大学入試問題を解くための質問応答システムにおける現状と課題

石下 円香[†1] 阪本浩太郎[†1†2] 渋木英潔[†2] 狩野芳伸[†3]
三田村 照子[†4] Di Wang[†4] 森辰則[†2] 神門典子[†1]

[†1]国立情報学研究所 [†2]横浜国立大学 [†3]静岡大学 [†4]Carnegie Mellon University
E-mail: {ishioroshi,kando}@nii.ac.jp, {sakamoto,shib,mori}@forest.eis.ynu.ac.jp,
kano@inf.shizuoka.ac.jp, {teruko+,diwang}@cs.cmu.edu

1 はじめに

近年，文書情報に対する情報要求は複雑化，高度化しており，そのような要求を満たすアクセス技術として質問応答が注目されている．質問応答とは，利用者の自然言語による質問に対して情報源となる文書集合から回答そのものを抽出する技術であり，複雑高度な情報要求を自然言語で表現できる点に特徴がある．しかしながら，従来の質問応答に関する研究では，「バラク・オバマとは誰ですか」といった比較的シンプルな形式の質問を扱うものが多かった．一方，現実世界においては，質問の核心に至るまでの背景や経緯を複数文にわたって説明したり，具体的な名称が思い出せずに「アメリカ初の黒人大統領」といった抽象的な表現を用いたりするなど，従来研究で想定されたのと異なる質問状況である場合がある．こういった背景から，我々は，現実世界における高精度かつロバストな質問応答の実現を目指して [1, 2]，NTCIR-11[1] において QA Lab タスクを提案した [3]．

QA Lab では，現実世界における質問応答への第一歩として，大学入試問題を解くことを目的としている．大学入試問題には上で述べたような，質問の背景を説明する記述や，「下線部の政策を行った人物」といった抽象的な記述を含む質問が多くあり，大学入試問題を対象とすることで現在の質問応答技術の精度と課題を明らかにできると考えられる．NTCIR-11 では，対象科目を世界史に限定し，センター試験（CT），および，東京大学，京都大学，北海道大学，早稲田大学教育学部，中央大学文学部の二次試験（SE）を対象とした[2]．センター試験は，回答方式こそ多肢選択だが，質問形式は正誤判断問題や穴埋め問題など多岐にわたり，二次試験では，これらに論述問題が加わる．全ての問題を英訳し，日本語と英語のどちらでも参加できる環境を整えた結果，国内海外合わせて 9 チームが参加した．本稿では，QA Lab に参加した各チームの結果をもとに，大学入試問題を対象とした質問応答技術の現状と課題を明確化し，現実世界での高度な質問応答を実現するための道筋を示す．

2 QA Lab の参加チーム

QA Lab は，2007 年センター試験，2003 年センター試験，2007 年二次試験の順にサブタスクを行っており，表 1 に示す 9 チーム[3] が参加した．各チームは複数の結果を提出することができ，表中の数字は提出した結果の数を示している．また，QA Lab では，各チームのシステムが単独で回答する End-to-End Run と，End-to-End の回答を組み合わせることで最終的な回答をする Combination Run の 2 種類を行っており．表中の括弧左の数字が End-to-End Run の提出数，括弧内の数字が Combination Run の提出数を示している．二次試験に関しては，論述問題を含むこと，センター試験と質問の表現が大きく異なることなどから，2 チームからしか結果が提出されなかった．そのため，二次試験の Combination Run は行わなかった．また，本稿の分析もセンター試験の結果を中心に行うこととする．

表 1: QA Lab の参加チームと提出数

チーム名	言語	2007 CT	2003 CT	2007 SE
CMUQA	英	3(2)	3(2)	-
DCUMT	日	1(0)	1(0)	1
FLL	日	3(0)	3(0)	-
Forst	日	1(0)	1(0)	1
FRDC_QA	英	1(0)	1(0)	-
KJP	日	1(0)	1(0)	-
nnlp	日	1(0)	1(0)	-
NUL	日	-(-)	2(0)	-
sJanta	英	-(-)	1(0)	-

