

Dialogue systems & chatbots

Pierre Lison

IN4080: Natural Language
Processing (Fall 2023)

17.10.2023



The next 3 weeks

What are they?
What applications?

How does (human-human)
dialogue actually *work*?



Dialogue systems

What are the core *components*
of dialogue systems?
Can they be learned from *data*?

How are dialogue
systems *designed,*
built and *evaluated*?

Plan

Lectures:

- ▶ October 17:
 - What is dialogue?
 - Basic chatbot models
- ▶ October 24:
 - Chatbots (cont') & NLU
 - Short intro to ASR
- ▶ October 31:
 - Dialogue management
 - System design & evaluation

Lab sessions:

- ▶ October 20:
 - Linguistic analysis
 - Basics of Numpy
- ▶ October 27:
 - Exercises on NLU & ASR
- ▶ November 3:
 - Exercises on dialogue management & RL
 - Help with assignment

Assignment

- ▶ Oblig 3 released this week
 - Deadline: november 6
- ▶ Three parts:
 - **Chatbots:** data-driven chatbot trained on movie and TV subtitles
 - **Speech processing:** implement a simple voice activity detector
 - **Dialogue management:** build a (simulated) talking elevator



Material

- ▶ The slides from the 3 lectures
- ▶ Chapter 24 of the upcoming version (v3) of Jurafsky & Martin's SLP book
 - & part of chapter 25 on phonetics
 - & dialog chapter from previous J&M edition
- ▶ + a few additional references listed in the weekly syllabus for the course

Plan for today

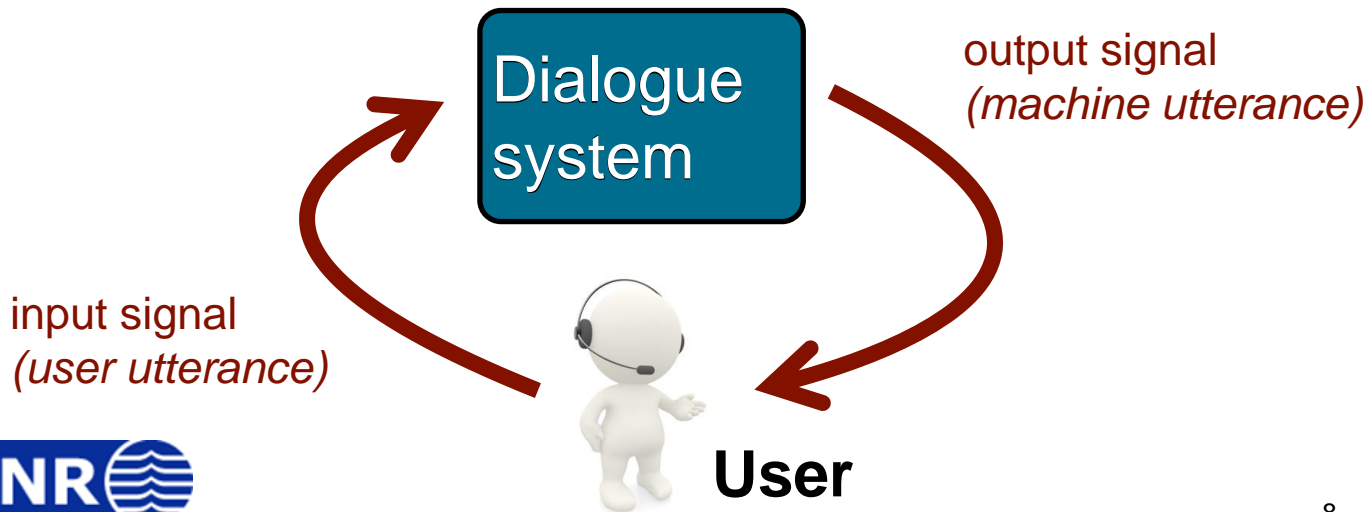
- ▶ A short intro to dialogue systems
- ▶ What is human dialogue?
- ▶ Basic chatbot models

Plan for today

- ▶ **A short intro to dialogue systems**
- ▶ What is human dialogue?
- ▶ Basic chatbot models

Dialogue systems?

A dialogue system is an artificial agent designed to interact with humans using *(spoken or text-based) natural language*



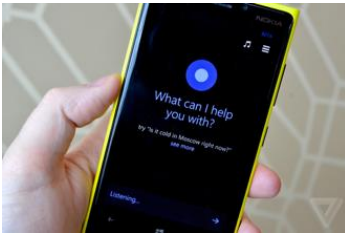
What for?

- ▶ **Highly intuitive:** no need for training or expertise: all you need is to talk/write!
- ▶ Touch-based interfaces may be inadequate, cumbersome or dangerous (car driving)
- ▶ Language is the ideal medium to express complex ideas in a flexible and efficient way



Applications

Mobile virtual assistants
(Siri, Cortana, etc.)



