Evaluating Rewards for Question Generation Models

Tom Hosking
University College London
thomas.hosking.17@ucl.ac.uk

Sebastian Riedel
University College London
sriedel@ucl.ac.uk

Abstract

Recent approaches to question generation have used modifications to a Seq2Seq architecture inspired by advances in machine translation. Models are trained using teacher forcing to optimise only the one-step-ahead prediction. However, at test time, the model is asked to generate a whole sequence, causing errors to propagate through the generation process (exposure bias). A number of authors have suggested that optimising for rewards less tightly coupled to the training data might counter this mismatch. We therefore optimise directly for various objectives beyond simply replicating the ground truth questions, including a novel approach using an adversarial discriminator that seeks to generate questions that are indistinguishable from real examples. We confirm that training with policy gradient methods leads to increases in the metrics used as rewards. We perform a human evaluation, and show that although these metrics have previously been assumed to be good proxies for question quality, they are poorly aligned with human judgement and the model simply learns to exploit the weaknesses of the reward source.

1 Introduction

Posing questions about a document in natural language is a crucial aspect of the effort to automatically process natural language data, enabling machines to ask clarification questions (Saeidi et al., 2018), become more robust to queries (Yu et al., 2018), and to act as automatic tutors (Heilman and Smith, 2010).

Recent approaches to question generation have used Seq2Seq (Sutskever et al., 2014) models with attention (Bahdanau et al., 2014) and a form of copy mechanism (Vinyals et al., 2015; Gulcehre et al., 2016). Such models are trained to generate a plausible question, conditioned on an input document and answer span within that document (Zhou et al., 2018; Du et al., 2017; Du and Cardie, 2018; Yuan et al., 2017).

There are currently no dedicated question generation datasets, and authors have used the context-question-answer triples available in SQuAD (Rajpurkar et al., 2016). Only a single question is available for each context-answer pair, and models are trained using teacher forcing (Williams and Zipser, 1989). This lack of diverse training data combined with the one-step-ahead training procedure exacerbates the problem of exposure bias (Ranzato et al., 2015). The model does not learn how to distribute probability mass over sequences that are valid but different to the ground truth; during inference, the model must predict the whole sequence, and may not be robust to mistakes during decoding.

Recent work has investigated training the models directly on a performance based objective, either by optimising for BLEU score (Kumar et al., 2018a) or other quality metrics (Yuan et al., 2017). By decoupling the training procedure from the ground truth data, the model is able to explore the space of possible questions and learn to recover from suboptimal predictions during decoding. While the metrics used seem to be intuitively good choices, there is an assumption that they are good proxies for question quality which has not yet been confirmed.

Our contributions are as follows. We perform fine tuning using a range of rewards, including a novel adversarial objective that directly estimates the probability that a question was generated or came from the ground truth data. We show that although fine tuning leads to increases in reward scores, the resulting models perform worse when evaluated by human workers. We also demonstrate that the generated questions exploit weaknesses in the reward models.

Context
although united methodist practices and interpretation of beliefs have evolved over time , these practices and beliefs can be traced to the writings of the church 's founders , especially **john wesley and charles wesley** (anglicans) , but also philip william otterbein and martin boehm (united brethren) , and jacob albright (evangelical association) .

Rewards	**Output**
Ground Truth Question	who were two of the founders of the united methodist church ?
No fine tuning	which two methodist can be traced to the church 's founders ?
LM	according to the writings of the church 's founders , according to the writings of the church 's founders , [...]
QA	who in anglicans ?
LM and QA	who are the writings of the church 's founders ?
Discriminator	who founded the church 's founders ?
Adversarial discriminator	who were two western methodist practices ?
LM, QA and adversarial discriminator	who are the anglicans of the church ?

Table 1: Example generated questions for various fine-tuning objectives. The answer is highlighted in bold. The model trained on a QA reward has learned to simply point at the answer and exploit the QA model, while the model trained on a language model objective has learned to repeat common phrase templates.

2 Background

Many of the advances in natural language generation have been led by machine translation (MT) (Sutskever et al., 2014; Bahdanau et al., 2014; Gulcehre et al., 2016).

