Appendix: AutoTemplate: A Simple Recipe for Lexically Constrained Text Generation

Hayate Iso
Megagon Labs
hayate@megagon.ai

Keywords: government , ability , companies , legal
Reference: Generally , the government has the ability to compel the cooperation of private companies and assure them legal immunity with a valid court order .
CBART: The government has restricted the ability of insurance companies to take legal action .
AutoTemplate: The government has the ability to force companies to comply with legal requirements, he said.

Table 1: Example generations for the keywords-to-sentence generation on One-billion-word.

Keywords: time , voters , primary , days
Reference: At the same time , he said the more he appears before voters , the better he does on primary days .
CBART: The last time , the voters were in the primary , two days before Nov .
AutoTemplate: At the same time , voters will be able to cast their ballots during the primary days , he said.

Table 2: Example generations for the keywords-to-sentence generation on One-billion-word.

A More qualitative examples

Table 1-4 show more qualitative examples of keywords-to-sentence generation task.

B Additional Experimental Details

B.1 Training details

Major hyper-parameters for training models are reported in Table 5 following the "Show-You-Work" style suggested by Dodge et al. (2019).

Keywords: experience , top , easily , driver
Reference: my experience with lv cans was top notch . cab was easily flagged just off the strip , the route was direct and the driver was very nice .
CBART: the whole experience was top notch , easily by the driver .
AutoTemplate: i had a great experience with this company. they were on top of everything. i was easily able to get a driver to pick me up at my hotel.

Table 3: Example generations for the keywords-to-sentence generation on Yelp.

C Experimental details of InstructGPT

We empirically evaluated the zero-shot capability of InstructGPT (Ouyang et al., 2022) for keywords-to-sentence generation task. We specifically used `text-davinci-003` checkpoint and the prompt: `"Please create a sentence that must contain the following keywords: {{', '.join(keywords)}}."` to generate sentences that includes the pre-specified keywords. To obtain deterministic output text, we use the temperature parameter 0.

D Full results of keywords-to-sentence generation

We show non-aggregated results of keywords-to-sentence generation in Table 6. The results show that the AutoTemplate consistently outperforms baseline models.

References

Jesse Dodge, Suchin Gururangan, Dallas Card, Roy Schwartz, and Noah A. Smith. 2019. Show your work: Improved reporting of experimental results. In *Proceedings of the 2019 Conference on Empirical*

Keywords: southern , fresh , made , friendly	
Reference: absolutely , the best pizza in southern nevada ! the pizza is always fresh , made fresh as ordered . the wait staff is very friendly and effecient !	
CBART: great southern food , fresh and made with friendly staff .	
AutoTemplate: this is the best southern food i have ever had. everything is fresh and made to order. the staff is very friendly and helpful. i will definitely be back.	

Table 4: Example generations for the keywords-to-sentence generation on Yelp.

Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP), pages 2185–2194, Hong Kong, China. Association for Computational Linguistics.

Junxian He, Wojciech Kryscinski, Bryan McCann, Nazneen Rajani, and Caiming Xiong. 2022. CTRLsum: Towards generic controllable text summarization. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 5879–5915, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Xingwei He. 2021. Parallel refinements for lexically constrained text generation with BART. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 8653–8666, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Mike Lewis, Yinhan Liu, Naman Goyal, Marjan Ghazvininejad, Abdelrahman Mohamed, Omer Levy, Veselin Stoyanov, and Luke Zettlemoyer. 2020. BART: Denoising sequence-to-sequence pre-training for natural language generation, translation, and comprehension. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 7871–7880, Online. Association for Computational Linguistics.

Ilya Loshchilov and Frank Hutter. 2019. Decoupled weight decay regularization. In *International Conference on Learning Representations, ICLR*.

Long Ouyang, Jeff Wu, Xu Jiang, Diogo Almeida, Carroll L. Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, John Schulman, Jacob Hilton, Fraser Kelton, Luke Miller, Maddie Simens, Amanda Askell, Peter Welinder, Paul Christiano, Jan Leike, and Ryan Lowe. 2022. Training language models to follow instructions with human feedback. *Preprint*, arXiv:2203.02155.

Christian Szegedy, Vincent Vanhoucke, Sergey Ioffe, Jon Shlens, and Zbigniew Wojna. 2016. Rethinking the inception architecture for computer vision. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 2818–2826.

