B6IM2027

修士論文

情報検索と文章読解を組み合わせた質問応答システム

鈴木 正敏

2018 年 2 月 13 日

東北大学 大学院
情報科学研究科 システム情報科学専攻

本論文は東北大学 大学院情報科学研究科 システム情報科学専攻に
修士 (工学) 授与の要件として提出した修士論文である。

鈴木 正敏

審査委員：

乾 健太郎 教授　（主指導教員）
木下 哲男 教授
周 暁 教授
岡崎 直観 教授　（副指導教員, 東京工業大学）

情報検索と文章読解を組み合わせた質問応答システム*

鈴木 正敏

内容梗概

質問応答システムの一般的な枠組みは、質問文に対する情報検索によって取得された文書群から解答候補を抽出するというものである。一方で近年、質問と関連文書の組の入力に対して、文書の読解によって答えをを出力するという読解タスクが提案され、大規模なデータセットとニューラルネットワークを用いた読解モデルが数多く提案されている。読解タスクの既存のモデル、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、質問応答システムへの読解モデルの適用を考えた場合には、読解によって答えを求められない文書に対して「解答不可能」という出力が可能な読解モデル、および訓練データセットが必要になる。本研究では、56651 件の質問・正解・文書の組に対して、読解による解答可能性の情報を付与した、読解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ベースに、解答可能性を判別する読解モデルを構築し、答えられない文書が存在する条件下での読解性能を実験により検証した。その結果、解答可能性を判別するタスクの難しさを示唆する実験結果が得られた。

キーワード

読解, 質問応答システム, クラウドソーシング

*東北大学 大学院情報科学研究科 システム情報科学専攻 修士論文, B6IM2027, 2018 年 2 月 13 日.

目 次

図 目 次

表 目 次

1　はじめに

本研究では、自然言語処理における読解タスクを、質問応答システムの構築に応用することを考える。

質問応答システムは、「日本で最も高い山は何？」というような自然言語による質問文の入力に対して、適切な解答を出力するシステムである。質問応答システムの構築手法は、情報検索に基づくアプローチと、知識ベースを利用したアプローチの 2 つに大別される。情報検索に基づくアプローチでは、情報検索の手法を用いて、質問文から複数の関連文書を検索し、取得された文書の中から解答候補の文字列を抽出し順位付けを行うことで、最終的な解答を決定する。一方、知識ベースを利用したアプローチでは、質問文を意味解析して検索クエリに変換し、知識ベースに問い合わせることで解答を得る。本研究では、前者の情報検索に基づく質問応答システムにおいて、検索された文書の中から解答候補を抽出する部分を、読解タスクのモデルで置き換えられる可能性に着目する。

読解タスクは、与えられる質問と文章の組に対して、文章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、近年盛んに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、一部の読解データセットに対する実験では、人間の能力に迫る読解性能を示すモデルが提案され始めている[1]。

情報検索に基づく質問応答は、世界全体、より現実的には、用意された文書集合全体を知識源として、質問に対する解答を求める問題である。一方、読解タスクは、各質問に付随する単一の文書を知識源とした、限定された条件下での質問応答の問題であるという見方ができる。そして、この問題設定においては、人間に匹敵する解答の性能が実現されつつある。そこで、読解タスクのモデルを、情報検索に基づく質問応答システムに導入することができれば、より性能の高い質問応答システムを実現できる可能性がある。

しかし、読解タスクを質問応答に応用するには、解決すべき課題がある。既存

[1] 2018 年 2 月現在、読解タスクのデータセットの 1 つである SQuAD[1] のリーダーズボード (`https://rajpurkar.github.io/SQuAD-explorer/`) において、一部の評価指標で人間によるテスト結果を上回る性能を示す読解モデルが複数報告されている。

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。一方、情報検索に基づく質問応答システムでは、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、読解で正解を導けない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。そのような解答可能性を判別できる読解モデルを訓練できれば、質問応答システムで、複数の文書から読解結果を集約するときに、解答可能だった結果のみを候補することで、解答候補の純度が上がり、結果として質問応答システムの性能改善に繋がると考えられる。しかし、解答可能性を判別できる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、56651 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元に、解答可能性を判別できる読解モデルを実験的に作成し、答えられない文書が存在する条件下での、読解モデルの解答可能性の判別性能を調査した。その結果、解答可能性の判別を伴う読解タスクは、既存の最先端の読解モデルにおいても難しいタスクであることを示唆する結果が得られた。そして、解答可能性を判別できるモデルを組み入れた質問応答システムを構築し、質問応答システムにおける、解答可能性を判別する読解モデルの有効性について検証した。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[2]。

