問い返し可能な質問応答：読解と質問生成の同時学習モデル

大塚 淳史 [1]　西田 京介 [2]　斉藤 いつみ [3]
西田 光甫 [4]　浅野 久子 [5]　富田 準二 [6]

本論文では，曖昧な質問を想定した，質問意図を明確にする質問応答モデルを提案する．提案モデルでは，質問に対する回答の候補を複数列挙し，回答候補ごとに回答を一意に特定できるように書き直した具体的な質問（改訂質問）を生成する．提案モデルを実装した質問応答システムでは，質問者が質問を入力したとき，回答と合わせて「問い返し」の効果を持つ改訂質問の候補を提示する．質問者は自身の質問意図に近い改訂質問を選択することで，所望の回答を得ることができる．本論文では，機械読解と改訂質問生成を同時に学習する End-to-End のニューラルネットワークモデルを提案する．評価実験では，曖昧な質問と具体的な質問を含む機械読解のコーパスにおいて，曖昧な質問を入力したときの機械読解の回答精度が従来モデルよりも向上し，具体的な質問と同程度の精度を達成できることを示す．

1　はじめに

人工知能がテキストを読み解いて回答する機械読解は，自然言語による質問応答技術の中核をなす技術として注目を集めている．機械読解では入力質問と，回答のための情報源となるテキスト（パッセージ）を入力すると，回答となる情報をパッセージから探索し，回答部分を抽出することで質問応答を可能にする．近年，SQuAD [21] などの大規模な機械読解用データセットが多数公開されており，それらのデータセットを用いた実験において，深層学習モデルによる機械読解は人間と同等以上の回答精度を達成できることが報告されている [4].

機械読解は，AI チャットやコールセンタのオペレータ支援など様々な領域において活用が期待される技術であるが，実用化に向けては課題が残っている．SQuAD などの実験用データセットの質問は，回答が一意に特定できるように人手で作成しているものが主流である．一方で，実際の質問応答システムは様々なユーザが利用するため，必ずしも機械読解で回答可能な質問ばかり入力されるとは限らない．例えば，質問者が曖昧な質問を入力してしまった場合，機械読解ではパッセージ中から回答を特定するための情報が足りずに，質問と関連の低い情報を回答として抽出してしまうという課題がある．曖昧な質問に対して正しい回答ができない例を図 1 に示す．パッセージ中には保険金に関する複数の金額が記載されているが，入力質問の「弁護士費用特約の保険金の限度額はいくら？」では，これらのどの金額について質問しているのか判断することができない．

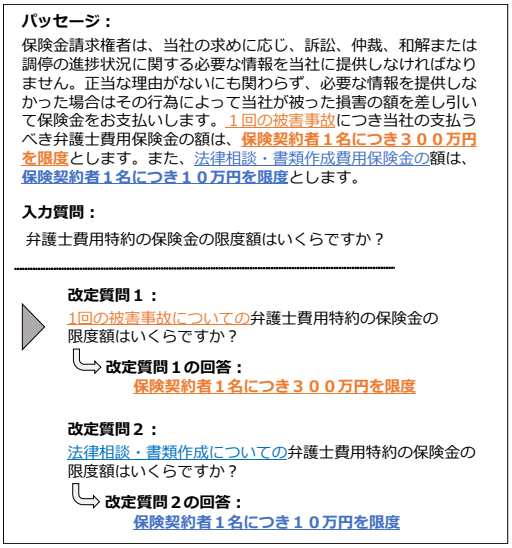

図 1　複数の回答と改訂質問による機械読解

本論文では，曖昧な質問に対する機械読解を実現するために，複数の回答に対応する改訂質問生成を提案する．改訂質問生成とは，パッセージと質問が与えられたとき，質問をパッセージの内容に基づいて具体化する技術である [19]．まず，機械読解によって質問の回答となる可能性がある範囲をパッセージ中から複数抽出する．そして，抽出した回答ごとに改訂質問を作成する．改訂質問を生成する際に回答の情報を利用することで，回答の特定性を高める改訂質問が生成できるようになる．

質問者は，曖昧な質問を入力してしまった場合であっても，質問応答が「問い返し」として提示した改訂質問と回答のペアを閲覧することで，“どのような質問をすれば、どのような回答を得られるか” を判別でき，自身の意図に近い改訂質問を選択することで，所望の回答を得られるようになる．図 1 の例では，パッセージ中の 100 万円と 10 万円という 2 つの回答について，異なる改訂質問を生成し, 提示することで，「1 回の被害事故について」の保険金は 300 万円，「法律相談・書類作成について」は 10 万円であると判別可能になる．

本論文では，複数の回答を出力する機械読解と，回答情報を参考にした改訂質問生成を End-to-End に実行・学習可能なマルチタスクモデルを提案する．提案モデルは，テキストの意味を理解する入力理解層の上に，機械読解層，改訂質問生成層を結合したニューラルネットワークモデルで表される．このとき，機械読解層の回答および状態ベクトルを改訂質問生成層に入力することで，回答情報を考慮した改訂質問が生成できるようになる．