3 大学入試問題の特徴

図 1 にセンター試験の例を，図 2 に二次試験の例をそれぞれ示す．どちらの例においても，核となる質問文の他に質問背景などを説明する記述が書かれていることが分かる．しかしながら，その書かれ方は異なっており，センター試験の例では大問の質問文や説明文のように構造化されているのに対し，二次試験の例では背景の記述から核となる質問文までひとまとまりで書かれている．このような差異も上手く処理できる能力が QA Lab では求められた．

表 2 に，我々が設定した，世界史の大学入試問題における質問形式の一覧と，その形式の質問がセンター試験および二次試験に含まれているかどうかを示す．Factoid 型の質問では「いつ」「どこ」「誰」などの疑問詞あるいは「人物の名」「王朝の名」といった語句

[1] http://research.nii.ac.jp/ntcir/ntcir-11/

[2] 訓練用データは，2009 年，2005 年，2001 年，1997 年のセンター試験，2009 年，2005 年の二次試験であり，テストデータは，2007 年，2003 年のセンター試験，2007 年の二次試験である．

[3] NUL と sJanta の両チームは 2007 年のセンター試験には参加していない．

表 2: 質問形式の一覧と有無

質問形式	CT	SE	例
Factoid	○	○	下線部 (3) に関連して，キューバ危機が起こったときのアメリカ合衆国大統領の名として正しいものを，次の 1-4 のうちから一つ選べ．
Slot Filling	○	○	下線部 (4) の時期に大西洋で行われた交易について述べた次の文中の空欄ア-ウに入れる語の組合せとして最も適当なものを，次の 1-4 のうちから一つ選べ．
True/False	○	○	下線部 (1) の国の歴史について述べた文として誤っているものを，次の 1-4 のうちから一つ選べ．
True/False Combo	○	○	下線部 (5) に関連して，世界史上見られた広域的な労働力の流れについて述べた次の文 a と b の正誤の組合せとして正しいものを，下の 1-4 のうちから一つ選べ．
Time	○	○	下線部 (3) の国が支配した地域の 20 世紀の歴史について述べた次の文 a-c が，年代の古いものから順に正しく配列されているものを，下の 1-6 のうちから一つ選べ．
Graph	○	○	次の地図中の地域 a-d のうち，下線部 (8) で「私たちの国」と呼ばれている地域を示すものとして最も適当なものを，下の 1-4 のうちから一つ選べ．
Essay	×	○	中国近代史において日中関係は大きな比重を占めるようになる．1911 年の辛亥革命から 1937 年の日中戦争開始までの時期における，日本と中国の関係について，300 字以内で述べよ．

