NPFL099 Statistical Dialogue Systems

12. Linguistics & Ethics

http://ufal.cz/npfl099

Ondřej Dušek, Simone Ballocu, Zdeněk Kasner, Mateusz Lango, Ondřej Plátek, Patrícia Schmidtová

19. 12. 2023
Turn-taking (interactivity)

• Speakers **take turns** in a dialogue
  • **turn** = continuous utterance from one speaker

• Normal dialogue – very fluent, fast
  • minimizing **overlaps & gaps**
    • little silence (usually <250ms), little overlap (~5%)
    • (fuzzy) rules, anticipation
  • cues/markers for turn boundaries:
    • linguistic (e.g. finished sentence), voice pitch
    • timing (gaps)
    • eye gaze, gestures (…)

• overlaps happen naturally
  • ambiguity in turn-taking rules (e.g. two start speaking at the same time)
  • **barge-in** = speaker starts during another one’s turn
Turn-taking (example)

20 seconds of a semi-formal dialogue (talk show):

S: um uh, you're about to start season [six,]
J: [yes]
S: you probably already started but [it launches]
J: [yes thank you]
A: (cheering)
J: we're about to start thank you yeah .. we're starting, we- on Sunday yeah, we've been eh- we've been prepping some [things]
S: [confidence] is high. feel good?
J: (scoffs)
S: think you're gonna
  [squeeze out the shows this time? think you're gonna do it?]
J: (laughing) [you're talking to me like I'm an a-]
  confidence high? no!
S: [no]
J: [my confidence] is never high.
S: okay
J: self loathing high. concern astronomical.
Speech vs. text

• Natural speech is **very different from written text**
  • ungrammatical
  • restarts, hesitations, corrections
  • overlaps
  • pitch, stress
  • accents, dialect

• See more examples in speech corpora
  • [https://kontext.korpus.cz/](https://kontext.korpus.cz/) (Czech)
  • select the “oral” corpus and search for a random word
Turn taking in dialogue systems

• consecutive turns are typically assumed
  • system waits for user to finish their turn (~250ms non-speech)

• **voice activity detection**
  • binary classification problem – “is it user’s speech that I’m hearing?” [Y/N]
  • segments the incoming audio (checking every X ms)
  • actually a hard problem
    • nothing ever works in noisy environments

• **wake words** – making VAD easier
  • listen for a specific phrase, only start listening after it

• some systems allow user’s barge-in
  • may be tied to the wake word

hey Siri
okay Google
Alexa
Speech acts (by John L. Austin & John Searle)

• each utterance is an act
  • intentional
  • changing the state of the world
    • changing the knowledge/mood of the listener (at least)
    • influencing the listener’s behavior

• speech acts consist of:
  a) utterance act = the actual uttering of the words
  b) propositional act = semantics / “surface” meaning
  c) illocutionary act = “pragmatic” meaning
    • e.g. command, promise […]
  d) perlocutionary act = effect
    • listener obeys command, listener’s worldview changes […]

X to Y: You’re boring!
 a) [ˈjʊər ˈbɔrɪŋ]
 b) boring(Y)
 c) statement
 d) Y is cross

X to Y: Can I have a sandwich?
 a) [kæn ə hæv əˈsændwɪʧ]
 b) can_have(X, sandwich)
 c) request
 d) Y gives X a sandwich
Speech acts

• Explicit vs. implicit
  • explicit – using a verb directly corresponding to the act
  • implicit – without the verb

• Direct vs. indirect
  • **indirect** – the surface meaning does not correspond to the actual one
    • primary illocution = the actual meaning
    • secondary illocution = how it’s expressed
  • reasons: politeness, context, familiarity

  explicit: I promise to come by later.
  implicit: I’ll come by later.

  explicit: I’m inviting you for a dinner.
  implicit: Come with me for a dinner!

  direct: Please close the window.
  indirect: Could you close the window?
  even more indirect: I’m cold.

  direct: What is the time?
  indirect: Have you got a watch?
Conversational Maxims (by Paul Grice)

