Dialogue Engineering

Dialogue Resources & Evaluation

Amandine Decker, amandine.decker@loria.fr

M2 NLP 2025–2026

Recap’

- Dialogue Systems have two primary dimensions: their task completion ability and their conversational ability;
- The complexity of dialogue must be taken into account when selecting resources, whether for building the system or evaluating it.

Today

- Why don’t we always collect real-life data suited for our task?
- What do we use instead?
- How do we evaluate our tasks?

Dialogue Data Collection

Recording Dialogue Data

- **Participant recruitment:** choosing relevant speakers, obtaining consent
- **Setup:** defining environment (lab, online, in-the-wild) and tools (audio, video, sensors)
- **Observation effect:**
  - Recording or experimental tasks can distort natural dialogue
  - Participants become self-conscious or adapt behaviour
- **Task design:**
  - Constrains spontaneity
  - Balances control and realism
- **Multimodal capture:**
  - Adds richness (gesture, gaze, prosody)
  - But raises privacy concerns → smaller datasets

Anonymity and Privacy

- **Anonymising dialogue data** is extremely challenging:
  - Natural conversation includes names, places, personal details
  - Linguistic style and voice can still reveal identity
- **Open-domain data:** impossible to fully anonymise casual speech without altering meaning or flow
- **Potential solutions:**
  - Synthetic or simulated data (e.g. crowd-sourced, scripted)
  - Careful metadata filtering and access restrictions
- Tension between **authenticity** and **ethical protection**.

Dialogue Data Collection – Summary

- Recording conditions shape conversational behaviour.
- Experimental control often reduces naturalness.
- Multimodal data enriches analysis but increases collection and privacy risk.
- Anonymity in open-domain data remains largely unsolved.

→ Data collection in dialogue is a balance between realism, control, and ethics.

Some Families of Resources

Dialogue Resource Example (I)

A1: I am wanting to know more about the Cambridge Museum of Technology.

B2: Sure, it's located in the centre area of town. The phone number is 01223368650. The entrance fee is 5 pounds.

A3: What is the address, please.

B4: The address is the old pumping station, cheddars lane. Did you want any more information about the museum?

A5: Yes, I need the entrance fee. I will also need a taxi to pick me up by 24:30. I need the contact number and car type please.

B6: The entrance fee is 5 pounds

A7: Can you book the taxi for me at 24:30 leaving from the museum please. I will need the contact number and the car type.

B8: and what is your destination?

A9: I will leave the Cambridge Museum at 24:30, going to the the Finches Bed and Breakfast.

B10: Your taxi has been booked. It will be a grey volvo and the contact number is 07239297218.

A11: Thanks so much. That is all I need for today. Bye.

Wizard of Oz Paradigm

- Principle:
  - A human secretly plays the role of the system (*the "wizard"*).
  - Participants believe they are interacting with an automated system.
- Purpose:
  - Collect realistic **task-oriented dialogues** before a system exists.
  - Often linked to database or information-retrieval tasks.
- Limitations:
  - Interactions not fully natural – participants "play a role".
  - Methodological issue today: many corpora likely included in LLM training → unreliable as evaluation data.

Dialogue Resource Example (II)

A1$_1$: I have one smalll rocket shaped tangram

A1$_2$: do u have this?

B2$_1$: Ok, so let me try describing them by what the kinda remind me of. My picture A looks like a square pacman, picture B is probably a cactus, and picture c looks like a standing person

B2$_2$: The closest to the rocket is B to me, but I don't think it is actually a rockets

A3$_1$: I think a square pacman maybe

A3$_2$: so its like you keep two rectangles on top of each other and shift any of them a bit

A3$_2$: so its like you keep two rectangles on top of each other and shift any of them a bit

A3$_3$: _---- something like this

A3$_4$: "_---"

A3$_5$: oops

A3$_6$: hard to make

B4: Are you describing a rocket or a pacman?

A5: pacman

B6: Pacman for me looks like a triangle with a rectangle on one side cut out (like a mouth piece) and a small square inside that mouth (like something it eats)

A7$_1$: ah yes this is there

A7$_2$: for me

Task-Oriented Dialogues

- Purpose:
  - Study collaboration and communication in specific tasks.
  - Convenient to design and control.
- Common Tasks:
  - *Tangrams:* developing shared references.
  - *Maze or map tasks:* coordination and clarification.
  - *Ethical reasoning tasks:* studying moral or pragmatic negotiation.
- Limitations:
  - Task structure can distort the very features being studied.
  - Limited generalisability to open-domain conversation.

Dialogue Resource Example (III)

Rachel Green: C'mon Daddy, listen to me! It's like, it's like, all of my life, everyone has always told me, 'You're a shoe! You're a shoe, you're a shoe, you're a shoe!'. And today I just stopped and I said, 'What if I don't wanna be a shoe? What if I wanna be a- a purse, y'know? Or a- or a hat! No, I'm not saying I want you to buy me a hat, I'm saying I am a ha- It's a metaphor, Daddy!

Ross Geller: You can see where he'd have trouble.

Rachel Green: Look Daddy, it's my life. Well maybe I'll just stay here with Monica.

Monica Geller: Well, I guess we've established who's staying here with Monica...

TV Shows as Dialogue Resources

- Examples:
  - *Friends*, *The Big Bang Theory*: multi-party, dialogue-driven shows.
- Advantages:
  - Large scale, easily available.
  - Rich multi-party interactions.
- Limitations:
  - Scripted: not spontaneous speech (but transcript of the real dialogue can be better).
  - Dialogue thought for an external audience.

Dialogue Resource Example (IV)

A1$_1$: What faction would you pick for this home slice?

