Improving the Robustness of QA Models to Challenge Sets with Variational Question-Answer Pair Generation

Kazutoshi Shinoda[1,2] Saku Sugawara[2] Akiko Aizawa[1,2]

[1] The University of Tokyo

[2] National Institute of Informatics

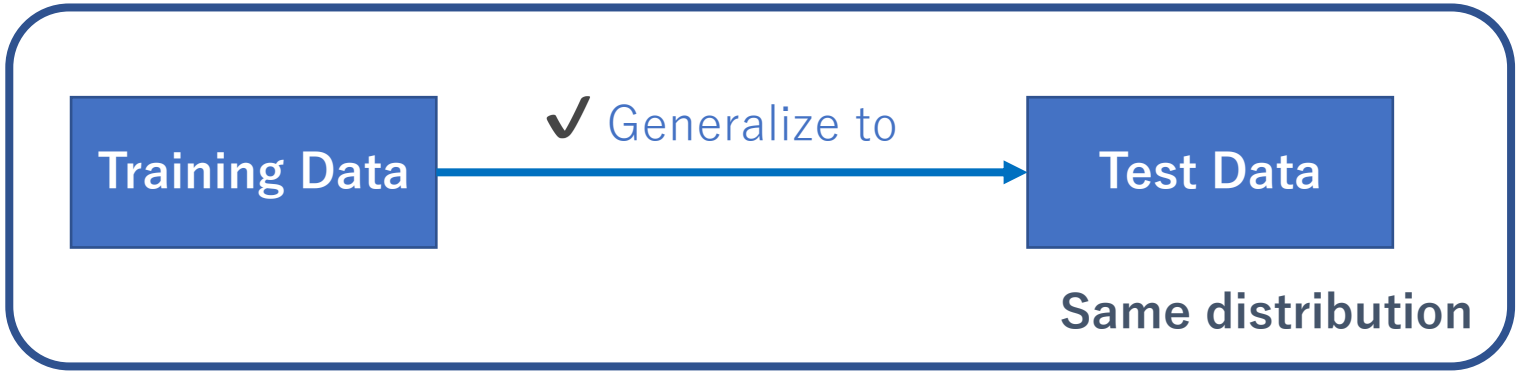
Question Answering (QA)

Context:
Beyoncé Giselle Knowles-Carter (born September 4, 1981) is an American singer, songwriter, record producer and actress. Born and raised in Houston, Texas, she performed in various singing and dancing competitions as a child, and rose to fame in the late 1990s as lead singer of R&B girl-group Destiny's Child. Managed by her father, Mathew Knowles, the group became one of the world's best-selling girl groups of all time. Their hiatus saw the release of Beyoncć's debut album, Dangerously in Love (2003), which established her as a solo artist worldwide, earned five Grammy Awards and featured the Billboard Hot 100 number-one singles "Crazy in Love" and "Baby Boy".
Question-answer pair:
What album made her a worldwide known artist? — Dangerously in Love What was the name of Beyoncé's first solo album? — Dangerously in Love

SQuAD dataset (Rajpurkar et al., 2016)

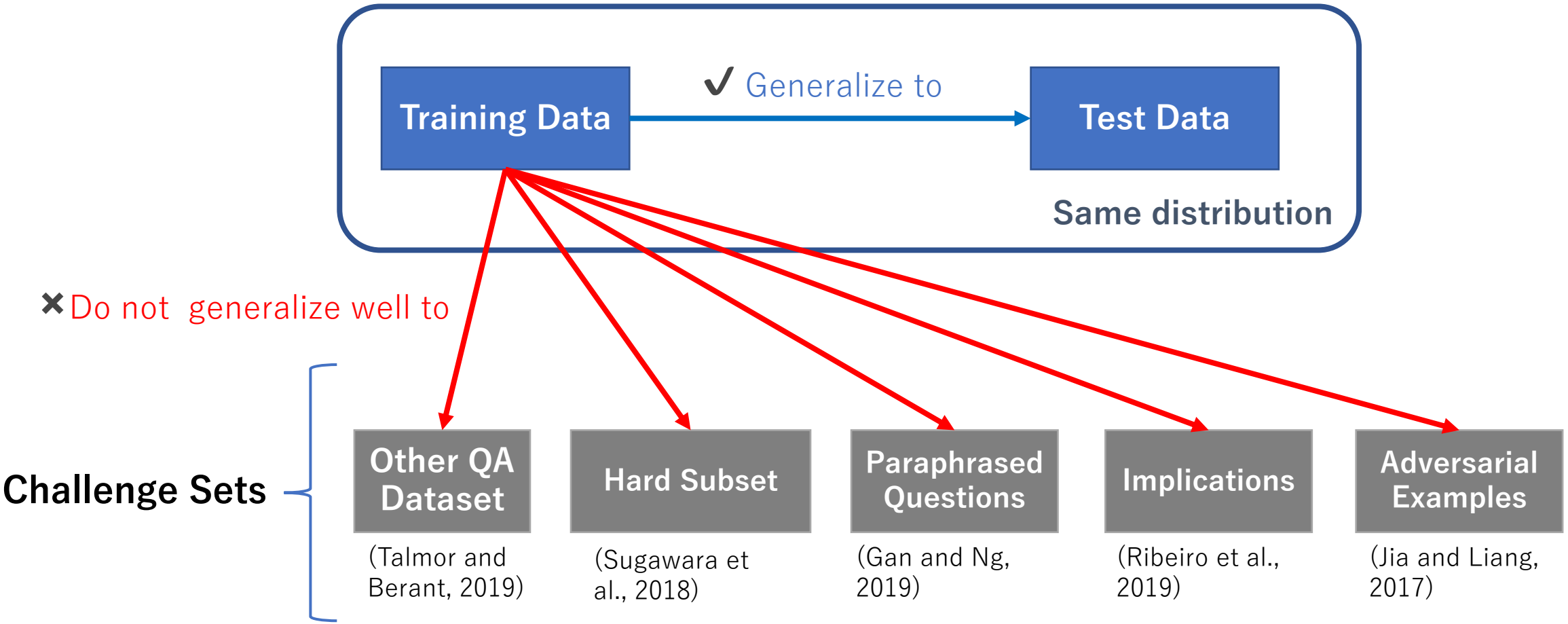
✔A central NLU task requiring broad NLP techniques

