NORSK REGNESENTRAL
NORWEGIAN COMPUTING CENTER

Dialogue systems (4)

Pierre Lison
plison@nr.no

IN4080, Autumn 2025, video lecture

Week 43

- The second lecture on dialogue systems will be done **remotely**, through a series of shorter videos
- Don’t hesitate to reach out to me (plison@nr.no) or on our Discourse channel if you have any questions!

Plan:

- Video 1: generative dialogue models
- Video 2: dialogue management
- **Video 3: system development and evaluation**

Obligatory assignment 4

- The final obligatory assignment of the IN4080 course is now out!
  - See github repo
- Deadline: **Friday, November 7 at 23:59**
- Two parts:
  - Implementation of a *talking elevator*
  - *Machine translation* experiments
- Don’t wait until the last minute!
  - And let us know if you experience problems



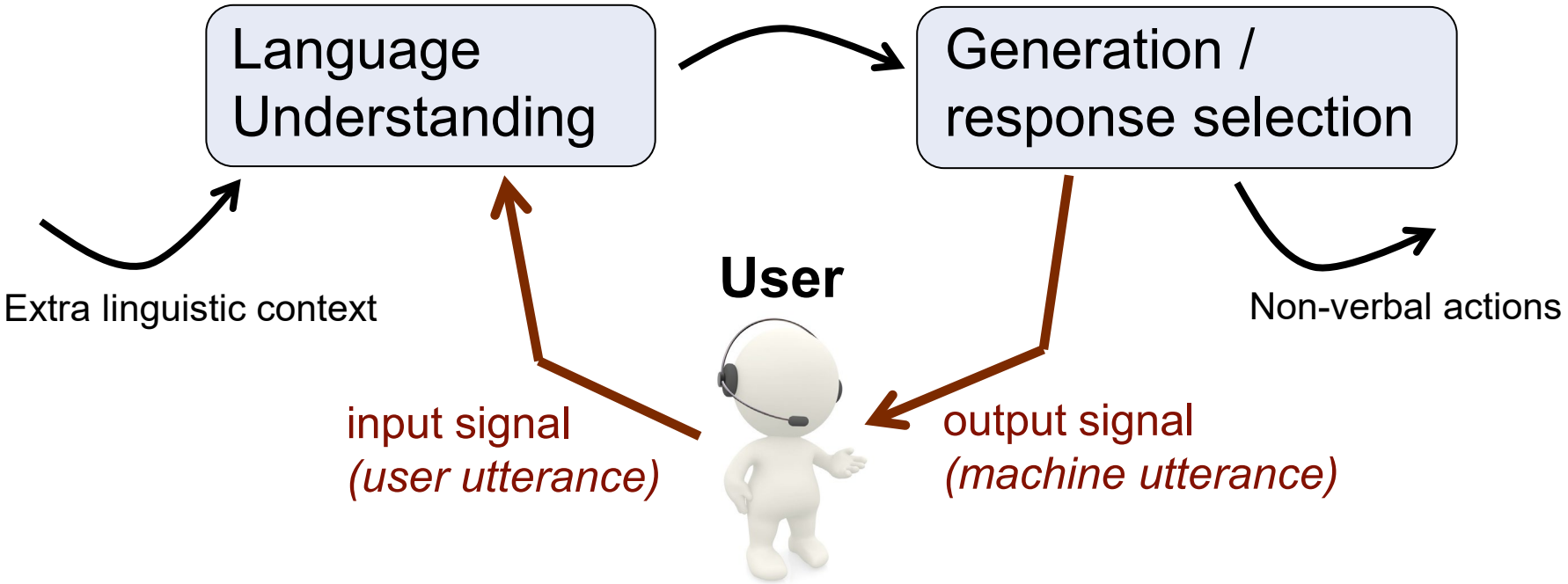
Outline

- System development
- Evaluation metrics

Outline

- **System development**
- Evaluation metrics

Quick recap


Language Understanding
Generation / response selection
Extra linguistic context
User
Non-verbal actions
input signal
(user utterance)
output signal
(machine utterance)

Pipeline architectures

- Components connected in processing chain
- Each component is a black box getting inputs from its predecessor and generating an output



