自動生成データを活用した機械読解モデルの汎用性の向上

谷口 元樹　　高橋 拓誠　　谷口 友紀　　大熊 智子
富士ゼロックス株式会社
{motoki.taniguchi, takahashi.takumi, taniguchi.tomoki, ohkuma.tomoko}@fujixerox.co.jp

1 はじめに

機械読解タスクの目的は、質問に対応する回答を与えれた文書（コンテキスト）から抽出することである。機械読解はニューラルネットワークモデルの言語理解のベンチマークタスクとして多くの研究がなされており、人間を上回る精度を達成している [2]。また、機械読解は対話システムや検索システムなどのアプリケーションに活用できる重要な要素技術である。

機械読解のモデルは学習データと同一のドメイン（in-domain）のデータでは人間を上回る精度を達成しているものの、学習データとは異なるドメイン（out–of-domain）では精度は低いことが報告されている [6]。このため、out-of-domain のデータでのモデルの精度をモデルの汎用性とみなして、これを改善する研究が近年注目を浴びている [6, 4]。汎用性が低いモデルを様々なドメインへ適用する場合には、対象ドメインごとに大量のコンテキスト C、質問 Q、回答 A の三つ組み（CQA トリプレット) を学習データとして追加学習する必要がある。しかし、大量の学習データを人手で作成するコストは非常に高いため、ドメインごとに学習データを作成することは現実的ではない。

機械読解モデルの in-domain における精度を改善する方法の一つに、人工的に生成した CQA トリプレットを用いて学習データを拡張する方法がある [1, 3, 5]。これらの研究では、まず回答抽出モデルを用いてコンテキストから回答になりえる範囲を抽出する。次に、質問生成モデルを用いて与えられたコンテキストと抽出された回答から回答に対応する質問を生成することで、CQA トリプレットを生成する。この生成した CQA トリプレットを用いることで、人手で作成した CQA トリプレットを拡張して機械読解モデルを学習する。しかし、これらの研究は、in-domain での機械読解モデルの精度向上を主な目的としており、汎用性に与える影響については調査されていない。

そこで、本論文では機械読解モデルの汎用性を向上するために、回答抽出モデルと質問生成モデルを用いて CQA トリプレットを生成することで学習データを拡張する手法について検討する。機械読解モデルの汎用性を効率的に改善するためには、CQA トリプレットの量と質が重要となる。特に質に関しては、低品質な CQA トリプレットが学習データに含まれると、精度を劣化させるおそれがある。そこで、機械読解モデルによる CQA トリプレットのフィルタと回答の修正を行い、低品質な CQA トリプレットの除外および品質の改善行う。本論文の貢献は以下の 2 点である。

- 生成した CQA トリプレットで事前学習することで、機械読解モデルの out-of-domain での精度が向上することを示す。
- 生成した CQA トリプレットの量と質が機械読解モデルの out-of-domain での精度に与える影響を分析し、影響を明らかにする。

2 関連研究

機械読解モデルの汎用性: [6] は機械読解の 10 個のデータセットを用いて、1 個のデータセットで学習した機械読解モデルの精度を残りの 9 個のデータセットで評価した。その結果、BERT[2] のような表現力が豊かなモデルであっても、out-of-domain の精度は高くはないことを示した。[4] は、機械読解モデルのシェアドタスクである Machine Reading for Question Answering (MRQA) を開催し、参加システムの特徴や精度を報告している。MRQA シェアドタスクでは 6 個の in-domain データセットで学習した機械読解モデルの汎用性を 12 個の out-of-domain のデータセットで評価した。

自動生成 CQA トリプレットによるデータ拡張: [1, 3] は、自動生成した CQA トリプレットで学習データを拡張することで、機械読解モデルの性能が向上することを示した。どちらの研究も in-domain で評価を行っており、out-of-domain の精度が向上するかは明らかではない。また、提案手法と同様に生成データのフィルタリングや修正を行っているが、その効果は定量的に評価されていない。

3 自動生成データによるデータ拡張

本節では、CQA トリプレットを生成することで機械読解の学習データを拡張する手法について説明する。

3.1 タスク設定

in-domain の N 個の機械読解データセット $S = \{s_1, s_2, \cdots, s_N\}$ が与えられ、out-of-domain の M 個のデータセット $T = \{t_1, t_2, \cdots, t_M\}$ を用いて機械読解モデルの精度を評価するタスクを考える。各データセット s_i には文書集合 C_{s_i}, 質問集合 Q_{s_i}, 回答集合 A_{s_i} があり、文書 $C_{s_i^j}$, 質問 $Q_{s_i^j}$, 回答 $A_{s_i^j}$ が j 番目の CQA トリプレットを構成している。

