ノウハウ機械読解における生成型モデルのマルチタスク学習

Multi-task Learning on How-to Tip Machine Reading Comprehension by Generative Model

李　廷軒† 田村　拓也† 西田　隼輔† 朱　福主† 宇津呂武仁††

† 筑波大学大学院 システム情報工学研究群 知能機能システム学位プログラム
〒 305–8577 茨城県つくば市天王台 1-1-1
†† 筑波大学 大学システム情報系 知能機能工学域
〒 305–8573 茨城県つくば市天王台 1-1-1

あらまし　ノウハウ質問回答事例を対象する機械読解の研究において，従来の抽出型機械読解の手法では，コンテキストの中で回答となる文字列が複数箇所にまたがる事例への対応が難しいという問題が存在している．本論文では，ノウハウサイト wikihow を情報源とする生成型データセットを構築し，生成型ノウハウ機械読解モデルを訓練した．さらに，抽出型と生成型のデータセットを一つのモデルに同時に学習させるマルチタスク学習の手法を提案した．評価実験の結果，生成型質問回答事例におけるマルチタスク学習のモデルの性能がシングルタスク学習のモデルの性能を上回る結果が得られた．

キーワード　質問応答，機械読解，生成型，ノウハウ，mT5，マルチタスク学習

1　はじめに

自然言語処理における機械読解タスクとは，図 1 に示すように，自然言語で記述される質問文とコンテキストを与え，コンテキストから質問文の回答となる部分を抽出するタスクである．近年，深層学習の発展および大規模データセットの整備により，機械読解の研究分野において成果が挙がっている．例を挙げると，英語 Wikipedia の記事から作成された機械読解用データセット SQuAD [10] に対して，人間を上回る読解性能を達成したことが報告されている [1]．また，コンテキスト中に質問の回答が含まれていない場合も考慮して，「回答不可能」の質問応答事例を含むデータセット SQuAD 2.0 [12] も作成されている．

機械読解タスクにおいては，固有名詞や数量などの事実を回答対象とした事実型問題を対象とする SQuAD [10], [12]，および，解答可能性付き読解データセット [14] の他に，物事のやり方や理由などの非事実を回答対象とした非事実型問題に対する研究も盛んに行われている．問題設定が比較的単純である事実型問題と比較すると，非事実型問題は難易度が高いと言える．非事実型問題の一つである日本語ノウハウ機械読解においては，文献 [2], [8] において，インターネット上のノウハウサイトから収集したコラムページを情報源として，ノウハウ機械読解モデルの訓練・評価事例を作成している．

ここで，通常の機械読解タスクのモデル，および，データセットにおいては，コンテキストから質問文の回答となる部分を抽出する方式に従っている．しかし，ノウハウ機械読解においては，コンテキスト中で回答となる文字列が複数箇所にまたがることも多い．機械読解タスクに適用されるモデルとしては，機械読解タスク用に fine-tuning された BERT [3] モデルが代表的であるが，抽出型モデルである BERT モデルはこのような事例への対応が難しい．そこで，本論文では，ノウハウに関する質問回答事例に対し，コンテキスト中の文字列を抽出する抽出型モデルではなく，生成型モデル mT5 [17] によって質問の回答を生成する方式を適用する．

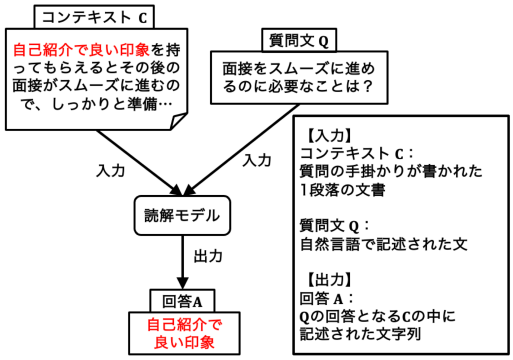

図 1　ノウハウ読解モデルの枠組み

2　機械読解における抽出型と生成型の違い

事実型機械読解の代表的データセットである SQuAD [10] における機械読解タスクは，コンテキストから質問文の回答となる部分を抽出するタスクである．文献 [2] で作成した日本語ノウハウ型質問応答データセットにおいても，コンテキストとして使用する文章に基づいて質問文を作成し，コンテキストから回答を抽出する方式によって，質問回答事例を作成している．

