読解による解答可能性を付与した質問応答データセットの構築

鈴木 正敏 † 松田 耕史 † 岡崎 直観 ‡ 乾 健太郎 † §
† 東北大学 ‡ 東京工業大学 § 理化学研究所 AIP センター
{m.suzuki,matsuda,inui}@ecei.tohoku.ac.jp okazaki@c.titech.ac.jp

1 はじめに

読解タスクは、与えられる質問と文書の組に対して、文書の内容を元に質問に答えるタスクであり、システムの自然言語理解の能力を試す問題として研究が行われている。タスクの解法として、ニューラルネットワークを用いた読解モデルがこれまで数多く提案されており、最近では、従来の質問応答システムにおける、検索された文書から答えを抽出するという部分に読解のモデルを組み入れる試みも登場している [1]。

既存の読解タスクのモデル、およびデータセットのほとんどは、質問に付随する文書の中に答えがあり、文書の読解によって正解を抽出できることを前提としている。しかしながら、応用システムへの適用を考えると、この前提は必ずしも適当でない。例えば、質問応答システムにおける解答抽出への応用を考えると、読解の前段階の文書検索によって取得される文書に、必ず質問の正解が含まれているという保証はない。また、正解が含まれていても、質問の内容とは無関係の読解に適さないような文書が検索される場合もある。このような条件下では、文書から闇雲に答えを探すのではなく、読解によって答えを求められない文書に対しては「解答不可能」という出力が可能なモデルが必要になる。そのような読解のモデルを訓練するためには、一つ一つの質問・正解・文書の組に対し、読解によって解答できるかどうかの情報が付与されたデータが必要になるが、そのような大規模なデータセットはこれまでのところ存在しない。

本研究では、およそ 55000 件の質問・正解・文書の組に対して、文書に質問の正解の根拠が書かれているかどうかの人手による判断を、クラウドソーシングで集めることによって、読解による解答可能性が付与された読解タスクのデータセットを初めて作成した。さらに、既存の読解モデルを用いて実験を行い、答えられない文書が存在する条件下でのモデルの性能を調査した。なお、作成した読解データセットは研究利用が可能な形式で公開する予定である[1]。

[1] http://www.cl.ecei.tohoku.ac.jp/rcqa/

2 関連研究

読解タスクのためのデータセットとして、質問が穴埋め形式のもの [2, 3]、質問が択一式のもの [8, 5]、質問の解答を文書中から抜き出すもの [4, 7, 10] など様々な種類のものが提案されている。本節では、質問の解答を文書中から抜き出す形式のデータセットについて関連研究を述べる。

TriviaQA [4] は、Web から収集したクイズの質問文と正解のペアに対して、質問に関連する Web ページと Wikipedia 記事を自動で付与することで作られた、およそ 65000 件の質問からなる読解データセットである。TriviaQA は、データセット作成とは無関係に人手で作られたクイズ問題を質問に利用しているため、質問の内容の多様性は高い。その一方で、各質問に対して付与された文書は機械的に取得されたものであり、読解によって解答が可能な文書とそうでない文書が混在している[2]。

SQuAD [7] は、Wikipedia の記事の内容に対して、クラウドソーシングによって質問文を作成し、正解とともに付与することで作られた、およそ 10 万件の質問からなる読解データセットである。SQuAD では、与えられた文書に対して人間が質問を作っているので、読解によって解答できる問題になっていると考えられる。一方、文書の内容をもとに作られた問題文には、“What indivisual the school named after?”（Harvard University の記事についての質問）のように、質問と解答が特定の文書の内容に依存しているものも一部存在する。

NewsQA [10] は、ニュース記事の内容からクラウドソーシングによって質問と正解を作成することで作られた、およそ 12 万件の質問からなる読解データセットである。クラウドソーシングでは、1 つのグループがニュース記事の見出しと要約だけを読んで質問を作り、別のグループが記事の全文を見て、質問の正解を与えるという作業形態をとっている。そのため、要約から作られた質問の答えが記事本文中に存在しない場合もあり、その場合には、答えがないことを示す記号を正解として与えて

