ウェブ上のコラムページを情報源とする回答不可能なノウハウ質問応答事例の作成*

陳 騰揚† 前田 竜冶† 李 宏宇† 銭 澤長† 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/理工学群 ‡(株) ログワークス

1 はじめに

質問応答とは，ユーザからの質問に対して，文書中から回答部分を探し出すタスクである．特に，自然言語で記述された質問文とコンテクストを与え，質問に対する回答を探すタスクを読解タスクと呼ぶ．読解タスクにおいては，近年，ニューラルネットワークを用いた読解モデルが提案され，代表的な読解タスク用データセットの一つである SQuAD [5] に対しては，人間の回答精度を上回る結果が得られることが報告されている [1]．ここで，質問応答のタスクは，固有名詞や数量等といった事実に関する内容を回答対象とする質問応答，および，目的や理由，方法等の非事実を回答対象とする質問応答の二種類に大別される．このうち，現状の研究動向としては，例えば Wikipedia に明示的に書かれるような事実を回答対象とする研究 [5, 6, 8] が大半を占める．そのため，非事実の典型として，ものごとのやり方を答えるノウハウや，意見を問う質問に対して的確に応答することは容易ではない．

その一方で，インターネット上には，物事のやり方に関するノウハウが多数掲載するウェブサイト (本論文では，ノウハウサイト [2] と呼ぶ) が多数存在し，それらのノウハウサイトを効率よく収集する手法 [2, 4] も提案されている．そこで本論文では，ニューラル読解モデル (図 1) の中でも，特にノウハウに関する質問応答を対象とするニューラル読解モデル訓練のための回答不可能な [6] 質問応答事例データセットを作成することを目的とする．具体的には，上記のノウハウサイトの中で，コラムページから回答可能なノウハウ質問応答事例を収集する手法 [3] を用いることにより，回答不可能なノウハウ質問応答事例を効率よく作成する方式を提案する．

*Developing an Unanswerable Example of Know-How Question Answering from Column Pages on the Web

†Tengyang Chen, Tatsuya Maeda, Hongyu Li, Zechang Qian, Takehito Utsuro, Graduate School of Systems and Information Engineering/School of Science and Engineering, University of Tsukuba
‡Yasuhide Kawata, Logworks Co., Ltd

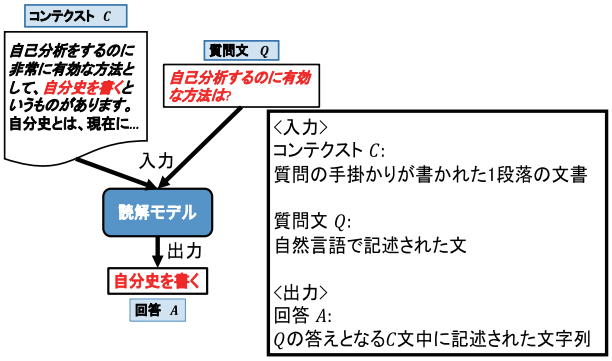

図 1: ニューラル読解モデルの概要

2 ノウハウ質問応答事例作成の情報源としての「コラムページ」 [3]

本論文では，クエリ・フォーカスとして，「就活」，「結婚」，「花粉症」，「マンション」，「虫歯」を対象として，文献 [3] に従い，図 2 の手順に沿って，クエリ・フォーカス (図 2 の例では「就活」) に関するノウハウを多く掲載するノウハウサイトを収集し，ノウハウサイト中のウェブページ集合をトピックに分類した結果を選定した[1]．次に，文献 [3] に基づき，ノウハウ質問応答事例作成の情報源として「コラムページ」が有効であるという分析結果を踏まえ，ノウハウ質問応答事例作成の情報源として用いるウェブページを選定した．

3 ノウハウ質問応答事例の作成手順

本節では，ノウハウ質問応答事例の作成手順 (図 3) について述べる．以下で定義されるコンテクスト C，質問 Q，および，回答 A の組を人手で作成する．

回答可能なコンテクスト C: 一段落で構成される短い部分文書であり，画像などの非テキスト情報を含まない．

[1]本論文においても，文献 [3] で選定した「就活」に関する計 14 のノウハウサイトをサイト集合を T とした．サイト集合 T のウェブページを使用し，確率最大となるトピック z_n を割り当てることにより，トピック z_n が割り当てられたウェブページの集合に対してサイトごとに確率値上位 5 ページのみを集めた部分集合 $D^{\inf}(z_n, T, \text{“}\leq 5\text{”})$ を以降の作業対象とする．