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

この方式に基づく機械読解を「抽出型機械読解」と呼ぶ．

抽出型機械読解タスクにおいて用いられる代表的な抽出型機械読解モデルとして，BERT [3] が挙げられる．BERT においては，Transformer [16] による双方向のエンコーダの仕組みにより，機械読解タスクにおいて回答の開始位置と終了位置を予測し，二つの位置に挟まれるスパンを質問文の回答として出力する．

一方，回答を生成する方式の機械読解を「生成型機械読解」と呼ぶ．生成型機械読解の研究事例においては，コンテキストが必要なく，事前訓練，および，fine-tuning されたモデルに質問を入力するだけで回答となる文を生成する事例 [1] も存在するが，本論文では，コンテキストの入力を仮定する設定で生成型機械読解を行う．

本論文では，生成型機械読解モデルにおいて回答を生成する能力として，Transformer によるデコーダに基づき，入力テキストを内部表現に変換する Encoder，および，Encoder の出力をまとめ，さらに可変長の出力をする Decoder の仕組みを持つモデルを生成型機械読解のタスクに適用する．そのような生成型モデルの代表として，T5 [11] が挙げられる．本論文では，多様なタスクに対する膨大な訓練事例を用いて事前訓練されたモデルに対して，下流タスクに対する fine-tuning を行う転移学習によって得られるモデルを利用する．これまでに，抽出型機械読解タスクにおいて，SQuAD [10] に対して T5 [11] を適用した結果においては，高い性能を達成したことが報告されている．

両モデルの特性の違いを示すため，本論文では，多言語テキストにより事前訓練された BERT [3](mBERT) および T5(mT5 [17]) の性能比較を行う．具体的には，抽出型モデルの方は文献 [2] で作成した 1,614 件の抽出型ノウハウ質問応答事例 (回答可能事例 807 件，および，回答不可能事例 807 件) によって mBERT の fine-tuning を行い，生成型モデルの方は 3 節で作成した生成型ノウハウ質問応答事例 10,000 件を用いて 4.2 節にて述べた手法で mT5 の fine-tuning を行った．さらに，生成型に関する質問回答事例の評価事例の質問文とコンテキストを両モデルに入力し，ノウハウ機械読解の性能を評価した．両モデルの評価の具体例を表 1 に示す．表 1(a) の例 1 により，回答となる部分を抽出する BERT に対し，T5 においては，コンテキスト中で複数箇所に点在する回答の断片をまとめる機能，または，語尾を補完する機能を示すことができた．表 1(b) の例 2 では，コンテキストの後半は否定的な意味を含むにもかかわらず，BERT においては，前半の肯定的な意味しか含まない部分を回答として出力するのに対し，T5 においては，コンテキストの全体的な文脈情報を参考にする回答を生成する特性が示された．

表 1　モデルが出力した予測回答の具体例

(a) 例 1

質問文	一時的なタトゥーシートを自作する方法について，どうすればいいですか？
コンテキスト	流動性に優れたゲルペンのインクは、デザインが完成したら、シールから肌へ簡単に付着します。トレーシングペーパーもしくはパーチメント紙上に自分の好きなデザインを鉛筆で描き、その後、ゲルペンで色を塗ります。線上にできるだけ多くのインクを出すようにしますが...
正答	濃い色のゲルペンを使ってデザインを描きます
mBERT の予測回答	パーチメント紙上に自分の好きなデザインを鉛筆で描き
mT5 の予測回答	ゲルペンを使ってトレーシングペーパーもしくはパーチメント紙上に自分の好みのデザインを描きましょう。