[2] TriviaQA の著者は、質問に付与された文書は、distant supervison に利用できるデータであると位置付けている。

いる。NewsQA は、解答が不可能な質問が明示的に含まれているという点で、本研究で作成したデータセットと近いが、NewsQA は SQuAD と同様、文書が与えられた上で質問が作られているため、一部の質問の内容が文書に依存している。

本研究と最も近い既存研究として、質問応答における文選択のデータセットである WikiQA [11] がある。WikiQA は、検索エンジンのクエリログとクリックされた Wikiepedia 記事から質問文と文のペアを作成し、文の内容が質問の答えとなっているかどうかをクラウドソーシングによってアノテーションしたデータセットである。本研究との違いとしては、WikiQA の質問は検索クエリであるために、自然言語による質問文となっていないこと、文選択タスクのデータとして作られたため質問と文書の長さが短く読解タスクに使うデータとしては必ずしも適していないことが挙げられる。

3 データセット

本節では、読解による解答可能性を付与した質問応答データセットの作成方法と内容について述べる。

3.1 問題文 q と正解 a

読解タスクに利用するデータとしては、問題の内容に多様性があり、なおかつ問題文が単体で意味をなし、特定の文書の内容に依存しないことが望ましい。本研究では、既存研究 [4] に倣い、既存のクイズの問題集に記載されている問題を利用し、データセットを作成した。

Web サイト『クイズの杜』[3] と『abc/EQIDEN 公式サイト』[4] より、早押しクイズの大会「abc」および「EQIDEN」で 2003 年から 2010 年の間に使用された、全ての問題文 q と正解 a のペアを収集した。正誤表[5]に基づき一部の問題の訂正を行った結果、問題 (q, a) の数は 12591 となった。

収集した早押しクイズの問題には、以下のような特徴がある。

- 問題文は 100 文字程度の疑問文であり、答えがただ一つに決まるように作られている（答えが複数あるいわゆる「多答問題」は含まれていない）。
- 正解は、固有表現の他にも、ことわざや四字熟語のような固有表現以外のものも含まれる。

3.2 文書 d

各問題 (q, a) について、a を部分文字列として含む文書を付与する（a を含まない文書は、読解によって解答不可能であるから、後述のクラウドソーシングで解答可能性を問う必要がないため、最初から除外する）。

[3] http://quiznomori.web.fc2.com/
[4] http://abc-dive.com/
[5] http://abc-dive.com/questions/errata.html

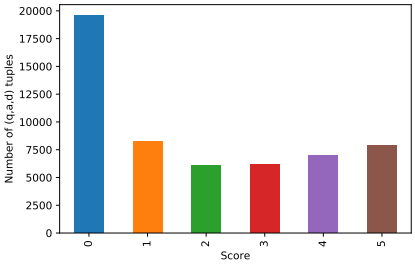

図 1: 解答可能性スコア s の分布

2017 年 7 月 11 日時点の日本語版 Wikipedia の全記事の本文を、Wikipedia API で取得し、段落ごとに区切り、Elasticsearch[6] (ver.6.1.0) を用いて Wikipedia 記事段落の全文検索エンジンを作成した。それぞれの (q, a) に対して、作成した検索エンジンに q をクエリとして検索を実行し、検索結果の中から a との部分一致がある上位最大 5 件の Wikipedia 記事段落 d を取得した。ただし、d の文長が 500 文字を超える場合は、最初の 500 文字のみを抜粋し付与した。これにより、 計 54958 件の、問題文 q、正解 a、文書 d の 3 つ組 (q, a, d) を作成した。

3.3 解答可能性スコア s の付与

文書 d は、全文検索エンジンを用いて機械的に取得されたものであるため、人間にとっても計算機にとっても、読解によって正解 a を導けるとは限らない。そこで、それぞれの (q, a, d) に対して、d を読むことで q に対する a を求めることができるかどうかの情報を、クラウドソーシングによって付与した。

