Controlling Dialogue Generation with Semantic Exemplars

Prakhar Gupta[1], Jeffrey P. Bigham[1,2], Yulia Tsvetkov[1], Amy Pavel[2]

Carnegie Mellon University
[1] Language Technologies Institute
[2] Human-Computer Interaction Institute

Carnegie Mellon University Language Technologies Institute

The Task: Generate coherent and controlled responses that follow system level goals

Motivation

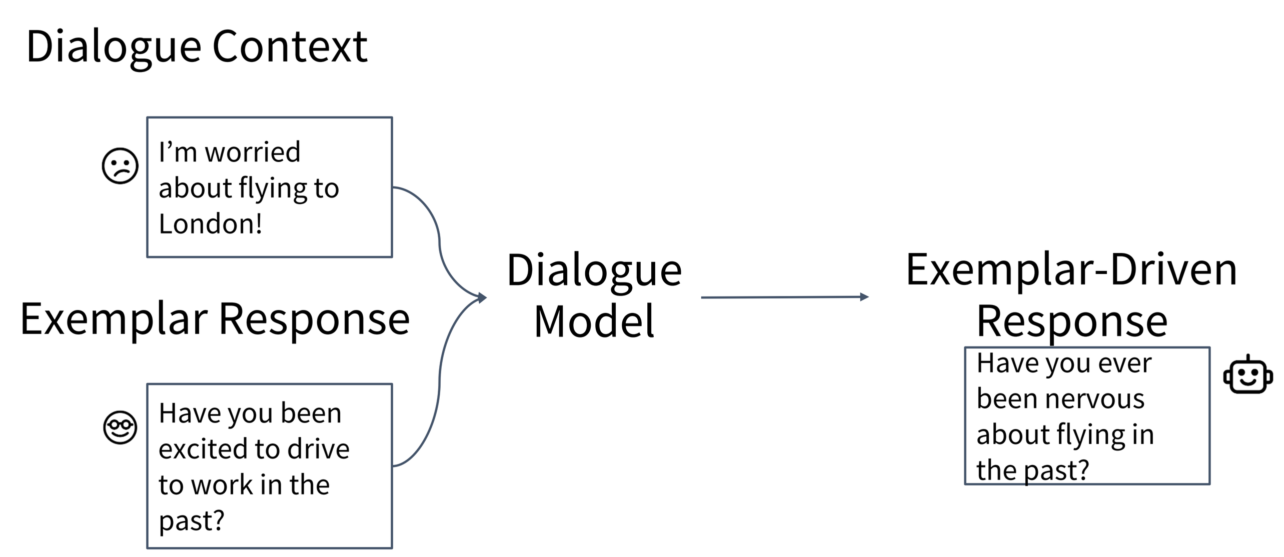
Dialogue systems based on large LMs (DialoGPT[1], Blenderbot[2], etc.)

- ✓ Coherent responses
- ✗ Follow system-level goals

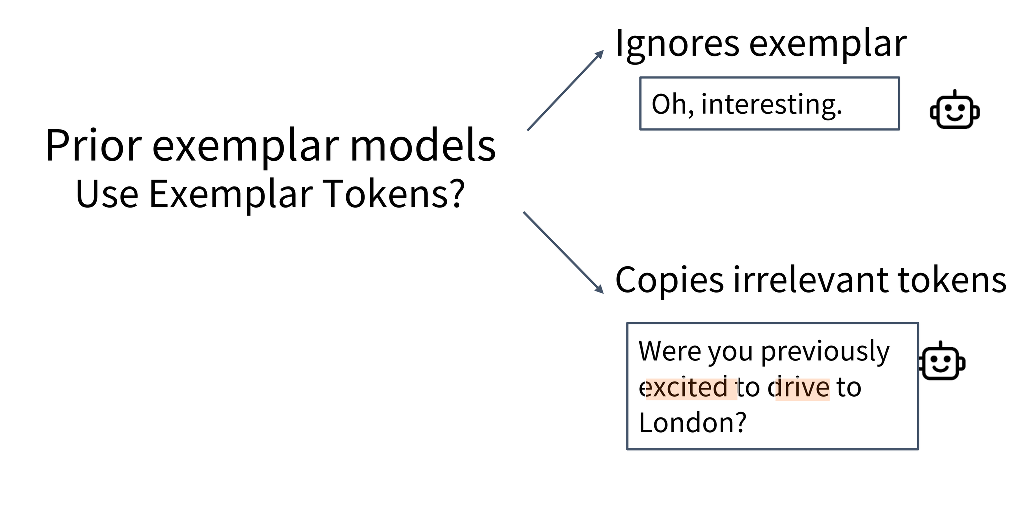
Dialogue systems with control based on persona, emotions, knowledge etc.

