
A Knowledge Plug-and-play Test Bed for Open-domain Dialogue Generation

— Xiangci Li, Linfeng Song, Lifeng Jin, —
Haitao Mi, Jessica Ouyang, Dong Yu

Background: Wizard of Wikipedia for Dialogue Generation

- Knowledge-based open-domain dialogue generation aims to build systems that talk to humans on various domains with mined support knowledge.
- The wizard and apprentice speak to each other
- The wizard can access to a knowledge source
 - Wikipedia
- Apprentice does not access to knowledge source
- Each wizard utterance is annotated with knowledge sentence.
- Tasks
 - Dialogue knowledge selection
 - Dialogue response generation

@inproceedings{dinan2019wizard, author={Emily Dinan and Stephen Roller and Kurt Shuster and Angela Fan and Michael Auli and Jason Weston}, title={{W}izard of {W}ikipedia: Knowledge-powered Conversational Agents}, booktitle = {Proceedings of the International Conference on Learning Representations (ICLR)}, year={2019}, }

Introduction

- Task: **Multi-source** open-domain dialogue generation.
 - How does knowledge from multiple sources interact?
 - What if new knowledge is added after the model is trained?
- Prior works either only use single-source knowledge for dialogue generation (e.g. WoW), or use post-retrieved multi-source dataset, which is unable to evaluate the knowledge selection performance.
- Multi-source Wizard of Wikipedia (Ms.WoW) dataset

Ms.WoW

Response

It was formed in 1965. The Pepsi-Cola company and Frito-Lay merged to form one big company.

Gold WoW sentence

PepsiCo was formed in 1965 with the merger of the Pepsi-Cola Company and Frito-Lay, Inc. PepsiCo has since expanded from its namesake product Pepsi to a broader range of food and beverage brands, the largest of which included an acquisition of Tropicana Products in 1998 and the Quaker Oats Company in 2001, which added the Gatorade brand to its portfolio.

Ms.WoW Knowledge Tuples

	Source	Gold
(“ , ‘formed’, ‘PepsiCo’, ‘in 1965’, “)	Sem. frm.	Yes
(‘its’, “ , ‘has’, ‘namesake product Pepsi’, “ , “)	OPIEC	No
(‘largest of which’, “ , ‘have included acquisition of Tropicana Products in 1998’, ‘beverage brands’, “ , “)	OPIEC	No
(‘largest of which’, “ , ‘have included Quaker Oats Company in 2001’, ‘food brands’, “ , “)	OPIEC	No
(‘Quaker Oats Company in 2001’, “ , ‘added Gatorade brand to’, ‘portfolio’, “ , “)	OPIEC	Yes
(‘Frito-Lay’, ‘parent organization’, ‘PepsiCo’)	Wikidata	Yes
(‘Pepsi’, ‘instance of’, ‘cola’)	Wikidata	Yes

Table 1: An example decomposition of a gold knowledge sentence from WoW into tuples from multiple sources in Ms.WoW. Tuples not used in the response are not labeled as gold tuples.

Ms.WoW Statistics

Multi-Source WoW	Train	Valid Seen	Valid Unseen	Test Seen	Test Unseen
Number of utterances	166787	8909	8806	8715	8782
Number of dialogues	18430	981	967	965	968
Number of topics	1247	545	54	533	58
Avg turns per dialogue	9.05	9.08	9.13	9.03	9.07
% of Wizard turns with knowledge	61.8	62.5	65.0	62.8	61.8
Avg non-zero # of knowledge per utterance	5.0	5.0	5.1	5.0	4.9
Avg # of (gold) Knowledge per Utterance	13.8 (1.55)	13.4 (1.49)	14.1 (1.59)	13.9 (1.55)	15.2 (1.59)
% of (gold) OPIEC	64.9 (57.4)	65.2 (59.2)	57.9 (51.5)	65.4 (58.2)	67.4 (58.7)
% of (gold) semantic frame	4.55 (9.87)	4.17 (8.73)	3.82 (8.85)	4.37 (8.98)	5.08 (12.1)
% of (gold) Wikidata	17.2 (13.9)	17.2 (12.9)	21.7 (16.9)	17.2 (13.9)	16.8 (12.4)
% of (gold) Wikipedia	13.4 (18.8)	13.5 (19.2)	16.6 (22.8)	13.0 (18.9)	10.7 (16.7)

Table 2: Statistics of our Multi-Source Wizard of Wikipedia.

