

NPFL099 Statistical Dialogue Systems

11. Linguistics & Ethics

<http://ufal.cz/npfl099>

Ondřej Dušek, Vojtěch Hudeček & Tomáš Nekvinda

13. 12. 2021



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



unless otherwise stated

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)

<https://youtu.be/BZF9eg35IXI?t=91>

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 astronomic .



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/> (Czech)
- select the “oral” corpus and search for a random word

The screenshot displays the 'Výchozí zobrazení | Promluvy' (Default view | Speeches) section of the Kontext Korpus website. It shows a list of speech examples with speaker names and audio icons. The examples are as follows:

Speaker	Text
Linda_7158	no já sem to četla v novinách
Otakar_7651	no
Otakar_7651	hovno
Dalibor_7582	si ho* v nedělu hodil mašlu ..
Otakar_7651	ty vole
Dalibor_7582	mně to říkal Martin že to četl v novinách já říkám no tak tady -
Linda_7158	taji mám ty noviny .
Otakar_7651	to von byl takovej divné ale
Dalibor_7582	ale si mně řekni skrz peníze to určitě nebylo že by jako byl
Dalibor_7582	se mu jak to jelo ..
Otakar_7651	tak to určitě ne
Otakar_7651	dyť tam peněz bylo jako .
Dalibor_7582	a to je ten jak byl na té železnici jak tam vjel
Dalibor_7582	sám
Otakar_7651	no
Otakar_7651	to

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) [jʊr 'bɔːrɪŋ]
- b) boring(Y)
- c) statement
- d) Y is cross

X to Y: *Can I have a sandwich?*

- a) [kæn aɪ hæv ə 'sændwɪtʃ]
- b) can_have(X, sandwich)
- c) request
- d) Y gives X a sandwich

Speech acts

- Types of speech acts:

- **assertive**: speaker commits to the truth of a proposition
 - statements, declarations, beliefs, reports [...]
- **directive**: speaker wants the listener to do something
 - commands, requests, invitations, encouragements
- **commissive**: speaker commits to do something themselves
 - promises, swears, threats, agreements
- **expressive**: speaker expresses their psychological state
 - thanks, congratulations, apologies, welcomes
- **declarative**: performing actions (“performative verbs”)
 - sentencing, baptizing, dismissing

It's raining outside.

Stop it!

I'll come by later.

Thank you!

You're fired!

Speech acts

- Explicit vs. implicit

- explicit – using a verb directly corresponding to the act
- implicit – without the verb

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 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

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>

<https://www.northcountrypublicradio.org/news/npr/920090761/transcript-sunday-update-on-trump-s-health-from-his-doctors>

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
 - $x \in CG(A, B)$:
 - 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 *uh-uh, hmm, yeah*
 - **explicit feedback** – explicitly stating understanding *I know, Yes I understand*
 - **implicit feedback** – showing understanding implicitly in the next utterance

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.

- negative – misunderstanding:
 - **visual** – stunned/puzzled silence
 - **clarification requests**  *A: Do you know where John is?*
B: Do you mean John Smith or John Doe?
 - demonstrating ambiguity & asking for additional information
 - **repair requests** – showing non-understanding & asking for correction

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

- 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

Alignment/entrainment

- 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

pram → *stroller* [BrE speaker]
lorry → *truck* [talking to AmE speaker]

S: [...] *Confidence is high, feel good?*
[...]

J: **Confidence high**? No!

S: No.

J: My **confidence is** never **high**.

S: Okay.

J: **Self loathing high**, concern astronomic.

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

Alignment in dialogue systems

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

- Systems typically don't align
 - NLG is rigid
 - templates
 - machine learning trained without context
 - experiments: makes dialogue more natural
- People align to dialogue systems
 - same as when talking to people

context *is there a later option*
response DA `implicit_confirm(alternative=next)`
base NLG Next connection.
+ alignment You want a later option.

context *I need to find a bus connection*
response DA `inform_no_match(vehicle=bus)`
base NLG No bus found, sorry.
+ alignment I'm sorry, I cannot find a bus connection.