(b) 例 2

質問文	小論文の冒頭に引用文を使う方法について，最適な引用文を見つける，どうすればいいですか？
コンテキスト	多くの人が利用している有名な引用文をそのまま使ったりすると、それだけで読者はうんざりしてしまいます。また、著者が真面目に取り組んでいないか、読者のことをあまり考えていないかのようにさえ思われるでしょう。
正答	決まり文句やよく目にする引用文は避ける
mBERT の予測回答	多くの人が利用している有名な引用文をそのまま使ったり
mT5 の予測回答	引用文をそのまま使うかどうかを考える

3　回答生成型データセット

本節では，wikihow の日本語版 [2]を情報源とする生成型機械読解方式に関する質問応答データセットの作成方法について述べる．

wikihow とは，車や家族や健康など全 19 トピックにまたがり，各分野に関するノウハウが集まっているウェブサイトである．wikihow の英語版から作成するデータセットを要約タスクへ適用した文献 [7] は存在するが，本研究では，タイトルとサブタイトルを繋げたものを質問文として利用し，記事本文のま

2 : https://www.wikihow.jp/

とめを回答として利用する．ウェブサイトに掲載しているノウハウコラムの構成の例は以下のようになる．

タイトル：色落ちしにくいジーンズの洗濯方法
サブタイトル：洗濯後の手入れ
記事内容まとめ：洗濯する代わりに霧吹きで水をかける
記事本文：ジーンズについた汗や汚れ、臭いが気になり出しても、すぐに洗濯機に入れてはいけません。まずは霧吹きで水をかけて臭いを落としましょう。ジーンズの洗濯は、4〜5 週間に 1 回程度で十分です。霧吹きを用意し、水とウォッカを 1：1 の割合で入れます。水を吹きつけたジーンズを一晩冷凍庫に入れておくと、さらに臭いを軽減できます。

上記のウェブコラムを本研究で使用する生成型機械読解方式に関する質問応答事例に整形すると，以下のようになる．

質問文：色落ちしにくいジーンズの洗濯方法について、洗濯後の手入れ、どうすればいいですか？
コンテキスト：ジーンズについた汗や汚れ、臭いが気になり出しても、すぐに洗濯機に入れてはいけません。まずは霧吹きで水をかけて臭いを落としましょう。ジーンズの洗濯は、4〜5 週間に 1 回程度で十分です。霧吹きを用意し、水とウォッカを 1：1 の割合で入れます。水を吹きつけたジーンズを一晩冷凍庫に入れておくと、さらに臭いを軽減できます。
回答：洗濯する代わりに霧吹きで水をかける

回答が難しい事例を削除するため，事例の長さによるフィルタリングを行った．ここの長さとは，一律に MeCab[3] により分かち書きにされる文の形態素数とする．フィルタリングの基準として，コンテキストは質問文の 2 倍以下，もしくは 10 倍以上の長さを持つ事例，または実験で使用する huggingface の mT5[4]モデルの仕様に合わせ，入力する文の長さ上限を超える事例を除いた．フィルタリング前後のデータセットの情報を表 2 に示す．フィルタリングにより，訓練される mT5 モデルが生成型に関する質問回答事例における性能は BLEU 値の 1.1 ポイント上昇した結果を得られた．

本節で作成したデータセットの事例数を表 3(a) に示す．

4 生成型ノウハウ機械読解におけるマルチタスク学習

マルチタスク学習とは，目的とするタスクに関係がある複数のタスクを一つのモデルに同時に学習を行わせることで精度を向上させる手法である．本節では，生成型ノウハウ機械読解における 2 通りのマルチタスク学習方式を提案する．

