BETOLD: A Task-Oriented Dialog Dataset for Breakdown Detection

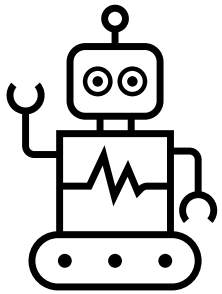
Silvia Terragni, Bruna Guedes, Andre Manso, Modestas Filipavicius, Nghia Khau, Roland Mathis

Telepathy Labs, Zürich

silvia.terragni@telepathy.ai
@TerragniSilvia

When creative AI meets conversational AI @COLING 2022

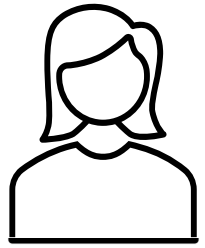
What is a Dialog Breakdown?



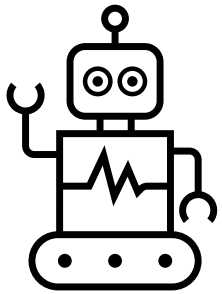
When would you like to bring in your phone device?

I'm sorry, there are no appointments available on Tuesday 18th October

I want to come Next Monday



What is a Dialog Breakdown?



When would you like to bring in your phone device?

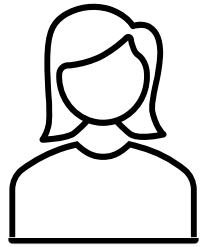
I'm sorry, there are no appointments available on Tuesday 18th October

When would you like to bring in your phone device?

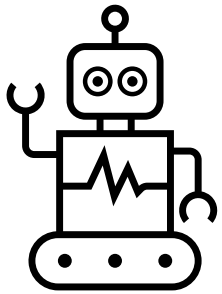
I'm sorry, there are no appointments available on Tuesday 18th October

I want to come Next Monday

I said next Monday



What is a Dialog Breakdown?



When would you like to bring in your phone device?

I'm sorry, there are no appointments available on Tuesday 18th October

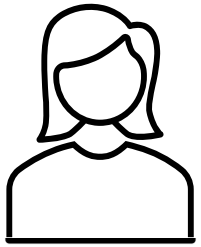
When would you like to bring in your phone device?

I'm sorry, there are no appointments available on Tuesday 18th October

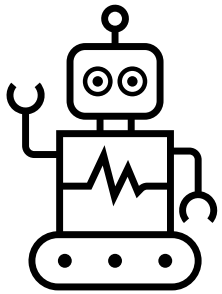
I want to come Next Monday

I said next Monday

I want to talk to human



What is a Dialog Breakdown?



When would you like to bring in your phone device?

I'm sorry, there are no appointments available on Tuesday 18th October

When would you like to bring in your phone device?

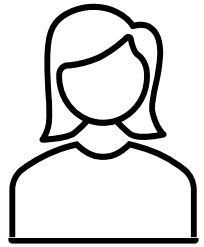
I'm sorry, there are no appointments available on Tuesday 18th October

breakdown

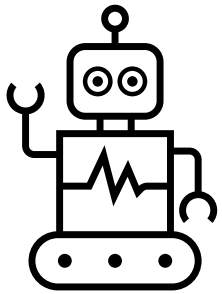
I want to come Next Monday

I said next Monday

I want to talk to human



Why Detecting Dialog Breakdowns?



When would you like to bring in your phone device?

I want to come Next Monday

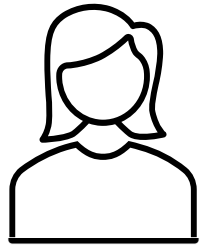
I'm sorry, there are no appointments available on Tuesday 18th October

When would you like to bring in your phone device?

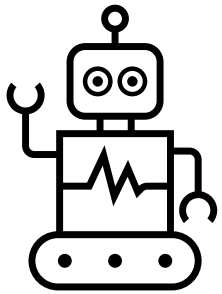
I said next Monday

I'm sorry, there are no appointments available on Tuesday 18th October

breakdown



Why Detecting Dialog Breakdowns?



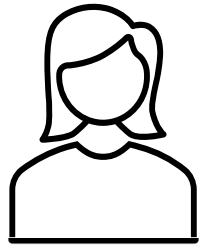
When would you like to bring in your phone device?

I want to come Next Monday

I'm sorry, there are no appointments available on Tuesday 18th October

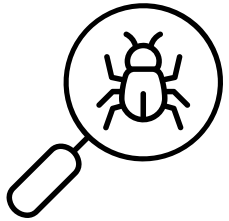
When would you like to bring in your phone device?

I said next Monday



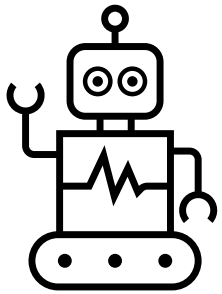
I'm sorry, there are no appointments available on Tuesday 18th October

breakdown



1) Report:
Prioritize conversations when breakdowns happen

Why Detecting Dialog Breakdowns?