表 1: ノウハウ質問応答事例 (コンテクスト C・質問 Q・回答 A の組) の例

コンテクスト C	質問 Q	回答 A
履歴書に短所を書く時は前向きにまとめるようにします.「工夫して克服した」「直すように努力している」などと書けば悪いイメージの短所で好印象を与えることも可能です. 自分の短所の中で努力すれば改善しそうなものを選ぶと書きやすいでしょう.	履歴書に短所を書く時のポイントは？	前向きにまとめる
結婚式の節約術の 1 つとして，ウエディングアイテムを持ち込むという方法があります. これは，結婚式に必要なものをネットやお店で自分で探して購入したり，自分自身で手作りしたりして当日使用するということです.	結婚式の費用を節約する方法は？	ウエディングアイテムを持ち込む
そのため花粉の季節は，室内の湿度を 50〜55%ほどに保てるように加湿器を使用しましょう.	花粉の季節に保つべき室内の湿度の目安は？	50〜55%

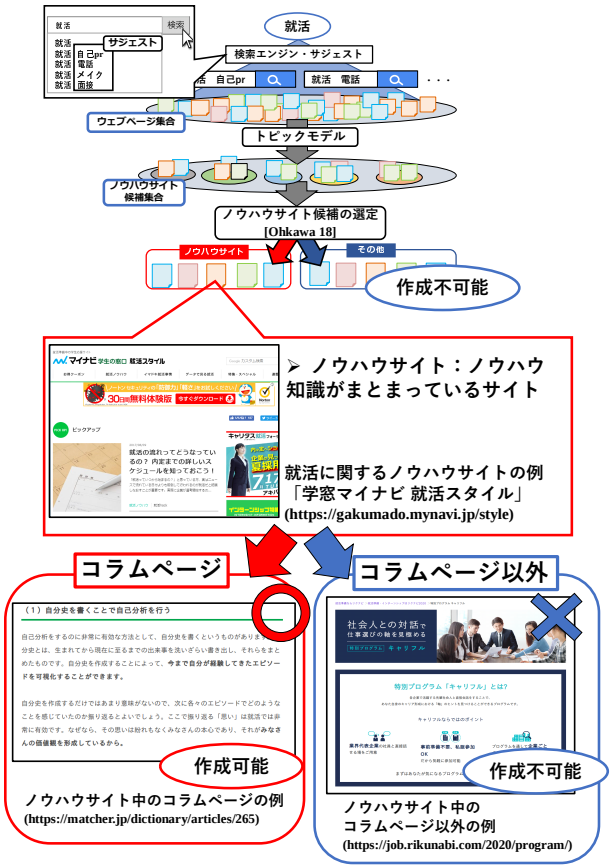

図 2: ノウハウ質問応答事例作成のためのウェブページ選定手順

質問 Q: 自然言語で記述された文.

回答 A: コンテクスト中に記述された任意の文字列.

具体的には，まず，回答可能事例として，一ページあたり最大 5 組を上限に [2] ページ中の各段落に対して，回答可能なコンテクスト C，質問 Q，および，回答 A の組の作成可否を判定し，作成可能な場合には，当該段落を回答可能なコンテクスト C とし，質問 Q および回答 A の組とともに記録する [3].

次に，回答可能事例で C として使用した段落の前後の段落のうち以下の条件を満たし，質問 Q に対して

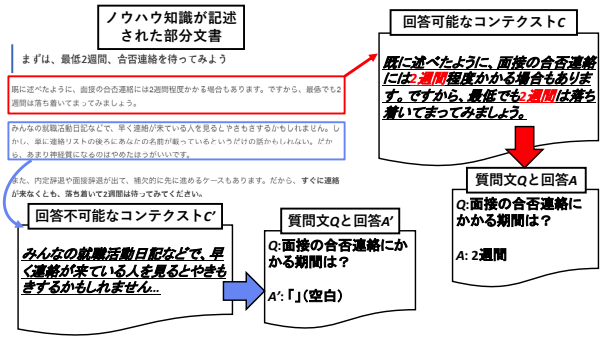

図 3: ノウハウ質問応答事例の作成例

回答不可能なコンテクストを C' とし，対応する Q とともに回答不可能事例として以下のように記録する [3].

回答不可能なコンテクスト C': 質問 Q に対する回答となりうる情報が含まれない文書. 一段落で構成される短い部分文書であり，画像などの非テキスト情報を含まない.

質問 Q: 自然言語で記述された文.

回答 A': 「」(空白)

一例として，ノウハウサイトの「コラムページ」[4] からノウハウ質問応答事例を作成する場合の模式図を図 3 に示す. この例では，回答可能なコンテクスト C として「既に述べたように，面接の合否連絡は 2 週間ほどかかる場合にもあります ...」という段落を抽出し，質問 Q として「面接の合否連絡にかかる時期は?」，Q に対する回答 A として「2 週間」を抽出した. この Q, A, C を回答可能事例として記録する. 次に，回答可能なコンテクスト C の段落の前後の段落 (この例では次の段落)「みんなの就職活動日記などで，早く連絡がきている人を見る ...」は回答になりうる情報を含まないと判断し，回答不可能なコンテクスト

[2] 一ページから作成された質問応答事例数の分布を表 2 に示す.

