DEIM Forum 2019 G7-4

前提知識を考慮した根拠の妥当性判断による自動回答手法

田上　諒† 木村　輔† 宮森　恒†

† 京都産業大学大学院先端情報学研究科　〒 603–8555 京都府京都市北区上賀茂本山
E-mail: †{i1788124,i1650047,miya}@cc.kyoto-su.ac.jp

あらまし　これまでの研究で提案されている，ファクトイド型質問応答の回答手法は，多くの場合，質問文に基づいて情報源から回答候補や回答候補を絞り込むための根拠を探索し，回答を導出するという手順がとられる．しかし，回答候補を絞り込むための根拠が，情報源上に直接的に役立つ形で記述されているとは限らず，複数の前提知識を結び付けなければ回答できない質問への対処が，課題となっている．本稿では，質問文と回答候補から生成される事実文と，情報源上の根拠文の対の妥当性を，複数の前提知識をもとに判断し，その結果をもとに回答を導出する手法を提案する．実験では，英語の自然科学系の四択問題を対象として，提案手法が正答率にどのような影響を及ぼすかについて明らかにする．

キーワード　自然言語処理, ファクトイド型質問応答, 自動回答, 妥当性

1 はじめに

近年，ユーザからの多様な情報要求を満たす技術として，検索エンジンなどの従来技術に代わり，質問応答などの自動回答技術が注目されている．大量に存在する情報源の中から，ユーザが必要な情報を得る一般的な手段として，関連するキーワードをクエリとして文書集合から検索し，検索結果となる複数文書から必要な情報を探し出す方法がある．しかし，この方法は，クエリ生成の過程や，複数文書の中から要求を満たす情報を選択する過程を，ユーザ自身が負担する必要がある．それに対し質問応答は，ユーザ自身の情報要求を自然言語で入力し，情報源などを参照して，1 つの正答を出力することを目指す技術である．ユーザの得たい情報を身近な言語で入力できる点，および，複数の情報を比較する必要がない点が特徴といえる．

質問応答が対象とする質問の一つに，多肢選択問題がある．多肢選択問題に対して，文書検索に基づく手法を用いて自動回答する場合，一般的には，回答候補ごとに，質問に対する回答として尤もらしいかを，情報源上の何らかの文書を根拠にして判断する．たとえば,「ジョージ・ワシントンはどこで生まれましたか?」という質問に対して,「ウェストモアランド郡」という回答候補があげられた場合,「ジョージ・ワシントンはウェストモアランド郡で生まれました」という仮の**事実**を考えることができる．情報源中に「ジョージ・ワシントンは 1732 年 2 月 22 日にウェストモアランド郡において誕生した」という文書が存在する場合，仮の**事実**は真であると言え，また，情報源中の当該文書は，事実に対する**根拠**と捉えることができる．この例では，単一の根拠で，仮の事実を真と認定することができている．

しかし，質問に回答するための根拠が，情報源上に直接的に役立つ形で記述されているとは限らない．たとえば,「スプレー缶に近づけてはいけないものは? 」という質問と，3 つの選択肢「（A）水 （B）火 （C）磁石」が与えられたとき，関連する情報源上の文書として「一般的なスプレー缶には LPG ガスが使われる」という文書が取得できたとする.「LPG ガスは可燃性で燃えやすい」という前提知識を持つ人間であれば，当該根拠と組み合わせることで,「（B）火」が回答であることを導けるかもしれないが，前提知識を持たない場合，正答することは困難になるであろう．このように，質問応答システムに入力される質問の中には，複数の前提知識を結び付けなければ回答できないような質問も想定でき，このような質問に対してどのように対処するかが，課題となっている．

本稿では，前提知識に基づいて，**事実**と**根拠**の対の妥当性を判断し，その結果をもとに回答を導出する手法を提案する．具体的には，質問文と回答候補から生成される事実文と，情報源上の根拠文の対が入力されると，各文を解釈するのに必要と思われる前提知識を参照したうえで，当該対の妥当性を判断するようなモデルを構築する．自動回答時には，本モデルを用い，一問の質問から想定される複数の事実と根拠の対の妥当性を判断したうえで，最終的な回答を導出する．ここでの妥当性とは，ある**根拠**が，ある**事実**に対しての根拠としてふさわしいかどうかを示す指標である．**事実**とは，実際に発生した事象あるいは現実に存在する事柄であり，**根拠**とは，その事実の発生または存在を示すのに必要な科学的知識に基づく理由とする．つまり，根拠から事実を導き出せる場合を「妥当である」とし，導き出せない場合を「妥当ではない」と表現する．

実験では，提案したモデルについて，モデルの構築条件の違いによる精度を比較する．また，使用するモデルの違いによる，自動回答の精度の変化についても調査する．

2 関連研究

近年では，質問応答に関連する様々な課題に応じたデータセットが公開されている．たとえば，SQuAD [1] は，前提となる文章（提示文脈）と，提示文脈についての最大 5 問の質問，およびその回答が含まれるデータセットである．提示文脈となる文章は，Wikipedia の記事内のある一段落であり，回答は，提示文脈中の該当区間をそのまま抽出する形式となっている．基本的には，提示文脈を読めば回答が可能であるため，機械読解と呼

ばれる分野でも，性能評価によく用いられている．模範回答とまったく同じ回答が出力されたかどうかの指標（Exact Match; EM）でみると，2018 年 12 月時点では，8 割以上の精度を誇る手法が複数提案されている．

読解力だけでなく，回答時に推論なども必要となるようなデータセットとして，アレン人工知能研究所（Allen Institute for Artificial Intelligence; AI2）[1]が公開している，AI2 Reasoning Challenge（ARC）[2] や Open Book QA [3] があげられる．ARC は，米国において，小学生レベルの試験として実際に使用されている，自然科学系の質問を含んだデータセットであり，すべて多肢選択型の質問となっている．質問は，Easy セットと Challenge セットに分類され，併せて，情報源となるコーパスも公開されている．特に Challenge セットに該当する質問は，従来の検索に基づく手法や，単語共起に基づく手法では，回答が誤りやすいものとなっている．Open Book QA は，質問データとは別に，あらかじめ，約 1,300 件の科学的事実が記述された open book と呼ばれる文の集合が用意されたデータセットである．実際の質問は，open book 内のいずれかの科学的事実が根拠となるように構成されている [2]．ARC と同様に，多肢選択型であり，open book の記述内容および質問は，小学生レベルのものである．大人の人間であれば，open book を参照すれば容易に回答できる質問ではあるが，システムが質問に正解するためには，「金属は電気を通しやすい」といったような open book 上の科学的事実だけではなく，「鎧は金属でできている」といった幅広い前提知識が必要とされるデータセットである．本章では，これらのデータセットが目的とするような，与えられた情報源のみでは回答が困難な質問に対する，自動回答手法を提案する．