In-car navigation & control



Chatbots



Tutoring
systems



Smart home
environments

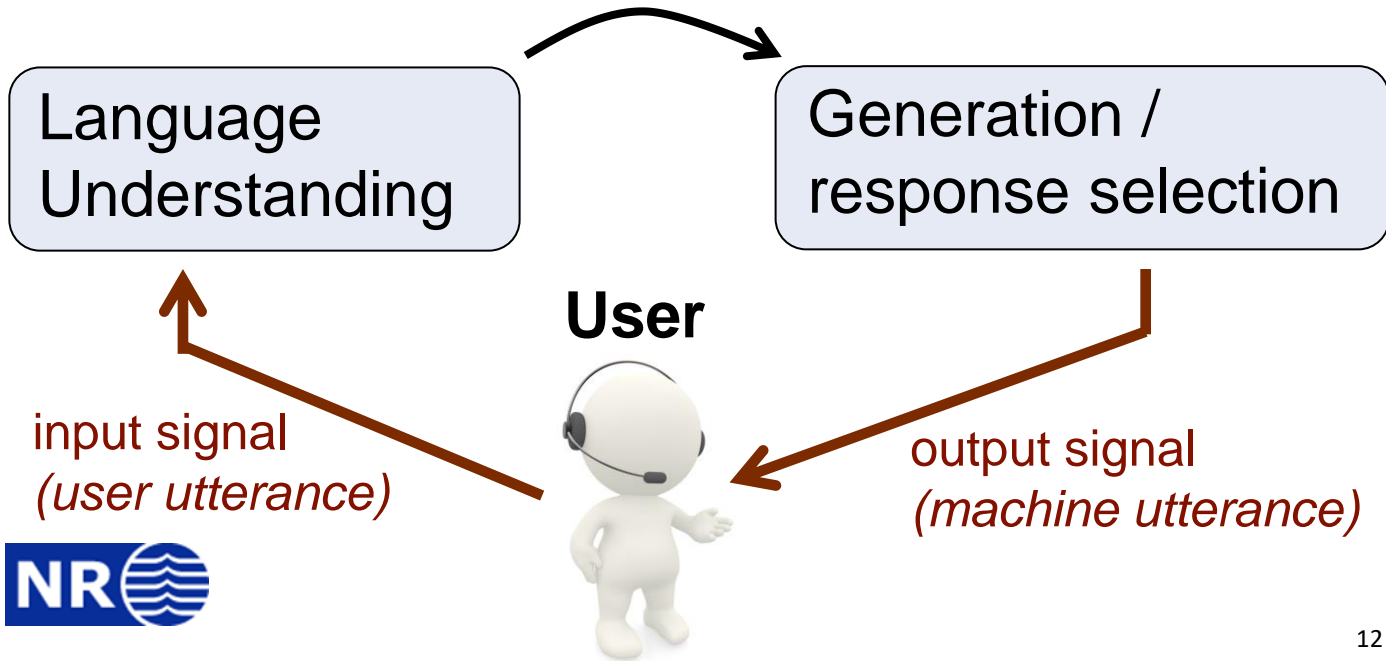


Service robots

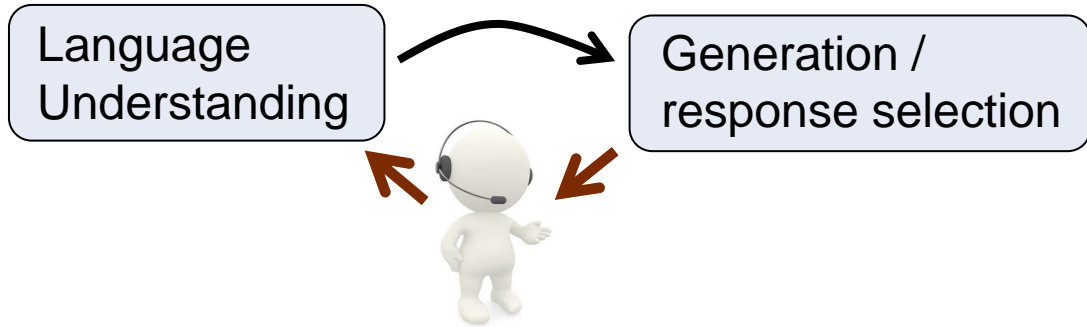


Basic architecture

Representation of user intent
(category, embeddings from LLM, etc.)



