ファクトイド型質問応答を用いた正誤判定問題の解決

金山 博
日本アイ・ビー・エム株式会社
東京基礎研究所
hkana@jp.ibm.com

宮尾 祐介
国立情報学研究所
yusuke@nii.ac.jp

1　はじめに

歴史の試験問題などに見られる客観式問題ではしばしば、記述された文の真偽を判定させることによって受験者の理解度を測る。この種の**真偽判定問題**を機械的に直接解こうとすると、閉世界を仮定できるだけの網羅性をもった知識が必要となるが、出題の幅広さに対応させることは現実的とはいえない。言い換えると、命題が「偽」である根拠を情報源から明示的に見つけることは困難である。そのためか、ファクトイド型を中心とした質問応答が成果を挙げている一方で、大規模な知識源を用いて真偽を判定する研究は少ない。

そこで、真偽判定問題を、定理証明のアプローチとは異なる見方で捉えるべく、以下の例をもとに[1]、知識を確認するための命題の作られ方について考えてみる。

(1) Chirac was the president of France in 2000.
*(2) Chirac was the president of Germany in 2000.

(2) のような偽の命題はしばしば、(1) のような真の命題をもとにして、一部の要素（この場合は国名）が入れ替わることによって作られる。本稿ではこれに着目し、命題の中に現れる語句を問うような質問文を生成し、それを解くことによって真偽の判定を行うアプローチを試みる。(1)(2) の例の場合、下線部を上位語で置き換えることにより、文 (3) を生成する。

(3) Chirac was the president of *this country* in 2000.

(3) を、斜体字の部分を問う**ファクトイド型**の質問[2]とみなし、既存の質問応答システムへの入力とする。France が第 1 位の解として出力されれば、命題 (1) が真である根拠となると同時に、(2) が偽であることが示唆される。

本論文では、2 節で述べる大学入試センター試験を題材として、真偽判定問題をファクトイド型質問応答に帰着させる方法の検証を行う。この手法の利点は、情報源へのアクセスの手段として、ある程度確立された質問応答の技術が活用できるとともに、単なる真偽の判定にとどまらない、題意を捉えるような処理が可能となる点である。3 節では前提となる質問応答システム DeepQA の動作について紹介する。4 節では命題を質問文に変換して DeepQA の結果をもとに真偽判定をする方法について述べ、5 節で実験および考察を行う。なお、実システムの構築よりも方法論の検証を重視するため、質問文の変換などの一部の処理は人手で行っている。

表 1: 真偽を問う問題の例（2009 年度センター試験 世界史 B より）。正答は 3 である。

唐代から宋代にかけての科挙の合格者である人物について述べた文として正しいものを次のうちから一つ選べ。	
1	欧陽脩や蘇軾は，唐代を代表する文筆家である。
2	顔真卿は，宋代を代表する書家である。
3	宋の王安石は，新法と呼ばれる改革を行った。
4	秦檜は，元との関係をめぐり主戦派と対立した。

2　大学入試センター試験

本研究では、大学入試センター試験[3]の世界史 B の問題を開発及び評価に用いる。その中では、表 1 に示すように、選択肢の中から真の（あるいは偽の）命題を選択させる設問が大半を占める。各選択肢は真偽が明確に求まるよう設計されていることと、多くは一般的な知識を用いて判定できることから、本稿で提案する手法の評価に適している。

以下の節では、国立情報学研究所人工知能プロジェクト[4][5] において XML 化され提供されているデータを利用する。但し、3 節にて述べるシステムが英語に基づいているため、実験では全問題文を専門家が英語に翻訳したものを用いる。元の日本語版データはテキスト間の含意・換言・矛盾の認識を目的とした NTCIR RITE タスク[5]においても利用されており [3]、手法の比較にも好適である。

3　DeepQA

オープンドメイン向け質問応答システム DeepQA[1] は、ファクトイド型の質問文を入力とし、複数の答えを 0〜1 の実数値で示される確信度とともに出力する。(3) のように、答えに相当する部分が「this + 上位語」または人称代名詞に置換された平叙文を受け付ける。