Sun ら [4] は，汎用的な言語モデルである OpenAI fine-tuned transformer（OFT）[5] に，自身らで提案した機械読解戦略を組み合わせた手法を提案しており，ARC および Open Book QA に対して，高成績を残している．OFT は，注意機構（Attention）のみを使用した機械翻訳モデルである Transformer [6] を元にしており，Sun らは，機械読解のモデルとして用いている．OFT へ入力する中間表現を生成する過程で，(1) 質問・選択肢・関連文書を読む順番を変えた中間表現を生成する，(2) 関連文書中に含まれる，質問と選択肢に関連するトークンの中間表現を強調させる，といった戦略を提案している．加えて，(3) 機械読解のベンチマークデータセットである RACE [7] から，自動で質問と回答候補を生成し，それらの質問に正解できるように学習（自己評価）した後で，ARC などのデータを学習させるという戦略もとっている．この戦略は，人間と同じように，ごく一般的な知識を学習させてから，より専門的な学習をさせるというプロセスとなる．本稿の提案手法も，事実と根拠の妥当性を判断する際には，前提知識を考慮したうえで判断させる流れとなっている．なお，Sun らはさらに，OFT を元にした BERT [8] と，以上の戦略を組み合わせた手法を提案しており，ARC および Open Book QA に対しては，2018 年 12 月時点で最も良い成績を残している．BERT は，様々な言語理解タスクで，高い成績を収めている，汎用言語モデルである．

1：アレン人工知能研究所: https://allenai.org

2：教科書などが持込可能な試験のことを，open book exam と呼ぶ．

3 提 案 手 法

本節では，事実と根拠の対の妥当性を判断するモデルの構築手法，および，当該モデルを用いた自動回答手法について述べる．図 1 は，それぞれの手法の概要を示したものである．学習時には，あらかじめ，妥当な事実と根拠の対，および，妥当ではない事実と根拠の対を大量に用意し，妥当性判断モデルを深層学習によって学習させる．自動回答で使用する際には，「質問文と回答候補から生成される事実」と，「情報源上の関連する根拠」間の妥当性をモデルに推論させ，その平均確率を元に最終的な回答を決定する．なお，本提案手法は，複数の回答候補から最も尤もらしい回答を選択することに焦点を当て，多肢選択問題のように，あらかじめ回答候補は用意されていることを想定する．

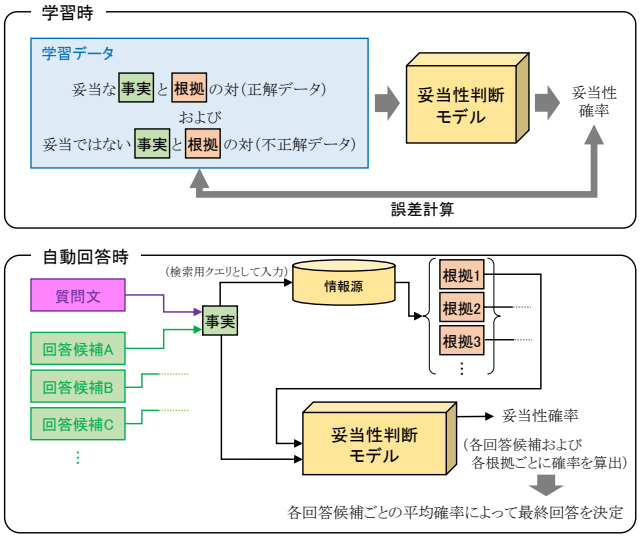

図 1　提案手法の概要

3.1 妥当性判断モデル

ここでは，1 章で定義した，「ある根拠が，ある事実を導き出せるかどうかの妥当性」を判断する**妥当性判断モデル**について述べる．たとえば，「火にスプレー缶を近づけると爆発のおそれがある」という事実と，「一般的なスプレー缶には LPG ガスが使われる」という根拠の対は，妥当である．反対に，同じ根拠に対して，「水にスプレー缶を近づけると爆発のおそれがある」という事実は，文法的破綻はないが，妥当とは言えない．このような判断を行わせるためのモデルを，本章では妥当性判断モデルと呼び，深層学習モデルとして構築する．

モデルの処理手順および構成を，図 2 に示す．事実文および根拠文の 2 つを入力すると，その対の妥当性を，確率値として出力する構成である．

3.1.1 入 力 変 数

入力変数として，事実文を表す F と，根拠文を表す R を定義する．F の s 番目の単語を f_s，R の t 番目の単語を r_t とすると，F と R はそれぞれ，$F = \{f_1, \cdots, f_s, \cdots, f_{|F|}\}$，$R = \{r_1, \cdots, r_t, \cdots, r_{|R|}\}$