[3] 前後の段落でも回答可能である場合は，ページ本文の冒頭や末尾に書かれることが多い要約段落を用いた. 本論文において，要約段落を用いた事例は、作成した回答可能な質問応答事例計 1,688 事例のうち 21 事例存在した.

[4] https://matcher.jp/dictionary/articles/12

C' として抽出する．回答不可能事例における回答 A' は「」（空白）とし，Q, A', C' を回答不可能事例として記録する．クエリ・フォーカスとして「就活」,「結婚」,「花粉症」を用いた場合のノウハウ質問応答事例各一例ずつを表 1 に示す．訓練・評価に用いた質問応答事例の事例数および平均単語数を表 3 に示す．クエリ・フォーカス「就活」の場合には，ノウハウ質問応答事例作成対象となる「コラムページ」 1,268 ページ中 352 ページを情報源とした.「結婚」の場合には「コラムページ」中 144 ページを情報源とした．一方，クエリ・フォーカス「マンション」については，100 事例,「花粉症」,「虫歯」,「食中毒」については合計 100 事例をそれぞれ作成した．

表 2: 一ページから作成された質問応答事例数

質問応答事例数	1	2	3	4	5	計
ページ数 (%)	131 (26.4)	102 (20.6)	64 (12.9)	34 (6.9)	165 (33.3)	496 (100)

表 3: 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 (回答可能/回答不可能)	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	27,427/28,742	95.4	28.2
評価事例	50/50	89.2	26.0

(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 (回答可能/回答不可能)	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	807/807	72.2	11.1
評価事例	50/50	73.8	10.9

(c) 「結婚」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 (回答可能/回答不可能)	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	481/481	49.3	12.2
評価事例	50/50	42.6	12.5

(d) 「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 (回答可能/回答不可能)	コンテクストの平均単語数	質問文の平均単語数
マンション	50/50	90.9	11.1
花粉症, 虫歯，食中毒	50/50	67.5	10.4

4 評価

4.1 評価手順

本論文で作成したノウハウに関する質問応答事例，および，事実に関する質問応答事例として，解答可能性付き読解データセット [7] を用いて BERT [1][5] を機械読解タスク用に fine-tuning する．訓練事例として，1)「事実に関する質問応答事例」，2)「就活活動」，および,「結婚」についてのノウハウに関する質問応答事例」，3)「事実・ノウハウ混合質問応答事例」，の三種類を用いてモデルを作成した．各モデルに対して，訓練事例とは別に用意した事実・ノウハウに関する質問応答事例に対する評価を行った．人手評価においては，モデルが出力した回答と参照用回答を比較し,「完全一致，部分一致，不一致」の三段階で人手評価し，完全一致，および，部分一致の事例数の割合を算出した (EM+PM)．また，出力回答の単語列と参照用回答の単語列に対する適合率と再現率から算出される F1 スコアも用いた．

4.2 評価結果

評価結果を図 4 に示す．事実に関する質問応答タスクにおいては,「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により最も高い性能が達成された．ノウハウ質問応答タスクにおいては,「事実・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．また，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で訓練・評価を交差させて，一方で訓練したモデルを他方で評価した場合，性能が低下することが分かった．一方，ノウハウ質問応答事例においては「就職活動」,「結婚」,「マンション」,「花粉症・虫歯・食中毒」といった話題に関係なく一定以上の性能が得られた．このこと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分かった．

「事実に関する質問応答事例」および,「ノウハウに関する質問応答事例」を訓練事例とした際の学習曲線を図 5 に示す．結果として，ノウハウ質問応答タスクにおいて，訓練事例数が事実に関する質問応答タスクの約 4.5%にも関わらず，事実に関する質問応答タスクと同等の性能となった．この結果から，ノウハウ質

[5]日本語実装として，TensorFlow 版 (`https://github.com/google-research/bert`) を用い，事前学習モデルには多言語モデル (Multilingual Cased model) を採用した．また日本語形態素解析においては Mecab (`http://taku910.github.io.mecab/`) を用いた．