A1$_2$: */Picture of board game layout (five hexagonal tiles with space background and planets)/*

> B2: Clan of Saar, park over industrex, and pump out warsuns as soon as you can
>
> > A3: I love this idea. Lock that bad boy up with chaos mapping. Use Lemus to get floating factory 2 pain free. Plus blue red breakthrough
> >
> > > B4: Had a buddy do that our first TE game about a month ago. Was super effective and super annoying 😂
> > >
> > > > A5: I can imagine. That planet sounds like a lot of fun. Haven't played with it in the board yet
>
> C6: Or Park in the entropic scar (but be careful your docks dont produce there) and research SD2 and then pump out a SD every turn :D until you have war sun researched and switch to that

Social Media Conversations

- Advantages:
  - Massive scale, public availability.
  - Natural expression of opinion, emotion, and stance.
- Limitations:
  - **Written modality:** lacks prosody and timing.
  - Constrained by platform: character limits, thread structure.
  - Demographic bias → not representative of all speakers.
  - Ethical and licensing considerations.
- **Takeaway:** a distinct communication genre with its own conventions.

Dialogue Resource Example (V)

Figure 1: Example from the KCDD dataset (Kim et al., 2024)

Handcrafted Resources (e.g. DailyDialog)

- Advantages:
  - No need for recording or transcription.
  - Easy to target specific phenomena or domains.
- Limitations:
  - Created by individuals or small teams → subjective, stylised.
  - May not reflect genuine conversational dynamics.
  - Quality and naturalness vary widely.
- Variants:
  - Professional writers or actors produce more realistic dialogues.
  - Some collect only next-turn completions or partial prompts.

Other Leads

- **Translation:** Cultural and politeness norms often do not transfer → requires native expertise.
- **Virtual Reality:** Some well-known dialogue effects fail to replicate in VR (Lücking et al., 2025); context changes behaviour.
- **LLM Generation:** Circularity problem – LLMs trained on prior data generating new "synthetic" dialogue → harder to evaluate authenticity.

Why Resources (Still) Matter

- **Evaluation is increasingly difficult:** Many public corpora already absorbed by large models → contaminated benchmarks.
- **Creating new, high-quality data** remains crucial:
  - Independent evaluation sets;
  - Datasets that reflect current conversational practices.
- **Studying real data** can helps recognise synthetic text.

  ⇒ Reliable resources underpin progress and trust in dialogue research.

Some Common Tasks

What we actually work on in the dialogue world

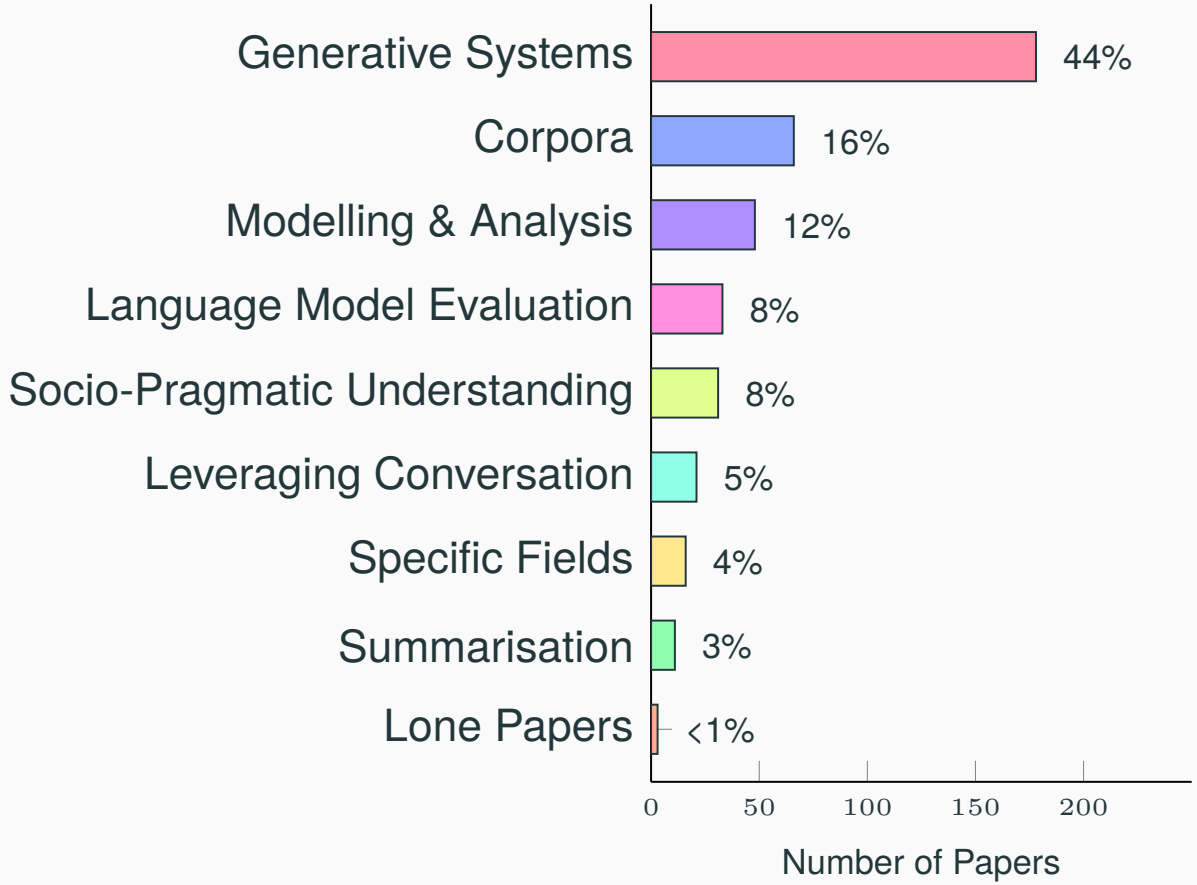


Figure 2: Distribution of dialogue related papers by **task category** in major NLP/CL publication venues in 2024 (over 407 papers)