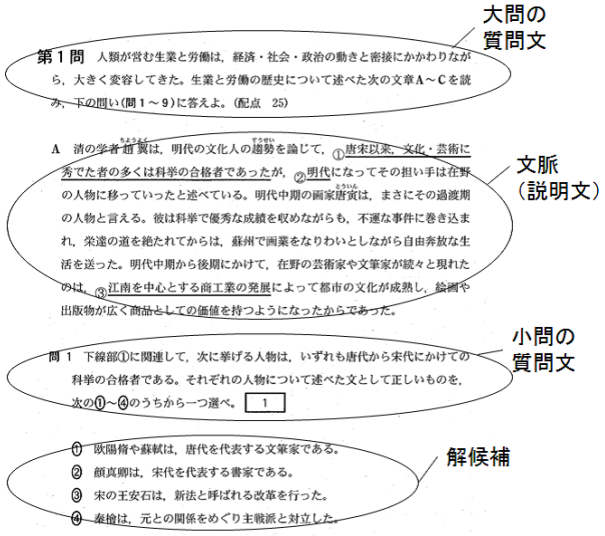

図 1: センター試験の例

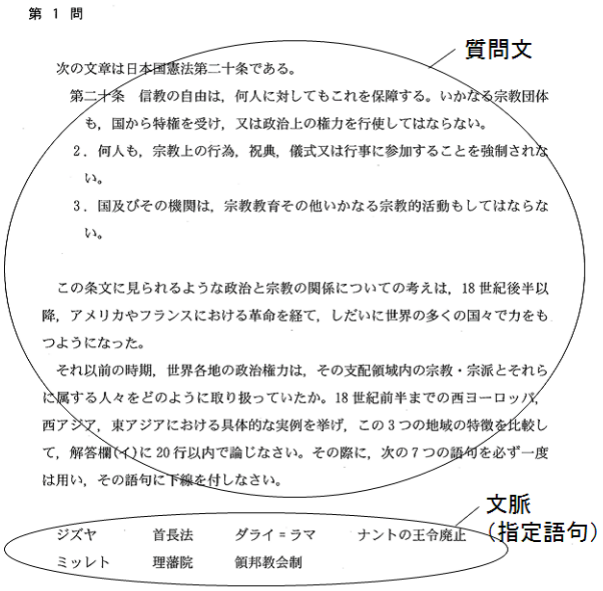

図 2: 二次試験の例

で解答のカテゴリーが明示的に示されているが，Slot Filling 型ではそのような語句は含まれていない．その代わり，Slot Filling 型では解答の語句が用いられる時の文脈が空欄の前後に明示されている．True/False 型と True/False Combo 型はどちらも命題の正誤判断を問うものであるが，True/False 型が命題集合の中から最も正しい (誤っている) と思われる命題を「相対的に」判断すればよいのに対し，True/False Combo 型は各命題の正誤を「絶対的に」判断しなくてはならない点で異なっている．Time 型は複数の出来事の間の時間的な関係を問うものであり，Graph 型は地図やグラフなどの非言語的な情報を伴うものである．Essay 型は数行または数百字にわたっての記述を要求するものであり，二次試験にしか存在しない．

4 各システムの結果

図 3 に，QA Lab で用いたセンター試験における質問形式の割合を示す．どの年でも True/False 型が最も多く 60%以上を占めている．他の質問形式に関しては，Slot Filling 型が比較的安定しているものの年によるばらつきが大きく，あって 5 問程度である．したがって，センター試験で高得点を狙うためには True/False 型に強いシステムを構築することが効果的である．しかしながら，現実世界でのロバストな質問応答を実現するという観点からは，他の質問形式に対してもきちんと回答できる必要がある．

そこで，質問形式ごとの出題数の差による影響を無視するために，質問形式 f ごとの正解率 $acc(f)$ を以下の式 (1) により求めた．

$$acc(f) = \frac{\text{質問形式 } f \text{ の正解数}}{\text{質問形式 } f \text{ の出題数}} \quad (1)$$

