5 minutes
 - labs homeworks/projects – building experimental systems (by agreement)
- Slides, news etc. at <http://ufal.cz/npfl099>
- vs. NPFL123: no ASR/TTS, more advanced
 - but also covering the basics, i.e. there's some overlap

About Us

Ondřej Dušek: lectures, course guarantor

- PhD at ÚFAL, 2 years at Heriot-Watt Uni Edinburgh, now back
- worked mostly on language generation
- also chatbots (HWU Alexa Prize team)



Vojtěch Hudeček: some labs, a bit of lectures

- PhD student at ÚFAL (4rd year)
- working on dialogue management & language understanding
- internships at Uber AI & UC Davis on dialogue systems



Course Syllabus

1. Introduction (today) ***
2. Machine learning techniques *
3. Evaluation **
4. Natural language understanding *
5. Dialogue state tracking *
6. Dialogue management *
7. Natural language generation *
8. End-to-end dialogue models (2 parts)
9. Chatbots **
10. Multimodal/visual dialogue
11. Ethics & Linguistics & Problems **

*/**/** = little/some/lot
of overlap with NPFL123

Recommended Reading

Primary:

- Jurafsky & Martin: Speech & Language processing. 3rd ed. draft 2019, Chap. 26 (<https://web.stanford.edu/~jurafsky/slp3/>) – basic, brief intro
- Gao et al.: Neural Approaches to Conversational AI, 2019 (<http://arxiv.org/abs/1809.08267>) – more advanced

Other (see also website):

- Janarthanam: Hands-On Chatbots and Conversational UI Development. Packt 2017
- McTear et al.: The Conversational Interface: Talking to Smart Devices. Springer 2016.
- Jokinen & McTear: Spoken dialogue systems. Morgan & Claypool 2010.
- Lemon & Pietquin: Data-Driven Methods for Adaptive Spoken Dialogue Systems. Springer 2012.
- Rieser & Lemon: Reinforcement learning for adaptive dialogue systems. Springer 2011.
- Psutka et al.: Mluvíme s počítačem česky. Academia 2006.

What's a dialogue system?

Definition:

- A (*spoken*) dialogue system is a **computer system designed to interact** with users **in (*spoken*) natural language**
- Wide definition – covers lots of different cases
 - “smart speakers” / phone OS assistants
 - phone hotline systems (even tone-dial ones)
 - in-car systems
 - assistive technologies: therapy, elderly care, companions
 - entertainment: video game NPCs, chatbots



Where are we now?

- Lots of talk about AI now
- Hype around Siri/Alexa/Google
- Sci-fi expectations – AI-complete
 - Star Trek – know-it-all <https://youtu.be/1ZXugicgn6U?t=3>
 - 2001 Space Odyssey – mutiny <https://youtu.be/qDrDUmuUBTo>
 - Her – personality https://youtu.be/6QRvTv_tpw0?t=27
- We're not there – probably for long
 - main bottleneck: understanding (not speech comprehension, meaning!)
 - ... more like the Red Dwarf talkie toaster

https://youtu.be/LRq_SAuQDec?t=71



HOWDY DOODLY DOO! IM TALKIE TOASTER!
YOUR CHIRPIE BREAKFAST COMPANION!

Example – Smart Speakers

- Google, Amazon, Apple & others, Mycroft: open-source
- Really good microphones
 - and not much else – they work online only
- Huge knowledge bases
 - Google: combined with web search
- Lots of domains programmed in, but all by hand
 - integration with a lot of services (calendar, music, shopping, weather, news...)
 - you can add your own (with limitations)
- Can keep some context
- Conversational capabilities limited



<https://www.lifehacker.com.au/2018/02/specs-showdown-google-home-vs-amazon-echo-vs-apple-homepod/>

<https://homealarmreport.com/smart-home/amazon-echo-vs-google-home/>

Why take interest in Dialogue Systems?

