ノウハウ質問応答におけるニューラル読解モデルの評価*

山本 航平† 前田 竜冶† 陳 騰揚† 川畑 修人† 大川 遥平† 宇津呂 武仁† 河田 容英‡
†筑波大学大学院 システム情報工学研究科 ‡(株) ログワークス

1 はじめに

本論文では，自然言語での質問に対してインターネット上の膨大の知識を用いて回答する，質問応答生成技術についての研究を行う．事実に関する質問応答生成技術の研究においては，ニューラルネットを用いた読解モデルが提案され，質問に対する回答の候補の探索範囲となるコンテクスト中に必ず回答が含まれるSQuAD1.1 タスクにおいて，人間の性能を超えるモデル [1] が提案される [1] など，研究が進んでいる．しかし，先行研究の大多数は，例えば Wikipedia に明示的に書かれるような事実に関する質問応答を訓練・評価事例とする読解モデルの研究である [1, 4, 5, 7]．そのため，物事のやり方を答えるノウハウや，意見を問う質問に対してはうまく応答出来ない．そこで本論文では，ノウハウに関する質問応答データセットを人手で作成し，読解モデルを適用することによって，ノウハウ質問応答モデルの訓練・評価を行う．

はじめに，日本語での就職活動のノウハウに関する質問応答を例題として，人手で質問応答データセットを作成する．次に，日本語における事実に関する質問応答約 20,000 事例 [6] で構成された訓練・評価事例，および，ノウハウに関する質問応答約 500 事例で構成された訓練・評価事例を用いて読解モデルの訓練を行った．事実に関する質問応答事例のドメイン，および，ノウハウに関する質問応答事例のドメインの 2 ドメイン間でドメインを横断する読解モデルの性能評価を行った．その結果，事実とノウハウの間では，ドメインを横断して読解モデルを適用した場合，性能が低下することがわかった．次に，学習曲線の評価を行った結果，事実に関する質問応答データセット，および，ノウハウに関する質問応答データセットのいずれも，論文のデータセットの規模では性能が飽和せず，訓練事例が不足する事がわかった．最後に，「就職活動」についてのノウハウ質問応答事例を教師データとして訓練した読解モデルを用いて，「花粉症」，「結婚」，「虫歯」，「食中毒」，「マンション」の全く異なる話題のノウハウ質問応答タスクへの適用可能性の評価を行った．教師データである「就職活動」についてのノウハウ質問応答タスクと比べるとやや劣るものの，ほぼ同等の性能であることがわかった．この結果から，ノウハウ質問応答タスクにおいては，異なる話題の間で読解モデルの横断的適用がある程度可能であることがわかった．

*Evaluation of Neural Machine Comprehension Model for Know-how Question Answering

†Kohei Yamamoto, Tatsuya Maeda, Tengyang Chen, Shuto Kawabata, Yohei Ohkawa, Takehito Utsuro, Graduate School of Systems and Information Engineering, University of Tsukuba
‡Yasuhide Kawata, Logworks Co., Ltd

[1] https://rajpurkar.github.io/SQuAD-explorer/

2 ニューラル質問応答モデル

読解タスクとは，質問文とコンテクストを与えて，コンテクストから質問文の回答を探し出してくるタスクである．タスクの解法として，ニューラルネットを用いたモデルが知られている．本論文では，その中でも，大規模の読解データである SQuAD [4] に適用されてきたモデルある BiDAF [5] を用いる．BiDAF は，質問文とコンテクストを入力として，コンテクスト中において，質問に対する回答の開始位置と終了位置を予測するモデルである．

3 質問応答事例のデータセット

3.1 事実に関する質問応答事例

事実に関する質問応答事例として，本論文では回答可能性付き読解データセット [6] を用いた．このデータセットは，早押しクイズ大会の質問文・回答と，回答の文字列が含まれている Wikipedia 記事の段落から抽出したコンテクストの三つ組から構成されるデータセットである．このデータセットのコンテクスト・質問文・回答の例を表 2(a) に示す．

回答可能性スコアは 0~5 の 6 段階で付けられており，スコアが 2 以上は回答可能な質問文とコンテクストの組，スコア 2 未満は回答不可能な質問文とコンテクストの組とされている [6]．本論文では，回答可能

Copyright(C) 2019 The Association for Natural Language Processing.
All Rights Reserved.

