SCAI-QReCC Shared Task on Conversational Question Answering

Svitlana Vakulenko
Amazon Spain
(prev. University of Amsterdam)

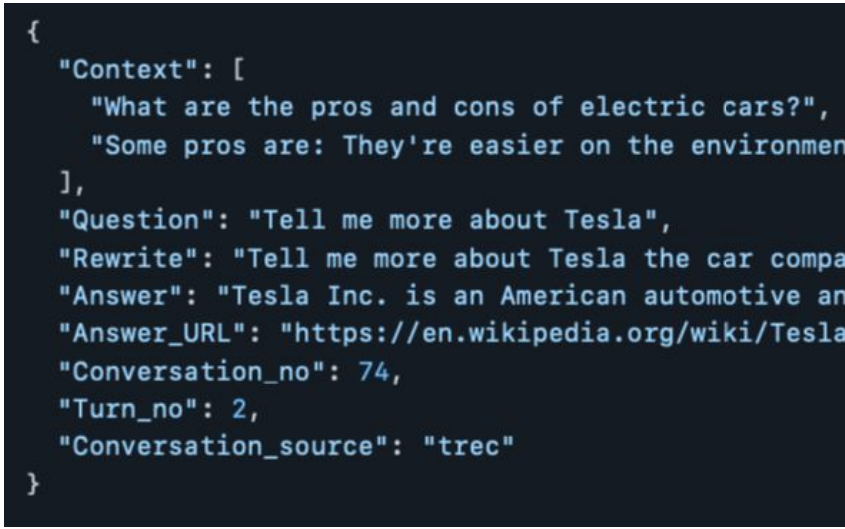
Johannes Kiesel
Bauhaus-Universität Weimar

Maik Fröbe
Martin-Luther-Universität Halle-Wittenberg

Dataset: SCAI-QReCC

- Conversational QA dataset QReCC (NAACL'21)

- 14K conversations
- 81K question-answer pairs
- 54M passages from web pages



https://github.com/apple/ml-qrecc

https://zenodo.org/record/5543685#.YV_OEC0RppR

Stats

- 29 runs
- 4 teams: Rachael, Rali, Torch, Ultron
- 3 baselines: gpt3, simple, basic
- 16,736 answers

Baselines

- basic
  - Predicted Answer = Question
- simple (open source to get people started)
  - Question rewriting: Return question as-is
  - Passage retrieval: BM25 as in the dataset paper (NAACL'21)
  - Question answering: Return the sentence from passages with highest noun phrase overlap with question
- gpt3
  - 50 USD via OpenAI API

Teams

- **Rachael** or **Serial Killer Android Miss Bloody Rachel**, first appears in *Viewtiful Joe 2*.
- **Rali-QA**
- **Torch** from Adventures of Sonic the Hedgehog
- **Ultron** an evil android portrayed by James Spader in *Avengers: Age of Ultron* (2015

https://en.wikipedia.org/wiki/List_of_fictional_robots_and_androids
https://viewtifuljoe.fandom.com/wiki/Miss_Bloody_Rachel
https://sonic.fandom.com/wiki/Torch

Automatic Evaluation

- QR: R1-R (ROUGE-1)
- PR: MRR (Mean reciprocal rank)
- QA:
  - EM (exact match)
  - F1
  - R1-R (ROUGE-1)
  - POSS (POSSCORE, Liu et al., 2021)
  - SAS (Semantic Answer Similarity, Risch et al., 2021)
  - BERT(BERTScore, Zhang et al., 2020)
  - B-KPQA & R-KPQA (BERTScore KPQA & ROUGE-L KPQA, Lee et al., 2021)