- It's *the ultimate natural interface* for computers
- Exciting & **active research topic**
 - some stuff works, but there's a long way to go
 - potential in many domains
 - integrates many different technologies
 - lots of difficult AI problems – **dialogue is hard!**
 - Turing test by dialogue – “proof” of general AI
- **Commercially viable**
 - interest & investment from major IT companies

Basic Dialogue System Types

Task-oriented

- focused on completing a certain task/tasks
 - booking restaurants/flights, finding bus schedules, smart home...
- most actual DS in the wild
 - also our main focus in this course
- “backend access” vs. “agent/assistant”

Non-task-oriented

- chitchat – social conversation, entertainment
 - getting to know the user, specific persona
- gaming the Turing test

Communication Domains

- “domain” = conversation topic / area of interest
- traditional: **single/closed-domain**
 - one well-defined area, small set of specific tasks
 - e.g. banking system on a specific phone number
- **multi-domain**
 - basically joining several single-domain systems (Google/Alexa/Siri)
- **open-domain**
 - “responds to anything” – the goal, but now mostly chitchat-only

Modes of Communication

- **text**

- most basic/oldest
- easiest to implement, most robust
- not completely natural

- **voice**

- more difficult, but can be more natural
 - emotions, tone, personality
- easy to deploy over the phone
- hands-free

- **multimodal**

- voice/text + graphics
- additional modalities: video – gestures, mimics; touch
- most complex

(Johnston et al., ACL 2002)

<https://www.aclweb.org/anthology/P02-1048/>



(Al Moubayed et al., 2012)

https://dl.acm.org/doi/10.1007/978-3-642-34584-5_9

https://www.eitdigital.eu/typo3temp/assets/_processed_/a/6/csm_FURHAT_ea50ba2bf9.jpg

Dialogue Initiative

- **system-initiative**

- “form-filling” (*“Hello. Please tell me your date of birth.”*)
- system asks questions, user must reply in order to progress
- traditional, most robust, but least natural

- **user-initiative**

- user asks, machine responds (*“Alexa, set the timer for two minutes”*)

- **mixed-initiative**

- system and user both can ask & react to queries
- most natural, but most complex

S: Hello. How may I help you?

U: I'm looking for a restaurant.

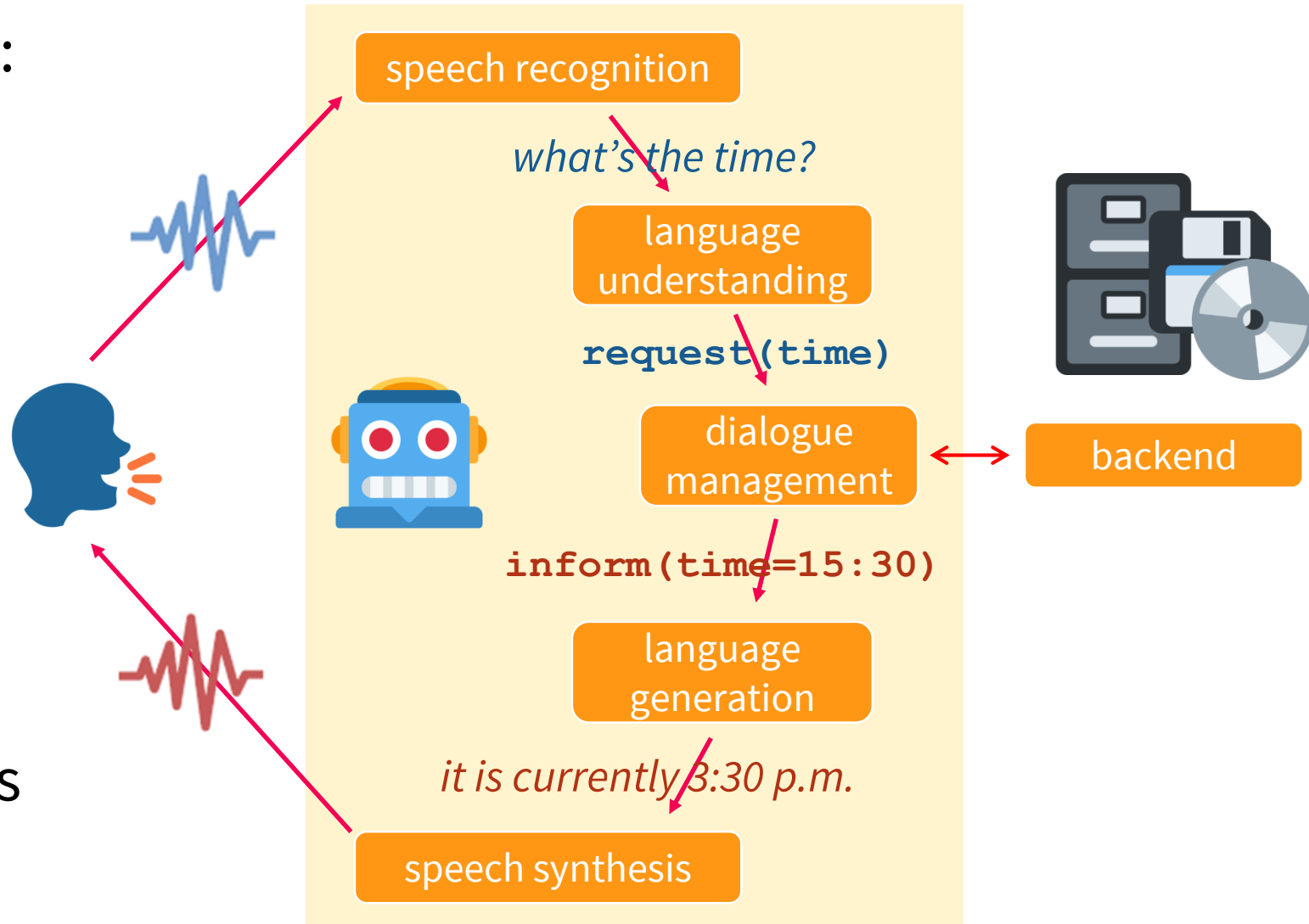
S: What price do you have in mind?

