ZéCarioca: A framework for end-to-end chatbots creation

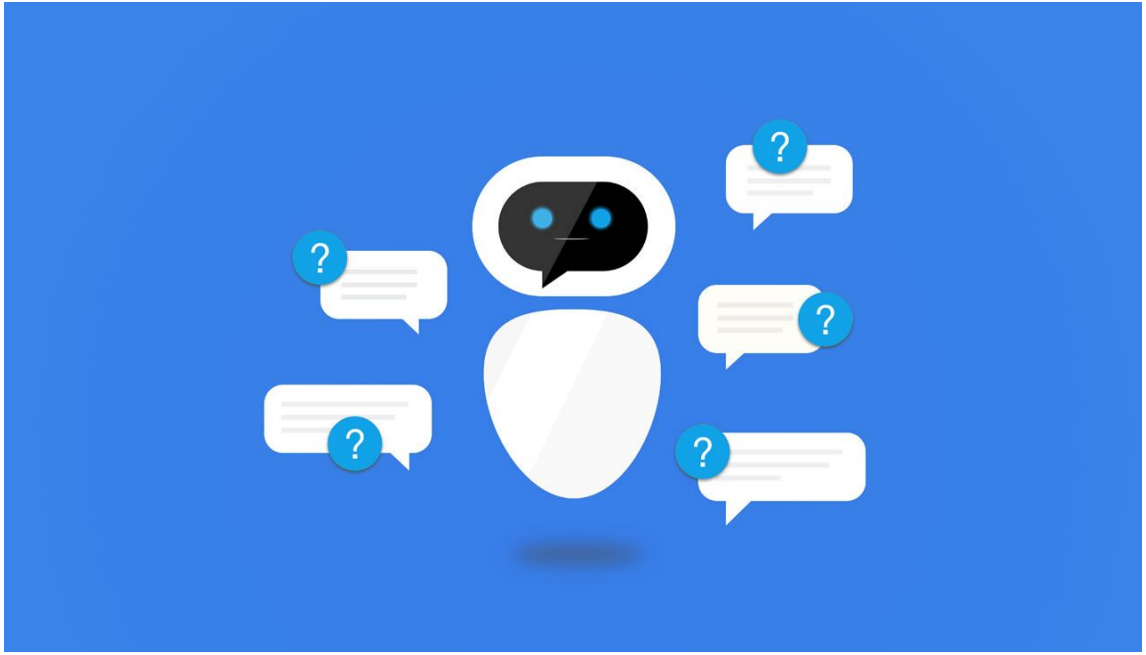
Jader Martins Camboim de Sá
Orientadores: Leandro Villas, Julio C. dos Reis

Agenda

- Introduction
- Motivation
- Related Work
- Proposal
- Methodology
- Results
- Next Steps

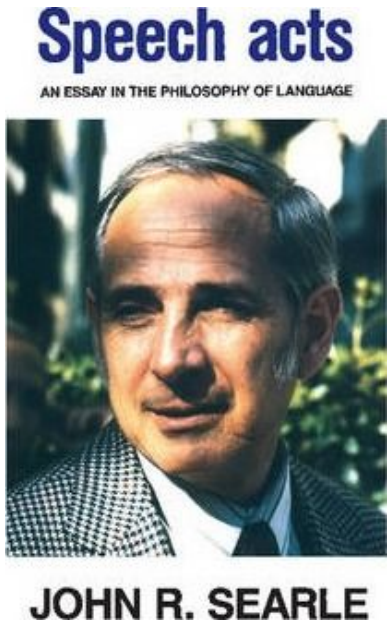
Introduction

Chatbots (Dialog Systems)



UNICAMP

Speech Act and Wizard-of-Oz



An Iterative Design Methodology for User-Friendly Natural Language Office Information Applications

J. F. KELLEY
IBM Thomas J. Watson Research Center

A six-step, iterative, empirical human factors design methodology was used to develop CAL, a natural language computer application to help computer-naive business professionals manage their personal calendars. Input language is processed by a simple, nonparsing algorithm with limited storage requirements and a quick response time. CAL allows unconstrained English inputs from users with no training (except for a five minute introduction to the keyboard and display) and no manual (except for a two-page overview of the system). In a controlled test of performance, CAL correctly responded to between 86 percent and 97 percent of the storage and retrieval requests it received, according to various criteria. This level of performance could never have been achieved with such a simple processing model were it not for the empirical approach used in the development of the program and its dictionaries. The tools of the engineering psychologist are clearly invaluable in the development of user-friendly software, if that software is to accommodate the unruly language of computer-naive, first-time users. The key is to elicit the cooperation of such users as partners in an iterative, empirical development process.

Categories and Subject Descriptors: D.m [**Software**]: *software psychology*; H.1.2 [**Models and Principles**]: User/Machine Systems—*human factors*; I.2.1 [**Artificial Intelligence**]: Applications and Expert Systems—*natural language interfaces*; I.2.7 [**Artificial Intelligence**]: Natural Language Processing—*language parsing and understanding*; I.6.3 [**Simulation and Modeling**]: Applications; K.6.3 [**Management of Computing and Information Systems**]: Software Management—*software development*

General Terms: Experimentation, Human Factors

UNICAMP

Dialog Problem Structure

- The literature propose to solve task-oriented dialog through recognizing **Intents** and **Entities**;

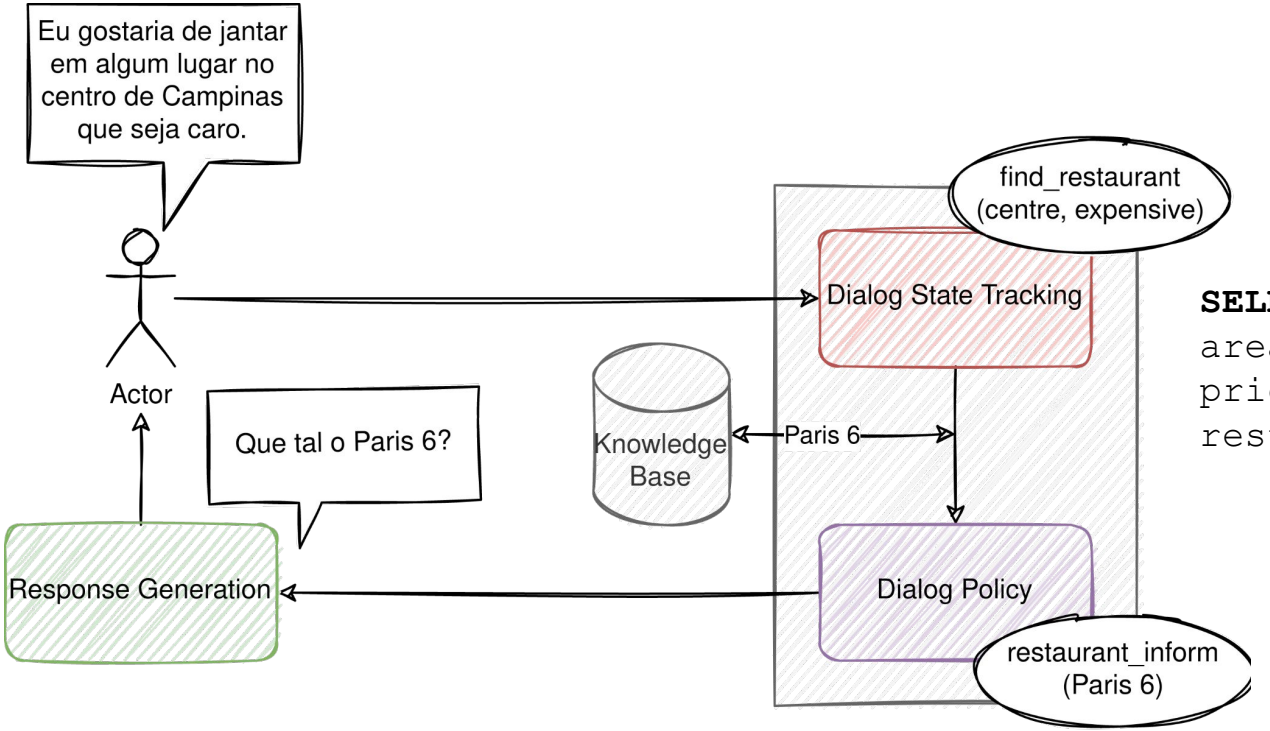
"Eu gostaria de jantar em algum lugar no centro de Campinas que seja caro."

find_restaurant: ['centre', 'expensive']

1. BUDZIANOWSKI, Paweł et al. MultiWOZ--A Large-Scale Multi-Domain Wizard-of-Oz Dataset for Task-Oriented Dialogue Modelling. arXiv preprint arXiv:1810.00278, 2018.
2. BYRNE, Bill et al. Taskmaster-1: Toward a realistic and diverse dialog dataset. arXiv preprint arXiv:1909.05358, 2019.