4.1 抽出型・生成型混合事例によるマルチタスク学習

T5 [11] においては，T5 モデルの構造上，text-to-text 的な入出力フォーマットの設定に基づき，タスクごとに異なる接頭辞 (prefix) を自分なりに定義して事例の先頭に加えることにより，複数のタスクを同時にマルチタスク学習できることが示された，多言語版の mT5 [17] においてもその機能が引き継がれているため，本論文では，mT5 によってマルチタスク学習を行った．

マルチタスク学習を行う対象を抽出型機械読解と生成型機械読解とする理由は，タスク間で共通する因子や有用な特徴を持っているからである．なお，生成型機械読解は抽出型よりタスク自体の難易度が高いであり，困難なタスクは容易なタスクから情報を得て学習が簡単になるのだと考えられる．また，使用するデータセットでは，共通の内容を持つデータから複数のタスクを行うことを望むため，3 節で作成した生成型質問応答事例を生成型タスクのデータとして使用した上で，生成型質問応答事例の正答をコンテキストの先頭に加え，コンテキストに回答を含む事例を抽出型タスクのデータとして使用する．両方のデータを同じ数で混ぜて生成型モデルに入力し，マルチタスク学習を行う．

4.2 文間意味的類似度・回答生成によるマルチタスク学習

文間意味的類似度尺度 (Semantic Similarity) は，機械翻訳の研究に多く導入されているが，近年，機械読解タスクにおいて利用する文献 [15] も存在している．ゆえに，mT5 による回答生成タスクに基づき，訓練事例の質問文・コンテキストと回答文の距離学習も行い，質問文・コンテキストの入力とより類似する回答を生成することを目標にする手法を提案する．

モデルの詳細を図 2 に示す．Sentence-BERT [13] は Siamese Network という枠組みを持つことにより，文間意味的類似度のタスクに対する高い性能を持つことが示された．mT5 の訓練途中で違うモデルの読み込みを挿入することは回答生成タスクにおける文の埋め込みの統一や訓練時間の遅さなど実装的な問題の存在により，mT5 のエンコーダの埋め込み層を利用し，Siamese Network と同様な構造を作り，文間意味的類似度のタスクを行う．

また，損失関数について，本節で提案した手法における文間意味的類似度タスクでは，訓練データの中の全ての事例を正例扱いするので，正例だけが存在する状況で使用する Multiple Negative Ranking Loss という損失関数を適用した．sentence-transformer のライブラリ内のモデルを使用しなかったため，Multiple Negative Ranking Loss は sentence-transformer のライブラリの関数 [5] をそのまま利用できなく，スクリプト内のアルゴリズムを参考して実装した．

一方，回答生成のタスクでは，mT5 モデルのデフォルトの Cross Entropy Loss を使用する．文間意味的類似度タスクの損失関数を L_m にし，回答生成タスクの損失関数を L_c にする．次式により，

$$Loss = \lambda_m L_m + \lambda_c L_c$$

二つのタスクの損失関数の重み付き和をモデルの損失関数とし

3：https://github.com/neologd/mecab-ipadic-neologd
4：https://huggingface.co/google/mt5-base
5：https://github.com/UKPLab/sentence-transformers/blob/master/sentence_transformers/losses/MultipleNegativesRankingLoss.py

表 2　フィルタリング前後の生成型ノウハウ質問応答データセットの情報

	事例総数（件）	コンテキスト 平均長さ（形態素数）	回答平均長さ （形態素数）	BLEU
フィルタリング前	25,681	108.75	9.92	8.0
フィルタリング後	11,744	67.01	11.53	9.1

て，生成型モデルのマルチタスク学習を行う．

5 評　　価

5.1 評 価 手 順

本節では，3 節で作成した生成型データセットを使用した実験と評価を述べる．

生成型モデルは T5 [11] の多言語版 mT5[6]を用いた．本論文では，3 節で作成した生成型ノウハウ機械読解データセットを用いて，4 節で述べた 2 通りの提案手法で実験を行った．