Computing infrastructure	NVIDIA A100
Training duration	4h
Search strategy	Manual tuning
Model implementation	`[MASK]`
Model checkpoint	`[MASK]`

Hyperparameter	**Search space**	**Best assignment**
# of training steps	50,000	50,000
validation interval	5,000	5,000
batch size	32	32

initial checkpoint for small models
initial checkpoint for base models
initial checkpoint for large models
	`google/t5-v1_1-small`
	`google/t5-v1_1-base`
	`google/t5-v1_1-large`
		`google/t5-v1_1-small`
		`google/t5-v1_1-base`
		`google/t5-v1_1-large`

label-smoothing (Szegedy et al., 2016)	*choice*[0.0, 0.1]	0.1
learning rate scheduler	linear schedule with warmup	linear schedule with warmup
warmup steps	5,000	5,000
learning rate optimizer	AdamW (Loshchilov and Hutter, 2019)	AdamW (Loshchilov and Hutter, 2019)
AdamW β_1	0.9	0.9
AdamW β_2	0.999	0.999
learning rate	5e-5	5e-5
weight decay	*choice*[0.0, 1e-3, 1e-2]	1e-2
max grad norm	0.1	0.1
beam width for keywords-to-sentence beam width for entity-guided summarization on CNNDM beam width for entity-guided summarization on XSUM	4 8 6	4 8 6

Table 5: AutoTemplate search space and the best assignments.

# of keywords = 1	One-Billion-Word						Yelp					
	B2	**B4**	**N2**	**N4**	**M**	**SR**	**B2**	**B4**	**N2**	**N4**	**M**	**SR**
CBART (He, 2021)	3.81	0.61	0.34	0.34	6.77	100.	5.71	1.66	0.31	0.32	8.33	100.
InstructGPT (Ouyang et al., 2022)	2.49	0.32	0.24	0.24	5.93	98.4	2.39	0.31	0.18	0.18	6.34	98.5
AutoTemplate												
w/ T5-small	5.56	0.88	1.23	1.23	9.04	100.	9.80	2.46	1.65	1.68	10.84	100.
w/ T5-base	6.01	1.01	1.36	1.36	8.82	100.	9.95	2.52	1.68	1.68	10.94	100.
w/ T5-large	**6.19**	**1.16**	**1.40**	**1.40**	**8.74**	100.	**9.78**	**2.44**	**1.67**	**1.69**	**10.99**	100.
# of keywords = 2	**B2**	**B4**	**N2**	**N4**	**M**	**SR**	**B2**	**B4**	**N2**	**N4**	**M**	**SR**
CBART (He, 2021)	7.25	1.91	0.68	0.68	10.02	100.	9.67	3.14	0.74	0.76	11.75	100.
InstructGPT (Ouyang et al., 2022)	4.57	0.84	0.48	0.49	8.68	95.2	3.94	0.66	0.30	0.30	8.89	95.0
AutoTemplate												
w/ T5-small	8.23	1.77	1.72	1.73	11.49	100.	13.46	3.94	2.14	2.18	13.09	100.
w/ T5-base	9.76	2.52	2.00	2.02	11.39	100.	13.71	4.16	2.18	2.22	13.36	100.
w/ T5-large	**10.06**	**2.59**	**2.05**	**2.06**	**11.35**	100.	**13.55**	**4.04**	**2.17**	**2.21**	**13.25**	100.
# of keywords = 3	**B2**	**B4**	**N2**	**N4**	**M**	**SR**	**B2**	**B4**	**N2**	**N4**	**M**	**SR**
CBART (He, 2021)	11.68	3.84	1.26	1.27	13.30	100.	16.03	6.48	1.73	1.77	15.75	100.
InstructGPT (Ouyang et al., 2022)	7.58	1.58	0.97	0.97	11.52	92.5	6.67	1.30	0.66	0.67	11.95	92.2
AutoTemplate												
w/ T5-small	13.20	3.73	2.60	2.62	13.76	100.	19.17	7.09	2.99	3.07	15.66	100.
w/ T5-base	15.26	5.13	2.85	2.88	14.08	100.	19.82	7.81	3.05	3.15	16.20	100.
w/ T5-large	**16.05**	**5.53**	**3.00**	**3.03**	**14.26**	100.	**20.20**	**8.11**	**3.09**	**3.19**	**16.01**	100.
# of keywords = 4	**B2**	**B4**	**N2**	**N4**	**M**	**SR**	**B2**	**B4**	**N2**	**N4**	**M**	**SR**
CBART (He, 2021)	17.67	7.07	2.31	2.34	16.92	100.	22.45	10.28	3.00	3.10	19.39	100.
InstructGPT (Ouyang et al., 2022)	11.29	3.09	1.81	1.82	14.52	91.6	10.35	2.68	1.46	1.48	15.19	90.1
AutoTemplate												
w/ T5-small	19.04	6.54	3.76	3.81	16.51	100.	25.84	10.77	3.96	4.10	18.30	100.
w/ T5-base	20.92	8.05	3.97	4.02	17.19	100.	26.87	12.26	4.02	4.21	19.03	100.
w/ T5-large	**21.23**	**8.58**	**4.01**	**4.08**	**17.29**	100.	**28.04**	**12.95**	**4.20**	**4.36**	**19.25**	100.
# of keywords = 5	**B2**	**B4**	**N2**	**N4**	**M**	**SR**	**B2**	**B4**	**N2**	**N4**	**M**	**SR**
CBART (He, 2021)	23.51	10.78	3.50	3.56	20.36	100.	27.97	13.80	4.12	4.28	22.73	100.
InstructGPT (Ouyang et al., 2022)	15.32	4.46	2.86	2.88	17.43	89.9	13.97	3.92	2.41	2.44	18.05	90.9
AutoTemplate												
w/ T5-small	23.47	9.76	4.33	4.40	19.58	100.	30.43	13.87	4.78	4.97	20.92	100.
w/ T5-base	25.97	12.03	4.68	4.78	20.44	100.	32.85	16.40	4.94	5.16	22.01	100.
w/ T5-large	**26.89**	**12.74**	**4.79**	**4.89**	**20.93**	100.	**33.11**	**16.71**	**5.05**	**5.28**	**22.18**	100.
# of keywords = 6	**B2**	**B4**	**N2**	**N4**	**M**	**SR**	**B2**	**B4**	**N2**	**N4**	**M**	**SR**
CBART (He, 2021)	29.93	15.38	4.83	4.93	23.72	100.	34.50	18.56	5.35	5.59	26.33	100.
InstructGPT (Ouyang et al., 2022)	19.50	6.71	3.93	3.97	20.20	86.4	18.33	5.76	3.50	3.55	21.01	86.3
AutoTemplate												
w/ T5-small	28.69	13.79	5.00	5.10	22.87	100.	36.31	18.99	5.53	5.80	24.03	100.
w/ T5-base	31.98	17.08	5.50	5.63	23.97	100.	38.85	21.73	5.80	6.10	25.36	100.
w/ T5-large	**33.20**	**18.18**	**5.66**	**5.80**	**24.42**	100.	**39.63**	**22.60**	**5.92**	**6.24**	**25.69**	100.