複数の語彙体系などの構造化情報に加えて、百科事典や新聞記事などのテキスト文書を知識源として蓄え

[1] * は偽であることを示す。
[2] (3) で国名を問うているように、固有表現等で単答できるタイプの質問。
[3] http://www.dnc.ac.jp/
[4] http://21robot.org/
[5] http://www.cl.ecei.tohoku.ac.jp/rite2/

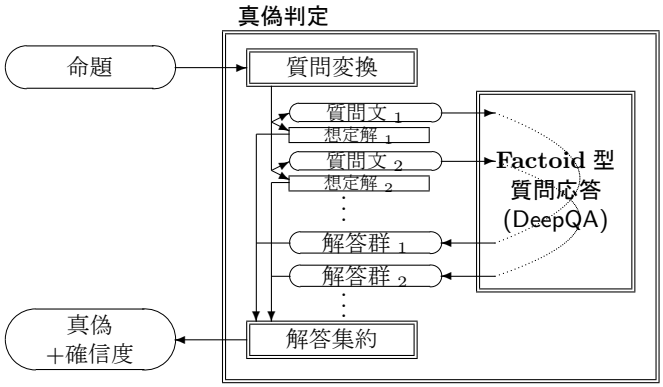

図 1: ファクトイド型質問応答システムを用いた真偽判定の流れ。

ている。これらの知識源を用いて、柔軟に解候補を列挙し、多様な構文・意味上の処理を用いて解候補と問題文のマッチングを行うため、特定の分野に依存せず高い正答率で問題を解くことができる。TREC 2002 [4] の質問応答のセットを用いた実験では、約 8,000 の質問と答えのペアを用いて学習した結果、約 67% という高い正答率を実現している。本研究においては、世界史分野に特化したチューニングや情報源の追加は一切行っていない。

4 質問変換による真偽判定手法

提案する真偽判定の手法は、図 1 に示すように、(1) **質問変換**（命題からファクトイド型の質問文を生成）、(2) **解答集約**（質問文を DeepQA で解いた結果を想定解と比較して真偽を判定）という 2 段階からなる。以下でこれらを順に述べる。さらなる詳細は文献 [2] を参照されたい。

4.1 質問変換

まず、命題中にある固有名詞を選択する。表 1 の選択肢 2 の英訳の場合、(4) の下線部のように 2 つの固有名詞がある。

(4) Yen Chen-ching was the best known chirographer in the Sung dynasty.

これらの固有名詞の型（上位語）を求め、this を付けた名詞句に変換する。人物の場合は、明白な場合はその型（職業など）を、そうでない場合は性別に応じて he や she を用いる。

(4a) *He* was the best known chirographer in the Sung dynasty.
(4b) Yen Chen-ching was the best known chirographer in *this dynasty*.

置換する前の語を**想定解**と呼ぶ。(4a) と (4b) の場合、(4) の下線部にあたる Yen Chen-ching と Sung dynasty がそれぞれの想定解である。

このような質問文を DeepQA に入力すると、表 2 に見られるように、複数の答えが確信度とともに返される。質問 q において、想定解が解答群の n 位に現れた時を $A(q) = n$ と書くと、表 2 の例では $A(q_1) = A(q_2) = 1$, $A(q_3) = 2$, $A(q_4) = \text{nil}$ となる。想定解と出力解のマッチングの際には、同義語や大文字・小文字の違いなどを柔軟に処理している。

上記の処理は自動的に行うこともできるが、本研究では、質問文の生成と同義語の処理を手動で行った。固有名詞の同定、上位語の検出や同義語の判定などにおける誤りを避けて、ファクトイド型の質問応答への帰着の有効性を検証するためである。この際、公平のために、元の命題の真偽は見ることなく想定解と出力解をもとに同義語リストを作成するなどの配慮をしている。

4.2 解答集約

上記で定義した $A(q)$ を用いて真偽を判定する方法について考える。直感的に、$A(q) = 1$ となる時に元の命題が真、そうでない時に偽であるといえるが、表 2 を見ると、q_1〜q_4 のどれを参照するかによって結果が異なる。すなわち、真偽判定に用いる最適な質問文の選択が解答集約のポイントとなる。