Lack of QA model robustness



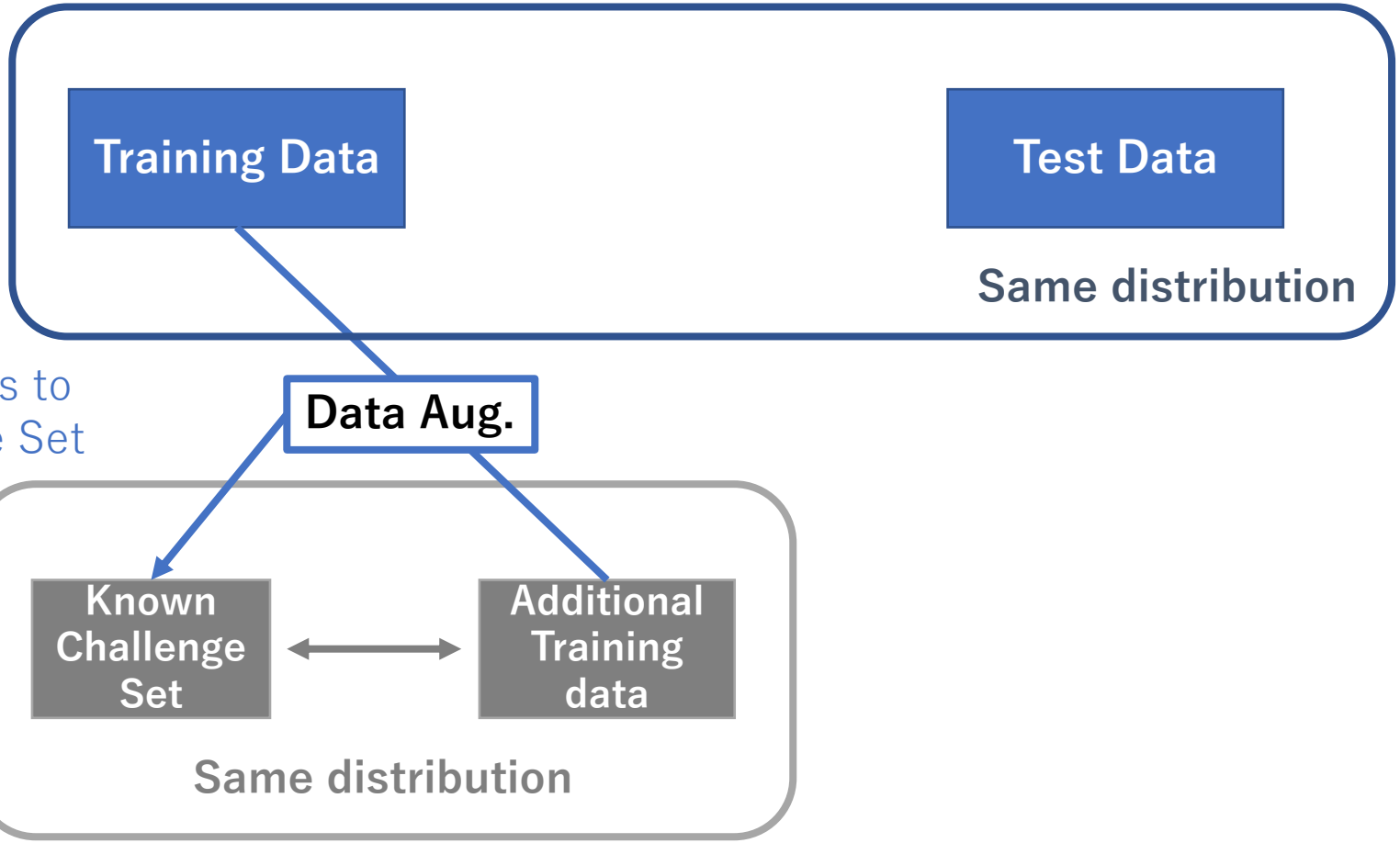
Introduction

Lack of QA model robustness



Adding Examples Similar to a Challenge Set

Existing Approach



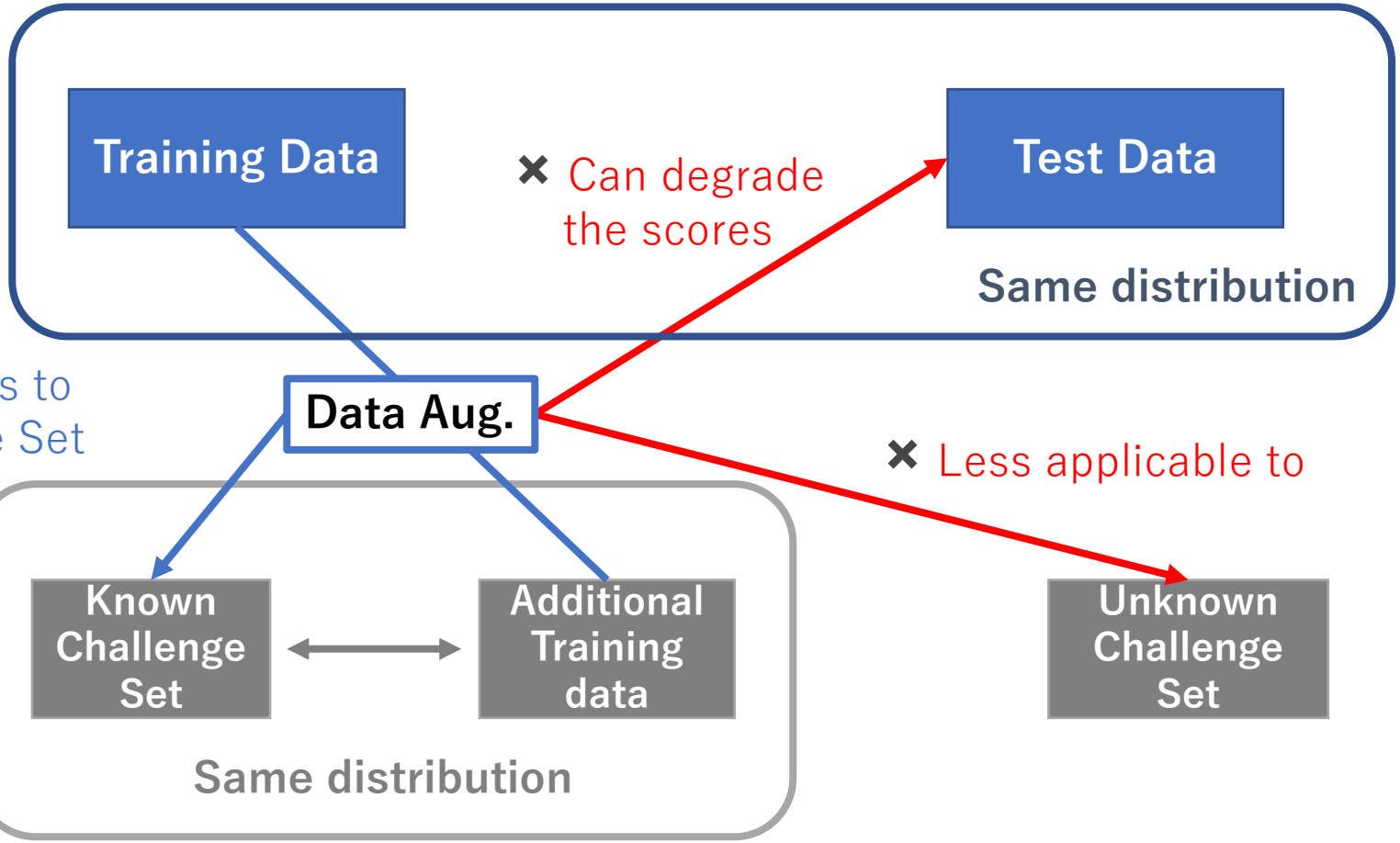
✓Improve Robustness to a Known Challenge Set