Dialogue Act Classification

- Description
  - Support task: Predicting the communicative function of each user or system utterance (e.g. question, request, acknowledgement).
- Use?
  - Fundamentally supports dialogue management.
  - Provides structured understanding of conversational flow.
- Training Data Needs
  - Utterances annotated with dialogue act labels.
  - Multi-turn conversational context.
- Evaluation
  - *Task*: accuracy, F1-score on act labels.
  - *Conversational*:

Conversational Search

- Description
  - Interactive search where the system clarifies needs, refines queries, and guides users to relevant information.
- Training Data Needs
  - User queries and reformulations.
  - Dialogue turns with clarification questions.
  - Relevance labels for retrieved items.
- Evaluation
  - *Task*: ranking quality (NDCG, MRR), Recall@k, success rate.
  - *Conversational*:

    Ideally: quality of clarifications, coherence, initiative handling;
    In practice: none or semantic similarity with reference

Conversational Recommender Systems

- Description
  - Dialogue systems that recommend items (movies, restaurants, etc.) while eliciting user preferences interactively.
- Training Data Needs
  - User–item interactions and preference annotations.
  - Dialogues containing constraints, critiques, confirmations.
- Evaluation
  - *Task*: hit-rate, recall@k, recommendation accuracy.
  - *Conversational*:

    Ideally: turn efficiency, naturalness, user satisfaction,...;
    In practice: none or lexical variety between user and system (Dist-n).

Persona-Based Dialogue Systems

- Description
  - Systems designed to display a specific personality/behaviour throughout dialogue.
- Applications
  - Casual conversations;
  - Empathetic conversations;
  - Persuasive conversations.
- Potential issues / risks
  - Increases the impression that these systems have sentience / can replace a human being for emotional counselling;
  - Should we really attempt to build replacements for human beings or aim for a society where isolation is not solved with robots?

A few great/original/fun tasks

SCIENCE ADVANCES | RESEARCH ARTICLE

SOCIAL SCIENCE

The CANDOR corpus: Insights from a large multimodal dataset of naturalistic conversation

Andrew Reece[1]*†, Gus Cooney[2]*†, Peter Bull[3], Christine Chung[3], Bryn Dawson[1], Casey Fitzpatrick[3], Tamara Glazer[3], Dean Knox[2], Alex Liebscher[1], Sebastian Marin[1]

Copyright © 2023 The Authors, some rights reserved; exclusive licensee American Association for the Advancement of Science. No claim to original U.S. Government Works. Distributed under a Creative Commons Attribution NonCommercial License 4.0 (CC BY-NC).

People spend a substantial portion of their lives engaged in conversation, and yet, our scientific understanding of conversation is still in its infancy. Here, we introduce a large, novel, and multimodal corpus of 1656 conversations recorded in spoken English. This 7+ million word, 850-hour corpus totals more than 1 terabyte of audio, video, and transcripts, with moment-to-moment measures of vocal, facial, and semantic expression, together with an extensive survey of speakers' postconversation reflections. By taking advantage of the considerable scope of the corpus, we explore many examples of how this large-scale public dataset may catalyze future research, particularly across disciplinary boundaries, as scholars from a variety of fields appear increasingly interested in the study of conversation.

INTRODUCTION

Conversation hardly needs introduction. It is a uniquely human act of cooperation that requires exquisite coordination across many levels of cognition (*1*–*4*). It is the seat of language acquisition (*5*). Its turn-taking system emerges early in development (*6*, *7*) and shows parallels in nonhuman primates and other animals (*8*, *9*). It is how group members absorb and transmit culture (*10*, *11*). It is the primary tool that humans use to form and maintain their social relationships (*12*, *13*). It has a substantial impact on people's mental and physical health (*14*, *15*), and more recently, generative models of conversation have emerged as a major milestone in artificial intelligence (*16*–*18*).

Despite its centrality, conversation's complexity has hampered its empirical study: Conversation is characterized by a strong degree of interdependence between speaking partners, in which one's words and behavior are adjusted rapidly in response to what one's partner is doing; conversation is staggeringly multimodal, involving information transmission across linguistic, paralinguistic, and visual channels simultaneously; and last, conversation is highly contextualized, in which people play certain social roles, pursue specific goals, and negotiate status and power hierarchies. In turn, this complexity presents numerous scientific challenges, from operationalization to measurement to statistical modeling. However, here, we demonstrate that recent technological advances have begun to offer solutions to these challenges, placing previously inaccessible research questions within reach and offering considerable opportunity for interdisciplinary collaboration.

Historically, progress on conversation research has been catalyzed by large public datasets, such as the Map Task Corpus (*19*), the Switchboard Corpus (*20*), or newer multimodal datasets, such as the MELD (*21*, *22*) and OMG-Empathy datasets (*23*) [for a review, see (*24*)]. While these datasets have advanced conversation science, none includes a large sample of naturalistic conversation, with full audio and video recordings, together with speakers' detailed postconversation reports.

We collected such a dataset of 1656 unscripted conversations over video chat that comprise more than 7 million words and 850+ hours of audio and video. Overall, our corpus includes more than 1 terabyte of raw and processed recordings. The corpus draws on a large and diverse sample of participants, aged 19 to 66, from all over the United States. Participants were paired using an automatic matching algorithm of our own design and were simply instructed to have a conversation with one another for at least 25 min, although many talked for much longer. The conversations occurred during 2020 and, thus, offer a unique perspective on one of the most tumultuous years in recent history, including the onset of a global pandemic and a hotly contested presidential election. The corpus is among the largest multimodal datasets of naturalistic conversation, which we refer to collectively as the CANDOR corpus (Conversation: A Naturalistic Dataset of Online Recordings).

