JointQA モデルによる質問応答精度向上の日本語データセット JAQKET への適用検討

竹下恭史朗
早稲田大学大学院情報生産システム研究科
tkstria@ruri.waseda.jp

1 はじめに

質問応答 (Question answering または Answer generation) タスクとは、与えられた文書などの知識をもとに、入力された質問や読解問題に正しく応答することを目指すタスクである。一方、問題生成 (Question generation) タスクとは、入力された言葉を解答とする問題を出力することを目指すタスクである。これらは問題、解答、背景知識の 3 つの要素で構成されており、2 つの要素から残り 1 つの要素を導くことができる。例えば質問応答においては、問題に対して背景知識を元に解答を導く。問題生成においては、解答になるべき語句に対して背景知識を用いて問題文を生成する。ここで、背景知識とは問題と解答の関係性を示す文書や、多くの文書から学習した単語の数学的表現 (word2vec など) が挙げられる。これらのタスクには多くのモデルが用いられ、これまでに多様な改善が試みられ、優れたモデルは社会に応用されてきた。

質問応答と問題生成は互いにその質を補完する関係が知られており、同一のモデルで質問応答と問題生成の同時に訓練したモデル (これを JointQA モデルと呼ぶ) は、個々のタスクに特化したモデルよりも教師データとの一致が高いという報告が見られる。そこで、本研究では、2020 年に発表された日本語の質問応答データセット JAQKET[12] を用いて、質問応答と問題生成の質の相関を観察した。

2 関連研究

WikiQA[11] や SQuAD[6] に代表される質問応答データセットを用いた研究として、問題から解答を導く質問応答のみならず、解答からそれを問う問題を作る問題生成が知られている。そしてこの問題生成は質問応答よりも長い単語列を出力しなくてはならないためタスクとして難しく、また、受け入れられうる問題が教師データとして与えられているもの以外にも多様に考えられるなど、多くの課題を抱えている。

ところで、質問応答・問題生成に関する興味深い事象として、JointQA モデルによる質問応答、問題生成の質の向上がある。JointQA モデルとは、質問応答・問題生成の双方を同時に訓練した深層学習モデルのことで、JointQA モデルの学習によって、単に一つのタスクのみを学習したモデルよりも精度の高い出力が得られるというものがある。

JointQA モデルによる質問応答の精度の向上は多くのデータセットで確認されており [8]、問題生成の訓練が、質問応答をする上で有効な Attention 機構の作用に寄与することが原因と見られている。これまでそういった JointQA モデルの研究の題材となっていたのは英語で書かれたデータセットによるものが多かった。

しかし、2020 年に鈴木らにより日本語で書かれた質問応答データセット JAQKET がリリースされ、構造化された日本語の質問応答データセットの供給により、日本語の問題生成や質問応答の研究が行えるようになった。

これまでに行われた質問応答と問題生成の精度の相関を見る研究は、英語で書かれたデータセットのものが多く、日本語での問題生成に着目した研究は見られていない。

3 研究目的

本研究では、JAQKET で与えられた問題文、解答、解答についての Wikipedia 記事、これら三つの要素を用いて、質問応答と問題生成をさまざまな条件で行い、JAQKET を質問応答のみならず問題生成についても活用することを試みる。

もっとも注目する点として、他のデータセットに見られていた JointQA モデル 1 による質問応答の一

_『 騎 士 団 長 殺 し 』『 1 Q 8 4 』『 ノ ル ウェイ の森 』 といった 小説 の作者 は誰 で しょう ？ _1 _村上 _春 樹 （ む ら か み _ はる き 、194 9 年 1 月 12 日 _- _ ）は 、 日本の 小説家 、 文学 翻 訳 家 。 京都府 京都 市 伏 見 区 に 生まれ 、 兵庫県 西 宮 市 ・ 芦 屋 市 に 育つ 。 _ 早 稲 田 大学 在 学 中に 喫 茶 を 開く 。……

