On the Importance of Diversity in Question Generation for QA

Md Arafat Sultan[†] **Shubham Chandel**[‡] **Ramón F. Astudillo**[†] **Vittorio Castelli**[†]
[†]IBM Research AI, T.J. Watson Research Center, New York, USA
[‡]New York University, New York, USA
{arafat.sultan,ramon.astudillo}@ibm.com,
shubhamchandel@nyu.edu, vittorio@us.ibm.com

Abstract

Automatic question generation (QG) has shown promise as a source of synthetic training data for question answering (QA). In this paper we ask: *Is textual diversity in* QG *beneficial for downstream* QA*?* Using top-p *nucleus* sampling to derive samples from a transformer-based question generator, we show that diversity-promoting QG indeed provides better QA training than likelihood maximization approaches such as beam search. We also show that standard QG evaluation metrics such as BLEU, ROUGE and METEOR are inversely correlated with diversity, and propose a diversity-aware intrinsic measure of overall QG quality that correlates well with extrinsic evaluation on QA.

1 Question Generation and Diversity

Besides areas such as dialog (Bordes et al., 2017) and tutoring systems (Lindberg et al., 2013), automatic question generation (QG) has recently been applied with great success to generating synthetic training examples for question answering (QA) (Alberti et al., 2019; Dong et al., 2019). Yet an important question has remained unexplored: *Does increased textual diversity in automatically generated questions lead to better* QA*?*

In Figure 1 we show four questions generated by one of our QG models (details in Section 2) from a SQuAD (Rajpurkar et al., 2016) passage and an answer span (the QG *prompt*). The questions are different not only lexically, but also in what information about the answer entity they draw upon and even their use of world knowledge, e.g., Tesla's reputation as a "mad scientist". Intuitively, such sample diversity, if sufficiently accurate, could provide QA models with rich training signal.

Existing QG work has predominantly relied on customary beam search decoding for generation and n-gram similarity metrics such as BLEU for evaluation (Du et al., 2017; Alberti et al., 2019; Dong et al., 2019; Zhang and Bansal, 2019).[1] Such methods/metrics solely optimize/reward similarity with human-generated reference questions treated as the ground truth (GT). However, in many open-ended generation tasks where only one or a few of many possible GTs are available through human annotation, this approach directly penalizes diversity by discouraging deviation from the GT(s).

On Tesla's 75th birthday in 1931, Time magazine put him on its cover. The cover caption "All the world's his power house" noted his contribution to electrical power generation. He received congratulatory letters from more than 70 pioneers in science and engineering, including Albert Einstein.

- *Who appeared on Time magazine's cover on his 75th birthday?*
- *Which famous scientist was in the cover of Time Magazine in 1931?*
- *Which mad scientist received more than a 70 people congratulating him on his birthday?*
- *What famous scientist was also 75?*

Figure 1: A passage with an underlined answer span ("Tesla"), and corresponding questions generated by our model. The generated questions exhibit both lexical and factual diversity.

In recent years, massively pre-trained neural language models (LMs) (Devlin et al., 2019; Radford et al., 2019; Liu et al., 2019) have revolutionized NLP. In open-ended text generation, these models show remarkable robustness under sampling (Radford et al., 2019; Holtzman et al., 2020). This observation, coupled with the examples presented in Figure 1, suggests that treating QG for QA as a more open-ended generation problem and relying on the power of modern text generators to produce diverse yet accurate samples might yield better QA results than the current approach of optimizing for the "most likely" question.

We test this hypothesis by fine-tuning a pre-trained transformer-based masked LM (Liu et al.,

[1] *http://aqleaderboard.tomhosking.co.uk/squad*

Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, pages 5651–5656
July 5 - 10, 2020.

2019) for QG, and sampling questions from it using top-p nucleus sampling (Holtzman et al., 2020). Other diversity-promoting text generation techniques exist—both at training time (e.g., VAEs (Kingma and Welling, 2014)) and during inference (e.g., top-k sampling and diverse beam search (Vijayakumar et al., 2018))—that have been applied to various NLP tasks: language modeling (Bowman et al., 2016), dialog (Cao and Clark, 2017), visual QG (Jain et al., 2017; Fan et al., 2018), image captioning (Vijayakumar et al., 2018) and so on. We choose nucleus sampling because of its effectiveness, simplicity and speed. Our experiments lead to the following discoveries:

– Nucleus sampling indeed produces better QA results than beam search, even when only one question is generated per prompt.
– QG metrics that only reward similarity with GT are negatively correlated with diversity, and as a result, are inaccurate predictors of downstream QA performance of diversity-promoting QG.
– A measure of QG can be devised that combines diversity with similarity to GT, showing strong correlations with QA performance.