(Gan and Ng, 2019)
(Ribeiro et al., 2019)
(Ravichander et al., 2021)

Introduction

Adding Examples Similar to a Challenge Set

Existing Approach

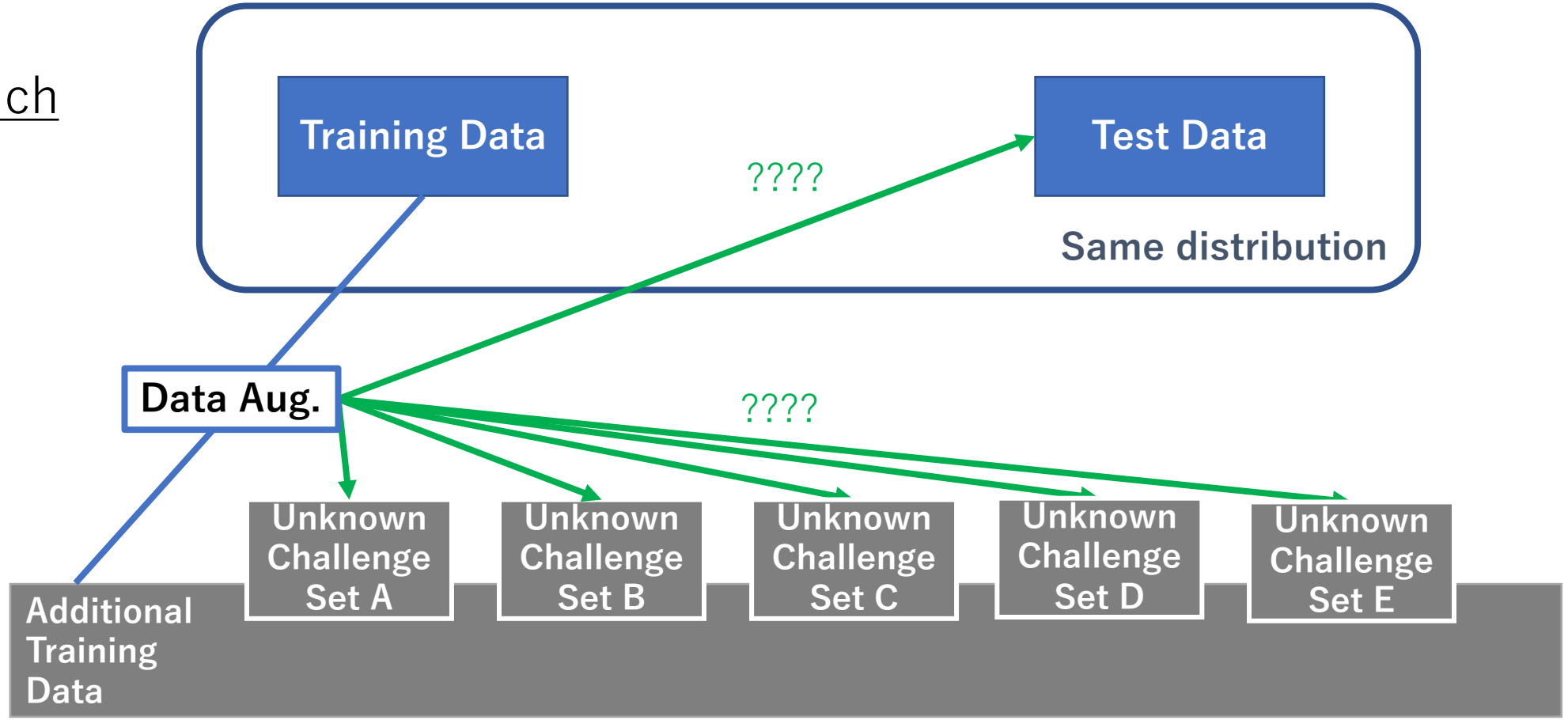


✓Improve Robustness to a Known Challenge Set

(Gan and Ng, 2019)
(Ribeiro et al., 2019)
(Ravichander et al., 2021)

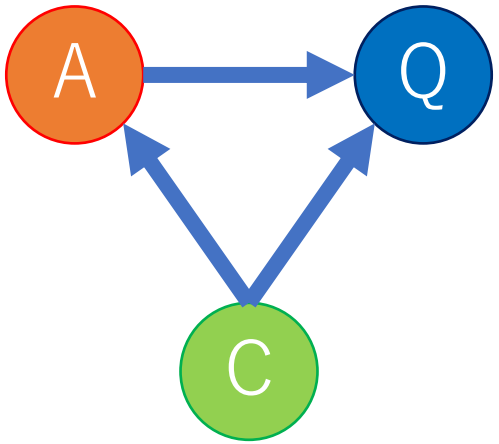
Improving the diversity of training data

Our Approach



QA Pair Generation for Question Answering

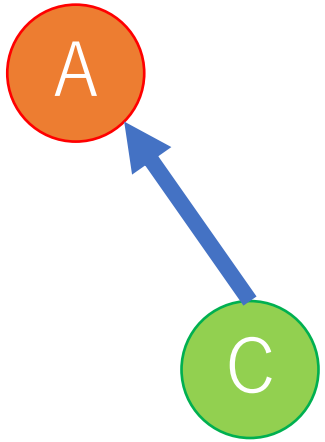
QA Pair Generation



(Du and Cardie, 2018)
(Lee et al., 2020)

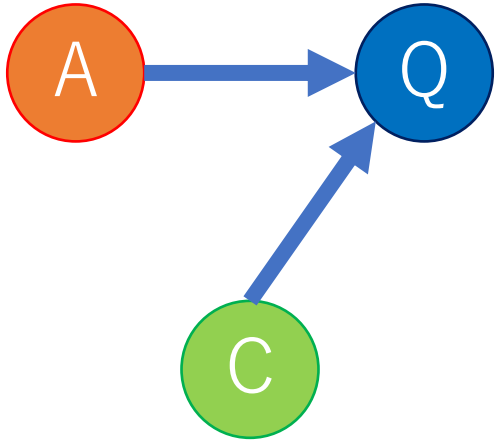
=

Answer Extraction



+

Question Generation



(Zhang and Bansal, 2019)

