原著論文 Original Paper

質問意図の明確化に着目した機械読解による質問応答手法の提案

Reading Comprehension based Question Answering technique by Focusing on Identifying Question Intention

大塚 淳史 Atsushi Otsuka
日本電信電話株式会社，NTT メディアインテリジェンス研究所
NTT Media Intelligence Laboratories, NTT Corporation
atsushi.otsuka.vs@hco.ntt.co.jp, https://researchmap.jp/a_otsuka/

西田 京介 Kyosuke Nishida
（同　上）
kyosuke.nishida.rx@hco.ntt.co.jp

斉藤 いつみ Itsumi Saito
（同　上）
itsumi.saito.df@hco.ntt.co.jp

浅野 久子 Hisako Asano
（同　上）
hisako.asano.fe@hco.ntt.co.jp

富田 準二 Junji Tomita
（同　上）
junji.tomita.xa@hco.ntt.co.jp

佐藤 哲司 Tetsuji Satoh
筑波大学 図書館情報メディア系
Faculty of Library, Information and Media Science
satoh@slis.tsukuba.ac.jp

keywords: Question answering, Reading comprehension, Question Generation, Deep neural network

Summary

The performance of reading comprehension, which is a question answering technique, by deep neural networks is now comparable to that of humans. However, there are still problems with the reading comprehension when given ambiguous questions. In this work, we propose a novel task called Specific Question Generation (SQG). SQG specifically revises the input question and suggests several specific question (SQ) candidates so that users can choose the SQ that is closest to their intent and obtain a highly accurate answer from the reading comprehension. We also propose a Specific Question Generation Model (SQGM) for facilitating the SQG. This model is based on an encoder-decoder model and uses two copy mechanisms (question copy and passage copy). The key idea here is that the missing information in the user-input question is described in the passage. Experimental results with public reading comprehension datasets demonstrated that our model generated specific questions that can improve reading comprehension accuracy.

1. は じ め に

人工知能がユーザの質問について回答する質問応答技術の必要性は，スマートフォンやスマートスピーカー，ロボットといったデバイスの普及に伴いますます高まっている．近年，機械読解と呼ばれる質問応答が特に注目を集めている [Hermann 15]．機械読解とは，システムが自然言語で記述された文書（パッセージ）を読み解き，そのパッセージに関する質問に対して，回答となる情報をパッセージ中から発見，抽出することで回答を可能とする質問応答技術である．

機械読解による質問応答は，深層学習に基づく手法により高い回答精度を達成することができることが知られている．例えば，Wikipedia の記事に対して，クラウドソーシングによって質問と回答を作成した機械読解のオープンデータセットである SQuAD [Rajpurkar 16] では，BiDAF [Seo 17] や QANet [Yu 18] といったニューラルネットワークによる End-to-End 型の機械読解モデルで人間に匹敵する回答精度を得られることが報告されている．

機械読解は自然言語による質問応答で高い精度を達成できる技術であるが，実際に機械読解を質問応答システムとして使用する場合を想定すると，依然として課題が存在する．従来の機械読解の回答精度は，SQuAD などのオープンデータセットを用いた実験によって得られたものである．しかしながら実際の質問応答システムでは，オープンデータセットの様に，回答を特定できるような明確な質問が必ずしも入力されるとは限らない．回答が

複数個考えられる曖昧な質問や，回答するための必要な情報が不足している短い質問が入力されたとき，現在の機械読解では十分な回答精度を得ることは難しい．

質問者が自身の質問意図を明確にできず，曖昧な質問をしてしまう場面は，人同士の質問応答でも発生する．例えば，コールセンターの質問者とオペレーターとのインタラクションでは，不慣れな質問者が，曖昧であったり，意図を特定できない質問をオペレーターに問い合わせてしまう場面がある．オペレーターは，マニュアル等の外部知識と質問者の質問を照らし合わせ，質問者の質問意図を推測し，質問者に確認する作業を行う．このときオペレーターは，質問者の質問をオペレーターが回答可能な具体的な質問に訂正し質問者に聞き返すことで，質問意図を確認する．本論文では，オペレーターと同様に，質問者の質問意図を推測し確認する仕組みを機械読解による質問応答システムに導入することで，曖昧な質問に対する自動応答を実現する．

質問応答において曖昧な質問が入力されたとき，質問者の質問意図の明確化を目的とするタスクとして「**改訂質問生成（Specific Question Generation：SQG）**」が提案されている [大塚 19]．改訂質問生成は，機械読解による質問応答において，パッセージと入力質問が与えられたとき，入力質問をパッセージの内容に基づいて，具体的な質問になるように書き換える．本論文では，改訂質問生成を用いた機械読解による質問応答手法を提案する．改訂質問生成を用いた質問応答を図 1 に示す．改訂質問生成により生成した具体化した質問である**改訂質問（Specific Question：SQ）**の候補を質問者に提示する．質問者は改訂質問の中から自身の質問意図に近い質問を選択することで，質問応答システムから精度の高い回答を得ることができる．

改訂質問生成を実行可能にするモデルとして，本論文では「**改訂質問生成モデル（Specific Question Generation Model: SQGM）**」を提案する．改訂質問生成モデルは，生成モデルである Encoder-Decoder モデル [Bahdanau 15] をベースに，機械読解モデル [Seo 17] とコピー機構 [Cao 17] を組み合わせた End-to-End 型のニューラルネットワークである．改訂質問生成モデルに入力質問と質問対象のパッセージを入力すると，改訂質問生成モデルは入力質問とパッセージの内容を機械読解モデルにより読み解き，関連する情報を抽出する．抽出した情報をアテンションによって注視しながら，生成モデルによって改訂質問を生成する．この際，コピー機構によって入力質問やパッセージ中の重要な単語や表現を改訂質問にコピーすることにより，パッセージの内容に即した改訂質問を生成している．本論文では改訂質問生成モデルの提案に加えて，文圧縮によって擬似的に短い質問を作成することによる学習データの自動生成手法についても提案する．

本論文の貢献：本論文では，以下の貢献を果たした．

- 改訂質問生成と機械読解を組み合わせた質問応答手法を提案した．パッセージと入力質問が与えられたとき，入力質問をパッセージの内容に基づいて具体化した改訂質問の候補を質問者に提示する．質問者は提示された改訂質問の中から自身の質問意図に近い質問を選択することで，質問応答システムから精度の高い回答を得ることができる．
- 改訂質問生成を実現する End-to-End のニューラルネットワークモデル（SQGM) を提案した．提案モデルは文生成と機械読解，コピー機構のニューラルネットワークを組み合わせることで，パッセージの内容に基づく改訂質問を生成する．
- 文圧縮により，機械読解コーパスの質問から曖昧な短い質問を作成することによる改訂質問生成モデルの学習データの自動生成手法を提案した．
- 機械読解のオープンデータセットである SQuAD および，日本語の機械読解質問応答コーパスに対して提案手法を適用し，入力質問が回答を特定できない程に短い場合に，機械読解の回答精度を改善できることを明らかにした．同時に，十分に長い入力質問の場合においても複数の改訂質問を提示することで，質問意図の異なる候補から，質問者の質問意図に沿った質問を選択できることを示した．

Passage: CBS provided digital streams of the game via CBSSports.com, and the CBS Sports apps on *tablets, Windows 10, Xbox One and other digital media players* (such as Chromecast and Roku). Due to Verizon Communications exclusivity, streaming on smartphones was only provided to *Verizon Wireless customers* via the NFL Mobile service. The ESPN Deportes Spanish broadcast was made available through WatchESPN.

Question: What app did viewers use to watch the game on their smartphones?
Answer of worker #1: NFL Mobile service
Answer of worker #2: the CBS Sports apps
Answer of worker #3: NFL

Specific Question #1: What app did viewers use to watch the game on their smartphones *of Verizon Wireless?*
Coressponding system answer: NFL Mobile service

Specific Question #2: What app did viewers use to watch the game on their *tablets?*
Coressponding system answer: CBS Sports apps

図 1 SQuAD における改訂質問生成のイメージ．改訂質問生成は，パッセージに関連する質問を生成するタスクであり，質問者は自身の質問意図に近い質問を選択することができる．

2. 問 題 定 義

本節では，本論文が取り組むタスク，関連する用語についての定義を行う．

[定義 1] 改訂質問生成（Specific Question Generation：SQG）は，入力質問とパッセージを入力とし，改訂質問（SQ）を出力するタスクである．改訂質問生成を用いた

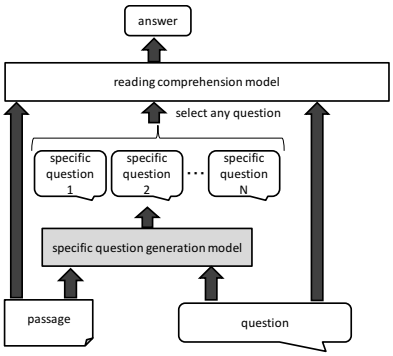


図 2　改訂質問生成と機械読解による質問応答.