そこで、質問文 q の**信頼度** $R(q)$ を用いて以下のように定式化する。

$$TF(Q) = \begin{cases} \mathsf{true} & A(q^*) = 1 \\ \mathsf{false} & \text{otherwise} \end{cases} \quad \text{where} \quad q^* = \operatorname*{argmax}_{q \in Q}(R(q)) \tag{5}$$

となる。信頼度の関数 $R(q)$ として、まずは最も簡単な $R_1(q)$ を、DeepQA の n 位の解に対する確信度 $C(q, n)$ を用いて、

$$R_1(q) = C(q, 1) \tag{6}$$

と定義する。表 2 の例では、解の確信度から $q^* = q_1$ となり、想定解と比較すると $A(q^*) = 1$、よって $TF(Q) = \mathsf{true}$ と、正しく真偽が判定される。

しかし、判定に失敗する例も残っている。DeepQA が質問文を正しく解釈できなかったり、正しい答えが情報源に見つからないケースだけでなく、2 つの本質的な問題点が見出されたので、以下にて対処する。

複数の解を持つ質問文　1 節で示した命題 (1) は充分な文脈を持っていたが、以下の命題 (7) を見てみよう。

(7) Chirac was the president of France.
(7a) *He* was the president of France.
(7b) Chirac was the president of *this country*.

質問文 (7b) は問題ないが、(7a) は Chirac の他にも Mitterand, Sarkozy, Hollande などの正解があり、DeepQA が Chirac を 1 位として出力する保証がない。このような場合は、1 位と 2 位の確信度が近接すると考えられる。そこで、確信度の比が閾値 θ_2（1 未満の数）を超える場合に、質問文の信頼度にペナルティ p_2

表 2: 'In China, Xuanzang traveled to India and brought home the Buddhist scriptures during the Tang period.' という真の命題に対して生成した 4 つの質問文に対し、DeepQA が返す解答と確信度。

	生成された質問文	想定解	1 位	2 位	3 位
q_1	In *this country*, Xuanzang traveled to India and brought home the Buddhist scriptures during the Tang period.	[China]	China 0.702	Nepal 0.513	Burma 0.129
q_2	In China, *he* traveled to India and brought home the Buddhist scriptures during the Tang period.	[Xuanzang]	Xuanzang 0.593	Tang 0.147	Buddhism 0.11
q_3	In China, Xuanzang traveled to *this country* and brought home the Buddhist scriptures during the Tang period.	[India]	monk 0.23	India 0.216	Faxian 0.135
q_4	In China, Xuanzang traveled to India and brought home the Buddhist scriptures during *this period*.	[Tang period]	Buddhism 0.441	monk 0.219	Faxian 0.127

を乗じることによって、複数の解を持つ質問文が選ばれにくくなるような信頼度 $R_2(q)$ を定義する。

$$R_2(q) = \begin{cases} p_2 R_1(q) & \frac{C(q,2)}{C(q,1)} > \theta_2 \\ R_1(q) & \text{otherwise} \end{cases} \quad (8)$$

属性を問う質問　偽の命題 (9) は 3 つの固有名詞を持つ。質問文 (9a) は理想的で、DeepQA は Philip of Spain を返し[6]、想定解の Carlos I と比較することにより正しく「偽」だと判定できる。しかし、(9b) のように、Carlos I と同格の表現となっている部分を問う質問文だと、命題全体の真偽に拘わらず、Spain が解として出力されてしまううえ、確信度が高くなる傾向があり、誤って「真」と判定する原因となる。

*(9) Carlos I, the King of Spain, was also the King of Portugal.
(9a) *He*, the King of Spain, was also the King of Portugal.
(9b) Carlos I, the King of *this country*, was also the king of Portugal.