Limitations:

- No feedback between components
- Rigid information flow
- Poor turn-taking behaviour (system does not react until the full pipeline has been traversed)

Acronyms:

ASR: Automatic Speech Recognition

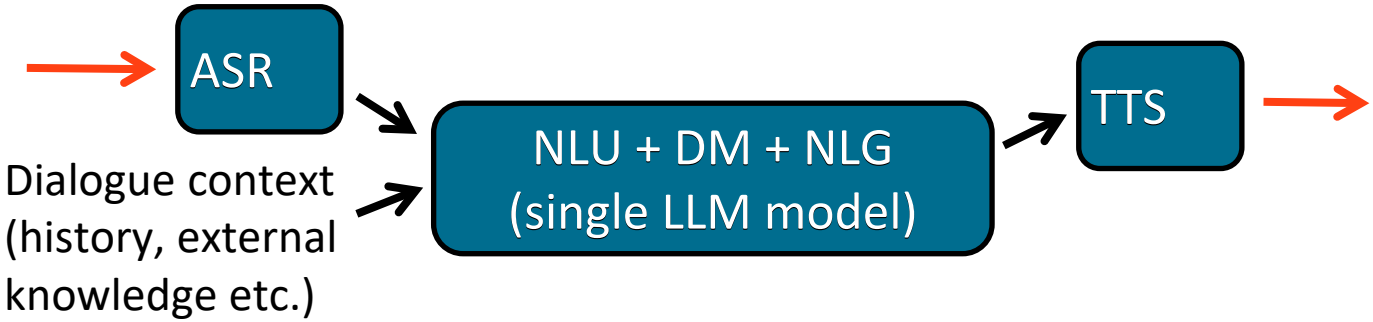
NLU: Natural Language Understanding

DM: Dialogue Management

NLG: Natural Language Generation

TTS: Text-to-speech Synthesis

End-to-end architectures



Single LLM mapping user inputs to responses

- **Pro**: no need for separate modules
- **Con**: less modular setup (harder to debug and interpret)

Dialogue management either done implicitly or by modifying the prompt

Incrementality

Humans process and produce language **incrementally:**

- When listening, we don't wait for an utterance to be fully pronounced to process it!
- We gradually refine our understanding as we go, phoneme by phoneme
- We also continuously provide *feedback signals*



Human-human dialogues are full of *interruptions*, speech *overlaps*, *backchannels*, and *co-completion* of utterances

Incrementality

- But many dialogue systems operate in «batch mode»
  - NLU expects full utterance as input
  - TTS waits for complete system response to start synthesis
- Leads to «ping-pong» turn-taking behaviour:
  - Alternating turns between user & system, one speaker at a time
- However, recent work has shown how LLMs can be adapted to operate in «full-duplex» mode



[Wang, P. et al (2024). A Full-duplex Speech Dialogue Scheme Based On Large Language Models. *NeurIPS 2024*]

How to collect data?

- "Chicken-and-egg" problem:
  - Need data to train models
  - But to collect in-domain data, we need a system that can interact with users
- One solution: **Wizard-of-Oz studies**

= experiment in which users interact with a seemingly autonomous system, but which is actually operated by an unseen human being



Outline

- System development
- **Evaluation metrics**

Evaluation

- Some dialogue processing tasks have standard evaluation metrics:
  - ASR: *Word Error Rate*
  - NLU: *[precision, recall, F-score]* for intent recognition and slot-filling
  - TTS: evaluation by human listeners on sound intelligibility and quality
- But how do we evaluate the end-to-end conversational behaviour of the system?



Evaluation

- One way to evaluate is via **user satisfaction ratings**
- The ratings can be obtained from surveys that users are asked to fill after interacting with the system:

TTS Performance	Was the system easy to understand ?
ASR Performance	Did the system understand what you said?
Task Ease	Was it easy to find the message/flight/train you wanted?
Interaction Pace	Was the pace of interaction with the system appropriate?
User Expertise	Did you know what you could say at each point?
System Response	How often was the system sluggish and slow to reply to you?
Expected Behavior	Did the system work the way you expected it to?
Future Use	Do you think you'd use the system in the future?