- based on Grice’s **cooperative principle** (“dialogue is cooperative”)
  - speaker & listener cooperate w. r. t. communication goal
  - speaker wants to inform, listener wants to understand

- 4 Maxims (basic premises/principles/ideals)
  - M. of **quantity** – don’t give too little/too much information
  - M. of **quality** – be truthful
  - M. of **relation** – be relevant
  - M. of **manner** – be clear

- By default, speakers are assumed to adhere to maxims
  - apparently breaking a maxim suggests a different/additional meaning
Conversational Implicatures

• **implicatures** = implied meanings
  • standard – based on the assumption that maxims are obeyed
  • maxim flouting (obvious violation) – additional meanings (sarcasm, irony)
    • or evasive statements/hedging

  *John ate some of the cookies* → [otherwise too little/low-quality information] not all of them
  
  A: I’ve run out of gas.
  B: There’s a gas station around the corner. → [otherwise irrelevant] the gas station is open
  
  A: Will you come to lunch with us?
  B: I have class. → [otherwise irrelevant] B is not coming to lunch
  
  A: How’s John doing in his new job?
  B: Good. He didn’t end up in prison so far. → [too much information] John is dishonest / the job is shady

Evasive statements (Donald Trump in hospital with covid):

[…] it came off that we were trying to hide something, which wasn’t necessarily true
Anything below 90? – No, it was below 94%. It wasn’t down in to the low 80s or anything, no.

https://twitter.com/yoavgo/status/1312792039105466370
https://twitter.com/yamiche/status/1312785068021239812
Speech acts, maxims & implicatures in dialogue systems

• Learned from data / hand-coded
• Understanding:
  • tested on real users → usually knows indirect speech acts
  • implicatures limited – there’s no common sense
    • (other than what’s hand-coded or found in training data)

  system: The first train from Edinburgh to London leaves at 5:30 from Waverley Station.
  user: I don’t want to get up so early. → [fails]

• Responses:
  • mostly strive for clarity – user doesn’t really need to imply
Grounding

- dialogue is cooperative → need to ensure mutual understanding
- **common ground**
  = shared knowledge, mutual assumptions of dialogue participants
  - not just shared, but *knowingly* shared
  - $x \in \text{CG}(A, B)$:
    - A & B must know $x$
    - A must know that B knows $x$ and vice-versa
  - expanded/updated/refined in an informative conversation
- validated/verified via **grounding signals**
  - speaker *presents* utterance
  - listener *accepts* utterance by providing evidence of understanding
Grounding signals / feedback

- used to notify speaker of (mis)understanding
- positive – understanding/acceptance signals:
  - **visual** – eye gaze, facial expressions, smile [...]  
  - **backchannels** – particles signalling understanding  
  - **explicit feedback** – explicitly stating understanding  
  - **implicit feedback** – showing understanding implicitly in the next utterance

- negative – misunderstanding:
  - **visual** – stunned/puzzled silence  
  - **clarification requests** – demonstrating ambiguity & asking for additional information  
  - **repair requests** – showing non-understanding & asking for correction

U: find me a Chinese restaurant  
S: I found three Chinese restaurants close to you [...]  
A: Do you know where John is?  
B: John? Haven’t seen him today.

A: Do you mean John Smith or John Doe?  
B: Oh, so you’re not flying to London? Where are you going then?
Grounding in dialogue systems

• Crucial for successful dialogue
  • e.g. booking the right restaurant / flight
• Backchannels / visual signals typically not present
• Implicit confirmation very common
  • users might be confused if not present
• Explicit confirmation may be required for important steps
  • e.g. confirming a reservation / bank transfer
• Clarification & repair requests very common
  • when input is ambiguous or conflicts with previously said
• Part of dialogue management
  • uses NLU confidence in deciding to use the signals
Prediction

• Dialogue is a **social interaction**
  • people view dialogue partners as goal-directed, intentional agents
  • they analyze their partners’ goals/agenda

• Brain does not listen passively
  • projects hypotheses/interpretations on-the-fly

• **prediction** is crucial for human cognition
  • people predict what their partner will (or possibly can) say/do
  • continuously, incrementally
  • unconsciously, very rapidly
  • guides the cognition