データセットを訓練，開発と評価データに分ける．4.1 節で述べた抽出型・生成型混合訓練事例によるマルチタスク学習で用いるデータのうち，抽出型に関する質問回答事例は，生成型に関する質問回答事例の正答をコンテキストの先頭に加えて作成される．各データの詳細な事例数を表 3 に示す．生成型質問応答事例だけを用いてシングルタスク学習方式で訓練される mT5 の評価結果をベースラインとし，提案手法との比較を行った．

4.2 節で述べた文間意味的類似度・回答生成によるマルチタスク学習の提案手法による実験では，提案した損失関数の式により，λ_m と λ_c の様々な組み合わせによる実験を行った．

自動評価には，BLEU [7]を使用した．また，BLEU による評価結果では，予測回答と正答との字面での一致度しか考慮しないため，評価事例 235 件からランダムサンプリングで選出した 100 件事例での人手評価も行った．生成文はコンテキストの回答として成立するか否かを人手評価の基準にした．

表 3　訓練・評価用質問回答事例の事例数
(a) 回答生成型事例

	事例総数（件）
訓練事例	10,000
開発事例	1,509
自動評価事例	235
人手評価事例	100

(b) 抽出型・生成型混合事例

	事例総数（件） (生成型/抽出型)
訓練事例	20,000(10,000/10,000)
開発事例	1,509(1,509/1,509)
自動評価事例	235(235/0)
人手評価事例	100(100/0)

6 : https://huggingface.co/google/mt5-base

7 : https://github.com/mjpost/sacrebleu

5.2 評 価 結 果

各手法で訓練されるモデルの評価結果を表 4 に示す．提案手法によって 2 通りのマルチタスク学習で訓練されるモデルにおいては，シングルタスク学習で訓練されるモデルを上回る性能が得られた．

BLEU での評価結果とはモデルの予測回答と正答との 1∼4-gram による一致度（表 4 では bleu-n にする）の幾何平均の上に，短い生成文に対するペナルティ(BP) をかけた値である．bleu-n の結果は文間意味的類似度・回答生成によるマルチタスク学習方式の方が勝るが，BP の影響により，BLEU 値は抽出型・生成型混合訓練事例によるマルチタスク学習方式の方が上回る．また，人手評価でも，抽出型・生成型混合訓練事例によるマルチタスク学習方式の方が優れている．

文間意味的類似度・回答生成によるマルチタスク学習方式で訓練されるモデルの評価結果では，λ_m と λ_c の様々な組み合わせによるモデルの評価結果の BLEU 値の上位 3 組の結果を挙げている．BLEU 値上位の結果は $\lambda_m < \lambda_c$ の範囲内に集中する． 4.2 節で提案した損失関数の式により，この理由は，訓練の収束時に L_m と L_c の関係は $L_m >>> L_c$ になるので，L_m の影響を減らすため，モデルを訓練する設定を $\lambda_m < \lambda_c$ とすると，文間意味的類似度タスクの損失 L_m の影響を抑えることができる。一方，λ_m をあまりにも小さくすると，L_m の影響がなくなるためモデルの性能が下がると考えられる．

6 関 連 研 究

関連研究として，Bing と Cortana の検索履歴を情報源とする大規模機械読解データセット MS MARCO [1] が挙げられる．抽出型機械読解のデータセットとは異なり，質問文の回答は，作業者がコンテキストを参照して作成している．MS MARCO を用いることにより，コンテキストに基づき回答を生成する機械読解タスクも提案され，生成型機械読解用評価基準として，ROGUE-L と BLEU-1 が用いられている．文献 [9] では，このタスクにおいて，Transformer による Encoder-Decoder モデルを提案し，発表当時のリーダーボード [8] 1 位を達成した (現時点では，人手によるベースラインを除いて 5 位となっており，それより上位の 4 個のモデルおよび論文は未公開である)．そのほか，回答を人手で作成した機械読解データセットとして，DuReader [4]，および，NarrativeQA [6] が存在する．