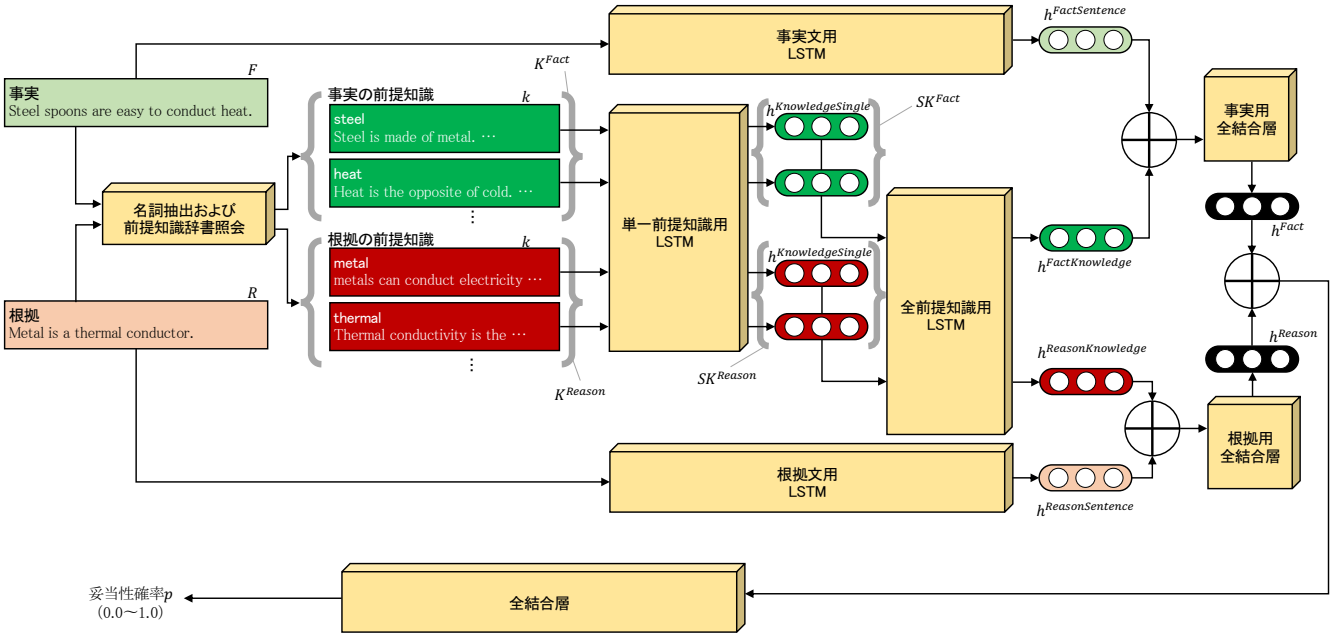

図 2　妥当性判断モデルの処理手順および構成

というシーケンスで表される．ただし，$|A|$ は，あるシーケンス A の要素数とする．

3.1.2　事実文と根拠文のエンコード

本モデルでは，妥当性を判断するための要素の一つとして，F および R を，Bidirectional LSTM（BiLSTM）[9] によって中間ベクトルへエンコードしたものを使用する．BiLSTM は，シーケンスデータをエンコードすることのできる Long-Short Term Memory（LSTM）[10] を，双方向に拡張したモデルである．

F のためのエンコーダを LSTM^{Fact}，R のためのエンコーダを LSTM^{Reason} とすると，最終的にエンコードされた $h^{FactSentence}$，$h^{ReasonSentence}$ は，式 1 および式 2 のとおり表される．関数 $e()$ は，入力された単語を，w 次元の単語ベクトル（単語の分散表現）へ変換する Word Embed である．

$$\begin{aligned}
\overrightarrow{v}_s^{Fact} &= \mathrm{LSTM}^{Fact}_{forward}\left(\overrightarrow{v}_{s-1}^{Fact}, e(f_s)\right) \\
\overleftarrow{v}_s^{Fact} &= \mathrm{LSTM}^{Fact}_{backward}\left(\overleftarrow{v}_{s+1}^{Fact}, e(f_s)\right) \\
h^{FactSentence} &= \left[\overrightarrow{v}_{|F|}^{Fact}; \overleftarrow{v}_1^{Fact}\right]
\end{aligned} \tag{1}$$

$$\begin{aligned}
\overrightarrow{v}_t^{Reason} &= \mathrm{LSTM}^{Reason}_{forward}\left(\overrightarrow{v}_{t-1}^{Reason}, e(r_t)\right) \\
\overleftarrow{v}_t^{Reason} &= \mathrm{LSTM}^{Reason}_{backward}\left(\overleftarrow{v}_{t+1}^{Reason}, e(r_t)\right) \\
h^{ReasonSentence} &= \left[\overrightarrow{v}_{|R|}^{Reason}; \overleftarrow{v}_1^{Reason}\right]
\end{aligned} \tag{2}$$

ここで，$\overrightarrow{v}$ はシーケンスを順方向に入力，$\overleftarrow{v}$ はシーケンスを逆方向に入力していることを示す．また，$[\cdot;\cdot]$ は，2 つのベクトルの結合を表す．

3.1.3　前提知識のエンコード

事実文と根拠文の妥当性をより正確に判断するには，様々な前提知識を持ち合わせていなければならない可能性が考えられる．たとえば，先の例の場合，「LPG ガスは可燃性で燃えやすい」という前提知識がなければ，妥当かどうかを正確に判断することはできない．よって，本モデルでは，あらかじめ辞書を用意し，それぞれの文中に存在する単語を説明するような文をエンコードして，判断に用いる．辞書はあらかじめ，Simple English Wikipedia[3] から生成する．記事名を単語と見なし，説明文として，各記事の一段落目を採用する．ここでは，それぞれの説明文のことを**前提知識**と呼ぶ．

ここで，F から得られた前提知識の集合を K^{Fact}，R から得られた前提知識の集合を K^{Reason} と定義する．K^{Fact} および K^{Reason} を得るため，F および R のそれぞれに対して形態素解析を行い，名詞のみを抽出する．そして，各名詞を辞書で照会し，当該名詞の前提知識 k を得る．ある文から u 番目に得られた 1 つの名詞の前提知識を k_u とすると，K^{Fact} および K^{Reason} は，全 $|K|$ 件の名詞から得られた前提知識からなる $K = \{k_1, \cdots, k_u, \cdots, k_{|K|}\}$ というシーケンスとなる．また，k_u の文の w 番目の単語を o_w とすると，k_u は $k_u = \{o_1, \cdots, o_w, \cdots, o_{|k_u|}\}$ というシーケンスで表される．ただし，文から 1 件も名詞が得られなかった場合，および，抽出したすべての名詞が辞書に存在しなかった場合，$K = \{\{unk\}\}$ とする．unk は，未知語を示す特殊な記号である．