Notation

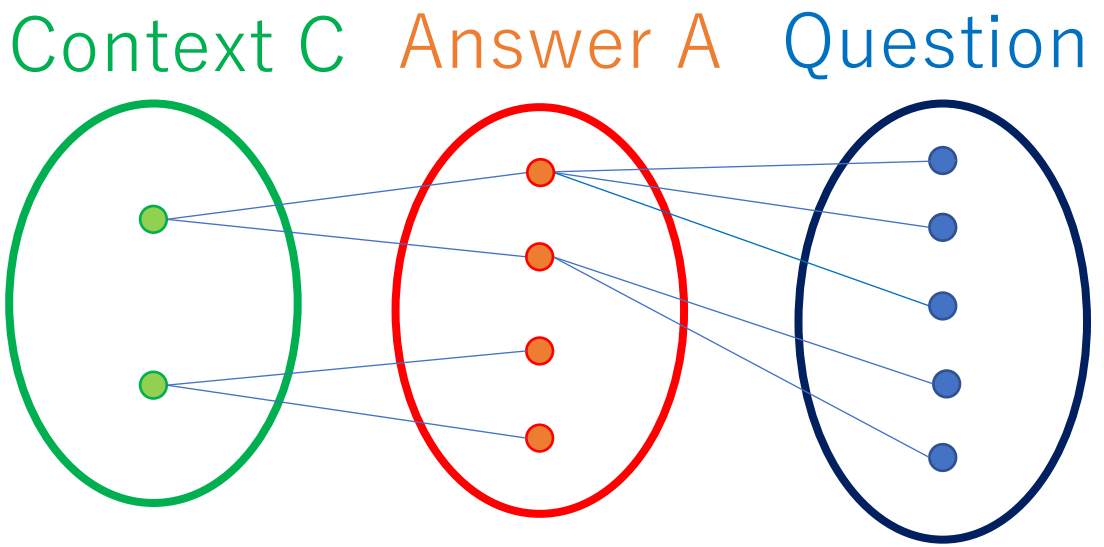
- Context: C
- Question: Q
- Answer: A

Two Levels of Diversity in QA Pairs

Context:
Beyoncé Giselle Knowles-Carter (born September 4, 1981) is an American singer, songwriter, record producer and actress. Born and raised in Houston, Texas, she performed in various singing and dancing competitions as a child, and rose to fame in the late 1990s as lead singer of R&B girl-group Destiny's Child. Managed by her father, Mathew Knowles, the group became one of the world's best-selling girl groups of all time. Their hiatus saw the release of Beyoncć's debut album, Dangerously in Love (2003), which established her as a solo artist worldwide, earned five Grammy Awards and featured the Billboard Hot 100 number-one singles "Crazy in Love" and "Baby Boy".
Question-answer pair:
What album made her a worldwide known artist? — Dangerously in Love
What was the name of Beyoncé's first solo album? — Dangerously in Love

SQuAD dataset (Rajpurkar et al., 2016)

Underlined phrases are used as answers in SQuAD

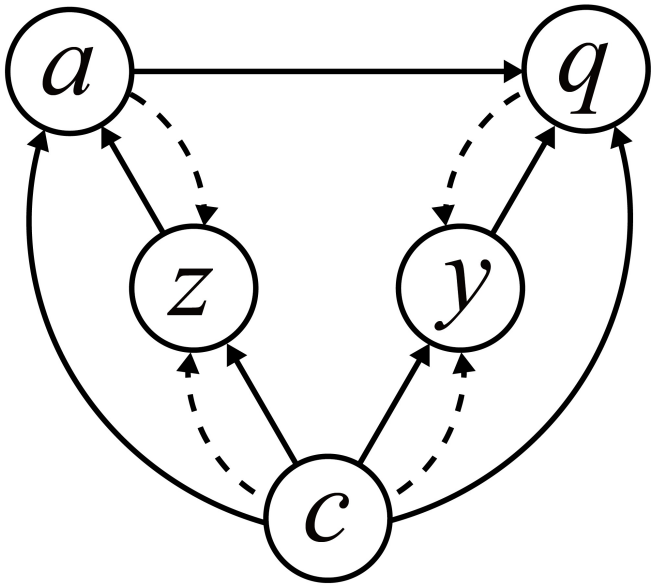


✓Multiple answer candidates can be extracted from a context

✓Multiple questions can be created from a context-answer pair

Approach

Variational Question-Answer Pari Generation



$$\log p_\theta(q, a|c) \geq \mathbb{E}_{q_\phi(z,y|q,a,c)}[\log p_\theta(q|y, a, c) + \log p_\theta(a|z, c)] - D_{\mathrm{KL}}(q_\phi(z|a, c)||p_\theta(z|c)) - D_{\mathrm{KL}}(q_\phi(y|q, c)||p_\theta(y|c)),$$

Generative model θ

Inference model ϕ

Mitigating the Posterior Collapse Issue

Posterior collapse:

a model generates almost the same output
from different latent variables (Bowman et al., 2016)

The modified objective function:

$$\begin{aligned}\mathcal{L} = \mathbb{E}_{q_\phi(z,y|q,a,c)}[&\log p_\theta(q|y,a,c) \\ &+ \log p_\theta(a|z,c)] \\ &- |D_{\mathrm{KL}}(q_\phi(z|a,c)||p_\theta(z|c)) - \mathrm{C_a}| \\ &- |D_{\mathrm{KL}}(q_\phi(y|q,c)||p_\theta(y|c)) - \mathrm{C_q}|, \quad (1)\end{aligned}$$

(Burgess et al., 2018)

Main Difference

- HarQG (Du and Cardie, 2018)
  - Supervised learning
- SemQG (Zhang and Bansal, 2019)
  - Using reinforcement learning to improve the quality of questions
- InfoHCVAE (Lee et al., 2020)
  - CVAE + Maximizing I(Q; A) to improve the consistency of QA pairs
- VQAG (ours)
  - CVAE + Explicitly controlling I(Q; Y|C) and I(A; Z|C) to further enhance the diversity

Result

Answer Extraction & Question Generation

	Relevance				Diversity
	Precision		Recall		Dist
	Prop.	Exact	Prop.	Exact	
NER	34.44	19.61	64.60	45.39	30.0k
HarQG	45.96	33.90	41.05	28.37	-
InfoHCVAE	31.59	16.18	78.75	59.32	70.1k
VQAG					
$C_a = 0$	**58.39**	**47.15**	21.82	16.38	3.1k
$C_a = 5$	30.16	13.41	**83.13**	**60.88**	71.2k
$C_a = 20$	21.95	5.75	72.26	42.15	**103.3k**