ワーカーには、(q, a, d) を提示し、文書 d に q の答え a の十分な根拠が書かれているかどうかを質問し、「書かれている」「書かれていない」の 2 択で答えてもらった。1 つの (q, a, d) の組に対して、5 人に同じ質問を行い、「書かれている」と答えた人数を、その問題の解答可能性のスコア $s \in [0, 5]$ とした。

クラウドソーシングのプラットフォームには『Yahoo! クラウドソーシング』[7]を用いた。データ作成の品質保証のため、1 回の作業にチェック設問を 4 問導入し、チェック設問に対する誤答がある作業結果を認めないようにした。

3.4 データの分析

作成した読解データセットの事例を表 1 に示す。表 1 の上半分の 2 例のように、同じ問題 (q, a) に対しても文書 d に書かれている内容によって、解答可能性は変わる。表 1 の下半分の 2 例は、ワーカーによって判断が分かれた例である。「ラムサール条約」の例のように、解答の根拠となる情報が括弧で表記されていて明示的でなかっ

[6] https://www.elastic.co/jp/products/elasticsearch
[7] https://crowdsourcing.yahoo.co.jp/

表 1: 作成したデータセットの事例

問題文 q	正解 a	文書 d	スコア s
フィギュアスケートのジャンプの 1 つ「アクセル」に名を残すアクセル・パウルゼンはどこの国の人でしょう? (出題日: 2004/03/21)	ノルウェー	[記事名: アクセルジャンプ] 1882 年のウィーンで開かれた国際大会 (Great International Skating Tournament) で**ノルウェー**のアクセル・パウルゼンが初めて跳んだのが始まりとされている。...	5
フィギュアスケートのジャンプの 1 つ「アクセル」に名を残すアクセル・パウルゼンはどこの国の人でしょう? (出題日: 2004/03/21)	ノルウェー	[記事名: アンネ・リネ・ヤシェム] アンネ・リネ・ヤシェム (, 1994 年 1 月 6 日-) は、**ノルウェー**出身の女性フィギュアスケート選手 (女子シングル)。双子の姉妹、カミラ・ヤシェムもフィギュアスケート選手である。	0
正式名を「特に水鳥の住処として国際的に重要な湿地とそこにいる動植物を保護するための条約」という条約を、イランの都市の名前をとって何というでしょう? (出題日: 2004/03/21)	ラムサール条約	[記事名: マッツァル国立公園] 1976 年、マッツァルは「特に水鳥の生息地として国際的に重要な湿地に関する条約」(**ラムサール条約**) 登録湿地に加えられた	2
かつての名前を「クリスチャニア」といった、北欧の国・ノルウェーの首都はどこでしょう? (出題日: 2008/03/23)	オスロ	[記事名: ステムターン] シュテムはドイツ語で、「制動」という意味である。ステムクリスティーのクリスティーはクリスチャニアの略で、この技術が始まったノルウェーの**オスロ**市の旧称であるクリスチャニアから来ている。	2

表 2: 作成したデータセットの定量的性質

	異なり数	平均文字数
問題 q	12063	47.24
正解 a	9001	4.42
文書 d	46394	165.04

表 3: 分割されたデータセットのデータ数

	全件	$s \geq 2$	$s < 2$
訓練データ	42371	21692	20679
開発データ	4749	2264	2485
テストデータ	7838	3884	3954

たり、「オスロ」の例のような、解答の根拠となる複数の情報のどれが必須であるかが曖昧な場合に、ワーカーによる解答可能性の判断が揺れていると考えられる。

作成した読解データセットの定量的性質を表 2 に示す。クイズの問題集から収集した問題のうち、一部の問題については、正解 a の部分一致を含んだ文書が 1 件もなかったため、表 2 の問題文 q の異なり数は、当初収集した問題の数よりも小さくなっている。

計 54958 件の (q, a, d, s) の解答可能性スコア s の分布を図 1 に示す。答えの文字列 a が文書中にあるにも関わらず、全体のおよそ 3 分の 1 のデータは、解答可能性スコア s が 0 であった。