図 1　入力する問題と文書を連結したもの
シアン=問題、黄=弁別のための値、緑=文書

_村上 _春 樹 _0 _村上 _春 樹 （ む ら か み _ はる き 、194 9 年 1 月 12 日 _- _ ）は 、 日本の 小説家 、 文学 翻 訳 家 。 京都府 京都 市 伏 見 区 に 生まれ 、 兵庫県 西 宮 市 ・ 芦 屋 市 に 育つ 。 _ 早 稲 田 大学 在 学 中 に 喫 茶 を 開く 。……

図 2　入力する解答と文書を連結したもの
マゼンタ=解答、黄=弁別のための値、緑=文書

致度の向上が、JAQKET においても見られるかを観察する。

4　実験

入力として、質問応答タスクの場合は問題と解答に対する情報を含んだ文書を入力し、深層学習モデルによって解答を出力することを目指す。問題生成タスクの場合は解答とその解答に関する情報を含んだ文書を入力し、教師データに近い問題を生成することを目指す。

4.1　入出力データ

データ資源として、2020 年に鈴木らが発表した JAQKET を用いる。JAQKET は各タプルに ID、問題、解答、真の解答と 19 のダミーを含んだ選択肢、正規化 (全角文字を半角にするなど) 前の問題文・解答を含んだデータセットで、別のデータが、各選択肢についての Wikipedia 記事全文を与えている。実際の JAQKET の一部を図 1 に示す。訓練用データとして 13,061、開発用データとして 997、評価用データとして 995 の問題－解答対がある。今回の研究では、JAQKET から問題と解答と Wikipedia 記事を sentencepiece で分かち書きしたものの先頭 500 サブワードを抽出し、質問応答の場合は問題文に、問題生成の場合は解答の語句にそれを連結した。この際、各タスクのどちらを行うかの弁別のために、問題あるいは解答と記事の間に、問題生成の場合は「0」を、質問応答の場合は「1」を付与し、モデルがどちらの学習を行っているか把握する上での助けとした。(図 1,2)

表 1.JAQKET に含まれるデータの一例

ID	ABC03-02-0319
問題	漫画『サザエさん』で、マスオさんの出身大学はどこでしょう?
解答	早稲田大学
20 個の選択肢	早稲田大学、拓殖大学、慶應義塾大学、神奈川大学、成城大学、一橋大学……
正規化前の問題	漫画『サザエさん』で、マスオさんの出身大学はどこでしょう?
正規化前の解答	早稲田大学

実験の条件として、解答と文書から問題 (q-gen のみ)、問題と文書から解答を導く (a-gen のみ) ものに加え、JointQA による精度の改善を観測するために、全データを半分に分けてそれぞれ a-gen と q-gen を一つのモデルで同時に行う a-gen ∨ q-gen、全データを a-gen にも q-gen にも用いる a-gen ∧ q-gen、これら四つの条件で実験を行った。各条件の入力における、訓練データ、開発データの数を表 2 に示す。

表 2. 各条件における入力

訓練 / 開発	質問応答	問題生成
a-gen のみ	13,061 / 997	0 / 0
q-gen のみ	0 / 0	13,061 / 997
a-gen ∨ q-gen	6,530 / 498	6,531 / 497
a-gen ∧ q-gen	13,061 / 997	13,061 / 997

このとき、a-gen ∨ q-gen は一つの問題－解答対は質問応答と問題生成のどちらかにしか使わない一方、a-gen ∧ q-gen は一つの問題－解答対を質問応答と問題生成の双方に使うことに注意されたい。

4.2　実験条件

先述の入出力を扱うため、ニューラル機械翻訳に用いられるフレームワークである OpenNMT[3] を用いて深層学習モデルを構築した。

具体的には、モデルとして LSTM を用い、SQuAD を用いた過去の研究および OpenNMT が推奨する設定を参考に、学習係数は 1.0、バッチサイズは 32、入出力次元は 500、ドロップアウトは 0.3 とした。最適化関数として、Adadelta を用いた。

4.3 評価の手法

既存の質問応答・問題生成の研究においては、データセットごとに推奨される評価の手法があり、たとえば SQuAD の場合は a-gen の場合 F1 スコアと EM(Exact match)、q-gen の場合 BLEU[5] を用いている。今回は a-gen と q-gen の評価について、SQuAD に倣い a-gen は EM と F1 スコア、q-gen は BLEU を用いた。