表 1: 訓練・評価用質問応答事例の事例数, 平均単語数

(a) 事実に関する質問応答事例

	コンテクスト, 質問文, 回答の組数	コンテクストの平均単語数	質問文の平均単語数
訓練事例	21,626	93.7	28.3
開発事例	3,024	93.6	29.2
評価事例	100	92.7	30.1

(b) 「就職活動」についてのノウハウに関する質問応答事例

	コンテクスト, 質問文, 回答の組数	コンテクストの平均単語数 (全データ)	質問文の平均単語数 (全データ)
訓練事例	512	77.4	10.0
開発事例	102		
評価事例	102		

(c) 「結婚」,「マンション」,「花粉症」,「虫歯」,「食中毒」のノウハウに関する質問応答事例

話題	コンテクスト, 質問文, 回答の組数	コンテクストの平均単語数	質問文の平均単語数
結婚	50	67.9	10.2
マンション	50	80.7	10.3
花粉症, 虫歯, 食中毒	50	75.3	9.5

性スコア 2 以上の質問応答事例のみを用いる．本論文で利用した「事実に関する質問応答事例」の事例数および平均単語数を表 1(a) に示す.

3.2 ノウハウに関する質問応答事例

ノウハウに関する質問応答事例は，ノウハウ知識が多く掲載されるノウハウサイトの候補となるサイトの段落から抽出したコンテクストを元に人手で作成した. ノウハウサイトの収集方法においては，ノウハウサイト自動選定手法 [2] においてノウハウサイト候補を選定する手順の結果，トピックモデルにおける複数のトピックにまたがるサイトを選定し，これを用いた.

「就職活動」,「結婚」,「マンション」,「花粉症」,「虫歯」,「食中毒」の 6 種類のクエリを対象として，まず, ノウハウサイトの候補を収集する．そして，ノウハウサイト候補のウェブページの段落から，コンテクストを抽出する．抽出したコンテクストから質問文と回答を人手で作成する．この時，コンテクストから質問文を作成する際には，コンテクストを読むことによって回答を求められるか否かに注意し，回答が求められる場合にのみ質問・回答を作成した．作成した質問応答事例の一部を表 2(b)，および，表 2(c) に示す．また, 各データセットの事例数・平均単語数を表 1(b)，および，表 1(c) に示す.「就職活動」については，716 事例,「結婚」および「マンション」については各 50 事例,「花粉症」,「虫歯」,「食中毒」については合計 50 事例をそれぞれ作成した.

3.3 事実・ノウハウ混合質問応答事例

事実に関する質問応答事例の数と同数の事実・ノウハウ混合質問応答事例を作成した．その際には，事実に関する質問応答事例集合から,「就職活動」についてのノウハウに関する質問応答事例 (訓練・開発事例) 数分の事例を除外した後，同数のノウハウに関する質問応答事例を追加することによって，事実・ノウハウ混合質問応答事例を作成した.

4 評価

4.1 評価手順

訓練事例として，1)「事実に関する質問応答事例 (訓練事例)」，2)「就職活動についてのノウハウに関する質問応答事例 (訓練事例)」，3)「事実・ノウハウ混合質問応答事例 (訓練事例)」，の三種類を用いて, BiDAF 読解モデル [5] の訓練 [2] を行い，事実に関する質問応答事例 (評価事例)，および，ノウハウに関する質問応答事例 (評価事例) に対する評価を行った．人手評価においては，モデルによって生成された回答を参照用回答と比較し,「完全一致，部分一致，不一致」の 3 段階で評価を行った.

4.2 評価結果

「事実に関する質問応答事例」，および,「就職活動についてのノウハウ質問応答事例」に対する評価結果を図 1 に示す．図 1 の結果から，事実に関する質問応答タスクにおいては,「事実に関する質問応答事例」を訓練事例とした読解モデル，および,「事実・ノウハウ混合質問応答事例」を訓練事例とした読解モデルの性能がほぼ同等となった．一方，ノウハウに関する質問応答タスクにおいては,「ノウハウに関する質問応答事例」を訓練事例とした読解モデルが最も高い性能となった [3].

[2] エポック数 12，ミニバッチサイズ 40 で訓練を行った．単語ベクトルとして，日本語版 Wikipedia 本文全文を用いて，Glove [3] によって訓練した 100 次元の分散表現を用いた.

[3] 日本語形態素解析においては MeCab (`http://taku910.github.io/mecab/`) を用いた．IPAdic および NEologd(`https://github.com/neologd/mecab-ipadic-neologd`) の性能を比較

Copyright(C) 2019 The Association for Natural Language Processing.
All Rights Reserved.