3.2 生成データの作成

以下の 4 つの手順で行うことで、生成データ G を作成する。

1. コンテキストの選定

CQA トリプレットの生成する元となる生コーパスとして 2017 年 3 月の Wikipedia ダンプデータ[1] からランダムにサンプリングした 20 万記事を用いる。各記事をパラグラフに分割し、80 単語以上 500 単語以下のパラグラフをコンテキスト C_G として用いる。

2. 回答の抽出

コンテキスト C_G を入力として回答抽出モデルを適用することで、回答 A_G を抽出する。回答抽出モデルは in-domain データセット S を用いて文書 C_S から回答 A_S を抽出するようにあらかじめ学習する。一般的に 1 つのパラグラフに対して複数の回答が候補として考えられるため、1 つのパラグラフから上位 30 の回答を抽出する。

3. 質問の生成

コンテキスト C_G と回答 A_G を入力として質問生成モデルを適用することで、質問 Q_G を生成する。質問生成モデルは in-domain データセット S を用いて文書 C_S と回答 A_S から質問 Q_S を生成するようにあらかじめ学習する。

4. フィルタリングと回答の修正

回答抽出、質問生成ともに自動的に行っているため、生成された CQA トリプレットには誤った回答や質問が含まれている可能性が高い。誤った CQA トリプレットが学習データに含まれていると、機械読解モデルの精度の低下をまねくおそれがるため、CQA トリプレットのフィルタリングと修正を行う。

あらかじめ S の文書 C_S, 質問 Q_S, 回答 A_S を用いて、機械読解モデルを学習する。この機械読解モデルを用いて、コンテキスト C_G と質問 Q_G から回答 A'_G を予測する。この予測回答 A'_G と回答 A_G のオーバーラップを文字ベースの F1 値で評価し、しきい値以下であれば誤った CQA トリプレットであるとして除外する。また、回答 A_G を予測回答 A'_G で置き換えたものであるコンテキスト C_G、質問 Q_G、予測回答 A'_G の CQA トリプレットを最終的な生成データ G とする。

3.3 機械読解モデルの学習

生成データ G と真の学習データ S を用いて、機械読解モデルを学習する。まず、生成データ G を用いて、機械読解モデルの事前学習を行う。そして、真の学習データ S を用いて追加学習することで、最終的な機械読解モデルを得る。

S_1	S_6	データセット	学習	開発	評価
✓	✓	SQuAD	76,079	10,507	10,509
	✓	NewsQA	69,947	4,212	4,213
	✓	TriviaQA	53,902	7,785	7,786
	✓	SearchQA	100,403	16,980	16,981
	✓	HotpotQA	67,010	5,904	5,902
	✓	NaturalQ	91,234	12,836	12,837

表 1: in-domain データセット S_1 と S_6 のサイズ

データセット	評価
DROP	1,503
RACE	674
BioASQ	1,504
TextbookQA	1,503
RelationExtraction	2,948
DuoRC	1,501

表 2: out-of-domain データセット T のサイズ

生成データ	しきい値	サイズ
$G_{\text{フィルタなし}}$	なし	11,593,276
$G_{F1 \geqq 0.2}$	$F1 \geqq 0.2$	8,206,836
$G_{F1 \geqq 0.4}$	$F1 \geqq 0.4$	5,917,182
$G_{F1 \geqq 0.6}$	$F1 \geqq 0.6$	3,767,634
$G_{F1 \geqq 0.8}$	$F1 \geqq 0.8$	2,156,352
$G_{F1=1}$	$F1 = 1$	1,564,836

表 3: 生成データ G のフィルタしきい値とサイズ

4 実験設定

実験には MRQA データセットの 6 つのデータセットを in-domain データセット S、また別の 6 つのデータセットを out-of-doman データセット T として用いた。表 1、2に用いた各データセットとその学習、開発、評価データのサイズを示す。in-domain データセットの量による比較を行うために、6 つ全てを用いる S_6、