質問応答では，質問者は複数の改訂質問の候補から自身の質問意図に合った改訂質問を選択することができる．

[定義 2] 機械読解（Reading Comprehension :RC）は，入力質問の単語トークン系列 q と，パッセージの単語トークン系列 x を入力とし，パッセージ中から回答の始点 s と終点 e を推定するタスクである．機械読解の出力となる回答はパッセージから，推定した回答範囲を抜き出した単語トークン系列となる．

[定義 3] 改訂質問生成モデル（Specific Question Generation Model : SQGM）は，入力質問の単語トークン系列 q と，パッセージの単語トークン系列 x を入力とし，改訂質問の単語トークン系列 y を出力するニューラルネットワークモデルである．ここで，各トークン系列の単語は改訂質問生成が扱う語彙集合（語彙数:V）に含まれるものとする．

[定義 4] 入力質問は，自然言語で記述された文である．入力質問は形態素解析等により，単語トークン列に分割し，one-hot ベクトルの系列である $q = \{q_1, q_2, ..., q_J\}$ として表す．one-hot ベクトルは，単語辞書のインデックスに対応する次元のみ 1，それ以外の次元を 0 とする V 次元のベクトルである．

[定義 5] パッセージは，自然言語で記述された文である．パッセージは数百語程度の単語から構成され，入力質問と同様に one-hot ベクトルで表される単語トークン列 $x = \{x_1, x_2, ..., x_T\}$ とする．ここで，パッセージは，入力質問の回答となる情報を含むものとする．

[定義 6] 改訂質問（Specific Question : SQ）は，入力質問の内容が具体化された文 $y = \{y_1, y_2, ..., y_K\}$ である．改訂質問は入力質問をパッセージ内容に沿って回答が一意に決定できるように具体化した質問である．

3. 提　案　手　法

本論文で提案する改訂質問生成と機械読解による質問応答を図 2 に示す．入力質問とパッセージを入力すると，機械読解モデルにより，回答がパッセージから抽出される．同時に，改訂質問生成モデルにより，入力質問とパッセージから複数の改訂質問の候補が生成される．質問者は機械読解が出力した回答に満足できない場合，改訂質問の候補のうち，いずれか 1 つを選択する．選択した改訂質問とパッセージを機械読解モデルに入力することで，新たな回答を得ることができる．本論文では，機械読解は任意の既存モデルを使用することを想定する．以降は，改訂質問生成モデルについて詳述する．

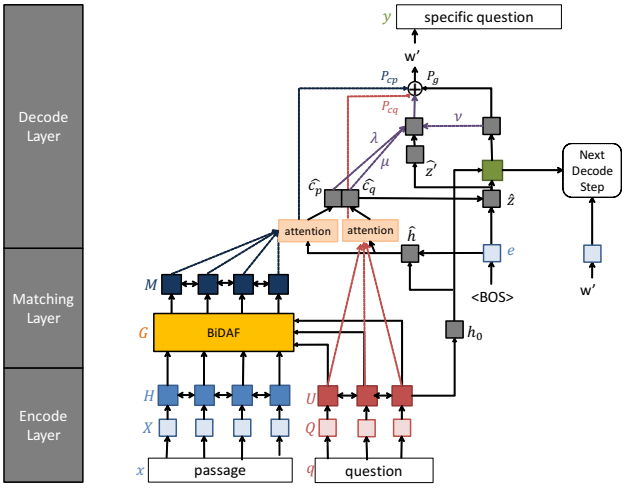


図 3　改訂質問生成モデル (specific question generation model : SQGM)．

本論文が提案する改訂質問生成モデルの構成を図 3 に示す．改訂質問生成モデルはアテンション付きのエンコーダ・デコーダモデルをベースに，機械読解モデルと 2 つのコピー機構を組み合わせたモデルである．[大塚 19] のモデルでは，コピー機構はパッセージからコピーを行う 1 つのコピー機構となっていたが，本論文で提案するモデルは入力質問からもコピーを行う．

改訂質問生成モデルは以下の 3 層によって構成される．

1. **Encoding layer** : パッセージと入力質問の単語トークン系列を，学習済みの単語埋め込みモデルにより連続値のベクトルに変換する．また，パッセージと入力質問のコンテキストを考慮したベクトル系列を作成する．
2. **Matching layer** : パッセージと入力質問の単語間の相互関係を捉えることにより，入力質問に依存した，パッセージのベクトル系列をモデリングする．
3. **Decoding layer** : アテンション付きリカレントニューラルネットワークと 2 つのコピー機構により，改訂質問を構成する単語トークン系列を生成する．

以降は各層について詳述する．

3·1 Encoding layer

Encoding layer の入力として，V 次元で表現される one-hot のベクトル系列で表されたパッセージ $x = \{x_1, ..., x_T\}$ と入力質問 $q = \{q_1, ..., q_J\}$ を与える．はじめに，one-hot ベクトルから，v 次元で表現される連続値ベクトルに変換する．単語の変換には学習済みのベクトル変換行列 $W_e \in \mathbb{R}^{v \times V}$ を用いる．変換されたベクトルは 2 層のハイ

ウェイネットワーク [Srivastava 15] を通して，最終的にパッセージと入力質問それぞれのベクトル系列 $X \in \mathbb{R}^{v \times T}$ と $Q \in \mathbb{R}^{v \times J}$ を得る．

連続値のベクトル系列に変換されたパッセージと入力質問を，リカレントニューラルネットワーク（RNN）に入力する．本論文では，RNN に 1 層の GRU [Chung 14] を使用する．ここで，GRU のパラメータはパッセージと入力質問で共有している．また，最初に入力したベクトルの影響が小さくなることを防ぐために，双方向の GRU（Bi-GRU）を使用する．Bi-GRU により，パッセージのコンテキスト行列 $H \in \mathbb{R}^{2d \times T}$ と入力質問のコンテキスト行列 $U \in \mathbb{R}^{2d \times J}$ を得る．ここで，d は隠れ状態の次元数である．

エンコーダの RNN の状態は，デコード時の RNN の初期状態を求めるためにも使用する．Decoding layer の RNN の初期状態として与えるベクトル $h_0 \in \mathbb{R}^{2d}$ は，入力質問のコンテキスト行列にアテンションを適用した以下の式により計算する．

$$h_0 = \textstyle\sum_{j<J} \alpha_j U_j, \tag{1}$$

ここで，$\alpha_j = \mathrm{softmax}_j(U_J^\top U_j)$ であり，$U_J^\top$ は，入力質問のコンテキスト行列の最終状態を表している．

3·2 Matching layer

Matching Layer では，Encoding Layer でエンコードしたパッセージと入力質問を照合し，パッセージ中から，入力質問に関連する領域を発見する．パッセージと入力質問の関連性を取得するモデルとして，本論文では，Bi-directional attention flow（BiDAF）[Seo 17] を使用する．BiDAF は，パッセージと入力質問の双方向からアテンションを計算し，最終的に入力質問に依存したパッセージのコンテキスト行列を作成する．

BiDAF ではまず，パッセージに関するコンテキスト行列 H と入力質問に関するコンテキスト行列 U により，類似度行列 $S \in \mathbb{R}^{T \times J}$ を以下の式により計算する．

$$S_{tj} = w_s^\top [H_t; U_j; H_t \odot U_j], \tag{2}$$

ここで，$w_s \in \mathbb{R}^{6d}$ は学習パラメータであり，$[;]$ は行成分のベクトルの連結を表している．次に，類似度行列を基に，**パッセージから入力質問へのアテンション**と，**入力質問からパッセージへのアテンション**の 2 方向のアテンションを計算する．

パッセージから入力質問へのアテンションでは，パッセージ中の単語について，入力質問の単語で重み付けしたベクトルを計算する．パッセージの t 番目の単語についてのアテンションベクトル $\check{U}_t \in \mathbb{R}^{2d}$ は次式によって計算する．

$$a_t = \mathrm{softmax}_j(S_t), \tag{3}$$

$$\check{U}_t = \textstyle\sum_j a_{tj} U_j, \tag{4}$$

入力質問からパッセージへのアテンションでは，入力質問のいずれかの単語に強く関連する単語に重みをつけたベクトル $\check{h}$ を，パッセージの系列長 T 分だけ並べた行列 $\check{H} \in \mathbb{R}^{2d \times T}$ を次式により計算する．

$$b = \mathrm{softmax}_t(\max_j(S)) \tag{5}$$

$$\check{h} = \sum_t b_t H_t \tag{6}$$

パッセージと入力質問の双方向のアテンションベクトルは，パッセージの各単語に対して，アテンションのベクトルを連結した次式により計算する．

$$G = [H; \check{U}; H \odot \check{U}; H \odot \check{H}] \in \mathbb{R}^{8d \times T}. \tag{7}$$