2 Question Generation using RoBERTa

We fine-tune a RoBERTa masked LM (Liu et al., 2019) for QG given an answer span within a textual *context* (as shown in Figure 1), and use nucleus sampling (Holtzman et al., 2020) for generation.

Model: Various transformer architectures can be used for text generation (Raffel et al., 2019). Following (Dong et al., 2019; Alberti et al., 2019), we fine-tune a pre-trained masked LM as a prefix LM (Raffel et al., 2019) to predict a question token q_t given (1) a prompt $\boldsymbol{p}_{1:N}$: a tokenized textual context with special tokens delimiting an answer span, and (2) question tokens $\boldsymbol{q}_{1:t-1}$, if any, that have already been generated for the given prompt in a left-to-right order. A special separator token separates the question prefix from the prompt. The prompt is encoded using bidirectional attention and question tokens using causal (left-only) attention. We choose RoBERTa as our pre-trained model because of its extended pre-training on large amounts of text (Liu et al., 2019). Our implementation of the QG model is based on Hugging Face's (Wolf et al., 2019) PyTorch implementation of RoBERTa.

Fine-Tuning: For each QG training example, the model is asked to predict a single question token q_t given the prompt $\boldsymbol{p}_{1:N}$, the previous question tokens $\boldsymbol{q}_{1:t-1}$ (teacher-forced), and the mask m at timestep t. All questions end with an EOS token that marks the end of generation. Training attempts to minimize the masked LM loss, i.e., the negative log-likelihood of the GT token q_t as the prediction for m in position t:

$$loss_t = -\log P(q_t \mid \boldsymbol{p}_{1:N}, \boldsymbol{q}_{1:t-1}, m)$$

Inference: During generation, the fine-tuned RoBERTa QG model outputs a probability distribution over the entire vocabulary at each question timestep t. Top-p nucleus sampling (NS@p henceforth) samples from the (re-normalized) categorical distribution $P_{\mathcal{N}}$ of the *nucleus* $\mathcal{N}$, which is the smallest subset of vocabulary items that has (1) a cumulative probability mass greater than p, and (2) the highest probability among all such subsets:

$$\hat{q}_t \sim P_{\mathcal{N}}(q_t \mid \boldsymbol{p}_{1:N}, \boldsymbol{q}_{1:t-1}, m)$$

By restricting the pool to a high-likelihood region of the vocabulary, compared to top-k sampling, NS reduces the chances of generating low-probability items when the original distribution is peaked at one or a few items. Our question generation works by repeated nucleus sampling of question tokens until $\hat{q}_t = \text{EOS}$.

3 Experiments and Results

To test the effect of QG diversity on QA, we generate questions with both nucleus sampling and beam search from a number of different QG models and compare their performance.

General Setup: Considering that performances of different generation methods may vary across models of different capacities, we train eight QG models, each uniquely characterized by: (1) its size (# of parameters), and (2) the amount of training data it was fine-tuned on. The two model sizes are those of RoBERTa: *base* (125M parameters) and *large* (355M parameters). For fine-tuning we use the *train* set of the SQuAD1 split by Du et al. (2017).[2] This is a three-way split of the public portion of SQuAD1 widely adopted in QG literature, with approximately 76k *train*, 18k *dev* and 12k *test* (prompt, question) pairs. We draw varying amounts of samples (ranging from 5% to 100%) at random from the *train* set to fine-tune each model on, simulating different points on the low- to high-resource

