

# AuGPT: End-to-end dialogue modeling with Pre-trained Language Models and Data Augmentation

Jonáš Kulháněk, Vojtěch Hudeček, Tomáš Nekvinda & Ondřej Dušek

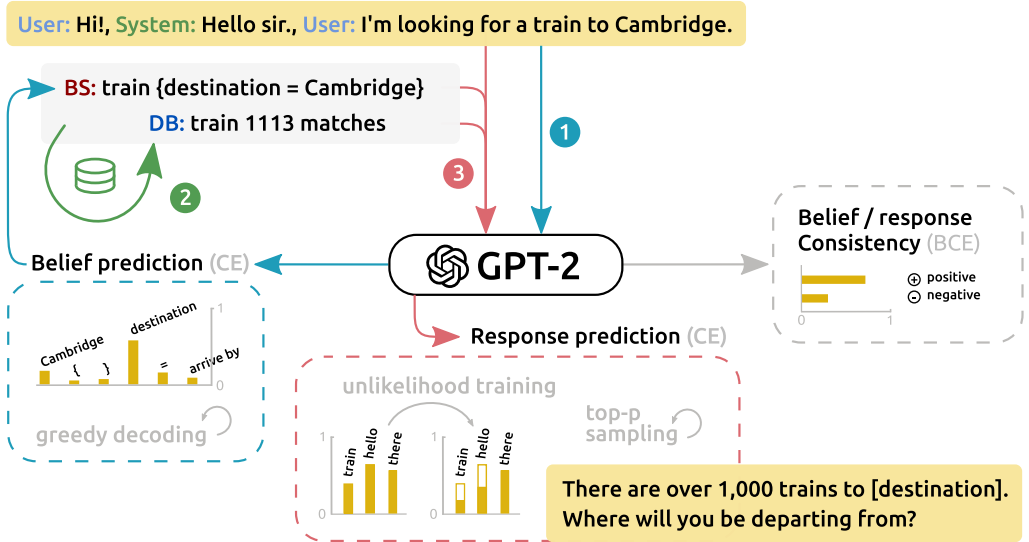
📅 February 8, 2021



Charles University  
Faculty of Mathematics and Physics  
Institute of Formal and Applied Linguistics

- model based on SOLOIST (Peng et al., 2020)
- auto-regressive language model used for decoding of (1) belief state and (2) response
- belief state also used as the query for the database
- finetuned GPT2 used as the language model

# AuGPT – architecture overview



# Belief state and response representations

- the **active** domain is determined by the order of domains in the belief state
- dropping domains in the delexicalized response

Belief state: train { leave at=15:30, arrive by=17:15 }, hotel { price range = cheap }  
DB: train 23 matches, hotel no match

**Example 1:** String format for AuGPT's belief state and database result count.

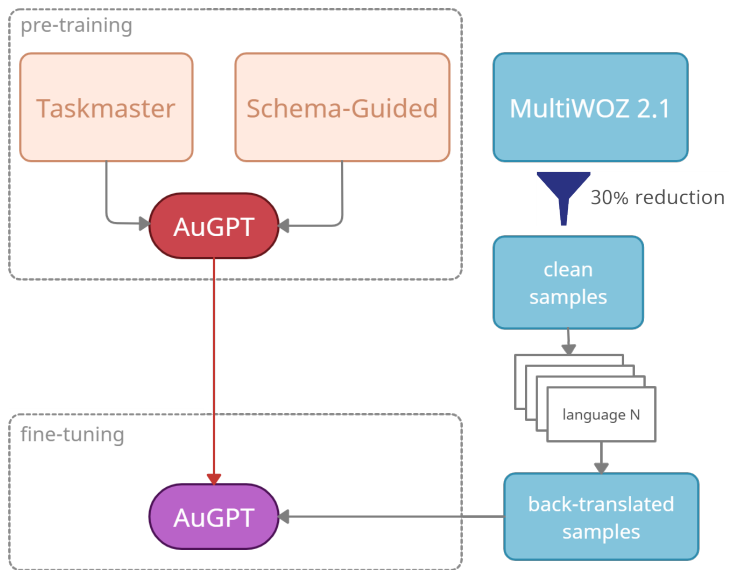
The phone number is [phone] and the postcode is [postcode].