Table 6: Comprehensive results of keywords-to-sentence generation on the One-Billion-Word and Yelp datasets. **Bold-faced** and underlined denote the best and second-best scores respectively. Baseline results are copied from He (2021). B2/4 denotes BLEU-2/4, N2/4 denotes NIST-2/4, M denotes METEOR-v1.5, and SR denotes the success rate of lexical constraint satisfaction.

Constrained Entities:	{ Amir Khan , Manny Pacquiao , Abu Dhabi , UAE , Khan , Floyd Mayweather Jr , Las Vegas , PacMan , Bob Arum , UAE , Khan , Muslim , Brit , the Money Man , PacMan , Khan , Chris Algieri , New York }
BART (Lewis et al., 2020):	Amir Khan has been linked with a fight with Manny Pacquiao . The fight could take place in Abu Dhabi in November or December. Khan is preparing to fight Chris Algieri in New York next month. Pacquiao is preparing to face Floyd Mayweather on May 2 in Las Vegas .
CTRLSum (He et al., 2022):	Amir Khan could face Manny Pacquiao in Abu Dhabi , UAE . Khan has been linked with a fight with Floyd Mayweather Jr in Las Vegas . The PacMan 's promoter Bob Arum is keen for a fight in the UAE .
AutoTemplate:	Amir Khan could face Manny Pacquiao in Abu Dhabi , UAE . Khan is preparing to face Floyd Mayweather Jr in Las Vegas on May 2. PacMan 's vintage promoter Bob Arum has to hand a treasure trove of an offer for a fight in the UAE this November or December. Khan is a hero of the Muslim world, the Brit would be a huge attraction there. Assuming that the Money Man wins his interim bout with PacMan next month, all that would appear to stand between him and his long-awaited mega-fight is the outside chance of a re-match. Khan is set to fight Chris Algieri in New York next month.
Reference:	Amir Khan could be set to face Manny Pacquiao in Abu Dhabi , UAE . Khan 's hopes of taking on Floyd Mayweather Jr in Las Vegas have faded. PacMan 's promoter Bob Arum has a mega offer for a UAE fight late in 2015. Khan is a hero of the Muslim world and his lure in the Middle East is clear. The Brit will be ringside when the Money Man fights the PacMan on May 2. Khan must first win interim bout with Chris Algieri in New York on May 29.

Table 7: Qualitative comparisons between CTRLSum and AutoTemplate. Constraint entities are extracted from the reference summary (oracle entities). Underlined entities are missed by the CTRLSum (He et al., 2022) while AutoTemplate can incorporate them into the generated summary.