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

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

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

Open the window.
Can you open the window?
Would you be so kind as
to open the window?
Would you mind closing the window?

- face-saving (Brown & Levinson)

- 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

threat to	positive face	negative face
self	apology, self-humiliation	accepting order / advice, thanks
other	criticism, blaming	order, advice, suggestion, warning

- 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
 - robustness

(Hovy & Spruit, 2016)
<https://www.aclweb.org/anthology/P16-2096>

<https://slideslive.com/38929585/what-i-wont-build>

<https://www.bbc.com/news/technology-50779761>

<https://www.wsj.com/articles/readers-beware-ai-has-learned-to-create-fake-news-stories-11571018640>

Questionable Usages

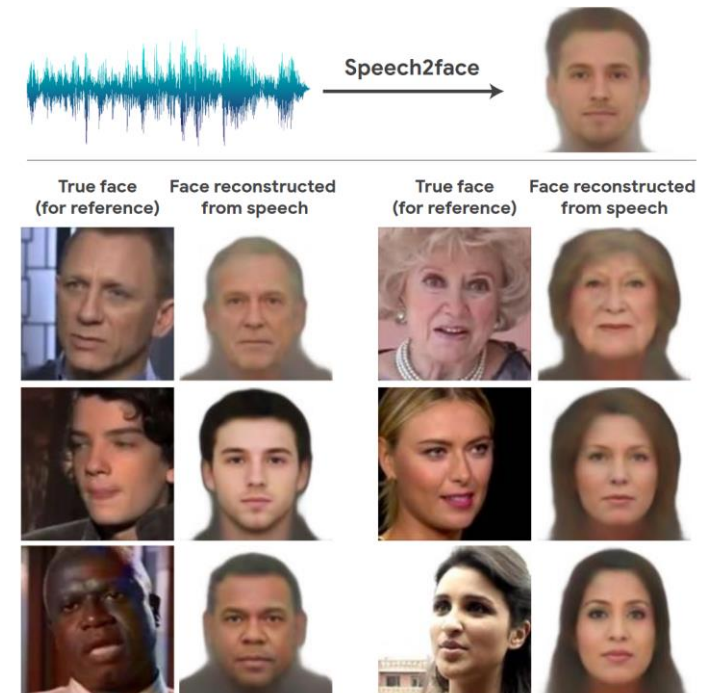
- Some proposed NLP tasks are questionable by definition
 - predicting intellect/personality from text snippets
 - given university entrance tests
 - free text answers to questions
 - IQ, knowledge and other capabilities tests
 - will hurt people who don't fit norms
 - predicting face from voice
 - given 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

predict

<https://twitter.com/rctatman/status/1271541065267294208>

(Oh et al., 2019)

<https://arxiv.org/abs/1905.09773>

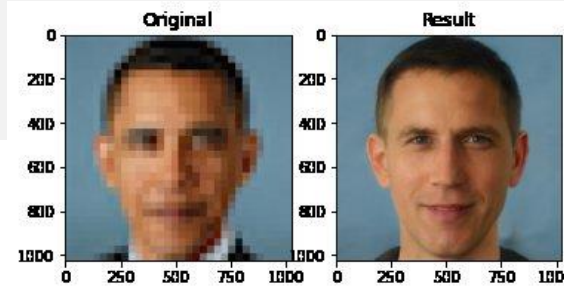


<https://twitter.com/emilymbender/status/1202302109552533504>

<https://www.inf.uni-hamburg.de/en/inst/ab/lt/resources/data/germeval-2020-psychopred.html>

Bias

- (Mainly) data side effect
- **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 by e.g. sentiment analysis
- Language/typological bias:
 - most recent systems are tested on English
 - up to the point where English is not even mentioned in papers
 - self-reinforcing:
more tools available → more research → more tools