**Example 2:** Delexicalized response.

# Auxiliary tasks

- we corrupt 50% of the samples and train a binary classifier to detect them, this is called **consistency detection** auxiliary task
- the binary classifier is attached to the last hidden representation of GPT2
- when constructing a negative sample, one of more of the following changes are applied with the same probability:
  - a different belief state is randomly sampled from the dataset
  - a different response is randomly sampled from the dataset
  - values in the belief state are uniformly randomly resampled (**new**)
- we also tried user intent and system action classifiers, but that did not help performance

# Training pipeline



- pre-trained on combined Taskmaster and Schema-Guided
- MultiWOZ 2.1 dataset is used for fine-tuning
- we filter the dataset to choose only 'clean' samples consistent with the database, i.e., dialogues with MultiWOZ success rate 1.0
- paraphrases are generated using backtranslations
- backtranslations: in-house MT system, with 10 languages chosen empirically out of 40 (Edunov et al., 2018, Macháček et al., 2020)

- massive **data augmentation** via backtranslation
- improved consistency detection **auxiliary training task**
- using **unlikelihood loss** for response to increase output diversity
- different decoding strategies for belief state and response
- training on '**clean**' samples only



## Results – MultiWOZ

method	MultiWOZ 2.0		
	inform	success	BLEU
Human	91.0	82.7	–
<b>AuGPT</b>	<b>90.2</b>	<b>75.5</b>	17.2
SOLOIST (Peng et al., 2020)	85.5	72.9	16.5
SimpleTOD (Hosseini-Asl et al., 2020)	84.4	70.1	15.1
DAMD (Zhang et al., 2020)	76.3	60.4	<b>18.6</b>
MD-Sequicity (Zhang et al., 2020)	86.6	71.6	16.8

Comparison with previous works on the MultiWOZ dataset

## Results – ConvLab 2

method	complete	success	book	inform F1	# turns
<b>AuGPT</b>	<b>89.4</b>	<b>60.1</b>	<b>85.7</b>	<b>70.3</b>	<b>14.6</b>
DAMD (Zhang et al., 2020)	39.5	34.3	51.4	56.3	29.8
Sequicity (Lei et al., 2018)	23.1	9.8	4.1	29.9	32.6

ConvLab evaluation comparison with other works

# Ablation study

method	MultiWOZ			ConvLab 2		
	inf	suc	BLEU	suc	book	inform F1
<b>AuGPT</b>	91.4	<b>72.9</b>	17.2	<b>60.1</b>	85.7	70.3
w/o. unlikelihood	90.8	70.4	16.9	59.3	<b>90.8</b>	69.5
w/o. clean	<b>91.6</b>	70.7	15.8	57.7	85.6	69.6
w/o. unlikelihood, clean	90.4	72.7	<b>17.5</b>	58.4	81.3	67.5
w/o. pre-training*	90.7	67.9	15.1	59.8	83.7	72.1
w/o. back-translations	89.1	67.9	15.2	58.2	87.4	<b>72.2</b>
w. old consistency	90.7	71.8	17.0	57.8	86.0	69.8
w/o. consistency	90.4	68.7	16.8	57.1	84.1	70.9

inf = inform, suc = success, book = book rate

\* picked for human evaluation

## Results – human evaluation

method	suc.	suc. w/ DB	suc. w/o DB	turns
Baseline	69.6	56.8	82.4	18.5
Team1 (winner)	<b>74.8</b>	<b>70.2</b>	79.4	18.5
Team2	<b>74.8</b>	68.8	80.8	19.4
Team7 (ours)	72.3	62.0	<b>82.6</b>	<b>17.1</b>
Team6	70.6	60.8	80.4	20.1
Team3	67.8	60.0	75.6	21

Human evaluation results on the DSTC9 Track 2.

We show a shortened version of the table, the full version is published online at <https://convlab.github.io>.

- In addition – our own detailed manual analysis
  - we followed MultiWOZ tasks and talked to the system ourselves
- 130 dialogues were evaluated in total.
- 50 erroneous dialogues, some errors recoverable
- 17 failed dialogues (i.e. 87 % success w/o DB grounding)
- Errors were categorized w.r.t. their cause

# Error types

Type	Count	Source	Description
Hallucinated values	21	BS/Policy	Used a slot value in the reply that is not grounded in the DB nor in the context
Wrong lexicalization	6	Policy	Repeats the same value in a list of choices during lexicalization
Missing information	5	Policy	Makes booking while not all information is specified
Ignored input	5	BS	Keeps asking for information that was provided
Bad domain	4	BS	Fails to focus on the correct domain
False response	4	Policy	States a different value of a slot than the value stored in DB
Repeated output	3	Policy	Repeats the same slot twice on the output
Failed booking	3	DB/Policy	Booking was unsuccessful due to DB mismatch
Other	10	BS/DB/P/Oth	(Various rare errors that could not be categorized)