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

2　関連研究

2.1　読解データセット

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 [4, 5]、質問の解答を文書中から抜き出すもの [1, 6, 7] 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の作成方法について関連研究を述べる。

TriviaQA [6] 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。各質問に関連する文書として、Web ページは、質問文をクエリとした検索エンジンの検索結果を、Wikipedia 記事は、質問文に固有表現抽出を適用した結果から取得した記事を付与している。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[3]。

SQuAD [1] 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手で質問が作られているので、ほとんどの質問が、読解によって解答できるもの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質問には、“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 [7] 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

[3]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。

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。

本研究のデータ作成方法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 [8] 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。

2.2 検索と読解に基づく質問応答

情報検索と読解タスクを組み合わせることで質問応答システムを構築する研究には、Chen ら [9] による先行研究がある。Chen らは、TF-IDF と bigram hashing に基づく情報検索モデルと再帰的ニューラルネットワークに基づく読解モデルの組み合わせにより、質問文に対して検索された Wikipedia 記事に対して読解モデルを適用することで、質問に対する答えを出力する質問応答システムを構築した。また、複数の質問応答のデータセットから、読解データセットを擬似的に作成し、読解モデルの訓練に利用することで、質問応答システムの性能が向上することを示した。最終的な質問応答システムの性能としては、SQuAD に対して 29.8%の正解率を報告している。

Chen らの既存研究では、読解による解答可能性を考慮せずに作成された、擬似的な読解データセットを読解モデルの訓練に用いることで、質問応答の性能が向上することを示している。これに対し、本研究は、質問と文書のペアに対し、読解による解答可能性を明示的に与え、解答可能性を読解モデルに判別させるこ

とで、質問応答の性能向上を目指すのものである。

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　質問 q と正解 a

読解タスクに利用するデータとしては、質問の内容に多様性があり、なおかつ質問が単体で意味をなし、曖昧性が排除されていることが望ましい。本研究では、既存研究 [6] 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』[4] と『abc/EQIDEN 公式サイト』[5]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、全ての質問 q と正解 a のペアを収集した。正誤表[6]に基づき一部の質問の訂正を行った結果、質問と正解のペア (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。

- 質問は 50 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。

収集した全ての質問と正解のペアに対して、以下の前処理を行った。

1. 質問および正解に含まれている、カッコとその内部の文字列を削除した。早押しクイズでは、問題を読み上げる人のために、問題文中の人名等にカッコで読み仮名が書かれている場合があるが、計算機での処理においては不要な情報であると判断し、削除した。

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

2. 問題によっては、正解の同義表現や別表記が複数与えられている場合があるが、全ての問題について過不足なく与えられている訳ではない。今回のデータセット作成にあたっては、正解として最初に与えられている文字列のみを残し、それ以外の正解の文字列は削除した。

3. 質問および正解中の全ての英数字を半角に正規化した。

以上の手順により、計 12591 件の、質問 q と正解 a のペア (q, a) を作成した。

3.2 文書 d

各質問・正解ペア (q, a) について、正解 a を部分文字列として含む文書の付与を行った（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）。付与する文書としては、Wikipedia 記事のテキストを用いた。

2017 年 10 月 1 日時点の日本語版 Wikipedia のダンプファイルから、WikiExtractor[7]を用いて、全記事の本文テキストを取得した。次に、取得した本文テキストを段落ごとに区切り、Elasticsearch[8] (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。そして、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 56651 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 解答可能性の付与

全文検索エンジンを用いて各質問・正解ペア (q, a) に付与した文書 d は、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