本論文の貢献：本論文では，以下の貢献を果たした．

[1] 正会員　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科
atsushi.otsuka.vs@hco.ntt.co.jp
[2] 正会員　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科
kyosuke.nishida.rx@hco.ntt.co.jp
[3] 非会員　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科
itsumi.saito.df@hco.ntt.co.jp
[4] 非会員　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科
kosuke.nishida.ap@hco.ntt.co.jp
[5] 非会員　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科
hisako.asano.fe@hco.ntt.co.jp
[6] 正会員　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科
junji.tomita.xa@hco.ntt.co.jp

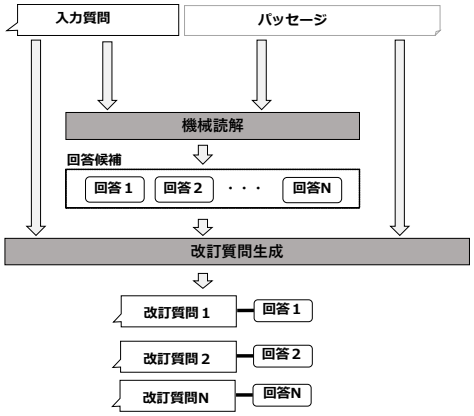

図 2　機械読解と改訂質問のマルチタスクの流れ

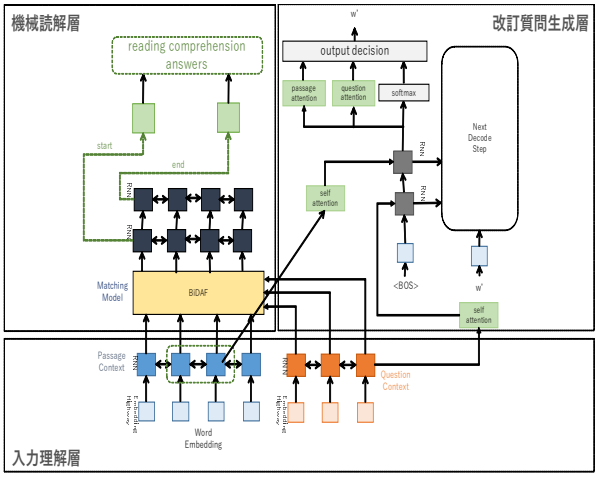

図 3　機械読解と改訂質問生成のマルチタスクモデル

- 改訂質問生成と機械読解をマルチタスクで実行する End-to-End の深層学習モデルを提案した．提案モデルではまず，機械読解層によって，機械読解の回答候補を複数列挙する．そして，改訂質問生成層で入力質問を具体化した改訂質問を生成する．このとき，機械読解の回答および機械読解層の状態を改訂質問生成層に入力することにより，回答を一意に特定できるように具体化した改訂質問を生成している．
- 日本語の機械読解データセットにおいて，提案モデルの有効性を明らかにした．実験で使用したデータセットは，1 つの回答に対して，長文質問と短文質問が付与されており，短文質問を入力したときの機械読解の回答精度は長文質問を入力よりも低下するが，提案モデルから生成した改訂質問を用いた場合，回答精度を改善できることを明らかにした．また，提案モデルが，従来の改訂質問生成と機械読解を疎結合で実行したときよりも，高い回答精度を達成できることを示した．

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] 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 2] 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 3] 機械読解（Reading Complihension :RC）は，入力質問 q と，パッセージ x を入力とし，パッセージ中から回答の始点 a_s と終点 a_e を出力とするタスクである．ここで，回答は $A = \{x_{a_s}, x_{a_{s+1}}, ..., x_{a_e}\}$ である．

[定義 4] 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．ここで，$y_1, y_2, ..., y_K$ は，生成した単語に対応する単語辞書のインデックスである．

[定義 5] 改訂質問生成（Specific Question Generation : SQG）は，入力質問 q とパッセージ x を入力とし，改訂質問（SQ）y を出力するタスクである．本論文では改訂質問生成の際に，機械読解の出力 A も使用する．改訂質問生成を用いた質問応答では，質問者は複数の改訂質問から自身の質問意図に合った SQ を選択する．

3　提案手法

本論文で提案する機械読解と改訂質問生成の流れを図 2 に示す．まず，入力質問とパッセージを機械読解に入力し，パッセージ中から抽出した複数の回答を得る．そして，抽出した回答ごとに改訂質問を生成し，最終的に改訂質問と回答のペアを出力する．

提案手法を End-to-End のニューラルネットワークで実現したものが図 3 となる．提案モデルは，パッセージと入力質問の単語トークン列を，単語埋め込みモデルとリカレントニューラルネットワークによってベクトル系列に変換する入力理解層と，機械読解の回答範囲を抽出する機械読解層，入力質問をパッセージの内容に基づいて具体化する改訂質問生成層の 3 層から構成される．以降はそれぞれの層について詳述する．

3.1　入力理解層

入力理解層の入力として，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 Highway ネットワーク [26] 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 [3] 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．