Table 1: Distribution of the most common error types

# Bad domain & Hallucination example

**user:** *I'm looking for a theatre in Cambridge*

belief: attraction {type = theatre}

**system:** There are 5 theatres in Cambridge. Is there an area of town you prefer?

**user:** *no I don't care*

belief: attraction {type = theatre}

**system:** I recommend Adc Theatre in the centre. Would you like more information?

**user:** *no, I'm also looking for some place to stay*

belief: attraction {type = theatre}

**system:** I have 5 theatres in the centre. Do you have a price range?

**user:** *I'm looking for a place to stay*

belief: attraction {type = theatre}

**system:** There are 5 theatres in the centre. Do you have a preference for area?

**user:** *I'm looking for a hotel*

belief: hotel {type = hotel}

attraction {type = theatre}

**system:** There are 9 hotels in the *north*. Do you have a price range?

## Conclusion

- We achieve state-of-the-art performance for end-to-end models on the MultiWOZ 2.1 dataset w.r.t. *inform* and *success* metrics
- We placed 3rd at the DSTC 9 Track 2 in the final human evaluation
- We performed detailed error analysis that provides important insight into the model behavior.

## Future Work

- We will use latent representation of the belief state and optimize it jointly with the language model
- We will replace the deterministic lexicalization with a trainable alternative



# Thank you!

- Source code and pre-trained models: <https://github.com/ufal/augpt>



Vojtěch Hudeček  
hudecek@ufal.mff.cuni.cz  
[ufal.mff.cuni.cz/vojtech-hudecek](https://ufal.mff.cuni.cz/vojtech-hudecek)



Jonáš Kulháněk  
jonas.kulhanek@cvut.cz  
[jkulhanek.github.io](https://jkulhanek.github.io)

# References

- Sergey Edunov, Myle Ott, Michael Auli, and David Grangier. Understanding back-translation at scale. In *Proceedings of the 2018 EMNLP*, pages 489–500, Brussels, Belgium, 2018.
- Ehsan Hosseini-Asl, Bryan McCann, Chien-Sheng Wu, Semih Yavuz, and Richard Socher. A simple language model for task-oriented dialogue. *arXiv preprint arXiv:2005.00796*, 2020.
- Wenqiang Lei, Xisen Jin, Min-Yen Kan, Zhaochun Ren, Xiangnan He, and Dawei Yin. Sequicity: Simplifying task-oriented dialogue systems with single sequence-to-sequence architectures. In *ACL*, Melbourne, Australia, 2018.
- Dominik Macháček, Jonáš Kratochvíl, Sangeet Sagar, Matúš Žilinec, Ondřej Bojar, Thai-Son Nguyen, Felix Schneider, Philip Williams, and Yuekun Yao. Elitr non-native speech translation at iwslt 2020. In *Proceedings of the 17th International Conference on Spoken Language Translation (IWSLT)*, page 200–208, Online, 2020.
- Baolin Peng, Chunyuan Li, Jinchao Li, Shahin Shayandeh, Lars Liden, and Jianfeng Gao. Soloist: Few-shot task-oriented dialog with a single pre-trained auto-regressive model. *arXiv preprint arXiv:2005.05298*, 2020.
- Yichi Zhang, Zhijian Ou, and Zhou Yu. Task-oriented dialog systems that consider multiple appropriate responses under the same context. In *AAAI*, New York, NY, USA, 2020.

# Problem formulation

- $r$  = response,  $b$  = belief state,  $c$  = context (past utterances),  $d$  = database results
- Query = the deterministic database lookup
- Prob. dist. over the responses given context can be factorized as:

$$\begin{aligned} p(r|c) &= \sum_d p(r|d, c)p(d|c) \\ &= \sum_d \sum_b p(r|d, b, c)p(d|b)p(b|c) \\ &= \sum_b p(r|Query(b), b, c)p(b|c), \end{aligned} \tag{1}$$

- Same model for belief state & response:

$$p(\bar{r}|d_c, b, c) \approx \hat{p}(\bar{r}|d_c, b, c, \theta) \tag{2}$$

$$p(b|c) \approx \hat{p}(b|\emptyset, \emptyset, c, \theta), \tag{3}$$

# Missing information example

[...]

**user:** *I want to make a booking at sitar tandoori*

**belief:** restaurant {name = sitar tandoori}

train {arrive by = 18:00, day = sunday, departure = cambridge, destination = london liverpool street, people = 1}

**system:** What day would you like the booking for?

**user:** *the same day, 13:30*

**belief:** restaurant {day = sunday, name = sitar tandoori, people = 1, time = 13:30}

train {arrive by = 18:00, day = sunday, departure = cambridge, destination = london liverpool street, people = 1}

**system:** Booking was successful. The table will be reserved for 15 minutes. Reference number is: 00000200.