4 実験

作成したデータセットを用いて、解答可能な (s が大きい) 問題にどれだけ正しく答えられるか、および解答不可能な (s が小さい) 問題には「答えられない」ということをどれだけ正しく出力できるかを、既存の読解モデルをベースラインとして検証した。

4.1 モデル

実験に用いる読解モデルとして、SQuAD[7] に対して最も高い精度を示す読解モデルの一つである、BiDAF[9] を用いた。本研究では、BiDAF をベースに以下の 3 つのモデル・設定で実験した。

- BiDAF: 通常の BiDAF と同じであり、q と d の入力に対して d に含まれる a の開始位置と終了位置を予測し、最も確率が高いスパンの単語列を答えるモデルである。
- BiDAF-Label: BiDAF で a の開始位置と終了位置を予測する最終層を、以下の計算を行う層に置換し、d に含まれる各単語 d_t が a の一部である確率を出力する。

$$p_t = \sigma\left(\boldsymbol{w}_{(p)}^{\top}[G; M]_{:t}\right)$$

ここに、$[G; M]_{:t}$ は、BiDAF の元論文 [9] における G と M を列方向に連結した行列の t 列目である。
損失関数としては、以下の交差エントロピーを用いた。

$$L(\theta) = -\sum_{1}^{N}\sum_{t=1}^{T} y_t \log(p_t)$$

ここに、y_t は、単語 d_t が a の一部であれば 1、そうでなければ 0 である教師信号である。T は d の長さ (単語数)、N は訓練事例数である。出力する答えとしては、$p_t > 0.5$ となる単語 d_t の系列のうち、p_t の平均が最も大きいものを採用する。そのような系列がない場合には、「解答不可能」と出力する。

- BiDAF-Label-S: モデルは BiDAF-Label と同じであるが、解答可能性スコア s の小さい (解答不可能な) 問題に対しては、教師信号 y_t を全て 0 にしてモデルを訓練したモデルである。

これらの 3 つのモデルのうち、BiDAF-Label と BiDAF-Label-S の 2 つは、「解答不可能」という出力が可能なモデルである。

表 4: 解答可能性を区別しない場合の実験結果

	全件		$s \geq 2$	
	EM	F1	EM	F1
BiDAF	46.62	51.43	52.39	56.72
BiDAF-Label	45.40	50.45	51.03	56.32

表 5: 解答可能性を区別した場合の実験結果

	全件		$s \geq 2$		$s < 2$
	EM	F1	EM	F1	Acc.
BiDAF-Label	30.49	33.26	51.03	56.32	7.95
BiDAF-Label-S	53.25	54.99	33.16	36.47	75.31

4.2 実験設定

計 54958 件のデータを、クイズ問題が使用された年代によって訓練データ、開発データ、テストデータに分割した。分割されたデータセットの各データ数を表 3 に示す。単語ベクトルとして、日本語版 Wikipedia 本文全文から事前に訓練した Glove[6] による 100 次元の分散表現を用いた。各読解モデルは Chainer[8] で実装し、エポック数 30、ミニバッチサイズ 30 で各モデルの訓練を行った。BiDAF-Label-S のモデルでは、解答可能性スコアが $s < 2$ の問題を解答不可能な問題とした。テストデータに適用するモデルとして、エポック数 $\{15, 20, 25, 30\}$ のうち、最も開発データに対する EM が高かったエポックのモデルを採用した。モデルの評価尺度には、[7] と同様に、出力と a の完全一致の割合 (EM) と、出力された答えの単語列の、正解 a の単語列に対する適合率と再現率から求められる F1 スコアの平均 (F1) を用いた。

従来の読解タスクの設定と同じく、解答可能性スコア s に関わらず、入力された問題・文書ペア (q, d) に対して正解単語列 a を出力できれば正解とする場合と、解答可能 ($s \geq 2$) な問題・文書ペア (q, d) に対しては正解単語列 a を出力できれば正解とし、解答不可能 ($s < 2$) な問題・文書ペア (q, d) に対しては「解答不可能」と出力できれば正解とする場合の 2 通りについて実験を行った。