このような誤判定を防ぐために、他の固有名詞を修飾している節（同格を含む）の中にある固有名詞を置き換えたものは「属性を問う質問文」であると判断し、1 より小さな値 p_3 を乗じることによってペナルティを与える。これを用いて式 (10) のように R_3 を定義する。

$$R_3(q) = \begin{cases} p_3 R_2(q) & q \text{ は属性を問う質問文} \\ R_2(q) & \text{otherwise} \end{cases} \quad (10)$$

システムは、真・偽の判定に加えて、その確信度 $C^*(Q)$ を出力する。これには、判定に用いた質問文の 1 位の解答の確信度を使う。すなわち、$C^*(Q) = C(q^*, 1)$ となる。

5 実験

5.1 環境と評価方法

大学入試センター試験の世界史 B の本試験問題うち、2009 年版を開発・観察のために、2007 年版をテストに用いた。それぞれ、真ないし偽の記述を 4 つの選択肢から選ばせる問題が 26 問、23 問あったため、2009 年版の 104 命題、2007 年版の 92 命題が利用できた。

[6] (9a) にこれを埋め込めば真の命題となる。

表 3: 2009 年本試験を用いたクローズドテストの結果。

手法	真偽判定	p	四択問題
ベースライン	65.4% (68/104)		25%
Model 1	77.8% (81/104)	.043	50% (13/26)
Model 2	79.8% (83/104)	.029	58% (15/26)
Model 3	81.7% (85/104)	.017	62% (16/26)

いくつかの場合は、選択肢の文に対して問題文の一部を補うことによって、命題の真偽が判定できるようになる。そのような場合には人手で前処理を行った。これは、含意関係認識のタスクに同じデータを用いた時 [6] と共通の方法である。

評価は、個々の選択肢の真偽判定と、四択問題の正解率判定の 2 種類で行う。四択問題の評価を行うために、命題の正しさの度合いを表現するスコア $CS(Q)$ を導入する。

$$CS(Q) = \begin{cases} C^*(Q) & TF(Q) = \mathsf{true} \\ (-1)C^*(Q) & TF(Q) = \mathsf{false} \end{cases} \quad (11)$$

「正しい選択肢を一つ選ぶ」という問題の場合は、$CS(Q)$ が最大となる選択肢を解答とする。逆に「誤った選択肢を一つ選ぶ」問題では、$CS(Q)$ が最小のものを解答とし、正解と比較する。

5.2 実験結果

表 3 は、開発データと同じ問題を使ったクローズドテストの結果である。ベースラインとは、真偽判定の場合は常に「偽」と答えた場合、四択問題の場合は鉛筆を転がした時の理論値 25% である。Model 1〜3 は、4.2 節で導入した信頼度の定義 R_1, R_2, R_3 に対応しており、R_1 は DeepQA の信頼度をそのまま用いたもの、R_2 は複数の解がある質問文の信頼度を減じたもの、R_3 はさらに属性を問う質問文にペナルティを与えたものである。

2009 年版の問題から経験的に、式 (8) の p_2, θ_2 の値はいずれも 0.5、式 (10) の p_3 は 0.1 に設定した。真偽判定の結果の t 検定の p 値を見ると、本手法の性能はベースラインと有意な差がある（$p < .05$）。データが少数のため Model 1〜3 相互の差は有意とはいえない。

次に、表 4 に、2007 年本試験データを使ったオープンテストの結果を示す。パラメータは 2009 年のデータを用いて設定した。ここでは、Model 3 でベースラインとの有意な差が出ており、真偽判定に使う質問文の選択手法の成果が現れている。

表 4: 2007 年本試験を用いたオープンテストの結果。

手法	真偽判定	p	四択問題
ベースライン	68.4% (63/92)		25%
Model 1	69.6% (64/92)	.500	57% (13/23)
Model 2	71.7% (66/92)	.418	57% (13/23)
Model 3	79.3% (73/92)	.046	65% (15/23)

5.3 失敗例

命題 (12) は真だが、生成された質問 (12a)〜(12c) で想定解が 1 位に現れたものはなかった。これは、February Revolution がフランスだけでなくロシアでも起こっているという曖昧性に起因する問題だった。

(12) French provisional government formed after the February Revolution created National Workshops.
(12a) Provisional government of *this country* formed after the February Revolution created National Workshops.
(12b) French provisional government formed after *this revolution* created National Workshops.
(12c) French provisional government formed after the February Revolution created *this place*.