U: Something in the city center please.

S: OK, city center. What price are you looking for?

Dialogue Systems Architecture

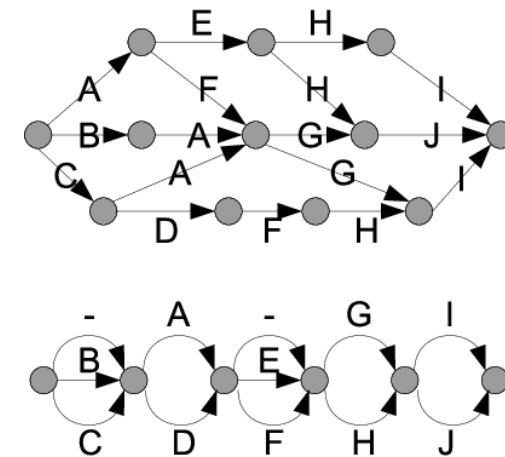
- traditional main DS pipeline:
 - voice → text
 - text → meaning
 - meaning → reaction
 - reaction → text
 - text → voice
- access to backend
 - for anything better than basic chit-chat
- multimodal systems need additional components



Automatic Speech Recognition (ASR)

- Converting **speech signal** (acoustic waves) **into text**
- Typically produces several possible hypotheses with confidence scores
 - **n-best list**
 - lattice
 - confusion network
- Very good in ideal conditions
- **Problems:**
 - noise, accents, longer distance, echo cancellation, channel (phone)...

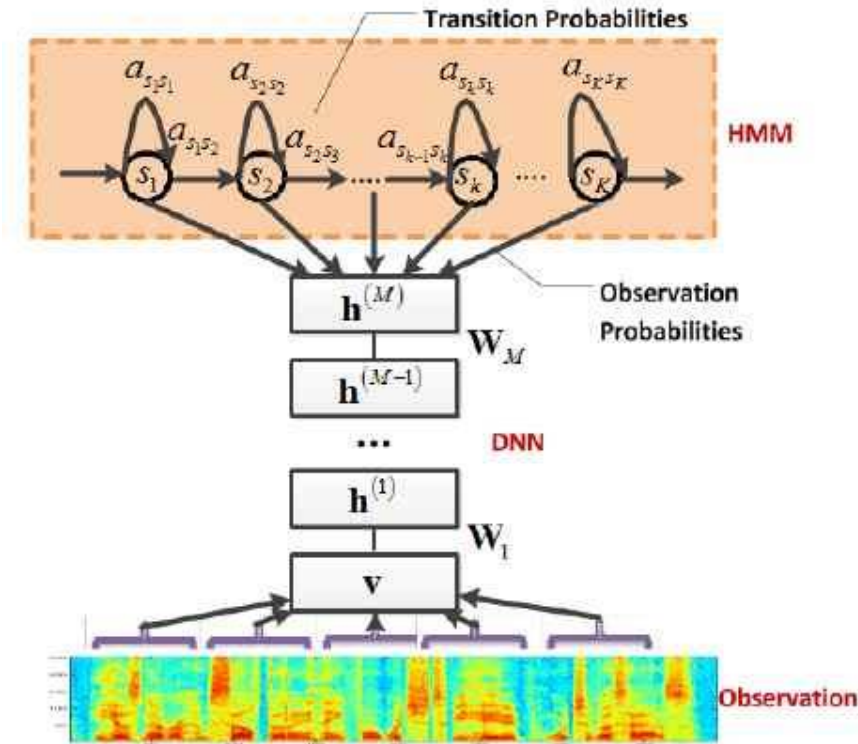
0.8 I'm looking for a restaurant
0.4 uhm looking for a restaurant
0.2 looking for a rest tour rant