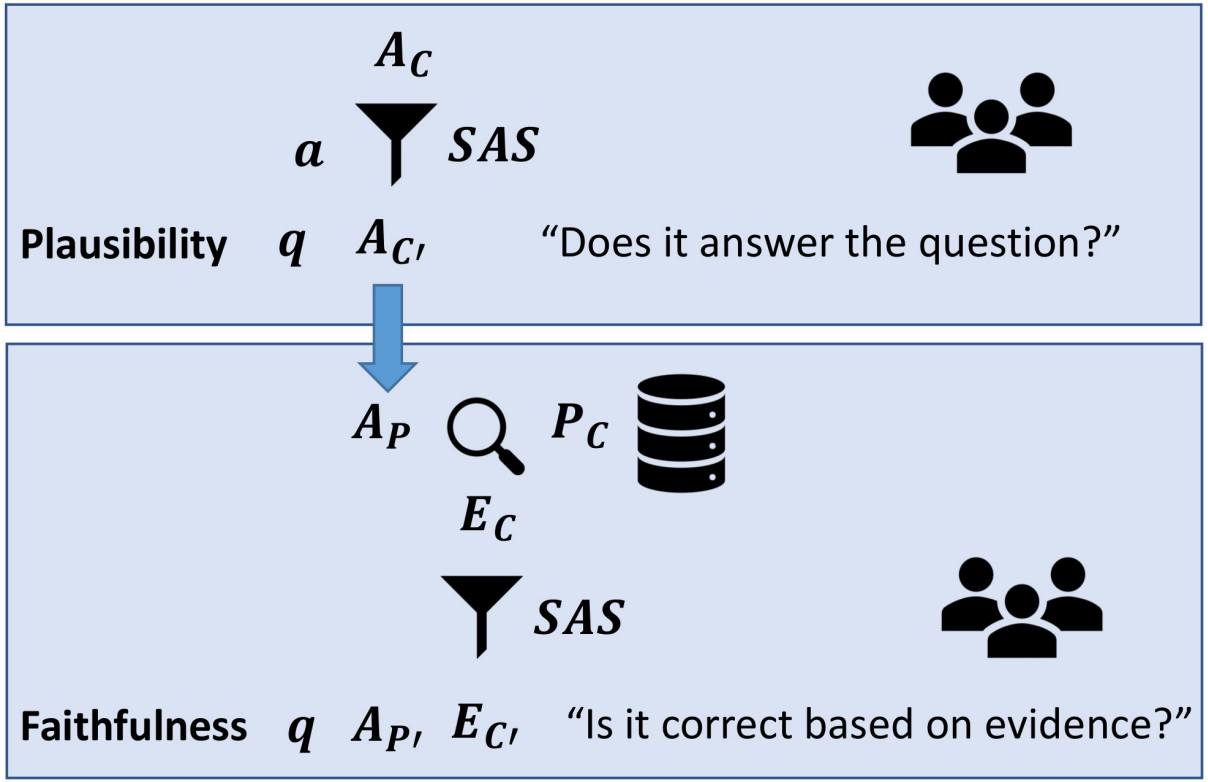
Results on the Original Dataset

Team	Run	QR	MRR	EM	F1	R1	POSS	SAS	BERT	BKPQA	RKPQA
Original questions											
-	Basic baseline	-	-	0.000	0.114	0.095	1.283	0.207	0.422	0.432	0.064
-	GPT3 baseline	-	-	0.001	0.149	0.148	1.305	0.264	0.448	0.467	0.134
-	Simple baseline	0.571	0.065	0.001	0.067	0.150	1.490	0.162	0.367	0.426	0.097
rachael	2021-09-04-10-38-07	-	0.056	0.002	0.138	0.193	**1.583**	0.163	0.410	0.476	0.135
rachael	2021-09-08-07-07-57	0.675	0.135	0.006	0.187	**0.226**	1.570	**0.277**	**0.452**	**0.498**	0.175
rachael	2021-09-08-07-09-57	0.682	0.128	0.006	0.186	0.226	1.558	0.269	0.448	0.494	0.175
rachael	2021-09-08-15-40-34	0.679	0.133	0.007	0.176	0.211	1.456	0.254	0.420	0.460	0.164
rachael	2021-09-08-21-49-44	0.681	0.130	0.008	0.177	0.211	1.461	0.246	0.422	0.462	0.167
rachael	2021-09-15-09-05-06	0.673	**0.158**	**0.011**	0.179	0.212	1.333	0.254	0.405	0.444	0.172
rachael	2021-09-15-09-06-44	0.681	0.150	0.010	0.179	0.211	1.369	0.249	0.408	0.449	0.169
rachael	2021-09-15-09-07-49	0.676	0.157	0.010	0.187	0.219	1.399	0.264	0.418	0.457	0.175
rachael	2021-09-15-09-08-40	**0.685**	0.149	0.010	**0.189**	0.222	1.458	0.259	0.428	0.470	**0.178**
torch	usi_T5_raw2	0.657	0.082	0.001	0.137	0.200	1.451	0.221	0.415	0.467	0.117