最終的に，双方向のアテンションベクトル G を 1 層の Bi-GRU に入力し，入力質問を考慮したパッセージのコンテキスト行列 $M \in \mathbb{R}^{2d \times T}$ を得る．

3·3 Decoding layer

Decoding layer では，Encoding layer と Matching layer の情報に基づいて，改訂質問を生成する．Decoding layer は，RNN 言語モデルに基づく生成モデルと，パッセージと入力質問のコピー機構を組み合わせたネットワークで構成される．

i. RNN 言語生成モデル

生成モデルは，アテンション付きの 1 層 GRU とソフトマックス層により構成される．改訂質問の単語列 $y = \{y_1, ..., y_s\}$ を入力したとき，生成モデルによって出力される，改訂質問の次の単語の確率分布 $P_g(y_{s+1}|y_{\leq s}, x, q)$ は次式によって表される．

$$P_g(y_{s+1}|y_{\leq s}, x, q) = \mathrm{softmax}(W_g h_{s+1} + b_g), \tag{8}$$

ここで，$W_g \in \mathbb{R}^{2d \times V_g}$ と $b_g \in \mathbb{R}^{V_g}$ は学習パラメータを表している．V_g は，生成モデルで生成する単語の語彙数を表しており，$V > V_g$ である．生成モデルから生成される単語は高頻度の単語のみとし，低頻度の単語はコピー機構により抽出的に生成することで，生成モデルのサイズを削減し，学習速度を向上させる．$h_{s+1} \in \mathbb{R}^{2d}$ は，GRU の $s+1$ 番目の隠れ状態であり，$h_{s+1} \leftarrow \mathrm{GRU}(h_s, \hat{z}_s)$ によって更新される．

GRU の入力ベクトル $\hat{z}_s$ は，生成モデルが出力した一つ前の単語 y_s に基づき，以下のように決定する（ここで，最初の入力は文の始端トークン $<BOS>$ とする）．y_s は，Encoding layer と同様に，one-hot ベクトル化した後，単語埋め込み層と 2 層のハイウェイネットワーク層によって連続値のベクトル $e_s \in \mathbb{R}^v$ となる．次に，アテンションを行うためのクエリとなるベクトル $\hat{h}_s \in \mathbb{R}^{2d}$ を，e_s と 1 つ前の GRU の隠れ状態 h_s を用いて次式により作成する．

$$\hat{h}_s = f(W_d[e_s; h_s] + b_d). \tag{9}$$

パッセージのアテンション $\alpha_{st} \in \mathbb{R}^T$ と，入力質問のアテンション $\beta_{sj} \in \mathbb{R}^J$ はクエリベクトルを用いて次式により計算する．

$$\alpha_{st} = \mathrm{softmax}_t(M_t \cdot \hat{h}_s), \tag{10}$$

$$\beta_{sj} = \mathrm{softmax}_j(U_j \cdot \hat{h}_s). \tag{11}$$

最終的に，GRU の入力 $\hat{z}_s \in \mathbb{R}^{v+4d}$ は次式により得る．

$$\hat{c}_{ps} = \textstyle\sum_t \alpha_{st} M_t, \tag{12}$$

$$\hat{c}_{qs} = \textstyle\sum_j \beta_{sj} U_j, \tag{13}$$

$$\hat{z}_s = [e_s; \hat{c}_{ps}; \hat{c}_{qs}], \tag{14}$$

ここで，$W_d \in \mathbb{R}^{(v+4d)\times 2d}$ と $b_d \in \mathbb{R}^{2d}$ は学習パラメータである．また，f は活性関数を表しており，本論文では PReLu[He 15] を使用する．

ii. コピー機構

コピー機構は，文生成の際に，入力の単語の一部をコピーすることにより，文章や対話などの文脈に一貫性を持たせた文を生成するための機構である．コピー機構を導入することで，入力質問やパッセージの内容に沿った改訂質問を生成する．要約や翻訳などでコピー機構が用いられる際はコピー元のテキストは 1 つであったが，本論文では入力質問とパッセージの両方をコピー元とする．

アテンションの重みを言語モデルの単語の生起確率と捉え [Cao 17]，コピー機構による単語の生成確率分布を次式により得る．

$$P_{cp}(y_{s+1}|y_{\leq s}, x, q) = \sum_t \mathbb{1}(y_{s+1} = x_t)\alpha_{(s+1)t}, \tag{15}$$

$$P_{cq}(y_{s+1}|y_{\leq s}, x, q) = \sum_j \mathbb{1}(y_{s+1} = q_j)\beta_{(s+1)j}, \tag{16}$$

ここで，$\mathbb{1}(y_s = x_t)$ は，$y_s = x_t$ のとき 1，それ以外のときは 0 となる関数である．$\mathbb{1}(y_s = q_j)$ も同様である．

iii. 言語生成モデルとコピー機構の統合

RNN 言語生成モデルとコピー機構では，それぞれで単語の生起確率の分布を得られる．最終的な出力単語を決定するための単語の確率分布 $P(y_{s+1}|y_{\leq s}, x, q)$ は，各々の確率分布の重み付き和によって計算される．

$$\begin{aligned} P(y_{s+1}|y_{\leq s}, x, q) = {} & \lambda_s P_g(y_{s+1}|y_{\leq s}, x, q) \\ & + \mu_s P_{cp}(y_{s+1}|y_{\leq s}, x, q) \\ & + \nu_s P_{cq}(y_{s+1}|y_{\leq s}, x, q), \end{aligned} \tag{17}$$

ここで，λ_s, μ_s, ν_s は $\lambda_s, \mu_s, \nu_s \in [0,1]$，$\lambda_s + \mu_s + \nu_s = 1$ の値をとる重みパラメータであり，以下の式に示す softmax 層の出力 $\gamma_s \in \mathbb{R}^3$ により決定する．

$$\hat{z}'_s = \mathrm{Highway}_2(\hat{z}_s), \tag{18}$$

$$\gamma_s = \mathrm{softmax}(W_c \hat{z}'_s + b_c), \tag{19}$$

$$\lambda_s = \gamma_{s0}, \quad \mu_s = \gamma_{s1}, \quad \nu_s = \gamma_{s2},$$

ここで，$\mathrm{Highway}_2()$ は 2 層のハイウェイネットワークを示している．また，$W_c \in \mathbb{R}^{(v+4d)\times 3}$ と $b_c \in \mathbb{R}^3$ は学習パラメータである．

Decoding layer が生成する確率分布をもとに，単語系列を生成することにより改訂質問を生成する．ここで，生成時にビームサーチを導入し，ビーム幅 b にしたがって，生成範囲を探索することで，複数の改訂質問の候補を生成する．

3·4 改訂質問生成モデル学習

改訂質問生成モデルの学習は，誤差関数 L を最小にするようにパラメータを更新することで行う．誤差関数 L は，負の対数尤度を用いた以下の式で計算する．

$$L = -\frac{1}{N} \sum_i \sum_s \log P(y^{(i)}_{s+1} | y^{(i)}_{\leq s}, x^{(i)}, q^{(i)}), \tag{20}$$

ここで，N はミニバッチのサイズであり，i はミニバッチの i 番目のサンプルを示すインデックスである．

3·5 文圧縮による学習データ作成

改訂質問生成モデルの学習には，入力となるパッセージと入力質問と，教師データとなる改訂質問を用意する必要がある．本論文では，SQuAD[Rajpurkar 16] などの機械読解コーパスから改訂質問生成モデルの学習データを作成する．機械読解コーパスの質問を，回答のための情報が十分に含まれている改訂質問とみなし，教師データとする．その上で，機械読解コーパスの質問から重要な情報を欠落させた質問を機械的に作成し，これを入力質問とする．

機械読解コーパスの質問を用いた学習データの作成手法として，[大塚 19] は，文節の結合を用いた手法を提案しているが，この手法は日本語の係り受け関係の特徴を利用したものであり，英語などには直接適用できない．本論文では，文の依存構造と整数計画法を用いた教師なしの文圧縮手法 [Filippova 08] により，英語等にも適用可能な学習データの作成手法を提案する．