[7] https://github.com/attardi/wikiextractor

[8] https://www.elastic.co/jp/products/elasticsearch

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。

作業者には、質問・正解・文書の組 (q, a, d) を提示し、文書 d に質問 q の正解 a の十分な根拠が書かれているかどうかを尋ね、「書かれている」「書かれていない」の 2 択で回答するよう教示した。作業者に対するタスクの教示の画面を図 1 に示す。作業者が回答を行う、タスクの実施画面を図 2 に示す。

1 つの (q, a, d) の組に対して、5 人の作業者に同じ質問を行い、「書かれている」と答えた人数を解答可能性の生のスコア s $(0 \leq s \leq 5)$ 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』[9]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は受け入れないようにした。1 作業あたりの設問数は、16 問（チェック設問を含む）とし、1 作業あたりの謝礼は 5 円相当とした。結果として、1929 人の作業者による回答結果が得られた。

[9] https://crowdsourcing.yahoo.co.jp/

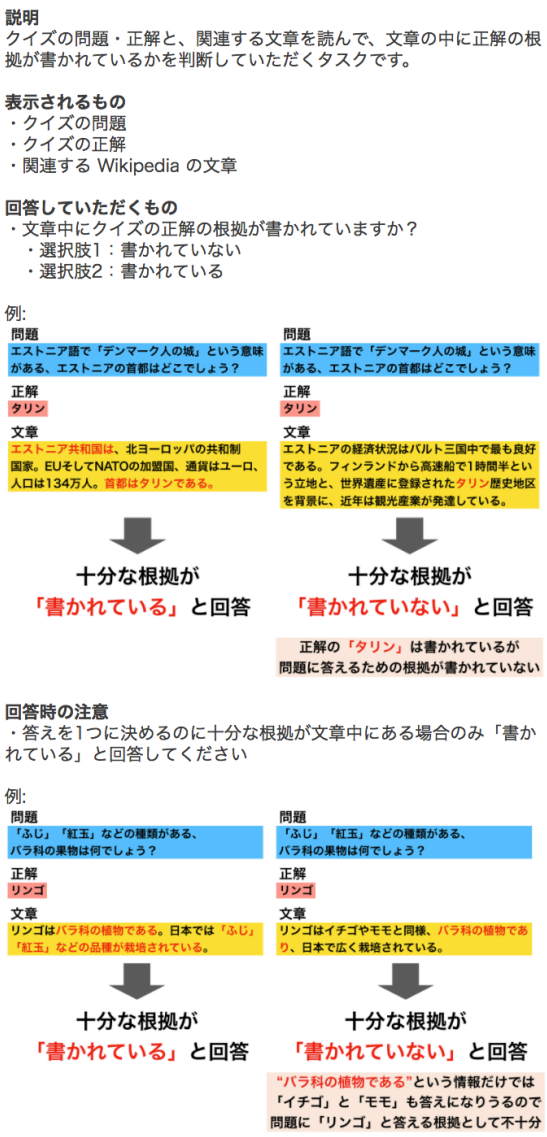

図 1: 作業者への教示画面

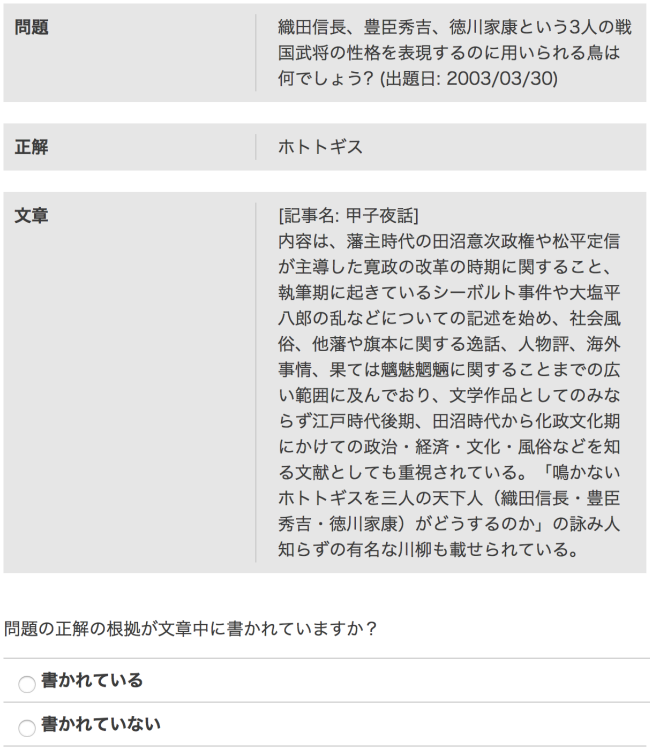

図 2: 作業者のタスク実施画面

表 1: 作成したデータセットの事例

質問 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

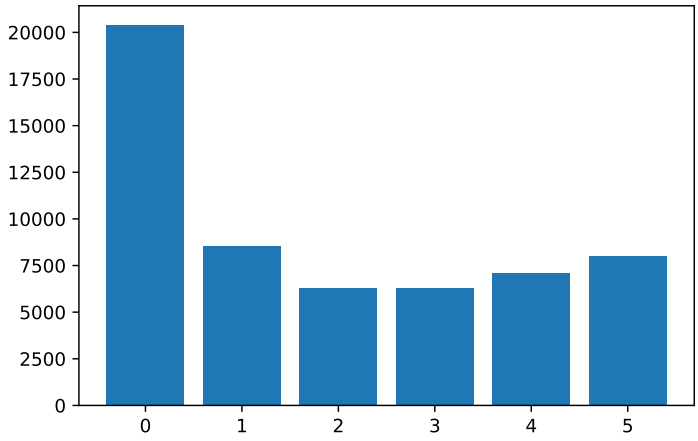

図 3: 解答可能性スコア s の分布

3.4 データの分析