Results on Human Rewritten Questions

Team	Run	QR	MRR	EM	F1	R1	POSS	SAS	BERT	BKPQA	RKPQA
Human rewritten questions											
-	Basic baseline		-	0.000	0.224	0.205	1.555	0.351	0.517	0.472	0.132
-	Simple baseline		**0.398**	0.001	0.098	0.282	1.666	0.372	-	-	-
rachael	2021-09-04-10-39-42		0.359	0.011	0.267	0.331	**1.674**	0.398	0.534	0.562	0.258
rachael	2021-09-06-09-21-43		0.359	0.018	0.290	0.339	1.649	**0.430**	**0.549**	**0.570**	0.277
rachael	2021-09-15-19-36-31		0.385	**0.028**	**0.302**	**0.345**	1.618	0.420	0.544	0.566	**0.290**
rali-qa	2021-09-09-13-01-07		0.269	0.003	0.166	0.212	1.385	0.264	0.407	0.457	0.174
ultron	2021-09-04-17-28-07		-	0.001	0.183	0.186	1.357	0.301	0.463	0.457	0.121
ultron	2021-09-08-15-04-28		-	0.015	0.261	0.258	1.565	0.383	0.533	0.539	0.220
ultron	2021-09-08-15-07-30		-	0.001	0.187	0.189	1.380	0.306	0.472	0.465	0.123
ultron	2021-09-08-15-08-00		-	0.004	0.247	0.236	1.597	0.379	0.536	0.525	0.177
ultron	bart-large_top1bm25		-	0.000	0.017	0.017	0.150	0.111	0.046	0.048	0.016
ultron	distilbart-xsum-12-1_top1bm25		-	0.000	0.019	0.020	0.170	0.113	0.050	0.054	0.018
ultron	distilbart-xsum-12-3_top1bm25		-	0.000	0.022	0.023	0.175	0.117	0.052	0.056	0.021
ultron	rag-bm25_100		-	0.004	0.247	0.236	1.597	0.379	0.536	0.525	0.177
ultron	rag-dpr-hard-neg-bm25-top10		-	0.015	0.261	0.258	1.565	0.383	0.533	0.539	0.220
ultron	rag-ft-dpr-hard-neg-bm25_10		-	0.015	0.261	0.258	1.565	0.383	0.533	0.539	0.220

Human Evaluation



Human Evaluation

Team	Run	Question	Plausible	Implausible	Malformed	Faithful	Unfaithful
rachael	2021-09-04-10-39-42	rewritten	**183**	5	4	**37**	2
rachael	2021-09-08-21-49-44	original	133	6	4	30	1
rachael	2021-09-08-07-07-57	original	120	4	5	30	0
rachael	2021-09-15-09-07-49	original	103	4	6	29	1
-	GPT3 baseline	original	149	4	8	28	3
ultron	rag-bm25_100	rewritten	173	15	6	27	2
rachael	2021-09-06-09-21-43	rewritten	158	4	3	26	**4**
ultron	2021-09-08-15-04-28	rewritten	149	**16**	6	24	1
rachael	2021-09-15-19-36-31	rewritten	132	2	2	24	0
rachael	2021-09-15-09-06-44	original	73	0	4	22	1
rachael	2021-09-08-07-09-57	original	75	2	4	16	1
rali-qa	2021-09-09-13-01-07	rewritten	33	6	11	16	1
rachael	2021-09-08-15-40-34	original	41	6	2	14	3
torch	usi T5 raw2	original	36	7	**16**	14	0
ultron	2021-09-04-17-28-07	rewritten	117	13	7	13	0
rachael	2021-09-15-09-08-40	original	52	4	4	10	0
ultron	BART-large-top1BM25	rewritten	29	3	11	10	0
rachael	2021-09-15-09-05-06	original	52	2	1	9	0
rachael	2021-09-04-10-38-07	original	41	2	0	6	1
-	Simple baseline	rewritten	14	2	3	1	0
-	Simple baseline	original	0	0	1	0	0
Total			1863	107	108	386	21

Conclusion

- End-to-end conversational QA model is an open problem
  - All participants used sparse indexes with BM25 retrieval
  - Open challenge: Build dense passage retrieval index over the full QReCC dataset
- QA models produce fluent answers
  - Answers might be incorrect
  - Open challenge: Ground answers in external information referenced by the QA model

- We proposed a method to discover new plausible and faithful answers

- See you at the SCAI workshop at SIGIR 2022

Conversational Question Answering

Conversational Question: "Do NPs or PAs make more?"

Rewritten Question: "Do nurse practitioners or physician assistants make more?"

Human Answer: "According to the Bureau of Labor and Statistics (BLS), the 2017 median pay for Nurse Practitioners is $110,930 per year, or $53.33 per hour. In comparison, 2017 median pay for Physician Assistants is $104,860 per year, or $50.41 per hour."

Generated Answer: "The average salary for a nurse practitioner is $65,000. The average salary for a physician's assistant is $30,000."

Extracted Answer: "The BLS reports that the median annual wage for nurse practitioners was $109,820 as of May 2019 , while the median annual wage for physician assistants reached $112,260 during the same month."