本モデルでは，まず，単一の前提知識である k ごとに，中間ベクトル $h^{KnowledgeSingle}$ へエンコードする．k のためのエンコーダを $\mathrm{LSTM}^{KnowledgeSingle}$ とすると，k_u から最終的にエンコードされた $h_u^{KnowledgeSingle}$ は，式 3 のとおり表される．なお，3.1.2 項と同様に，BiLSTM を用いる．

$$\begin{aligned}
\overrightarrow{v}_w^{KnowledgeSingle} &= \mathrm{LSTM}^{KnowledgeSingle}_{forward}\left(\overrightarrow{v}_{w-1}^{KnowledgeSingle}, e(o_w)\right) \\
\overleftarrow{v}_w^{KnowledgeSingle} &= \mathrm{LSTM}^{KnowledgeSingle}_{backward}\left(\overleftarrow{v}_{w+1}^{KnowledgeSingle}, e(o_w)\right) \\
h_u^{KnowledgeSingle} &= \left[\overrightarrow{v}_{|k_u|}^{KnowledgeSingle}; \overleftarrow{v}_1^{KnowledgeSingle}\right]
\end{aligned} \tag{3}$$

この時点で，各前提知識ごとの中間ベクトル $h^{KnowledgeSingle}$ が生成されているため，F および R のそれぞれの

3：Simple English Wikipedia: https://simple.wikipedia.org/

前提知識の中間ベクトルをまとめた，SK^{Fact}，SK^{Reason} というシーケンスが定義できる．SK は，$SK = \left\{h_1^{KnowledgeSingle}, \cdots, h_u^{KnowledgeSingle}, \cdots, h_{|K|}^{KnowledgeSingle}\right\}$ と表される．続いて，SK^{Fact} および SK^{Reason} をエンコードし，中間ベクトル $h^{FactKnowledge}$ および $h^{ReasonKnowledge}$ を生成する．SK のためのエンコーダを $LSTM^{KnowledgeAll}$ とすると，$h^{FactKnowledge}$ または $h^{ReasonKnowledge}$ は，式 4 のとおり表される．なお，3.1.2 項と同様に，BiLSTM を用いる．

$$\begin{aligned}\overrightarrow{v}_u^{KnowledgeAll} &= \mathrm{LSTM}_{forward}^{KnowledgeAll}\left(\overrightarrow{v}_{u-1}^{KnowledgeAll}, h_u^{KnowledgeSingle}\right)\\ \overleftarrow{v}_u^{KnowledgeAll} &= \mathrm{LSTM}_{backward}^{KnowledgeAll}\left(\overleftarrow{v}_{u+1}^{KnowledgeAll}, h_u^{KnowledgeSingle}\right)\\ h^{FactKnowledge|ReasonKnowledge} &= \left[\overrightarrow{v}_{|k_u|}^{KnowledgeAll}; \overleftarrow{v}_1^{KnowledgeAll}\right]\end{aligned} \tag{4}$$

以上により，事実文のための前提知識を表す $h^{FactKnowledge}$ と，根拠文のための前提知識を表す $h^{ReasonKnowledge}$ が生成される．

3.1.4 デコード

まず，事実に関する中間ベクトル h^{Fact}，および，根拠に関する中間ベクトル h^{Reason} を，式 5 および式 6 のとおり生成する．ただし，W は，全結合層の重みを表し，簡単のため，バイアス項 b は W に含まれるものとする．また，活性化関数 f には ELU 関数 [11] を用いる．

$$h^{Fact} = \mathrm{f}\left(W_{Fact2} \times \mathrm{f}\left(W_{Fact1} \times \left[h^{FactSentence}; h^{FactKnowledge}\right]\right)\right) \tag{5}$$

$$h^{Reason} = \mathrm{f}\left(W_{Reason2} \times \mathrm{f}\left(W_{Reason1} \times \left[h^{ReasonSentence}; h^{ReasonKnowledge}\right]\right)\right) \tag{6}$$

そして，妥当性確率 p は，式 7 に従って算出される．確率値は，0.0〜1.0 の実数値となる．

$$p = \mathrm{sigmoid}\left(W_{Final2} \times \mathrm{f}\left(W_{Final1} \times \left[h^{Fact}; h^{Reason}\right]\right)\right) \tag{7}$$

3.1.5 損失関数

本モデルの最適化対象の損失関数は，式のとおりである．ただし，データセットは全 N 件であり，θ は，モデル中の全パラメータを表す．y は正解ラベルであり，「妥当である」場合は $y = 1.0$，「妥当ではない」場合は $y = 0.0$ となる．

$$E(\theta) = \sum_{n=1}^{N} \|y - M\left(F_n, R_n, K_n^{Fact}, K_n^{Reason}; \theta\right)\|^2 \tag{8}$$

3.2 自動回答

自動回答時には，先の妥当性判断モデルが出力する確率値を指標として，複数の回答候補の中から，最終的な回答を出力する．

まず，質問文と回答候補から，事実文 F をルールベースによって生成する．たとえば，「Which should not be brought near the spray can?」という質問に対して，「(a) water (b) fire (c) magnet」の回答候補が存在する場合，回答候補 i から生成される事実文を F_i とすると，以下のような 3 つの事実文が生成される．

F_a water should not be brought near the spray can.

F_b fire should not be brought near the spray can.

F_c magnet should not be brought near the spray can.

次に，F ごとに，検索エンジンを用いて，情報源上の関連する文書のうち，上位 d 件を取得する．この情報源は，3.1 節で述べた，前提知識辞書とは異なる．これらの文書はすべて，事実文 F に対する根拠文 R として取り扱う．この時点で R は，F と表層的に類似した文書であり，真に妥当な根拠であるかは分からない．F_i をクエリとして取得された上位 n 件の文書のうち，j 番目の文書を R_{ij} とする．この時点で，事実文 F ごとに d 個の根拠文 R が存在することとなる．最後に，妥当性判断モデルを用いて，各回答候補 i が正解である可能性を示すスコア S_i を算出する．妥当性判断モデルを $Model$(事実文, 根拠文) とすると，S_i は式 9 によって導かれる．つまり，1 つの回答候補 i につき，妥当性判断モデルから q 件の確率値が得られるため，それらの平均値を，当該回答候補が正解である可能性を示すスコアとする．

$$S_i = \frac{1}{q}\sum_{j=1}^{q} Model(F_i, R_{ij}) \tag{9}$$

回答候補ごとにスコア S を得たうえで，この値が最も大きい回答候補を，自動回答の最終的な回答として出力する．

4 実験

本節で使用する，妥当性判断モデルの学習データ，自動回答時の質問データ，および，前提知識の使用言語は，英語のみである．そのため，形態素解析，共参照解決，構文解析などの自然言語処理には，Stanford CoreNLP [12][4]を使用する．また，情報源から関連文を取得する処理においては，検索エンジンとして，オープンソースの全文検索システムである Apache Solr を使用する．