クラウドソーシングによって作成したデータの事例を表 1 に示す。表 1 の上半分の 2 例のように、同じ質問 q 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、作業者によって解答可能性の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、作業者による解答可能性の判断が揺れていると考えられる。

計 56651 件の事例 (q, a, d, s) の解答可能性スコア s の分布を図 3 に示す。正解の文字列 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。

3.5 データの後処理

クラウドソーシングの作業者の回答の統合を行い、回答統合の結果を読解可能性の確率とした。回答統合のアルゴリズムには、作業者の能力と正しい回答を EM アルゴリズムにより推定する回答統合の手法 [10] を用いた。回答統合によって求

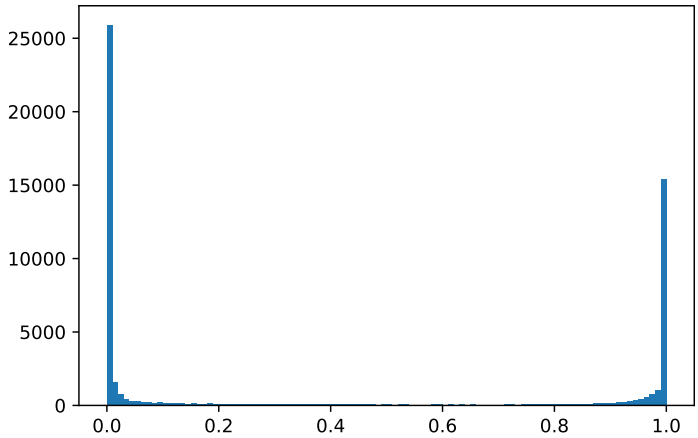

図 4: 解答可能性の確率 p の分布

表 2: 分割されたデータセットのデータ数

	解答可能	解答不可能	計
訓練データ	17221	26389	43610
開発データ	2916	3947	6863
テストデータ	2498	3680	6178
合計	22635	34016	56651

められた、解答可能性の確率 p が 0.5 以上の事例を解答可能、p が 0.5 未満の事例を解答不可能とした。解答可能性の確率 p の分布を図 4 に示す。

最終的に作られた計 56651 件の問題・文書ペアを、問題がクイズの大会で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 2 に示す。

4　実験

作成したデータセットを用いて、解答可能性の判断を行う読解モデルの訓練を行った。従来の読解モデルにはない、解答可能性の判断によって、解答可能な問題に対する解答性能にどのような影響がもたらされるかを検証した。また、訓練した読解モデルを用いて、質問応答システムを構築し、読解モデルに解答可能性を判断させることで、質問応答システムの解答の精度の向上に寄与できるかを検証した。