文圧縮ではまず，入力文の依存構造を取得する．次に，解析結果から取得した文の構成単位（単語）について，文の先頭から順にインデックス番号を付与し，このインデックス番号を基に整数計画問題の定式化を行う．文圧縮の整数計画問題は以下の式により定義される．

$$\begin{aligned} \underset{a_i \in \{0,1\}}{\textbf{maximize}} \quad & \textstyle\sum_{i \leq L} w_i a_i \\ \textbf{subject to} \quad & \textstyle\sum_{i \leq L} a_i \leq l \\ & a_{\mathrm{parent}(i)} - a_i \geq 0 \ \ (1 \leq i \leq L) \\ & a_{\mathrm{argmax}(w_i)} = 0 \end{aligned} \tag{21}$$

ここで，$a_i = 1$ のとき，文中の i 番目の単語は選択される．$a_i = 0$ のとき i 番目の単語は選択されず，文圧縮によって削除される．L は文圧縮前の文の長さ（単語数）であり，l は文圧縮後の文の最大単語数である．$\mathrm{parent}(i)$ は，依存構造上で i 番目の単語の親ノードに当たる単語を示している．w_i は単語重みを表し，本論文では $w_i = \log(F/F(a_i))$

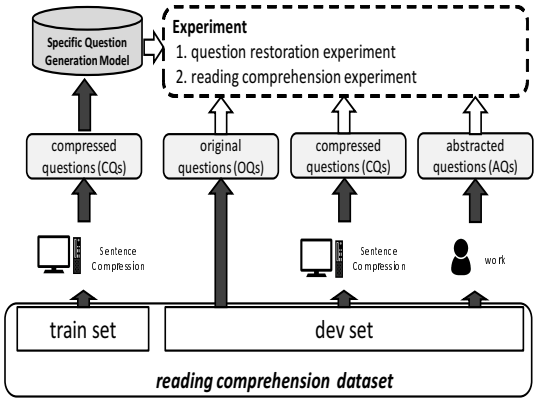


図 4 実験で使用する質問セットの概要.

で計算する．このとき，F は学習コーパス上の全単語の出現頻度の合計，$F(a_i)$ は i 番目の単語の出現頻度である．

文圧縮では，$\sum_{i \leq L} a_i \leq l$ の制約によって圧縮後の文の長さを決定する．本論文では，長さを $3 \leq l \leq L-2$ の範囲で変化させた時のそれぞれの長さで質問を圧縮する．$a_{\mathrm{parent}(i)} - a_i \geq 0$ の制約は依存構造で，親ノードの単語が選択されていないとき，子ノードの単語も選択されないようにするためのものである．$a_{\mathrm{argmax}(w_i)} = 0$ の制約は単語重みが最大の語を敢えて選択しないようにするための制約である．本論文では，回答のための情報が不足している質問を生成するために，質問中で重要と思われる単語を除外するように文圧縮を行っている．

文圧縮は，依存構造を取得できる言語であれば適用可能である．英語の場合は単語間の依存構造によって文圧縮を行うが，日本語の場合でも文節間の依存構造（係り受け関係）を用いることで同様に文圧縮を適用できる．

4. 評 価 実 験

本節では，改訂質問生成に関する評価実験について述べる．本論文では，**1. 質問の復元実験**と，**2. 改訂質問による機械読解実験**の 2 つの実験を行うことで提案手法の有効性について検証する．

本節では，まず評価実験に使用する質問について説明し，次に 2 つの実験について詳述する．

4·1 質 問 セット

本論文では，機械読解のオープンデータセットを用いた実験を行う．機械読解のオープンデータセットには，パッセージと質問，そして質問に対する回答が含まれる．本論文では，機械読解データセットの質問から，以下に示す 3 種類の質問を用意する．

1. **オリジナル質問（Original Questions : OQs）** 無加工の機械読解のデータセットの質問.

2. **圧縮質問（Compressed Questions : CQs）** オリジナル質問に対して，文圧縮によって機械的に短縮した質問.

表 1 質問の復元実験で使用する機械読解データセット．N はパッセージ数，質問数を示しており L はパッセージまたは質問の単語長である，

	train		dev		
	OQs	CQs	OQs	CQs	AQs
SQuAD					
N. passage	18,896	18,896	2,067	2,067	278
N. questions	87,599	1,235,000	10,570	32,208	350
L. passage	142.8	142.8	144.5	144.5	140.8
L. question	13.7	7.9	13.3	7.7	6.6
NewsQA					
N. passages	3,072	3,072	1,900	1,900	–
N. questions	10,510	69,959	3,000	16,569	–
L. passage	237.9	237.9	297.5	144.5	–
L. question	10.4	6.3	10.7	6.3	–

3. **要約質問（Abstracted Questions : AQs）** オリジナル質問に対して，人手によって，最低限の質問意図が特定できるまで短縮した質問．オリジナル質問から修飾語等の単語を削除していき，最終的に英語の基本 5 文型の範囲に収まるまで質問を短縮することで要約質問を作成している．このとき，短縮の結果質問意図は判別可能だが，回答が一意に定まらなくなる質問が作成されることも許容する．

各質問の利用に関する概要を図 4 に示す．質問改訂モデルの学習には，機械読解データセットの学習セットの質問を利用する．学習セットのオリジナル質問から圧縮質問を作成し，圧縮質問からオリジナル質問を復元するように改訂質問生成モデルを学習する．ここで，要約質問は改訂質問生成モデルの学習では使用しない．評価実験では，公開されている機械読解の開発セットをテストセットとして使用する．オリジナル質問から圧縮質問，要約質問を作成し各実験で使用する．

4·2 改訂質問生成モデルによる質問の復元実験

本節では，質問の復元実験について述べる．本実験ではまず，圧縮質問（CQs）および要約質問（AQs）を改訂質問生成モデルに入力し，改訂質問（SQs）を得る．次に，生成した改訂質問とオリジナル質問（OQs) の類似度を計算する．オリジナル質問と類似度が高い改訂質問を生成することができれば，改訂質問生成モデルは，入力質問に不足した情報を適切に補った改訂質問を生成できるモデルであるといえる．

以降はまず，実験で作成したデータセットについて述べる．次に改訂質問生成モデルの作成方法について説明した後，評価尺度と比較モデルについて詳述し，最後に実験結果を示す．

§1 データセット

質問の復元実験には，機械読解のオープンデータセットである SQuAD [Rajpurkar 16] [*1] と NewsQA [Trischler 17] [*2] を使用する．SQuAD は Wikipedia の記事をもとに作成された，機械読解の基本的なデータセットの一つ

*1 https://rajpurkar.github.io/SQuAD-explorer/
*2 https://datasets.maluuba.com/NewsQA

である．NewsQA はニュース記事によって作成された機械読解データセットであり，SQuAD の質問よりも短く，抽象的な質問が多いという特徴を持つ．

データセットの詳細を表 1 に示す．ここで，各圧縮質問は文圧縮により，単語長の制約を 3 から $L-2$（L は入力文の最大単語長）まで変化させることで，1 オリジナル質問に対して，複数の圧縮質問を作成している．文圧縮の際に必要となる依存構造は，言語解析器 spaCy[*3] によって取得した．テストセットについては，オリジナル質問が長いほど多くの圧縮質問を生成することから，評価の偏りが発生するため，10 個以上の圧縮質問が生成された場合は，10 個になるようにランダムに抽出を行った．また，NewsQA には学習セットとテストセットの区別が存在しないため，学習セットとパッセージが重複しない 3000 質問をランダムで抽出して作成している．

§2 改訂質問生成モデル学習

質問改訂モデルの Encoding layer に入力する，学習済みの単語埋め込みモデルとして，Wikipedia の全記事と SQuAD，NewsQA の全パッセージ，質問を結合したコーパスから作成した，300 次元の fastText [Bojanowski 17] のモデルを使用する．

質問改訂生成モデルの学習は，2 つの GPU（Quadro P6000）を使用した. 学習のためのミニバッチ数は 96 に設定し，隠れ状態の次元数 d は 128 とした. 生成モデルの出力語彙数 V_g は 5000 とし，SQuAD と NewsQA のそれぞれの質問で出現回数上位 5000 単語（質問の全語彙の 90% をカバー）を生成する. 学習の際のドロップアウト [Srivastava 14] の確率は，ハイウェイネットワークの入力層のみ 0.2 とし，残りの層を 0.5 に設定した. パラメータの最適化には Adam を使用し，学習の反復回数は SQuAD が 5 回（全 64,300 学習ステップ），NewsQA が 20 回（全 14,560 学習ステップ）である．