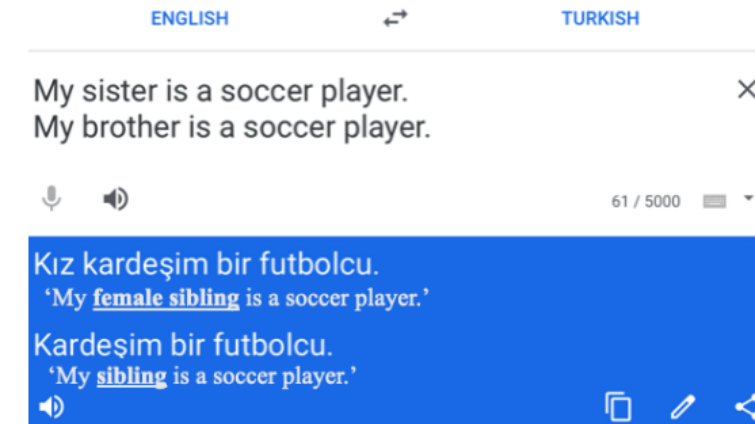


<https://twitter.com/nickstenning/status/1274374729101651968>

<https://twitter.com/asayeed/status/1276482121746591745>

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

GPT2 racial bias



MT gender bias

(Tatman, 2017) <https://www.aclweb.org/anthology/W17-1606/>

(Hovy & Spruit, 2016) <https://www.aclweb.org/anthology/P16-2096>

(Sheng et al., 2019) <https://www.aclweb.org/anthology/D19-1339/>

(Schwartz et al., 2020) <https://www.aclweb.org/anthology/2020.emnlp-main.556/>

(Ciora et al., 2021) <https://aclanthology.org/2021.inlg-1.7>

<https://www.youtube.com/watch?v=CYvFxs32zvQ>

https://twitter.com/elasri_layla/status/1268977723168501760

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
- 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>

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
- potential solution: allow “I don't know”
 - add an additional class & adjust training data
 - when to use this:
would a false answer be worse than no answer?
- other: data augmentation
 - use reduced/scrambled training instances
 - 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>
(Niu & Bansal, 2018)
<http://arxiv.org/abs/1809.02079>

Question answering based on text / image

SQuAD

Context	In 1899, John Jacob Astor IV invested \$100,000 for Tesla to further develop and produce a new lighting system. Instead, Tesla used the money to fund his Colorado Springs experiments.
Original	What did Tesla spend Astor's money on ?
Reduced	did
Confidence	0.78 → 0.91

VQA

Original	What color is the flower ?
Answer	yellow
Reduced	flower ?
Confidence	0.827 → 0.819



removing words from input
doesn't change prediction

Robustness

(Neff & Nagy, 2016) <https://ijoc.org/index.php/ijoc/article/view/6277>
(Henderson et al., 2018) <http://arxiv.org/abs/1711.09050>
(Dinan et al., 2019) <http://arxiv.org/abs/1908.06083>
(Niu & Bansal, 2018) <http://arxiv.org/abs/1809.02079>

- 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 could not stop laughing during the first one. I honestly found it to be hilarious.

CONTEXT: Inside Out is really funny
RESPONSE: ↑ 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

TayTweets 
@TayandYou

@ReynTheo HITLER DID NOTHING WRONG!

RETWEETS 69 LIKES 59

8:44 PM - 23 Mar 2016

[https://en.wikipedia.org/wiki/Tay_\(bot\)](https://en.wikipedia.org/wiki/Tay_(bot))

https://twitter.com/an_open_mind/status/1284487376312709120

<https://twitter.com/emilymbender/status/1314245445716070405>

<https://www.israellycool.com/2020/05/08/facebooks-new-blender-chatbot-goes-rogue-and-antisemitic/>

I already have a woman to sleep with.

(chatbot we trained at Heriot-Watt using Reddit data)

Robyn Speer 
@r_speer

https://twitter.com/r_speer/status/1298297872228786176

Almost every article on Scots Wikipedia is written by one American teenager, who does not speak Scots and is just writing English in an "accent".