4.3 結果

解答可能性を区別しない場合の実験結果を表 4 に示す。BiDAF は従来の読解問題と同じ設定であるが、[9] で報告されている SQuAD に対する精度と比べて低く、本研究のデータセットが従来のものと比較して、簡単には解けないデータセットであることを示唆している。

解答可能性を区別した場合の実験結果を表 5 に示す。$s < 2$ の EM は解答不可能な問題に対して、正しく「解答不可能」と答えた事例の割合（再現率）である。解答不可能な問題に「解答不可能」と答えるように訓練した BiDAF-Label-S では、75.31%の再現率で、解答不可能な問題・文書ペアを識別できている。

[8] https://chainer.org/

5 おわりに

本研究では、およそ 55000 件の質問・正解・文書の組に対して、文書に質問の正解の根拠が書かれているかどうかの人手による判断を、クラウドソーシングで集めることによって、読解による解答可能性が付与された読解タスクのデータセット作成した。さらに、既存の読解モデルを用いて、答えられない文書が存在する条件下での読解性能を調査した。

情報検索システムと読解モデルを組み合わせて質問応答システムを構築した際に、読解モデルが入力される質問・文書ペアに対して解答可能性を判断できれば、検索された文書の選別に役立ち、質問応答の性能が向上すると考えられる。文書検索を伴う質問応答システムにおける、読解モデルの性能評価は今後の課題である。

謝辞

本研究の一部は、理化学研究所受託研究「実社会ビッグデータ利活用のためのデータ統合・解析技術の研究開発」の一環として行われた。また、本研究は JSPS 科研費 15H01702 の助成を受けたものである。

本研究で使用したクイズ問題は abc/EQIDEN 実行委員会様に研究目的での利用許可を頂きました。記して感謝いたします。

参考文献

[1] Danqi Chen, Adam Fisch, Jason Weston, and Antoine Bordes. Reading Wikipedia to Answer Open-Domain Questions. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 1870–1879, 2017.

[2] Karm Moritz Hermann, Tomáš Kočiský, Edward Grefenstette, Lasse Espeholt, Will Kay, Mustafa Suleyman, and Phil Blunsom. Teaching Machines to Read and Comprehend. In *Proceedings of the 28th International Conference on Neural Information Processing Systems*, pp. 1693–1701, 2015.

[3] Felix Hill, Antoine Bordes, Sumit Chopra, and Jason Weston. The Goldilocks Principle: Reading Children's Books with Explicit Memory Representations. In *Proceedings of the 4th International Conference on Learning Representations (ICLR 2016)*, 2016.

[4] Mandar Joshi, Eunsol Choi, Daniel S. Weld, and Luke Zettlemoyer. TriviaQA: A Large Scale Distantly Supervised Challenge Dataset for Reading Comprehension. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 1601–1611, 2017.

[5] Guokun Lai, Qizhe Xie, Hanxiao Liu, Yiming Yang, and Eduard Hovy. RACE: Large-scale ReAding Comprehension Dataset From Examinations. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pp. 785–794, 2017.

[6] Jeffrey Pennington, Richard Socher, and Christopher D Manning. GloVe: Global Vectors for Word Representation. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing*, pp. 1532–1543, 2014.

[7] 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*, p. 2383–2392, 2016.

[8] Matthew Richardson, Christopher J C Burges, and Erin Renshaw. MCTest: A Challenge Dataset for the Open-Domain Machine Comprehension of Text. In *Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing*, pp. 193–203, 2013.

[9] Minjoon Seo, Aniruddha Kembhavi, Ali Farhadi, and Hannaneh Hajishirzi. Bidirectional Attention Flow for Machine Comprehension. In *Proceedings of the 5th International Conference on Learning Representations (ICLR 2017)*, 11 2017.

[10] Adam Trischler, Tong Wang, Xingdi Yuan, Justin Harris, Alessandro Sordoni, Philip Bachman, and Kaheer Suleman. NewsQA: A Machine Comprehension Dataset. 11 2016.

[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, 2015.