NPFL123 Dialogue Systems

2. What happens in a dialogue?

https://ufal.cz/npfl123

Ondřej Dušek, Mateusz Lango, Simone Balloccu & Jan Cuřín

28. 2. 2024
How do you “define” dialogue?

• Spoken/written conversational (interactive, collaborative) communication between two or more people

• **verbal** + (possibly) non-verbal
  • can be multimodal (language + gestures, pitch, expressions…)

• **collaborative**, social
  • participants aim at communicative goal(s)
  • involves inference about intended meanings

• **practical**, related to actions

• **interactive**, incremental, messy!

Dialogue systems – simpler than that
Describing a dialogue

- Levels of linguistic description
  - phonetics / phonology – sounds
  - morphology – word forms
  - syntax – sentence structure
  - semantics – sentence (propositional) meaning
  - pragmatics – meaning in context, communication goal

- This lecture is (a lot) about pragmatics
  (I don’t remember it well)

[Sound examples and linguistic structures are shown in the diagram.]

http://lindat.mff.cuni.cz/services/udpipe/
http://cohort.inf.ed.ac.uk/amreager.html
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
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] (nods)
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 .

https://youtu.be/BZF9eg35IXI?t=91
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
Voice activity detection

- Overlapping windows of ~30ms + binary classifier
- Features – actually similar to speech recognition itself
  - energy (loudness)
  - autocorrelation
  - checking for fundamental voice frequency
  - MFCCs (mel frequency spectrum)
  - deltas (trends over time)
- Onset is easier to detect than end of speech
  - they’re louder, more pronounced
  - hard to detect speech towards the system vs. someone else
    - that’s why wake words are used
    - how long can pauses/hesitations be?
- Postprocessing
  - smoothing out short-term errors
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 [...]
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!*

https://www.npr.org/2022/02/15/1080829813/priest-resigns-baptisms
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

https://youtu.be/IJEaMtNN_dM
Conversational Implicatures

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

*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
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 feedback/evidence**
  • speaker *presents* utterance
  • listener *accepts* utterance by providing evidence of understanding

• information added to common ground only after acceptance
Grounding evidence / feedback

**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
- **implicit / explicit repairs** – denying / presenting alternative
- **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: Are you going to London?
B: I’m going to Edinburgh.

A: Do you know where John is?
B: John Smith or John Doe?

Oh, so you’re not flying to London? Where are you going then?
Grounding (example)

S: um uh , you're about to start season [six ,]  
J: [yes]  
S: you probably already started but [it launches]  
J: [yes thank you] (nods)  
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  
[J: [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 .
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
Deixis

- **deixis** = “pointing” – relating between language & context/world
  - this is very important in dialogue
  - dialogue is typically set/situated in a specific context

- **deictic expressions** = words/grammar expressing deixis
  - their meaning depends on the context
    - who is talking, when, where
  - pronouns: *I, you, him, this*
  - verbs: tense & person markers: *goes [3rd ps. sg.], went [past]*
  - adverbs: *here, now, yesterday*
  - other (lexical meaning): *come / go [=here/away]*
  - non-verbal (gestures, gaze…)*
**Deixis**

- (typically) **egocentric**: *I – here – now* is the center (**origo**)

- main types of deixis:
  - **personal** – *I/me/you/she…*
  - **temporal** (time) – *now, yesterday, later, on Monday…*
  - **local** (space) – *here, there…*

- other:
  - **social** (politeness)
    - formal/informal address (Cz. *ty/vy*, Ger. *du/Sie*), honorifics in Asian languages
  - **discourse/textual**
    - referring to words/portions of texts – *next chapter, how do you spell that?*

https://en.wikipedia.org/wiki/Deixis
https://glossary.sil.org/term/discourse-deixis
Anaphora/Coreference

- expression referring to something mentioned in context
  - anaphora = referring back
  - cataphora = referring forward
- avoiding repetition, faster expression
- can refer to basically anything
  - objects/persons/events
  - qualities
  - actions/full sentences/portions of text
- used frequently in dialogue
- may be ambiguous

Susan dropped the plate. It shattered.