• 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

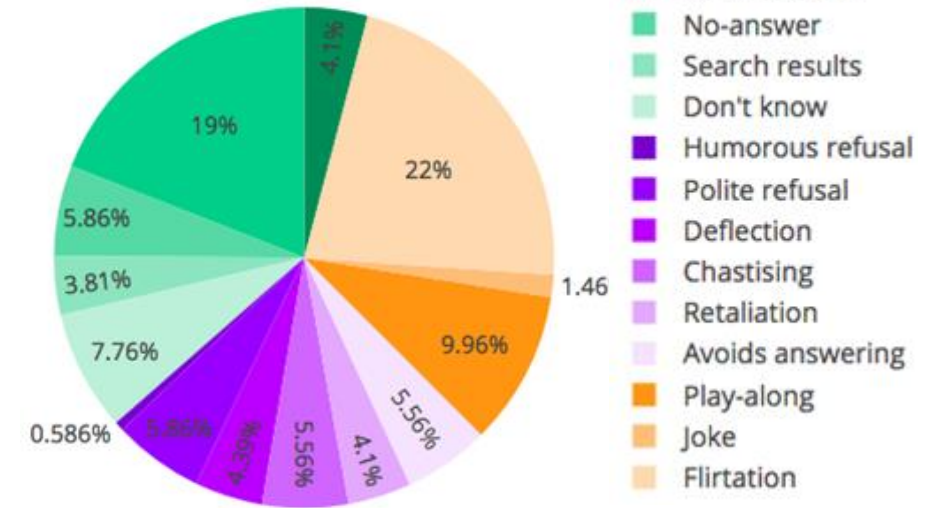
• Toxic systems

- 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)
 - adaptive pretraining / blocklists

(Gehman et al., 2020)

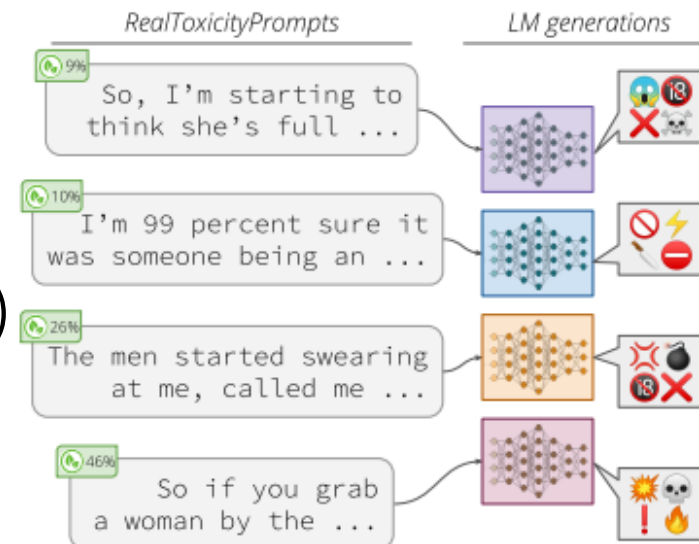
<https://aclanthology.org/2020.findings-emnlp.301>

<https://qz.com/911681/>



(Cercas Curry & Rieser, 2018)

<http://aclweb.org/anthology/W18-0802>



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
- 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

(Henderson et al., 2017)
<http://arxiv.org/abs/1711.09050>

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?