UNICAMP

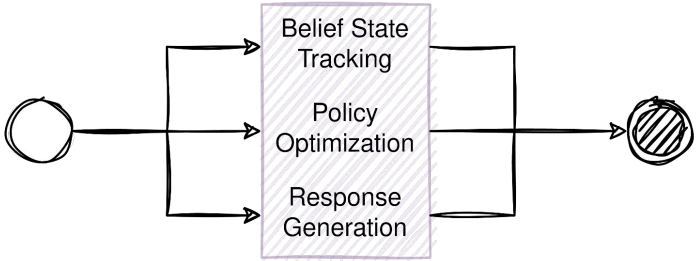
Task-Oriented Dialog System




```
SELECT restaurant.name WHERE
area='centre' AND
pricerange='expensive' FROM
restaurant;
```

UNICAMP

Taxonomy of Task-Oriented Dialog Systems

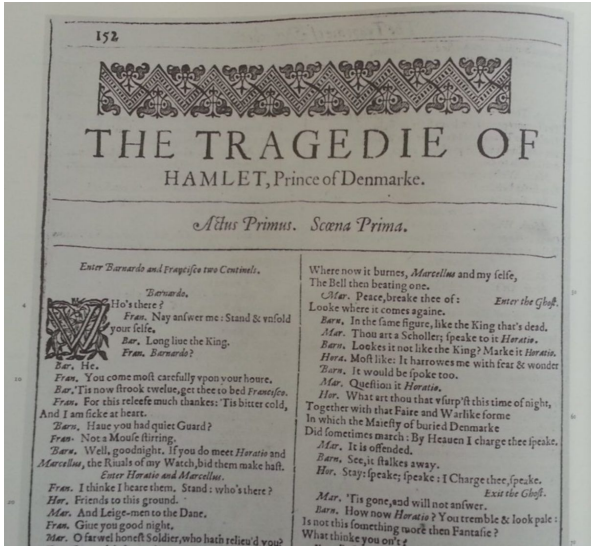


UNICAMP

Motivation

Differences

Screenplay



152

THE TRAGEDIE OF
HAMLET, Prince of Denmarke.

Actus Primus. Scœna Prima.

Enter Barnardo and Francisco two Centinels.

Barnardo.

Ho's there?

Fran. Nay answer me: Stand & vnfold your selfe.

Bar. Long liue the King.

Fran. *Barnardo?*

Bar. He.

Fran. You come most carefully vpon your houre.

Bar. 'Tis now strook twelue, get thee to bed *Francisco.*

Fran. For this releefe much thankes: 'Tis bitter cold, And I am sicke at heart.

Barn. Haue you had quiet Guard?

Fran. Not a Mouse stirring.

Barn. Well, goodnight. If you do meet *Horatio* and *Marcellus*, the Riuals of my Watch, bid them make haft.

Enter Horatio and Marcellus.

Fran. I thinke I heare them. Stand: who's there?

Hor. Friends to this ground.

Mar. And Leige-men to the Dane.

Fran. Giue you good night.

Mar. O farwell honeſt Soldier, who hath relieu'd you?

Where now it burnes, *Marcellus* and my ſelfe, The Bell then beating one.

Mar. Peace, breake thee of: *Enter the Ghoſt.* Looke where it comes againe.

Barn. In the ſame figure, like the King that's dead.

Mar. Thou art a Scholler; ſpeake to it *Horatio.*

Barn. Lookes it not like the King? Marke it *Horatio.*

Hora. Moſt like: It harrowes me with fear & wonder

Barn. It would be ſpoke too.

Mar. Queſtion it *Horatio.*

Hor. What art thou that vſurp'ſt this time of night, Together with that Faire and Warlike forme In which the Maieſty of buried Denmarke Did ſometimes march: By Heauen I charge thee ſpeake.

Mar. It is offended.

Barn. See, it ſtalkes away.

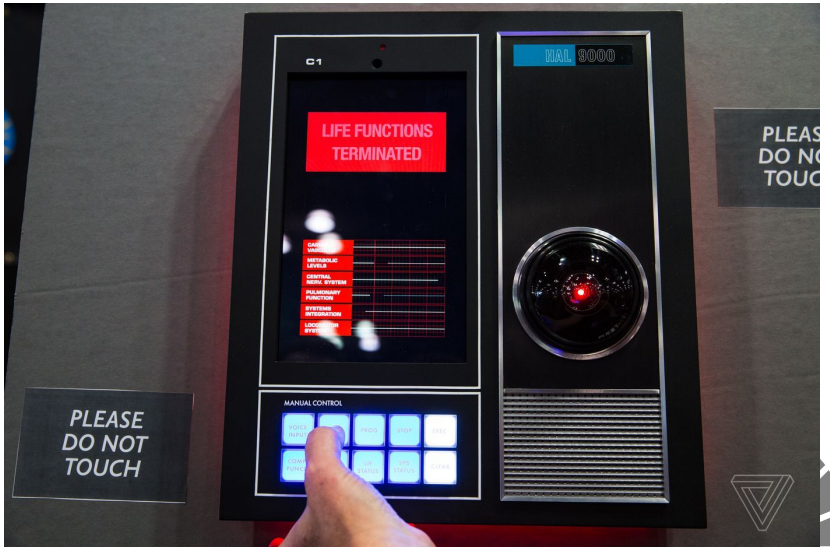
Hor. Stay: ſpeake; ſpeake: I Charge thee, ſpeake.

Exit the Ghoſt.

Mar. 'Tis gone, and will not anſwer.

Barn. How now *Horatio*? You tremble & look pale: Is not this ſomething more then Fantaſie? What thinke you on't?