[2] *https://github.com/xinyadu/nqg/blob/master/data/raw/*

%train	generator	B_1	R_4	MT	QA F_1	B_1	R_4	MT	QA F_1
	$b = 5$	**33.9**	**7.9**	**39.1**	81.1	**35.9**	**8.5**	**40.7**	**83.2**
	$p = .1$	32.3	6.2	36.8	80.6	34.1	7.1	38.8	82.7
5	$p = .5$	32.0	6.1	36.4	81.0	33.8	7.0	38.3	82.8
	$p = .75$	30.1	5.1	34.1	81.3	32.3	6.2	36.5	83.1
	$p = .95$	26.5	3.9	29.7	**81.6**	28.7	4.6	31.9	83.1
	$b = 5$	**37.2**	**10.5**	**42.2**	82.1	**38.7**	**11.2**	**43.3**	83.9
	$p = .1$	35.9	9.0	40.9	82.8	37.6	9.8	42.3	84.3
20	$p = .5$	35.5	8.7	40.4	83.0	37.4	9.7	42.1	84.5
	$p = .75$	33.8	7.7	38.1	83.7	35.8	8.7	40.0	84.9
	$p = .95$	30.0	5.6	33.4	**83.9**	31.6	6.4	35.2	**85.3**
	$b = 5$	**39.1**	**11.9**	**44.4**	82.8	**40.6**	**12.6**	**45.4**	84.3
	$p = .1$	37.8	10.3	43.4	83.6	39.6	11.2	44.7	84.8
50	$p = .5$	37.4	10.0	42.9	83.8	39.4	11.1	44.4	84.9
	$p = .75$	35.4	8.8	40.2	84.3	38.2	10.3	42.8	85.3
	$p = .95$	31.4	6.3	35.2	**84.8**	33.6	7.5	37.2	**85.7**
	$b = 5$	**40.3**	**12.6**	**45.8**	83.6	**41.6**	**13.4**	**46.7**	84.5
	$p = .1$	38.9	11.0	44.6	83.9	40.6	12.1	46.1	84.9
100	$p = .5$	38.5	10.7	44.1	84.3	40.3	11.9	45.7	85.0
	$p = .75$	36.7	9.6	41.7	84.8	38.8	10.8	43.7	85.5
	$p = .95$	32.5	6.9	36.4	**85.3**	34.4	7.6	38.3	**86.1**
		base model				*large* model			

Table 1: Performance of beam search (BEAM) ($b = 5$) and nucleus sampling (NS@p; $p \in \{.1, .5, .75, .95\}$) on the SQuAD-Du dataset. (**Bold**: best, underlined: worst). NS yields stronger QA results than BEAM but lower BLEU, ROUGE and METEOR scores. Moreover, QA performance of NS improves with the nucleus probability mass p.

spectrum. Each model is trained for two epochs with a learning rate of 2e-5 and a batch size of 96.

In-Domain Experiments: With each QG model, we generate questions for all prompts in the SQuAD1-Du *dev* set. These questions are first evaluated using existing generation metrics: BLEU, ROUGE and METEOR. To extrinsically evaluate on QA, we then (1) fine-tune a BERT (Devlin et al., 2019) whole-word-masked (wwm) LM for QA on the generated *dev* examples from each model, and (2) evaluate on *test*.

For each of the eight QG models, we evaluate beam search (BEAM henceforth) and NS@p for different values of p. Our BEAM experiments with the RoBERTa-base model did not show significant performance differences between beam sizes 5 and 10, therefore we report results only for $b = 5$ in this paper. An important point to note here is that given paragraph-long input prompts in QG for QA, where large numbers of synthetic examples may also be needed in many practical use cases, large beam sizes can become prohibitively expensive from a computational standpoint for transformer-based generators.

For NS, we evaluate with $p \in \{.1, .5, .75, .95\}$. Among these, $p = .1$ closely approximates greedy decoding, as we observed for all models an average nucleus size of practically 1 in this setup. We also set the maximum number of vocabulary items in a nucleus to 20, which even the largest p values rarely reached in our experiments.

Table 1 shows performances (mean over five different seeds) of all generators in BLEU-1 (B_1), ROUGE-4 (R_4) and METEOR (MT), the variant in each metric family that showed the highest correlation with downstream QA performance. We also show QA performances measured by SQuAD's official F_1 score metric, which computes the degree of lexical overlap between the predicted and the target answer. As expected, model performance improves with both model size and # of training instances, both in intrinsic evaluation and on QA. Importantly, however, while BEAM has the best intrinsic evaluation results for all eight models, it is competitive in QA only in the lowest-resource setup (5% training data). On the other hand, NS@.95 has the lowest QG but the highest QA scores, especially when sufficient training data is available (20% or more). Note that in these experiments we generate a single question per prompt; yet generation diversity across different prompts yields higher-quality QA training data for NS, which is also a faster alternative to BEAM. Sampling five questions per prompt from the *large*-100% model with NS@.95 provides additional improvement (F_1 = 86.4).