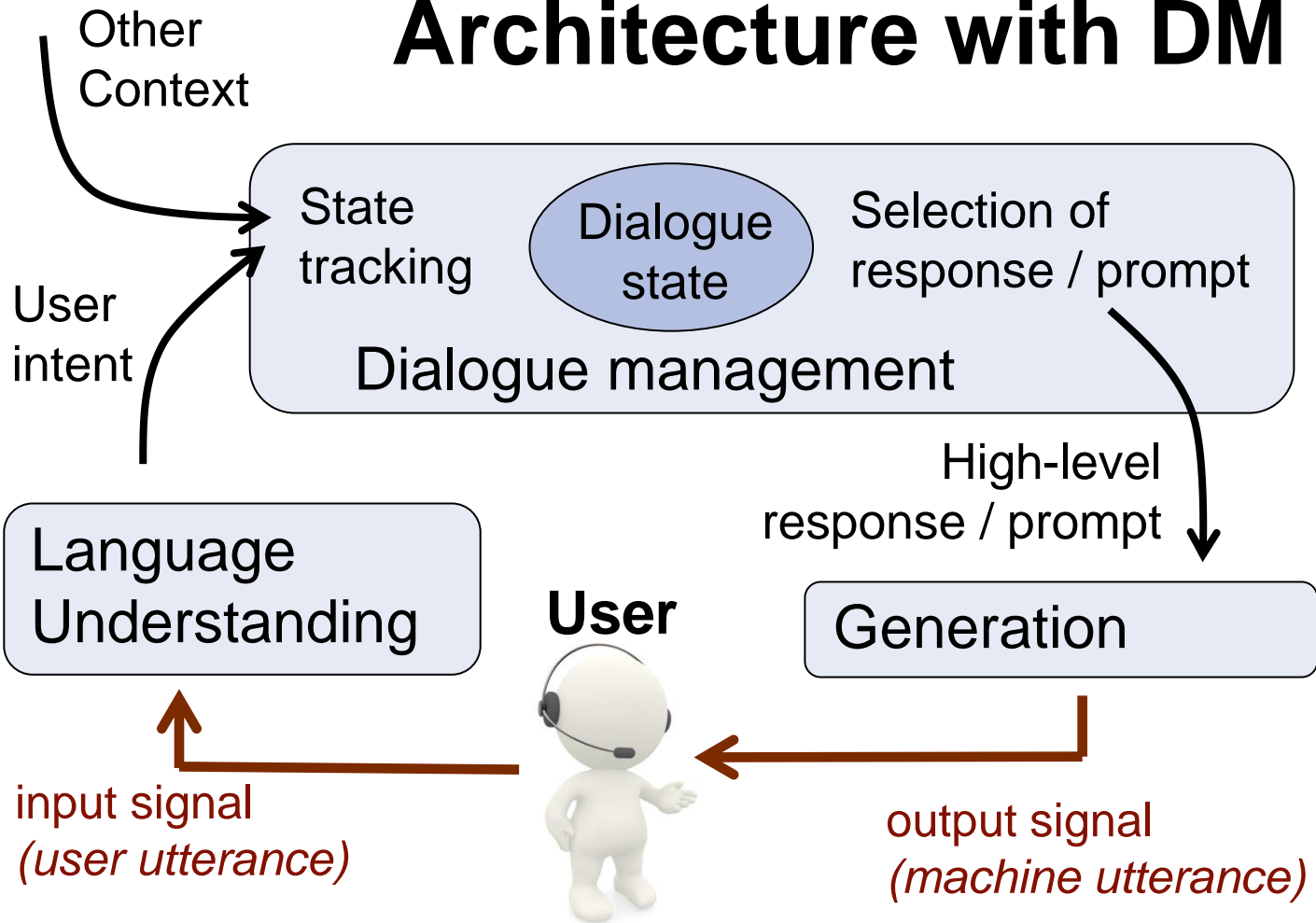
Basic architecture



This pipeline is often used for chatbots

- **Main limitation:** no management of the dialogue itself (beyond local context)
- Most appropriate for short interactions

Architecture with DM



Outline

- ▶ In two weeks, we'll look at dialogue management in more details
 - How to integrate the external context?
 - How to handle multiple (i.e. non-verbal) modalities?
 - How to design, build and evaluate dialogue systems?

- ▶ *But let's first have a look at how human conversation actually works*

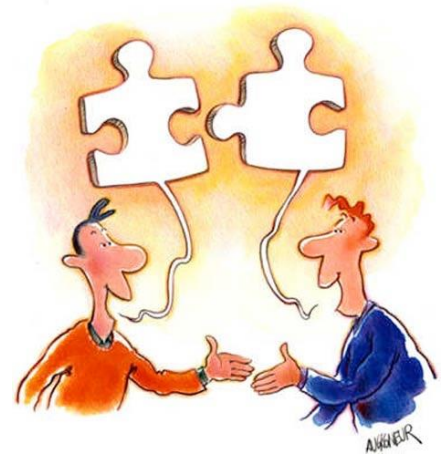


Plan for today

- ▶ A short intro to dialogue systems
- ▶ **What is human dialogue?**

What is dialogue?

- Spoken (“verbal”) + possibly non-verbal interaction between two or more participants
- Dialogue is a joint, social *activity*, serving one or several purposes for the participants
- What does it mean to view dialogue as a **joint activity**?



Turn-taking

- ▶ Dialogue participants take *turns*
 - Turn = continuous contribution from one speaker
 - Turn-taking is a *resource allocation problem*
- ▶ Surprisingly fluid in normal conversations:
 - Minimise both gaps (no speaker) and overlaps (more than one speaker)
 - Interval between speakers is around 250 ms

[Duncan (1972): «Some Signals and Rules for Taking Speaking Turns in Conversations», in *Journal of Personality and Social Psychology*]

Turn-taking

- ▶ How are turns taken or released?
- ▶ Markers for turn boundaries:
 - Complete syntactic/semantic unit?
 - Dialogue structure (greetings → greetings, question → answer)
 - Intonation (falling intonation signals that speaker is finished)
 - Non-verbal cues (eye gaze, gestures)
 - Silence & hesitation markers (unfilled pauses ≠ filled pauses)
 - Social conventions



My Turn



Your Turn

Example of turn-taking

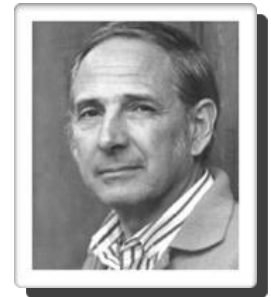
Speaker 1:	han vil bo i skogen ?
Speaker 2:	# altså hvis jeg hadde kommet og sagt " skal vi flytte i skogen ? " så hadde han sagt ja
Speaker 1:	mm
Speaker 2:	men jeg vil ikke bo i skogen
Speaker 1:	nei det skjønner jeg
Speaker 2:	så vi må jo finne et sted som er mellomting og det jeg vil ikke bo utpå landet # i hvilken som helst (uforståelig) ...
Speaker 1:	* men det kommer jo an på hvor i skogen da