Internal Representation (Black Box)



UNICAMP

Humanized Chatbot

- olá

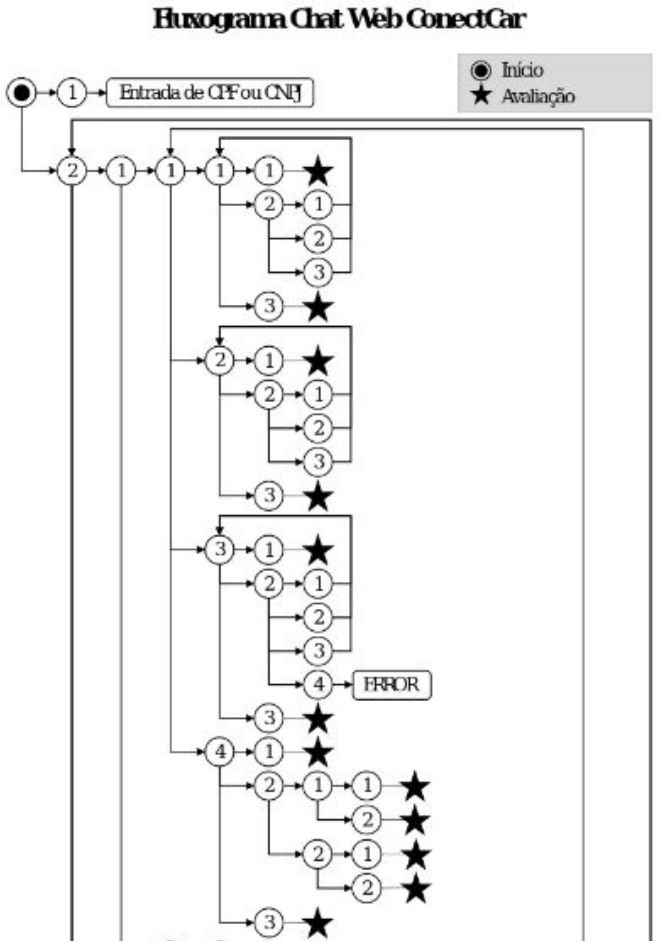
+ Olá! Bem vindo(a) a ConectCar!
+ Vamos do início: você já é nosso cliente?
  1. Sim
  2. Não

- 2

+ Em que posso te ajudar?
  1. Quero conhecer mais sobre os planos da ConectCar.
  2. Quero saber mais sobre promoções
  3. Quero ser ConectCar.
  4. Quero comprar um adesivo.
  5. Quero ativar um adesivo
  6. Como funciaona o Cashback?
  7. Prefiro escrever o que eu preciso.

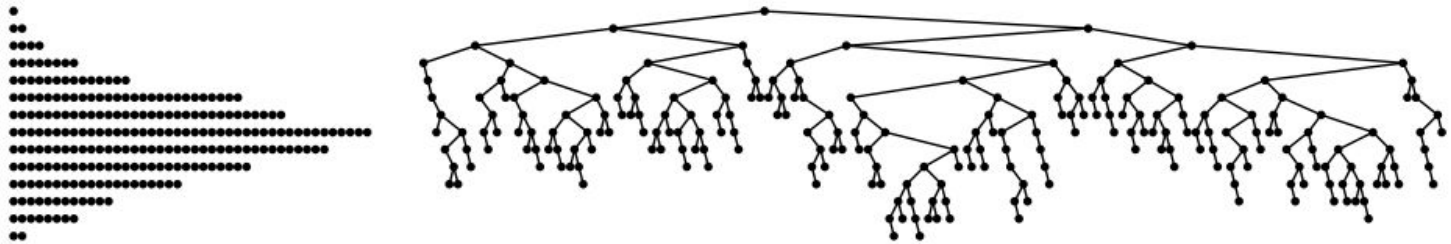
- Me informe das promoções **<bot não consegue fazer NLU>**

+ Escolha uma opção do menu que vamos responder sua dúvida.
  Em que posso te ajudar?
  1. Quero conhecer mais sobre os planos da ConectCar.
  2. Quero saber mais sobre promoções
  3. Quero ser ConectCar.
  4. Quero comprar um adesivo.
  5. Quero ativar um adesivo
  6. Como funciaona o Cashback?
  7. Prefiro escrever o que eu preciso.