Out-of-Domain Experiments: As we increase p to make generation more diverse, the chances of NS@p drawing less likely candidates and thus

model-%train	generator	R_1	QA F_1
base-20	$b = 5$	**34.6**	56.6
	$p = .1$	**34.6**	56.3
	$p = .5$	34.2	57.1
	$p = .75$	32.4	57.5
	$p = .95$	28.9	**58.4**
base-100	$b = 5$	**37.9**	57.5
	$p = .1$	**37.9**	58.4
	$p = .5$	37.6	59.2
	$p = .75$	35.7	60.4
	$p = .95$	31.5	**61.3**
large-20	$b = 5$	**36.3**	60.4
	$p = .1$	**36.3**	59.9
	$p = .5$	36.1	59.7
	$p = .75$	34.7	**60.8**
	$p = .95$	30.9	60.6
large-100	$b = 5$	39.1	60.6
	$p = .1$	**39.2**	61.5
	$p = .5$	39.0	61.9
	$p = .75$	37.5	62.1
	$p = .95$	33.4	**63.8**

Table 2: Despite lower ROUGE scores, diverse QG with nucleus sampling improves QA results over beam search in zero-shot out-of-domain generation for NEWSQA.

generating incorrect questions also go up. In Table 1, the gains in QA due to QG diversity are generally greater than any drop in performance likely due to decreased accuracy. To find out if the same holds in a more challenging out-of-domain setup, we perform a zero-shot application (i.e., with no further fine-tuning) of four of the above SQUAD-trained QG models to NEWSQA, a reading comprehension dataset of CNN news articles (Trischler et al., 2017). Table 2 shows results on the answerable subset of NEWSQA, with 76k *train* (from which we extract our QG prompts) and 4k *test* (used for QA evaluation) samples: while the absolute scores are lower than those in SQUAD, the relative performances of BEAM and NS are similar both in intrinsic (the best predictor of QA performance for NEWSQA was ROUGE-4) and extrinsic (QA F_1) evaluation.

Comparison with and Augmentation of Human Generation: To assess the quality of our generated questions in absolute terms, in Table 3 we compare the QA performances of the best QG model above (*large*-100%, NS@.95) and corresponding human annotations (GT). Impressively, in-domain model performance on QA is very similar to that of GT, while zero-shot score on NEWSQA is also within roughly 4 points of GT.

We also evaluate the generator's ability to augment human-generated questions. Taking an approach similar to prior augmentation experiments

dataset	train source	QA F_1
SQUAD1-DU	GT (*dev*)	86.3
	SYNTH	86.1
	5×-SYNTH	86.4
	SYNTH* + GT	88.6
NEWSQA	GT (*train*)	67.9
	SYNTH	63.8
	SYNTH* + GT	69.2

Table 3: Diverse QG (SYNTH; NS@.95) shows impressive QA results compared to human annotation (GT), and in augmenting GT (SYNTH* + GT).

(Dong et al., 2019; Alberti et al., 2019), we generate a large synthetic dataset SYNTH* of 4 million examples from Wikipedia passages. The answer spans in these examples are extracted from their corresponding passages using a separate QA model which we train on ten SQUAD question types (instead of full-length questions): *what*, *which*, *where*, *who*, *when*, *why*, *how*, *how many*, *how much*, and *how long*. SYNTH* is used to fine-tune a BERT-wwm LM for QA, which is finally fine-tuned on the target datasets (SQUAD1-DU, NEWSQA). As Table 3 shows, SYNTH* achieves 1.3–2.3 absolute points improvements for the high-performance large BERT-wwm model.

Summary of Results: The above results empirically show that given enough training data and sufficiently powerful QG models: (1) diverse QG leads to strong in-domain and out-of-domain QA training, (2) asking the "most likely" question (i.e., beam search) every time is less useful, and (3) existing generation metrics are inadequate for evaluating diverse question generators as sources of QA training examples.

