A Supplementary Material

A.1 Experimental Setup for MS MARCO

Model configurations. We trained our model on a machine with eight NVIDIA P100 GPUs. Our best model was jointly trained with the two answer styles in the ALL set for a total of eight epochs with a batch size of 80, where each batch consisted of multi-style answers that were randomly sampled. The training took roughly six days. The hidden size d was 304, and the number of attention heads was 8. The inner state size of the feed-forward networks was 256. The numbers of shared encoding blocks, modeling blocks for a question, modeling blocks for passages, and decoder blocks were 3, 2, 5, and 8, respectively. We used the pre-trained uncased 300-dimensional GloVe (Pennington et al., 2014)[1] and the original 512-dimensional ELMo (Peters et al., 2018)[2]. We used the spaCy tokenizer, and all input words were lowercased except the input for ELMo. The output words were also lowercase. The number of common words in V_{ext} in the extended vocabulary was 5,000. Each passage and each answer were truncated to 100 words for training.

Optimizer. We used Adam (Kingma and Ba, 2015) with $\beta_1 = 0.9$, $\beta_2 = 0.999$, and $\epsilon = 10^{-8}$. The weights were initialized using $N(0, 0.02)$, except that the biases of all the linear transformations were initialized with zero vectors. The learning rate was increased linearly from zero to 2.5×10^{-4} in the first 2,000 steps and then annealed to 0 by using a cosine schedule. All parameter gradients were clipped to a maximum norm of 1. An exponential moving average was applied to all trainable variables with a decay rate of 0.9995. The balancing factors for joint learning, λ_{rank} and λ_{cls}, were set to 0.5 and 0.1, respectively.

Regularization. We used a modified version of the L_2 regularization proposed in (Loshchilov and Hutter, 2017), with $w = 0.01$ on all non-bias. We additionally used a dropout (Srivastava et al., 2014) rate of 0.3 for all highway networks and residual and scaled dot-product attention operations in the multi-head attention mechanism. We also used one-sided label smoothing (Szegedy et al., 2016) for the passage relevance and answer possibility labels. We smoothed only the positive labels to 0.9.

Ensemble model. The ensemble model consisted of six training runs with identical architectures and hyperparameters but with different weight initializations. The final answer was decided with a weighted majority, where we used the ROUGE-L score for the dev. set as the weight of each model.

Evaluation settings. We used the official evaluation script. The answers were normalized by making words lowercase.

A.2 Experimental Setup for NarrativeQA

Model configurations. Our best model was jointly trained with the NarrativeQA and MS MARCO NLG datasets for a total of seven epochs with a batch size of 64, where each batch consisted of multi-style answers that were randomly sampled. For efficient multi-style learning, each summary in the NarrativeQA dataset was divided into ten passages (size of 130 words) with sentence-level overlaps such that each sentence in the summary was entirely contained in a passage. Each passage from MS MARCO was also truncated to 130 words. The rest of the configuration was the same as in the MS MARCO experiments.

Evaluation settings. An official evaluation script is not provided, so we used the evaluation script created by Bauer et al. (2018)[3]. The answers were normalized by making words lowercase and removing punctuation marks.

A.3 Output Examples Generated by Masque

Tables 1 and 2 list the generated examples for questions from MS MARCO 2.1 and NarrativeQA, respectively. We can see from the examples that our model could control answer styles appropriately for various question and reasoning types. We did find some important errors: style errors, yes/no classification errors, copy errors with respect to numerical values, grammatical errors, and multi-hop reasoning errors.

[1] https://nlp.stanford.edu/projects/glove/

[2] https://allennlp.org/elmo

[3] https://github.com/yicheng-w/CommonSenseMultiHopQA/

(a) Question: why your body would feel like it is shaking
Relevant Passage: Shaking is a symptom in which a person has tremors (shakiness or small back and forth movements) in part or all of his body. Shaking can be due to cold body temperatures, rising fever (such as with infections), neurological problems, medicine effects, drug abuse, etc. ...Read more.
Reference Answer (Q&A): Shaking can be due to cold body temperatures, rising fever (such as with infections), neurological problems, medicine effects, drug abuse, etc.
Prediction (Q&A): because of cold body temperatures , rising fever , neurological problems , medicine effects , drug abuse . ✓
Reference Answers (NLG): Shaking can be due to cold body temperatures, rising fever, neurological problems, medicine effects and drug abuse. / Body would feel like it is shaking due to cold body temperatures, rising fever, neurological problems, medicine effects, drug abuse.
Prediction (NLG): your body would feel like it is shaking because of cold body temperatures , rising fever , neurological problems , medicine effects , drug abuse . ✓

(b) Question: _____ is the name used to refer to the era of legalized segregation in the united states
Relevant Passage: Jim Crow law, in U.S. history, any of the laws that enforced racial segregation in the South between the end of Reconstruction in 1877 and the beginning of the civil rights movement in the 1950s.
Reference Answer (Q&A): Jim Crow
Prediction (Q&A): jim crow ✓
Reference Answer (NLG): Jim Crow is the name used to refer to the era of legalized segregation in the United States.
Prediction (NLG): jim crow is the name used to refer to the era of legalized segregation in the united states . ✓