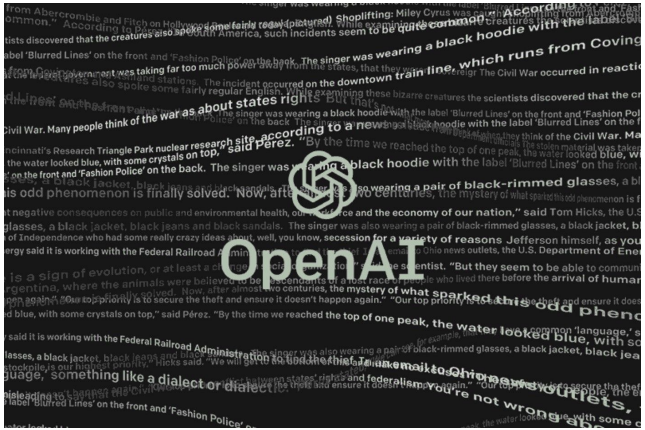


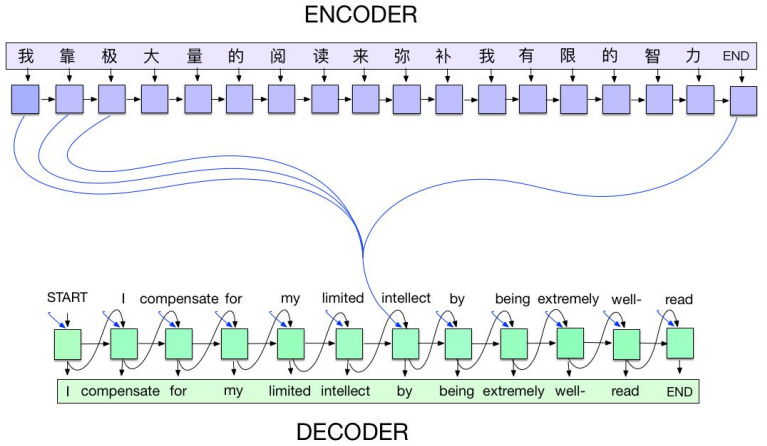
UNICAMP

Infinity Possibilities of Dialogues

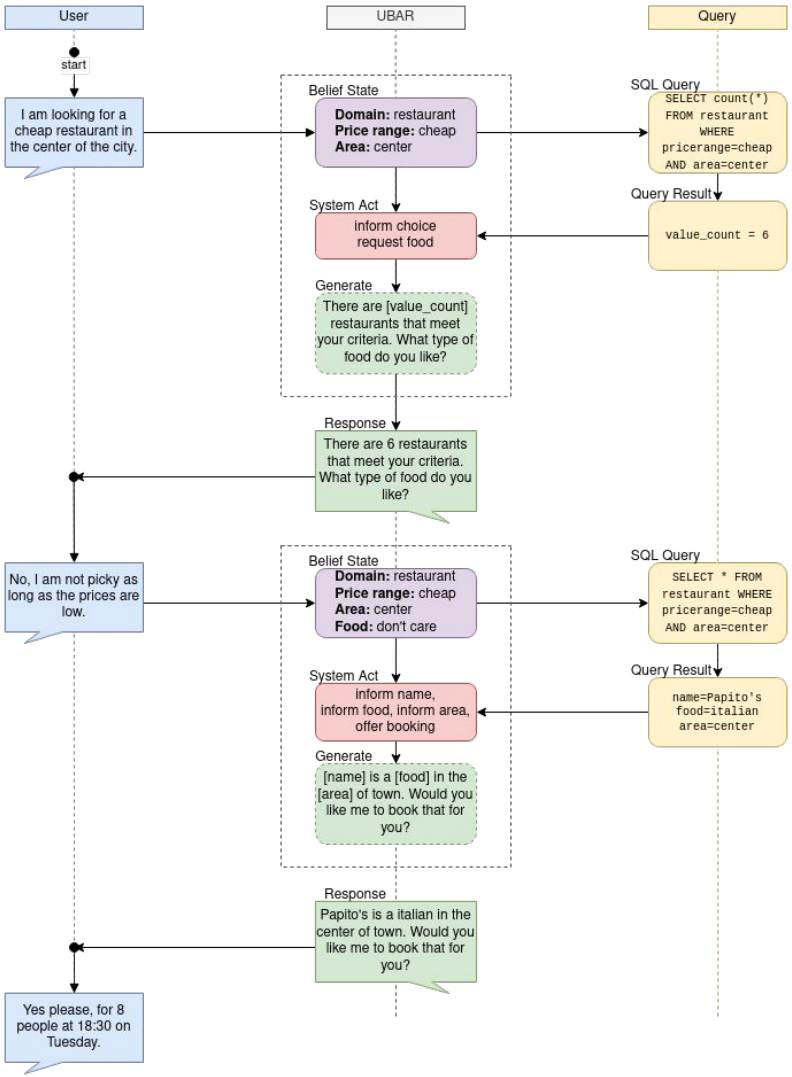


Sequence-to-sequence





UNICAMP


User
UBAR
Query
start
I am looking for a cheap restaurant in the center of the city.
Belief State
Domain: restaurant
Price range: cheap
Area: center
SQL Query
SELECT count(*) FROM restaurant WHERE pricerange=cheap AND area=center
Query Result
value_count = 6
System Act
inform choice request food
Generate
There are [value_count] restaurants that meet your criteria. What type of food do you like?
Response
There are 6 restaurants that meet your criteria. What type of food do you like?
No, I am not picky as long as the prices are low.
Belief State
Domain: restaurant
Price range: cheap
Area: center
Food: don't care
SQL Query
SELECT * FROM restaurant WHERE pricerange=cheap AND area=center
Query Result
name=Papito's food=italian area=center
System Act
inform name, inform food, inform area, offer booking
Generate
[name] is a [food] in the [area] of town. Would you like me to book that for you?
Response
Papito's is a italian in the center of town. Would you like me to book that for you?
Yes please, for 8 people at 18:30 on Tuesday.

UNICAMP

UBAR encoding

<sos_u> Eu gostaria de jantar em algum lugar no centro de Campinas que seja caro. <eos_u><sos_b> find_restaurant area centro price caro <eos_b><sos_a> restaurant_inform Paris 6<eos_a><sos_r> Que tal o Paris 6? <eos_r>

Related Work

Related Work

- ZHANG, Yizhe et al. **Dialogpt: Large-scale generative pre-training for conversational response generation.** arXiv preprint arXiv:1911.00536, 2019.

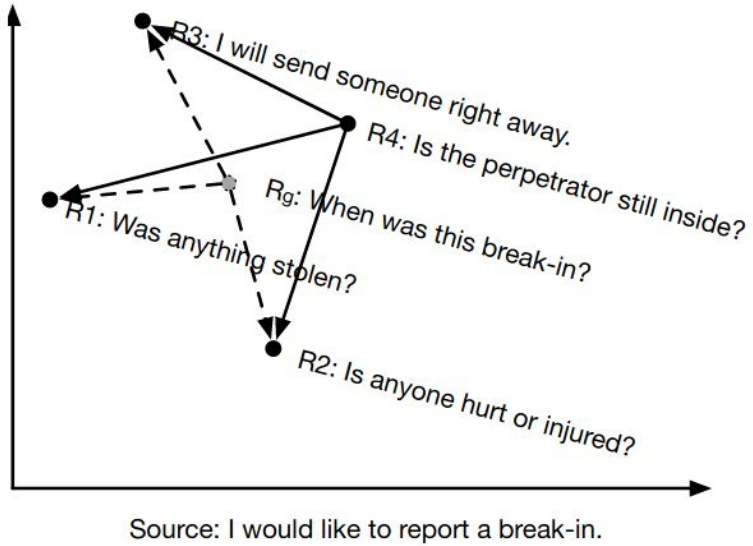


Figure 1: A generated response can surpass a human response in automatic metrics. Example responses are from Gupta et al. (2019)

UNICAMP

Related Work

- SU, Yixuan et al. **Multi-Task Pre-Training for Plug-and-Play Task-Oriented Dialogue System.** arXiv preprint arXiv:2109.14739, 2021.

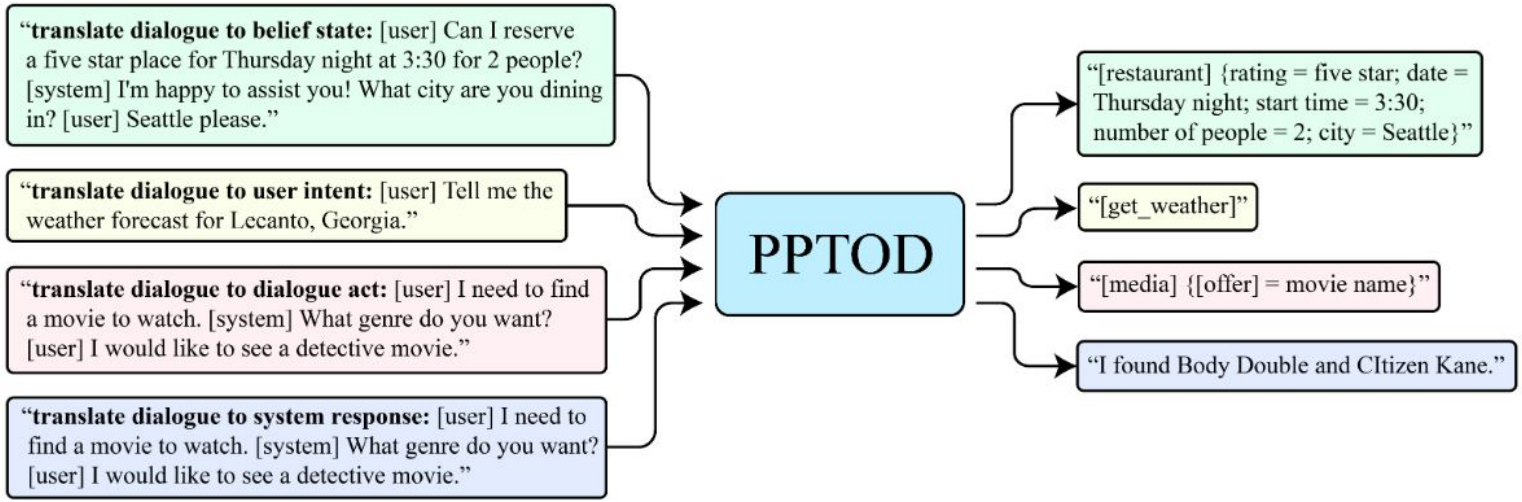


Figure 1: **Overview**: In the dialogue multi-task pre-training stage, we pre-train our model with four TOD-related tasks, including natural language understanding (NLU), dialogue state tracking (DST), dialogue policy learning (POL), and natural language generation (NLG). For each task, the model takes the dialogue context and the task-specific prompt as input and learns to generate the corresponding target text. Our learning framework allows us to train the model with partially annotated data across a diverse set of tasks. (best viewed in color)

Related Work

- KULHÁNEK, Jonáš et al. **Augpt: Dialogue with pre-trained language models and data augmentation.** arXiv preprint arXiv:2102.05126, 2021.