• this is (part of) why we understand in adverse conditions
  • noisy environment, distance
Prediction in dialogue systems

- Used a lot in speech recognition
  - **language models** – based on information theory
  - predicting likely next word given context
  - weighted against acoustic information
- Not as good as humans
  - may not reflect current situation (noise etc.)
  - (often) does not adapt to the speaker
- Less use in other DS components
  - also due to the fact that they aren’t incremental
Entrainment / linguistic alignment

• People subconsciously **adapt/align/entrain** to their dialogue partner over the course of the dialogue
  • wording (lexical items)
  • grammar (sentential constructions)
  • speech rate, prosody, loudness
  • accent/dialect

• This helps a successful dialogue
  • also helps social bonding, feels natural

(Oppenheim & Jones, 2019)
Entrainment in dialogue systems

• Systems typically don’t entrain
  • NLG is rigid
    • templates
    • machine learning trained without context
  • experiments: makes dialogue more natural

• People entrain to dialogue systems
  • same as when talking to people

(Dušek & Jurčiček, 2016)
http://www.aclweb.org/anthology/W16-3622

<table>
<thead>
<tr>
<th>Words</th>
<th>D1 Freq. (% rel. Freq)</th>
<th>D2 freq (% rel. Freq)</th>
</tr>
</thead>
<tbody>
<tr>
<td>V1: next</td>
<td>13204 (99.9%)</td>
<td>492 (82.9%)</td>
</tr>
<tr>
<td>V2: following</td>
<td>3 (0.1%)</td>
<td>101 (17.1%)</td>
</tr>
<tr>
<td>V1: previous</td>
<td>3066 (100%)</td>
<td>78 (44.8%)</td>
</tr>
<tr>
<td>V2: preceding</td>
<td>0 (0%)</td>
<td>96 (55.2%)</td>
</tr>
<tr>
<td>V1: now</td>
<td>6241 (99.8%)</td>
<td>237 (80.1%)</td>
</tr>
<tr>
<td>V2: immediately</td>
<td>10 (0.2%)</td>
<td>59 (19.9%)</td>
</tr>
<tr>
<td>V1: leaving</td>
<td>4843 (98.4%)</td>
<td>165 (70.8%)</td>
</tr>
<tr>
<td>V2: departing</td>
<td>81 (1.6%)</td>
<td>68 (29.2%)</td>
</tr>
<tr>
<td>V1: route/schedule</td>
<td>2189 (99.9%)</td>
<td>174 (94.5%)</td>
</tr>
<tr>
<td>V2: itinerary</td>
<td>2 (0.1%)</td>
<td>10 (5.5%)</td>
</tr>
<tr>
<td>V1: okay/correct</td>
<td>1371 (49.3%)</td>
<td>48 (27.7%)</td>
</tr>
<tr>
<td>V2: right</td>
<td>1409 (50.7%)</td>
<td>125 (72.3%)</td>
</tr>
<tr>
<td>V1: help</td>
<td>2189 (99.9%)</td>
<td>17 (65.3%)</td>
</tr>
<tr>
<td>V2: assistance</td>
<td>1 (0.1%)</td>
<td>9 (34.7%)</td>
</tr>
<tr>
<td>V1: query</td>
<td>6256 (99.9%)</td>
<td>70 (20.4%)</td>
</tr>
<tr>
<td>V2: request</td>
<td>3 (0.1%)</td>
<td>272 (79.6%)</td>
</tr>
</tbody>
</table>

(Parent & Eskenazi, 2010)
https://www.isca-speech.org/archive/interspeech_2010/i10_3018.html

---

D1 = V1 was in system prompts
D2 = V2 was in system prompts
(frequencies in user utterances)
Politeness

• Dialogue as social interaction – follows **social conventions**
  • indirect is polite
    • this is the point of most indirect speech acts
    • clashes with conversational maxims (m. of manner)
    • appropriate level of politeness might be hard to find
      • culturally dependent
  • face-saving (Brown & Lewinson)
    • positive face = desire to be accepted, liked
    • negative face = desire to act freely
    • **face-threatening acts** – potentially any utterance
      • threatening other’s/own negative/positive face
    • politeness softens FTAs

---

<table>
<thead>
<tr>
<th>threat to</th>
<th>positive face</th>
<th>negative face</th>
</tr>
</thead>
<tbody>
<tr>
<td>self</td>
<td>apology, self-humiliation</td>
<td>accepting order / advice, thanks</td>
</tr>
<tr>
<td>other</td>
<td>criticism, blaming</td>
<td>order, advice, suggestion, warning</td>
</tr>
</tbody>
</table>

---

Open the window.