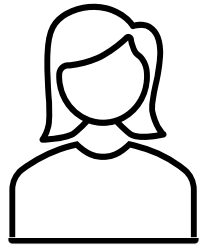
When would you like to bring in your phone device?

I want to come Next Monday

I'm sorry, there are no appointments available on Tuesday 18th October

When would you like to bring in your phone device?

I said next Monday



breakdown

2) Recovery strategies

Sorry, could you rephrase it?

Dialog Breakdown Detection Challenge (DBDC) Dataset

- Small dataset (615 dialogs)
- Manually annotated
- Open domain dialogs
- Textual utterances

Proposed dataset: BETOLD

DBDC

- Small dataset (615 dialogs)
- Manually annotated
- Open domain dialogs
- Textual utterances

BETOLD

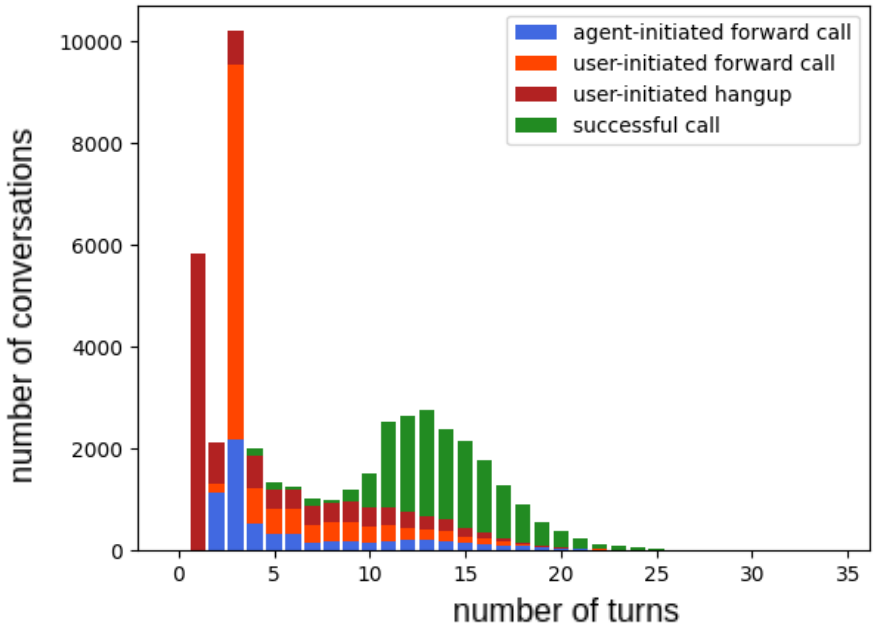
- Large dataset (13k dialogs)
- Automatically annotated
- Task-oriented Dialogs
- NLG and NLU intents and entities for privacy preservation

BETOLD Dataset Creation

- 4 types of call endings:
  - **Successful calls**
  - Agent-initiated forwarded calls
  - **User-initiated forwarded calls**
  - **User-initiated hang-ups**

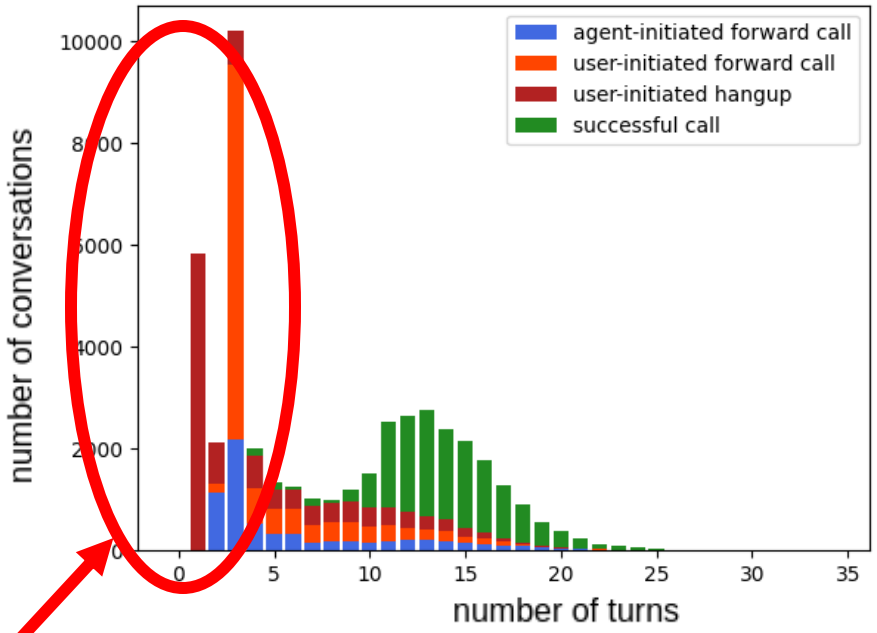
BETOLD Dataset Creation

- 4 types of call endings:
  - **Successful calls**
  - Agent-initiated forwarded calls
  - **User-initiated forwarded calls**
  - **User-initiated hang-ups**