[M. Walker et al. (2001), «Quantitative and Qualitative Evaluation of Darpa Communicator Spoken Dialogue Systems», *Proceedings of ACL*]

Evaluation

- However, user evaluation surveys are expensive and time-consuming
  - Not feasible to conduct after each system change!
  - Can we **automate** the evaluation process?
- Possible solution: rely on metrics that can be extracted from interaction logs & *are known to correlate with user satisfaction*
  - Hypothesis: improving these observable metrics should therefore increase user satisfaction

[M. Walker et al. (1997), "PARADISE: A general framework for evaluating spoken dialogue agents", Proceedings of ACL]

Evaluation

Criteria	Description	Possible metrics
Task completion success	How often did the system complete its task successfully?	- κ agreement on slots - completion ratio
Efficiency costs	How efficient was the system in executing its task?	- nb of turns (from user, system, or both) - total elapsed time
Quality costs	How good was the system interaction?	- nb of ASR rejection prompts - nb of user barge-ins - nb of error messages

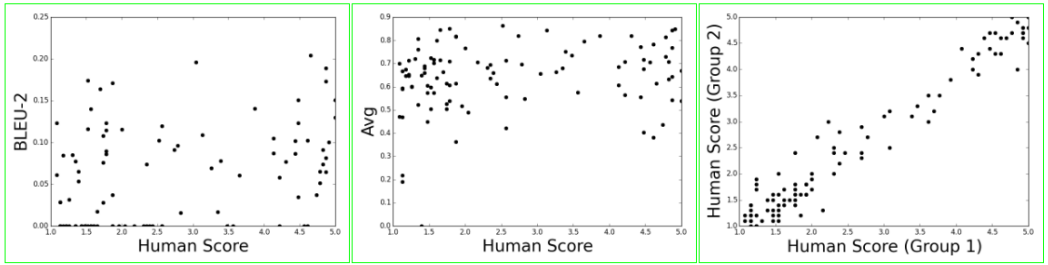
NB: this list of metrics is of course not exhaustive!

Evaluation

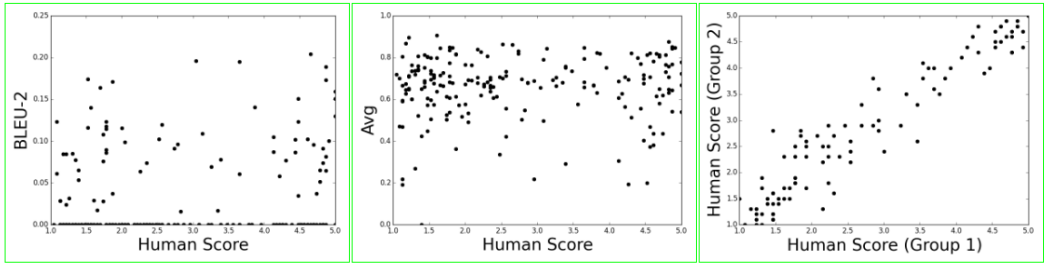
Can we use overlap-based metrics (BLEU, ROUGE etc.) to compare system outputs with human responses?

- Benefit: straightforward if you have access to high-quality responses

But: weak correlation between BLEU scores and human judgments!



(a) Twitter



(b) Ubuntu

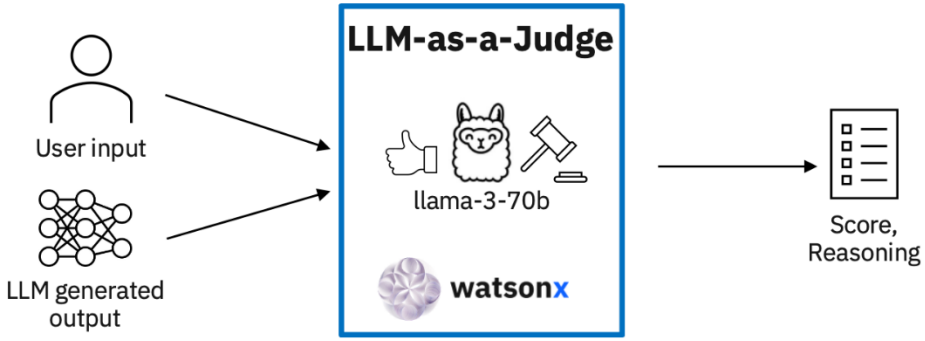
[Liu et al (2016). How NOT To Evaluate Your Dialogue System: An Empirical Study of Unsupervised Evaluation Metrics for Dialogue Response Generation. In *EMNLP*.]