Previous work on question generation has made extensive use of MT techniques. Du et al. (2017) use a Seq2Seq based model to generate questions conditioned on context-answer pairs, and build on this work by preprocessing the context to resolve coreferences and adding a pointer network (Du and Cardie, 2018). Similarly, Zhou et al. (2018) use a part-of-speech tagger to augment the embedding vectors. Both authors perform a human evaluation of their models, and show significant improvement over their baseline. Kumar et al. (2018a) use a similar model, but apply it to the task of generating questions without conditioning on a specific answer span. Song et al. (2018) use a modified context encoder based on multi-perspective context matching (Wang et al., 2016).

Kumar et al. (2018b) propose a framework for fine tuning using policy gradients and perform a human evaluation showing promising results. However, they use as rewards various similarity metrics that are still coupled to the ground truth. Yuan et al. (2017) describe a Seq2Seq model with attention and a pointer network, with an additional encoding layer for the answer. They also describe a method for further tuning their model using policy gradients, with rewards given by an external language model and question answering (QA) system. Unfortunately they do not perform any human evaluation to determine whether this tuning led to improved question quality.

For the related task of summarisation, Paulus et al. (2017) propose a framework for fine tuning a summarisation model using reinforcement learning, with the ROUGE similarity metric used as the reward.

3 Experimental setup

The task is to generate a natural language question, conditioned on a document and the location of an answer within that document. For example, given the input document “this paper investigates rewards for question generation” and answer “question generation”, the model should produce a question such as “what is investigated in the paper?”

3.1 Model description

We use the model architecture described by Yuan et al. (2017). Briefly, this is a Seq2Seq model (Sutskever et al., 2014) with attention (Bahdanau et al., 2014) and copy mechanism (Vinyals et al., 2015; Gulcehre et al., 2016). Yuan et al. (2017) also add an additional answer encoder layer, and initialise the decoder with a hidden state constructed from the final state of the encoder. Beam search (Graves, 2012) is used to sample from the model at inference time. We train the model using maximum likelihood before fine tuning. Our implementation achieves a BLEU-4 score (Papineni et al., 2002) of 13.5 on the test set used by Du et al. (2017), before fine tuning.

Features				Metrics				
QA reward	LM reward	Discriminator reward	Adversarial discriminator	NLL	BLEU	QA	LM	Discriminator
-	✓	-	-	-0.7	-1.9	-3.7	-13.4	+1.5
✓	-	-	-	+1.7	-4.5	**+3.9**	+226	+5.4
✓	✓	-	-	-0.5	-2.6	+2.0	**-16.3**	+2.9
-	-	✓	-	**-0.8**	-1.8	-2.1	-9.4	+2.5
-	-	✓	✓	+6.4	-2.7	-2.5	-1.0	**+10.8**
✓	✓	✓	✓	+1.0	-2.4	+1.3	-6.2	+10.0

Table 2: Changes in automatic evaluation metrics after models were fine tuned on various objectives. QA refers to the F1 score obtained by a question answering system on the generated questions. LM refers to the perplexity of generated questions under a separate language model. The discriminator reward refers to the percentage of generated sequences that fooled the discriminator. Lower LM and NLL scores are better. BLEU scores decreased in all cases.

Model	Fluency	Relevance
No fine tuning	**3.34**	**3.12**
+QA, LM rewards	3.05	2.75
+QA, LM, discriminator rewards +Adversarial discriminator	2.89	2.82
Ground Truth	4.67	4.72

Table 3: Summary of human evaluation of selected models

3.2 Fine tuning

Generated questions should be formed of language that is both *fluent* and *relevant* to the context and answer. Following (Yuan et al., 2017), we perform fine tuning on a trained model, using rewards given either by the negative perplexity under a LSTM language model, or the F1 score attained by a question answering (QA) system, or a weighted combination of both. The language model is a standard recurrent neural network formed of a single LSTM layer. For the QA system, we use QANet (Yu et al., 2018) as implemented by Kim (2018).