Large amounts of raw data alone are not sufficient to advance the study of conversation. In other domains, growth in computational power, the use of crowdsourcing platforms, and technological advances in machine learning, e.g., language and signal-processing algorithms such as Word2Vec, BERT, and ResNet, have proven to be yet another catalyst of scientific advancement, enabling discovery and inference at scale (*25*–*29*). In this spirit, we applied an elaborate computational pipeline to quantify features of conversation such as overlaps and pauses, second-by-second variation in facial features, and full transcripts with accompanying prosodic characteristics of speech. Last, we collected a battery of psychological measures from the participants, including trait-level measures such as personality, as well as people's opinions about their conversation partner and their feelings about the overall conversation.

We explore the corpus in five sections. First, we use the corpus to

A few great/original/fun tasks

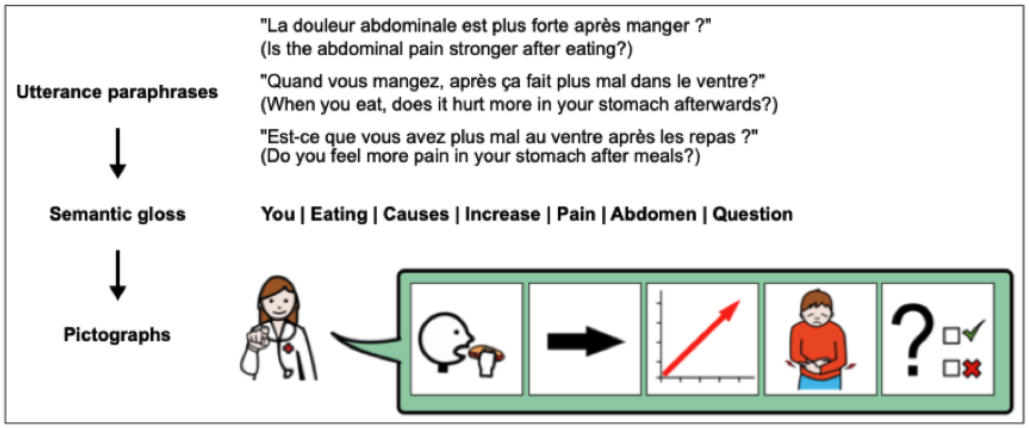
A Concept Based Approach for Translation of Medical Dialogues into Pictographs

Johanna Gerlach[1], Pierrette Bouillon[1], Jonathan Mutal[1] and Hervé Spechbach[2]
[1] TIM/FTI, University of Geneva, Geneva, Switzerland
[2] HUG, Geneva University Hospitals, Geneva, Switzerland
{johanna.gerlach, pierrette.bouillon, jonathan.mutal}@unige.ch
herve.spechbach@hcuge.ch

Abstract

Pictographs have been found to improve patient comprehension of medical information or instructions. However, tools to produce pictograph representations from natural language are still scarce. In this contribution we describe a system that automatically translates French speech into pictographs to enable diagnostic interviews in emergency settings, thereby providing a tool to overcome the language barrier or provide support in Augmentative and Alternative Communication (AAC) contexts. Our approach is based on a semantic gloss that serves as pivot between spontaneous language and pictographs, with medical concepts represented using the UMLS ontology. In this study we evaluate different available pre-trained models fine-tuned on artificial data to translate French into this semantic gloss. On unseen data collected in real settings, consisting of questions and instructions by physicians, the best model achieves an F0.5 score of 86.7. A complementary human evaluation of the semantic glosses differing from the reference shows that 71% of these would be usable to transmit the intended meaning. Finally, a human evaluation of the pictograph sequences derived from the gloss reveals very few additions, omissions or order issues (<3%), suggesting that the gloss as designed is well suited as a pivot for translation into pictographs.

Keywords: pictographs, medical communication, pre-trained models, UMLS



A few great/original/fun tasks

Ontologically Faithful Generation of Non-Player Character Dialogues

Nathaniel Weir[1] **Ryan Thomas**[2] **Randolph d'Amore**[2] **Kellie Hill**[2]
Benjamin Van Durme[1,2] **Harsh Jhamtani**[2]
[1]Johns Hopkins University [2]Microsoft
nweir@jhu.edu, hjhamtani@microsoft.com

Abstract

We introduce a language generation dataset grounded in a popular video game. KNUDGE (**KN**owledge Constrained **U**ser-NPC **D**ialogue **GE**neration) requires models to produce trees of dialogue between video game characters that accurately reflect quest and entity specifications stated in natural language. KNUDGE is constructed from side quest dialogues drawn directly from game data of Obsidian Entertainment's *The Outer Worlds*, leading to real-world complexities in generation: (1) utterances must remain faithful to the game lore, including character personas and backstories; (2) a dialogue must accurately reveal new quest details to the human player; and (3) dialogues are large trees as opposed to linear chains of utterances. We report results for a set of neural generation models using supervised and in-context learning techniques; we find competent performance but room for future work addressing the challenges of creating realistic, game-quality dialogues.

Figure 1: An example non-player character (NPC) dialogue from *The Outer Worlds* by Obsidian. NPCs must speak faithfully to a granular ontology of **quest specifications** and **game lore**.

1 Introduction

Player interactions with non-player characters (NPCs) in role-playing games (RPGs) often serve to flesh out backstories while allowing the player to progress through engaging quest storylines (Onuczko et al., 2007). Figure 1 shows a dialogue turn, taken from *The Outer Worlds* (Ob-

NPC says D, which is important for completing the quest…) and to serve a storytelling role, espousing details to the player about the game world. NPC interactions often take the form of complex trees that can have dozens of nodes, and creating these branching structures according to the many specifications of dialogue authoring can be time-

A few great/original/fun tasks

Analysis of Sensation-transfer Dialogues in Motorsports

Takeru Isaka, Atsushi Otsuka, Yoko Tokunaga*, Iwaki Tosima
NTT Digital Twin Computing Research Center
Tokyo Japan
{takeru.isaka, atsushi.otsuka, iwaki.toshima}@ntt.com

Abstract

Clarifying the effects of subjective ideas on group performance is essential for future dialogue systems to improve mutual understanding among humans and group creativity. However, there has been little focus on dialogue research on quantitatively analyzing the effects of the quality and quantity of subjective information contained in dialogues on group performance. We hypothesize that the more subjective information interlocutors exchange, the better the group performance in collaborative work. We collected dialogues between drivers and engineers in motorsports when deciding how the car should be tuned as a suitable case to verify this hypothesis. Our analysis suggests that the greater the amount of subjective information (which we defined as "sensation") in the driver's utterances, the greater the race performance and driver satisfaction with the car's tuning. The results indicate that it is essential for the development of dialogue research to create a corpus of situations that require high performance through collaboration among experts with different backgrounds but who have mastered their respective fields.