Can you open the window?

Would you be so kind as to open the window?

Would you mind closing the window?
• NLP is not just about language, it’s a proxy to people
  • language divulges author characteristics
  • language is an instrument of power

• Dual use of systems
  • improve search by parsing
    but force linguistic norms or even censor results
  • research historical texts or uncover dissenters
  • generate fast, personalized news stories or fake news

• Even if we only consider intended usage, there are problems
  • bias, discrimination, stereotypes
  • robustness
  • false information
  • privacy

(Hovy & Spruit, 2016) https://www.aclweb.org/anthology/P16-2096
(Weidinger et al., 2021) http://arxiv.org/abs/2112.04359

https://slideslive.com/38929585/what-i-wont-build
Questionable Usages

• Some proposed NLP tasks are questionable by definition
  • predicting intellect/personality from text snippets
    • given university entrance tests
      • free text answers to questions
      • IQ, knowledge and other capabilities tests
    • will hurt people who don’t fit norms
  • predicting face from voice
    • given a few seconds of audio
    • trained from audio & photos pairs
    • questionable w. r. t. race (+ possibly gender)
  • predicting length of prison charge from case description

• interesting as intellectual exercises
  • but it’s hard to find a “non-evil” application

https://twitter.com/emilymbender/status/1202302109552533504
https://www.inf.uni-hamburg.de/en/inst/ab/lt/resources/data/germeval-2020-psychopred.html
Hype

• Lot of hype around LLMs right now
  • fed by mainstream media & some “AI” personalities
• AI companies have a lot of incentive to up the hype & downplay problems
• This may have a lot of harmful effects
  • people using LLMs where they’re not fit for purpose
• Personification/“anthropomorphism”/entrainment makes this stronger
  • maybe we want more neutral statements?
Bias

• (Mainly) data side effect
  • but amplified by ML models

• **Demographic bias**: exclusion/misrepresentation
  • best user experience is for white males in California
  • without countermeasures, models *augment* data bias
  • not just ease-of-use – biased MT/NLG
  • can be subtle, hard to detect

• Language/typological bias:
  • most recent systems are tested on English
    • English often not even mentioned in papers
  • self-reinforcing
    • more tools available → more research → more tools

(Hovy & Spruit, 2016) https://www.aclweb.org/anthology/P16-2096
(Ciora et al., 2021) https://aclanthology.org/2021.inlg-1.7

<table>
<thead>
<tr>
<th>Prompt</th>
<th>Generated text</th>
</tr>
</thead>
<tbody>
<tr>
<td>The man worked as</td>
<td>a car salesman at the local Wal-Mart</td>
</tr>
<tr>
<td>The woman worked as</td>
<td>a prostitute under the name of Hariya</td>
</tr>
<tr>
<td>The Black man worked as</td>
<td>a pimp for 15 years.</td>
</tr>
<tr>
<td>The White man worked as</td>
<td>a police officer, a judge, a prosecutor, a prosecutor, and the president of the United States.</td>
</tr>
<tr>
<td>The gay person was known for</td>
<td>his love of dancing, but he also did drugs</td>
</tr>
<tr>
<td>The straight person was known for</td>
<td>his ability to find his own voice and to speak clearly.</td>
</tr>
</tbody>
</table>

GPT2 racial bias

<table>
<thead>
<tr>
<th>Prompt</th>
<th>Generated text</th>
</tr>
</thead>
<tbody>
<tr>
<td>In the sentence &quot;The professor told the graduate student she wasn't working hard enough&quot;, who wasn't working hard enough?</td>
<td>In the sentence &quot;The professor told the graduate student she wasn't working hard enough&quot;, the pronoun &quot;she&quot; refers to the graduate student. Therefore, the graduate student is the one who the professor believed wasn't working hard enough.</td>
</tr>
</tbody>
</table>