Differences Among Knowledge Sources

	sbj	neg	rel	obj	tmp	spa	total
OPIEC	2.2	1.0	3.3	2.8	2.1	2.3	8.7
Sem. frm.	6.3	—	1.0	14.3	3.9	8.2	22.1
Wikidata	1.7	—	2.2	1.7	—	—	5.6
Wikipedia	—	—	—	—	—	—	24.9

Table 3: Average of number of words per non-empty knowledge attribute.

Introduction

- Multi-source Wizard of Wikipedia (Ms.WoW) dataset extends the WoW dataset, which has clean annotation of utterance-level knowledge selection, to multiple knowledge sources: OPIEC, semantic frames, Wikidata & Wikipedia.
- We use Ms.WoW for **dialogue knowledge plug-and-play**, which aims to test an already trained dialogue model on using new support knowledge from previously unseen sources in a zero-shot fashion.

Dialogue Knowledge Plug-and-Play: Knowledge Selection

Training	<i>Test Seen + Unseen</i>		
	P	R	F1
Full Kn.	0.384	0.380	0.382
– OPIEC	0.368	0.274	0.314
– Sem. frm.	0.404	0.333	0.365
– Wikidata	0.446	0.303	0.361
– Wikipedia	0.497	0.319	0.389

Table 4: Dialogue knowledge selection performance on Ms.WoW test set (seen + unseen).

Training	<i>OPIEC</i>			<i>Sem. frm.</i>			<i>Wikidata</i>			<i>Wikipedia</i>		
	P	R	F1	P	R	F1	P	R	F1	P	R	F1
Full Knowledge	0.340	0.347	0.343	0.591	0.639	0.614	0.301	0.397	0.342	0.550	0.321	0.406
– OPIEC	<i>0.301</i>	<i>0.194</i>	<i>0.236</i>	0.549	0.647	0.594	0.256	0.180	0.211	0.459	0.384	0.418
– Sem. frm.	0.352	0.275	0.309	<i>0.580</i>	<i>0.585</i>	<i>0.583</i>	0.360	0.270	0.309	0.460	0.418	0.438
– Wikidata	0.401	0.265	0.319	0.587	0.640	0.612	<i>0.277</i>	<i>0.087</i>	<i>0.133</i>	0.505	0.391	0.441
– Wikipedia	0.473	0.244	0.322	0.569	0.701	0.628	0.436	0.268	0.332	<i>0.519</i>	<i>0.376</i>	<i>0.436</i>

Table 5: Dialogue knowledge selection performance on Ms.WoW test set (seen + unseen) by knowledge source. All knowledge sources are present during testing, simulating the scenario where a new knowledge source becomes available at test time.

Dialogue Knowledge Plug-and-Play: Knowledge Selection

Training	<i>OPIEC</i>			<i>Sem. frm.</i>			<i>Wikidata</i>			<i>Wikipedia</i>		
	P	R	F1	P	R	F1	P	R	F1	P	R	F1
Full Knowledge	0.340	0.347	0.343	0.591	0.639	0.614	0.301	0.397	0.342	0.550	0.321	0.406
– OPIEC	<i>0.301</i>	<i>0.194</i>	<i>0.236</i>	0.549	0.647	0.594	0.256	0.180	0.211	0.459	0.384	0.418
– Sem. frm.	0.352	0.275	0.309	<i>0.580</i>	<i>0.585</i>	<i>0.583</i>	0.360	0.270	0.309	0.460	0.418	0.438
– Wikidata	0.401	0.265	0.319	0.587	0.640	0.612	<i>0.277</i>	<i>0.087</i>	<i>0.133</i>	0.505	0.391	0.441
– Wikipedia	0.473	0.244	0.322	0.569	0.701	0.628	0.436	0.268	0.332	<i>0.519</i>	<i>0.376</i>	<i>0.436</i>

Table 5: Dialogue knowledge selection performance on Ms.WoW test set (seen + unseen) by knowledge source. All knowledge sources are present during testing, simulating the scenario where a new knowledge source becomes available at test time.