Keywords: Sensation, Collaborative Dialogue, Group Performance

1. Introduction

Humans are social creatures and share their innermost thoughts through dialogue. The active use of subjective ideas is essential to develop a dialogue system that improves mutual understanding and creativity among humans. Since subjective ideas often contain ambiguity, it is conceivable that poor comprehension conditions could lead to confusion or misdirection within the group. Understanding subjective information is also more difficult when the positions and roles of the interlocutors differ. Therefore, we examined the effect of conveying subjective ideas among people in different positions and roles on group performance during collaborative work.

For such verification, it is essential to have a dialogue resource where several people collaboratively work and actively express their subjective ideas for solving problems that cannot be solved from objective facts alone. An example of such a rare dialogue is between drivers and engineers in motorsports. Group performance in motorsports can be replaced with race performance. Motorsport drivers are professionals who drive their cars under extreme conditions (Reid and Lightfoot, 2019; Reid, 2022), under which the maximum speed of the car can reach over 300 km/h. They exhibit different neuroscientific (Bernardi et al., 2013), cognitive (Land and Tatler, 2001; Lappi, 2022), and sensorimotor (Van Leeuwen et al., 2017; Nishizono et al., 2023) characteristics while driving than non-racing drivers. They can understand the increase or decrease in lap time in 0.05-s increments, depending on how good or bad their driving is [1].

The dialogue we focus on has the following four advantages from the perspective of research execution.

- High Resolution - High-resolution sensations that cannot be expressed through basic emotion classification are expressed.
- Reproducibility - The same person has the same feeling when placed in the same situation. ·.·Drivers and engineers are consistent in their utterances as experts.
- Evaluability - Some indicators can be used to evaluate the results of sensation transfer objec-

Dialogue Evaluation

Semantic Similarity Metrics

- Common metrics:
  - BLEU, ROUGE, METEOR – compare surface forms (n-gram overlap);
  - BERTScore, BLEURT – capture semantic proximity via embeddings.
- Advantages:
  - Quick, automatic, reproducible.
- Limitations:
  - Do not account for **interactional dynamics** (turn-taking, grounding, etc.);
  - Surface similarity ≠ good conversational response.

LLMs as Evaluators

- Advantages:
  - Scalable, cost-effective;
  - Virtually infinite set of aspects can be evaluated.
- Limitations:
  - Validation of one feature (e.g. fluency) does not generalise to others (e.g. coherence, engagement);
  - Extrinsic evaluation.
- Open Questions:
  - **Representation:** whose language and norms are reflected?
  - **Transparency:** training data contamination and circular testing.

Human Intrinsic Evaluation

Ask users to provide feedback while/after using the system:

- Advantages:
  - As close as possible to the actual use of the system;
  - Multiple features can be evaluated (both task- and conversation related).
- Limitations
  - Expensive (time and money);
  - Requires clear definitions and formalised evaluation schemes;
  - Ethical considerations w.r.t. crowdwork platforms (Amazon Mechanical Turk).

Human Extrinsic Evaluation

Ask human evaluators to rate a dialogue:

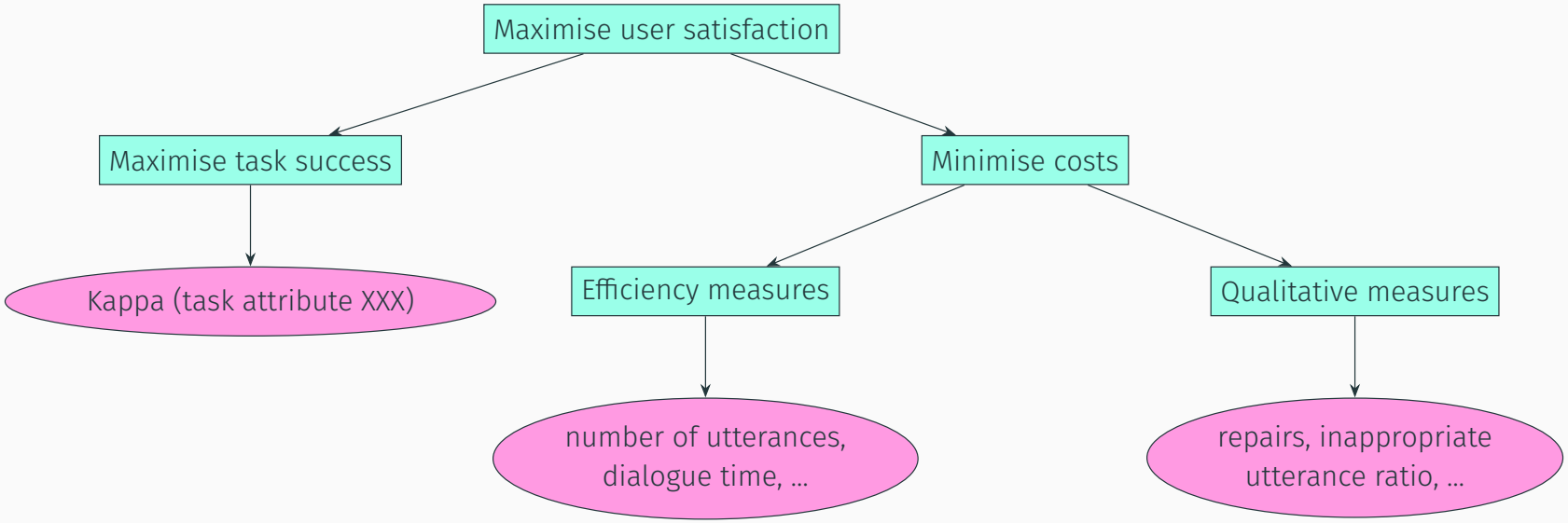
- Advantages:
  - Can have an intuition of the actual use of the system;
  - Multiple features can be evaluated (both task- and conversation related).
- Limitations:
  - Expensive (time and money);
  - Requires clear definitions and formalised evaluation schemes;
  - Ethical considerations w.r.t. crowdwork platforms (e.g. Amazon Mechanical Turk).
- **Open question:** Are humans really good at extrinsic evaluation?
  - Lack of exposure to conversation transcripts;
  - Good adaptation capability.