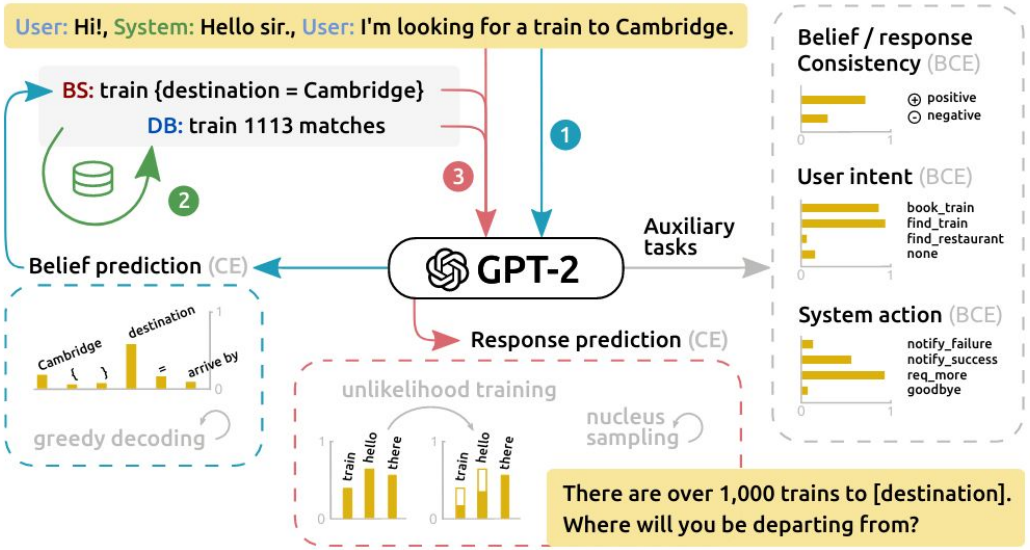
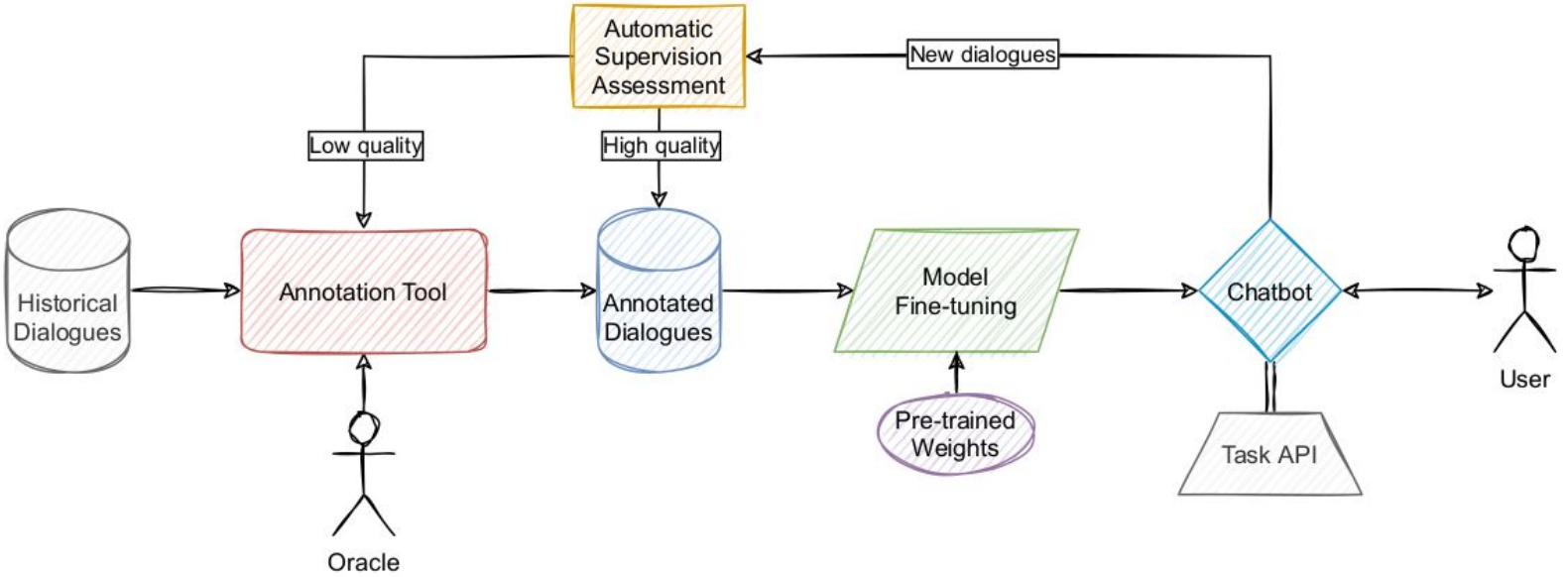


Figure 1: The architecture of AuGPT. The pipeline runs in two stages. First, a finetuned GPT-2 LM is used to predict a belief. Then the database results are obtained and everything is passed to the GPT-2 again to predict a final delexicalized response, along with possible auxiliary tasks (belief consistency, intent classification, system action classification). Unlikelihood loss is used for response prediction training.

UNICAMP

Proposal

ZéCarioca



UNICAMP

Multi-Sequential Transfer Learning

Selecting and ordering tasks addressing the final objective.

Scaling Hypothesis

“The loss scales as a power-law with model size, dataset size, and the amount of compute used for training, with some trends spanning more than seven orders of magnitude.”

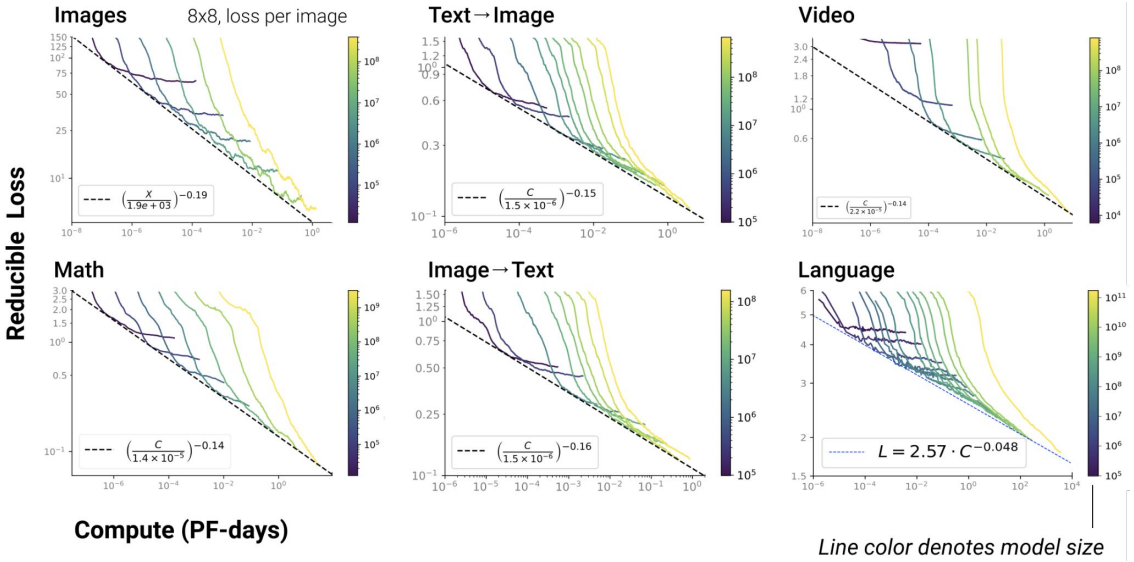
-- KAPLAN, Jared et al. Scaling laws for neural language models. arXiv preprint arXiv:2001.08361, 2020.

Scaling Hypothesis

"[...] Once we find a scalable architecture like self-attention or convolutions, which like the brain can be applied fairly uniformly, we can simply train ever larger NNs and ever more sophisticated behavior will emerge naturally as the easiest way to optimize for all the tasks & data. More powerful NNs are 'just' scaled-up weak NNs, in much the same way that human brains look much like scaled-up primate brains."

-- Gwern https://www.gwern.net/Scaling-hypothesis

Scaling Laws



HENIGHAN, Tom et al. Scaling laws for autoregressive generative modeling. **arXiv preprint arXiv:2010.14701**, 2020.