5 結果

各条件の a-gen、q-gen の結果を表 3 に示す。

表 3. 各条件で訓練したモデルの出力の評価

入力条件	a-gen(EM)	a-gen(F1)	q-gen(BLEU)
a-gen のみ	37.81	75.00	0.0
q-gen のみ	0	0	3.9
a-gen ∨ q-gen	11.33	32.68	2.2
a-gen ∧ q-gen	52.76	75.00	2.8

結果から、a-gen(EM) においては、a-gen のみの入力よりも a-gen ∧ q-gen の入力のほうが大きい値を得ており、JointQA による精度の向上が見られた。一方、a-gen(F1) においては、a-gen のみも a-gen ∧ q-gen も同じ値を得た。出力の一例を付録中の表 5,6 に示す。

また、q-gen に関して、過去の研究で Sachan ら [8] がデータセットとして SQuAD を用いて報告した値を表 4 に示す。

表 4.SQuAD を用いた既存研究の q-gen の評価 [8] より引用

入力条件	q-gen(BLEU)
IR[7]	1.07
MOSES[4]	0.31
DirectIn	11.25
H&S[2]	11.23
Tang ら,2017[10]	5.03
Du ら,2017[1]	12.28
Ensemble[9]	14.37

IR は編集距離を元にした生成アルゴリズム、MOSES は統計的機械翻訳に用いられるモデル、DirectIn は longest sub-sequence を問題として出力するモデル、H&S はあらかじめ設定されたルールベースで問題を多く生成し、ランキングするというモデルである。「Tang ら, 2017」は GRU を用いたニューラルネットワークを用いたモデルで、「Du ら, 2017」は LSTM を用いたニューラルネットワークを用いたモデルである。

出力の例を付録の

6 考察

6.1 a-gen の評価の SQuAD との比較

a-gen に関して、a-gen のみでの訓練で 37.81 の EM、0.7500 の F1 スコアを得た。また、a-gen に加えて q-gen の訓練をすることで 52.76 の EM、0.7500 の F1 を得た。本研究で参考としている SQuAD の公式サイトでは、コンペティションの形式で各モデルの EM と F1 スコアを載せるリーダーボードを公開している[1)]。そこでは末席である 63 位のモデル (匿名) でも、EM は 53.698 を記録しており、当研究で得られた最大の EM である 52.76 を上回っている。一方で、当研究で記録した F1 スコアの 75.00 に関しては、リーダーボードの 55 位と 56 位の間に位置している。なお、57 位には、当研究でも実装した LSTM を含む、CNN モデルが入っている。

低いスコアを記録した一因として JAQKET が SQuAD に比べて新しく、探索がされていないデータセットであるという以外にも、JAQKET が本来 20 の選択肢の中から正しい選択肢を選択する課題のためのデータセットとしてリリースされたため、問題から自由に解答を出力する形式のタスクとしては困難なことが考えられる。

6.2 q-gen の BLEU の低さ

q-gen に関して、本研究の BLEU を考察する。本研究では LSTM を用いているため、その他のハイパーパラメータの影響はあるにせよ、表 4 に挙げたなかでのモデルでは、LSTM を用いて SQuAD を処理した Du らの実装により近いといえるだろう。Du らの結果 12.28 に比べ、本研究では q-gen のみの訓練で 3.9、a-gen ∧ q-gen の条件で 2.8 とそれを大きく下回る結果となった。

原因として考えられるのは、一つは単にハイパーパラメータやモデルの選択に余地がある可能性である。今回の学習では、特に q-gen, a-gen ∨ q-gen, a-gen ∧ q-gen の学習において過学習の傾向がある。より適切に実装を行うことで、精度が改善する可能性がある。

他にもデータセットの特徴が影響している可

1) https://rajpurkar.github.io/SQuAD-explorer/

能性がある。例えば SQuAD は英語で書かれた、Wikipedia の一段落に基づいて作成された問題－解答対を含むデータセットである。対して JAQKET は日本語で書かれた、あらかじめ存在したクイズ問題文に Wikipedia の全文を文書として与え、それが情報として十分であることが保証されていないデータセットである。