表 2: 訓練・評価用質問応答事例

(a) 事実に関する質問応答事例

コンテクスト H	質問文 U	回　答 A
日本全国の百名山が複数あり、特に読売文学賞を受賞した深田久弥の『日本百名山』はよく知られている。『日本百名山』の本は多数重版され、改版や新装版も出版されている。深田久弥自身の著書の他に、深田久弥に関するものや、日本百名山のすべての山を解説する登山ガイド本などが多数出版されている。他にも、ビデオ・DVD なども、多数出版されている。	登山ブームのきっかけとなった山岳紀行集「日本百名山」などの著作で知られる作家は誰でしょう?	深田久弥

(b) 「就職活動」についてのノウハウに関する質問応答事例

コンテクスト H	質問文 U	回答 A
一般的な履歴書のサイズは A 4 か B 5 なのでどちらの履歴書でも折らずに入れられる角型２号という封筒を選びましょう。コンビニでは売っていないケースもあるので文房具屋で購入するのが確実です。	履歴書を送る際の封筒のサイズは？	角型２号

(c) 「花粉症」についてのノウハウに関する質問応答事例

コンテクスト H	質問文 U	回答 A
シソには花粉症の原因でもあるアレルギー症状を抑え、免疫機能の動きを正常に戻してくれる働きがあります。またシソに含まれる α-リノレン酸は、悪玉コレステロールを減少させる働きがあります。	花粉症に効果のある食べ物は?	シソ

また,「就職活動」,「結婚」,「マンション」,「花粉症」,「虫歯」,「食中毒」についての「ノウハウに関する質問応答事例」を評価事例とした場合の評価結果を図 2 に示す．この結果から分かるように，訓練事例が「就職活動についてのノウハウに関する質問応答事例」であるにも関わらず，いずれの話題の「ノウハウに関する質問応答事例」においても,「就職活動についてのノウハウに関する質問応答事例」と同等かそれ以上の性能となった．特に,「結婚」および「マンション」についての「ノウハウに関する質問応答事例」を評価事例とした場合に，性能が 10%以上高くなった．この結果から，ノウハウ質問応答タスクの評価においては，訓練事例におけるノウハウ質問応答の話題に関係なく一定上の性能が期待できる可能性があることが分かった．

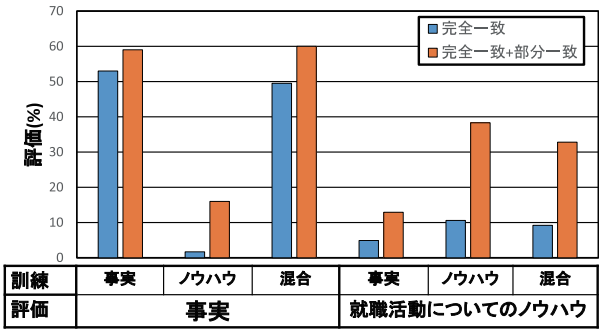


図 1: 「事実に関する質問応答事例」・「就職活動についてのノウハウ質問応答事例」に対する評価結果

5　関連研究

機械読解タスクのデータセットに関する関連研究として，SQuAD [4] は，英語 Wikipedia 記事をコンテクストとして，コンテクストをもとに人手で質問文および回答を作成したデータセットであり，約 10 万件の質問・回答組から構成される大規模読解データセットである．WikiQA [7] は，検索エンジンのクエリログ，および，検索結果の英語 Wikipedia 記事から質問文・コンテクスト・回答の組を作成したデータセットである．回答可能性付き読解データセット [6] は，約 12,000 件の早押しクイズの質問文と回答に対して，関連する日本語 Wikipedia 記事の段落からコンテクストを作成した読解データセットである．以上の関連研究における読解データセットは，いずれも，事実に関する質問応答事例から構成されているため，ノウハウに関する質問応答タスクにおいて利用可能であるか否かが不明である．そこで本論文では，実際にノウハウに関する質問応答事例を作成し，それらの事例に対して既存の読解モデルが適用し，一定以上の性能が達成できることを示した．

した結果では,「事実に関する質問応答事例」に対しては IPAdic の方が高い性能を示した．一方,「ノウハウ質問応答事例」に対しては両者はほぼ同等の性能となった．

Copyright(C) 2019 The Association for Natural Language Processing.
All Rights Reserved.