BETOLD Dataset Creation

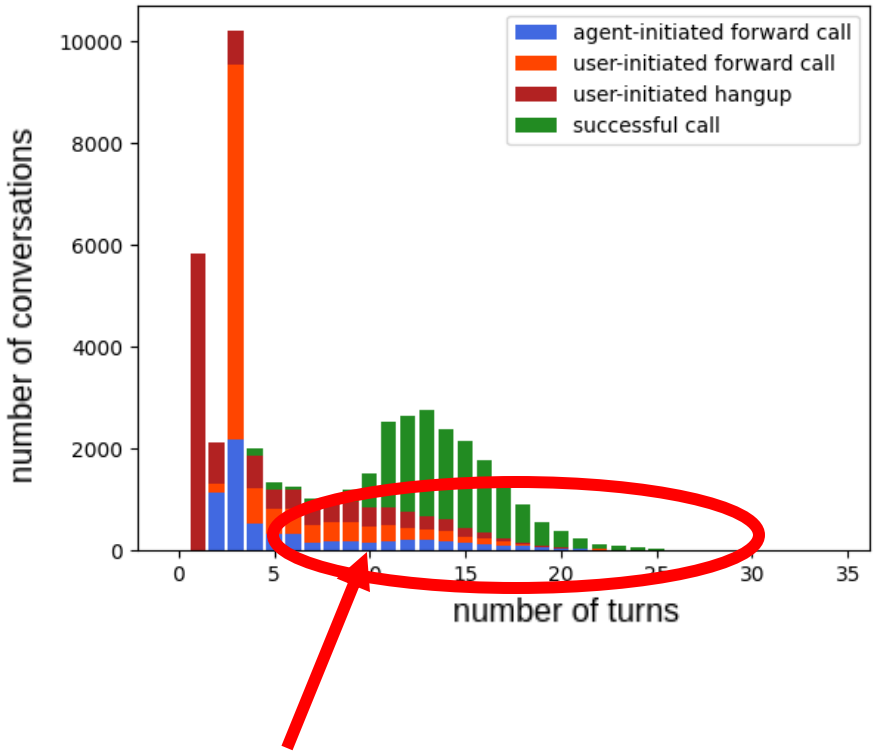
- 4 types of call endings:
  - **Successful calls**
  - Agent-initiated forwarded calls
  - **User-initiated forwarded calls**
  - **User-initiated hang-ups**



no way to prevent the caller
from hanging up in the initial turns

BETOLD Dataset Creation

- 4 types of call endings:
  - **Successful calls**
  - Agent-initiated forwarded calls
  - **User-initiated forwarded calls**
  - **User-initiated hang-ups**

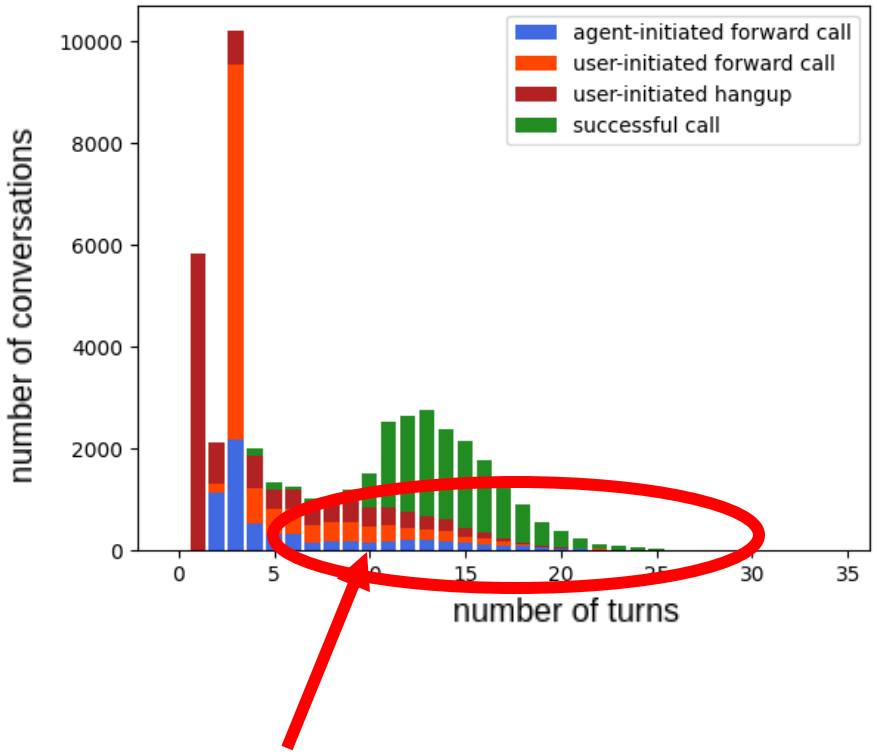


LUHFs: Late User-initiated Hang-ups or Forward calls

We will focus on this category of dialog breakdowns

BETOLD Dataset Creation

- 4 types of call endings:
  - **Successful calls** → **Not LUHFS**
  - Agent-initiated forwarded calls
  - **User-initiated forwarded calls** → **LUHFS**
  - **User-initiated hang-ups** → **LUHFS**