[1] https://s3-eu-west-1.amazonaws.com/fever.public/wiki-pages.zip

	事前学習データ	追加学習データ	in-domain		out-of-domain	
			F1	EM	F1	EM
学習：真のデータ S	-	S_1	80.1	66.7	46.9	36.4
		S_6	81.2	68.0	56.3	45.9
事前学習：生成データ G	$G_{\text{フィルタなし}}$	-	48.4	28.5	20.1	16.9
	$G_{F1 \geqq 0.2}$		77.1	62.9	44.8	39.0
	$G_{F1 \geqq 0.4}$		76.7	62.7	46.6	35.2
	$G_{F1 \geqq 0.6}$		75.5	61.5	45.3	33.8
	$G_{F1 \geqq 0.8}$		73.6	58.7	39.4	31.4
	$G_{F1=1}$		72.7	58.3	42.0	30.1
事前学習：生成データ G 追加学習：真のデータ S	$G_{\text{フィルタなし}}$	S_1	79.3	66.2	42.6	32.1
	$G_{F1 \geqq 0.2}$		**81.6**	**68.4**	51.2	40.1
	$G_{F1 \geqq 0.4}$		81.2	68.0	50.6	39.9
	$G_{F1 \geqq 0.6}$		81.3	68.1	50.9	41.9
	$G_{F1 \geqq 0.8}$		81.3	67.9	48.9	38.3
	$G_{F1=1}$		81.3	68.1	48.7	38.3
	$G_{\text{フィルタなし}}$	S_6	80.4	67.3	54.0	43.5
	$G_{F1 \geqq 0.2}$		80.4	67.3	54.0	43.5
	$G_{F1 \geqq 0.4}$		77.4	63.9	51.3	41.2
	$G_{F1 \geqq 0.6}$		81.4	68.2	56.5	45.9
	$G_{F1 \geqq 0.8}$		81.3	67.9	**56.6**	**46.0**
	$G_{F1=1}$		81.4	68.3	56.5	**46.0**

表 4: 評価データセットにおける機械読解モデルの精度。上段は真のデータ S のみで学習したモデル、中段は生成データ G で事前学習したモデル、下段は生成データ G で事前学習し、真のデータ S で追加学習したモデルを表す。太字は各評価指標で最も精度が高いモデルを表す。

SQuAD 1 つに限定したものを S_1 の 2 つの条件で学習を行った。

回答抽出モデルおよび機械読解モデルには $\text{BERT}_{\text{base}}$[2]、質問生成モデルには CopyNet+ [7] を用いた。各モデルのハイパーパラメータは各論文に記載の値を用いた。計算の効率化のため、回答抽出モデルおよび質問生成モデルの学習には S_1 を用いた。

機械読解モデル、回答抽出モデル、質問生成モデルのチューニングには in-domain の開発データを用いた。機械読解モデルの事前学習には 1 エポック、追加学習には 1 エポック学習を行った。また、計算には産総研の AI 橋渡しクラウド (ABCI) を利用した。

3章で説明した手順を用いて CQA トリプレット G を作成した。フィルタリングに用いた F1 のしきい値を変化させた時の生成データサイズを表 3に示す。

評価にあたっては予測回答と正解回答の文字ベースのオーバーラップを F1 と EM(Exact Match）の 2 つの指標を用いた。また、複数のデータセットを用いて評価した場合はマクロ平均した値を用いた。

5 精度評価および分析

生成データ G による事前学習した機械読解モデル、さらに真のデータ S による追加学習した機械読解モデルの精度の評価結果を表 4に示す。

5.1 生成データ G による事前学習

表 4の中段に生成データ G による事前学習したモデルの精度を示す。$G_{\text{フィルタなし}}$ と比較して、フィルタによって out-of-domain の F1 は最大 26.5pt、EM は 22.1pt 向上しており、汎用性向上には生成データの質が重要なことがわかる。また、真のデータ S_1 のみで学習したモデル (表 4の 1 行目）と比較すると、$G_{F1 \geqq 0.2}$ や $G_{F1 \geqq 0.4}$ で学習したモデルは out-of-domain の精度でおおよそ同程度の精度を達成している。また、表 1から真のデータ S_1 のデータサイズは約 7.5 万であり、表 3から $G_{F1 \geqq 0.2}$ や $G_{F1 \geqq 0.4}$ のデータサイズは 600-800 万程度である。これらの結果から、真のデータ S と比較して、生成データ G はおおおよそ 1/108 から 1/78 の精度向上効果があることが推定できる。

5.2 真のデータ S による追加学習

表 4の下段に真のデータ S による追加学習したモデルの精度を示す。$G_{F1 \geqq 0.4}$ で事前学習し、S_6 で追加学習したモデルが out-of-odmain の精度が最も高くなっており、F1 で 56.6、EM で 46.0 を達成している。真のデータ S_6 のみで学習したモデル（2 行目）と比較してみると、生成データ G の事前学習による精度の向上は F1 で 0.3pt、EM で 0.1pt の向上にとどまっている。一方で、6 つの in-domain のデータセットのうち 1 つだけを真のデータとして追加学習した場合 (S_1)