大規模機械読解の研究として，文献 [5] では, 検索結果上位のコンテキストと質問文を Encoder に入力し，全てのコンテキストの文脈情報を結合して Decoder に入力する手法を提案しており，大規模機械読解においても，生成型機械読解方式が重要で

8 : https://microsoft.github.io/msmarco/

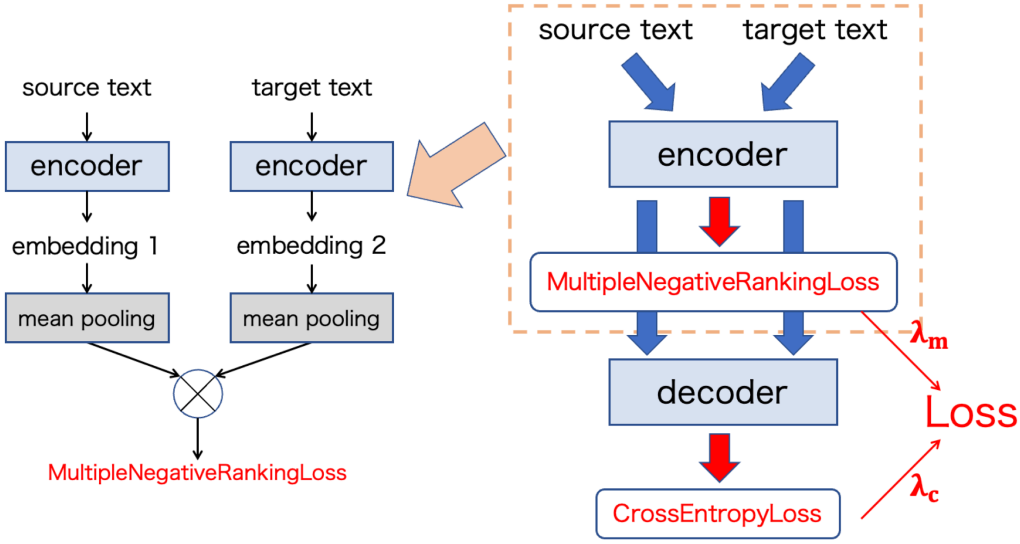

図 2　文間意味的類似度・回答生成のマルチタスク学習を行うモデルの枠組み

表 4　質問応答事例 (評価事例) における生成型モデルの評価結果

(a) 回答生成型データセットによる生成型モデルの評価結果

fine-tuning 用データセット	fine-tuning 方式	bleu-1	bleu-2	bleu-3	bleu-4	BP	BLEU	人手評価 正解/評価
回答生成型	シングルタスク	42.6	19.6	10.5	6.0	0.599	9.1	38/100

(b) 抽出型・生成型混合事例によるマルチタスク学習で訓練される生成型モデルの評価結果

fine-tuning 用データセット	fine-tuning 方式	bleu-1	bleu-2	bleu-3	bleu-4	BP	BLEU	人手評価 正解/評価
抽出型・生成型混合	マルチタスク	41.2	20.2	11.4	**6.7**	**0.917**	**14.6**	**60/100**

(c) 文間意味的類似度・回答生成によるマルチタスク学習で訓練される生成型モデルの評価結果

fine-tuning 用データセット	ロスに施した重み付け λ_m,λ_c	bleu-1	bleu-2	bleu-3	bleu-4	BP	BLEU	人手評価 正解/評価
回答生成型	0.5 , 0.5	42.8	20.8	11.4	6.6	0.842	13.6	55/100
	0.25 , 0.75	**43.7**	**21.6**	**12.1**	**6.7**	0.855	14.2	55/100
	0.10 , 0.90	42.5	20.7	11.8	6.5	0.878	14.1	55/100