Dialogue acts

- ▶ Each utterance is an *action* performed by the speaker
 - The speaker has a specific **goal** (which might be only to establish or maintain *rapport* with the listeners)
 - The utterance produces specific **effects** upon the listeners, or the world at large
 - «*Language as action*» perspective



J.L. Austin (1911-1960)
philosopher of language



J. Searle (1932, -)
philosopher of language

Dialogue acts



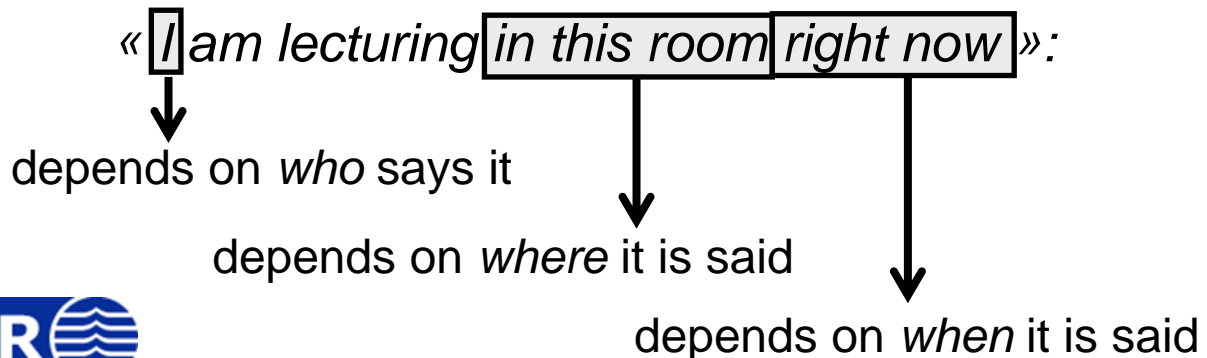
- ▶ The mother reaction has a specific **purpose**
 - Communicating her surprise/anger, and stop Calvin
- ▶ Her question will trigger some **effects**:
 - A psychological reaction from Calvin (e.g. surprise)
 - Possibly a real-world effect as well (Calvin stopping his action)

Searle's taxonomy

- ▶ **Assertives:** committing the speaker to the truth of a proposition. E.g.: «*The exam will take place on November 25*»
- ▶ **Directives:** attempts by the speaker to get the addressee to do something. E.g. : «*could you please clean up your room?*»
- ▶ **Commissives:** committing the speaker to some future course of action. E.g.: «*I promise I'll clean up my room*».
- ▶ **Expressives:** expressing the psychological state of the speaker. E.g.: «*thanks for cleaning up your room*».
- ▶ **Declaratives:** bringing about a different state of the world by the utterance. E.g.: «*You're fired*».

Deixis

- ▶ Dialogue often *referential* to a **spatio-temporal context**
- ▶ Such references are called deictics
 - Related concepts: *indexicals*, *anaphora*
- ▶ The meaning of a deictic depends on the context in which it is uttered (including the speaker perspective)



Deictic markers

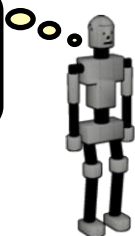


- *Pronouns*: «I», «you», «my», «yours»
- *Adverbs of time and place*: «now», «yesterday», «here», «there»
- *Demonstratives*: «this», «that»
- *Tense markers*: «he just left»
- *Others*: «the mug to your right», «go away!», «the other one»
- *Non-verbal signs*, based on gestures, gaze, etc.

Deixis

- ▶ Deictics can refer to virtually anything:
 - Objects: «take that mug»
 - Events: «don't do that», «this car accident was awful»
 - Persons: «You're being an idiot»
 - Abstract entities: «This methodology is flawed»
- ▶ Perspective is important:

behind the guy
= in front of me!

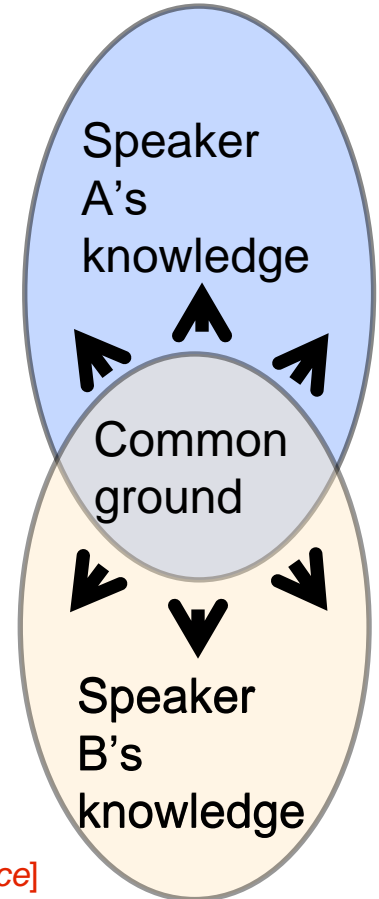


The table is
behind me!