には、out-of-odmain の精度は生成データ G の事前学習によって、F1 で 4.2pt、EM で 5.5pt の改善が見られた。したがって、学習データが十分なドメインをカバーしておらず、量も十分ではない場合には、自動生成による学習データの拡張には汎用性の向上効果があることがわかる。

in-domain の精度を着目してみると、生成データ $G_{F1 \geqq 0.2}$ で事前学習し、真のデータ S_1 で追加学習したモデルが最も精度が高く、真のデータ S_1 だけで学習したモデルよりも F1 で 1.5pt、EM で 1.7pt 向上している。in-domain と out-of-domain の精度の改善を比較すると、out-of-domain の方が生成データ G による事前学習の効果は大きいことがわかる。

5.3 生成データ G の量と質

生成データ G の量と質がモデルの精度に与える影響を調べるために、生成データ G を 10%, 50%にランダムサンプリングしたものを用いて事前学習を行った。精度評価結果を図 1を示す。$G_{フィルタなし}$ ではデータサイズが小さくなるほど、精度が向上している。これは、フィルタなしでは低品質な生成データが多いため、モデルの精度を低下させているためであると考えられる。一方で、フィルタリングを行ったモデルでは、最適なサイズとフィルタリング条件（品質）のバランスが存在することがわかる。$G_{F1 \geqq 0.2}$ を 50%にサンプリングしたもので学習したモデルの精度が最も高く、真のデータ S_1 のみを使って学習したモデルよりも F1 が 4.9pt 向上している.

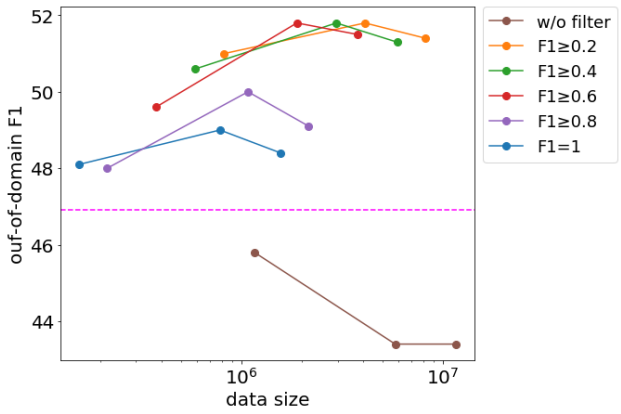

図 1: 事前学習に用いる生成データ G のデータサイズを変化させた時の out-of-domain の F1。凡例は生成データのフィルタ条件を表す。点線は 真のデータ S_1 のみで学習したモデルの精度を表す。

6 おわりに

機械読解モデルの汎用性を向上するために、回答抽出モデルと質問生成モデルを用いて CQA トリプレットを生成することで学習データを拡張する手法について検討した。実験の結果から、in-domain の学習データが十分にある場合にはデータ拡張の効果はほとんどないが、in-domain の学習データが少ない場合にはデータ拡張によって out-of-domain における F1 が 4.2pt、EM が 5.5pt 向上することを示した。また、生成した CQA トリプレットの量と質が機械読解モデルの out-of-domain での精度に与える影響を調査し、影響を明らかにした。

参考文献

[1] Jacob Devlin. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. https://nlp.stanford.edu/seminar/details/jdevlin.pdf, 2019.

[2] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pp. 4171–4186, Minneapolis, Minnesota, June 2019. Association for Computational Linguistics.

[3] Li Dong, Nan Yang, Wenhui Wang, Furu Wei, Xiaodong Liu, Yu Wang, Jianfeng Gao, Ming Zhou, and Hsiao-Wuen Hon. Unified language model pre-training for natural language understanding and generation. In *33rd Conference on Neural Information Processing Systems (NeurIPS 2019)*, 2019.

[4] Adam Fisch, Alon Talmor, Robin Jia, Minjoon Seo, Eunsol Choi, and Danqi Chen. MRQA 2019 shared task: Evaluating generalization in reading comprehension. In *Proceedings of the 2nd Workshop on Machine Reading for Question Answering*, pp. 1–13, Hong Kong, China, November 2019. Association for Computational Linguistics.

[5] Patrick Lewis, Ludovic Denoyer, and Sebastian Riedel. Unsupervised question answering by cloze translation. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pp. 4896–4910, Florence, Italy, July 2019. Association for Computational Linguistics.

[6] Alon Talmor and Jonathan Berant. MultiQA: An empirical investigation of generalization and transfer in reading comprehension. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pp. 4911–4921, Florence, Italy, July 2019. Association for Computational Linguistics.

[7] Qingyu Zhou, Nan Yang, Furu Wei, Chuanqi Tan, Hangbo Bao, and Ming Zhou. Neural Question Generation from Text: A Preliminary Study. *arXiv:1704.01792 [cs]*, April 2017. arXiv: 1704.01792.