LUHFs: Late User-initiated Hang-ups or Forward calls

We will focus on this category of dialog breakdowns

BETOLD Dataset Statistics

- We consider the LUHFs (with more than 8 turns) out of about 35k conversations
- We sample from the successful conversations to obtain the not LUHFs

Labels	LUHF	4508
	not LUHF	9016
Turns	min	8
	avg	10
	max	34

Example of a conversation

Phone Repairing Booking Agent
Can I book your service appointment under the phone number ending in <2385>?
Yeah, that's correct
What is the brand, model and year of your phone device?
It's a <phonepink> why 100 <2022>
What is the model of your phone device?
It is <Y100>
What is the battery health percentage of your phone device?
<zero>
Great, what is your name?
My name's <John>

The user may release **sensitive information**, such as the name and phone number

Example of a conversation

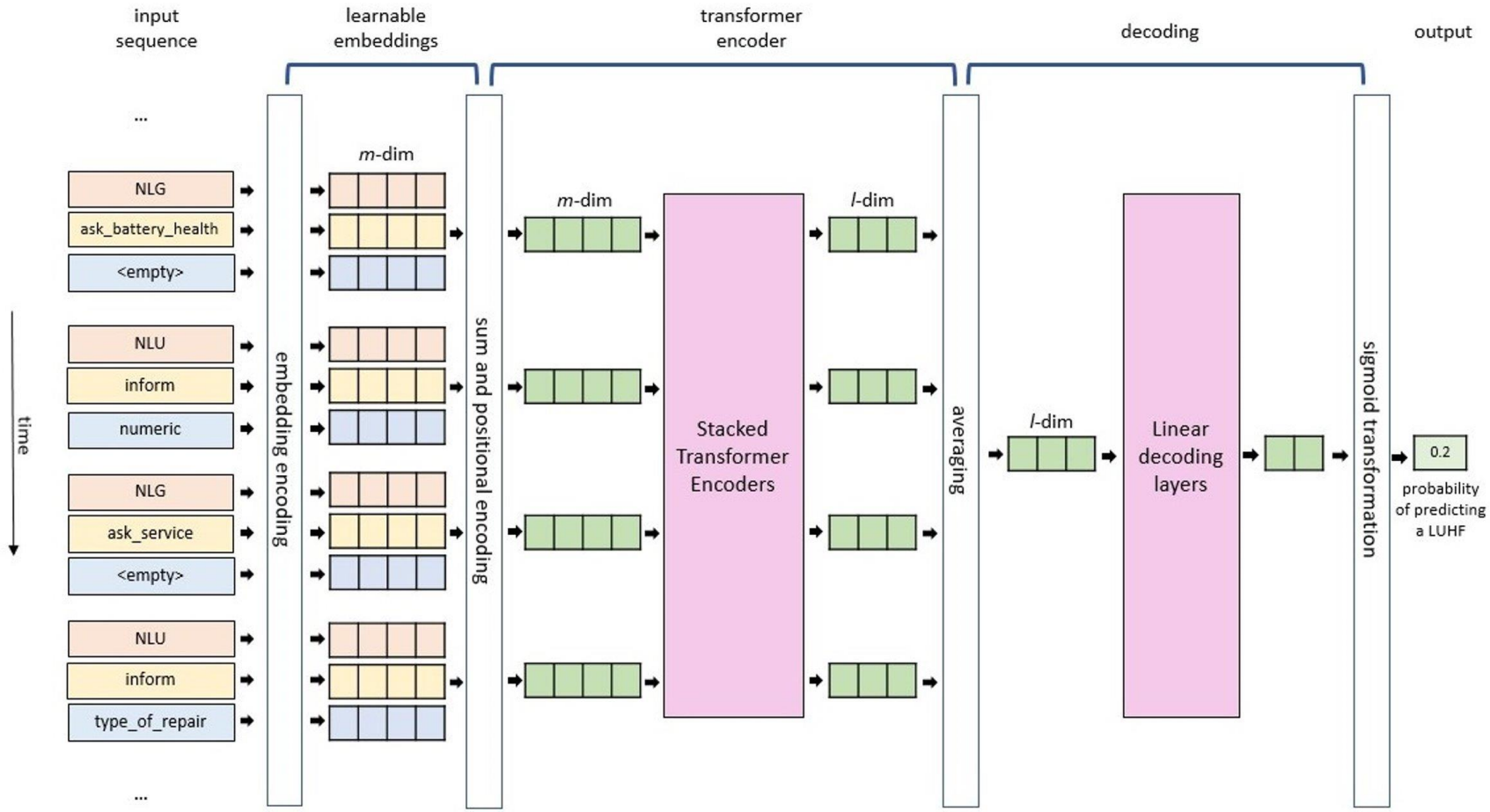
Phone Repairing Booking Agent	Caller	Intents	Entities
Can I book your service appointment under the phone number ending in <2385>?	NLG	confirm_phone_number	userphone_suffix
Yeah, that's correct	NLU	confirm	
What is the brand, model and year of your phone device?	NLG	new_user_profile_branch_model_year	
It's a <phonepink> why 100 <2022>	NLU	inform	brand_device, year
What is the model of your phone device?	NLG	ask_device_model	
It is <Y100>	NLU	inform	model_device
What is the battery health percentage of your phone device?	NLG	ask_battery_health	
<zero>	NLU	inform	numeric
Great, what is your name?	NLG	ask_first_name	
My name's <John>	NLU	inform	client_name