4.1 実験 1：妥当性判断モデルの精度

4.1.1 目的

本実験では，提案した妥当性判断モデルの精度について調査する．使用する学習データ，および，前提知識の考慮の有無の組み合わせによって，複数条件でモデルを構築し，どの程度正確に推定できるかについて比較する．

4.1.2 学習用データセット

ここでは，妥当性判断モデルの学習用データセットについて述べる．

まず，表 1 は，用意した学習用データの一覧である．データの作成タイプは，**タイプ A** と**タイプ B** の 2 つに分かれ，作成方法が大きく異なる．

4：Stanford CoreNLP Natural language software: https://stanfordnlp.github.io/CoreNLP/

5：Wikibooks: https://en.wikibooks.org/

表 1　学習用データの一覧

データ名	作成タイプ	作成元	正例数	負例数
Wiki	タイプ A	Wikipedia	401,317	400,790
SimpleWiki	タイプ A	Simple English Wikipedia	9,512	9,469
Wikibooks	タイプ A	Wikibooks[5]	24,186	24,192
OpenBookQA	タイプ B	Open Book QA [3]	24,613	73,850
ARC-Easy	タイプ B	AI2 Reasoning Challenge [2] Easy セット	11,255	33,755
ARC-Challenge	タイプ B	AI2 Reasoning Challenge Challenge セット	5,590	16,785

a）タイプ A

質問応答とは関係ない，Wikipedia などの一般的な文書から作成するタイプであり，以下の手順で作成される．

（1） 作成元データを段落ごとに区切り，一段落ごとに共参照解決を行う．

（2） 各段落を一文ごとに区切り，単語「because」が含まれる文のみを抽出する．ただし，「because of」の文節が含まれる文は対象外とする．

（3） 抽出された各文に対して，句構造に基づく構文解析を行い，「because」に含まれる文節をすべて根拠文，それ以外の文節を事実文として抽出し，正例の対とする．

（4） 手順 3 で取得した各正例対ごとに，当該事実文と表層的に類似した他の対の事実文を探索し，正例の根拠文と対とすることで，負例を作成する．

手順 1 では，共参照の解決を行う．特に英語の場合，一度出現した名詞は，2 回目以降，代名詞に置き換わる場合が多いため，共参照解決によって，それら代名詞などをすべてもとの単語に戻す．

手順 2 および 3 では，単語「because」を含む文から，事実文と根拠文の正例対を抽出する．これは，「because」を含む文が，「fire should not be brought near the spray can because LPG gas is used for common spray cans.」といったように，事実と根拠の両方を含む可能性が高いと考えられるためである．たとえば，先の例文の場合，「fire should not be brought near the spray can」という事実文と，「LPG gas is used for common spray cans」という根拠文が抽出される．ただし，「because of」の文節が含まれる文は，意味が異なるため，抽出対象外とする．

手順 4 では，正例対をもとに，疑似的に負例対を作成する．ここでは，同じ根拠文に対して，正例の事実文に表層的に類似している事実文を，負例として取り扱う．類似の指標として，ジャロ・ウィンクラー距離（Jaro-Winkler distance）[13] を用いる．この距離は，0.0〜1.0 の実数値となり，同一の文を比較すると，距離 1.0 となる．類似する文を探索する際には，距離が限りなく 1.0 に近い文を選択する．

以上により，正例と負例が同じ件数だけ作成されるが，すべてを学習用データにはせず，9 割を学習データとする．残りの 1 割のデータは，開発用データとして用いる．

b）タイプ B

多肢選択問題の質問応答データセットから作成するタイプである．質問応答データセット内には，質問文と，正解および不正解からなる複数の回答候補が存在し，本実験用の学習データは以下の手順で作成される．

（1） 質問ごとに，選択肢の個数分の事実文を作成する．

（2） データセットに付随のコーパスを情報源として，事実文ごとに，情報源に対して検索を行い，関連文を 5 件取得する．それらをすべて根拠文として取り扱う．

（3） 質問の正解の選択肢から得られた対を正例，不正解の選択肢から得られた対を負例とする．

手順 1 では，質問文と選択肢（回答候補）から，事実文を作成する．作成には，3.2 節で述べたものと同じルールを用いる．

手順 2 では，根拠文を取得する．Open Book QA ならびに ARC には，それぞれ，質問のデータセットとは別に，質問を解く際の情報源となるコーパスが用意されているため，それらを根拠として取り扱う．クエリには，事実文を用いる．ただし，クエリによっては，必ずしも 5 件の関連文を取得できるとは限らないため，そのような場合は，取得できた件数分だけで同じ処理を行う．

以上の手順によって，仮に，データセット内の k 件の多肢選択問題がすべて，正解が 1 択のみの l 択問題 ($l \geq 2$) であり，常に関連文が 5 件取得できた場合，手順 3 で得られる正例は $k \times 1 \times 5$ 件，負例は $k \times (l-1) \times 5$ 件となる．

なお，OpenBookQA および ARC の配布データは，全質問データを「学習用」「評価用」「テスト用」の 3 種類に分けて提供しているため，本実験では，「学習用」のみを学習用データとする．

4.1.3 評価用データセット

続いて，本実験で使用する評価用データセットについて述べる．

評価用データには，Open Book QA で提供されている，「評価用」データの Additional データを用いる．Additional データには，通常のデータに加えて，当該質問が作られるもととなった，付随コーパス内の 1 つの根拠文がすでに用意されている．よって，以下の手順により，評価用データを作成する．