6.3 a-gen ∨ q-gen における値の低さ

a-gen ∨ q-gen の入力条件について、単一のタスクを訓練したモデルや a-gen ∧ q-gen を訓練したモデルよりも、各タスクにおいて低い数値を得ることがわかった。

この研究において、a-gen ∨ q-gen は一つのデータセットを半分に分け、片方を質問応答、もう片方を問題生成のみに用いる入力である。つまり、質問応答だけに使われる問題と、問題生成だけに使われる解答が存在することになる。一方、a-gen ∧ q-gen はデータセットに含まれるデータすべてを質問応答にも問題生成にも用いる入力である。つまり、すべての問題・解答が、質問応答にも問題生成にも用いられることになる。向上が a-gen ∧ q-gen だけに見られたのだとすれば、一つの問題－解答対の両方を学習に用いることが必要になるということである。

JointQA による精度の向上のメカニズムは未だ議論されているが、一説には文書への attention 機構がより効果的に訓練されることが述べられている。本研究での a-gen ∨ q-gen で向上が起こらなかったのには、ある一つの文書を質問応答・問題生成のどちらかにしか使わず、文書への attention への訓練を、質問応答と問題生成の両方で訓練しなかったことが一つの原因として考えられる。

6.4 a-gen ∧ q-gen における a-gen のみとの比較

a-gen ∧ q-gen について、a-gen(EM) については 52.76、a-gen(F1) については 75.00 の値を得た。a-gen のみと比較して、F1 スコアは同程度のものを得たが、EM については、q-gen の訓練を行うことで 37.81 から 52.76 への向上を得た。q-gen については、q-gen のみの訓練が 3.9 に対して、a-gen ∧ q-gen は 2.8 を得たため、q-gen については向上を得られなかった。また、表 4 における Du らの結果は LSTM を用いたモデルであるから、本研究で実装したものと近いものであるが、それは SQuAD においては 12.28 の BLEU を記録しており、それに比べると低い値を得たことがわかった。これは SQuAD における q-gen のみの訓練をしたモデルの結果だが、これが本研究と同じ LSTM を用いて実装されたものであることを考えると、その他のハイパーパラメータの影響を無視すれば、この差はデータセットによるものだと考えられる。

7 おわりに

今回の研究では JAQKET を用いて質問応答・問題生成、およびそれらの相互関係について踏み込む研究を行った。今後の課題として、問題生成を SQuAD などの既存のデータセットを用いた研究と同程度の精度で出力を得ること、単に正解データとの一致度だけでなく、問題生成の解答可能性や多様性を評価することなどが挙げられる。

参考文献

[1]Xinya Du, Junru Shao, and Claire Cardie. Learning to ask: Neural question generation for reading comprehension, 2017.

[2]Michael Heilman. *Automatic Factual Question Generation from Text*. PhD thesis, USA, 2011.

[3]Guillaume Klein, Yoon Kim, Yuntian Deng, Jean Senellart, and Alexander Rush. OpenNMT: Open-source toolkit for neural machine translation. In *Proceedings of ACL 2017, System Demonstrations*, pp. 67–72, Vancouver, Canada, July 2017. Association for Computational Linguistics.

[4]Philipp Koehn, Hieu Hoang, Alexandra Birch, Chris Callison-Burch, Marcello Federico, Nicola Bertoldi, Brooke Cowan, Wade Shen, Christine Moran, Richard Zens, Chris Dyer, Ondřej Bojar, Alexandra Constantin, and Evan Herbst. Moses: Open source toolkit for statistical machine translation. In *Proceedings of the 45th Annual Meeting of the Association for Computational Linguistics Companion Volume Proceedings of the Demo and Poster Sessions*, pp. 177–180, Prague, Czech Republic, June 2007. Association for Computational Linguistics.

[5]Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. Bleu: a method for automatic evaluation of machine translation. In *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics*, pp. 311–318, Philadelphia, Pennsylvania, USA, July 2002. Association for Computational Linguistics.