Grounding

- ▶ Dialogue is a *joint, collaborative process* between the participants
 - Need to ensure mutual understanding
- ▶ Gradual expansion and refinement of **common ground**
 - Common ground = shared knowledge



Grounding

- ▶ Grounding is the process of gradually augmenting the common ground during the interaction
 - Variety of signals and strategies
- ▶ Multiple levels:
 - Contact (attention to interlocutor)
 - Perception (detection of utterance)
 - Understanding (comprehension of utterance)
 - Attitudinal reactions



Herbert H. Clark
psycholinguist



Jens Allwood
linguist

Grounding acts

- ▶ Backchannels: «uh-uh», «mm», «yeah»
- ▶ Explicit feedback: «*ja det skjønner jeg*»
- ▶ Implicit feedback: A: «*I want to fly to Rome*» → B: «*there are two flights to Rome on Wednesday: ...* »
- ▶ Clarification strategies: «*Did you mean to Rome or to Goa?*», «*could you confirm that ...*»
- ▶ Repair strategies: «*OK, you're not going to Goa. Where do you want to go then?*»

Examples of grounding

Speaker 1:	vi vasker den hver dag vi # vi har mopp
Speaker 2:	mm ## ja det er fort og faren til M27 legger nytt teppe han # det er gjort på to timer ## så det er fort gjort
Speaker 1:	ja ## da er ikke noe sak
Speaker 2:	vi har skifta teppe tre ganger allerede han gjør det gratis
Speaker 1:	hæ ?
Speaker 2:	vi har skifta teppe tre ganger og # han han ...
Speaker 1:	* jeg skjønner ikke hvorfor dere har teppe
Speaker 2:	jeg syns det var rart jeg òg # men e # (sibilant)

Examples of grounding

Speaker 1:	e # nei det er ikke mange	
Speaker 2:	ja * nei	
Speaker 1:	men heldigvis så var ikke Petter Rudi tatt ut denne gangen da	
Speaker 2:	ja # jeg skjønner ikke hva han skal på landslaget å gjøre	
Speaker 1:	* nei han har ingen ting på landslaget	
Speaker 2:	nei # definitivt	
Speaker 1:	å gjøre # han er ubrukelig	
Speaker 2:	* moldensere	implicit feedback (repetition of <i>landslaget</i>)
Speaker 1:	hm?	
Speaker 2:	ja disse moldenserne	clarification requests
Speaker 1:	en gang til?	
Speaker 2:	disse moldenserne	
Speaker 1:	* å ja (fremre klikkelyd) # unnskyld # jeg hørte ikke hva du sa	



Grounding

- ▶ Common ground is more than «knowledge that happens to be shared by all participants»
 - The participants must also know that it is shared (i.e. know that the others know it as well)
- ▶ Given two speakers A and B, the common ground CG can be defined as :

$$\begin{aligned} \forall x, CG(x) \rightarrow & \textit{knows}(A, x) \\ & \wedge \textit{knows}(B, x) \\ & \wedge \textit{knows}(A, \textit{knows}(B, x)) \\ & \wedge \textit{knows}(B, \textit{knows}(A, x)) \\ & \wedge \textit{knows}(A, \textit{knows}(B, \textit{knows}(A, x))) \\ & \wedge \dots \end{aligned}$$

Conversational implicatures

- ▶ Very often, part of the meaning of utterance is not explicitly stated, but only implied

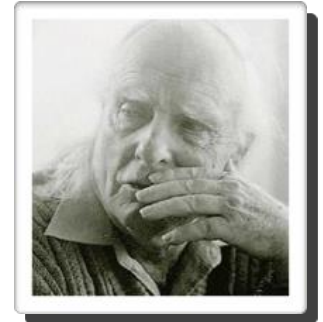
A: «Is William working today?»

B: «He has a cold»

- ▶ How can we retrieve this «suggested» meaning, and go beyond literal interpretations?
 - Need to make some *assumptions* about the speaker to help us infer the hidden part

Conversational implicatures

- ▶ Same idea again: dialogue as *a collaborative process*
- ▶ Grice's *Cooperative Principle*:
 - Maxim of Quality: «be truthful»
 - Maxim of Quantity: «be exactly as informative as required»
 - Maxim of Relation: «be relevant»
 - Maxim of Manner: «be clear»



Paul Grice (1913-1988)
philosopher of language

Conversational implicatures

- ▶ Based on the cooperative principle, one can draw *conversational implicatures*
 - All participants are assumed to adhere to the maxims
 - If an utterance initially seems to deliberately violate a maxim, the listener will then infer additional hypotheses required to make sense of the utterance



Conversational implicatures

A: «Is William working today?»

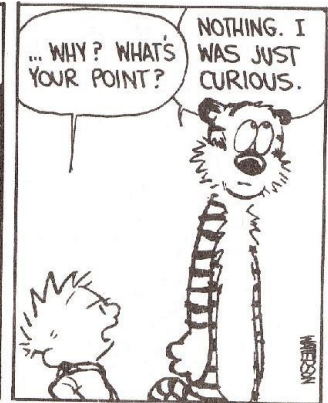
B: «He has a cold»

- ▶ At first glance, B seems to violate the maxim of relevance - he does not directly answer A's question
- ▶ But looking at the utterance more closely, we can read it as implying that (due to his cold) he is probably at home, and thus not working today
- ▶ This is because we assume that B is cooperative and wouldn't have uttered «he has a cold» if it didn't help answering A's question

Conversational implicatures



LOOK AT ME, I'M SMART! I DON'T NEED 11½ MORE YEARS OF SCHOOL! IT'S A COMPLETE WASTE OF MY TIME!