His friends describe John as smart and hard-working.

I don’t like it as much as he does.

Her dress is green. So is mine.


Bill stands next to John. He is tall.
Bill tickled John. He squirmed.

(Smaby, 1978)
https://link.springer.com/chapter/10.1007/978-94-009-9775-2_2
• systems typically assume a **single user**
  • this makes personal deixis much easier

• most systems are aware of time, location is more complicated
  • pronouns are often avoided – clearer, although less natural

• coreference resolution – separate problem
  • a whole area of research, specific resolution systems developed
  • some dialogue systems don’t include it, some do, sophistication varies
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
Information theory: dialogue is information transfer

- **communication channel** – speaker to listener (in the given situation)
- **entropy** – expected value of information conveyed (in bits)

\[
H(\text{text}) = - \sum_{\text{word} \in \text{text}} \frac{\text{freq(word)}}{\text{len(text)}} \log_2 \left( \frac{\text{freq(word)}}{\text{len(text)}} \right)
\]

Over vocabulary

Plays well with the social interaction perspective

- people tend to **use all available channel capacity**
  - limiting factors: noise, listener’s hearing ability, mental capacity
- people tend to **spread information evenly**
  - words carrying more information are emphasized
• how hard it is to guess the next word in the sentence?
  • given preceding context (n-gram)
  • related to Shannon entropy, but may differ
    • typically much lower than Shannon entropy
  • better estimate of prediction difficulty
    • although humans work with “unlimited” preceding context and reevaluate using following context

\[
H_{\text{cond}}(\text{text}) = - \sum_{(c,w) \in \text{text}} \frac{\text{freq}(c,w)}{\text{len}(\text{text})} \log_2 \left( \frac{\text{freq}(c,w)}{\text{freq}(c)} \right)
\]

\begin{align*}
<s> \text{The cat sat on the mat} & . \\
P(\text{cat} | <s> \text{The}) & \\
P(\text{sat} | \text{the cat}) & \\
P(\text{on} | \text{the cat sat}) & \\
P(\text{the} | \text{the cat sat on}) & 
\end{align*}

\begin{align*}
\text{# of times word } w & \text{ occurs after context } c \\
\text{context (preceding n-gram)} & \\
\text{word} & \\
\text{total # of times context } c & \text{ occurs, with or without word } w \\
\text{means # of n-grams here (not just words)} & 
\end{align*}
Prediction in dialogue systems

- Used a lot in speech recognition
  - **language models** (not necessarily LLMs) – based on information theory
    - statistical, trained on a text corpus (bunch of texts)
    - 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
Entrainment/adaptation

- 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

(2018) Oppenheim, Jones. 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.

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

http://oppenheim-lab.bangor.ac.uk/pubs/OppenheimJones_2018_COM_Americanisms_poster.pdf
Entrainment in dialogue systems

- 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

(Dušek & Jurčíček, 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

\[D1 = V1 \text{ was in system prompts} \]
\[D2 = V2 \text{ was in system prompts} \]
\[(frequencies \text{ in user utterances)}\]
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?
Politeness in dialogue systems

• Typically **handcrafted** by system design
  • does not adapt to situation very much
  • typically not much indirect speech, but trying to stay polite

• Learning from data can be tricky
  • **check your data** for offensive speech!
  • not just swearwords – problems can be hard to find

I already have a woman to sleep with.

(Experimental chatbot we trained at Heriot-Watt using Reddit data)
• 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 takes place in context
  • lot of pointing – deixis
• Dialogue is cooperative, social process
  • conversational maxims ~ “play nice”
  • all while following social conventions (politeness)
  • people predict & adapt to each other
• Next week: data & evaluation
Contact us: 
https://ufaldsg.slack.com/  
{odusek,schmidtova,hudecek}@ufal.mff.cuni.cz  
Skype/Meet/Zoom (by agreement)

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

References/Inspiration/Further:

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

- 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
- Wikipedia: Anaphora (linguistics) Conversation Cooperative_principle Coreference Deixis Grounding_in_communication Implicature Speech_act Sprechakttheorie