あることを示している．

7　おわりに

本論文では，Transformer による Encoder-Decoder 型モデルである mT5 [17] を生成型機械読解モデルとして利用して，ノウハウ型機械読解へ適用した．wikihow の日本語版から，回答生成型機械読解方式に関する質問応答データセットを作成した．また，生成型モデルを fine-tuning するには，2 通りのマルチタスク学習方式の手法を提案した．評価結果により，生成型に関する質問応答事例において，2 通りのマルチタスク学習の手法とも，シングルタスク学習を上回る性能を達成した．

文　献

[1] P. Bajaj, D. Campo, N. Craswell, L. Deng, J. Gao, X. Liu, R. Majumder, A. McNamara, B. Mitra, T. Nguyen, M. Rosenberg, X. Song, A. Stoica, S. Tiwary, and T. Wang. MS MARCO: A human generatfed machine reading comprehension dataset. *arXiv preprint arXiv:1611.09268*, 2016.

[2] 陳騰揚, 前田竜治, 李宏宇, 銭澤長, 宇津呂武仁, 河田容英. 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. 言語処理学会第 26 回年次大会論文集, pp. 315–318, 2020.

[3] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova. BERT: Pre-training of deep bidirectional Transformers for language understanding. In *Proc. NAACL-HLT*, pp. 4171–4186, 2019.

[4] W. He, K. Liu, J. Liu, Y. Lyu, S. Zhao, X. Xiao, Y. Liu, Y. Wang, H. Wu, Q. She, X. Liu, T. Wu, and H. Wang. DuReader: a Chinese machine reading comprehension dataset from real-world applications. In *Proc. MRQA*, pp. 37–46, 2018.

[5] G. Izacard and E. Grave. Leveraging passage retrieval with generative models for open domain question answering. In *Proc. 16th EACL*, pp. 874–880, 2021.

[6] T. Kočiský, J. Schwarz, P. Blunsom, C. Dyer, K. Hermann, G. Melis, and E. Grefenstette. The NarrativeQA reading comprehension challenge. *Transactions of the Association for Computational Linguistics*, Vol. 6, pp. 317–328, 2018.

[7] Koupaee, Mahnaz, and W. Yang Wang. Wikihow: A large scale text summarization dataset. *arXiv preprint arXiv:1810.09305*, 2018.

[8] 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.

[9] K. Nishida, I. Saito, K. Nishida, K. Shinoda, A. Otsuka,

H. Asano, and J. Tomita. Multi-style generative reading comprehension. In *Proc. 57th ACL*, pp. 2273–2284, 2019.

[10] 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.

[11] C. Raffel, N. Shazeer, A. Roberts, K. Lee, S. Narang, M. Matena, Y. Zhou, W. Li, and P J. Liu. Exploring the limits of transfer learning with a unified text-to-text transformer. *Journal of Machine Learning Research*, pp. 1–67, 2020.

[12] P. Rajpurkar, R. Jia, and P. Liang. Know what you don't know: Unanswerable questions for SQuAD. In *Proc. 56th ACL*, pp. 784–789, 2018.

[13] N. Reimers and I. Gurevych. Sentence-BERT: Sentence embeddings using Siamese BERT-networks. In *Proc. 9th EMNLP-IJCNLP*, pp. 3982–3992, 2019.

[14] 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.

[15] K. Tymoshenko and A. Moschitti. Cross-pair text representations for answer sentence selection. In *Proc. EMNLP*, pp. 2162–2173. Association for Computational Linguistics, 2018.

[16] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. Gomez, L. Kaiser, and I. Polosukhin. Attention is all you need. In *Proc. 30th NIPS*, pp. 5998–6008, 2017.

[17] L. Xue, N. Constant, A. Roberts, M. Kale, R. Al-Rfou, A. Siddhant, A. Barua, and C. Raffel. mT5: A massively multilingual pre-trained text-to-text transformer. In *Proc. NAACL*, pp. 483–498, 2021.