$\rightarrow$ Parallels with psychological test design: reliability and bias control.

The PARADISE Framework

- **Principle:**
  - Model-based evaluation framework (Walker et al., 1997) for **task-oriented dialogue systems**.
  - Predicts **user satisfaction** from measurable features.
- **Core Idea:**
  - Dialogue quality = *Task success* + *Conversational efficiency*.
  - Efficiency measured via dialogue costs: time, number of turns, repair rates, etc.
- **Limitations:**
  - Suited to structured, goal-oriented systems – less to open dialogue.
  - Not initially designed for subjective and social aspects of conversation.

The PARADISE Framework



Graph from the original PARADISE papers (Walker et al., 1997)

The PARADISE Framework - Performance function (simplified)

$$P = \alpha * \clubsuit + \sum_i w_i * \diamondsuit_i$$

- $\clubsuit$: task efficiency
- $\diamondsuit$: costs (as many as you want)
- α, w_i: weights to estimate with real user provided performance values

The PARADISE Framework - Task efficiency

Example

A1: Hello, I would like to order a **Toasted** sandwich please.
B2: Sure, medium or large?
A3: **Medium**.
B4: Where would you like to pick it up?
A5: At **ARTEM** please.
B6: No problem, what time?
A7: **12:30**
B8: Noted. A medium toasted at ARTEM for 12:30. Thank you!

The PARADISE Framework - Task efficiency

Example

A1: Hello, I would like to order a **Toasted** sandwich please.
B2: Sure, medium or large?
A3: **Medium**.
B4: Where would you like to pick it up?
A5: At **ARTEM** please.
B6: No problem, what time?
A7: **12:30**
B8: Noted. A large toasted at ARTEM for 12:30. Thank you!

The PARADISE Framework - Task efficiency

	Key											
	Sandwich			Size		Location				Time		
Data	V_1	V_2	V_3	V_1	V_2	V_1	V_2	V_3	V_4	V_1	V_2	V_3
V_1	**23**	2	1	0	0	0	0	0	0	0	1	1
V_2	2	**10**	3	0	0	0	0	0	0	0	0	0
V_3	1	2	**15**	0	1	1	0	0	1	0	0	0
V_1	0	0	0	**45**	3	0	0	0	0	0	0	0
V_2	1	0	0	1	**18**	0	0	0	0	0	0	0
V_1	0	0	0	0	0	**22**	0	1	5	0	0	0
V_2	0	0	0	0	0	0	**4**	2	0	0	0	0
V_3	0	0	0	2	0	0	0	**21**	0	0	0	0
V_4	0	0	0	0	0	4	0	0	**9**	0	0	0
V_1	0	0	0	0	0	0	0	0	0	**15**	1	3
V_2	0	0	0	0	0	0	0	0	0	1	**24**	1
V_3	0	0	0	0	0	0	0	0	0	4	1	**18**
Sum	27	14	19	48	22	27	4	24	15	20	27	23

The PARADISE Framework - Cost functions

	Dialogue	Length	Repairs
A1:	Hello, I would like to order a **Toasted** sandwich please.	1	0
B2:	Sure, medium or large?	1	0
A3:	**Medium**.	1	0
B4:	Where would you like to pick it up?	1	0
A5:	At **ARTEM** please.	1	0
B6:	No problem, what time?	1	0
A7:	**12:30**	1	0
B8:	Noted. A large toasted at ARTEM for 12:30. Thank you!	1	0
A9:	Actually I asked for a medium one.	1	1
B10:	Oh sorry, medium toasted at ARTEM for 12:30.	1	1
	Sum:	10	2/10

The PARADISE Framework - Let’s try it

- Our task: Order a sandwich to pick-up from one of the MadeInFrance restaurants in Nancy.

The PARADISE Framework - Let's try it

- Our task: Order a sandwich to pick-up from one of the MadeInFrance restaurants in Nancy.
- Our attributes:

Attribute	Possible values
sandwich	88, Toasted, Salmon
size	medium, large
pickup point	St-Epvre, Velodrome, ARTEM, St-Georges
pickup time	11:30, 12:00, 12:30

The PARADISE Framework - Let's try it

Dialogue 1

A1: Hello! I would like to order a sandwich please.
B2: Sure! What sandwich would you like?
A3: Uh, an 88 please, and in medium size.
B4: Okay where do you want to pick it up?
A5: Uh... At Velodrome.
B6: Oh but I'm sorry you can't do this, its under construction right now.
A7: Ah okay. So at ARTEM please.
B8: Okay and when do you want to pick it up?
A9: At 12:00.
B10: Okay so we have a medium 88 at 12:00 at ARTEM. Thank you!
A11: Thank you, bye.