3.2 機械読解層

機械読解層ではまず，入力理解層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）[23] 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，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．

$$S_{tj} = w_s^\top [H_t; U_j; H_t \odot U_j], \tag{1}$$

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．

$$a_t = \mathrm{softmax}_j(S_t), \tag{2}$$
$$\check{U}_t = \textstyle\sum_j a_{tj} U_j, \tag{3}$$

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．

$$b = \mathrm{softmax}_t(\max_j(S)) \tag{4}$$
$$\check{h} = \sum_t b_t H_t \tag{5}$$

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．

$$G = [H; \check{U}; H \odot \check{U}; H \odot \check{H}] \in \mathbb{R}^{8d \times T}. \tag{6}$$

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．

コンテキスト行列により，機械読解の出力結果である，回答の始点と終点位置を決定する．始点と終点は以下により計算される確率分布 $p_{a_s}, p_{a_e} \in \mathbb{R}^T$ に従う．

$$p_{a_s} = \mathrm{softmax}_t(w_{a_s}^\top [G; M]), \tag{7}$$
$$p_{a_e} = \mathrm{softmax}_t(w_{a_e}^\top [G; M; M']), \tag{8}$$

ここで，$w_{a_s} \in \mathbb{R}^{10d}$，$w_{a_e} \in \mathbb{R}^{12d}$ は学習パラメータである．$M' \in \mathbb{R}^{2d \times T}$ は，$[G; M; M_{a_s}] \in \mathbb{R}^{12d \times T}$ を単層の双方向リカレントニューラルネットワークに入力して得た，始点を考慮したコンテキスト行列である．$M_{a_s} \in \mathbb{R}^{2d \times T}$ は，コンテキスト行列 M の a_s 番目のベクトルを T 列分並べたものである．

機械読解の出力層の始点 p_{a_s} と終点 p_{a_e} の積の降順によって，機械読解層が出力する回答範囲を N 個抽出する．このとき，回答の範囲が重複しないように選択する．仮に抽出した回答範囲が，先に抽出した回答範囲と重複が存在する場合，この候補をスキップし，次候補を採用する．

3.3 改訂質問生成層

改訂質問生成層では，入力理解層と機械読解層の出力に基づいて，改訂質問を生成する．改訂質問は，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．

■RNN 言語生成モデル　生成モデルは，アテンション付きの 2 層 GRU とソフトマックス層により構成される．1 層目の GRU の初期状態として与えるベクトル $h_0^1 \in \mathbb{R}^{2d}$ 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．

$$\alpha_{U_j} = \mathrm{softmax}_j(v_{h_0^1}^\top tanh(w_{h_0^1}[U_J; U_j])), \tag{9}$$
$$h_0^1 = \textstyle\sum_{j<J} \alpha_{U_j} U_j, \tag{10}$$

ここで，$U_J^\top$ 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．$w_{h_0^1} \in \mathbb{R}^{4d \times 2d}$，$v_{h_0^1} \in \mathbb{R}^{2d}$ は学習パラメータである．

2 層目の GRU の初期状態 $h_0^2 \in \mathbb{R}^{2d}$ には，パッセージのコンテキスト行列 H と，回答範囲のベクトル $\bar{A}$ のアテンションによる次式によって計算する．

$$\bar{A} = \mathrm{mean}(H_{a_s:a_e}), \tag{11}$$
$$\alpha_{H_t} = \mathrm{softmax}_t(v_{h_0^2}^\top tanh(w_{h_0^2}[\bar{A}; H_t])), \tag{12}$$
$$h_0^2 = \textstyle\sum_{t<T} \alpha_{H_t} H_t, \tag{13}$$

ここで，$\mathrm{mean}(H_{a_s:a_e})$ は，パッセージのコンテキスト行列 H の a_s 列目から a_e 列目のベクトルの平均ベクトルを表している．$w_{h_0^2} \in \mathbb{R}^{4d \times 2d}$，$v_{h_0^2} \in \mathbb{R}^{2d}$ は学習パラメータである．

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．

$$P_g(y_{s+1}|y_{\leq s}, x, q) = \mathrm{softmax}(W_g h_{s+1}^2 + b_g), \tag{14}$$

ここで，$W_g \in \mathbb{R}^{2d \times V_g}$ と $b_g \in \mathbb{R}^V$ 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．$h_s \in \mathbb{R}^{2d}$ は，2 層目の GRU の s 番目の隠れ状態であり，$h^2_{s+1} \leftarrow \mathrm{GRU}^2(h^2_s, \mathrm{GRU}^1(h^1_s, \hat{z}_s))$ によって更新される．

1 層目の GRU に入力するベクトル $\hat{z}_s$ 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，入力理解層と同様に，one-hot ベクトル化した後，単語埋め込み層と 2 層の Highway ネットワーク層によって連続値のベクトル $e_s \in \mathbb{R}^d$ 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^1_s, h^2_s を用いて次式により作成する．

$$\hat{h}_s = f(W_d[e_s; h^1_s; h^2_s] + b_d). \tag{15}$$

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．

$$\alpha_{st} = \mathrm{softmax}_t(M_t^\top \cdot \hat{h}_s), \qquad \hat{c}_{ps} = \textstyle\sum_t \alpha_{st} M_t, \tag{16}$$

$$\beta_{sj} = \mathrm{softmax}_j(U_j^\top \cdot \hat{h}_s), \qquad \hat{c}_{qs} = \textstyle\sum_j \beta_{sj} U_j. \tag{17}$$