(c) Question: average height nba player
Relevant Passage: The average height of an NBA player is around 6 feet 7 inches. The tallest NBA player ever was Gheorghe Muresan, who was 7 feet 7 inches tall. In contrast, the shortest NBA player ever was Tyrone Muggsy Bogues, who was 5 feet 3 inches tall.
Reference Answer (Q&A): Around 6 feet 7 inches
Prediction (Q&A): 6 feet 7 inches ✓
Reference Answers (NLG): The average height of NBA players is around 6 feet, 7 inches. / The height of NBA player is around 6 feet 7 inches.
Prediction (NLG): the average height of an national basketball association player is 6 feet 7 inches . ✓

(d) Question: population of henryetta ok
Relevant Passage: With it's population of 5,927 people, Henryetta represented 14.79% of the county's total. In addition, where the city of Henryetta has a population density of 891 people per square mile [344.2 people/km2], Okmulgee County averages 57.5 people per square mile [22.19 people/km2].
Reference Answer (Q&A): 5927
Prediction (Q&A): the population of henryetta , oklahoma is 5,927 . ✓ (content is OK, but style is NG ✗)
Reference Answer (NLG): The population of Henryetta, Oklahoma is 5,927.
Prediction (NLG): the population of henryetta , oklahoma is 5,927 . ✓

(e) Question: does gameplay programmer need math skill
Relevant Passage: A good computer programmer is more of a problem solver and logical thinker than a math buff. And besides, the industry is peppered with many successful computer programmers who do not really know much about mathematics.
Reference Answer (Q&A): No
Prediction (Q&A): yes ✗
Reference Answers (NLG): No, a gameplay programmer doesn't need math skill. / No, gameplay programmer do not need an math skill.
Prediction (NLG): no , gameplay programmer does not need math skill . ✓

(f) Question: how long does a freezer take to cool down
Relevant Passage: Quick Answer. It takes anywhere from three to 24 hours for a refrigerator to reach safe temperatures for storing food, depending on the size and type of unit. When the refrigerator compartment reaches 40 degrees Fahrenheit and the freezer reaches 5 degrees Fahrenheit, it is safe to transfer food items. Keep Learning.
Reference Answer (Q&A): 24 hours
Prediction (Q&A): 4 to 5 hours ✗
Reference Answers (NLG): A freezer takes 24 hours to cool down. / A freezer take to cool down is 24 hours.
Prediction (NLG): a freezer takes 4 to 12 hours to cool down . ✗

Table 1: Output examples generated by Masque from MS MARCO. The model was trained with the Q&A and NLG styles. The relevant passage is one that an annotator selected to compose the reference answer. The model could control answer styles appropriately for (a) natural language, (b) cloze-style, and (c) keywords questions. (d) The answer style was incorrect. (e) The answers were not consistent between the styles. (f) Copying from numerical words worked poorly. There were some grammatical errors in the generative answers, which are underlined.

(a) Question: Where does Mark broadcast his radio station?

Summary: Mark Hunter (Slater), a high school student in a sleepy suburb of Phoenix, Arizona, starts an FM pirate radio station that broadcasts from the basement of his parents' house. Mark is a loner, an outsider, whose only outlet for his teenage angst and aggression is his unauthorized radio station. His pirate station's theme song is "Everybody Knows" by Leonard Cohen and there are glimpses of cassettes by such alternative musicians as The Jesus and Mary Chain, Camper Van Beethoven, Primal Scream, Soundgarden, Ice-T, Bad Brains, Concrete Blonde, Henry Rollins, and The Pixies. By day, Mark is seen as a loner, hardly talking to anyone around him; by night, he expresses his outsider views about what is wrong with American society. When he speaks his mind about what is going on at his school and in the community, more and more of his fellow students tune in to hear his show. (...)

Reference Answers: In his parent's basement. / His parents' basement.

Prediction (NQA): the basement of his parents ' house ✓

Prediction (NLG): mark broadcast his radio station in the basement of his parents ' house . ✓

(b) Question: Fletch is a reporter for what newspaper?

Summary: Los Angeles Times reporter Irwin "Fletch" Fletcher (Chase) is writing an article exposing drug trafficking on the beaches of Los Angeles. Posing as an addict during his investigation, he is approached by Boyd Aviation executive vice president Alan Stanwyk (Matheson) who mistakenly assumes Fletch is a junkie. Stanwyk claims to have bone cancer, with only months left to live, and wishes to avoid the pain and suffering. Stanwyk offers $50,000 for Fletch to come to his mansion in a few days time, kill him, and then escape to Rio de Janeiro, staging the murder to look like a burglary. Fletch, while not completely convinced on the truth of Stanwyk's story, reluctantly agrees to the plan. Along with his colleague Larry (Davis), he begins investigating Stanwyk instead of completing his drug trafficking expos, much to the disapproval of his overbearing editor Frank Walker (Libertini). Disguised as a doctor, Fletch accesses Stanwyk's file at the hospital and learns Stanwyk lied about having cancer. (...)