Example of a conversation

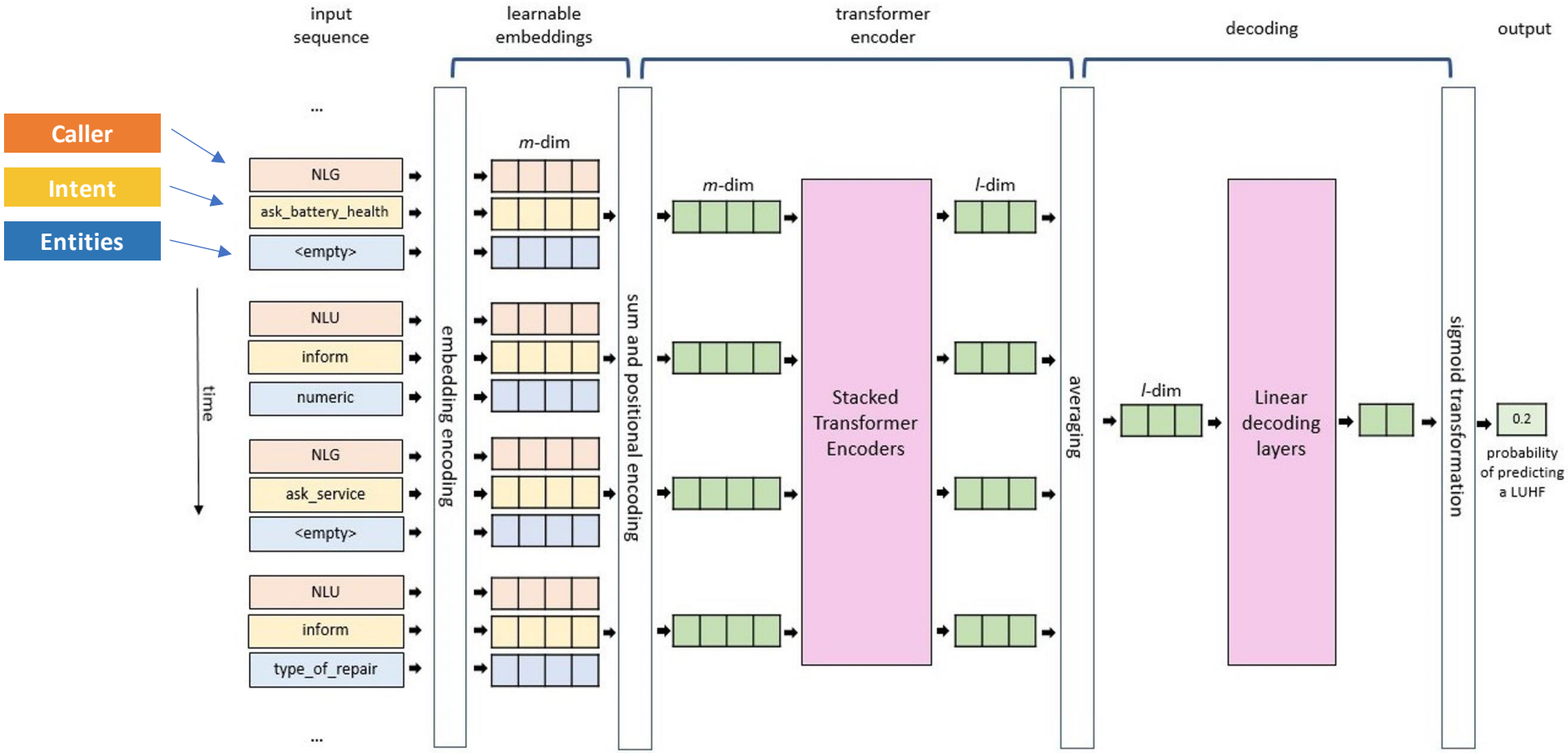
Phone Repairing Booking Agent	Caller	Intents	Entities
Can I book your service appointment under the phone number ending in <2385>?	NLG	confirm_phone_number	userphone_suffix
Yeah, that's [illegible]rrect	NLU	confirm	
What is the b[illegible] model and yea[illegible] your phone device?	NLG	new_user_profile_branch_model_year	
It's a <phonepink> [illegible]100 <[illegible]2>	NLU	inform	brand_device, year
What is the model of yo[illegible]one device?	NLG	ask_device_model	
It is <Y100>	NLU	inform	model_device
What is the batte[illegible]ealth per[illegible]age of your phone device?	NLG	ask_battery_health	
<zero>	NLU	inform	numeric
Great, what is your name?	NLG	ask_first_name	
My name's <John>	NLU	inform	client_name

Can we predict a LUHF with only NLU and NLG annotations?