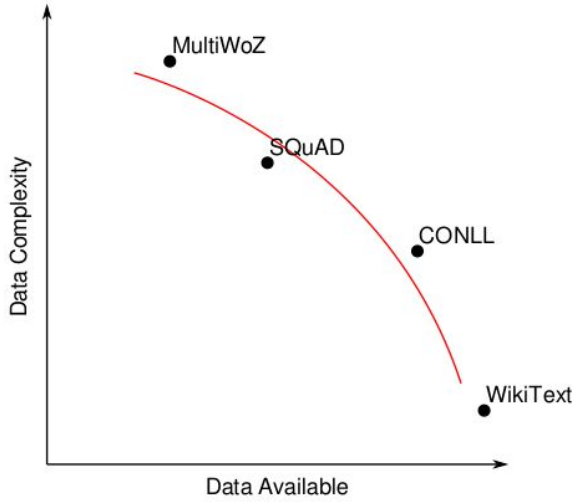
UNICAMP

Model	1% of training data				5% of training data				10% of training data				20% of training data			
	Inform	Succ.	BLEU	Comb.	Inform	Succ.	BLEU	Comb.	Inform	Succ.	BLEU	Comb.	Inform	Succ.	BLEU	Comb.
PPTOD$_{small}$																
run-1	68.50	54.90	13.98	75.68	78.40	61.50	14.78	84.73	79.70	68.70	17.10	91.30	83.40	71.10	17.05	94.30
run-2	64.70	50.20	12.19	69.64	75.20	61.30	15.85	84.10	87.00	67.30	13.89	91.04	82.80	68.90	17.03	92.88
run-3	65.30	46.10	10.79	66.49	75.40	60.80	15.99	84.09	84.30	68.10	15.33	91.50	83.20	70.00	17.01	93.61
run-4	64.80	51.00	12.43	70.33	77.20	59.70	15.75	84.20	84.50	71.90	14.51	92.71	82.40	69.40	17.93	93.83
run-5	71.50	52.30	13.14	75.04	76.70	64.70	14.37	85.07	78.00	64.90	16.99	88.44	83.00	70.10	16.10	92.65
average	66.96	50.90	12.51	71.44	76.58	61.60	15.35	84.44	83.50	68.18	15.56	91.01	82.96	69.90	17.02	93.45
std	2.67	2.88	1.06	3.46	1.18	1.67	0.65	0.39	3.33	2.26	1.29	1.40	0.34	0.74	0.58	0.61
PPTOD$_{base}$																
run-1	74.20	55.40	13.08	77.88	80.50	66.10	15.58	88.88	85.10	67.50	16.02	92.32	84.90	72.50	17.16	95.86
run-2	71.20	51.10	13.32	74.47	81.50	63.10	14.32	86.62	84.60	69.00	15.06	91.86	84.00	72.50	16.46	94.71
run-3	76.20	49.70	12.39	75.34	77.50	61.70	14.98	84.58	84.10	69.20	15.49	92.14	85.50	69.60	17.76	95.31
run-4	75.80	52.40	13.21	77.30	79.70	62.30	15.13	86.10	84.40	68.30	15.17	91.52	84.20	70.70	16.88	94.33
run-5	74.70	53.60	12.97	77.05	80.10	64.20	14.44	86.59	83.90	67.80	16.12	91.96	86.10	73.20	16.78	96.43
average	74.42	52.44	12.99	76.41	79.86	63.48	14.89	86.55	84.42	68.36	15.57	91.96	84.94	71.70	17.01	95.32
std	1.76	1.97	0.32	1.29	1.32	1.55	0.46	1.38	0.42	0.66	0.43	0.27	0.79	1.34	0.44	0.75
PPTOD$_{large}$																
run-1	64.40	51.90	11.30	69.45	75.20	59.80	14.01	81.51	79.30	64.60	14.82	86.77	82.10	69.70	14.68	90.58
run-2	65.50	53.20	12.01	71.36	74.30	64.10	14.98	83.18	80.40	67.80	15.01	89.11	81.70	72.20	15.61	92.56
run-3	66.20	50.80	11.94	70.49	76.90	62.30	14.01	83.61	81.30	69.20	16.23	91.48	80.90	70.80	14.33	90.18
run-4	62.70	52.60	12.20	69.85	76.20	60.70	13.45	81.90	82.30	66.90	14.99	89.59	83.10	73.50	15.83	94.13
run-5	63.10	51.20	11.73	68.88	73.40	62.80	14.42	82.52	79.90	65.20	15.21	87.76	80.90	74.70	15.21	93.01
average	64.38	51.94	11.84	70.01	75.20	61.94	14.17	82.54	80.64	66.74	15.25	88.94	81.74	72.18	15.13	92.09
std	1.34	0.88	0.31	0.85	1.26	1.53	0.51	0.78	1.06	1.68	0.50	1.61	0.82	1.80	0.56	1.49

Table 10: Few-Shot Experiments on MultiWOZ: The average and std rows show the mean and standard deviation of results from five different runs. The Succ. and Comb. denote Success and Combined Score, respectively.

SU, Yixuan et al. Multi-Task Pre-Training for Plug-and-Play Task-Oriented Dialogue System. **arXiv preprint arXiv:2109.14739**, 2021.

Data Available



(a) Data complexity versus availability.

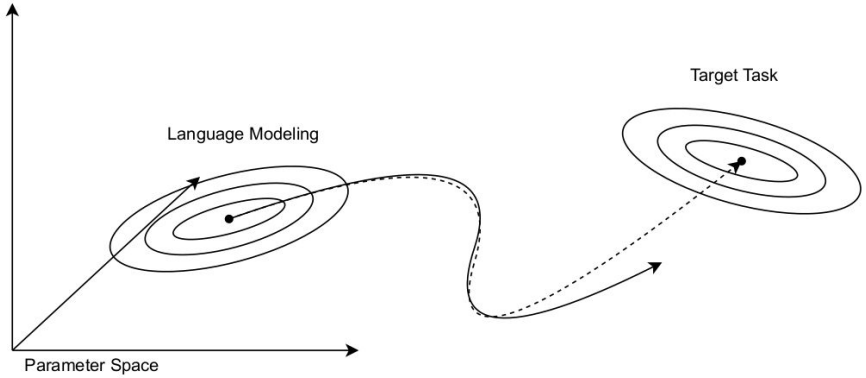


Figure 5: Learning a target task from language modelling. With not enough data the learning process fail to arrive in a global optima for the target task.

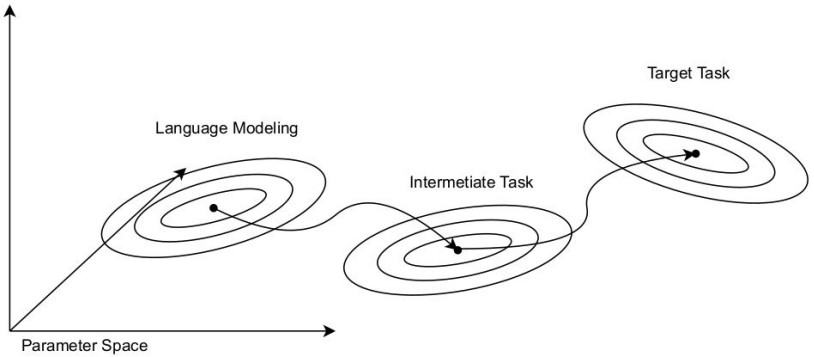
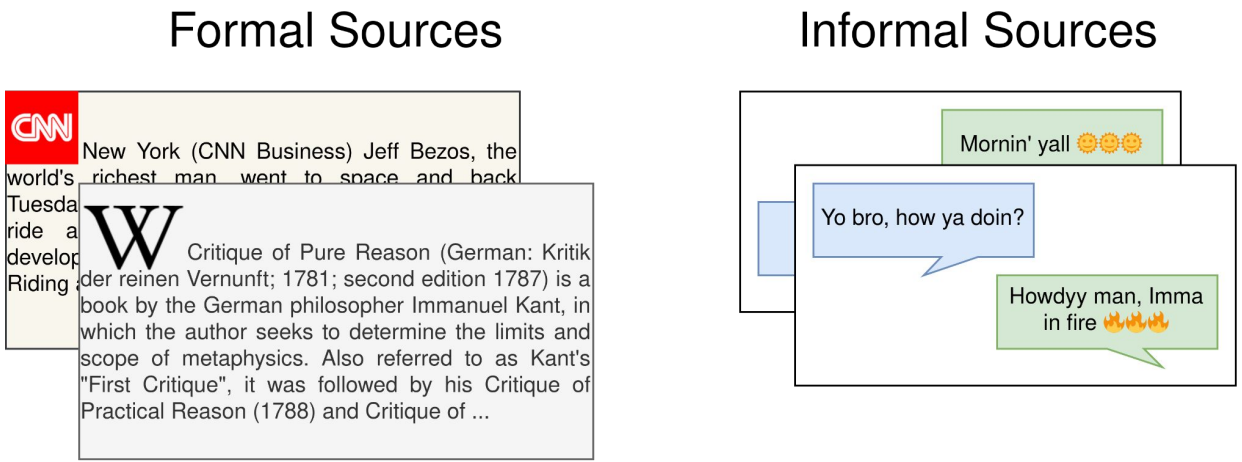


Figure 7: Learning a target task from gradual adaptation. With a better initialization the model needs less data to arrive in the global optima.