(Kazemian et al., ICMR 2008)
<https://doi.org/10.1145/1460096.1460112>

Speech Recognition

- Also: voice activity detection
 - detect when the user started & finished speaking
 - wake words (“*OK, Google*”)
- ASR implementation: mostly neural networks
 - take acoustic features (frequency spectrum)
 - compare with previous
 - emit phonemes/letters
- Limited domain: use of language models
 - some words/phrases more likely than others
 - previous context can be used
 - this can improve the experience **a lot!**
 - problem: out-of-vocabulary



<https://www.i-programmer.info/images/stories/News/2011/AUG/DNNspeech.jpg>

Natural/Spoken Language understanding (NLU/SLU)

- **Extracting the meaning** from the (now textual) user utterance
- Converting into a structured semantic representation
 - **dialogue acts:**
 - act type/intent (*inform, request, confirm*)
 - slot/attribute (*price, time...*)
 - value (*11:34, cheap, city center...*)
 - typically intent detection + slot-value tagging
 - other, more complex – e.g. syntax trees, predicate logic
- **Specific steps:**
 - **named entity resolution** (NER)
 - identifying task-relevant names (*London, Saturday*)
 - **coreference resolution**
 - (“*it*” → “*the restaurant*”)

inform(food=Chinese, price=cheap)
request(address)

Language Understanding

- Implementation varies
 - (partial) **handcrafting** viable for limited domains
 - keyword spotting
 - regular expressions
 - handcrafted grammars
 - **machine learning** – various methods
 - intent classifiers + slot/value extraction
- Can also provide n-best outputs
- Problems:
 - recovering from bad ASR
 - ambiguities
 - variation

S: Leaving Baltimore. What is the arrival city?

U: fine Portland [ASR error]

S: Arriving in Portland. On what date?

U: No not Portland Frankfurt Germany

[On a Tuesday]

U: I'd like to book a flight from London to New York for next Friday

U: Chinese city center

U: uhm I've been wondering if you could find me a restaurant that has Chinese food close to the city center please

Dialogue Manager (DM)

- Given NLU input & dialogue so far, responsible for **deciding on next action**
 - keeps track of what has been said in the dialogue
 - keeps track of user profile
 - interacts with backend (database, internet services)
- Dialogue so far = **dialogue history**, modelled by **dialogue state**
 - managed by **dialogue state tracker**
- System actions decided by **dialogue policy**

Dialogue state / State tracking

- Stores (a summary of) dialogue history
 - User requests + information they provided so far
 - Information requested & provided by the system
 - User preferences
- Implementation
 - **handcrafted** – e.g. replace value per slot with last-mentioned
 - good enough in some circumstances
 - **probabilistic** – keep an estimate of per-slot preferences based on SLU output
 - more robust, more complex