The PARADISE Framework - Let's try it

Dialogue 2

A1: Hi! I'm Bob, I would like to order a sandwich.
B2: Sure what sandwich would you like?
A3: Do you have one without cheese?
B4: Hum we have the 88 and the salmon.
A5: So I'll take an 88.
B6: Okay. What size do you want? We have medium and large.
A7: Uhm medium.
B8: Where do you want to pick it up?
A9: Uh is ARTEM possible?
B10: Yes sure.
A11: So at ARTEM. I would like to pick it up at 11:30.
B12: Alright noted. A medium 88 at ARTEM at 11:30.

The PARADISE Framework - Let's try it

Dialogue 3

A1: Hello, do you have toilets?
B2: I'm just receiving the orders for the different restaurants in town. But they have toilets accessible from outside at St-Epvre.
A3: Okay. Good good good. What is the cheapest sandwich?
B4: We have the toasted sandwich.
A5: What?
B6: Toasted.
A7: What's inside?
B8: Cheese and... cheese.
A9: Okay. Uhhh. Okay. I'll take that please.
B10: What size do you want?
A11: I want the cheapest sandwich.
B12: Medium then. Where do you want to pick it up? It's the same price.
A13: St-Epvre.
B14: Okay when do you want to pick it up?
A15: As soon as possible.
B16: Alright, a med-
A17: [Thank you bye.

The PARADISE Framework - Let's try it

Dialogue 4

A1: Hi! I would like to order a sandwich. I would like a salmon sandwich size medium, and I would like to pick it up at 12:00 at St-Georges please.

B2: Okay good, do you want anything else?

A3: No thank you.

B4: Okay, medium salmon, St-Georges, 12:00. Thank you.

The PARADISE Framework - Let's try it

Dialogue 5

A1: Hello, do you sell pizzas?

B2: Hi. No we only have sandwiches, do you want to order one?

A3: No thank you. Bye.

B4: Bye.

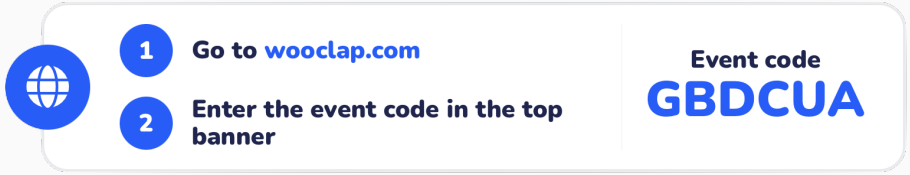
The PARADISE Framework - Let's try it

Dialogue 6

A1: Hello, I'd like a medium salmon sandwich please.
B2: Which location would you pick it up from?
A3: St-Georges.
B4: And the pickup time?
A5: 11:30.
B6: Got it, salmon sandwich at St-Epvre at 11:30.

The PARADISE Framework - Let’s try it

Let’s grade them!



The PARADISE Framework - Let's try it

Test Dialogue

A1: Hi, I'd like a toasted sandwich.
B2: Sure! Which pickup point would you like?
A3: ARTEM.
B4: And at what time?
A5: 12:30.
B6: Okay, toasted sandwich at ARTEM at 12:00.
A7: Actually I said 12:30.
B8: Oh my apologies! Toasted sandwich at ARTEM at 12:30.

1. Task efficiency?
2. Costs (length, repairs)?

The PARADISE Framework - Limitations

- Expensive:
  - Requires initial fitting corpus;
  - Requires human performance evaluation.
- More realistic model $\rightarrow$ more cost functions $\rightarrow$ larger fitting corpus;
- Method to get human performance grade?
  - One global grade $\rightarrow$ meaning?
  - Several sub-grades $\rightarrow$ combination?

    $\Rightarrow$ **In the end we still don't know what humans really grade**.

Comparing Evaluation Approaches

Approach	Focus	Advantages	Limitations
Semantic Similarity	Textual overlap / embedding similarity	Fast, automatic, reproducible	Ignores interaction, context, and intent
LLM Benchmarks	Virtually any	Easy to scale, often multi-metric	Bias, lack of transparency, unclear what "good" means
Human Intrinsic Evaluation	Virtually any: coherence, grammar, interaction, engagement, empathy, ...	Use-based, closer to actual in-interaction perception	Costly, requires a lot of time and organisation
Human Extrinsic Evaluation		Rich insights, possible links to real dialogue theory	Still costly, reflects subjective perception of what a dialogue should be
PARADISE	Task success + efficiency (user satisfaction model)	Quantitative, interpretable (somewhat)	Costly, limited to structured tasks, cost functions not always so simple

Conclusion

Summary

- Dialogue evaluation remains a major open challenge:
  - Current automatic (non LLM-based) metrics capture form, not function;
  - Precise definitions of dialogue features are necessary to build reliable evaluation methods...
  - It is all the more important when using opaque evaluators such as LLMs to make sure that we evaluate what we wanted to;
  - But this point is also valid for human evaluation.
- LLMs create a new data issue:
  - No matter how good a resource is, it will be absorbed by the fast-evolving models and will no longer serve as a clean benchmark for future evaluation;
  - This calls for new efficient resource collection methods.

Overall Take-home Message

- Dialogue systems are spreading everywhere for many kinds of tasks... including many we can perfectly solve without dialogue systems.
- Building a good quality dialogue system calls for:
  - Assessing actual needs to precisely define expectations and requirements;
  - Finding relevant inspiration/training resources;
  - Building evaluation approaches that reflect real expectations and use.
- This requires to have an idea of what a good conversation is in a given situation...
  - But there is no established reference point for this yet.
  - We do not have a formalised, widely accepted evaluation framework – everyone does their own thing, often without verifying that the evaluation actually measures what matters.