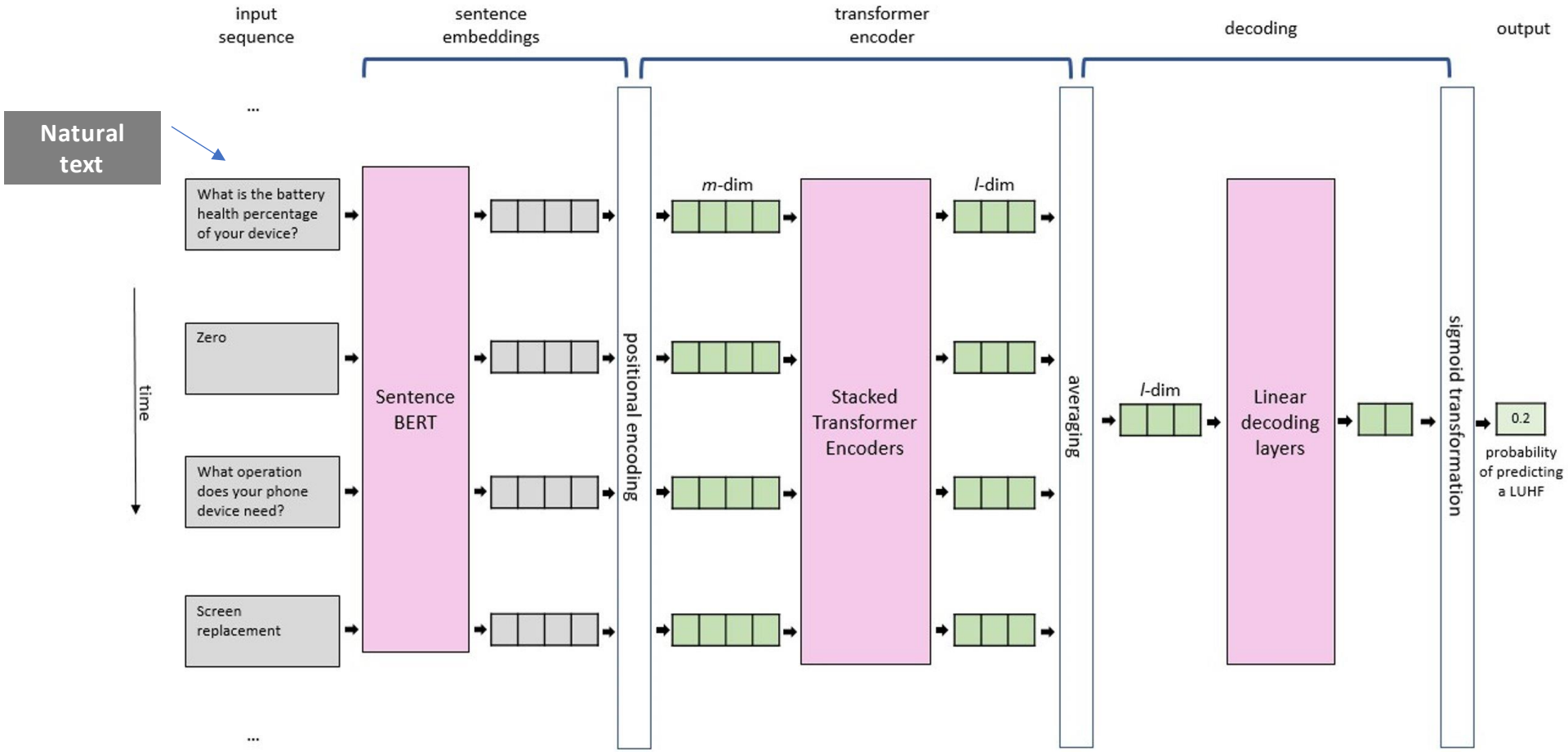
Proposed Model Architecture



Proposed Model Architecture



Baseline: Text-only Model

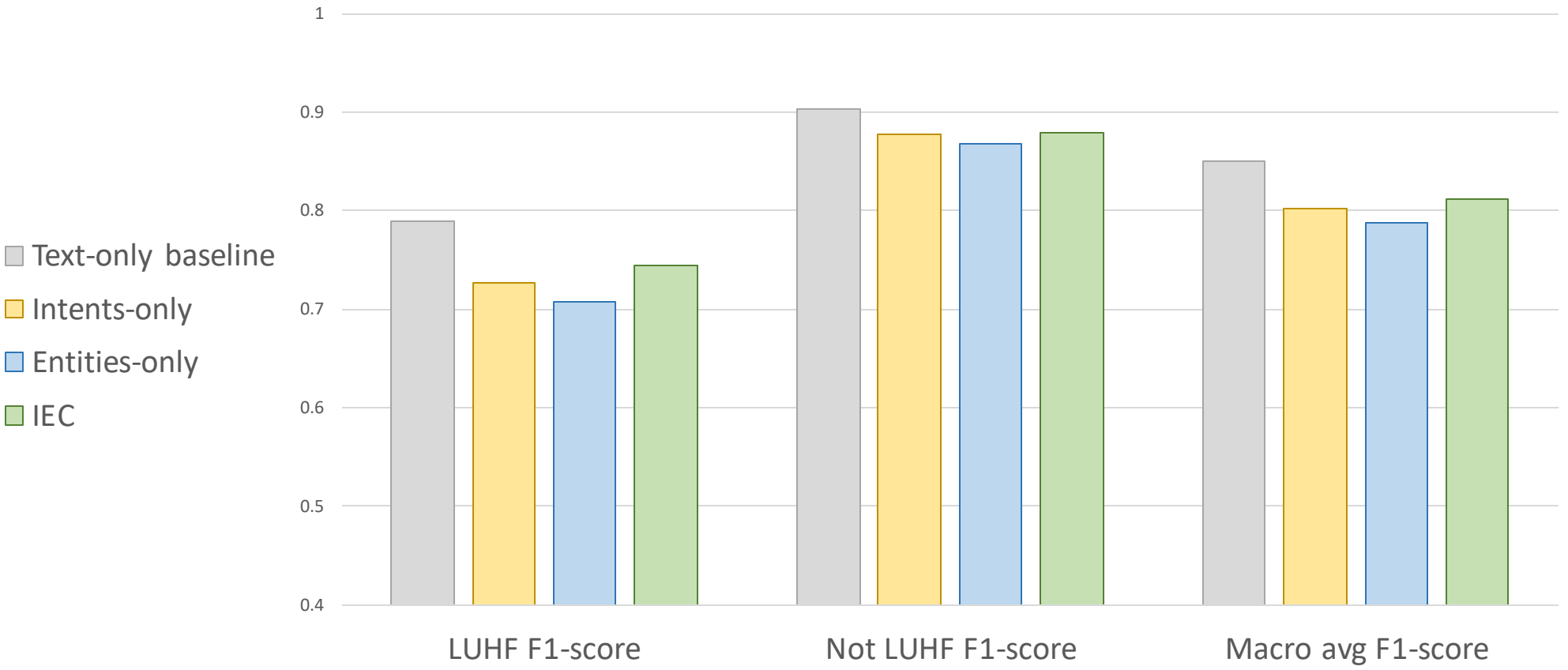


Experimental Setting

- Models:
  - Text-only baseline
  - Entities only
  - Intents only
  - IEC (Entities+Intents+Callers)

- Grid search to determine the optimal hyperparameters

Classification Results



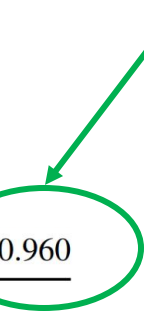
Generalization capabilities

- LUHF annotation on BETOLD applies to the overall call (not applied to each step of the conversation)
- Can the model predict if a breakdown happened at each step of a conversation?

Sample conversation

Step	Caller	Intent	Entities	Probability of LUHF
9	NLG	ask_for_battery_health		
10	NLU	inform	numeric	
11	NLG	ask_first_name		
12	NLU	inform	client_name	
13	NLG	ask_last_name		
14	NLU	inform	client_name	
15	NLG	ask_desired_service		
16	NLU	user_initial_request	type_of_repair	
17	NLG	ask_additional_service		
18	NLU	inconclusive		
19	NLG	transportation_of_device	ask_to_schedule, ask_means_of_transportation	
20	NLU	confirm		
21	NLG	inform_schedule_inspection		
22	NLG	propose_date	transportation_type_selection, available_slot_to_schedule	
23	NLU	negate		
24	NLG	ask_time_preference		
25	NLU	user_proposed_date	time_range_indication	
26	NLG	time_asked_unavailable_propose_new	transportation_type_selection, available_slot_to_schedule, user_request_start_time	