[6]Pranav Rajpurkar, Jian Zhang, Konstantin Lopyrev, and Percy Liang. SQuAD: 100,000+ questions for machine comprehension of text. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pp. 2383–2392, Austin, Texas, November 2016. Association for Computational Linguistics.

[7]Alexander M. Rush, Sumit Chopra, and Jason Weston. A neural attention model for abstractive sentence summarization. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*, pp. 379–389, Lisbon, Portugal, September 2015. Association for Computational Linguistics.

[8]Mrinmaya Sachan and Eric Xing. Self-training for jointly learning to ask and answer questions. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, pp. 629–640, 2018.

[9]Linfeng Song, Zhiguo Wang, Wael Hamza, Yue Zhang, and Daniel Gildea. Leveraging context information for natural question generation. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 2 (Short Papers)*, pp. 569–574, New Orleans, Louisiana, June 2018. Association for Computational Linguistics.

[10]Duyu Tang, Nan Duan, Tao Qin, Zhao Yan, and Ming Zhou. Question answering and question generation as dual tasks, 2017.

[11]Yi Yang, Wen-tau Yih, and Christopher Meek. WikiQA: A challenge dataset for open-domain question answering. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*, pp. 2013–2018, Lisbon, Portugal, September 2015. Association for Computational Linguistics.

[12]鈴木正敏, 鈴木潤, 松田耕史, 西田京介, 井之上直也. JAQKET: クイズを題材にした日本語 QA データセットの構築. 言語処理学会第 26 回年次大会 (NLP2020), 2020.

A 付録

A.1 質問応答・問題生成出力の例と正解データとの比較

各入力条件の出力結果と、目標となる正解データを表5、表6に示す。

表5. 質問応答の出力の例

入力	正解	a-gen	q-gen	a-gen ∨ q-gen	a-gen ∧ q-gen
__和名をハダカカメガイといい、実は巻き貝の一種とされている、その姿から「流氷の天使」と呼ばれる動物は何でしょう？	__クリオネ	__クリオネ	__和名を「ハルボット」という、バスケットボールで、グリーン球に投与するのはどれでしょう？	__クリオ	__クリオネ
__作家のルスティケロが、マルコ・ポーロから聞いた話をまとめた作品といえば何でしょう？	__東方見聞録	__東方見聞録	__オペラ『ダメル』や『オボット』などの作品が収められている、アメリカの作家は誰でしょう？	__ウァギズ	__東方見聞録
__『騎士団長殺し』『1Q84』『ノルウェイの森』といった小説の作者は誰でしょう？	__村上春樹	__村上春樹	__主人公・黒澤明の監督である女優は誰でしょう？	__森田哲樹	__村上春樹

表6. 問題生成の出力の例

入力	正解	a-gen	q-gen	a-gen ∨ q-gen	a-gen ∧ q-gen
__クリオネ	__和名をハダカカメガイといい、実は巻き貝の一種とされている、その姿から「流氷の天使」と呼ばれる動物は何でしょう？	(空出力)	__和名を「ハエカカメガイ」という、その姿に含まれる曲は何でしょう？	__和名を「ハカカカイイ」という、その姿から「イカカカカカカカ」という名前は何でしょう？	__和名で「ハダカメガイ」という意味がある、モイドカシイとイイ科の果物は何でしょう？
__東方見聞録	__作家のルスティケロが、マルコ・ポーロから聞いた話をまとめた作品といえば何でしょう？	__アッキーシ	__正式名称を『世界の記述』という、代表作に『ユメン』シリーズの舞台は誰でしょう？	__正式名称を『世界の記述』という、ルス・フォンの短編小説は何でしょう？	__ノコ・ポーロ
__村上春樹	__『騎士団長殺し』『1Q84』『ノルウェイの森』といった小説の作者は誰でしょう？	__東京山樹	__昨年2月20日に86歳で亡くなった、『楡家の人』や『西洋の物語』などで知られる作家は誰でしょう？	__昨年8月24日に亡くなった、『カガラの女人』などで知られる日本の作家は誰でしょう？	__今年1月にシングル『ノルン・カフカミー』でデビューした、日本の詩人は誰でしょう？