§3 評価尺度

改訂質問とオリジナル質問の類似度を計算指標として，**精度**，**再現率**，**F 値**，**意味的類似度**を用いる．精度は，改訂質問の全単語のうち，オリジナル質問の単語が含まれている割合を示しており，再現率は，オリジナル質問の全単語に対する，改訂質問の単語の網羅度を示している．F 値は，精度と再現率の調和平均である．意味的類似度は，改訂質問とオリジナル質問の文ベクトルを作成したときのコサイン距離により計算する．ここで，文ベクトルの作成には universal sentence encoder [Cer 18] を用いる．

オリジナル質問と比較する改訂質問は，改訂質問生成モデルのビームサーチのビーム幅を 1 としたときに生成される，1 つの改訂質問である．

§4 比較モデル

質問の復元実験において，提案モデル（**proposed**）と比較するモデルとして，提案モデルから，一部の層を除外したモデルを使用する．まず，コピー機構を除外したモデルとして，Decoding layer のすべてのコピー機構を除外したモデル (**w/o copy**)，パッセージからのコピーのみを除外したモデル (**w/o p_copy**)，質問からのコピーのみを除外 (**w/o q_copy**）した 3 つのモデルを用意する．ここで，質問からのコピーを除外したモデル (w/o q_copy) は，[大塚 19] の提案モデルとほぼ同一のモデルとみなすことができる．次にコピー機構はそのままにし，Matching layer の BiDAF モデルを除外したモデル (**w/o BiDAF**）を用意する．本モデルでは，デコード時のアテンションとして，提案モデルで使用した BiDAF のモデリング層 M の代わりに，Encoding layer のパッセージのコンテキスト行列 H を使用する．

*3 https://spacy.io/

表 2 質問の復元実験に関する実験結果.

	Model	Precision	Recall	F1	Sim
SQuAD	CQs	1.00	0.586	0.715	0.758
	: w/o copy	0.214	0.203	0.205	0.256
	: w/o p_copy	0.346	0.301	0.319	0.361
	: w/o q_copy	0.669	0.539	0.619	0.770
	: w/o BiDAF	**0.856**	0.678	0.747	0.814
	: **Proposed**	0.824	**0.718**	**0.758**	**0.834**
	AQs	1.00	0.626	0.754	0.794
	: w/o copy	0.265	0.282	0.260	0.345
	: w/o p_copy	0.365	0.363	0.353	0.436
	: w/o q_copy	0.707	0.570	0.647	0.621
	: w/o BiDAF	**0.900**	0.650	0.735	0.810
	: **Proposed**	0.844	**0.716**	**0.759**	0.**837**
NewsQA	CQs	1.00	0.581	0.715	0.715
	: w/o copy	0.281	0.231	0.250	0.232
	: w/o p_copy	0.702	0.597	0.637	0.695
	: w/o q_copy	0.623	0.541	0.568	0.658
	: w/o BiDAF	**0.832**	0.631	0.711	0.738
	: **Proposed**	0.805	**0.690**	**0.731**	**0.766**

§5 実験結果

質問の復元実験における実験結果を表 2 に示す．ここで，各モデルの CQs，AQs に該当するカラムの結果は，入力の圧縮質問または要約質問と，正解であるオリジナル質問を比較した結果である．

SQuAD，NewsQA のどちらのデータセットにおいても，精度を除くすべての項目で提案手法（proposed）が他の比較モデルよりも，有意に高いスコアを示している (t 検定; $p < .01$)．精度のみ，BiDAF を使用してないモデルのほうが良い結果となっている．BiDAF を使用していないモデルは，精度が高い一方で，再現率が低くなっている．これは，比較的短い改訂質問しか生成できていないことを表している．しかし，意味的類似度（Sim）では，提案手法のほうが高いスコアになることから，提案手法が最もオリジナル質問に近い改訂質問を生成できているといえる．

実験結果では，コピーを使用していない場合，入力の圧縮質問や要約質問よりも低いスコアとなっている．このことから，改訂質問生成において，コピー機構が重要な役割を果たしていることがわかる. SQuAD では，パッセージからのコピーを除外（w/o p_copy）したときのス

表 3 改訂質問による機械読解で使用する日本語機械読解データセット．N はパッセージ数，質問数を示しており L はパッセージまたは質問の単語長である．

	train		dev	
	LoQs	ShQs	LoQs	ShQs
jpWiki				
N. passage	1,236	1,236	290	290
N. questions	4,176	4,176	933	933
L. passage	295.3	295.3	300.2	300.2
L. question	18.8	9.7	19.1	9.8

コアの低下が大きい，これは，表 1 に示す通り，SQuAD の質問は比較的長い質問が多く，質問内容もパッセージ内容に沿っていることからパッセージからのコピーが重要な役割と果たしたといえる．一方で，NewsQA では，質問からのコピーを除外（w/o q_copy）したときの影響が大きくなっている．これは，NewsQA の質問が SQuAD ほど長くない（オリジナル質問と圧縮質問の単語長の差が小さい）ことに加え，抽象的な質問が多いため，パッセージからのコピーが SQuAD ほど有効に機能しなかったためであると考えられる．

4·3 改訂質問による機械読解実験

本節では，改訂質問を実際の機械読解モデルに入力したときの，機械読解の回答精度に関する実験を行う．まず，オリジナル質問，圧縮質問，要約質問を機械読解モデルに入力し，出力された回答の正解率を得る．次に，各質問を改訂質問生成モデルに入力し，改訂質問を作成する．作成した改訂質問を機械読解モデルに入力し，回答の正解率を得る．改訂前の質問（オリジナル質問，圧縮質問，要約質問）と改訂後の改訂質問の機械読解の正解率を比較し，改訂質問による正解率の変化について検証する．また本実験では，日本語の Wikipedia から作成した質問（短文質問，長文質問）を用いた実験も行う．

以降はまず実験で使用するデータセットについて説明する．次に，実験で使用する機械読解モデル，評価尺度について述べ最後に実験結果について詳述する．

§1 データセット

本実験では，英語と日本語の機械読解データセットを使用する．英語の実験には SQuAD を用いる．SQuAD を用いた実験では表 1 に示した，SQuAD のテストセットの全 2,067 パッセージ中の 10,570 のオリジナル質問とオリジナル質問から作成した 32,208 の圧縮質問，そしてランダムに抽出した質問から作成した 350 の要約質問を使用する．

日本語の実験として，Wikipedia の日本語版から独自に機械読解のデータセット（jpWiki）を作成した．日本語版 Wikipadia の中からランダムに 600 記事を抽出し，1 記事に対して最大 10 パッセージ程度になるように記事をパッセージ単位に分割した．そして，各パッセージに対し，機械読解用の質問と回答をクラウドソーシングによって作成した．

本実験用の質問として，人手で 1 から作成した短い質問である短文質問も作成する．圧縮質問と要約質問はオリジナル質問から作成した質問である．特に，圧縮質問は文圧縮により機械的に生成した質問であり，非文等も含まれる可能性があるため，実際に質問者が入力する質問と乖離がある可能性がある．そこで，短文質問を作成することで，より実際に質問で改訂質問生成の効果を検証する．意図的に短い質問と長い質問を作成するために，以下の手順により質問と回答を作成した．

(1) 作業者 A がある 1 パッセージを読み，質問と回答を作成する．このとき，作業者 A には，できるだけ曖昧な質問を，文字数制限以内（25 文字）で作成するように教示を行う．
(2) 作業者 B は，作業者 A と同じパッセージに対して，作業者 A が作成した質問と回答を閲覧して，回答が同じになり，かつ閲覧した質問よりも具体的になるような質問を作成する．このとき，作業 B が作成する質問は文字数の下限（30 文字）を設定し，下限以上の文字数となるようにする．上限は設定しない．

ここで，作業者 A が作成した質問を短文質問（Short Questions ：ShQs)，作業者 B が作成した質問を長文質問（Long Questions ：LoQs) とする．機械読解実験で作成した日本語データセットについて，表 3 に示す．作成したデータセットのうち，ランダムに 100 記事を抽出し，抽出した記事に含まれるパッセージと質問，回答をテストセットとし，残りを学習セットとした．学習セットは後述する機械読解モデルおよび改訂質問生成モデルの学習に使用する．なお，jpWiki はデータセットに短文質問と長文質問が含まれるため，機械的に短い質問を生成する文圧縮によるデータ作成は行わない．

§2 機械読解モデル・改訂質問生成モデル

本実験では，学習済みの機械読解モデルを使用する．英語の機械読解モデルとして，Web でオープンに公開されている SQuAD で学習した BiDAF の機械読解モデル [Gardner 18]*4を使用する．日本語については，[Nishida 18] が作成した日本語のニュース記事（jpNews）に関する機械読解データに，表 3 の学習セットのデータを加えたデータによって学習した BiDAF モデルを用いる．ここで，機械読解の学習には長文質問を用いた．

英語では 4·2·2 節で作成した SQuAD の改訂質問生成モデル（**proposed**）を使用する．日本語の改訂質問モデルとして，jpNews から圧縮質問を作成したものに，jpWiki の学習セットから短文質問を入力，長文質問を教師としたときのデータを加えた学習データ（全 1,673,000 入力質問）を用いて学習を行った．改訂質問生成モデルおよび学習パラメータは 4·2·2 節と同様である．

§3 評価尺度

機械読解の回答の正解率を測る指標として，完全一致率を使用する．完全一致率は，機械読解モデルが出力し

*4 http://allennlp.org/models

表 4 SQuAD における機械読解の回答精度の比較結果.