27	NLU	negate		
28	NLG	ask_time_preference		0.960

Probability to predict a LUHF for the overall call



Sample conversation – how to generalize

Step	Caller	Intent	Entities	Probability of LUHF
9	NLG	ask_for_battery_health		
10	NLU	inform	numeric	
11	NLG	ask_first_name		
12	NLU	inform	client_name	

We can split iteratively the conversation
in steps – until step *n*
(with *n*=1, …, length of the conversation)

Sample conversation – how to generalize

Step	Caller	Intent	Entities	Probability of LUHF
9	NLG	ask_for_battery_health		
10	NLU	inform	numeric	
11	NLG	ask_first_name		
12	NLU	inform	client_name	0.002

We can split iteratively the conversation in steps – until step *n* (with *n*=1, …, length of the conversation)

And then predict the probability of a LUHF given the conversation until step *n*

Sample conversation – how to generalize

Step	Caller	Intent	Entities	Probability of LUHF
9	NLG	ask_for_battery_health		
10	NLU	inform	numeric	
11	NLG	ask_first_name		
12	NLU	inform	client_name	0.002
13	NLG	ask_last_name		0.296

Sample conversation – how to generalize

Step	Caller	Intent	Entities	Probability of LUHF
9	NLG	ask_for_battery_health		
10	NLU	inform	numeric	
11	NLG	ask_first_name		
12	NLU	inform	client_name	0.002
13	NLG	ask_last_name		0.296
14	NLU	inform	client_name	0.016