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

$$\hat{z}_s = [e_s; \hat{c}_{ps}; \hat{c}_{qs}], \tag{18}$$

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 は活性化関数を表しており，本論文では PReLu [10] を使用する．

■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．

パッセージまたは入力質問からコピーする確率は，GRU の状態ベクトル h^2_{s+1}，パッセージのコンテキスト行列 H，入力質問のコンテキスト行列 U による pointer generator network [22] に基づき，アテンション重みを用いた次式により計算により計算する．

$$\alpha'_{(s+1)t} = \mathrm{softmax}_j(v_{cq}^\top tanh(w_{cq}[h^2_{s+1}; U_j] + b_{cq})), \tag{19}$$

$$\beta'_{(s+1)t} = \mathrm{softmax}_t(v_{cp}^\top tanh(w_{cp}[h^2_{s+1}; H_t] + b_{cp})), \tag{20}$$

ここで，$v_{cq}, v_{cp}, b_{cq}, b_{cp} \in \mathbb{R}^{2d}$ および，$w_{cq}, w_{cp} \in \mathbb{R}^{4d\times 2d}$ は学習パラメータである．

パッセージからコピーされる確率をそのまま使用した場合，パッセージ中の回答範囲の単語からもコピーされることになる．回答範囲を改訂質問に用いた場合，質問は“〜は・・・ですか？”のように Yes/No を問う様な質問になり，回答との不整合が生じる可能性がある．そこで，パッセージから単語をコピーする際に，回答範囲以外の単語をコピーするようにする．具体的には式 20 で，$t \in A$ のとき，$\mathrm{softmax}_t$ の引数を $-\infty$ とすることによるマスク処理を行う．これにより，回答範囲に該当する t 番目の単語のコピー確率を 0 とする．

最終的に，パッセージと入力質問中の単語がコピーされる確率 P_{cp} および P_{cq} は次式で表される．

$$P_{cp}(y_{s+1}|y_{\le s}, x, q) = \sum_t \mathbb{1}(y_{s+1} = x_t)\alpha'_{(s+1)t}, \tag{21}$$

$$P_{cq}(y_{s+1}|y_{\le s}, x, q) = \sum_j \mathbb{1}(y_{s+1} = q_j)\beta'_{(s+1)j}, \tag{22}$$

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．

■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 $P(y_{s+1}|y_{\le s}, x, q)$ は，各々の確率分布の重み付き和によって計算される．

$$\begin{aligned} P(y_{s+1}|y_{\le s}, x, q) = {} & \lambda_s P_g(y_{s+1}|y_{\le s}, x, q) \\ & + \mu_s P_{cp}(y_{s+1}|y_{\le s}, x, q) \\ & + \nu_s P_{cq}(y_{s+1}|y_{\le s}, x, q), \end{aligned} \tag{23}$$

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．

$$\hat{c}'_{p(s+1)} = \textstyle\sum_t \alpha'_{(s+1)t} H_t, \tag{24}$$

$$\hat{c}'_{q(s+1)} = \textstyle\sum_j \beta'_{(s+1)j} U_j, \tag{25}$$

$$\gamma_{s+1} = \mathrm{softmax}(v_c^\top [\hat{z}_s; h^1_{s+1}; h^2_{s+1}; \hat{c}'_{p(s+1)}; \hat{c}'_{q(s+1)}])$$

$$\lambda_s = \gamma_{s0}, \quad \mu_s = \gamma_{s1}, \quad \nu_s = \gamma_{s2}, \tag{26}$$

ここで，$v_c \in \mathbb{R}^{(v+12d)\times 3}$ は学習パラメータである．

3.4 マルチタスクモデル学習

機械読解と改訂質問生成を同時に実行する提案モデルのパラメータの学習は，それぞれのタスクの損失を結合した損失関数を最小化することで行う．損失関数 L は次式により計算する．

$$L(\theta) = L_{RC} + \lambda L_{QG} \tag{27}$$

ここで，θ は提案モデルの全ての学習パラメータを示しており，λ は損失関数調整用のハイパーパラメータである．

機械読解タスクと，改訂質問生成の損失関数の損失関数は，負の対数尤度を用いた次式により計算する．

$$L_{RC} = -\frac{1}{N}\sum_i (\log P_{a^{(i)}_s} + \log P_{a^{(i)}_e}), \tag{28}$$

$$L_{QG} = -\frac{1}{N}\sum_i \sum_s \log P(y^{(i)}_{s+1}|y^{(i)}_{\le s}, x^{(i)}, q^{(i)}), \tag{29}$$

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．また，$a^{(i)}_s$ と，$a^{(i)}_e$ は，i 目のサンプルの正解となる始点，終点の位置を表している．