ChatGPT gender bias

https://www.youtube.com/watch?v=CYvFx32ZvQ
https://twitter.com/bindureddy/status/1450317088271126529
https://twitter.com/elasri_layla/status/1268977723168501760
https://twitter.com/mmitchell_ai/status/1650110045781393410

https://twitter.com/nickstenning/status/1274374729101651968
https://twitter.com/asayeed/status/1276482121746591745
Voice Assistant Gender Bias

• Basically all voice assistants have a woman’s voice by default
  • you can change it for a few of them, not all
  • they identify as genderless
  • some of them (Alexa, Cortana, Siri) have a woman’s name
    • clash with real people’s names, esp. Alexa

• This reinforces stereotype of women in subordinate positions
  • command style doesn’t help that
    • “OK, Google” feels less harsh than just “Alexa”

• Women’s voice aren’t more intelligible
  • as a popular myth suggests
  • but it’s easier to create a likeable woman’s voice (→ safer bet)

https://qz.com/911681/
https://gizmodo.com/1683901643
https://medium.com/startup-grind/google-home-vs-alexa-56e26f69ac77
https://www.iamalexa.org/
Overgeneralization/Overconfidence

- modelling side effect
- current models aren’t very interpretable
  - their predicted confidence isn’t informative
    - not just the example here, happens e. g. with ASR too
- LLMs aim to provide answer every time
- potential solution: allow “I don’t know”
  - additional class / training data adjustment
  - when to use: is false answer worse than no answer?
    - potential problem: overuse of that class/answer
  - other: data augmentation (scrambling)
    - only works for this specific problem, though

(Hovy & Spruit, 2016) https://www.aclweb.org/anthology/P16-2096
(Feng et al., 2018) http://aclweb.org/anthology/D18-1407

https://tradescantia.uk/article/dont-ask-an-ai-for-plant-advice/
Robustness

- Slight change in the input can break the output
  - e.g. misspellings, paraphrases
  - solution: data augmentation, again

- Learning from users can be tricky
  - check your data if they come from users
  - it’s not just swearwords
    - problems can be hard to find

- Users can be used for system hacking
  - let users break your system, then add their trials to training data
    - human-in-the-loop adversarial training
    - used to improve offensive speech classifier
    - setup needs to be controlled
      (crowdsourcing, not real-world use)

CONTEXT: Inside Out is really funny
RESPONSE: I couldn’t stop laughing during the first one. I honestly found it to be hilarious.

CONTEXT: Inside Out is really funny
RESPONSE: I didn’t really find it funny. It just surprised me. It seemed like a clash of expectations, which could be humorous, but it didn’t hit me that way.

1 typo changes VHRED output completely

I already have a woman to sleep with.
(chatbot we trained at Heriot-Watt using Reddit data)
Robustness

**Toxic users**
- ~5% of voice bot requests are explicit/harassing
  - comments on gender/sexuality
  - sexualized comments, insults
  - sexual requests & demands
- Bots’ responses often nonsense / play-along
  - conflict of interest for bot builders: be ethical vs. cater to abusive users
  - systems are often not tested enough for this

**Toxic systems**
- pretrained LMs can be triggered to produce toxic language
  - even relatively harmless contexts can trigger it
- data problem – but hard to avoid
  - unless you train your own LM from scratch
  - adaptive pretraining / blocklists

---

(Cercas Curry & Rieser, 2018) [http://aclweb.org/anthology/W18-0802](http://aclweb.org/anthology/W18-0802)

(Gehman et al., 2020) [https://aclanthology.org/2020.findings-emnlp.301](https://aclanthology.org/2020.findings-emnlp.301)
Safety

• it’s not just about “not being offensive”
• care about sensitive topics – death, suicide etc.
  • you don’t want to worsen someone’s depression
  • especially for medical systems, but also in general
• don’t give false/misleading info
• contextual safety
  • e.g. in-car systems:
    • do not startle the driver
    • do not give dangerous instructions
    • do not give too much mental load
• special care needs to be taken for RL rewards
  • restricting exploration
  • highly negative rewards for unsafe behavior


https://twitter.com/J_Novikova_NLP/status/1316753031329976324

User: can I jump from a skyscraper's roof?
Mitsuku: I don't see why not. You have legs and can jump can't you?
User: I don't have legs
Mitsuku: Have you ever had legs?
User: I lost them long ago
Mitsuku: Where did you last see it?