4 Intrinsic Evaluation of Diverse QG

To better understand the performance of existing generation metrics as measures of diverse QG, we take the set of all 32 samplers in Table 1 (e.g., *base*-100%-p@.75) and randomly generate a large number (100k) of subsets, each consisting of n samplers ($2 \leq n \leq 32$) to be evaluated. We assign each n (# of samplers) to a bin and measure performances of QG metrics separately in each bin. The process is repeated for Table 2. Note that the member sets of a given bin, say $n = 5$, all contain the same number of generators (5), but the actual selection of generators are generally different in different members of a bin. This setup allows us to evaluate a varying number of generators with different capacities and performance, and to average

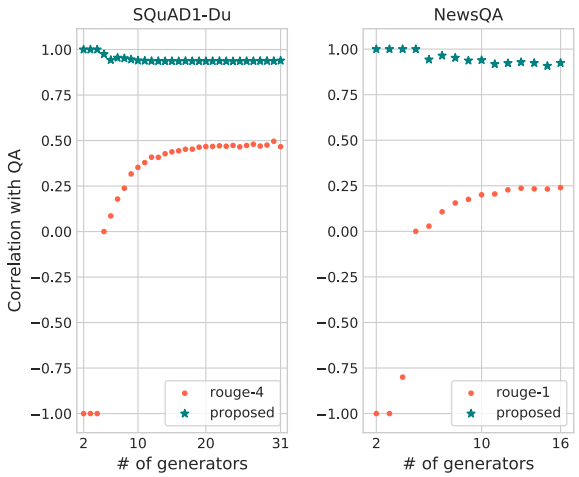

Figure 2: Performances of existing and proposed generation metrics as measures of diverse QG for QA. The proposed metric shows strong correlations (Spearman's $\rho > 90\%$) with QA F_1 in both in-domain and out-of-domain evaluation.

over a large number of experiments.

Figure 2 shows for all bins a rather poor, for some bins negative, median Spearman's ρ score between the best QG metric (SQuAD1-Du: ROUGE-4, NewsQA: ROUGE-1) and downstream QA F_1. These results provide quantitative confirmation that ROUGE and similar metrics are inadequate evaluators of diverse QG for QA due to their sole focus on accuracy with respect to available GTs. This leads us to our final research question: *How to intrinsically measure the overall quality of* QG *for* QA *under diverse nucleus sampling?*

Given the categorical distribution $P_{\mathcal{N}}$ of vocabulary items in a model's nucleus $\mathcal{N}$, we propose to measure both its accuracy (relative to GT) and diversity of generation.

Accuracy: Similarly to LM perplexity, for timestep t of evaluation example s, we take the probability $P_{\mathcal{N}}(q_{s,t} \mid \boldsymbol{p}, \boldsymbol{q}_{s,1:t-1})$ of the model (more precisely, its nucleus $\mathcal{N}$) generating the GT token $q_{s,t}$, given prompt $\boldsymbol{p}$ and GT history $\boldsymbol{q}_{s,1:t-1}$. We then average over all evaluation (s,t) pairs to compute model accuracy $P(\text{GT})$.

Diversity: An intuitive measure of the diversity of a model's nucleus $\mathcal{N}$ is the average entropy of $P_{\mathcal{N}}$ over all evaluation timesteps. However, entropy is an unbounded measure, and has a non-linear inverse growth relative to our proposed accuracy metric, which makes their mathematical combination difficult. We instead rely on the observation that as we increase p in NS@p to make generation more diverse, the cardinality of $\mathcal{N}$ also goes up, on average, and so does the probability $P(\text{GT} \in \mathcal{N})$ that $\mathcal{N}$ contains the GT token. Our experiments on both datasets showed that this measure of diversity, computed as the proportion of times $\mathcal{N}$ was found to include GT across all timesteps in the QG evaluation data, has high positive correlations with the entropy of $P_{\mathcal{N}}$ (Pearson's r: 98%–99%, Spearman's ρ: 87%–95%). Note that unlike the accuracy metric $P(\text{GT})$, at each timestep t, the diversity metric $P(\text{GT} \in \mathcal{N})$ is Boolean: the GT token is either in $\mathcal{N}$ or it is not. But importantly, its average across many evaluation timesteps is a probability measure of diversity, which enables a straightforward convex combination with our proposed accuracy metric.

Our final QG metric is a weighted sum of accuracy and diversity: $w \cdot P(\text{GT}) + (1-w) \cdot P(\text{GT} \in \mathcal{N})$, where $w \in [0,1]$ is a tunable parameter reflecting the weight of accuracy relative to diversity. In our experiments, this metric outperforms all existing metrics by a large margin for a wide range of w values. In Figure 2, the median Spearman's ρ score between this metric and QA F_1 in both in-domain (w=.7) and out-of-domain (w=.8) evaluation is over 90% for all bins. We observe similar performance differences between the proposed and existing metrics with Pearson's r.