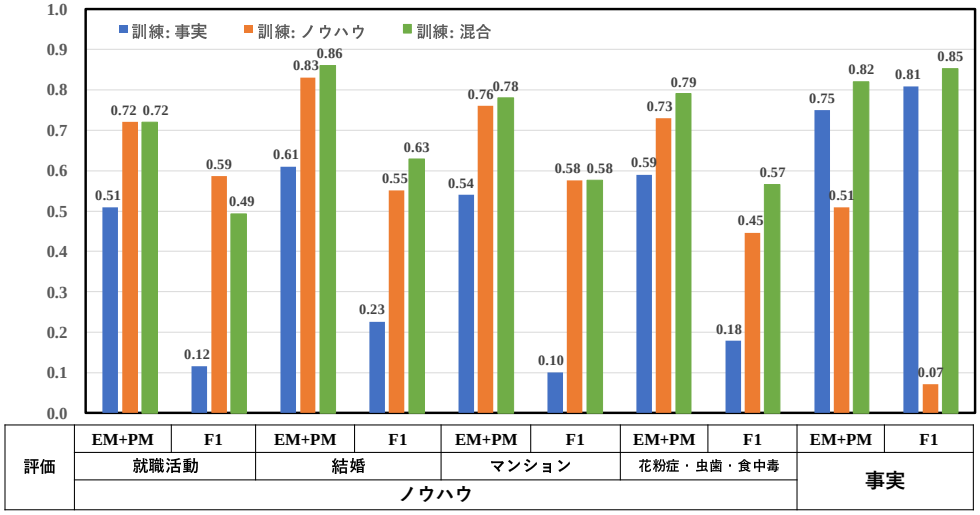

図 4: 評価結果 (完全一致+部分一致 (EM+PM)，および，F1 スコアの平均 (F1))

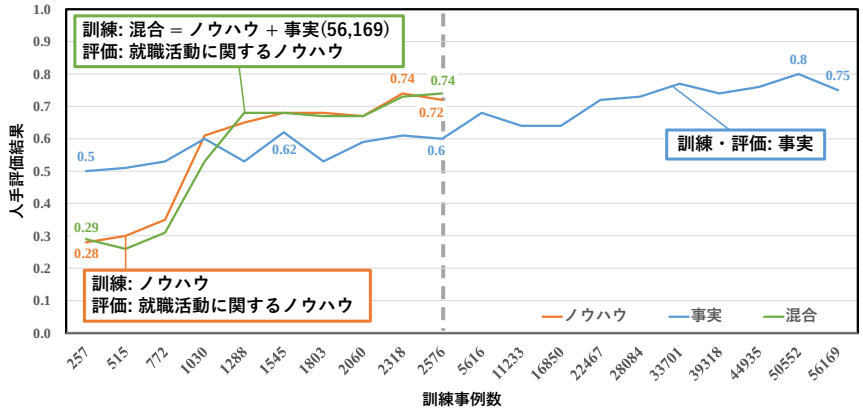

図 5: 学習曲線の比較 (完全一致+部分一致)

問応答タスクは事実に関する質問応答タスクに比べてはるかに少ない訓練事例数で一定以上の性能が得られることが分かった．

5 関連研究

本論文に関連して，事実に関する読解タスク用データセットとして，英語を対象とした SQuAD [5, 6]，および，日本語を対象とした解答可能性付き読解データセット [7] が挙げられる．SQuAD [5,6] は，英語版 Wikipedia 記事中の段落をコンテクストとして，クラウドソーシングにより質問・回答を付与した約 10 万件の質問応答事例から構成されたデータセットである．一方，解答可能性付き読解データセット [7] は，日本語の早押しクイズ大会の約 12,000 件の質問文・回答に対して，回答の文字列が含まれる日本語版 Wikipedia 記事中の段落から選定したコンテクストを付与した質問応答事例から構成されたデータセットである．

6 おわりに

本論文では，ノウハウサイトにおいてノウハウが掲載されたウェブページを情報源として，ニューラル読解モデル訓練のための回答不可能な質問応答事例を作成する方式を提案した．また，作成したノウハウ質問応答事例に対して読解モデル BERT [1] を適用し，結果として，ノウハウ質問応答タスクは事実に関する質問応答タスクに比べて相対的に極めて少量の訓練事例のもとで一定以上の性能が得られることが分かった．

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] 李佳奇, 趙辰, 林友超, 丁易, 川畑修人, 宇津呂武仁, 河田容英. トピックモデルおよび分類器学習を用いたノウハウサイトの同定. 第 10 回 DEIM フォーラム論文集, 2018.

[3] 前田竜治, 陳騰揚, 大川遥平, 宇津呂武仁, 河田容英. ウェブ上のコラムページからのノウハウ質問応答事例の収集. 第 33 回人工知能学会全国大会論文集, 2019.

[4] 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.

[5] 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.

[6] R. Pranav, J. Robin, and L. Percy. Know what you don't know: Unanswerable questions for SQuAD. In *Proc. 56th ACL*, pp. 784–789, 2018.

[7] 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.

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