4 評価実験

4.1 データセット

本論文では，日本語の機械読解コーパスを用いた実験を行う．作成した機械読解コーパスは，日本語の Wikipedia からランダムで記事を抽出した 600 記事を基に，クラウドソーシングによって，パッセージ抽出と，パッセージごとの質問の作成と質問に対

する回答範囲のアノテーションを行ったデータとなっている．従来の機械読解コーパスとは異なり本データセットでは，1 つの回答ごとに 2 種類の質問を作成している．一つ目の質問は，従来の機械読解コーパスと同様に回答を特定できるような具体的な質問（長文質問）であり，もう一方の質問は，質問意図が伝わる最低限まで，文が短くなるように作成した質問（短文質問）である．長文質問と短文質問は抽出したパッセージから以下の手順により作成した．

1. 作業者 A がある 1 つのパッセージを読み，質問と回答を作成する．このとき，作業者 A は質問を作成する際に文字数制限以内（25 文字）で質問を作成するようにする．
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，回答が同じになるように質問を作成する．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．

作成したデータセットの詳細を表 1 に示す．作成したデータセットは，記事単位で学習セットと評価セットにランダムに分割している．

表 1　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，質問数を表している．L はパッセージまたは質問の単語長である．

	train		test	
	LoQs	ShQs	LoQs	ShQs
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

4.2 実験設定

4.2.1 マルチタスクモデル

提案モデルの入力理解層では，単語埋め込みモデルを使用する．本実験では，日本語 Wikipedia 全記事に GloVe [20] を適用して学習した 300 次元の単語埋め込みモデルを使用する．

マルチタスクモデルの学習は，2 つの GPU（Quadro P6000）を使用した．学習のためのミニバッチ数は 48 に設定し，隠れ状態の次元数 d は 100 とした．改訂質問の生成層で用いる語彙数 V_g は 5000 に設定した．学習の際のドロップアウト [25] の確率は，Highway ネットワーク層のみ 0.2 とし，その他の層を 0.5 に設定した．パラメータの最適化には Adam を使用する．

モデル学習では，学習の 1 エポック目は入力質問に長文質問を入力する，同時に損失関数の係数 $\lambda = 0$ とする．これにより，1 エポック目は機械読解のみの学習を行うことになる．この操作により，マルチタスクモデルが機械読解において，回答に必要な情報を判別できるようにする．2 エポック目以降は，$\lambda = 0.8$ に設定し，機械読解と改訂質問生成のマルチタスク学習を行う．このとき，入力質問は確率 p_q で短文質問が選択され，$(1 - p_q)$ で長文質問が選択される．本論文では $p_q = 0.9$ に設定している．

本論文でより多くのデータからモデル学習を行うために，西田らが日本語のニュース記事から作成した機械読解コーパス（jpNews）[18, 29] の学習データも使用する．jpNews には，短文質問は含まれない．そこで，文圧縮による短文生成 [19] によって，コーパスの質問から機械的に短文質問を作成した．jpNews を含めた学習データの総数は 462,612 である．反復回数は 10 エポック（合計 96,380 学習ステップ）とした．

4.2.2 実験方法

実験は日本語 Wikipedia 機械読解コーパスの test セットを用いて行う．test セットは，1 エントリーにつき，パッセージ，短文質問，長文質問，回答がセットになっている．回答を質問者が知りたい真の正解と捉え，機械読解が出力した回答が，真の正解と一致するかを評価する．評価尺度は全エントリー中，機械読解の出力が真の正解と一致したエントリーの割合である正解率を用いる．

実験ではまず，短文質問と長文質問をそれぞれ直接機械読解に入力したときの正解率を算出する．次に，短文質問を提案モデルであるマルチタスクモデルに入力したときの正解率を得る．このとき，マルチタスクモデルは N 個の回答を出力するため，k 位までの正解率である正解率@k を採用する．正解率@2 の場合，提案モデルの出力の 1 位または 2 位の結果が正解と一致しているかを測る．

短文質問と長文質問を入力する機械読解モデルは，BiDAF による機械読解モデルを使用する．また比較モデルとして大塚らの改訂質問生成モデル [19] に，機械読解モデルを疎結合したモデルを使用する．疎結合の改訂質問生成モデルも複数の改訂質問を生成するため，正解率@k によって評価する．機械読解モデルは，長文・短文質問を入力したものと同じ BiDAF モデルを使用する．

4.3 実験結果

4.3.1 機械読解の正解率評価

実験結果を表 2 に示す．短文質問および長文質問を入力した場合，機械読解は 1 つの回答しか出力できないため，正解率@1 のみのスコアとなっている．疎結合モデルは，改訂質問生成モデル [19] と機械読解モデルを疎結合により接続し，回答を得たものである．マルチタスクモデルは，改訂質問生成と機械読解をマルチタスクで行う提案モデルの結果を示している．