References i

Banerjee, Satanjeev and Alon Lavie (June 2005). **"METEOR: An Automatic Metric for MT Evaluation with Improved Correlation with Human Judgments".** In: *Proceedings of the ACL Workshop on Intrinsic and Extrinsic Evaluation Measures for Machine Translation and/or Summarization*. Ed. by Jade Goldstein et al. Ann Arbor, Michigan: Association for Computational Linguistics, pp. 65–72.

Kim, Minju, Heuiyeen Yeen, and Myoung-Wan Koo (Mar. 2024). **"Towards Context-Based Violence Detection: A Korean Crime Dialogue Dataset".** In: *Findings of the Association for Computational Linguistics: EACL 2024*. Ed. by Yvette Graham and Matthew Purver. St. Julian's, Malta: Association for Computational Linguistics, pp. 603–623.

Lin, Chin-Yew (July 2004). **"ROUGE: A Package for Automatic Evaluation of Summaries".** In: *Text Summarization Branches Out*. Barcelona, Spain: Association for Computational Linguistics, pp. 74–81.

References ii

 Papineni, Kishore et al. (July 2002). **“Bleu: a Method for Automatic Evaluation of Machine Translation”.** In: *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics*. Ed. by Pierre Isabelle, Eugene Charniak, and Dekang Lin. Philadelphia, Pennsylvania, USA: Association for Computational Linguistics, pp. 311–318.

 Walker, Marilyn A. et al. (July 1997). **“PARADISE: A Framework for Evaluating Spoken Dialogue Agents”.** In: *35th Annual Meeting of the Association for Computational Linguistics and 8th Conference of the European Chapter of the Association for Computational Linguistics*. Madrid, Spain: Association for Computational Linguistics, pp. 271–280.

Some Dialogue Resources i

- Anantha, Raviteja et al. (June 2021). **“Open-Domain Question Answering Goes Conversational via Question Rewriting”.** In: *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*. Ed. by Kristina Toutanova et al. Online: Association for Computational Linguistics, pp. 520–534.
- Budzianowski, Paweł et al. (2018). **“MultiWOZ - A Large-Scale Multi-Domain Wizard-of-Oz Dataset for Task-Oriented Dialogue Modelling”.** In: *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*. Ed. by Ellen Riloff et al. Brussels, Belgium: Association for Computational Linguistics, pp. 5016–5026.
- Busso, Carlos et al. (2008). **“IEMOCAP: Interactive emotional dyadic motion capture database”.** In: *Language resources and evaluation* 42.4, pp. 335–359.

Some Dialogue Resources ii

- Dinan, Emily et al. (2019). **“Wizard of Wikipedia: Knowledge-powered Conversational Agents”.** In: *Proceedings of the International Conference on Learning Representations (ICLR)*. ICLR.
- Eric, Mihail et al. (May 2020). **“MultiWOZ 2.1: A Consolidated Multi-Domain Dialogue Dataset with State Corrections and State Tracking Baselines”.** eng. In: *Proceedings of the Twelfth Language Resources and Evaluation Conference*. Ed. by Nicoletta Calzolari et al. Marseille, France: European Language Resources Association, pp. 422–428.
- Gerlach, Johanna et al. (May 2024). **“A Concept Based Approach for Translation of Medical Dialogues into Pictographs”.** In: *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*. Ed. by Nicoletta Calzolari et al. Torino, Italia: ELRA and ICCL, pp. 233–242.

Some Dialogue Resources iii

 Lee, Harrison et al. (June 2022). **“SGD-X: A Benchmark for Robust Generalization in Schema-Guided Dialogue Systems”.** In: *Proceedings of the AAAI Conference on Artificial Intelligence* 36.10, pp. 10938–10946.

 Li, Yanran et al. (Nov. 2017). **“DailyDialog: A Manually Labelled Multi-turn Dialogue Dataset”.** In: *Proceedings of the Eighth International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*. Ed. by Greg Kondrak and Taro Watanabe. Taipei, Taiwan: Asian Federation of Natural Language Processing, pp. 986–995.

Some Dialogue Resources iv

 Ou, Jiao et al. (June 2024). **“DialogBench: Evaluating LLMs as Human-like Dialogue Systems”.** In: *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*. Ed. by Kevin Duh, Helena Gomez, and Steven Bethard. Mexico City, Mexico: Association for Computational Linguistics, pp. 6137–6170.

 Poria, Soujanya et al. (July 2019). **“MELD: A Multimodal Multi-Party Dataset for Emotion Recognition in Conversations”.** In: *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*. Ed. by Anna Korhonen, David Traum, and Lluís Màrquez. Florence, Italy: Association for Computational Linguistics, pp. 527–536.

Some Dialogue Resources v

- Rastogi, Abhinav et al. (Apr. 2020). **“Towards Scalable Multi-Domain Conversational Agents: The Schema-Guided Dialogue Dataset”.** In: vol. 34. 05. AAAI, pp. 8689–8696.
- Reece, Andrew et al. (2023). **“The CANDOR corpus: Insights from a large multimodal dataset of naturalistic conversation”.** In: *Science Advances* 9.13, eadf3197. eprint: `https://www.science.org/doi/pdf/10.1126/sciadv.adf3197`.
- Weir, Nathaniel et al. (Nov. 2024). **“Ontologically Faithful Generation of Non-Player Character Dialogues”.** In: *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*. Ed. by Yaser Al-Onaizan, Mohit Bansal, and Yun-Nung Chen. Miami, Florida, USA: Association for Computational Linguistics, pp. 9212–9242.

Some Dialogue Resources vi

 Zang, Xiaoxue et al. (July 2020). **“MultiWOZ 2.2 : A Dialogue Dataset with Additional Annotation Corrections and State Tracking Baselines”.** In: *Proceedings of the 2nd Workshop on Natural Language Processing for Conversational AI*. Ed. by Tsung-Hsien Wen et al. Online: Association for Computational Linguistics, pp. 109–117.