命題 (13) も興味深い例である。これは偽の命題であるが、質問文 (13a)(13b) の両方に対して想定解が返されてしまった。

*(13) The prices of agricultural products were lowered by the Agricultural Adjustment Act in United States.
(13a) The prices of agricultural products were lowered by *this law* in United States.
(13b) The prices of agricultural products were lowered by the Agricultural Adjustment Act in *this country*.

この場合、raised の部分を lowered とした命題が正しい。解が必ず存在するという仮定の下で相対的に解を見出す DeepQA にとって、物価の上昇・下降に拘わらず Agricultural Adjustment Act や United States よりもふさわしい名詞句が無かったからである。

6 考察

6.1 真偽判定を超えた処理

偽の命題の中の語句を DeepQA の解で置き換えれば真の命題が得られた (9a) の例ように、このシステムは真偽判定をするだけでなく、偽の命題を正しく修正すること、いわば題意を読むことができる。開発データ中の 68 個の偽の命題のうち 11 個については、信頼度 1 位の質問の 1 位の答えに正しい語句が提示された。さらに、68 個のうち 24 個において、3 位までの質問文の 5 位以内の答えの中に、正しい語句を見出すことができた。一方で、時代や「○世紀」が誤りのポイントである場合、DeepQA が正しい解を提示できないことがほとんどであった。この点は既存のシステムを使う限界であるが、イベントが行われた時間にまつわる構造化情報を使うことによって改善できるであろう。

6.2 含意関係認識との比較

真偽判定は含意関係認識と関連が深い。センター試験の問題を含意関係認識の手法で解く試みがあり [5]、各命題に関連する事実が書かれた文（真の命題に対してはそれを含意できるような文）を与えて、世界史 B の四択問題の正解率が最高で 58% であったと報告されている。これは日本語の実験であり、英訳を用いた本手法とは直接の比較はできないとはいえ、本手法では含意関係を保つような文を準備する必要が無いことを考えると、表 4 にある 65% の正解率は、含意関係認識の技術を用いる時よりも充分に高いといえ、提案手法の効果が現れている。

7 まとめ

命題の真偽を判定するために、ファクトイド型質問応答の仕組みを介して大量の情報源にアクセスすることの効果を示した。特に、質問文を複数生成して、その中から真偽判定に適したものを選択する手法を提案するとともに、誤った命題のポイントを見出して正しく修正する能力を示した。本手法は世界史の分野や四択の問題形式に依存しておらず、一般的な知識と人間の誤解によって生成される偽の命題とを弁別するものと考えられ、記述や発話の誤りを自動的に発見するなどといった知的処理への応用が期待される。

謝辞 「大学入試センター験問題データベース センター Ten2011 通常版 全教科セット」の研究目的の利用を許諾していただいた株式会社ジェイシー教育研究所様に感謝いたします。

参考文献

[1] D. A. Ferrucci. Introduction to “This is Watson”. *IBM Journal of Research and Development*, Vol. 56, No. 3.4, pp. 1:1–1:15, 2012.
[2] Hiroshi Kanayama, Yusuke Miyao, and John Prager. Answering Yes/no questions via question inversion. In *Proceedings of the 24th International Conference on Computational Linguistics*, 2012.
[3] Yusuke Miyao, Hideki Shima, Hiroshi Kanayama, and Teruko Mitamura. Evaluating textual entailment recognition for university entrance examinations. *ACM Transactions on Asian Language Information Processing*, 2012.
[4] Ellen M. Voorhees. Overview of the TREC 2002 question answering track. In *Proceedings of the 11th Text REtrieval Conference (TREC)*, pp. 115–123, 2002.
[5] 宮尾祐介, 川添愛. 「大学入試問題を解く」ことから見える言語，知識，世界理解に関する研究課題. 人工知能学会誌, 2012.
[6] 宮尾祐介, 嶋英樹, 金山博, 三田村照子. 大学入試センター試験を題材とした含意関係認識技術の評価. 言語処理学会第 18 回年次大会, 2012.