1. Introduction

Ondřej Dušek & Ondřej Plátek & Jan Cuřín

ufal.cz/npfl123

19. 2. 2019
Organizational
NPFL123 – 2/2 Z+ZK – 5 Credits

• Lecture (Tue 10:40am S11) + labs (Wed 9:00am SU1)
• Lecture: intro, theory
• Labs: practical examples, hands-on exercises
• To pass the course:
  • Written exam – freeform questions, as covered by the lectures
  • Lab exercises (best to come there)
  • Small personal projects (make your own system, by agreement)
• Slides, news etc. at ufal.cz/npfl123
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)

Ondřej Plátek: labs
- founded Oplatai
- R&D in startups and Apple Siri team
- MSc. at ÚFAL 2014 on speech recognition

Jan Cuřín: speech lectures, dialog authoring tools
- IBM – Manager at IBM Prague AI R&D Lab – IBM Watson Assistant Service
- PhD at ÚFAL in 2006 (machine translation)
- dialog systems and applications, speech recognition, machine translation
Course Syllabus (1)

1. Introduction (today)
2. What happens in a dialogue?
3. Dialogue system data & how to evaluate
4. Assistants (Alexa, Siri, Google etc.), question answering
5. Dialogue authoring/tooling systems
6. Language understanding
7. Dialogue state tracking
8. Dialogue management
9. Language generation
Course Syllabus (2)

10. Automatic speech recognition
11. Speech synthesis
12. Chatbots
Recommended Reading

• There’s nothing ideal (active research topic!)

Primary (brief):
Jurafsky & Martin: Speech & Language processing. 3rd ed. draft 2018, Chap. 24-25
(https://web.stanford.edu/~jurafsky/slp3/)

Other (see also website):
• Janarthanam: Hands-On Chatbots and Conversational UI Development. Packt 2017
(http://www.speech.kth.se/~gabriel/thesis/chapter2.pdf)
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
“AI”: sci-fi vs. reality

• Lots of talk about AI now
• Hype around Siri/Alexa/Google
• Sci-fi expectations – AI-complete
  • Star Trek – know-it-all (youtu.be/1ZXugicgn6U?t=3)
  • 2001 Space Oddysssey –mutiny (youtu.be/9W5Am-a_xWw)
  • Her – personality (youtu.be/6QRvTv_tpw0?t=27)
• We’re not there – probably for long
  • main bottleneck: understanding
    (not speech comprehension, meaning!)
  • … more like Red Dwarf talkie toaster (youtu.be/LRq_SAuQDec?t=71)
Real Dialogue System Examples

• “Smart speakers” / conversational assistants
  • Alexa, Siri, Google (+ others)
• Phone systems
  • even basic ones (DMTF)
  • voice-based ones deployed now
• Computer games
• Chatbots
• Assistive technologies
• Research systems (skylar.speech.cs.cmu.edu)
Example: Google Assistant

• Handling call for a client (Google IO 2018 demo)
  • very natural speech
  • show’s what’s possible now in a limited domain
  • redirects to a human if it can’t handle the shop’s request

https://youtu.be/d40jgFZ5hXk
Example: Alana Chatbot (Heriot-Watt University)

- Open-domain
Possible Areas of Use

• Information retrieval
  • Let’s go / Buses: http://www.speech.cs.cmu.edu/letsgo/example.html
  • CLASSiC / Restaurants: https://youtu.be/lHfLr1MF7DI

• Navigation
  • SpaceBook: https://youtu.be/qQZnwrOyeTE?t=65

• Cars

• Task completion / home automation

• Assistive technologies
  • therapy, elderly care

• Language learning

• Robotics
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!*

• **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
• “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

• **open-domain**
  • “responds to anything” – mostly chitchat
Application Areas

• **phone** (traditional)
  - users call a phone number & a dialogue system picks up

• **apps**
  - assistant apps for your phone/computer
  - companions (Xiaolce)

• **smart speakers**
  - home automation, assistants (Alexa/Google Home)

• **appliances**
  - voice operated TVs
  - other devices connect to smart speakers

Application Areas

• cars
  • hands-free car-specific functions
  • Android Auto, Apple CarPlay, vendor-specific solutions

• web
  • search assistants (IKEA)
  • Facebook Messenger chatbots

• embodied (robots)
  • information assistants

• virtual characters
  • computer games
Modes of Communication

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

• **voice**
  - more difficult, but can be more natural
  - easy to deploy over the phone

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

• main loop:
  • voice → text
  • text → meaning
  • meaning → reaction
  • reaction → text
  • text → voice

• access to backend

• multimodal systems: additional components

speech recognition

what’s the time?

language understanding

request(time)

dialogue management

inform(time=15:30)

language generation

it is currently 3:30 p.m.

speech synthesis

backend
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, distance, 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
DOI 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 letters

• Limited domain: use of language models
  • some words/phrases more likely than others
  • previous context can be used
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* …)
  • 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")
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)

https://www.w3.org/2004/Talks/05-www2004-voice/dialog.png
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, ACL-SRW 2002
White, ENLG 2011
Wen, INLG 2016
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

tei k bas nəmbə θri: æt fɔːr əʊ fɔːr ei ɛm
Speech Synthesis

• TTS Methods:
  • **Formant**-based: phoneme-specific frequencies
    • oldest, not very natural, but works on limited hardware
  • **Concatenative**
    • record a single person, cut into phoneme transitions (diphones), glue them together
  • **Hidden Markov Models**
    • phonemes in context modelled as hidden Markov models
    • Model parameters estimated from data (machine learning)
  • **Neural networks**
    • HMMs swapped for a recurrent neural network
    • can go directly from text, no need for phoneme conversion

https://en.wikipedia.org/wiki/MBROLA
http://homepages.inf.ed.ac.uk/jyamagis/
https://google.github.io/tacotron/
https://youtu.be/9Avlhm55kv0?t=379
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
Dialogue Systems Research

• Multi/open domains
  • reusability, domain transfer
• Joint models (“end-to-end”, all in one neural network)
• Multimodality
  • adding video (input/output)
• Context dependency
  • understand/reply in context (grounding, speaker alignment)
• Incrementality
  • don’t wait for the whole sentence to start processing
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
• It’s an active and interesting research topic!
• Next week: what happens in dialogue and why it’s hard
Thanks

Contact me:
  odusek@ufal.mff.cuni.cz
  room 424 (but email me first)

Get the slides here:
  http://ufal.cz npfl123

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
- 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/

Come to labs!
Tomorrow 9:00 SU1

Talk to me about
Ph.D./MSc./BSc. theses!

NPFL123 L1 2019