4.1　モデル

実験に用いる読解モデルとして、SQuAD[1] 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[11] と、解答可能性を判別できるように BiDAF を改変したモデルである BiDAF-opt の 2 つを実装した。このうち、BiDAF-opt は、関係知識の獲得タスクの既存研究 [12] で用いられた手法の再現である。

- BiDAF: N 単語からなる文書 d に対して、解答 a の開始位置と終了位置の確信度スコアのベクトル $\boldsymbol{z}^{start}, \boldsymbol{z}^{end} \in \mathbb{R}^N$ が計算され、softmax 層への入力により、開始位置と終了位置の確率分布を表すベクトル $\boldsymbol{p}^{start}, \boldsymbol{p}^{end} \in \mathbb{R}^N$ にそれぞれ変換される。モデルが予測する解答としては、最も確率が高いスパンの単語列を出力する。

- BiDAF-opt: BiDAF を改変し、解答可能性の判断機構を導入したモデルである。通常の BiDAF における、解答の開始位置と終了位置の確信度スコアのベクトル $\boldsymbol{z}^{start}, \boldsymbol{z}^{end} \in \mathbb{R}^N$ のそれぞれに、解答不可能であることの確信度を表す $N+1$ 次元目の要素を、訓練可能なパラメータとして追加して $\tilde{\boldsymbol{z}}^{start}, \tilde{\boldsymbol{z}}^{end} \in \mathbb{R}^{N+1}$ とする。$\tilde{\boldsymbol{z}}^{start}$ と $\tilde{\boldsymbol{z}}^{end}$ は、softmax 層への入力により、それぞれ確率分布を表すベクトル $\tilde{\boldsymbol{p}}^{start}, \tilde{\boldsymbol{p}}^{end} \in \mathbb{R}^{N+1}$ に変換される。モデルが予測する解答不可能である確率は、$\tilde{\boldsymbol{p}}^{start}, \tilde{\boldsymbol{p}}^{end} \in \mathbb{R}^{N+1}$ の新たに追加した $N+1$ 次元目の要素同士の積 $P(a=\emptyset) = \tilde{\boldsymbol{p}}^{start}_{N+1}\tilde{\boldsymbol{p}}^{end}_{N+1}$ で表される。モデルが予測する解答としては、BiDAF と同様、最も確率が高いスパンの単語列を

表 3: 解答可能な問題に対する実験結果

	EM	F1
BiDAF	65.53	70.20
BiDAF-opt	27.53	28.46

表 4: 解答不可能な問題に対する実験結果

	Precision	Recall	F1
BiDAF-opt	66.48	89.28	76.22

出力するが、どのスパンの確率よりも $P(a = \emptyset)$ が高い場合は、「解答不可能」と出力する。

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[13] による 100 次元の分散表現を用いた。

各読解モデルは Chainer[10] で実装し、エポック数 20、ミニバッチサイズ 60 でモデルの訓練を行った。

4.2 解答可能性の判別を伴う読解の実験

読解モデルに解答可能性を判別させることで、解答不可能な問題を正しく識別できるか、および、解答可能な問題に対するモデルの本来の読解性能にどのような影響があるかを実験により検証した。

設定　データセット中の解答可能な問題のみを用いて、BiDAF を訓練した場合と、データセット中の全ての問題を用いて、解答可能な問題には正解の文字列を、解答不可能な問題には「解答不可能」と出力するように BiDAF-opt を訓練した場合とで、テストデータに対するそれぞれの読解モデルの読解性能を比較した。

評価尺度　解答可能な問題に対しては、SQuAD[1] と同様に、出力と正解 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

[10] `https://chainer.org/`

適合率と再現率から求められる F1 スコアの平均 (F1) を評価尺度に用いた。解答不可能な問題に対しては、解答不可能であることを正しく判別できれば正解とする二値分類タスクとみなし、適合率 (Precision)、再現率 (Recall)、F 値 (F1) を評価尺度に用いた。

結果　解答可能な問題に対する 2 つのモデルの読解性能を表 3 に示す。解答可能な問題に対しては、元の BiDAF では 70%近くの精度で正解を出力することができていたが、解答可能性を判断する機構を導入した BiDAF-opt では、読解の精度は 20%台まで大きく低下した。