Evaluation

- To get a more “semantic” view on the similarity between the system outputs and human references, metrics such as BERTScore are preferable
  - And ideally with more than one reference response…
- Many reference-free metrics are also available:
  - ADEM, DynaEval, USR, etc.
  - Typically optimized to correlate with human ratings, by identifying features that characterize “good responses”

Yeh, Y. T., Eskenazi, M., & Mehri, S. (2021). A Comprehensive Assessment of Dialog Evaluation Metrics. In *The First Workshop on Evaluations and Assessments of Neural Conversation Systems*.

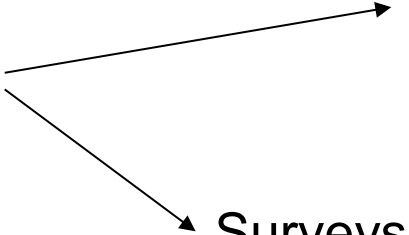
LLM-as-a-judge



- **Idea**: Ask an LLM to judge the quality of the system responses
  - Preferably using another model than the one employed to generate the response
  - Need to describe precisely the evaluation criteria (along with examples)
- Can give useful insights…
- **But**: *beware of large variances*, depending on the prompt (in particular how the evaluation criteria are described), model, and dialogue domain!

Bavaresco, A et al (2024). LLMs instead of human judges? A large-scale empirical study across 20 NLP evaluation tasks. *arXiv*.

Human evaluations



Selection/ranking of preferred responses (to compare systems)

Surveys after the dialogue is complete

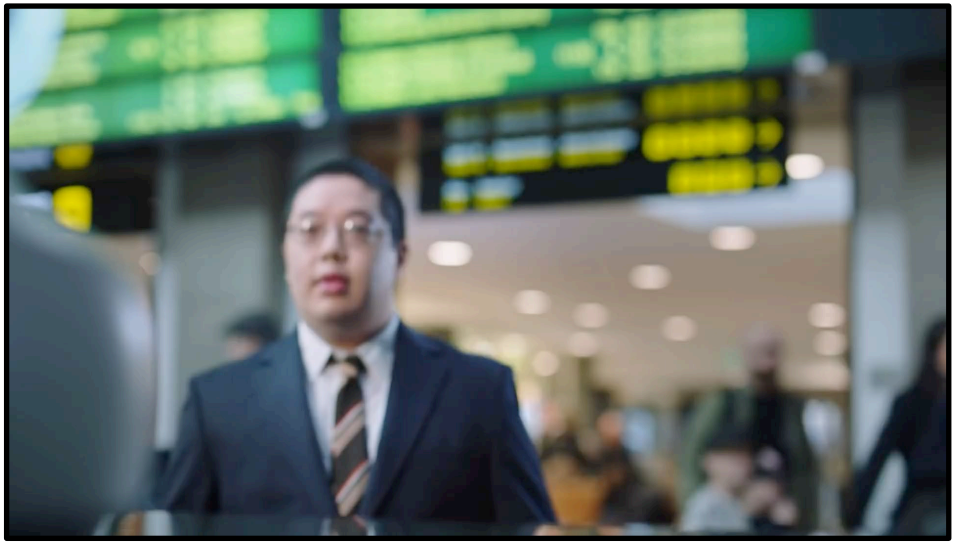
➔ Need to provide human evaluators with **well-defined criteria** they will need to assess along with a **scale** (i.e. Likert)

➔ You should include **several evaluators** in your study and compute their inter-annotator agreement

➔ When comparing systems, compute the **statistical significance** of your findings (using e.g. the *Wilcoxon signed-rank test***)**

What we haven't covered

- **Speech**
  (lecture next week)
- **Multimodal** &
  situated systems



Furhat robot (initially developed at KTH, Stockholm), see www.furhatrobotics.com