Hobbes' question is *suggesting* something about Calvin's need for schooling, without stating it explicitly

We can understand it because we assume that Hobbes' contribution is cooperative and thus relevant to the discussion

(shared) intentionality

- ▶ Humans naturally view each other as goal-directed, *intentional agents*
 - Understand other agents in terms of belief, desires and intentions (*theory of mind*)
- ▶ But there's more: humans can *jointly attend* to external entities and establish *shared intentions*

[Dennett, D (1996), *The intentional stance*.]

[Tomasello, M (2018), *Becoming Human: A Theory of Ontogeny*]



Daniel Benett (1942, -)
philosopher of mind



Michael Tomasello (1950, -)
developmental psychologist

Alignment

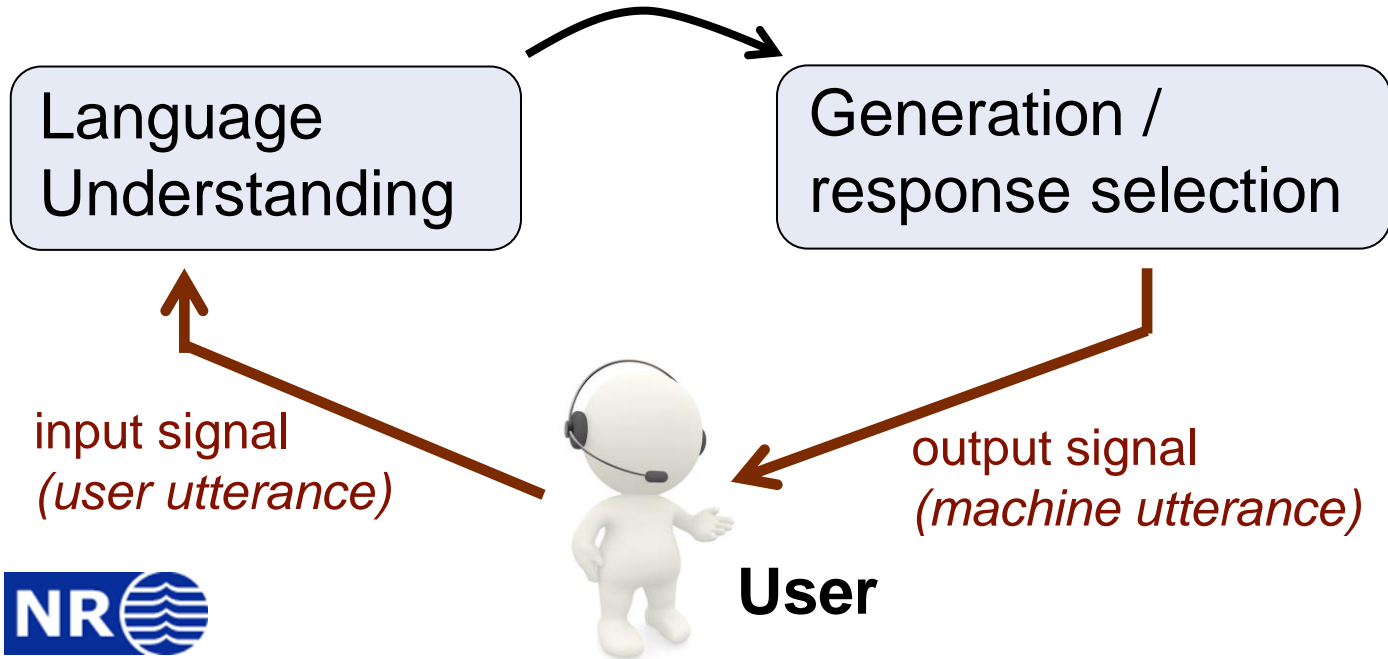
- ▶ Participants in a dialogue continuously **align** their mental representations
 - Notion of common ground discussed earlier
- ▶ But dialogue participants also align at a deeper level, by unconsciously *imitating* each other
- ▶ As the interaction unfolds, the participants automatically align their *wording*, *pronunciation*, *speech rate*, and *gestures*

Plan for today

- ▶ A short intro to dialogue systems
- ▶ What is human dialogue?
- ▶ **Basic chatbot models**

Chatbots

High-level representation of user intent
(category, embedding, etc.)



Rule-based models

- ▶ Pattern-action rules

(0 YOU 0 ME) [*pattern*]

→

(WHAT MAKES YOU THINK I 3 YOU) [*transform*]

- ▶ For instance:

You hate me

WHAT MAKES YOU THINK I HATE YOU



IR models

- ▶ Alternatively, one can adopt a data-driven approach and learn how to respond to the user based on a *dialogue corpus*
- ▶ Key idea:
 - Given a user input \mathbf{q} , find the utterance \mathbf{t} in the dialogue corpus that is most similar to \mathbf{q}
 - Then return as response the utterance \mathbf{r} following \mathbf{t} in the corpus

IR models

$$r = \text{response} \left(\underset{t \in C}{\operatorname{argmax}} \frac{q^T t}{\|q\| \|t\|} \right)$$

- ▶ How to determine which utterance is «most similar» to the actual user utterance?
 - Cosine similarity over some vectors
 - The vectors can be TF-IDF weighted words
 - Or utterance-level embeddings

Example

TF vectors:

Corpus:

1. hei ! →
2. hei ! har du det bra ? ⇨
3. ja , hva med deg ? →
4. bare bra →
5. har du spist ? →
6. ja →

	bare	bra	deg	det	du	ja	har	hei	hva	med	spist	,	!	?
1. hei !								1					1	
2. hei ! har du det bra ?		1		1	1		1	1					1	1
3. ja , hva med deg ?			1			1			1	1		1		1
4. bare bra	1	1												
5. har du spist ?					1		1				1			
6. ja						1								

Example

$$\log(6) \approx 0.78$$

$$\log\left(\frac{6}{2}\right) \approx 0.48$$

TF-IDF vectors:

Corpus:

1. hei ! \longrightarrow
2. hei ! har du det bra ? \curvearrowright
3. ja , hva med deg ? \longrightarrow
4. bare bra \longrightarrow
5. har du spist ? \longrightarrow
6. ja \longrightarrow

	bare	bra	deg	det	du	ja	har	hei	hva	med	spist	,	!	?
1. hei !								.48					.48	
2. hei ! har du det bra ?		.48		.78	.48		.48	.48					.48	.48
3. ja , hva med deg ?			.78			.48			.78	.78		.78		.48
4. bare bra	.78	.48												
5. har du spist ?					.48		.48				.78			
6. ja						.48								

New user utterance q : "går det bra med deg?"

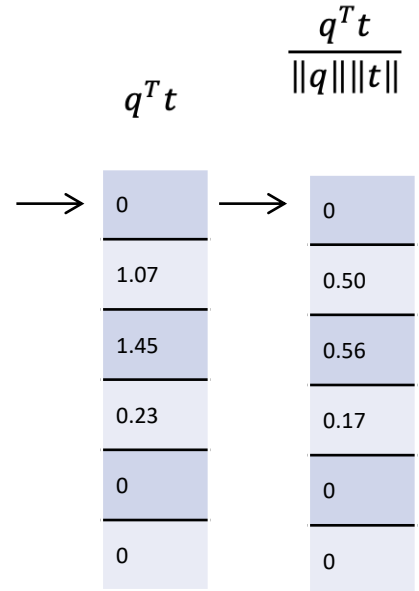
TF-IDF vector:

	.48	.78	.78							.78				.48
--	-----	-----	-----	--	--	--	--	--	--	-----	--	--	--	-----



Example

	bare	bra	deg	det	du	ja	har	hei	hva	med	spist	,	!	?
1.								.48					.48	
2.		.48		.78	.48		.48	.48					.48	.48
3.			.78			.48			.78	.78		.78		.48
4.	.78	.48												
5.					.48		.48				.78			
6.						.48								



	.48	.78	.78							.78				.48
--	-----	-----	-----	--	--	--	--	--	--	-----	--	--	--	-----

Example

$$\frac{q^T t}{\|q\| \|t\|}$$

Corpus:

1. hei !	→	0
2. hei ! har du det bra ?	→	0.50
3. ja , hva med deg ?	→	0.56
4. bare bra	→	0.17
5. har du spist ?	→	0
6. ja	→	0

☐ The utterance closest to q in our corpus is utterance 3: "ja, hva med deg?"

☐ the system should choose as response utterance 4

New user utterance q : "går det bra med deg?"

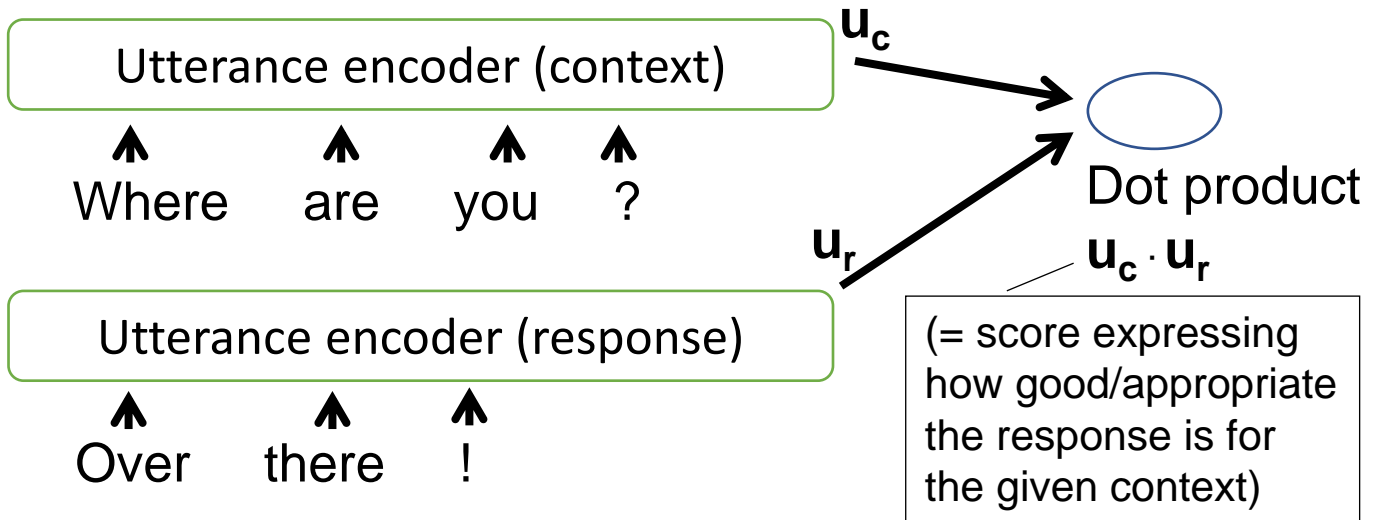
↙
↓
System response: "bare bra"



Dual encoders

Another type of IR-based chatbots

- ▶ We compute here the dot product between the user input (called "*context*") and a possible *response*



Dual encoders

The encoders are typically deep neural networks based on e.g. transformers

Utterance encoder (context)

Where are you ?

u_c

Utterance encoder (response)

Over there !