Answer Extraction

✓Improve diversity while preserving recall-oriented scores

	Relevance				Diversity		
	B1-R	ME-R	RL-R	Token	D1	E4	SB4
SemQG	**62.32**	**36.77**	**62.87**	7.0M	15.8k	18.28	91.44
VQAG							
$C_q = 0$	35.57	18.31	33.92	7.6M	14.4k	17.33	97.61
$C_q = 5$	44.19	25.84	45.18	11.5M	19.0k	19.71	82.59
$C_q = 20$	48.19	25.29	48.26	4.9M	**22.4k**	**19.72**	**44.41**

Question Generation

✓Improve diversity while degrading recall-oriented scores

✓Different C values correspond to different output distributions

Synthetic dataset construction

Synthetic Datasets	(C_a, C_q)
$\mathcal{D}_{5,5}$	(5, 5)
$\mathcal{D}_{5,20}$	(5, 20)
$\mathcal{D}_{20,20}$	(20, 20)

✔Different C values correspond to different output distributions
→ Combine them to further enhance the diversity of QA pairs

Result

Human Evaluation

Experiments		SemQG	$\mathcal{D}_{5,5}$	$\mathcal{D}_{20,20}$	SQuAD
Question is well-formed	No	2.9%	23.1%	27.8%	2.3%
	Understandable	34.5%	16.0%	17.0%	10.5%
	Yes	62.6%	60.9%	55.1%	87.2%
Question is relevant	No	2.5%	9.5%	11.5%	4.0%
	Yes	97.5%	90.5%	88.5%	96.0%
Answer is correct	No	2.8%	28.8%	30.5%	7.5%
	Partially	21.8%	28.1%	26.6%	11.8%
	Yes	75.4%	43.2%	42.9%	80.6%
Answer is important	No	1.5%	10.0%	5.0%	6.0%
	Yes	98.5%	90.0%	95.0%	94.0%

✖Our synthetic datasets contains many noisy examples

✓90% of our questions are relevant to the contexts
✓90% of our answers are worth being asked about

Heatmap of extracted answers & Generated QA pairs

beyoncé 's vocal range spans [four octaves]. [jody rosen] highlights her tone and timbre as particularly distinctive , describing her voice as " one of the most compelling instruments in popular music " . while another critic says she is a " vocal acrobat , being able to sing long and complex melismas and vocal runs effortlessly , and in key . [her vocal abilities] mean she is identified as the centerpiece of destiny 's child . [the daily mail] calls beyoncé 's voice " [versatile] " , capable of exploring power ballads , soul , rock belting , operatic flourishes , and [hip hop]. jon pareles of the new york times commented that her voice is " velvety yet [tart], with an insistent flutter and reserves of soul belting " .

Q: how can one find her vocal abilities in key music ?	A: she is identified as the centerpiece of destiny 's child
Q: how many octaves spans beyoncé 's vocal range ?	A: spans four
Q: how many octaves 's vocal range spans the beyoncé hop vocal range ?	A: four
Q: who commented that her voice is tart yet tart ?	A: jon pareles

: darker words are more likely to be extracted by our model

: phrases that are used as the ground-truth answers in SQuAD

✓Our model can extract diverse phrases including not only noun phrases but also clauses and adjectives, while extracting the ground-truth answers.

✓The generated questions are not very natural but relevant to the context and answers.

Main Result

QA performance on Challenge Sets

	Training Data (Size)	Challenge Sets												
		$\text{SQuAD}^{\text{Du}}_{\text{test}}$	News	NQ	Quo	Para	APara	Hard	Imp	Add	AddO	MT	ASR	KB
Single	$\text{SQuAD}^{\text{Du}}_{\text{train}}$ (76k)	83.5	49.2	67.7	30.1	85.7	50.2	75.6	64.7	62.9	71.8	79.7	67.5	80.1
	+HarQG (1,205k)	*83.3*	*48.5*	*66.2*	31.3	*85.2*	56.5	*73.0*	*63.5*	65.1	73.1	*78.6*	70.0	80.3
	+SemQG (1,204k)	84.7	50.5	69.8	***34.5***	86.2	51.8	*75.0*	65.1	***66.5***	74.3	*79.5*	71.0	80.7
	+InfoHCVAE (824k)	**84.8**	**51.3**	**71.2**	33.8	85.6	53.3	**77.7**	64.8	66.1	**74.5**	**81.3**	**71.6**	**82.8**
	+VQAG (432k)	84.5	49.2	70.1	32.0	**86.7**	**59.0**	76.1	**66.3**	64.8	73.9	79.9	70.5	81.1
Ensemble	$\{\text{SQuAD}^{\text{Du}}_{\text{train}}\}$*3	84.2	50.4	69.4	31.3	86.4	53.2	76.6	65.7	63.6	72.6	80.3	68.7	81.2
	{+SemQG}*3	85.5	51.8	71.3	**35.1**	87.5	57.8	78.2	66.5	67.0	75.1	80.8	72.9	82.5
	{+InfoHCVAE}*3	85.3	52.0	**72.2**	34.0	88.0	56.9	**79.0**	65.7	**67.7**	**75.9**	81.4	73.1	**83.2**
	{+VQAG}*3	84.9	50.9	70.1	32.3	**88.1**	**58.6**	77.3	**67.5**	64.9	73.9	80.8	71.2	81.6
	{+Sem,+Info,+V}	**85.8**	**52.1**	72.0	34.2	88.0	55.1	78.8	67.0	66.3	74.7	**82.2**	**73.5**	83.0
If challenge set is known		-	62.9	83.0	66.9	88.6	73.9	-	-	-	-	80.8	75.9	82.6