	OQs			CQs			AQs		
	Success@k	Accuracy	Improvement	Success@k	Accuracy	Improvement	Success@k	Accuracy	Improvement
Input	0.645	0.645	–	0.347	0.347	–	0.445	0.445	–
SQ1	0.663	**0.595**	0.0520	0.482	**0.389**	**0.207**	0.501	**0.455**	0.108
SQ2	0.684	0.555	0.0811	0.519	0.374	0.194	0.542	0.422	0.134
SQ3	0.698	0.544	0.0869	0.541	0.363	0.189	0.571	0.402	0.134
SQ4	0.710	0.542	**0.100**	0.558	0.363	0.191	0.580	0.403	**0.160**
SQ5	**0.721**	0.522	0.0957	**0.571**	0.343	0.181	**0.589**	0.420	0.134

表 5 jpWiki における機械読解の回答精度の比較結果.

	LoQs			ShQs		
	Success@k	Accuracy	Improvement	Success@k	Accuracy	Improvement
Input	0.668	0.668	–	0.509	0.509	–
SQ1	0.680	**0.584**	0.0355	0.628	**0.547**	0.242
SQ2	0.700	0.553	0.0710	0.694	0.535	0.306
SQ3	0.721	0.507	0.0903	0.729	0.537	**0.308**
SQ4	0.734	0.473	0.0903	0.744	0.482	0.249
SQ5	**0.744**	0.437	**0.100**	**0.750**	0.423	0.236

た回答と，正解データの回答の文字列が完全一致した質問の割合である．機械読解の評価では，文字列の部分一致の指標（F 値）を用いる場合もあるが，本実験では，改訂質問により，回答の特定性が高まることを検証するため，より厳密に正解を判定できる完全一致率を採用する．

本論文が提案する改訂質問生成では，1 つの入力質問に対して，複数の改訂質問を提示する．本実験では，改訂質問生成モデルのビームサーチのビーム幅を 5 に設定し，5 つの改訂質問を生成したときの完全一致率を計測する．複数の改訂質問の完全一致率を評価する指標として，**success@k**，**accuracy**，**improvement** を導入する．success@k は，評価セットの全質問中，入力質問と上位 k 位までの改訂質問のうちどれか一つでも正解と完全一致となった回答が出力された質問の割合である．**accuracy** は，入力質問および各順位の改訂質問単体での完全一致率である．**improvement** は，入力質問の機械読解の回答が不正解のときの，各順位の改訂質問の回答結果の accuracy である．success@k と improvement により，本来の入力質問では完全一致できなかった質問について，改訂質問により完全一致可能となった質問の割合を計測する．

§4 実験結果

改訂質問による機械読解実験の結果を表 4，表 5 に示す．SQuAD データにおける実験結果である表 4 では，圧縮質問（CQs）または要約質問（AQs）を機械読解に入力したときの完全一致率（accuracy）は，オリジナル質問（OQs）を入力したときから低下している．各質問から改訂質問を作成した後の機械読解では，圧縮質問と要約質問の第 1 位の改訂質問（SQ1）が高い accuracy となっている．一方で，オリジナル質問を入力したとき，改訂質問はいずれの順位においても accuracy が低下している．しかし，入力質問および改訂質問のいずれかで完全一致となる success@k では，いずれの入力質問の場合でも，入力質問のみのときから success@5 で 10 ポイント以上の向上が見られた．また，入力質問が完全一致しないときの，各順位での改訂質問の accuracy である improvement では，圧縮質問が順位の高い改訂質問で高いスコアとなった一方で，オリジナル質問と要約質問では，下位の改訂質問のスコアが高くなるという結果となった．

日本語データである jpWiki を用いた実験結果（表 5）においても，SQuAD と同様の傾向が見られた．短文質問を入力した際には，短文質問そのものよりも，短文質問から作成した改訂質問（SQ1）の方が accuracy が高くなった．一方で，長文質問から作成した改訂質問では accuracy が低下する結果となった．success@k において短文質問を入力したとき，success@2 で長文質問の accuracy を上回り，success@5 では 75%の正解率となっている．

機械読解の回答精度に関して，圧縮質問（CQs），要約質問（AQs），短文質問（ShQs）と，改訂質問（SQ1）の accuracy について，正解・不正解に関する McNemar 検定を実施した．圧縮質問と短文質問については，それぞれ $p < .01$，$p < .05$ で SQ1 との accuracy に有意差があることを確認した．一方，要約質問については，SQ1 との accuracy の差について，$p > .1$ で有意な差があるとは認められなかった．そこで要約質問について，SQ1 と SQ2 のいずれかが正解であるときの accuracy と比較した結果 $p < .05$ で有意差があることが認められた．

§5 短文質問に対する改訂質問生成に関する考察

英語（SQuAD），日本語（jpWiki）どちらのデータセットにおいても，圧縮質問，要約質問，短文質問を機械読解に入力すると，オリジナル質問，長文質問を入力したときよりも回答精度（accuracy）が低下している．このことから，機械読解では，質問が短かったり，質問内に含まれる情報が不十分なとき，十分な性能が発揮できていないといえる．圧縮質問・要約質問（短文質問）から作成した改訂質問により，機械読解の回答精度が改善されている．これは改訂質問生成モデルが，曖昧な入力質

なお、ちり・ほこり・しみの主な原因は、人間、ペット、観葉植物などやその活動である。また、衣服・カーテン・カーペット類の劣化による繊維ごみもある。人やペットの体毛、皮脂、汗、糞尿。また食品の残滓やハネ。他にもインク類のハネこぼれや子供の落書き、区域外から持ち込まれたり舞い込んだりした土埃、粉塵、花粉などがある。ちり・ほこり・しみ等は、ダニ・ノミ・雑菌類の繁殖培地となり、虫刺され、かぶれ、喘息や、各種炎症・感染症の原因となるほか、**カビの発生、腐敗等**により悪臭の原因や什器類の劣化の原因となる。地方によっては**蚊の発生が**マラリアなどの深刻な感染症の原因となるため、観葉植物用の潅水容器、屋外の雨どいや排水設備、防火水槽、エアコン室外機などに溜まった雨水などにボウフラが発生しないよう、とりわけ注意が必要である。電気まわりの汚れ・粉塵などは絶縁不良、トラッキングを引き起こし、火災の原因となるので掃除を行う。コンピュータやテレビ、換気扇など電気製品、作業機械や工作機械などが粉塵に汚染された状態で放置されていると排熱不良や稼動部の消耗、咬合不良などにより故障の原因となるので掃除を行う。
Ground-truth answers: 蚊の発生

Short and specific questions (ShQ, SQ)	**Answers from BiDAF**
ShQ1 :何が深刻な感染症の原因となる？	**カビの発生、腐敗等**
SQ1:地方によっては何がマラリアなどの深刻な感染症の原因となる？	**蚊の発生**

Copied from passage Copied from question Generated

図 5 jpWiki における改訂質問による機械読解の例.

Following their loss in the divisional round of the previous season 's playoffs , the Denver Broncos underwent numerous coaching changes , including a mutual parting with head coach **John Fox** (who had won four divisional championships in his four years as Broncos head coach) , and the hiring of **Gary Kubiak** as the new head coach. (..)
Ground-truth answers: Gary Kubiak, Gary Kubiak, Kubiak
OQ: who is the head coach of the broncos ?