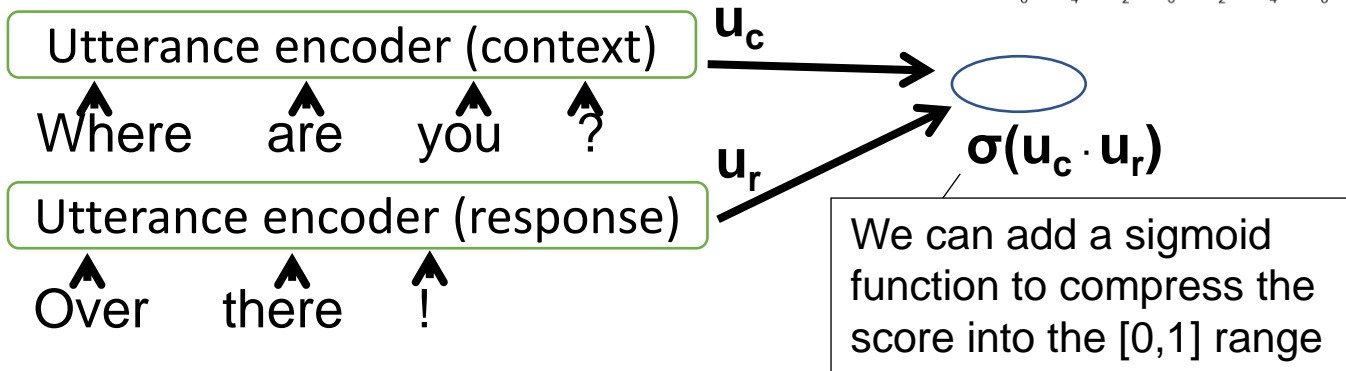
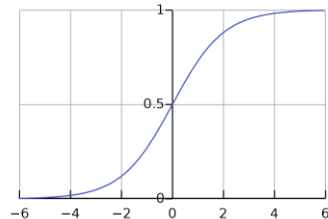
u_r

$u_c \cdot u_r$

The two encoders often rely on a shared neural network, apart from a last transformation step that is specific for the context or response

Dual encoders

$$\sigma(x) = \frac{1}{1+e^{-x}}$$



Dual encoders are trained with both *positive* and *negative* examples:

- ▶ *Positive*: actual consecutive pairs of utterances observed in the corpus \rightarrow output=1
- ▶ *Negative*: random pairs of utterances \rightarrow output=0

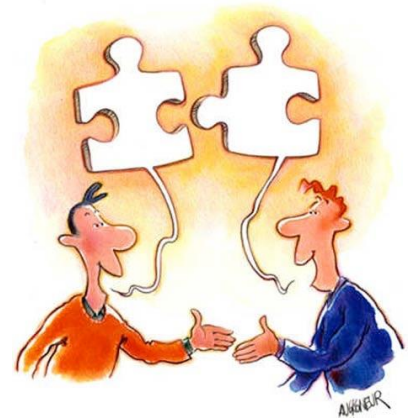
Plan for today

- ▶ A short intro to dialogue systems
- ▶ What is human dialogue?
- ▶ Basic chatbot models
- ▶ **Wrap up**

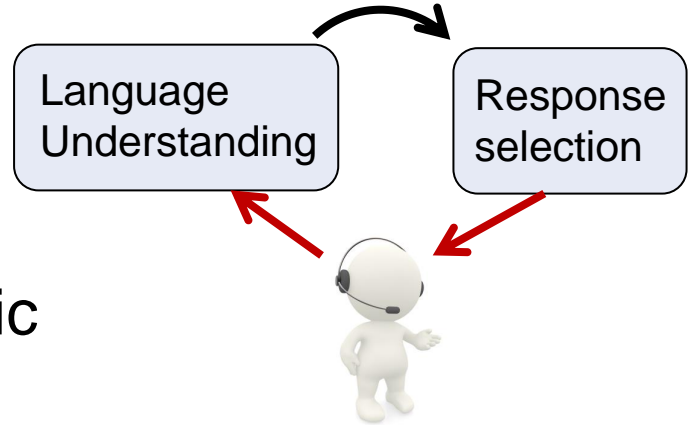
Summary (1)

Dialogue = **joint social activity**

- ▶ Dialogue participants take *turns*
- ▶ Each turn is composed of one or several *dialogue acts*
- ▶ Cooperation to ensure mutual understanding (gradual expansion of *common ground*)
- ▶ Cooperative interpretation of each other's utterances (*conversational implicatures*)
- ▶ Takes place in a *context* which is crucial for making sense of the interaction (cf. *deictics*)



Summary (2)



We also looked at basic models for chatbots:

- **Rule-based systems**, which map *conditions* (e.g. surface patterns on the user utterance) to *responses*
- **IR-based systems** searching for the most similar utterance in a dialogue corpus, and then selecting the utterance after it

Next week

- ▶ In the next lecture, we'll look at more advanced chatbot models
 - NLU-based approaches (intent & slot recognition)
 - Generative models
- ▶ + short intro to phonetics & speech recognition!