Sample conversation – how to generalize

Step	Caller	Intent	Entities	Probability of LUHF
9	NLG	ask_for_battery_health		
10	NLU	inform	numeric	
11	NLG	ask_first_name		
12	NLU	inform	client_name	0.002
13	NLG	ask_last_name		0.296
14	NLU	inform	client_name	0.016
15	NLG	ask_desired_service		0.344

Sample conversation – how to generalize

Step	Caller	Intent	Entities	Probability of LUHF
9	NLG	ask_for_battery_health		
10	NLU	inform	numeric	
11	NLG	ask_first_name		
12	NLU	inform	client_name	0.002
13	NLG	ask_last_name		0.296
14	NLU	inform	client_name	0.016
15	NLG	ask_desired_service		0.344
16	NLU	user_initial_request	type_of_repair	0.033

Qualitative Analysis

Step	Caller	Intent	Entities	Probability of LUHF
9	NLG	ask_for_battery_health		0.002
10	NLU	inform	numeric	0.000
11	NLG	ask_first_name		0.109
12	NLU	inform	client_name	0.002
13	NLG	ask_last_name		0.296
14	NLU	inform	client_name	0.016
15	NLG	ask_desired_service		0.344
16	NLU	user_initial_request	type_of_repair	0.033
17	NLG	ask_additional_service		0.262
18	NLU	inconclusive		0.012
19	NLG	transportation_of_device	ask_to_schedule, ask_means_of_transportation	0.253
20	NLU	confirm		0.014
21	NLG	inform_schedule_inspection		0.012
22	NLG	propose_date	transportation_type_selection, available_slot_to_schedule	0.269
23	NLU	negate		0.033
24	NLG	ask_time_preference		0.378
25	NLU	user_proposed_date	time_range_indication	0.153
26	NLG	time_asked_unavailable_propose_new	transportation_type_selection, available_slot_to_schedule, user_request_start_time	0.922
27	NLU	negate		0.764
28	NLG	ask_time_preference		0.960

Qualitative Analysis

Step	Caller	Intent	Entities	Probability of LUHF
9	NLG	ask_for_battery_health		0.002
10	NLU	inform	numeric	0.000
11	NLG	ask_first_name		0.109
12	NLU	inform	client_name	0.002
13	NLG	ask_last_name		0.296
14	NLU	inform	client_name	0.016
15	NLG	ask_desired_service		0.344
16	NLU	user_initial_request	type_of_repair	0.033
17	NLG	ask_additional_service		0.262
18	NLU	inconclusive		0.012
19	NLG	transportation_of_device	ask_to_schedule, ask_means_of_transportation	0.253
20	NLU	confirm		0.014
21	NLG	inform_schedule_inspection		0.012
22	NLG	propose_date	transportation_type_selection, available_slot_to_schedule	0.269
23	NLU	negate		0.033
24	NLG	ask_time_preference		0.378
25	NLU	user_proposed_date	time_range_indication	0.153
26	NLG	time_asked_unavailable_propose_new	transportation_type_selection, available_slot_to_schedule, user_request_start_time	0.922
27	NLU	negate		0.764
28	NLG	ask_time_preference		0.960

the LUHF probability often decreases after a user input

Qualitative Analysis

Step	Caller	Intent	Entities	Probability of LUHF
9	NLG	ask_for_battery_health		0.002
10	NLU	inform	numeric	0.000
11	NLG	ask_first_name		0.109
12	NLU	inform	client_name	0.002
13	NLG	ask_last_name		0.296
14	NLU	inform	client_name	0.016
15	NLG	ask_desired_service		0.344
16	NLU	user_initial_request	type_of_repair	0.033
17	NLG	ask_additional_service		0.262
18	NLU	inconclusive		0.012
19	NLG	transportation_of_device	ask_to_schedule, ask_means_of_transportation	0.253
20	NLU	confirm		0.014
21	NLG	inform_schedule_inspection		0.012
22	NLG	propose_date	transportation_type_selection, available_slot_to_schedule	0.269
23	NLU	negate		0.033
24	NLG	ask_time_preference		0.378
25	NLU	user_proposed_date	time_range_indication	0.153
26	NLG	time_asked_unavailable_propose_new	transportation_type_selection, available_slot_to_schedule, user_request_start_time	0.922
27	NLU	negate		0.764
28	NLG	ask_time_preference		0.960

I'm sorry there isn't an opening <tomorrow>. Our closest available appointment is <on Wednesday 15th>, for <dropoff>. Will that work for you?

Conclusions

- Simple way to automatically generate a breakdown detection dataset
- Released an NLG/NLU-annotated dataset
  BETOLD: github.com/telepathylabsai/BETOLD_dataset
- Proposed an attention-based classifier trained on
  BETOLD: github.com/telepathylabsai/dialog_breakdown_detection
- NLU and NLG annotations are useful for the prediction of a breakdown while preserving the privacy of the data

Thank you for your attention :)

silvia.terragni@telepathy.ai
@TerragniSilvia