解答不可能な問題に対する BiDAF-opt の解答可能性の判別性能を表 4 に示す。解答不可能な問題に対しては、およそ 90%の再現率で解答不可能であることを正しく判別できた一方、適合率は約 66%と低く、解答可能である問題に対しても過剰に「解答不可能」と出力している。実際に、BiDAF-opt が誤った解答を出力した事例を調査したところ、誤りのほとんどが、解答可能性判断における false positive（解答可能な問題に対して「解答不可能」と誤判定してしまう）に起因するものであった。

今回実装した解答可能性を判別できる読解モデルは、既存の読解モデルの単純な拡張ではあるものの、解答可能性を判別することによって、読解の性能が大きく低下することは、解答可能性の判別を伴う読解タスクの難しさを示唆している。

4.3　質問応答システムの実験

読解モデルが解答可能性を判別できれば、質問応答システムで検索された複数の文書からの読解結果を集約するときに、解答可能だった結果のみを解答の候補することで、解答候補の純度が上がり、結果として質問応答システムの性能改善に繋がると考えられる。本実験では、4.2 で訓練した 2 つの読解モデルを用いて質問応答システムを構築し、質問応答の性能にどのような差が生じるかを検証した。

設定　データセットのテストデータに含まれる各問題文 q に対して、Wikipedia 記事段落の全文検索エンジンを用いて、関連する文書を 20 件ずつ取得した。次に、問題文と各文書のペア 20 組に対して、前節の実験で訓練した BiDAF または BiDAF-opt を適用した。そして、読解により出力された計 20 個の解答を集約し、

表 5: 質問応答システムの実験結果

	正解率	無回答による誤り	読解の誤り	解答集約の誤り
BiDAF	27.21	-	72.79	0.00
BiDAF-opt	18.79	29.71	51.50	0.00

もっとも多く出力された解答[11]を、問題文 q に対する解答として出力した。ただし、BiDAF-opt を読解モデルに用いた場合には、「解答不可能」という出力は棄却し、解答可能なものとして出力された解答だけで集約を行った。

評価尺度　質問応答システムが出力する解答文字列と、問題の正解文字列 a が完全に一致する場合のみを正解とみなし、テストデータ中の全ての問題に対する正解率を評価尺度として用いた。

結果　2 つの読解モデルによる質問応答システムの実験結果を表 5 に示す。解答可能性の判断によって、解答候補の純度が上がるという予想に反し、BiDAF-opt を読解に用いた場合の正解率は BiDAF を用いた場合を下回った。

読解モデルを利用した質問応答システムでは、無回答による誤り（「解答不可能」と誤って判断し、解答が出力されない）、読解の誤り（誤った文字列を解答として出力する）、解答集約の誤り（解答候補中に正解が存在するにも関わらず集約の結果選ばれない）の 3 つが考えられる。表 5 には、2 つの読解モデルを用いた場合のこれらの誤りの内訳を併記した。BiDAF-opt では、全体のおよそ 30%の問題で、解答可能性の判断の誤りにより、問題の解答を出力することができなかった。これは、BiDAF-opt による解答可能性判別の適合率の低さが現れた結果であるといえる。

[11]同数の解答が複数ある場合は、出力確率の最大値がもっとも大きかったものを採用した。

5　おわりに

本研究では、56651 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ベースに、解答可能性を判別できる読解モデルを実験的に作成し、答えられない文書が存在する条件下での、読解モデルの解答可能性の判別性能を調査した。そして、解答可能性を判別できるモデルを組み入れた質問応答システムを構築し、質問応答システムの解答集約における、解答可能性を判別することの有効性について検証した。実験の結果、解答可能性を判別することによって、解答可能な問題に対する読解の性能は低下し、解答可能性の判別を伴う読解タスクの難しさを示唆する結果となった。

本研究で構築した、解答可能性を判別する読解モデルは、既存のモデルの単純な拡張であり、読解可能性の判別を伴う読解タスクにおいて、モデルを改良する余地は十分にある。解答可能・不可能の判別をより高い精度で行うことができ、質問応答システム等への応用に耐え得る読解モデルを構築することは今後の課題である。