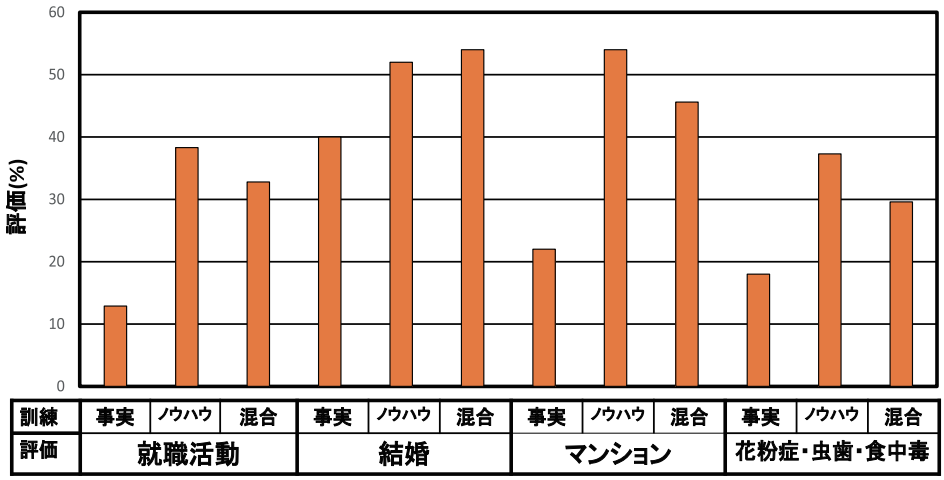
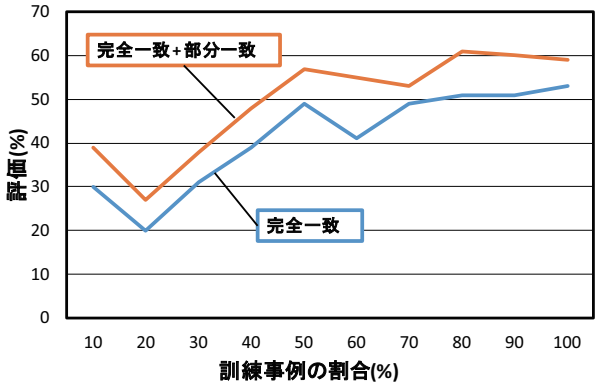


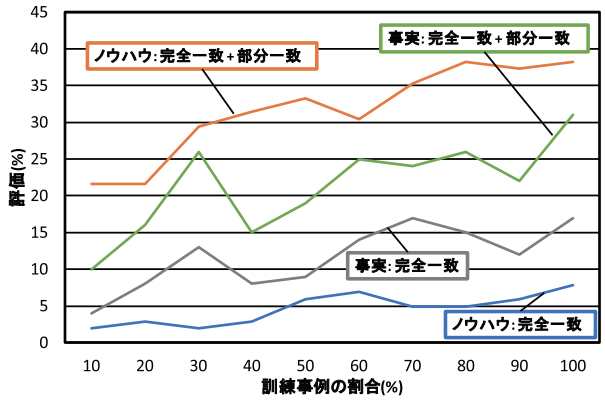
図 2: 「就職活動」,「結婚」,「マンション」,「花粉症」,「虫歯」,「食中毒」についてノウハウを評価事例とした場合の評価結果 (完全一致 + 部分一致.「事実に関する質問応答事例」,「就職活動についてのノウハウに関する質問応答事例」,「事実・ノウハウ混合質問応答事例」をそれぞれ訓練事例とする．)

6 おわりに

本論文では，ノウハウに関する質問応答事例を人手で作成し，既存の読解モデルを適用することにより，一定レベルの性能でノウハウ質問応答タスクが実現できることを示した．今後の課題として，学習曲線の評価結果をふまえると，今後は，訓練事例を効率よく増やす必要があると言える．その他，転移学習等のドメイン適用手法を導入し，事実に関する質問応答とノウハウに関する質問応答の間の異ドメイン間を横断するモデルを確立する必要がある．



(a) 評価事例: 事実に関する質問応答 21,626 事例



(b) 評価事例:「就職活動」についてのノウハウに関する質問応答 512 事例・事実に関する質問応答 512 事例

図 3: 学習曲線

参考文献

[1] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova. BERT: Pre-training of deep bidirectional transformers for language understanding. In *CoRR*, Vol. abs/1810.04805, 2018.

[2] Y. Ohkawa, S. Kawabata, C. Zhao, W. Niu, Y. Lin, T. Utsuro, and Y. Kawada. Identifying tips Web sites of a specific query based on search engine suggests and the topic distribution. In *Proc. 3rd ABCSS*, pp. 4347–4353, 2018.

[3] J. Pennington, R. Socher, and C. D. Manning. Glove: Global vectors for word representation. In *Proc. EMNLP*, pp. 1532–1543, 2014.

[4] R. Pranav, Z. Jian, L. Konstantin, and L. Percy. SQuAD: 100,000+ questions for machine comprehension of text. In *Proc. EMNLP*, pp. 2383–2392, 2016.

[5] M. J. Seo, A. Kembhavi, A. Farhadi, and H. Hajishirzi. Bidirectional attention flow for machine comprehension. In *Proc. 5th ICLR*, 2017.

[6] 鈴木正敏, 松田耕史, 岡崎直観, 乾健太郎. 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. 言語処理学会第 24 回年次大会論文集, pp. 702–705, 2018.

[7] Y. Yi, Y. Wen-tau, and M. Christopher. WikiQA: A challenge dataset for open-domain question answering. In *Proc. EMNLP*, pp. 2013–2018, 2015.

Copyright(C) 2019 The Association for Natural Language Processing.
All Rights Reserved.