Main Result

QA performance on Challenge Sets

	Training Data (Size)	Challenge Sets												
		$\text{SQuAD}^{\text{Du}}_{\text{test}}$	News	NQ	Quo	Para	APara	Hard	Imp	Add	AddO	MT	ASR	KB
Single	$\text{SQuAD}^{\text{Du}}_{\text{train}}$ (76k)	83.5	49.2	67.7	30.1	85.7	50.2	75.6	64.7	62.9	71.8	79.7	67.5	80.1
	+HarQG (1,205k)	*83.3*	*48.5*	*66.2*	31.3	*85.2*	56.5	*73.0*	*63.5*	65.1	73.1	*78.6*	70.0	80.3
	+SemQG (1,204k)	84.7	50.5	69.8	***34.5***	86.2	51.8	*75.0*	65.1	***66.5***	74.3	*79.5*	71.0	80.7
	+InfoHCVAE (824k)	**84.8**	**51.3**	**71.2**	33.8	*85.6*	53.3	**77.7**	64.8	66.1	**74.5**	**81.3**	**71.6**	**82.8**
	+VQAG (432k)	84.5	49.2	70.1	32.0	**86.7**	**59.0**	76.1	**66.3**	64.8	73.9	79.9	70.5	81.1
Ensemble	$\{\text{SQuAD}^{\text{Du}}_{\text{train}}\}$*3	84.2	50.4	69.4	31.3	86.4	53.2	76.6	65.7	63.6	72.6	80.3	68.7	81.2
	{+SemQG}*3	85.5	51.8	71.3	**35.1**	87.5	57.8	78.2	66.5	67.0	75.1	80.8	72.9	82.5
	{+InfoHCVAE}*3	85.3	52.0	**72.2**	34.0	88.0	56.9	**79.0**	65.7	**67.7**	**75.9**	81.4	73.1	**83.2**
	{+VQAG}*3	84.9	50.9	70.1	32.3	**88.1**	**58.6**	77.3	**67.5**	64.9	73.9	80.8	71.2	81.6
	{+Sem,+Info,+V}	**85.8**	**52.1**	72.0	34.2	88.0	55.1	78.8	67.0	66.3	74.7	**82.2**	**73.5**	83.0
	If challenge set is known	-	62.9	83.0	66.9	88.6	73.9	-	-	-	-	80.8	75.9	82.6

Red : the score is degraded
Bold: the score is the best

Other QA datasets | Variants of SQuAD | Adversarial SQuAD | NoiseQA

Main Result

QA performance on Challenge Sets

	Training Data (Size)	Challenge Sets												
		$SQuAD^{Du}_{test}$	News	NQ	Quo	Para	APara	Hard	Imp	Add	AddO	MT	ASR	KB
Single	$SQuAD^{Du}_{train}$ (76k)	83.5	49.2	67.7	30.1	85.7	50.2	75.6	64.7	62.9	71.8	79.7	67.5	80.1
	+HarQG (1,205k)	*83.3*	*48.5*	*66.2*	31.3	*85.2*	56.5	*73.0*	*63.5*	65.1	73.1	*78.6*	70.0	80.3
	+SemQG (1,204k)	84.7	50.5	69.8	***34.5***	86.2	51.8	*75.0*	65.1	***66.5***	74.3	*79.5*	71.0	80.7
	+InfoHCVAE (824k)	**84.8**	**51.3**	**71.2**	33.8	*85.6*	53.3	**77.7**	64.8	66.1	**74.5**	**81.3**	**71.6**	**82.8**
	+VQAG (432k)	84.5	49.2	70.1	32.0	**86.7**	**59.0**	76.1	**66.3**	64.8	73.9	79.9	70.5	81.1
Ensemble	$\{SQuAD^{Du}_{train}\}$*3	84.2	50.4	69.4	31.3	86.4	53.2	76.6	65.7	63.6	72.6	80.3	68.7	81.2
	{+SemQG}*3	85.5	51.8	71.3	**35.1**	87.5	57.8	78.2	66.5	67.0	75.1	80.8	72.9	82.5
	{+InfoHCVAE}*3	85.3	52.0	**72.2**	34.0	88.0	56.9	**79.0**	65.7	**67.7**	**75.9**	81.4	73.1	**83.2**
	{+VQAG}*3	84.9	50.9	70.1	32.3	**88.1**	**58.6**	77.3	**67.5**	64.9	73.9	80.8	71.2	81.6
	{+Sem,+Info,+V}	**85.8**	**52.1**	72.0	34.2	88.0	55.1	78.8	67.0	66.3	74.7	**82.2**	**73.5**	83.0
If challenge set is known		-	62.9	83.0	66.9	88.6	73.9	-	-	-	-	80.8	75.9	82.6

✖Lack of the robustness of a QA model to the challenge sets

Main Result

QA performance on Challenge Sets

	Training Data (Size)	Challenge Sets												
		$\text{SQuAD}^{\text{Du}}_{\text{test}}$	News	NQ	Quo	Para	APara	Hard	Imp	Add	AddO	MT	ASR	KB
Single	$\text{SQuAD}^{\text{Du}}_{\text{train}}$ (76k)	83.5	49.2	67.7	30.1	85.7	50.2	75.6	64.7	62.9	71.8	79.7	67.5	80.1
	+HarQG (1,205k)	*83.3*	*48.5*	*66.2*	31.3	*85.2*	56.5	*73.0*	*63.5*	65.1	73.1	*78.6*	70.0	80.3
	+SemQG (1,204k)	84.7	50.5	69.8	***34.5***	86.2	51.8	*75.0*	65.1	***66.5***	74.3	*79.5*	71.0	80.7
	+InfoHCVAE (824k)	**84.8**	**51.3**	**71.2**	33.8	*85.6*	53.3	**77.7**	64.8	66.1	**74.5**	**81.3**	**71.6**	**82.8**
	+VQAG (432k)	84.5	49.2	70.1	32.0	**86.7**	**59.0**	76.1	**66.3**	64.8	73.9	79.9	70.5	81.1
Ensemble	$\{\text{SQuAD}^{\text{Du}}_{\text{train}}\}$*3	84.2	50.4	69.4	31.3	86.4	53.2	76.6	65.7	63.6	72.6	80.3	68.7	81.2
	{+SemQG}*3	85.5	51.8	71.3	**35.1**	87.5	57.8	78.2	66.5	67.0	75.1	80.8	72.9	82.5
	{+InfoHCVAE}*3	85.3	52.0	**72.2**	34.0	88.0	56.9	**79.0**	65.7	**67.7**	**75.9**	81.4	73.1	**83.2**
	{+VQAG}*3	84.9	50.9	70.1	32.3	**88.1**	**58.6**	77.3	**67.5**	64.9	73.9	80.8	71.2	81.6
	{+Sem,+Info,+V}	**85.8**	**52.1**	72.0	34.2	88.0	55.1	78.8	67.0	66.3	74.7	**82.2**	**73.5**	83.0
If challenge set is known		-	62.9	83.0	66.9	88.6	73.9	-	-	-	-	80.8	75.9	82.6