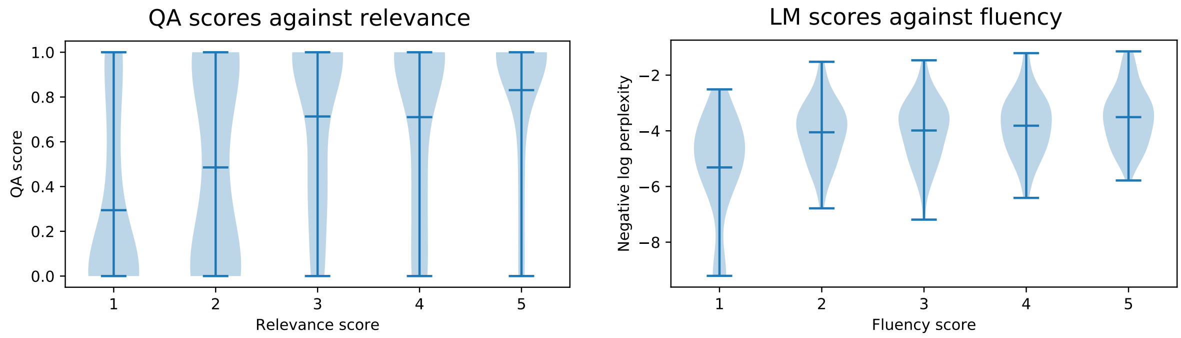
3.3 Adversarial training

Additionally, we propose a novel approach by learning the reward directly from the training data, using a *discriminator* detailed in Appendix A. We generate questions for each context-answer pair in the training set using a generator trained by maximum likelihood, and train the discriminator to predict whether an input question was generated by our model, or originated from the training data. Keeping the discriminator fixed, we then fine-tune the generator, using as reward the probability estimated by the discriminator that a generated question was in fact real. In other words, the generator is rewarded for successfully fooling the discriminator. We also experiment with interleaving updates to the discriminator within the fine tuning phase, allowing the discriminator to become adversarial and adapt alongside the generator.

The rewards described above are used to update the model parameters via the REINFORCE policy gradient algorithm (Williams, 1992). We teacher force the decoder with the generated sequence to reproduce the activations calculated during beam search, to enable backpropagation. All rewards are normalised with a simple form of PopArt (Hasselt et al., 2016), with the running mean μ_R and standard deviation σ_R updated online during training. We continue to apply a maximum likelihood training objective during this fine tuning.

3.4 Evaluation

We report the negative log-likelihood (NLL) of the test set under the different models, as well as the corpus level BLEU-4 score (Papineni et al., 2002) of the generated questions compared to the ground truth. We also report the rewards achieved on the



(a) QA scores plotted against human relevance scores for all rated questions.

(b) LM scores plotted against human fluency scores for all rated questions.

Figure 1: Comparison of human and automatic metrics.

test set, as the QA, LM and discriminator scores.

For the human evaluation, we follow the standard approach in evaluating machine translation systems (Koehn and Monz, 2006), as used for question generation by Du and Cardie (2018). We ask three workers to rate 300 generated questions between 1 (poor) and 5 (good) on two separate criteria: the fluency of the language used, and the relevance of the question to the context document and answer.

4 Results

Table 2 shows the changes in automatic metrics for models fine tuned on various combinations of rewards, compared to the model without tuning. In all cases, the BLEU score reduces, as the training objective is no longer closely coupled to the training data. In general, models achieve better scores on the metrics on which they were fine tuned. Jointly training on a QA *and* LM reward results in better LM scores than training on only a LM reward; the LM score did not increase smoothly when used as the sole objective, and we believe the additional QA reward acts as a form of regularisation. We conclude that fine tuning using policy gradients can be used to attain higher rewards, as expected.

Table 3 shows the human evaluation scores for a subset of the fine tuned models. The model fine tuned on a QA and LM objective is rated as significantly worse by human annotators, despite achieving higher scores in the automatic metrics. In other words, the training objective given by these reward sources does not correspond to true question quality, despite them being intuitively good choices.

The model fine tuned using an adversarial discriminator has also failed to achieve better human ratings, with the discriminator model unable to learn a useful reward source. Although the training process was stable and robust to different initialisations, and the outputs do not appear to be significantly worse, we conclude that the discriminator was unable to learn a sufficiently useful distinction between generated and real questions, and the additional fine tuning procedure simply added unwanted noise to the model predictions.