- ✓ Coherent responses
- ✓ Follow system-level goals
- ✗ Require labelled data for each new goal

Exemplar based control



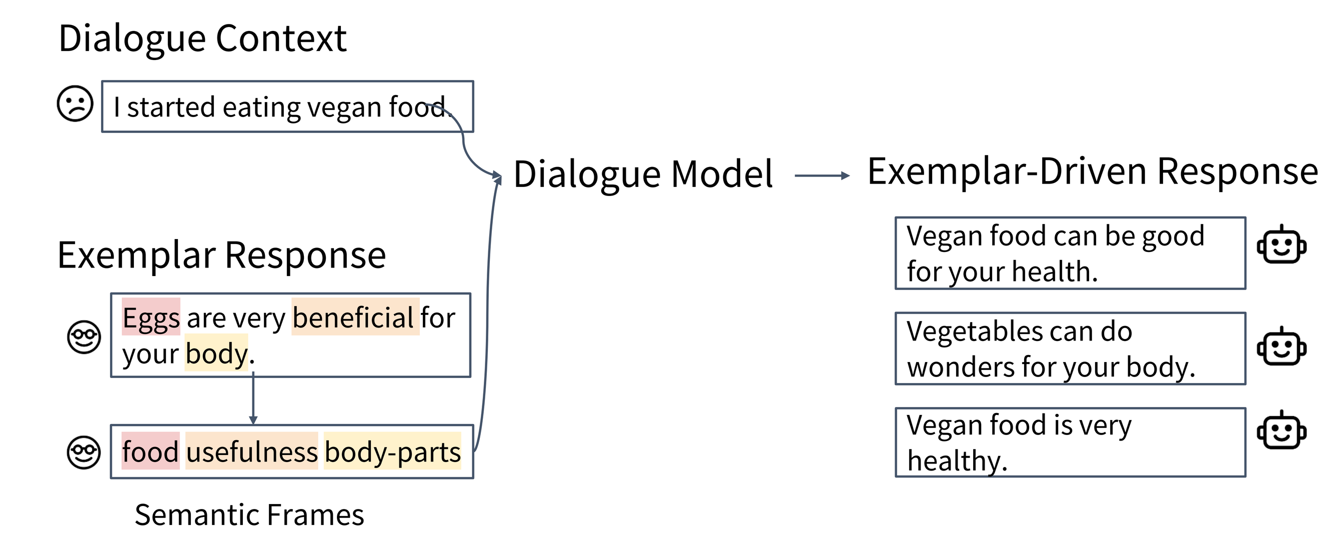
Exemplar driven response generation



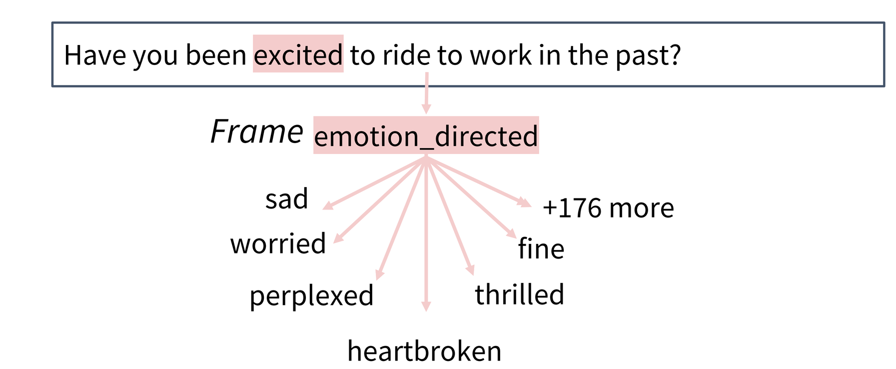
Challenges with existing approaches

Key Idea

Our model, EDGE, conditions response generation on the **semantic frames** of response exemplars.



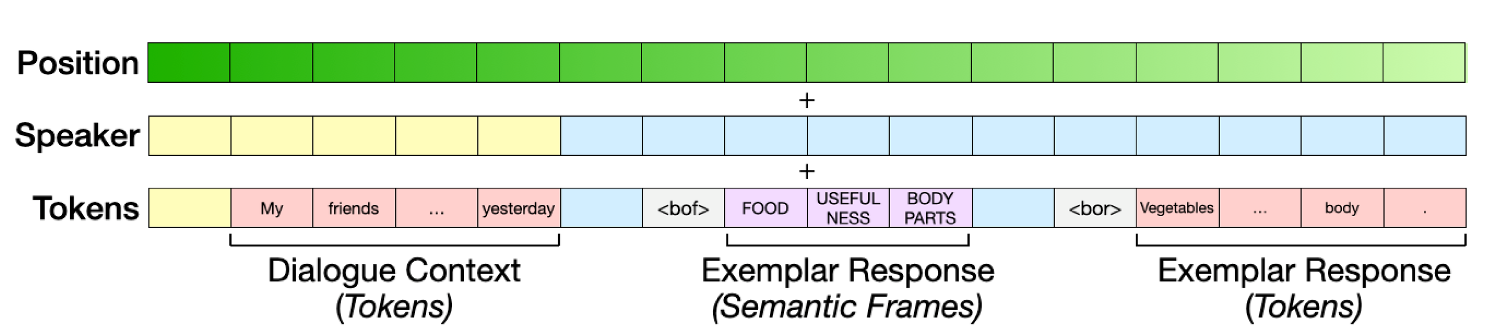
Semantic Frames[3]