謝辞

本研究を進めるにあたり、多くの方々のご協力、ご助言をいただきましたことに、ここに心より感謝申し上げます。主指導教員である乾健太郎教授には、ご多忙の中、研究活動だけでなく進路や研究室での生活に関することなど多くのご指導、ご助言を頂きましたことに心より感謝申し上げます。副指導教員である岡崎直観教授には、同じく研究活動に関して多くのご助言を頂きましたことに、心より感謝申し上げます。研究室の松田耕史さんには、日頃から研究の内容や方針はもとより、研究室生活のあらゆる面でのご助言、ご協力を頂きましたことに、心より感謝申し上げます。ご多忙の中審査委員をお引き受けくださいました、木下哲男教授、周暁教授に心より感謝申し上げます。研究会や日々の議論におきまして、多くのアドバイスを頂きました乾研究室の皆様に感謝申し上げます。最後になりましたが、これまでの学校生活におきまして関わってくださいましたすべての皆様に感謝致します。

参考文献

[1] 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. 23832392, 2016.

[2] Karm Moritz Hermann, Tom 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] 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.

[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] 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.

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

[8] 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.

[9] 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.

[10] A. P. Dawid and A. M. Skene. Maximum Likelihood Estimation of Observer Error-Rates Using the EM Algorithm. *Applied Statistics*, Vol. 28, No. 1, p. 20, 1979.

[11] 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.

[12] Omer Levy, Minjoon Seo, Eunsol Choi, and Luke Zettlemoyer. Zero-Shot Relation Extraction via Reading Comprehension. In *Proceedings of the 21st Conference on Computational Natural Language Learning (CoNLL 2017)*, pp. 333–342, 2017.

[13] 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.

発表文献一覧

受賞一覧

- 情報処理学会東北支部学生奨励賞, 2016 年 3 月 12 日

学術論文誌

- Masatoshi Suzuki, Koji Matsuda, Satoshi Sekine, Naoaki Okazaki and Kentaro Inui. A Joint Neural Model for Fine-Grained Named Entity Classification of Wikipedia Articles. IEICE TRANSACTIONS on Information and Systems, Vol. E101-D, No.1, pp.73-81, January 2018. (DOI: 10.1587/transinf.2017SWP0005)

国際会議論文

1. Masatoshi Suzuki, Koji Matsuda, Satoshi Sekine, Naoaki Okazaki and Kentaro Inui. Neural Joint Learning for Classifying Wikipedia Articles into Fine-grained Named Entity Types. In Proceedings of the 30th Pacific Asia Conference on Language, Information and Computation (PACLIC 30), pp. 535-544, October 2016.

2. Masatoshi Suzuki, Koji Matuda, Satoshi Sekine, Naoaki Okazaki, and Kentaro Inui. Multi-label Classification of Wikipedia Articles into Fine-grained Named Entity Types. In Proceedings of the IEEE/WIC/ACM International Conference on Web Intelligence, pp. 483-486, October 2016.

国内会議・研究会論文

- 伊藤拓海, 鈴木正敏, 田然, 山口健史, 岡崎直観, 乾健太郎. 自治体 QA サービスのための FAQ の自治体間の横断的解析. 第 12 回 NLP 若手の会 シンポ

ジウム (YANS), September 2017.

- 鈴木正敏, 松田耕史, 岡崎直観, 乾健太郎. Wikipedia を知識源に用いた文書検索と読解によるクイズ解答システム. 第 12 回 NLP 若手の会 シンポジウム (YANS), September 2017.

- 関根聡, 安藤まや, 松田耕史, 鈴木正敏, 乾健太郎. 「拡張固有表現+Wikipedia」データ. pp.41-44, 言語処理学会第 22 回年次大会予稿集, pp.41-44, March 2016.

- 鈴木正敏, 松田耕史, 関根聡, 岡崎直観, 乾健太郎. Wikipedia 記事に対する拡張固有表現ラベルの多重付与. 言語処理学会第 22 回年次大会予稿集, pp.797-800, March 2016.

- 鈴木正敏, 松田耕史, 関根聡, 岡崎直観, 乾健太郎. Wikipedia エントリの拡張固有表現階層への自動分類. 第 10 回 NLP 若手の会 シンポジウム, September 2015.