✔QA pair generation improved the in-distribution accuracy in general

Main Result

QA performance on Challenge Sets

	Training Data (Size)	$\text{SQuAD}^{\text{Du}}_{\text{test}}$	Challenge Sets											
			News	NQ	Quo	Para	APara	Hard	Imp	Add	AddO	MT	ASR	KB
Single	$\text{SQuAD}^{\text{Du}}_{\text{train}}$ (76k)	83.5	49.2	67.7	30.1	85.7	50.2	75.6	64.7	62.9	71.8	79.7	67.5	80.1
	+HarQG (1,205k)	*83.3*	*48.5*	*66.2*	31.3	*85.2*	56.5	*73.0*	*63.5*	65.1	73.1	*78.6*	70.0	80.3
	+SemQG (1,204k)	84.7	50.5	69.8	***34.5***	86.2	51.8	*75.0*	65.1	***66.5***	74.3	*79.5*	71.0	80.7
	+InfoHCVAE (824k)	**84.8**	**51.3**	**71.2**	33.8	85.6	53.3	**77.7**	64.8	66.1	**74.5**	**81.3**	**71.6**	**82.8**
	+VQAG (432k)	84.5	49.2	70.1	32.0	**86.7**	**59.0**	76.1	**66.3**	64.8	73.9	79.9	70.5	81.1
Ensemble	$\{\text{SQuAD}^{\text{Du}}_{\text{train}}\}$*3	84.2	50.4	69.4	31.3	86.4	53.2	76.6	65.7	63.6	72.6	80.3	68.7	81.2
	{+SemQG}*3	85.5	51.8	71.3	**35.1**	87.5	57.8	78.2	66.5	67.0	75.1	80.8	72.9	82.5
	{+InfoHCVAE}*3	85.3	52.0	**72.2**	34.0	88.0	56.9	**79.0**	65.7	**67.7**	**75.9**	81.4	73.1	**83.2**
	{+VQAG}*3	84.9	50.9	70.1	32.3	**88.1**	**58.6**	77.3	**67.5**	64.9	73.9	80.8	71.2	81.6
	{+Sem,+Info,+V}	**85.8**	**52.1**	72.0	34.2	88.0	55.1	78.8	67.0	66.3	74.7	**82.2**	**73.5**	83.0
If challenge set is known		-	62.9	83.0	66.9	88.6	73.9	-	-	-	-	80.8	75.9	82.6

✔Different QAG methods have different benefits.

Main Result

QA performance on Challenge Sets

	Training Data (Size)	Challenge Sets												
		$SQuAD^{Du}_{test}$	News	NQ	Quo	Para	APara	Hard	Imp	Add	AddO	MT	ASR	KB
Single	$SQuAD^{Du}_{train}$ (76k)	83.5	49.2	67.7	30.1	85.7	50.2	75.6	64.7	62.9	71.8	79.7	67.5	80.1
	+HarQG (1,205k)	*83.3*	*48.5*	*66.2*	31.3	*85.2*	56.5	*73.0*	*63.5*	65.1	73.1	*78.6*	70.0	80.3
	+SemQG (1,204k)	84.7	50.5	69.8	***34.5***	86.2	51.8	*75.0*	65.1	***66.5***	74.3	*79.5*	71.0	80.7
	+InfoHCVAE (824k)	**84.8**	**51.3**	**71.2**	33.8	85.6	53.3	**77.7**	64.8	66.1	**74.5**	**81.3**	**71.6**	**82.8**
	+VQAG (432k)	84.5	49.2	70.1	32.0	**86.7**	**59.0**	76.1	**66.3**	64.8	73.9	79.9	70.5	81.1
Ensemble	{$SQuAD^{Du}_{train}$}*3	84.2	50.4	69.4	31.3	86.4	53.2	76.6	65.7	63.6	72.6	80.3	68.7	81.2
	{+SemQG}*3	85.5	51.8	71.3	**35.1**	87.5	57.8	78.2	66.5	67.0	75.1	80.8	72.9	82.5
	{+InfoHCVAE}*3	85.3	52.0	**72.2**	34.0	88.0	56.9	**79.0**	65.7	**67.7**	**75.9**	81.4	73.1	**83.2**
	{+VQAG}*3	84.9	50.9	70.1	32.3	**88.1**	**58.6**	77.3	**67.5**	64.9	73.9	80.8	71.2	81.6
	{+Sem,+Info,+V}	**85.8**	**52.1**	72.0	34.2	88.0	55.1	78.8	67.0	66.3	74.7	**82.2**	**73.5**	83.0
If challenge set is known		-	62.9	83.0	66.9	88.6	73.9	-	-	-	-	80.8	75.9	82.6

✔Different benefits can be combined in the Ensemble setting

Main Result

QA performance on Challenge Sets

	Training Data (Size)	Challenge Sets												
		$\text{SQuAD}^{\text{Du}}_{\text{test}}$	News	NQ	Quo	Para	APara	Hard	Imp	Add	AddO	MT	ASR	KB
Single	$\text{SQuAD}^{\text{Du}}_{\text{train}}$ (76k)	83.5	49.2	67.7	30.1	85.7	50.2	75.6	64.7	62.9	71.8	79.7	67.5	80.1
	+HarQG (1,205k)	*83.3*	*48.5*	*66.2*	31.3	*85.2*	56.5	*73.0*	*63.5*	65.1	73.1	*78.6*	70.0	80.3
	+SemQG (1,204k)	84.7	50.5	69.8	***34.5***	86.2	51.8	*75.0*	65.1	***66.5***	74.3	*79.5*	71.0	80.7
	+InfoHCVAE (824k)	**84.8**	**51.3**	**71.2**	33.8	*85.6*	53.3	**77.7**	64.8	66.1	**74.5**	**81.3**	**71.6**	**82.8**
	+VQAG (432k)	84.5	49.2	70.1	32.0	**86.7**	**59.0**	76.1	**66.3**	64.8	73.9	79.9	70.5	81.1
Ensemble	$\{\text{SQuAD}^{\text{Du}}_{\text{train}}\}$*3	84.2	50.4	69.4	31.3	86.4	53.2	76.6	65.7	63.6	72.6	80.3	68.7	81.2
	{+SemQG}*3	85.5	51.8	71.3	**35.1**	87.5	57.8	78.2	66.5	67.0	75.1	80.8	72.9	82.5
	{+InfoHCVAE}*3	85.3	52.0	**72.2**	34.0	88.0	56.9	**79.0**	65.7	**67.7**	**75.9**	81.4	73.1	**83.2**
	{+VQAG}*3	84.9	50.9	70.1	32.3	**88.1**	**58.6**	77.3	**67.5**	64.9	73.9	80.8	71.2	81.6
	{+Sem,+Info,+V}	**85.8**	**52.1**	72.0	34.2	88.0	55.1	78.8	67.0	66.3	74.7	**82.2**	**73.5**	83.0
If challenge set is known		-	62.9	83.0	66.9	88.6	73.9	-	-	-	-	80.8	75.9	82.6

✔VQAG did not degrade the scores on the 12 challenge sets while improving the in-distribution accuracy