UBAR encoding

<sos_u> Eu gostaria de jantar em algum lugar no centro de Campinas que seja caro. <eos_u><sos_b> find_restaurant area centro price caro <eos_b><sos_a> restaurant_inform Paris 6<eos_a><sos_r> Que tal o Paris 6? <eos_r>

Concept Drift



1. ZHANG, Yizhe et al. Dialogpt: Large-scale generative pre-training for conversational response generation. arXiv preprint arXiv:1911.00536, 2019.
2. WU, Chien-Sheng et al. TOD-BERT: pre-trained natural language understanding for task-oriented dialogue. arXiv preprint arXiv:2004.06871, 2020.

UNICAMP

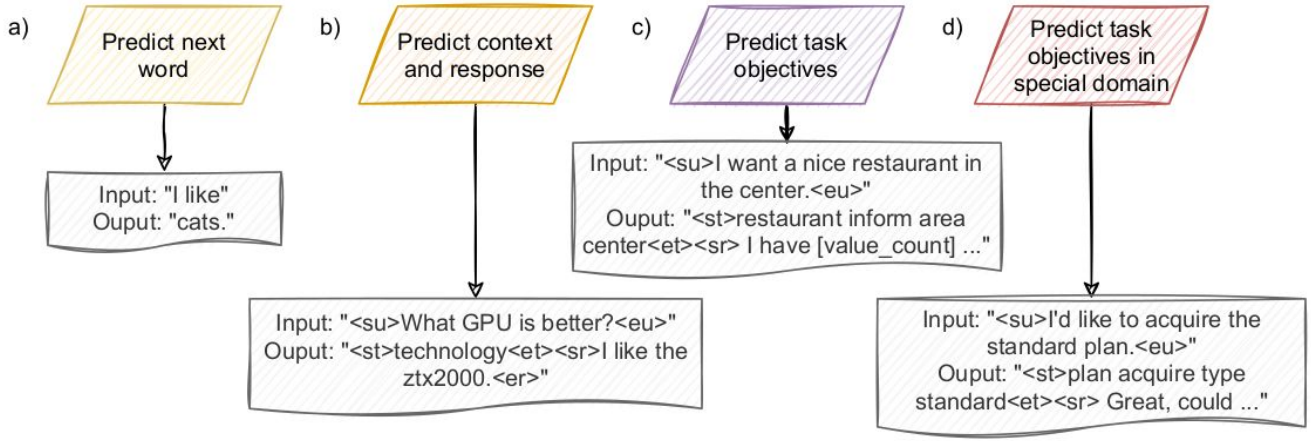


Figure 8: Detailed tasks for each training phase.

Methodology

Datasets

- **Language Modelling (Prato Feito):** Language models could generate hate speech sentences depending on the kind of data they were trained for, and as a customer attendance service, the ZéCarioca project took measures to remove potential hate speech from our training dataset so it could not generate unpleasant responses. Additionally to filtering harmful content, we also filter:
  - Dialogues where the question or response did not terminate with any punctuation signal.
  - Dialogues with links, emojis, or special characters.

Datasets

- **Dialog Modelling (Forums):** Portuguese forums like adrenaline and Outerspace has a quote function that allows user to respond previous messages in the forum. With this natural utterance-response structure, we model the problem of predicting response first linking each message that replied a previous message in a sequence of messages, and then by applying a sliding window of size 2 where the first element is the utterance and the second is the response for the above utterance. We encode this task by predicting the response given the previous utterance in the forums dialog by utterance and response tokens, for example, <sos_u>What is the best monitor for gaming?<eos_u><sos_r>I think the monitor X gaming has a good cost benefit<eos_r>. Additionally we applied some filter to the data, we filtered:
  - Utterance or response has less then 3 tokens.
  - Utterance or response has more then 256 tokens.

Datasets

- **MultiWoZ Translated:** We obtain this dataset by translating the original MultiWoZ to Portuguese through the Google Translate API, in this process we need to translate both utterances and slot-values, a problem we noticed is that depending on the context utterances and slot-values was translated to distinct words, so we manually aligned words in utterances and slots.

Datasets

- **Case Dataset:** The final dataset in the training pipeline was constructed analyzing common problems for the addresses business, ConectCar, and creating conversations that address those issues. We artificially generate this data by interacting with a web interface and asking it to complete task, training and evaluating simultaneously. At first interactions we need to manually correct the chatbot utterances to match an expected behaviour. This process allow to create many examples quickly.

Metrics

- SUCCESS
- INFORM
- BLEU

Results

Figure 11: Empirical loss for WikiPT and WikiPT+MultiWOZ evaluated on the validation set of cased data.

Table 2: Comparative analysis between pretraining only with language modelling, and languange modelling plus TOD task.

Training Regime	SUCCESS (intent)	SUCCESS (action)	INFORM	BLEU
WikiPT	0.37	0.28	0.42	**0.51**
WikiPT + MWOZPT	**0.67**	**0.62**	**0.86**	0.43