Original and specific questions (OQ, SQ)	**Answers from BiDAF**
SQ1 : who is the head coach of the denver broncos ?	**John Fox**
SQ2 : who is the head coach of the broncos ?	**John Fox**
SQ3: who is the head coach of the denver ?	**John Fox**
SQ4 : who is the new head coach of the broncos ?	**Gary Kubiak**
SQ5 : who is the new head coach of the denver broncos ?	**Gary Kubiak**

Copied from passage Copied from question Generated

図 6 SQuAD における改訂質問による機械読解の例．‘(...)’ は省略したパッセージを表している．

問の質問意図を適切に読み取り，回答に必要となる情報を補完できているからであると考えられる．一方，入力質問が短いほど，改訂質問生成による具体化のバリエーションは増加すると考えられる．圧縮質問・要約質問（短文質問）のように入力質問が短い場合，改訂質問を 1 つ提示するだけでも，入力質問による機械読解よりも高い回答精度が得られることが期待できるが，必ずしも質問者の質問意図どおりに質問が改訂されるとは限らないため，改訂質問生成を用いる場合，2 個以上の改訂質問を提示し，質問者に選択させるほうが望ましいと考えられる．

§6 長文質問に対する改訂質問生成に関する考察

オリジナル質問や長文質問から改訂質問を作成した場合，機械読解の回答精度（accuracy）は，入力質問をそのまま使用したときよりも低下している．これは，改訂質問が回答に不必要な情報を追加してしまったり，回答自体を改訂質問に追加してしまったことが原因であると考えられる．一方で，入力質問を使用したときの機械読解の回答が不正解だったとき，ビームサーチにより生成した下位の改訂質問の方が回答精度の向上に貢献（improvement）した．上位で生成する改訂質問は，入力質問の質問意図を補強するような改訂質問を生成するが，下位の改訂質問で，異なる回答となる可能性がある改訂質問を提示していると考えられる．

§7 改訂質問生成に関する実例

図 5 は jpWiki の短文質問について，機械読解と改訂質問生成を適用した例である．短文質問を入力したとき，機械読解は比較的正解に近い回答を抽出しているが，改訂質問生成により “マラリア” など特徴的な単語を含んだ改訂質問に変換されたことで，本来の正解となる回答を機械読解が抽出できるようになっている．

図 6 は，SQuAD のオリジナル質問を機械読解と改訂質問生成に入力したときの結果である．1 位の改訂質問（SQ1) では，チーム名が正式名称である “denver broncos” に変化しているが，回答は “John Fox” のまま変化していない．しかし，4，5 位での改訂質問では “the new head coach” という情報が追加されたことにより機械読解の回答が “Gary Lubiak” に変化している．

5. 関 連 研 究

5·1 質問の書き換え・パラフレーズに関する研究

質問応答の回答精度の向上を目的に，質問文を書き換える研究では，Buck らの研究がある [Buck 18]．Buck らは，質問者と質問応答システムとの間にブラックボックスで質問書き換えモデルを設置し，入力質問を質問応答システムが回答可能な質問に書き換えてから質問応答システムに入力することで，回答精度を高めている．質問書き換えモデルには，生成モデル（Seq2Seq）を用いて，回答精度を最大化するように強化学習を行うことでモデルパラメータの学習を行う．Buck らの手法では，事前学習済みの質問の書き換えモデルが必要となる．しかしながら，モデル学習のための言い換えコーパスを用意することが難しいという課題がある．また，書き換えモデルを学習したコーパスに含まれていない話題に関する質問

が入力されたとき，後段の質問応答システムに無関係な質問が生成されてしまうという可能性がある．本論文が提案する改訂質問生成モデルは，言い換えコーパス等の別のデータセットが不要であるという点が異なる．また，後段の機械読解で使用するパッセージ中の単語をコピーすることにより，改訂質問を生成するため，パッセージと無関係な質問が生成されにくいという特徴がある．

ユーザの入力を書き換える研究は，クエリ拡張として情報検索において多く取り組まれている．情報検索型の質問応答では，入力質問を解析し，質問に対する適合文書を発見することで，回答となる情報を抽出する．このとき，より文書に適合しやすくするために，クエリ拡張やクエリの変形などの処理を行う．しかし，これらの処理は知識源の冗長性に依存して行われるので [Brill 02, Lin 07]，ある特定の知識について質問を書き換える必要のある機械読解では有効に機能しない．

言い換え文を作成するパラフレーズも，質問の書き換えに関するタスクの一つといえる．Dong らは，言い換え文生成と質問応答を同時に行うモデルを構築し，学習を行うことで入力質問と同じ回答が得られるような言い換え文を生成する手法を提案している [Dong 17]．Gupta らの手法では，Variational Autoencoder（VAE）を用いて，言い換え文の変換スタイルを学習させることで，文生成時には，スタイルを変えることで様々なバリエーションの言い換え文を生成している [Gupta 18]．パラフレーズは質問の意味を変えないまま，異なる語彙を使用した質問を作成するタスクであり，質問を具体化する改訂質問生成のタスクにパラフレーズを適用することはできない．

5·2 質問生成に関する研究

質問生成は，主に教育分野で活用することを目的に，機械読解の逆問題として，パッセージと回答を与えたときに質問を自動生成するタスクとして取り組まれている．Du らは回答を含んだ文を入力すると，自然言語の質問を生成する手法を提案している [Du 17]．また，文中に BIO タグと共参照関係のタグを埋め込むことで，パラフレーズ単位での回答に対する質問生成手法を提案している [Du 18]．Duan らは，パッセージを入力したときに，回答となる文を選択するタスクを質問生成と同時に扱うことで質問生成の性能を向上させる手法を提案している．

質問生成は，パッセージなどの自然文だけでなく，画像を入力したときに，画像に関する質問を生成するタスク [Mostafazadeh 16, Jain 18] や，知識ベースから質問を作成するタスク [Serban 16, ElSahar 18] も行われている．改訂質問生成は，知識源となるパッセージの他に，質問文自体を入力に与えるという違いがある．

5·3 適合性フィードバックに関する研究

適合性フィードバックは，情報検索における手法の一つであり，ユーザからのフィードバックをもとに検索結果やユーザのクエリを書き換える手法である [Manning 08]．適合性フィードバックを利用するためには，ユーザが自身の情報要求や質問意図に合致する文書をシステムにフィードバックするか，事前に用意したクエリログが必要となる．しかし，改訂質問生成はユーザからのフィードバックやクエリログを必要とせず，入力質問を理解し，具体的に書き換えることができるという違いがある．

6. ま と め

本論文では，改訂質問生成を用いた，機械読解による質問応答手法を提案した．改訂質問生成では，入力質問をパッセージの内容を元に具体化し，改訂質問として質問者に提示する．質問者は，曖昧であったり短い質問を質問応答に入力した場合でも，情報が補完された改訂質問の候補から，自身の質問意図に近い改訂質問を選択することで，質問意図に適合する回答を得ることができる．

機械読解コーパスに提案手法を適用した評価実験では，短い質問を入力した際には，改訂質問生成モデルが出力する改訂質問を使用することにより，機械読解の回答精度を改善できることを示した．また，十分に具体的な質問であっても，改訂質問生成が異なる質問意図の改訂質問を生成するため，自身の質問意図に沿った質問に対する回答を選択可能になることを確認した．

今後は，改訂質問の生成精度の向上や，機械読解と改訂質問生成モデルを End-to-End で行うモデルの検討を行う予定である．

◇ 参 考 文 献 ◇

[Bahdanau 15] Bahdanau, D., Cho, K., and Bengio, Y.: Neural Machine Translation by Jointly Learning to Align and Translate, in *International Conference on Learning Representations (ICLR)* (2015)

[Bojanowski 17] Bojanowski, P., Grave, E., Joulin, A., and Mikolov, T.: Enriching Word Vectors with Subword Information, *Transactions of the Association for Computational Linguistics (TACL)*, Vol. 5, pp. 135–146 (2017)

[Brill 02] Brill, E., Dumais, S. T., and Banko, M.: An Analysis of the AskMSR Question-Answering System, in *Empirical Methods in Natural Language Processing (EMNLP)*, pp. 257–264 (2002)

[Buck 18] Buck, C., Bulian, J., Ciaramita, M., Gajewski, W., Gesmundo, A., Houlsby, N., and Wang, W.: Ask the Right Questions: Active Question Reformulation with Reinforcement Learning, in *International Conference on Learning Representations (ICLR)* (2018)

[Cao 17] Cao, Z., Luo, C., Li, W., and Li, S.: Joint Copying and Restricted Generation for Paraphrase, in *Association for the Advancement of Artificial Intelligence (AAAI)*, pp. 3152–3158 (2017)

[Cer 18] Cer, D., Yang, Y., Kong, S., Hua, N., Limtiaco, N., John, R. S., Constant, N., Guajardo-Cespedes, M., Yuan, S., Tar, C., Sung, Y., Strope, B., and Kurzweil, R.: Universal Sentence Encoder, *CoRR*, Vol. abs/1803.11175, (2018)