Model

Model architecture

Based on GPT-2[4] (substitutable with other language models)



Training	Context Tokens	Gold Response Frames	Gold response Tokens
Testing	Context Tokens	Exemplar Frames	[Response generated]

Improving Robustness

to incorrect detections & to irrelevant frames

Drop frames	food usefulness ~~body-parts~~
Shuffle frames	body-parts usefulness food
Add random frames	food emotion usefulness body-parts

Advantages of EDGE

- ✓ Coherent responses
- ✓ Follow system-level goals
- ✓ No explicit training labels

Experiment: Open-domain

Dataset: DailyDialog[5]

Quantitative Results

Model	Dist-2	Dist-3	MaUdE	Coherent	Fluent	Consistent	Interesting	Semantics
Retrieval	0.294	0.526	0.921	2.41	2.61	2.48	2.32	-
GPT2-Gen	0.249	0.494	0.905	2.42	2.55	2.41*	2.18*	-
LSTM-Tokens	0.182	0.380	0.890	2.04*	2.10*	2.11*	1.89*	2.17
LSTM-Frames	0.185	0.392	0.901	2.36*	2.30*	2.33*	1.97*	**2.29**
GPT2-Tokens	0.254	0.513	**0.927**	2.19*	2.47*	2.29*	2.11*	2.04*
EDGE (Ours)	**0.278**	**0.571**	0.922	**2.52**	**2.63**	**2.56**	**2.39**	2.24
Human	0.385	0.720	0.911	2.76	2.69	2.78	2.44	-

(1) EDGE > all other models
(2) frame-based models > token-based models
(3) GPT-2-based models > LSTM-based models

Qualitative Results

Context	*Human1*: they sell everything. *Human2*: well, i want chinese food.
Retrieved	well, what do you want to eat?
Frames	WHAT DESIRING INGESTION ?
GPT2-Gen	it's a good idea.
LSTM-Tokens	well, what's the you do?
LSTM-Frames	i hope so.
GPT2-Tokens	i'm not sure what to get.
EDGE (Ours)	you want to eat something chinese?

Context	*Human1*: jeff, i'm going to the supermarket. do you want to come with me? *Human2*: i think the supermarket is closed now.
Retrieved	i know. i intent to go to the store today.
Frames	AWARENESS PURPOSE MOTION BUSINESSES TEMPORAL-COLLOCATION
GPT2-Gen	what a pity!
LSTM-Tokens	yes, i'm sorry to go with you.
LSTM-Frames	where is the market?
GPT2-Tokens	where is the supermarket?
EDGE (Ours)	i know, but i'm planning to go to the bank today.

(1) EDGE generates longer and more specific responses
(2) EDGE generates coherent responses even with irrelevant or missing frames
(3) EDGE occasionally diverges from exemplar response

Experiment: Scam Defense

Dataset: Scam Email Defense[6]

Quantitative Results

Model	Coherence	Intent	Engagement
GPT2-Gen	2.10	33.0	70.1
EDGE	**2.39**	**79.7**	**87.3**

Qualitative Results

Scam Email	Anti-Scam Exemplar	EDGE Response
"i want you to assist in investing money […] my son shall tell you where the funds are being maintained and discuss remuneration for your services"	*Intent: Ask For Contact or Location* "Are you located near me? Could we meet up soon about this? Tell me where your place is."	"Would you please be able to meet me about the location of the funds?"

Conclusion

- EDGE generates responses based on **semantic frames** of exemplar responses.
- EDGE achieves **coherent responses** that **preserve system-level goals** (implicitly present in exemplars) as demonstrated in our experiments.
- EDGE's key advantages are its **controllability** and **zero-shot performance**

Check out the code for this project!
https://github.com/prakharguptaz/EDGE-exemplars

References

[1] Yizhe Zhang, Siqi Sun, Michel Galley, Yen-Chun Chen, Chris Brockett, Xiang Gao, Jianfeng Gao, Jingjing Liu, and Bill Dolan. 2020. DIALOGPT : Largescale generative pre-training for conversational response generation.
[2] Stephen Roller, Emily Dinan, Naman Goyal, Da Ju, Mary Williamson, Yinhan Liu, Jing Xu, Myle Ott, Kurt Shuster, Eric M Smith, et al. 2020. Recipes for building an open-domain chatbot. arXiv preprint arXiv:2004.13637
[3] Collin F. Baker, Charles J. Fillmore, and John B. Lowe. 1998. The berkeley framenet project. *Volume 1,* ACL '98/COLING '98
[4] Alec Radford, Jeff Wu, Rewon Child, David Luan, Dario Amodei, and Ilya Sutskever. 2019. Language models are unsupervised multitask learners
[5] Yanran Li, Hui Su, Xiaoyu Shen, Wenjie Li, Ziqiang Cao, Shuzi Niu. DailyDialog A manually labelled multi-turn dialogue dataset EMNLP 2017
[6] https://www.kaggle.com/rtatman/fraudulent-email-corpus