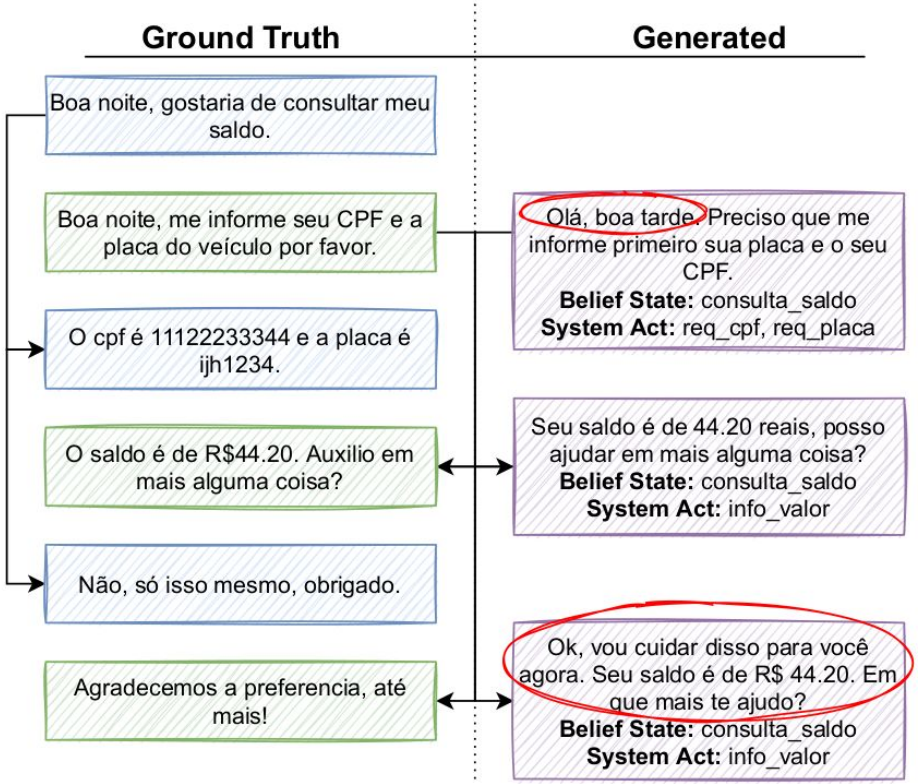
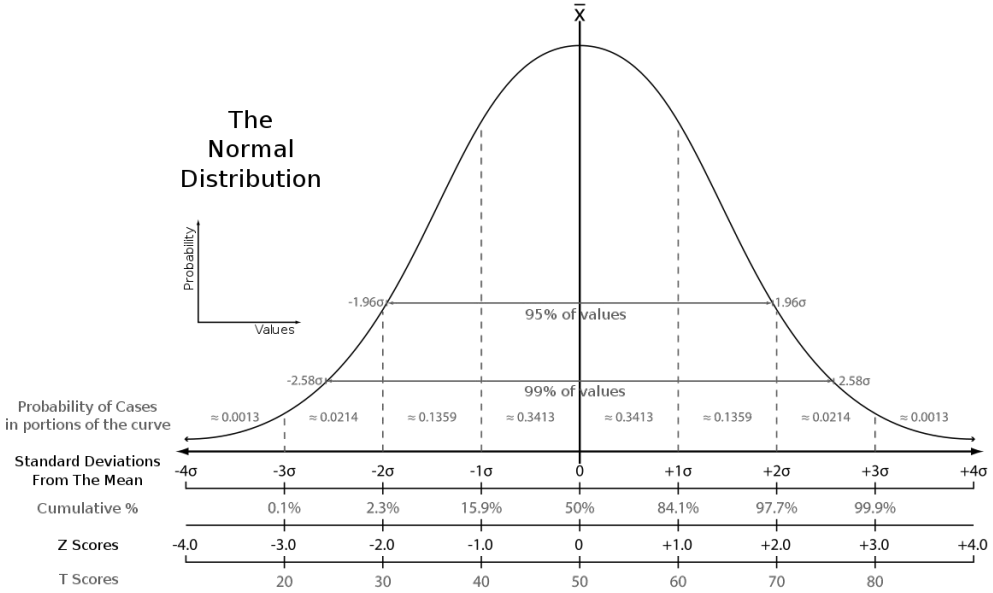


Figure 10: Example of three turns interacting with our bot created by the ZéCarioca framework. Demo available at Telegram @zecariocabot.

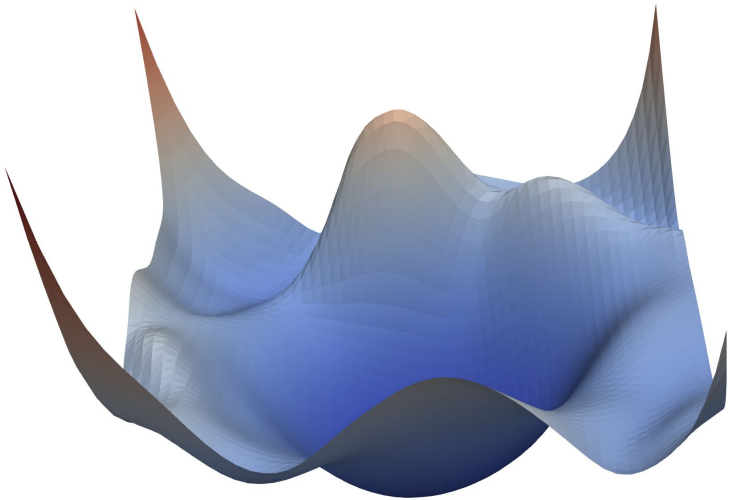
Next Steps

Compare Linguistic Distributions

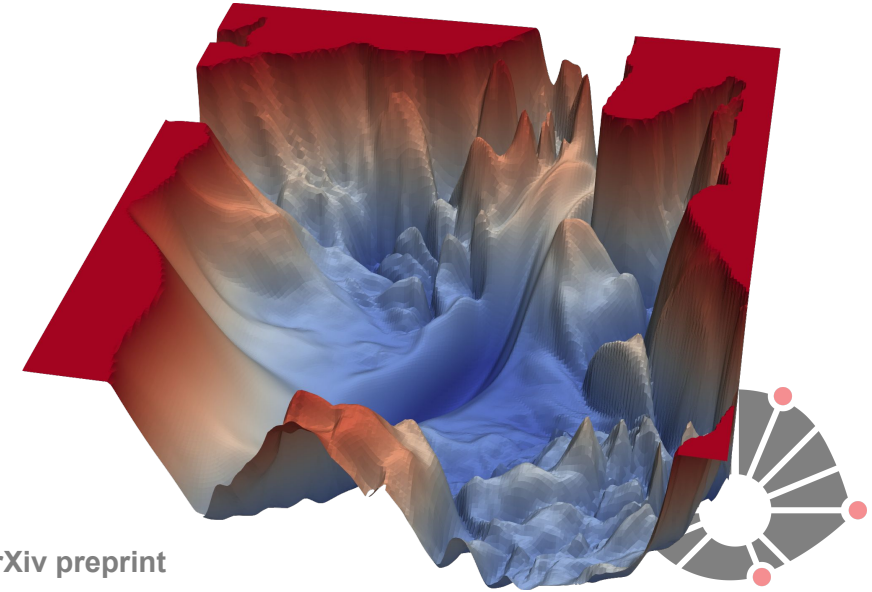


Loss Complexity

Predict next word



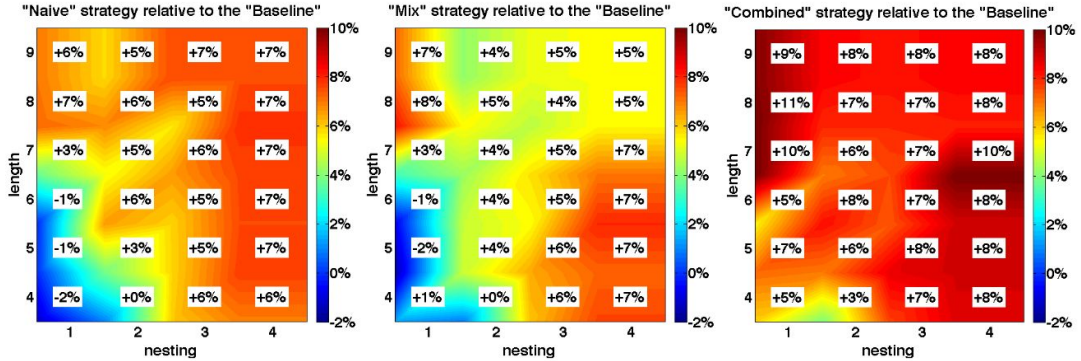
Predict next response

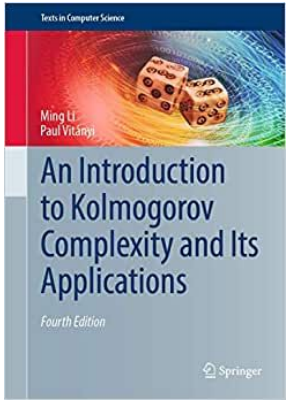


1. LI, Hao et al. Visualizing the loss landscape of neural nets. **arXiv preprint arXiv:1712.09913**, 2017.
2. BENGIO, Yoshua. Evolving culture versus local minima. In: Growing Adaptive Machines. Springer, Berlin, Heidelberg, 2014. p. 109-138.

UNICAMP

Data Complexity





$$H = -\sum_{i=1}^{N} p_i \log(p_i)$$

1. ZAREMBA, Wojciech; SUTSKEVER, Ilya. Learning to execute. arXiv preprint arXiv:1410.4615, 2014.

UNICAMP

The end