[Chung 14] Chung, J., Gülçehre, Ç., Cho, K., and Bengio, Y.: Empirical Evaluation of Gated Recurrent Neural Networks on Sequence Modeling, *CoRR*, Vol. abs/1412.3555, (2014)

[Dong 17] Dong, L., Mallinson, J., Reddy, S., and Lapata, M.: Learning to Paraphrase for Question Answering, in *Empirical Methods in Natural Language Processing (EMNLP)*, pp. 875–886 (2017)

[Du 17] Du, X., Shao, J., and Cardie, C.: Learning to Ask: Neural Question Generation for Reading Comprehension, in *Association for*

Computational Linguistics (ACL), pp. 1342–1352 (2017)
[Du 18] Du, X. and Cardie, C.: Harvesting Paragraph-level Question-Answer Pairs from Wikipedia, in *Association for Computational Linguistics (ACL)*, pp. 1907–1917 (2018)
[ElSahar 18] ElSahar, H., Gravier, C., and Laforest, F.: Zero-Shot Question Generation from Knowledge Graphs for Unseen Predicates and Entity Types, in *North American Association for Computational Linguistics (NAACL)*, pp. 218–228 (2018)
[Filippova 08] Filippova, K. and Strube, M.: Dependency Tree Based Sentence Compression, in *International Natural Language Generation Conference (INLG)*, pp. 25–32 (2008)
[Gardner 18] Gardner, M., Grus, J., Neumann, M., Tafjord, O., Dasigi, P., Liu, N. F., Peters, M. E., Schmitz, M., and Zettlemoyer, L.: AllenNLP: A Deep Semantic Natural Language Processing Platform, *CoRR*, Vol. abs/1803.07640, (2018)
[Gupta 18] Gupta, A., Agarwal, A., Singh, P., and Rai, P.: A Deep Generative Framework for Paraphrase Generation, in *Association for the Advancement of Artificial Intelligence (AAAI)*, pp. 5149–5156 (2018)
[He 15] He, K., Zhang, X., Ren, S., and Sun, J.: Delving Deep into Rectifiers: Surpassing Human-Level Performance on ImageNet Classification, in *IEEE International Conference on Computer Vision (ICCV)*, pp. 1026–1034 (2015)
[Hermann 15] Hermann, K. M., Kočiský, T., Grefenstette, E., Espeholt, L., Kay, W., Suleyman, M., and Blunsom, P.: Teaching Machines to Read and Comprehend, in *International Conference on Neural Information Processing Systems (NIPS)*, pp. 1693–1701 (2015)
[Jain 18] Jain, U., Lazebnik, S., and Schwing, A. G.: Two can play this Game: Visual Dialog with Discriminative Question Generation and Answering, in *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 5754–5763 (2018)
[Lin 07] Lin, J.: An exploration of the principles underlying redundancy based factoid question answering, *ACM Transactions on Information Systems (TOIS)*, Vol. 25, No. 2, pp. 1–55 (2007)
[Manning 08] Manning, C. D., Raghavan, P., and Schtze, H.: Introduction to Information Retrieval, pp. 177–194, Cambridge University Press, Cambridge, UK (2008)
[Mostafazadeh 16] Mostafazadeh, N., Misra, I., Devlin, J., Mitchell, M., He, X., and Vanderwende, L.: Generating Natural Questions About an Image, in *Association for Computational Linguistics (ACL)*, pp. 1802–1813 (2016)
[Nishida 18] Nishida, K., Saito, I., Otsuka, A., Asano, H., and Tomita, J.: Retrieve-and-Read: Multi-task Learning of Information Retrieval and Reading Comprehension, in *ACM International Conference on Information and Knowledge Management (CIKM)*, pp. 647–656 (2018)
[Rajpurkar 16] Rajpurkar, P., Zhang, J., Lopyrev, K., and Liang, P.: SQuAD: 100,000+ Questions for Machine Comprehension of Text, in *Empirical Methods in Natural Language Processing (EMNLP)*, pp. 2383–2392 (2016)
[Seo 17] Seo, M., Kembhavi, A., Farhadi, A., and Hajishirzi, H.: Bidirectional Attention Flow for Machine Comprehension, in *International Conference on Learning Representations (ICLR)* (2017)
[Serban 16] Serban, I. V., García-Durán, A., Gülçehre, Ç., Ahn, S., Chandar, S., Courville, A. C., and Bengio, Y.: Generating Factoid Questions With Recurrent Neural Networks: The 30M Factoid Question-Answer Corpus, in *Association for Computational Linguistics (ACL)*, pp. 588–598 (2016)
[Srivastava 14] Srivastava, N., Hinton, G. E., Krizhevsky, A., Sutskever, I., and Salakhutdinov, R.: Dropout: a simple way to prevent neural networks from overfitting, *Journal of Machine Learning Research*, Vol. 15, No. 1, pp. 1929–1958 (2014)
[Srivastava 15] Srivastava, R. K., Greff, K., and Schmidhuber, J.: Highway Networks, *CoRR*, Vol. abs/1505.00387, (2015)
[Trischler 17] Trischler, A., Wang, T., Yuan, X., Harris, J., Sordoni, A., Bachman, P., and Suleman, K.: NewsQA: A Machine Comprehension Dataset, in *Workshop on Representation Learning for NLP (RepL4NLP@ACL)*, pp. 191–200 (2017)
[Yu 18] Yu, A. W., Dohan, D., Luong, M.-T., Zhao, R., Chen, K., Norouzi, M., and Le, Q. V.: QANet: Combining Local Convolution with Global Self-Attention for Reading Comprehension, in *International Conference on Learning Representations (ICLR)* (2018)
[大塚 19] 大塚淳史, 西田京介, 斉藤いつみ, 浅野久子, 富田準二：質問の意図を特定するニューラル質問生成モデル, 日本データベース学会和文論文誌, Vol. 17-J, No. 6, pp. 1–8 (2019)

〔担当委員：鍛治 伸裕〕

2019 年 1 月 24 日 受理

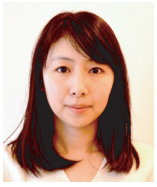
著 者 紹 介

大塚淳史（正会員）

2011 年筑波大学情報学群卒業，2013 年同大大学院図書館情報メディア研究科博士前期課程修了．同年日本電信電話株式会社入社．現在，NTT メディアインテリジェンス研究所研究員．筑波大学大学院図書館情報メディア研究科博士後期課程在籍．情報検索，対話処理，質問応答に関する研究開発に従事．情報処理学会，日本データベース学会各会員．

西田京介

2004 年北海道大学工学部情報高校学科卒．2006 年同大大学院情報科学研究科修士課程了．2008 年同博士課程了．博士（情報科学）．2006-2009 年日本学術振興会特別研究員．2009 年日本電信電話株式会社入社．現在，NTT メディアインテリジェンス研究所特別研究員．2017 年日本データベース学会上林奨励賞受賞．人工知能，自然言語処理，Web・位置情報マイニングに関する研究開発に従事．ACM，言語処理学会，情報処理学会，電子情報通信学会，日本データベース学会各会員．



斉藤いつみ

2010 年早稲田大学理工学部卒業，2012 年東京大学大学院工学系研究科博士前期課程修了．同年日本電信電話株式会社入社．現在，NTT メディアインテリジェンス研究所研究員．自然言語処理に関する研究開発に従事．情報処理学会，言語処理学会各会員

浅野久子

1991 年横浜国立大学工学部卒業．現在，日本電信電話株式会社 NTT メディアインテリジェンス研究所主幹研究員．自然言語処理に関する研究開発に従事．情報処理学会，言語処理学会各会員．

富田準二

1995 年慶應義塾大学理工学部計測工学科卒業．1997 年同大大学院計算機科学専攻修了．同年日本電信電話（株）入社．2005 年米国ワシントン大学客員研究員．博士（工学，2012 年慶應義塾大学）．2006 年から 2017 年まで NTT レゾナント（株）所属．現在，NTT メディアインテリジェンス研究所主幹研究員．自然言語処理，情報検索，対話処理関連の研究開発に従事．情報処理学会会員．日本データベース学会理事．

佐藤哲司

1980 年山梨大学工学部電子工学科卒業．同年日本電信電話公社武蔵野電気通信研究所に入所．データベース，マルチメディアデータベース，情報検索・情報共有などに関する研究・開発に従事．1994 年工学博士 (大阪大学) 取得．2007 年より現職．情報検索・知識発見，社会ネットワーク分析，社会インタラクションに興味を持つ．情報処理学会，日本データベース学会各会員．