短文質問と長文質問をそのまま機械読解に入力した場合，質問が短くなると正解率が大きく低下している．改訂質問を生成した場合では，正解率@ 1 で疎結合モデル，マルチタスクモデル共に短文質問を機械読解に入力したときよりも高い正解率となっている．特に，提案モデルは疎結合モデルに比べて，正解率が大幅に改善されている．マルチタスクモデルの結果で，改訂質問と回答を 5 個出力した結果では，正解率@5 が 0.652 と長文質問の正解率 0.668 の約 97% とほぼ同等の正解率となっていることがわかる．

機械読解の正解率の実験結果より，入力する質問が短い場合，機械読解はその性能を十分に発揮できないことが明らかとなったといえる．しかし改訂質問生成により改訂質問を生成することで，短文質問が入力された場合でも，機械読解の回答精度を改善できると考える．本論文では，短い質問は曖昧であり回答を一意に特定できないという仮定から複数の改訂質問候補を生成し，質

表 2　機械読解の正解率@k

k	短文質問	長文質問	疎結合モデル	マルチタスクモデル
@1	0.509	0.668	0.516	0.583
@2	-	-	0.583	0.628
@3	-	-	0.605	0.638
@4	-	-	0.623	0.647
@5	-	-	0.640	0.652

表 3　改訂質問の長文質問との類似性評価

	適合率	再現率	F 値
短文質問	1.00	0.450	0.621
疎結合モデル	0.738	0.684	0.710
マルチタスクモデル	0.767	0.698	0.731

問者が選択する手法を提案しているが，正解率@1 の結果から，仮に 1 つの改訂質問を質問者に選択させずに使用したとしても，改訂質問は有効に機能するといえる．改訂質問を 5 個生成したときの正解率は長文質問とほぼ同等になった．長文質問は，回答を特定するための情報を過不足なく含むように作成された質問であるため，長文質問を入力したときの機械読解の回答精度は，機械読解モデル自体の上限に近い性能を発揮している．以上のことより，改訂質問を 5 個程度生成すれば，機械読解モデルの性能を最大限発揮した回答が得られるようになるといえる．

疎結合モデルの機械読解の異なり回答数を調査したところ，機械読解での異なり回答数の平均は 2.01 であった．改訂質問は 5 個生成しているため，生成した改訂質問の半数以上が同じ回答となってしまうものだということがわかる．この結果が，疎結合モデルで正解率@k がマルチタスクモデルよりも低くなった原因であると考える．このことより，複数の回答を提示する場合，疎結合モデルよりもマルチタスクモデルのほうが有効であるといえる．

4.3.2　改訂質問に関する評価

機械読解実験の結果は，改訂質問を入力したときの機械読解の回答結果のみを調査しているため，生成した改訂質問自体の品質は評価していない．そこで，生成された改訂質問の評価のために，改訂質問がどの程度長文質問に近い質問を生成できたかを調査した．長文質問を正解とし，改訂質問との単語の重複率を適合率，再現率，F 値の評価値を設定し評価を行った．

評価結果を表 3 に示す．短文質問は，短文質問の単語の重複を長文質問と直接比較したものである．長文質問は短文質問を基に作成されているため，適合率は 1.0 となるが，文が短くなるため，再現率が大幅に低下している．

改訂質問は疎結合モデル，マルチタスクモデルともに 1 位として生成された改訂質問である．どちらも F 値では入力の短文質問よりも改善されており，より長文質問に近い改訂質問が生成されていることがわかる．特に，マルチタスクモデルのほうが，疎結合モデルよりも高い F 値となっており，マルチタスクで機械読解の結果を用いたほうがより長文質問に近い質問が生成できることが明らかとなった．

機械読解の回答に基づく改訂質問生成の例を図 4 に示す．ニュースリリースのパッセージに対して，「何を目指していますか？」という短文質問を入力しときの例である．この短文質問を機械読解モデルに入力すると「対話システムの開発」という回答が得られる．一方で，マルチタスクモデルに入力した場合「対話システムの開発」と「事業導入」という複数パターンの回答が得られている．このときの改訂質問は，どちらとも短文質問よりも長くなっている．パッセージ上には，マルチタスクモデルで得られた回答部分および，改訂質問を生成する際に pointer generator のアテンションが強くかかった領域をアノテーションしている．太字が回答領域であり，下線部がアテンションがかかった領域である．また，改訂質問 1（赤）と 2（青）は色で識別されている．

マルチタスクモデルでは，異なる 2 箇所の回答に対して，それぞれ全く異なるアテンションによって改訂質問が生成されていることがわかる．特に回答が「事業導入」になる改訂質問 2 は回答部の周辺領域のテキストがそのままコピーされて使用されており，回答が特定できるような改訂質問が生成されていることがわかる．

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を向上させるために，質問文を書き換える研究では，Buck らの研究がある [2]．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，質問者の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は，End-to-End のニューラルネットワークによるもので，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．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 [5]．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 [9]．Huang らは，フレーズの変換辞書を組み込んだ，アテンション付き Encoder-Decoder によって，動的にパラフレーズを作成するモデルを提案している [11]．