price: cheap
food: Chinese
area: riverside

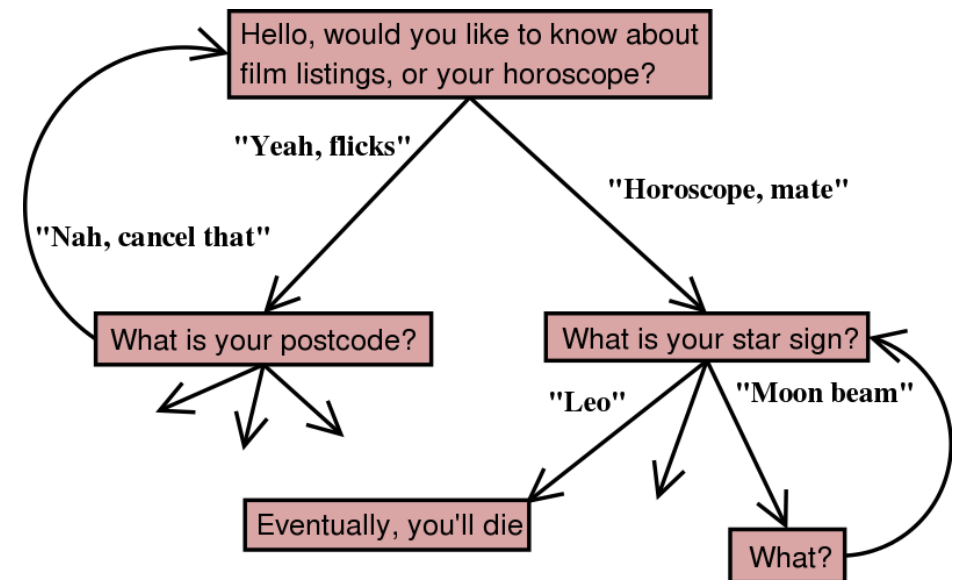
price: 0.8 cheap
0.1 moderate
0.1 <null>
food: 0.7 Chinese
0.3 Vietnamese
area: 0.5 riverside
0.3 <null>
0.2 city center

Dialogue Policy

- Decision on next system action, given dialogue state
- Involves backend queries
- Result represented as system dialogue act
- Handcrafted:
 - **if-then-else** clauses
 - **flowcharts** (e.g. VoiceXML)
- Machine learning
 - often trained with **reinforcement learning**
 - POMDP (Partially Observable Markov Decision Process)
 - recurrent neural networks

confirm(food=Chinese)

inform(name=Golden Dragon,
food=Chinese, price=cheap)



Natural Language Generation (NLG) / Response Generation

- Representing system dialogue act in natural language (text)
 - reverse NLU
- How to express things might depend on context
 - Goals: fluency, naturalness, avoid repetition (...)
- Traditional approach: **templates**
 - Fill in (=lexicalize) values into predefined templates (sentence skeletons)
 - Works well for limited domains

inform(name=Golden Dragon, food=Chinese, price=cheap)

+

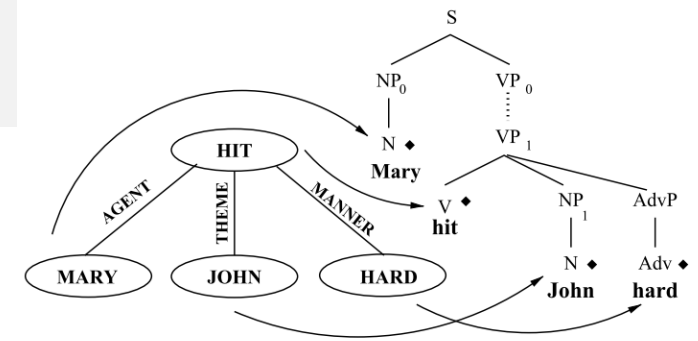
<name> is a <price>-ly priced restaurant serving <food> food

=

Golden Dragon is a cheaply priced restaurant serving Chinese food.

Natural Language Generation

- Grammar-based approaches
 - grammar/semantic structures instead of templates
 - NLG **realizes** them (=converts to linear text) by applying syntactic transformation rules
- Statistical approaches
 - most prominent: **recurrent neural networks**
 - generating word-by-word
 - input: encoded semantics + previous words