（1） 質問ごとに，選択肢の個数分の事実文を作成する．

（2） 当該質問のすべての事実文に対して，Additional データ上で用意されている根拠文を対にさせる

（3） 質問の正解の選択肢から得られた対を正例，不正解の選択肢から得られた対を負例とする．

（4） Open Book QA では，質問ごとに，正例が 1 つ，負例が 3 つ作成される．正例と負例の割合を同一とするため，質問ごとに，ランダムに 2 件の負例を消去する．

4.1.4 方　　法

本実験では，表 2 に示す 6 種類の条件で，それぞれ妥当性判断モデルを構築し，モデルの精度を比較する．

表 2　妥当性判断モデルの構築条件の一覧

条件番号	使用する学習データ	前提知識を考慮するか
C1	タイプ A	しない
C2	タイプ A	する
C3	タイプ B	しない
C4	タイプ B	する
C5	タイプ A とタイプ B	しない
C6	タイプ A とタイプ B	する

各条件の違いは，学習時にどのデータを使用するか，および，

前提知識を考慮するかどうかである．学習データは，4.1.2 項で述べたように，作成方法の違いで 2 種類に分かれるため，どちらか片方を使用するか，両方とも使用するかで，条件が分かれる．また，前提知識を考慮するかどうかについては，3.1 節で示したモデルの学習時に，妥当性判断の要素として，$h^{FactKnowledge}$ と $h^{ReasonKnowledge}$ を使用するか否かの違いである．もし，使用しない場合は，式 5 で示した h^{Fact} の生成式は式 10 へ，式 6 で示した h^{Reason} の生成式は式 11 へ置き換えられる．

$$h^{Fact} = \mathrm{f}\left(W_{Fact2} \times \mathrm{f}\left(W_{Fact1} \times h^{FactSentence}\right)\right) \tag{10}$$

$$h^{Reason} = \mathrm{f}\left(W_{Reason2} \times \mathrm{f}\left(W_{Reason1} \times h^{ReasonSentence}\right)\right) \tag{11}$$

モデル構築時のパラメータは，すべての条件において，次のとおり統一する．

- 事実文用 LSTM，根拠文用 LSTM，単一前提知識用 LSTM の 3 つにおいて使用する，単語の分散表現は，学習済みの 100 次元の GloVe モデル [14] を使用する．各 LSTM の出力次元数は，各方向において，事実文用 LSTM および根拠文用 LSTM は 32 次元，単一前提知識用 LSTM は 64 次元，全前提知識用 LSTM は 32 次元とする．
- 全結合層の出力次元数は，W^{Fact1} と $W^{Reason1}$ は 16 次元，W^{Fact2} と $W^{Reason2}$ は 8 次元，W^{Final1} は 8 次元とする．
- 最終層以外のすべての LSTM および全結合層には，学習時に一定割合のノードを不活性化させる Dropout [15] を導入する．不活性化割合は，$LSTM^{Fact}$，$LSTM^{Reason}$，$LSTM^{KnowledgeSingle}$，$LSTM^{KnowledgeAll}$ は 0.25，W^{Fact1} と $W^{Reason1}$ は 0.25，W^{Fact2} と $W^{Reason2}$ は 0.50，W^{Final1} は 0.75 とする．
- W^{Fact1}，$W^{Reason1}$，W^{Fact2}，$W^{Reason2}$ および W^{Final1} には，バッチ正則化（Batch Normalization）[16] を適用する．
- 最適化アルゴリズムには Adam [17] を使用し，各パラメータは $\alpha = 0.001$，$\beta 1 = 0.9$，$\beta 2 = 0.999$，$\epsilon = 10^{-8}$ とする．
- 学習全体のエポック数は 50，ミニバッチサイズは 512 とする．

モデルの精度の指標としては，評価用データを入力させた際の，「妥当である / 妥当ではない」の 2 値判断の正答率を用いる．モデルの出力 p は確率値であるため，$p \geqq 0.5$ の場合は「妥当である」，$p < 0.5$ の場合は「妥当ではない」と判断したこととする．精度は，モデルが最終エポックまで学習し終えた時点で評価する．

4.1.5 結　　果

各構築条件におけるモデルの精度は，表 3 のとおりとなった．

表 3　妥当性判断モデルの各構築条件における精度

条件番号	精度
C1	0.507
C2	0.506
C3	0.553
C4	0.551
C5	0.546
C6	0.544

4.2 実験 2：自動回答の正答率

4.2.1 目　　的

本実験では，4.1 節の実験 1 で構築された妥当性判断モデルを，3.2 節で提案した自動回答時に使用することで，与えられた質問に対しどの程度正答するかを調査する．同時に，モデルの構築条件の違いによって，正答率にどの程度影響を与えるかについて比較する．

4.2.2 データセット

質問のデータセットとして，表 4 に示す 3 種類の多肢選択問題を使用する．ただし，いずれのデータセットも，「テスト用」として提供されているもののみを使用する．

表 4　自動回答させる質問のデータセットの一覧

データ名	質問数	情報源として使用するコーパス	コーパスの文数
OpenBookQA	500	Open Book QA に付随のコーパス（open book）	1,326
ARC-Easy	2,376	AI2 Reasoning Challenge に付属のコーパス	14,621,856
ARC-Challenge	1,172	（同上）	（同上）

表 4 に示すとおり，質問のデータセットによって，根拠文を取得する元となる情報源は変更する．

OpenBookQA は，すべて 4 択の多肢選択問題である．ARC は，3 択や 5 択の問題が数件存在するものの，ほぼすべてが，4 択の多肢選択問題である．

4.2.3 方　　法

本実験では，4.1 節の実験 1 で構築したモデルを用いて，データセットの質問に自動回答する．自動回答時のパラメータは，$d = 5$ とする．

実験結果としては，使用したモデル，および，自動回答させたデータセットごとに，自動回答の正答率を算出する．与えられた多肢選択問題に対して，本来正解である選択肢を，システムが回答とした場合は，システムはその問題に対して正解したとみなし，それ以外の場合は，不正解であったとみなす．正答率は，全問題数に対して，システムが正解した問題数の割合である．

4.2.4 結　　果

使用したモデル，および，自動回答させたデータセットごとの正答率は，表 5 のとおりである．ただし，表の下段は，同じデータセットを用いた，他の手法の正答率である．

表 5　自動回答の正答率

モデルの条件番号または手法名	自動回答させたデータセット		
	OpenBookQA	ARC-Easy	ARC-Challenge
C1	0.240	0.244	0.255
C2	0.244	0.252	0.244
C3	0.352	0.271	0.272
C4	0.386	0.301	0.253
C5	0.344	0.278	0.232
C6	0.354	0.283	0.256
Sun ら [4] の手法 [6]	0.696	0.764	0.538

6 : 2018 年 12 月 31 日時点.