Table 1 shows an example where fine tuning has not only failed to improve the quality of generated questions, but has caused the model to exploit the reward source. The model fine tuned on a LM reward has degenerated into producing a loop of words that is evidently deemed probable, while the model trained on a QA reward has learned that it can simply point at the location of the answer. This observation is supported by the metrics; the model fine tuned on a QA reward has suffered a catastrophic worsening in LM score of +226.

Figure 1 shows the automatic scores against human ratings for all rated questions. The correlation coefficient between human relevance and automatic QA scores was 0.439, and between fluency and LM score was only 0.355. While the automatic scores are good indicators of whether a question will achieve the lowest human rating or not, they do not differentiate clearly between the higher ratings: training a model on these objectives will not necessarily learn to generate better questions. A good question will likely attain a high QA and LM score, but the inverse is not true; a sequence may exploit the weaknesses of the metrics and achieve a high score *despite* being unintelligible to a human. We conclude that fine tuning a question generation model on these rewards does not lead to better quality questions.

5 Conclusion

In this paper, we investigate the use of external reward sources for fine tuning question generation models to counteract the lack of task-specific training data. We show that although fine tuning can be used to attain higher rewards, this does not equate to better quality questions when rated by humans. Using QA and LM rewards as a training objective causes the generator to expose the weaknesses in these models, which in turn suggests a possible use of this approach for generating adversarial training examples for QA models. The QA and LM scores are well correlated with human ratings at the lower end of the scale, suggesting they could successfully be used as part of a reranking or filtering system. We plan to research overgenerating questions and using the reward signals to rerank the outputs, thereby including the inductive bias the rewards represent without allowing the model to exploit them.

References

Dzmitry Bahdanau, Kyunghyun Cho, and Yoshua Bengio. 2014. Neural Machine Translation by Jointly Learning to Align and Translate.

Xinya Du and Claire Cardie. 2018. Harvesting Paragraph-Level Question-Answer Pairs from Wikipedia.

Xinya Du, Junru Shao, and Claire Cardie. 2017. Learning to Ask: Neural Question Generation for Reading Comprehension. pages 1342–1352.

Alex Graves. 2012. Sequence Transduction with Recurrent Neural Networks.

Caglar Gulcehre, Sungjin Ahn, Ramesh Nallapati, Bowen Zhou, and Yoshua Bengio. 2016. Pointing the Unknown Words.

Hado Van Hasselt, Arthur Guez, Matteo Hessel, and David Silver. 2016. Learning functions across many orders of magnitudes. *arXiv*, (Nips):1–19.

Michael Heilman and Noah A Smith. 2010. Good question! Statistical ranking for question generation. *NAACL HLT 2010 - Human Language Technologies: The 2010 Annual Conference of the North American Chapter of the Association for Computational Linguistics, Proceedings of the Main Conference*, pages 609–617.

Min Sang Kim. 2018. Qanet. `https://github.com/NLPLearn/QANet`.

Philipp Koehn and Christof Monz. 2006. Manual and Automatic Evaluation of Machine Translation between European Languages. In *Proceedings of the Workshop on Statistical Machine Translation*, pages 102–121.

Vishwajeet Kumar, Kireeti Boorla, Yogesh Meena, Ganesh Ramakrishnan, and Yuan Fang Li. 2018a. Automating reading comprehension by generating question and answer pairs. In *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, volume 10939 LNAI, pages 335–348.

Vishwajeet Kumar, Ganesh Ramakrishnan, and Yuan-Fang Li. 2018b. A Framework for Automatic Question Generation from Text using Deep Reinforcement Learning. Technical report.

Kishore Papineni, Salim Roukos, Todd Ward, and Wj Zhu. 2002. BLEU: a method for automatic evaluation of machine translation. *. . . of the 40Th Annual Meeting on . . .*, (July):311–318.

Romain Paulus, Caiming Xiong, and Richard Socher. 2017. A DEEP REINFORCED MODEL FOR ABSTRACTIVE SUMMARIZATION. (i):1–12.