Main Result

QA performance on Challenge Sets

	Training Data (Size)	$\text{SQuAD}^{\text{Du}}_{\text{test}}$	Challenge Sets											
			News	NQ	Quo	Para	APara	Hard	Imp	Add	AddO	MT	ASR	KB
Single	$\text{SQuAD}^{\text{Du}}_{\text{train}}$ (76k)	83.5	49.2	67.7	30.1	85.7	50.2	75.6	64.7	62.9	71.8	79.7	67.5	80.1
	+HarQG (1,205k)	*83.3*	*48.5*	*66.2*	31.3	*85.2*	56.5	*73.0*	*63.5*	65.1	73.1	*78.6*	70.0	80.3
	+SemQG (1,204k)	84.7	50.5	69.8	***34.5***	86.2	51.8	*75.0*	65.1	***66.5***	74.3	*79.5*	71.0	80.7
	+InfoHCVAE (824k)	**84.8**	**51.3**	**71.2**	33.8	85.6	53.3	**77.7**	64.8	66.1	**74.5**	**81.3**	**71.6**	**82.8**
	+VQAG (432k)	84.5	49.2	70.1	32.0	**86.7**	**59.0**	76.1	**66.3**	64.8	73.9	79.9	70.5	81.1
Ensemble	$\{\text{SQuAD}^{\text{Du}}_{\text{train}}\}$*3	84.2	50.4	69.4	31.3	86.4	53.2	76.6	65.7	63.6	72.6	80.3	68.7	81.2
	{+SemQG}*3	85.5	51.8	71.3	**35.1**	87.5	57.8	78.2	66.5	67.0	75.1	80.8	72.9	82.5
	{+InfoHCVAE}*3	85.3	52.0	**72.2**	34.0	88.0	56.9	**79.0**	65.7	**67.7**	**75.9**	81.4	73.1	**83.2**
	{+VQAG}*3	84.9	50.9	70.1	32.3	**88.1**	**58.6**	77.3	**67.5**	64.9	73.9	80.8	71.2	81.6
	{+Sem,+Info,+V}	**85.8**	**52.1**	72.0	34.2	88.0	55.1	78.8	67.0	66.3	74.7	**82.2**	**73.5**	83.0
If challenge set is known		-	62.9	83.0	66.9	88.6	73.9	-	-	-	-	80.8	75.9	82.6

✔Generating too many examples may be more likely to induce the trade-off between the scores.

Main Result

QA performance on Challenge Sets

	Training Data (Size)	$\text{SQuAD}^{\text{Du}}_{\text{test}}$	Challenge Sets											
			News	NQ	Quo	Para	APara	Hard	Imp	Add	AddO	MT	ASR	KB
Single	$\text{SQuAD}^{\text{Du}}_{\text{train}}$ (76k)	83.5	49.2	67.7	30.1	85.7	50.2	75.6	64.7	62.9	71.8	79.7	67.5	80.1
	+HarQG (1,205k)	*83.3*	*48.5*	*66.2*	31.3	*85.2*	56.5	*73.0*	*63.5*	65.1	73.1	*78.6*	70.0	80.3
	+SemQG (1,204k)	84.7	50.5	69.8	***34.5***	86.2	51.8	*75.0*	65.1	***66.5***	74.3	*79.5*	71.0	80.7
	+InfoHCVAE (824k)	**84.8**	**51.3**	**71.2**	33.8	*85.6*	53.3	**77.7**	64.8	66.1	**74.5**	**81.3**	**71.6**	**82.8**
	+VQAG (432k)	84.5	49.2	70.1	32.0	**86.7**	**59.0**	76.1	**66.3**	64.8	73.9	79.9	70.5	81.1
Ensemble	$\{\text{SQuAD}^{\text{Du}}_{\text{train}}\}$*3	84.2	50.4	69.4	31.3	86.4	53.2	76.6	65.7	63.6	72.6	80.3	68.7	81.2
	{+SemQG}*3	85.5	51.8	71.3	**35.1**	87.5	57.8	78.2	66.5	67.0	75.1	80.8	72.9	82.5
	{+InfoHCVAE}*3	85.3	52.0	**72.2**	34.0	88.0	56.9	**79.0**	65.7	**67.7**	**75.9**	81.4	73.1	**83.2**
	{+VQAG}*3	84.9	50.9	70.1	32.3	**88.1**	**58.6**	77.3	**67.5**	64.9	73.9	80.8	71.2	81.6
	{+Sem,+Info,+V}	**85.8**	**52.1**	72.0	34.2	88.0	55.1	78.8	67.0	66.3	74.7	**82.2**	**73.5**	83.0
If challenge set is known		-	62.9	83.0	66.9	88.6	73.9	-	-	-	-	80.8	75.9	82.6

✖There is still a significant room for improvement in some challenge sets

Analysis

Ablation Study

Training Data (Size)	EM	F1
VQAG (432k)	81.49	88.61
$-\mathcal{D}_{5,5}$ (251k)	81.04	88.39
$-\mathcal{D}_{5,20}$ (113k)	81.00	88.48
$-\mathcal{D}_{20,20}$ (68k)	81.14	88.52

Evaluation on the SQuAD-Du dev set

✓Each synthetic dataset contributes to the QA performance

Analysis

Distributions of Question Types

Dataset	what	how	who	which	when	where	why
$\text{SQuAD}^{\text{Du}}_{\text{train}}$	58.3	10.4	10.3	6.7	6.7	4.2	1.5
$\text{SQuAD}^{\text{Du}}_{\text{test}}$	56.5	12.1	11.5	8.6	6.0	3.8	0.8
HarQG	61.3	7.8	13.8	0.7	10.1	5.8	0.5
SemQG	71.1	8.1	12.8	1.3	3.6	2.7	0.2
InfoHCVAE	77.1	6.6	5.0	1.6	5.6	3.3	0.5
VQAG							
$\mathcal{D}_{5,5}$	36.6	54.9	4.9	0.5	0.3	0.5	2.3
$\mathcal{D}_{5,20}$	9.5	35.5	3.6	49.2	1.2	0.9	0.0
$\mathcal{D}_{20,20}$	28.2	36.7	6.3	23.2	0.2	1.6	3.9
							(%)

✓Our dataset is more diverse than other datasets

✓Our method is unique in controlling the distribution using different configurations

Summary

Contributions

- We proposed the variational QA pair generation (VQAG) model with explicit KL control to generate diverse QA pairs from contexts.
- We showed that our noisy but diverse synthetic datasets are effective to improve the robustness of a QA model while improving the in-distribution score, even though the distributions of the challenge sets are not known a priori.

Future work

- We will identify the reason why noisy examples are effective to improve the QA performance.

Thank you for listening!

Our data and code are available here.

(https://github.com/KazutoshiShinoda/VQAG)