Reference Answers: Los Angeles Times / Los Angeles

Prediction (NQA): los angeles times ✓

Prediction (NLG): fletch is a reporter for los angeles times . ✓

(c) Question: How long approximately was the voyage from London to Thailand supposed to take?

Summary: (...) The story is set twenty-two years earlier, when Marlow was 20. With two years of experience, most recently as third mate aboard a crack clipper, Marlow receives a billet as second mate on the barque Judea. The skipper is Captain John Beard, a man of about 60. This is Beard's first command. The Judea is an old boat, belonging to a man "Wilmer, Wilcox or something similar", suffering from age and disuse in Shadewell basin. The 400-ton ship is commissioned to take 600 tons of coal from England to Thailand. The trip should take approximately 150 days. The ship leaves London loaded with sand ballast and heads north to the Senn river to pick up the cargo of coal. On her way, the Judea suffers from her ballast shifting aside and the crew go below to put things right again. The trip takes 16 days because of inclement weather, and the battered ship must use a tug boat to get into port. The Judea waits a month on the Tyne to be loaded with coal. The night before she ships out she is hit by a steamer, the Miranda or the Melissa. The damage takes another three weeks to repair. Three months after leaving London, the Judea ships off for Bangkok. The Judea travels through the North Sea and Britain. 300 miles west of the Lizard a winter storm, 'the famous winter gale of twenty-two years ago', hits. (...)

Reference Answers: Approximately 150 days / 150 days

Prediction (NQA): 150 days ✓

Prediction (NLG): the voyage from london to thailand was supposed to take 150 days . ✓

(d) Question: Why does Jamie start avoiding Landon?

Summmary: (...) During these functions, Landon notices Jamie Sullivan, a girl he has known since kindergarten and who has attended many of the same classes as him, and is also the local minister's daughter. Since he's one of the in-crowd, he has seldom paid any attention to Jamie, who wears modest dresses and owns only one sweater. Jamie is labeled an outsider and a geek. She makes no attempt to wear make-up or otherwise improve her looks or attract attention to herself. Landon has trouble learning his lines for the play. Jamie, who is also in the play, agrees to help him on one condition: Jamie warns Landon not to fall in love with her; he laughs it off and dismisses it as a foolish idea. Landon and Jamie begin practicing together at her house after school. They get to know each other and a spark of affection arises between them. On the opening night of the play, Jamie astounds Landon and the entire audience with her beauty and her voice. Onstage at the peak of the ending to the play, Jamie sings. When Jamie finishes, Landon improvises and kisses her which is not a part of the play. Afterwards, Jamie avoids Landon, and it is not until Landon's friends play a cruel prank on Jamie and he protects her in opposition to his friends that she warms up to him again. Landon asks Jamie on a date soon after, but Jamie says her father doesn't allow her to date. (...)

Reference Answers: Because he kissed her in the play. / He kisses her

Prediction (NQA): he is not a part of the play ✗

Prediction (NLG): he is not a part of the play ✗

Table 2: Output examples generated by Masque from NarrativeQA. The model was trained with the NarrativeQA (NQA) and MS MARCO (NLG) styles. It could control answer styles appropriately for questions that required (a,b) single-sentence reasoning and (c) multi-sentence reasoning. (d) Example of an error in multi-sentence reasoning. There were some grammatical errors in the generative answers, which are underlined.

References

Lisa Bauer, Yicheng Wang, and Mohit Bansal. 2018. Commonsense for generative multi-hop question answering tasks. In *Empirical Methods in Natural Language Processing (EMNLP)*, pages 4220–4230.

Diederik P. Kingma and Jimmy Ba. 2015. Adam: A method for stochastic optimization. In *International Conference on Learning Representations (ICLR)*.

Ilya Loshchilov and Frank Hutter. 2017. Fixing weight decay regularization in adam. *Computing Research Repository (CoRR)*, arXiv:1711.05101. Version 1.

Jeffrey Pennington, Richard Socher, and Christopher D. Manning. 2014. Glove: Global vectors for word representation. In *Empirical Methods in Natural Language Processing (EMNLP)*, pages 1532–1543.

Matthew E. Peters, Mark Neumann, Mohit Iyyer, Matt Gardner, Christopher Clark, Kenton Lee, and Luke Zettlemoyer. 2018. Deep contextualized word representations. In *North American Chapter of the Association for Computational Linguistics: Human Language Technologies (NAACL-HLT)*, pages 2227–2237.

Nitish Srivastava, Geoffrey E. Hinton, Alex Krizhevsky, Ilya Sutskever, and Ruslan Salakhutdinov. 2014. Dropout: a simple way to prevent neural networks from overfitting. *Journal of Machine Learning Research*, 15(1):1929–1958.

Christian Szegedy, Vincent Vanhoucke, Sergey Ioffe, Jonathon Shlens, and Zbigniew Wojna. 2016. Rethinking the inception architecture for computer vision. In *Computer Vision and Pattern Recognition (CVPR)*, pages 2818–2826.