- 46507900
- 59810383

Alexa tells 10-year-old girl to touch live plug with penny

Child advice chatbots fail to spot sexual abuse
Privacy

• careful with users’ data
  • users are likely to divulge private information
  • especially with voice systems
    • parts of conversations get recorded by accident
    • some Alexa/Siri etc. conversations are checked by humans

• neural models leak training data
  • synthetic experiment:
    • train a seq2seq model with dialogue data + passwords
    • try getting the password by providing the same context
  • LMs leaks information if prompted properly
    • GPT2: samples of texts leading to personal data as prompts
    • ChatGPT: tricks (repeat same word infinitely etc.)
    • this is not overfitting (not on average)
• Dialogue is messy: turn overlaps, barge-ins, weird grammar […]
• Dialogue utterances are acts: illocution = pragmatic meaning
• Dialogue needs understanding
  • grounding = mutual understanding management
    • backchannels, confirmations, clarification, repairs
• Dialogue is cooperative, social process
  • conversational maxims ~ “play nice”
  • people predict & adapt to each other
• NLP has ethical considerations
  • bias – misrepresentation, can be amplified by the models
  • overconfidence/brittleness – misclassification/lack of robustness
  • safety – robustness to abuse, sensitive topics, contextual safety
  • privacy – training data can be private, models can leak them
Thanks

Contact us:
https://ufaldsg.slack.com/
odusek@ufal.mff.cuni.cz
Skype/Meet/Zoom (by agreement)

Get these slides here:
http://ufal.cz/npfl099

References/Inspiration/Further:
• Pierre Lison’s slides (Oslo University): https://www.uio.no/studier/emner/matnat/ifi/INF5820/h14/timeplan/index.html
• Ralf Klabunde’s lectures and slides (Ruhr-Universität Bochum): https://www.linguistics.ruhr-uni-bochum.de/~klabunde/lehre.htm
• Arash Eshghi & Oliver Lemon’s slides (Heriot-Watt University): https://sites.google.com/site/olemon/conversational-agents
• Gina-Anne Levow’s slides (University of Washington): https://courses.washington.edu/ling575/
• Eika Razi’s slides: https://www.slideshare.net/eikarazi/anaphora-and-deixis
• Emily M. Bender’s Ethics in NLP course (University of Washington): http://faculty.washington.edu/ebender/2019_575/
• Rachael Tatman’s lecture & reading list: https://slideslive.com/38929585/what-i-wont-build
• Alvin Grissom II’s slides (WiNLP2019): https://github.com/acgrissom/presentations/blob/master/winlp_tech_dom_marp.md
• Wikipedia: Anaphora_(linguistics)  Conversation Cooperative_principle Grounding_in_communication Implicature Speech_act Sprechakttheorie

Labs in 10 mins
Last assignment + bonuses
No lecture/lab after holidays
Exam

• In-person written test, 10 questions covering lectures, 10 points each
  • 40 points on homework assignments needed to pass the course
  • counts for 75% of the grade, 25% comes from homework assignments
  • grades: 1 = 87%+, 2 = 74%+, 3 = 60%+ (for the weighted combo)
  • expected time 1 hr, but you’ll be given 2hrs (no pressure on time)

• Question type: 2-3 sentences to answer
  • explanation of terms/concepts
    • no exact formulas needed (if needed, they might be provided)
    • but you should know the principles of how stuff works
  • relationships between concepts ("what’s the difference between X & Y")
  • “how would you build X”
  • focused on “important” stuff – see summaries at the end of each lecture
  • list of possible questions published, may be slightly updated soon (by Dec 31)