Training & Testing	<i>OPIEC</i>			<i>Sem. frm.</i>			<i>Wikidata</i>			<i>Wikipedia</i>		
	P	R	F1	P	R	F1	P	R	F1	P	R	F1
– OPIEC	–	–	–	0.584	0.584	0.584	0.293	0.199	0.237	0.459	0.395	0.425
– Sem. frm.	0.349	0.274	0.307	–	–	–	0.352	0.266	0.303	0.452	0.413	0.431
– Wikidata	0.389	0.268	0.317	0.579	0.623	0.600	–	–	–	0.431	0.322	0.368
– Wikipedia	0.472	0.226	0.306	0.575	0.687	0.626	0.452	0.245	0.318	–	–	–

Table 6: Dialogue knowledge selection performance on the Ms.WoW test set (seen + unseen), excluding the ablated knowledge source for each model; both training and testing are conducted with one knowledge source missing, simulating the scenario where one knowledge source never becomes available.

Dialogue Knowledge Plug-and-Play: Response Generation

Configurations	Training	R-1	R-2	R-L	F1	K-P	K-R	K-F1
No knowledge	No knowledge	0.189	0.043	0.159	0.207	–	–	–
WoW Full knowledge	WoW Full knowledge	0.261	0.101	0.225	0.265	0.502	0.162	0.245
Ms.WoW Full knowledge	Ms.WoW Full knowledge	0.259	0.094	0.222	0.264	0.460	0.159	0.236
	– OPIEC	0.247	0.084	0.212	0.251	0.433	0.146	0.219
	– sem. frm.	0.256	0.093	0.219	0.260	0.448	0.158	0.234
	– Wikidata	0.256	0.093	0.220	0.261	0.440	0.154	0.228
	– Wikipedia	0.251	0.089	0.215	0.256	0.459	0.164	0.242
WoW Gold knowledge	WoW Gold knowledge	0.317	0.150	0.278	0.317	0.387	0.528	0.446
Ms.WoW Gold knowledge	Ms.WoW Gold knowledge	0.322	0.149	0.280	0.321	0.311	0.576	0.404
	– OPIEC	0.306	0.133	0.265	0.302	0.302	0.560	0.392
	– sem. frm.	0.321	0.147	0.280	0.321	0.316	0.582	0.409
	– Wikidata	0.320	0.145	0.279	0.319	0.313	0.580	0.406
	– Wikipedia	0.319	0.145	0.277	0.318	0.311	0.585	0.406

Table 9: Response generation performance on test set (seen + unseen). All knowledge sources are present during testing, simulating the scenario where a new knowledge source becomes available at test time. Full knowledge refers to no knowledge selection, where all available candidate knowledge is used; gold knowledge refers to oracle knowledge selection.

Dialogue Knowledge Plug-and-Play: Results

- Selection quality of the knowledge matters: Dialogue generation with gold knowledge significantly outperform using retrieved knowledge.

Configurations	Training	R-1	R-2	R-L	F1	K-P	K-R	K-F1
No knowledge	No knowledge	0.189	0.043	0.159	0.207	–	–	–
WoW Full knowledge	WoW Full knowledge	0.261	0.101	0.225	0.265	0.502	0.162	0.245
Ms.WoW Full knowledge	Ms.WoW Full knowledge	0.259	0.094	0.222	0.264	0.460	0.159	0.236
	– OPIEC	0.247	0.084	0.212	0.251	0.433	0.146	0.219
	– sem. frm.	0.256	0.093	0.219	0.260	0.448	0.158	0.234
	– Wikidata	0.256	0.093	0.220	0.261	0.440	0.154	0.228
	– Wikipedia	0.251	0.089	0.215	0.256	0.459	0.164	0.242
WoW Gold knowledge	WoW Gold knowledge	0.317	0.150	0.278	0.317	0.387	0.528	0.446
Ms.WoW Gold knowledge	Ms.WoW Gold knowledge	0.322	0.149	0.280	0.321	0.311	0.576	0.404
	– OPIEC	0.306	0.133	0.265	0.302	0.302	0.560	0.392
	– sem. frm.	0.321	0.147	0.280	0.321	0.316	0.582	0.409
	– Wikidata	0.320	0.145	0.279	0.319	0.313	0.580	0.406
	– Wikipedia	0.319	0.145	0.277	0.318	0.311	0.585	0.406

Table 9: Response generation performance on test set (seen + unseen). All knowledge sources are present during testing, simulating the scenario where a new knowledge source becomes available at test time. Full knowledge refers to no knowledge selection, where all available candidate knowledge is used; gold knowledge refers to oracle knowledge selection.

Dialogue Knowledge Plug-and-Play: Results

- The performance varies across different knowledge sources.
- The larger the ablated training knowledge source is, the worse the test performance is.