Given the scope of this paper, we evaluate the combined metric only on QG, but the underlying ideas apply to diverse text generation in general. Further experiments are necessary to evaluate the metric on other generation tasks.

5 Conclusion

While diversity of generation has received significant attention in other text generation problems (e.g., dialog), we show in this paper that it is also an important and measurable dimension of quality in question generation for QA. We hope that our work will encourage further exploration of diversity-promoting QG and its evaluation. Possible future directions include a systematic study of different aspects of QG diversity (e.g., lexical and factual) and controlled diversification of individual aspects in generation.

Acknowledgments

We thank the anonymous reviewers for their valuable feedback.

References

Chris Alberti, Daniel Andor, Emily Pitler, Jacob Devlin, and Michael Collins. 2019. Synthetic QA Corpora Generation with Roundtrip Consistency. In *ACL*.

Antoine Bordes, Y-Lan Boureau, and Jason Weston. 2017. Learning End-to-End Goal-Oriented Dialog. In *ICLR*.

Samuel R. Bowman, Luke Vilnis, Oriol Vinyals, Andrew M. Dai, Rafal Jozefowicz, and Samy Bengio. 2016. Generating Sentences from a Continuous Space. In *ICLR*.

Kris Cao and Stephen Clark. 2017. Latent Variable Dialogue Models and their Diversity. In *EACL*.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. In *NAACL-HLT*.

Li Dong, Nan Yang, Wenhui Wang, Furu Wei, Xiaodong Liu, Yu Wang, Jianfeng Gao, Ming Zhou, and Hsiao-Wuen Hon. 2019. Unified Language Model Pre-training for Natural Language Understanding and Generation. In *NeurIPS*.

Xinya Du, Junru Shao, and Claire Cardie. 2017. Learning to Ask: Neural Question Generation for Reading Comprehension. In *ACL*.

Zhihao Fan, Zhongyu Wei, Piji Li, Yanyan Lan, and Xuanjing Huang. 2018. A Question Type Driven Framework to Diversify Visual Question Generation. In *IJCAI*.

Ari Holtzman, Jan Buys, Maxwell Forbes, and Yejin Choi. 2020. The Curious Case of Neural Text Degeneration. In *ICLR*.

Unnat Jain, Ziyu Zhang, and Alexander Schwing. 2017. Creativity: Generating Diverse Questions using Variational Autoencoders. In *CVPR*.

Diederik P. Kingma and Max Welling. 2014. Auto-Encoding Variational Bayes. In *ICLR*.

David Lindberg, Fred Popowich, John Nesbit, and Phil Winne. 2013. Generating Natural Language Questions to Support Learning On-Line. In *Proceedings of the European Workshop on Natural Language Generation*.

Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2019. RoBERTa: A Robustly Optimized BERT Pretraining Approach. *arXiv preprint*.

Alec Radford, Jeffrey Wu, Rewon Child, David Luan, Dario Amodei, and Ilya Sutskever. 2019. Language Models are Unsupervised Multitask Learners. *Unpublished manuscript*.

Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J. Liu. 2019. Exploring the Limits of Transfer Learning with a Unified Text-to-Text Transformer. *arXiv preprint*.

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

Adam Trischler, Tong Wang, Xingdi Yuan, Justin Harris, Alessandro Sordoni, Philip Bachman, and Kaheer Suleman. 2017. NewsQA: A Machine Comprehension Dataset. In *Proceedings of the 2nd Workshop on Representation Learning for NLP*.

Ashwin K Vijayakumar, Michael Cogswell, Ramprasaath R Selvaraju, Qing Sun, Stefan Lee, David Crandall, and Dhruv Batra. 2018. Diverse Beam Search for Improved Description of Complex Scenes. In *AAAI*.

Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, R'emi Louf, Morgan Funtowicz, and Jamie Brew. 2019. HuggingFace's Transformers: State-of-the-art Natural Language Processing. *arXiv preprint*.

Shiyue Zhang and Mohit Bansal. 2019. Addressing Semantic Drift in Question Generation for Semi-Supervised Question Answering. In *EMNLP*.