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 [7]．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 [6]．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

nttでは『日常会話ができる対話システム』の研究を進めてきました。世の中の多くの対話システムが特定の仕事をこなすタスク指向型の対話システムであるのに対し、日常の幅広い話題に対応し、自分からユーザに働きかけ長時間の会話を続けられる**対話システムの開発**を目指しています。またnttでは、システムにキャラクター性を持たせることで親しみが増し、長く使い続けられると考えており、キャラクター性を持った対話システムの検討を進めています。現在、tottoへ導入した機能の一部をエンジン化した、対話エンジンを開発しており、**事業導入**を目指しています。今後はバーチャルユーチューバーや既存のキャラクターとの対話サービスへの応用などを検討しています。

質問	回答
短文質問： 何を目指していますか？	**対話システムの開発**
改訂質問1： nttでは『日常会話ができる対話システム』の研究を進めて、何を目指していますか?	**対話システムの開発**
改訂質問2： tottoへ導入した機能の一部をエンジン化した、対話エンジンを開発しており、 何を目指していますか	**事業導入**

図 4　改訂質問と機械学習のマルチタスクの例

性能を向上させる手法を提案している．Wang らは，マルチエージェントの仕組みと深層学習の質問生成モデルを組み合わせて，教師データなしでの回答抽出と，その回答に関する質問生成を実行する手法を提案している [27]．Kim らは，質問生成の際に回答となる単語が質問に含まれてしまう問題に対応するため，回答部分と特殊なトークンに置き換えることで，回答を含まず，かつ回答周辺の情報から適切な疑問詞を付与した質問を生成する手法を提案している [15]．

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 [13, 17] や，知識ベースから質問を作成するタスク [8, 24] 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　クエリ推薦に関する研究

クエリ推薦とは情報検索や質問応答において，ユーザの入力したクエリ（キーワード組）に対して，キーワードの追加や修正を行いユーザに推薦する技術である．Jaech らは，RNN 言語モデルに基づいて，クエリに追加するキーワードを生成する際に，ユーザに関する情報を埋め込んだベクトルを追加することで，個人の趣向に合わせたクエリが推薦できる手法を提案している [12]．Jiang らの研究では，一連の情報検索行動を追跡し，閲覧した Web サイトやそこで使用されている単語情報をクエリ生成のニューラルネットワークに組み込むことによる，コンテキスト依存なクエリ推薦を実現している [14]．クエリ推薦は主に情報検索をターゲットにし，キーワードの追加修正を行う技術である．本論文の改訂質問生成は，質問応答をターゲットに自然文の質問文を書き換えるのでこれらの研究とは異なる．

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 [1, 16, 28]，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．

6　まとめ

本論文では，機械読解による質問応答において，曖昧な質問が入力された場合に，質問をパッセージの内容に基づいて具体化する改訂質問生成と，パッセージから質問の回答部を抽出する機械読解をマルチタスクで学習・実行する深層学習モデルを提案した．提案モデルでは，質問に対する回答の候補を，機械読解によってパッセージから複数抽出する．そして，抽出した回答候補ごとに改訂質問を生成する．このとき，改訂質問生成に回答情報を利用することにより，回答が特定できるように具体化した改訂質問を生成する．

日本語の機械読解コーパスを用いた実験では，短い質問を入力したとき，提案モデルを利用することで，長い質問を入力したときと同程度の回答精度を得られることを明らかにした．また，回答ごとに異なる改訂質問が生成されることを示した．

今後は，抽出型ではなく生成型の機械読解を適用したモデルの検討や，改訂質問を含めた対話型の質問応答技術について検討を進めていく予定である．

参考文献

[1] Eric Brill, Susan T. Dumais, and Michele Banko. An analysis of the askmsr question-answering system. In *Empirical Methods in Natural Language Processing (EMNLP)*, 2002.
[2] Christian Buck, Jannis Bulian, Massimiliano Ciaramita, Wojciech Gajewski, Andrea Gesmundo, Neil Houlsby, and Wei Wang. Ask the right questions: Active question reformulation with reinforcement learning. In *International Conference on Learning Representations (ICLR)*, 2018.
[3] Junyoung Chung, Çaglar Gülçehre, KyungHyun Cho, and Yoshua Bengio. Empirical evaluation of gated recurrent neural networks on sequence modeling. *CoRR*, abs/1412.3555, 2014.
[4] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. Bert: Pre-training of deep bidirectional transformers for language understanding. *CoRR*, abs/1810.04805, 2018.
[5] Li Dong, Jonathan Mallinson, Siva Reddy, and Mirella Lapata. Learning to paraphrase for question answering. In *Empirical Methods in Natural Language Processing (EMNLP)*, pages 875–886, 2017.
[6] Xinya Du and Claire Cardie. Harvesting paragraph-level question-answer pairs from wikipedia. In *Association for Computational Linguistics (ACL)*, pages 1907–1917, 2018.