5 考　　察

本節では，4 章の実験結果について考察する．

5.1 考察 1：妥当性判断モデルの精度

4.1 節の実験 1 では，本章で提案した妥当性判断モデルの精度について調べた．本実験の精度は，正例と負例の件数が同一の評価データに対する，2 値判断の結果により算出されているため，ランダムに判定した場合の精度は 0.50 であるといえる．実験では，複数の条件でモデルを構築したが，評価データに対する精度は，どの条件においても 0.50〜0.55 程度であった．

まず，使用した学習データの違いによる点からみると，各モデル間で差があることが分かる．Wikipedia などから生成したタイプ A の学習データのみを用いて学習させた場合，評価データに対する妥当性判断は，精度が 0.50 程度であり，ランダムに判定した精度と同等である．対して，OpenBookQA などの質問応答のデータセットから生成したタイプ B を用いた場合，精度は 0.55 程度となった．これより，Wikipedia 等などの記事上に存在する，単語「because」を含む文から生成される事実と根拠の対と，OpenBookQA から生成される対（評価データ）の間には，大きな隔たりがあり，タイプ A を学習させても，本実験で期待する妥当性判断は行えないことが分かった．実際に，4.1.2 項で述べた，タイプ A 作成時に残した考察用評価データで，C1 および C2 の精度を評価したところ，精度は 0.60 程度となり，タイプ A 内で完結させた場合は，より良い精度となった．タイプ B のみを用いた C3 および C4 の，評価データに対する精度は，0.55 程度であったが，学習データに対する精度を調べたところ，0.90 を上回っており，過学習[7]を起こしていることがわかった．この原因の一つは，学習用データが少量であることが考えられる．タイプ A とタイプ B の両方を用いた C5 および C6 は，過学習は抑えられたものの，先に述べたように，タイプ A の学習データが評価データに対して有効に働かなかったため，精度向上にはつながらなかった．

次に，前提知識を考慮するか否かの違いでみると，精度の向上にはつながらなかったことが分かる．原因の一つとして，用意した辞書の構築手法が挙げられる．本章では，3.1.3 項で述べたように，Simple English Wikipedia から，前提知識用の辞書を構築している．単純な構築手法であるため，同じ表層の単語に複数の意味がある場合や，複数単語で一つの意味をなすような場合に，現時点では対応できていない．つまり，事実文や根拠文の文脈に適した前提知識を取得できていない可能性がある．また，提案手法では，妥当性判断モデルの学習時に，前提知識をうまく活用した学習ができていない可能性がある．妥当性判断の際には，事実文や根拠文，前提知識内のどこに着目すべきかを考慮する必要があると考えられるため，注意機構などの導入を検討する必要がある．

最後に，各モデルの詳細な判断結果と，適合率および再現率を，表 6 に示す．ここで，適合率とは，モデルが「妥当である（妥当ではない）」と判断した対のうち，正解も「妥当である（妥当ではない）」となる対の割合である．また，再現率とは，正解が「妥当である（妥当ではない）」となっている対のうち，モデルが「妥当である（妥当ではない）」と判断した対の割合である．C3〜C6 の結果をみると，モデルは「妥当ではない」という判定をしやすい傾向であることが分かる．これは，タイプ B の学習データの正例:負例の割合が，およそ 1:3 であるためだと考えられる．

表 6　妥当性判断モデルの判断結果と適合率および再現率

		モデルの条件番号	正解件数 妥当ではない	正解件数 妥当である	合計件数	適合率
判断結果件数	妥当ではない	C1	285	277	562	0.507
		C2	225	218	443	0.507
		C3	388	335	723	0.536
		C4	384	333	717	0.535
		C5	384	337	721	0.532
		C6	359	314	673	0.533
	妥当である	C1	215	223	438	0.509
		C2	275	282	557	0.506
		C3	112	165	277	0.595
		C4	116	167	283	0.590
		C5	116	163	279	0.584
		C6	141	186	327	0.568
合計件数		C1–C6	500	500	1000	
再現率		C1	0.570	0.446		
		C2	0.450	0.564		
		C3	0.776	0.330		
		C4	0.768	0.334		
		C5	0.768	0.326		
		C6	0.718	0.372		

5.2 考察 2：自動回答の正答率

4.2 節の実験 2 では，実験 1 で構築した妥当性判断モデルを使用して，本章で提案した自動回答手法の精度について調べた．本実験は，ほぼすべてが 4 択である多肢選択問題に対する正答率を，結果として算出しているため，ランダムに自動回答した場合の正答率は 0.25 程度といえる．

C1 および C2 は，ランダムな自動回答の精度と同等であり，C3〜C6 は，0.30 を超えたデータセットが存在する．また，妥当性判断モデルの精度の傾向とは異なり，前提知識を考慮したモデルを用いた自動回答のほうが，正答率が高い．ただし，いずれのモデルを使用したしたとしても，Sun ら [4] の手法による自動回答の正答率を，大幅に下回っている．よって，妥当性判断モデルの構築手法を含む，自動回答手法の大幅な見直しが必要であると考えられる．

C6 において，提案手法が正解した問題例を，図 3 に示す．質問は，「舗装道路のすぐそばに植えられたオークの木が成長し，根が伸びたときにどうなるか」といった内容であり，正解は「(C) コンクリートを破壊する」である．自動回答処理内で，妥当性判断を行った結果を見ると，正解である選択肢（C）から生成された事実と，検索で得られた根拠の対の妥当性確率が，他の選択肢と比較して高いことが分かる．しかし，根拠として得られた実際の文をみてみると，根拠としては不適当であると思われる．他の問題においても，正解した問題，不正解した問題ともに，妥当性判断の結果が不適切であったものが多く見ら