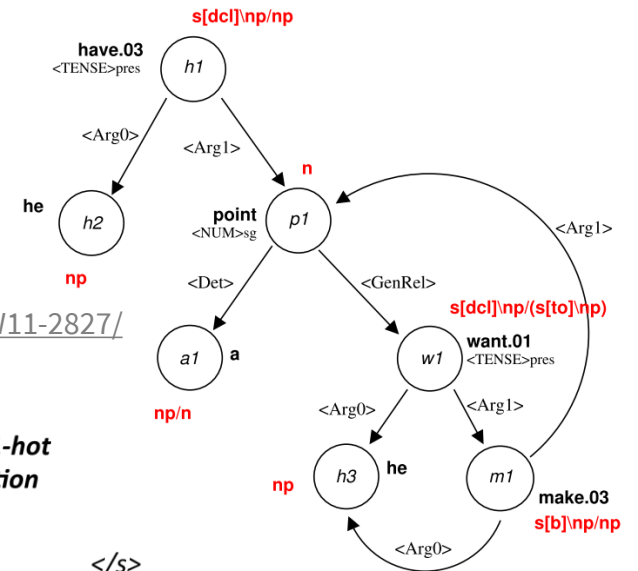


(Kozlowski, 2002)

<http://www.eecis.udel.edu/~mccoy/publications/2002/Kozlowski-ACL-Stu.ps>

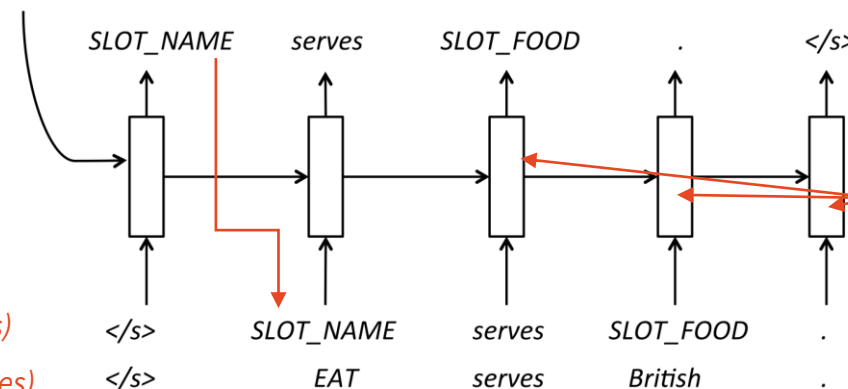
(White, 2011)

<https://www.aclweb.org/anthology/W11-2827/>



Inform(name=EAT, food=British)
 [0, 0, 1, 0, 0, ..., 1, 0, 0, ..., 1, 0, 0, 0, 0, 0...]

dialog act 1-hot representation



delexicalized (generates templates)

after lexicalization (filling in templates)

(Wen et al. 2015)

<http://aclweb.org/anthology/W15-4639>

Text-to-speech (TTS) / Speech Synthesis

- Generate a speech signal corresponding to NLG output
 - text → sequence of **phonemes**
 - minimal distinguishing units of sound (e.g. [p], [t], [ŋ] “ng”, [ə] “eh/uh”, [i:] “ee”)
 - + pitch/intonation, speed, pauses, volume/accents
- Standard pipeline:
 - text normalization
 - abbreviations
 - punctuation
 - numbers, dates, times
 - pronunciation analysis (**grapheme → phoneme conversion**)
 - intonation/stress generation
 - waveform synthesis





take bus number 3 at 5:04am

take bus number three at five o four a m

t eɪ k b ʌ s n ʌ m b ə θ r iː æ t f aɪ v ə ʊ f ɔː r eɪ ɛ m

Speech Synthesis

- TTS Methods:

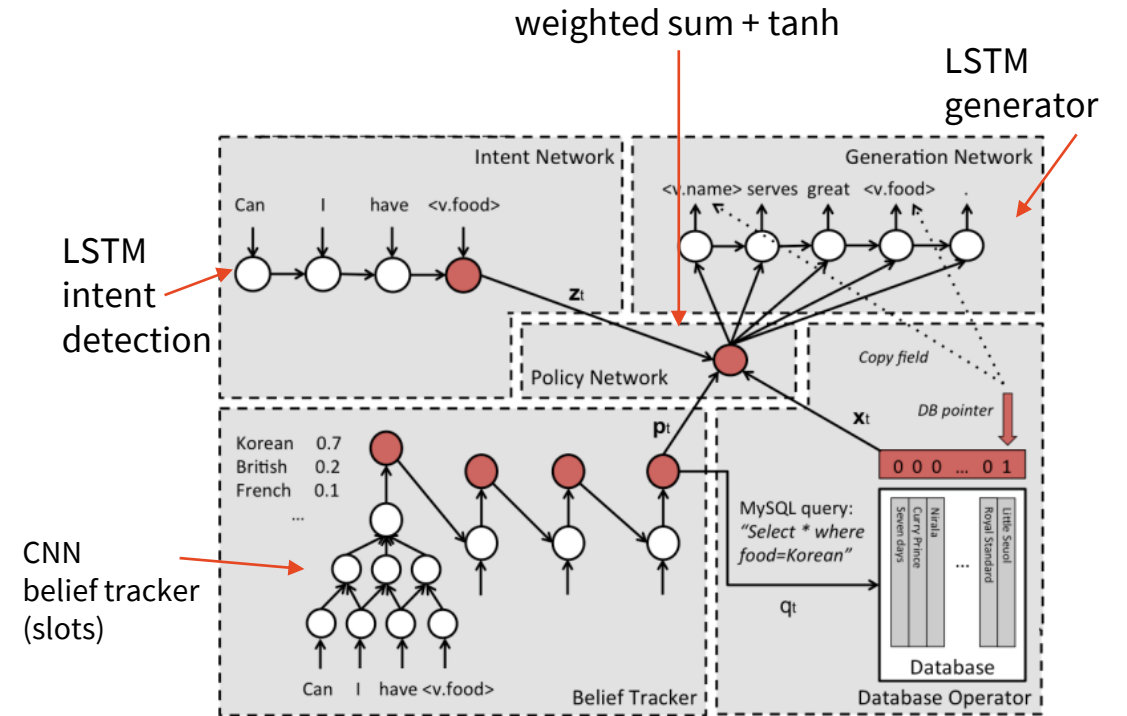
- **Formant-based**: phoneme-specific frequencies  <http://www.festvox.org/history/klatt.html> (example 35)
 - oldest, not very natural, but works on limited hardware
- **Concatenative**  <https://en.wikipedia.org/wiki/MBROLA>
 - record a single person, cut into phoneme transitions (diphones), glue them together
- **Hidden Markov Models**  <http://homepages.inf.ed.ac.uk/jyamagis/>
 - phonemes in context modelled as hidden Markov models
 - Model parameters estimated from data (machine learning)
- **Neural networks**  <https://google.github.io/tacotron/>
 - HMMs swapped for a recurrent neural network
 - also can go directly from text, no need for phoneme conversion

Organizing the Components

- Basic: pipeline
 - ASR → NLU → DM → NLG → TTS
 - components oblivious of each other
- Interconnected
 - read/write changes to dialogue state
 - more reactive (e.g. incremental processing), but more complex
- Joining the modules (experimental)
 - ASR + NLU
 - NLU + state tracking
 - NLU & DM & NLG

End-to-End Systems

- now typical for non-task-oriented
 - single network, trained e.g. on movie subtitles
- task oriented – very experimental
- the whole system (NLU/DM/NLG) is a single neural network
 - joint training (“end-to-end”)
 - more elegant
 - potentially easily retrainable
- typically still needs annotation
 - same as individual modules
 - can be less predictable
- connecting the database is a problem



(Wen et al., 2017)

<https://www.aclweb.org/anthology/E17-1042/>

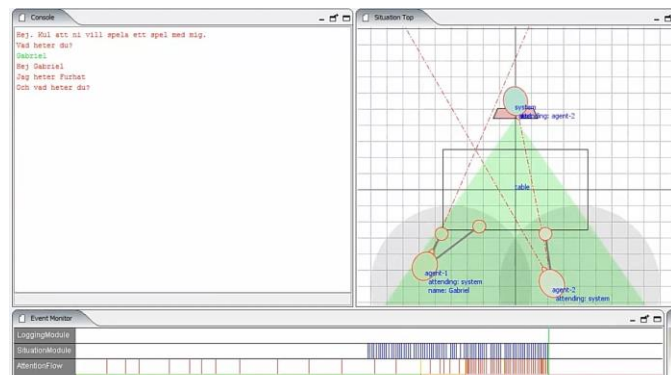
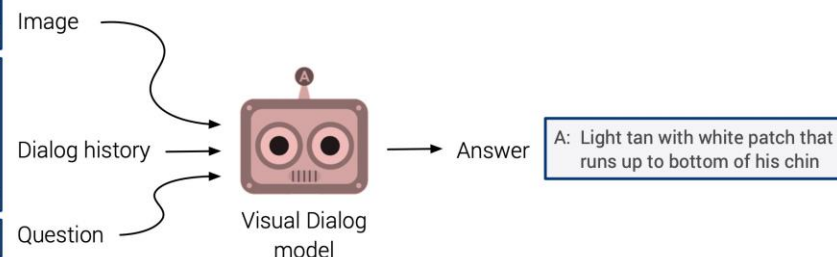
Multimodal/Visual Dialogue

