

変分質問回答ペア生成による質問応答モデルの汎化性能と頑健性の向上

篠田一聡 (東大/NII)
shinoda@is.s.u-tokyo.ac.jp

菅原朔 (NII)
saku@nii.ac.jp

相澤彰子 (東大/NII)
aizawa@nii.ac.jp

1. 導入

背景

質問応答モデルは訓練時と同じ分布のテストセットには汎化するが**違う分布のテストセットには汎化しない** [1]

目的

質問応答モデルの**分布外データセットへの汎化性能と難しいテストセットへの頑健性の向上**

手法

変分質問回答ペア生成によるデータ拡張によって**訓練セット中の質問回答ペアの多様性を向上**する

2. 変分質問回答ペア生成

- 潜在変数 $y \rightleftharpoons$ 質問 q , 潜在変数 $z \rightleftharpoons$ 回答 a の写像を学習することで多様な質問と回答の生成が可能
- 出力が潜在変数に依存しない Posterior Collapse 問題を回避するために KL 項を C_a, C_q によって制御

↑ 質問生成 + 回答抽出

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

✓ C_a, C_q が 0 より大きい時にこの問題を回避 (表1, 2)

3. 実験

訓練セット: SQuAD-Du (Wikipedia)

ベースライン:

- HarQG (BiLSTM-CRF + seq2seq)
- SemQG (BiLSTM-CRF + 強化学習)

質重視

質問応答モデル: BERT-base

提案データ拡張手法: $(C_a, C_q) = (5, 5), (5, 20), (20, 20)$ で訓練/生成

✓ (C_a, C_q) によって出力の分布が異なり (表1, 2) それぞれが精度向上に寄与 (付録 表5)

4. 実験結果

	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	-
Ours					
$C_a = 0$	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

表1 テストデータにおける回答抽出の結果。

N: 生成文の数
Token: 単語数
B1: Bleu1
ME: Meteor
RL: Rouge-L
D1: Dist-1
E4: Ent-4
SB4: Self-B4
-R: 再現率

回答の評価指標
Prop.: 部分一致
Exact: 完全一致

- ✓ 人手データやNERよりも**多様な回答の抽出**
- ✓ 既存手法よりも**高い再現率**

	Dev	Test
SQuAD-Du	80.12/87.85	72.69/84.08
+HarQG	79.49/87.05	72.32/83.31
+SemQG	81.02/88.53	73.59/84.72
+Ours	81.49/88.61	73.11/84.53

表4 半教師あり学習

- ✓ 分布内データセットにおいて提案手法は既存手法に匹敵する精度

評価指標
EM/F1スコア
EM: 完全一致
F1: 部分一致

	Relevance				Diversity			
	N	B1-R	ME-R	RL-R	Token	D1	E4	SB4
SemQG	50	62.32	36.77	62.87	7.0M	15.8k	18.28	91.44
Ours								
$C_q = 0$	50	35.57	18.31	33.92	7.6M	14.4k	17.33	97.61
$C_q = 5$	50	44.19	25.84	45.18	11.5M	19.0k	19.71	82.59
$C_q = 20$	50	48.19	25.29	48.26	4.9M	22.4k	19.72	44.41

表2 テストデータにおける質問生成の結果。

✓ ベースラインよりも**多様な質問の生成**

分布外データセット:

- NewsQA (ニュース記事)
- TriviaQA (雑学)
- NQ (検索クエリ+Wiki)

難しいテストセット:

- non-Adv/Adv (質問のパラフレーズ)
- Easy/Hard (簡単/難しい質問)
- Implications (一貫性を要する質問)

Data	Generalization to OOD QA Datasets			Robustness to Challenge Test Sets				
	NewsQA	TriviaQA	NQ	non-Adv	Adv	Easy	Hard	Implications
SQuAD-Du	32.81/49.21	37.40/47.57	55.35/67.70	78.15/85.73	42.86/50.16	82.49/90.13	67.43/75.59	49.43/64.72
+HarQG	32.85/48.46	36.42/45.84	54.97/66.20	76.65/85.15	51.79/56.52	82.00/89.67	66.04/73.01	49.24/63.47
+SemQG	33.86/50.51	37.56/47.50	58.19/69.81	78.91/86.21	46.43/51.82	83.50/91.05	67.73/75.02	49.72/65.08
+Ours	32.81/49.25	38.19/47.72	58.02/70.06	79.00/86.73	51.79/59.00	83.87/90.94	68.75/76.10	50.63/66.26
Target	42.61/62.90	55.80/61.66	74.19/83.03	-	-	-	-	-

表3 質問応答モデルの汎化性能と頑健性の評価

- ✓ 分布外データセットにおいて特に**NQ**に対して**汎化性能が向上**
- ✓ 難しいテストセットではほぼ一貫して**最も高い精度**を達成

5. 結論

✓ 質問回答ペアの多様性の向上により質問応答モデルの汎化性能・頑健性が向上

✓ 提案手法は**ターゲットの分布が未知**にも関わらず分布内と分布外の双方での精度向上が可能

(ターゲットの分布が既知の設定でのデータ拡張は分布内テストセットでの精度を犠牲にしようとする報告している既存研究もある)

✓ 生成データの質の改善と外部コーパスの利用が今後の課題

参考文献

[1] Talmor+ "MultiQA: An empirical investigation of generalization and transfer in reading comprehension". In ACL (2019)