Pranav Rajpurkar, Jian Zhang, Konstantin Lopyrev, and Percy Liang. 2016. SQuAD: 100,000+ Questions for Machine Comprehension of Text.

Marc'Aurelio Ranzato, Sumit Chopra, Michael Auli, and Wojciech Zaremba. 2015. Sequence Level Training with Recurrent Neural Networks.

Marzieh Saeidi, Max Bartolo, Patrick Lewis, Sameer Singh, Tim Rocktaschel, Mike Sheldon, Guillaume Bouchard, and Sebastian Riedel. 2018. Interpretation of natural language rules in conversational machine reading. In *Empirical Methods in Natural Language Processing (EMNLP)*.

Linfeng Song, Zhiguo Wang, Wael Hamza, Yue Zhang, and Daniel Gildea. 2018. Leveraging Context Information for Natural Question Generation. pages 569–574.

Ilya Sutskever, Oriol Vinyals, and Quoc V. Le. 2014. Sequence to Sequence Learning with Neural Networks. *Nips*, page 9.

Oriol Vinyals, Meire Fortunato, and Navdeep Jaitly. 2015. Pointer Networks. *Proceedings of the 28th International Conference on Neural Information Processing Systems - Volume 2*, pages 2692–2700.

Zhiguo Wang, Haitao Mi, Wael Hamza, and Radu Florian. 2016. Multi-Perspective Context Matching for Machine Comprehension.

Ronald J. Williams. 1992. Simple Statistical Gradient-Following Algorithms for Connectionist Reinforcement Learning. *Machine Learning*, 8(3):229–256.

Ronald J. Williams and David Zipser. 1989. A Learning Algorithm for Continually Running Fully Recurrent Neural Networks. *Neural Computation*, 1(2):270–280.

Adams Wei Yu, David Dohan, Minh-Thang Luong, Rui Zhao, Kai Chen, Mohammad Norouzi, and Quoc V. Le. 2018. QANet: Combining Local Convolution with Global Self-Attention for Reading Comprehension.

Xingdi Yuan, Tong Wang, Caglar Gulcehre, Alessandro Sordoni, Philip Bachman, Sandeep Subramanian, Saizheng Zhang, and Adam Trischler. 2017. Machine Comprehension by Text-to-Text Neural Question Generation.

Qingyu Zhou, Nan Yang, Furu Wei, Chuanqi Tan, Hangbo Bao, and Ming Zhou. 2018. Neural question generation from text: A preliminary study. In *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, volume 10619 LNAI, pages 662–671. Springer, Cham.

A Discriminator architecture

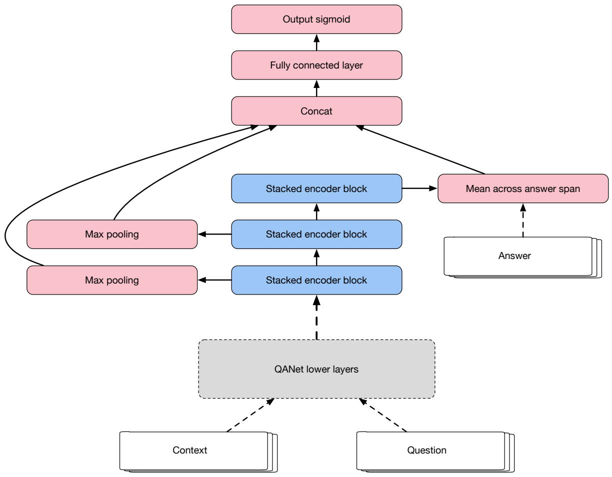


Figure 2: Discriminator architecture diagram.

We use an architecture based on a modified QANet as shown in Figure 2, replacing the output layers of the model to produce a single probability. Since the discriminator is also able to consider a full context-question-answer triple as input (as opposed to a context-question pair for the QA task), we fuse this information in the output layers.

Specifically, we apply max pooling over time to the output of the first two encoders, and we took the mean of the outputs of the third encoder that formed part of the answer span. These three reduced encodings were concatenated, a 64 unit hidden layer with ReLU activation applied, and the output passed through a single unit sigmoid output layer to give the estimated probability that an input context-question-answer triple originated from the ground truth dataset or was generated.