- adding other modalities
- specific components
 - parallel to NLU
 - vision – image classification networks
 - face identification/tracking
 - parallel to NLG
 - mimics/gesture generation
 - gaze
 - image retrieval
 - vision – typically CNN
 - often off-the-shelf stuff
 - specific classifiers/rules



C: A dog with goggles is in a motorcycle side car.
 Q: Is motorcycle moving or still?
 A: It's parked
 Q: What kind of dog is it?
 A: Looks like beautiful pit bull mix
 Q: What color is it?

<http://demo.visualdialog.org/>



SHOPPER: Hello
AGENT: Hi, please tell me what i can help you with today?
SHOPPER: show me few of your top large sized rubber type upper material clogs that is mostly light pink in colored that i would like .
AGENT: Of course. Just wait a few seconds while i browse through my catalog
AGENT: Sorry i dont have any in pink but would you like to see some in



other color

SHOPPER: Please show me something similar to the 1st image but in a different upper material

AGENT: The similar looking ones are



SHOPPER: I like the 4th result . Show me something like it but in material as in the 1st image from what you had previously shown me in clogs



<https://youtu.be/5fhjuGu3d0I?t=137>

<https://vimeo.com/248025147>

(Agarwal et al., 2018)
<http://aclweb.org/anthology/W18-6514>

Further Research Areas

- Multi/open domains
 - reusability, domain transfer
 - training from little data
 - pretraining with “generic” data
 - connecting task-oriented systems and chatbots
- Context dependency
 - understand/reply in context (grounding, speaker alignment)
- Incrementality
 - don't wait for the whole sentence to start processing
 - not much stuff going on at the moment, but would help
- Evaluation
 - checking if the system does well is actually non-trivial

Summary

- We're far from AI sci-fi dreams, but it still works a bit
 - dialogue is hard
- DSs have many forms & usage areas
 - **task-oriented vs. non-task-oriented**
 - **closed vs. open domain**
 - system vs. user initiative
- Main components: **ASR → NLU → DM → NLG → TTS**
 - implementation varies
 - sometimes things are joined together
- It's an active and interesting research topic!
- Next week: machine learning bits and pieces

Thanks

Contact us:

[https://ufaldsg.slack.com/
{odusek,hudecek}@ufal.mff.cuni.cz](https://ufaldsg.slack.com/{odusek,hudecek}@ufal.mff.cuni.cz)
Troja N231/N233 (by agreement)

Next Tuesday:
Labs 9:50am
Lecture 10:40am

Get the slides here:

<http://ufal.cz/npfl099>

References/Inspiration/Further:

Apart from materials referred directly, these slides are based on slides and syllabi by:

- Pierre Lison (Oslo University): <https://www.uio.no/studier/emner/matnat/ifi/INF5820/h14/timeplan/index.html>
- Oliver Lemon & Verena Rieser (Heriot-Watt University): <https://sites.google.com/site/olemon/conversational-agents>
- Filip Jurčiček (Charles University): <https://ufal.mff.cuni.cz/~jurcicek/NPFL099-SDS-2014LS/>
- Milica Gašić (University of Cambridge): <http://mi.eng.cam.ac.uk/~mg436/teaching.html>
- David DeVault & David Traum (Uni. of Southern California): <http://projects.ict.usc.edu/nld/cs599s13/schedule.php>
- Luděk Bártek (Masaryk University Brno): <https://is.muni.cz/el/1433/jaro2018/PA156/um/>
- Gina-Anne Levow (University of Washington): <https://courses.washington.edu/ling575/>