7：過学習: 学習用データに対して過度に適応している状態のこと．

れた．これは，5.1 節の考察でも述べたように，妥当性判断モデルの精度が不十分であるためである．また，現在の手法の場合，根拠文を検索する際に，真に根拠となる文が取得できる保証はない．よって，確実に根拠を得る手法や，検索に頼らない手法を検討する必要がある．

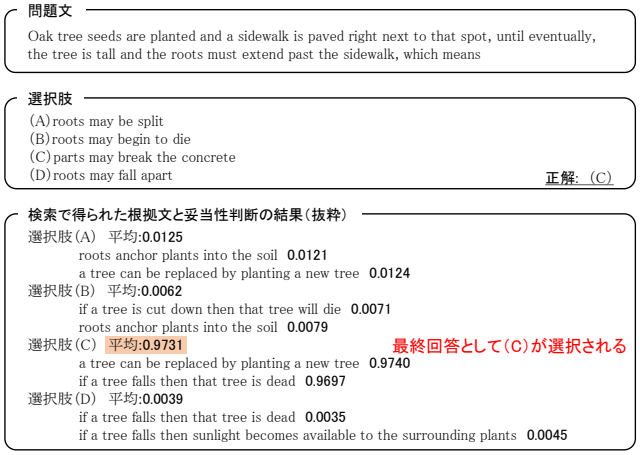

図 3　自動回答の正解例

さらに，自動回答された問題を筆者が確認したところ，同じ事実文に対して，どのような根拠文を対として妥当性を判断させたとしても，妥当性判断モデルが出力する妥当性確率の値は，C1～C6 のいずれにおいても，ほぼ同値になりやすい傾向がみられた．この原因としては，構築した妥当性判断モデルが，事実文のみから妥当性を判断しようとしている可能性が考えられる．よって，モデル内での計算方法や，学習データの見直しを検討する必要がある．

6　ま　と　め

本章では，多肢選択問題に焦点を当て，前提知識に基づき，**事実**と**根拠**の対の妥当性を判断することによって，回答を導出する手法を提案した．具体的には，質問文と回答候補から生成される事実文と，情報源上の根拠文の対が入力されると，各文を解釈するのに必要と思われる前提知識を参照したうえで，当該対の妥当性を判断するようなモデル（妥当性判断モデル）を構築した．自動回答時には，本モデルを用い，一問の質問から想定される複数の事実と根拠の対の妥当性を判断したうえで，最終的な回答を導出した．

妥当性判断モデルの精度を調査した結果，どのような条件でモデルを構築したとしても，本章で想定していた妥当性を，高い精度で判断することはできなかった．また，当該モデルを用いた自動回答においても，2018 年時点で最も成績の良い手法による正答率を，大幅に下回る結果となった．

今後，事実と根拠の妥当性判断の精度向上のためには，モデル構造や学習データの見直しが必要であると思われる．また，自動回答時には，検索によって根拠を得るのではなく，前提知識を含め，あらかじめ大量の根拠を妥当性判断モデルが学習したうえで，妥当性を判断させるような処理を検討する必要がある．

謝　　辞

本研究の一部は科研費 18K11557 の助成を受けたものです．ここに記して感謝の意を表します．

文　　献

[1] Pranav Rajpurkar, Robin Jia, and Percy Liang. Know what you don't know: Unanswerable questions for squad. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pp. 784–789. Association for Computational Linguistics, 2018.

[2] Peter Clark, Isaac Cowhey, Oren Etzioni, Tushar Khot, Ashish Sabharwal, Carissa Schoenick, and Oyvind Tafjord. Think you have solved question answering? try arc, the ai2 reasoning challenge. *arXiv preprint arXiv:1803.05457*, 03 2018.

[3] Todor Mihaylov, Peter Clark, Tushar Khot, and Ashish Sabharwal. Can a suit of armor conduct electricity? a new dataset for open book question answering. In *EMNLP*, 2018.

[4] Kai Sun, Dian Yu, Dong Yu, and Claire Cardie. Improving machine reading comprehension with general reading strategies. *CoRR*, Vol. abs/1810.13441, , 2018.

[5] Alec Radford, Karthik Narasimhan, Tim Salimans, and Ilya Sutskever. Improving language understanding by generative pre-training. 2018.

[6] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. In *Advances in Neural Information Processing Systems*, pp. 5998–6008, 2017.

[7] Guokun Lai, Qizhe Xie, Hanxiao Liu, Yiming Yang, and Eduard Hovy. Race: Large-scale reading comprehension dataset from examinations. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pp. 785–794, 2017.

[8] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*, 2018.

[9] Mike Schuster and Kuldip K. Paliwal. Bidirectional recurrent neural networks. *IEEE Transactions on Signal Processing*, Vol. 45, No. 11, pp. 2673–2681, 11 1997.

[10] Sepp Hochreiter and Jürgen Schmidhuber. Long short-term memory. *Neural computation*, Vol. 9, No. 8, pp. 1735–1780, 1997.

[11] Djork-Arné Clevert, Thomas Unterthiner, and Sepp Hochreiter. Fast and accurate deep network learning by exponential linear units (elus). *arXiv preprint arXiv:1511.07289*, 2015.

[12] Christopher D. Manning, Mihai Surdeanu, John Bauer, Jenny Finkel, Steven J. Bethard, and David McClosky. The Stanford CoreNLP natural language processing toolkit. In *Association for Computational Linguistics (ACL) System Demonstrations*, pp. 55–60, 2014.

[13] William E Winkler. String comparator metrics and enhanced decision rules in the fellegi-sunter model of record linkage., 1990.

[14] Jeffrey Pennington, Richard Socher, and Christopher Manning. Glove: Global vectors for word representation. In *Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP)*, pp. 1532–1543, 2014.

[15] Nitish Srivastava, Geoffrey Hinton, Alex Krizhevsky, Ilya Sutskever, and Ruslan Salakhutdinov. Dropout: a simple way to prevent neural networks from overfitting. *The Journal of Machine Learning Research*, Vol. 15, No. 1, pp. 1929–1958, 2014.

[16] Sergey Ioffe and Christian Szegedy. Batch normalization: Accelerating deep network training by reducing internal covariate shift. *arXiv preprint arXiv:1502.03167*, 2015.

[17] Diederik P Kingma and Jimmy Ba. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*, 2014.