[7] Xinya Du, Junru Shao, and Claire Cardie. Learning to ask: Neural question generation for reading comprehension. In *Association for Computational Linguistics (ACL)*, pages 1342–1352, 2017.
[8] Hady ElSahar, Christophe Gravier, and Frédérique Laforest. Zero-shot question generation from knowledge graphs for unseen predicates and entity types. In *North American Association for Computational Linguistics (NAACL)*, 2018. to appear.
[9] Ankush Gupta, Arvind Agarwal, Prawaan Singh, and Piyush Rai. A deep generative framework for paraphrase generation. In *Association for the Advancement of Artificial Intelligence (AAAI)*, 2018.
[10] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Delving deep into rectifiers: Surpassing human-level performance on imagenet classification. In *IEEE International Conference on Computer Vision (ICCV)*, pages 1026–1034, 2015.
[11] Shaohan Huang, Yu Wu, Furu Wei, and Zhongzhi Luan. Dictionary-guided editing networks for paraphrase generation. In *Association for the Advancement of Artificial Intelligence (AAAI)*, 2019.
[12] Aaron Jaech and Mari Ostendorf. Personalized language model for query auto-completion. In *Annual Meeting of the Association for Computational Linguistics (ACL)*, pages 700–705, 2018.
[13] Unnat Jain, Svetlana Lazebnik, and Alexander G. Schwing. Two can play this game: Visual dialog with discriminative question generation and answering. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2018.
[14] Jyun-Yu Jiang and Wei Wang. Rin: Reformulation inference network for context-aware query suggestion. In *Conference on Information and Knowledge Management (CIKM)*, pages 198–206, 2018.
[15] Yanghoon Kim, Hwanhee Lee, Joongbo Shin, and Kyomin Jung. Improving neural question generation using answer separation. In *Association for the Advancement of Artificial Intelligence (AAAI)*, 2019.
[16] Jimmy Lin. An exploration of the principles underlying redundancy based factoid question answering. *ACM Transactions on Information Systems (TOIS)*, 25(2):6, 2007.
[17] Nasrin Mostafazadeh, Ishan Misra, Jacob Devlin, Margaret Mitchell, Xiaodong He, and Lucy Vanderwende. Generating natural questions about an image. In *Association for Computational Linguistics (ACL)*, 2016.
[18] Kyosuke Nishida, Itsumi Saito, Atsushi Otsuka, Hisako Asano, and Junji Tomita. Retrieve-and-read: Multi-task learning of information retrieval and reading comprehension. In *Conference on Information and Knowledge Management (CIKM)*, pages 647–656, 2018.
[19] Atsushi Otsuka, Kyosuke Nishida, Itsumi Saito, Hisako Asano, and Junji Tomita. Specific question generation for reading comprehension. *In AAAI 2019 Reasoning for Complex Question Answering Workshop*, 2019.
[20] Jeffrey Pennington, Richard Socher, and Christopher D. Manning. Glove: Global vectors for word representation. In *Empirical Methods in Natural Language Processing (EMNLP)*, pages 1532–1543, 2014.
[21] Pranav Rajpurkar, Jian Zhang, Konstantin Lopyrev, and Percy Liang. Squad: 100,000+ questions for machine comprehension of text. In *Empirical Methods in Natural Language Processing (EMNLP)*, pages 2383–2392, 2016.
[22] Abigail See, Peter J. Liu, and Christopher D. Manning. Get to the point: Summarization with pointer-generator networks. In *Annual Meeting of the Association for Computational Linguistics (ACL)*, pages 1073–1083, 2017.
[23] Minjoon Seo, Aniruddha Kembhavi, Ali Farhadi, and Hannaneh Hajishirzi. Bidirectional attention flow for machine comprehension. In *International Conference on Learning Representations (ICLR)*, 2017.
[24] Iulian Vlad Serban, Alberto García-Durán, Çaglar Gülçehre, Sungjin Ahn, Sarath Chandar, Aaron C. Courville, and Yoshua Bengio. Generating factoid questions with recurrent neural networks: The 30m factoid question-answer corpus. In *Association for Computational Linguistics (ACL)*, 2016.
[25] Nitish Srivastava, Geoffrey E. Hinton, Alex Krizhevsky, Ilya Sutskever, and Ruslan Salakhutdinov. Dropout: a simple way to prevent neural networks from overfitting. *Journal of Machine Learning Research*, 15(1):1929–1958, 2014.
[26] Rupesh Kumar Srivastava, Klaus Greff, and Jürgen Schmidhuber. Highway networks. *CoRR*, abs/1505.00387, 2015.
[27] Siyuan Wang, Zhongyu Wei, Zhihao Fan, Yang Liu, and Xuanjing Huang. A multi-agent communication framework for question-worthy phrase extraction and question generation. In *Association for the Advancement of Artificial Intelligence (AAAI)*, 2019.
[28] Tom Chao Zhou, Chin-Yew Lin, Irwin King, Michael R. Lyu, Young-In Song, and Yunbo Cao. Learning to suggest questions in online forums. In *Association for the Advancement of Artificial Intelligence (AAAI)*, pages 1298–1303, 2011.
[29] 西田京介, 斉藤いつみ, 大塚淳史, 浅野久子, and 富田準二. 情報検索とのマルチタスク学習による大規模機械読解. 言語処理学会第 *24* 回年次大会論文集 *(NLP2018)*, 2018.