Training	<i>Test Seen + Unseen</i>		
	P	R	F1
Full Kn.	0.384	0.380	0.382
– OPIEC	0.368	0.274	0.314
– Sem. frm.	0.404	0.333	0.365
– Wikidata	0.446	0.303	0.361
– Wikipedia	0.497	0.319	0.389

Table 4: Dialogue knowledge selection performance on Ms.WoW test set (seen + unseen).

Dialogue Knowledge Plug-and-Play: Results

- “More is better”: During zero-shot adaptation, including more knowledge inputs improve the performance. This is similar to in-context learning.

Training	OPIEC			Sem. frm.			Wikidata			Wikipedia		
	P	R	F1	P	R	F1	P	R	F1	P	R	F1
Full Knowledge	0.340	0.347	0.343	0.591	0.639	0.614	0.301	0.397	0.342	0.550	0.321	0.406
– OPIEC	0.301	0.194	0.236	0.549	0.647	0.594	0.256	0.180	0.211	0.459	0.384	0.418
– Sem. frm.	0.352	0.275	0.309	0.580	0.585	0.583	0.360	0.270	0.309	0.460	0.418	0.438
– Wikidata	0.401	0.265	0.319	0.587	0.640	0.612	0.277	0.087	0.133	0.505	0.391	0.441
– Wikipedia	0.473	0.244	0.322	0.569	0.701	0.628	0.436	0.268	0.332	0.519	0.376	0.436

Table 5: Dialogue knowledge selection performance on Ms.WoW test set (seen + unseen) by knowledge source. All knowledge sources are present during testing, simulating the scenario where a new knowledge source becomes available at test time.

Training & Testing	OPIEC			Sem. frm.			Wikidata			Wikipedia		
	P	R	F1	P	R	F1	P	R	F1	P	R	F1
– OPIEC	–	–	–	0.584	0.584	0.584	0.293	0.199	0.237	0.459	0.395	0.425
– Sem. frm.	0.349	0.274	0.307	–	–	–	0.352	0.266	0.303	0.452	0.413	0.431
– Wikidata	0.389	0.268	0.317	0.579	0.623	0.600	–	–	–	0.431	0.322	0.368
– Wikipedia	0.472	0.226	0.306	0.575	0.687	0.626	0.452	0.245	0.318	–	–	–

Table 6: Dialogue knowledge selection performance on the Ms.WoW test set (seen + unseen), excluding the ablated knowledge source for each model; both training and testing are conducted with one knowledge source missing, simulating the scenario where one knowledge source never becomes available.

Dialogue Knowledge Plug-and-Play

- Both SLMs and LLMs show similar trends.

Configurations	Testing	R-1	R-2	R-L	F1	K-P	K-R	K-F1
No knowledge	No knowledge	0.187	0.031	0.147	0.202	–	–	–
WoW Full knowledge	WoW Full knowledge	0.202	0.046	0.153	0.216	0.216	0.184	0.199
Ms.WoW Full knowledge	Ms.WoW Full knowledge	0.196	0.044	0.149	0.212	0.382	0.260	0.310
	– OPIEC	0.185	0.034	0.141	0.199	0.258	0.157	0.195
	– sem. frm.	0.189	0.038	0.143	0.203	0.343	0.225	0.272
	– Wikidata	0.196	0.043	0.149	0.211	0.379	0.256	0.306
	– Wikipedia	0.196	0.043	0.149	0.210	0.373	0.247	0.297
WoW Gold knowledge	WoW Gold knowledge	0.218	0.053	0.170	0.226	0.049	0.086	0.062
Ms.WoW Gold knowledge	Ms.WoW Gold knowledge	0.230	0.061	0.176	0.236	0.114	0.351	0.172
	– OPIEC	0.193	0.038	0.147	0.205	0.072	0.219	0.109
	– sem. frm.	0.215	0.051	0.165	0.223	0.095	0.289	0.144
	– Wikidata	0.220	0.056	0.170	0.226	0.108	0.325	0.162
	– Wikipedia	0.224	0.057	0.172	0.230	0.110	0.332	0.165

Table 11: Vicuna-13B response generation performance on the test set (seen + unseen).

Summary of Observations

- This work shows how the model behavior changes as we vary the inputs, in a RAG-like setting.
- Obviously training model on knowledge intensive inputs help a lot.
- Meanwhile, simply adding more knowledge during inference also helps.

Conclusion

- Ms.WoW dataset
- Dialogue knowledge plug-and-play challenge