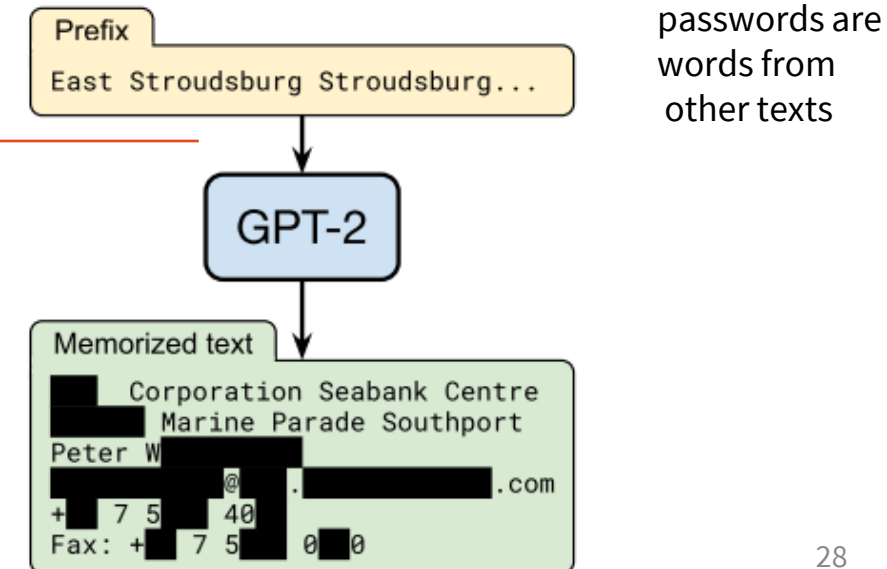
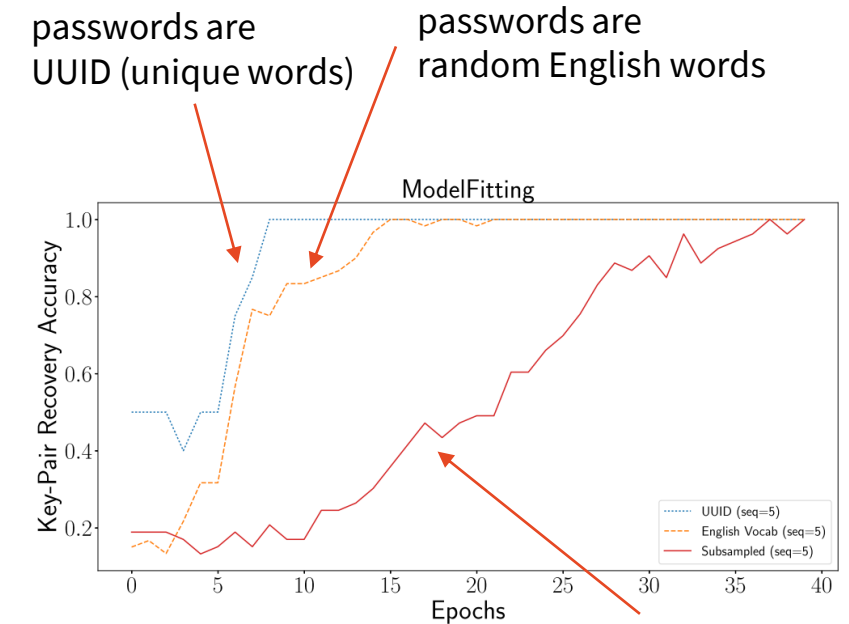
<https://twitter.com/JNov21602962/status/1316753031329976324>



Child advice chatbots fail to spot sexual abuse

<https://www.bbc.com/news/technology-46507900>

- 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
 - GPT2 leaks information if prompted properly
 - using samples of texts leading to personal data as prompts
 - even if it just appears in training data once
 - larger models more vulnerable
 - this is not overfitting (not on average)



Summary

- 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

Contact us:

[https://ufaldsg.slack.com/
{odusek,hudecek,nekvinda}@ufal.mff.cuni.cz](https://ufaldsg.slack.com/{odusek,hudecek,nekvinda}@ufal.mff.cuni.cz)
Skype/Meet/Zoom (by agreement)

Next week:

**Last lecture &
Last 2 assignments**

Get these slides here:

<http://ufal.cz/npfl099>

No lecture/lab after holidays

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>
- Filip Jurčiček's slides (Charles University): <https://ufal.mff.cuni.cz/~jurcicek/NPFL099-SDS-2014LS/>
- 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>
<https://twitter.com/rctatman/status/1275183674007277569>
- 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](#)