参加システム全て [4] における質問形式ごとの正解率を図 4 に示す．また，2007 年と 2003 年のセンター試験

[4] End-to-End Run の結果のみであり、Combination Run の結果は含んでいない．

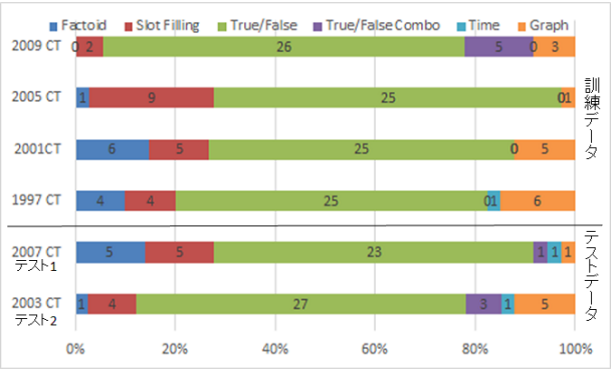

図 3: センター試験における質問形式の割合

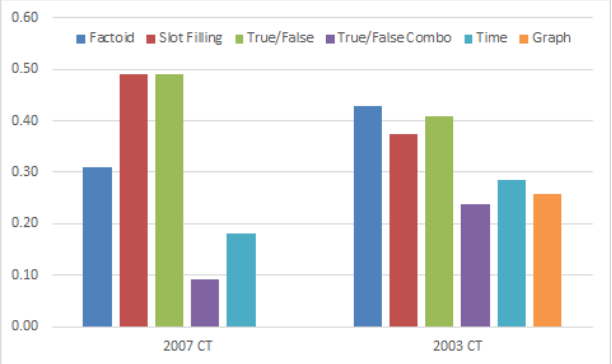

図 4: 全参加システムの質問形式ごとの正解率

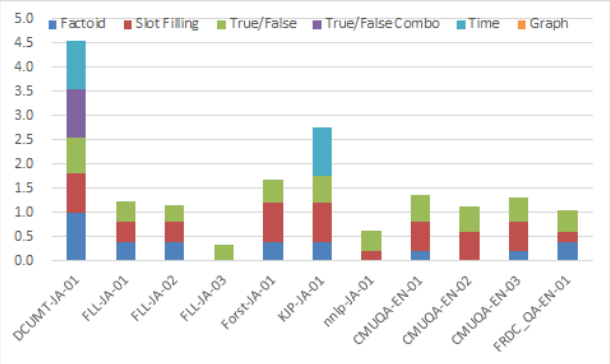

図 5: 参加システムごとの各質問形式の正解率の和（2007 CT）

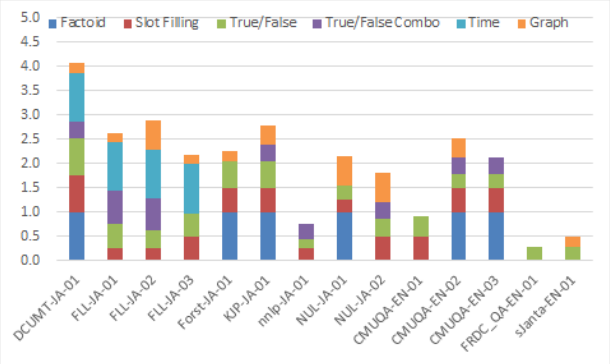

図 6: 参加システムごとの各質問形式の正解率の和（2003 CT）

における参加システムごとに各質問形式の正解率を合算したものを図 5 と図 6 にそれぞれ示す．各質問形式の正解率は 0 から 1 の範囲であり，センター試験に含まれる質問形式は 6 種類であるから，図 5 と図 6 に示したグラフの最大値は 6 となる．

図 4 から，True/False 型，Slot Filling 型が比較的高い正解率であるのに対して，True/False Combo 型，Graph 型の正解率は低かった．同じ正誤判断を問う質問であっても，「相対的判断」か「絶対的判断」かで大きく差がついたのは興味深い．図 5 と図 6 を見ると，参加システムごとの正解率の差が顕著だったのは Time 型の質問である．Time 型を解いたシステムには時系列に特化した処理が組み込まれており，例えば，FLL のシステムは，2007 年センター試験のサブタスクでは時系列処理を行っておらず，2003 年センター試験のサブタスクにおいて時系列処理を加えている．図 5 と図 6 の FLL の結果を比較するとこのことが分かる．

以上の事柄から，今後精度向上のために，質問形式ごとに特化した処理を汎用的な処理に組み込むアプローチがとられるようになる可能性がある．一般に質問応答システムは，質問文解析、文書検索、回答候補抽出、回答選択という 4 段階の処理が「水平」に統合された枠組みをもっている．質問形式ごとに特化した処理を組み込むためには，例えば，図 7 に示すような，各処理を「垂直」に統合するような枠組みが必要となるかもしれない．

5 事例分析

本節では，参加システムの多くが解けなかった質問を中心に，解くためにはどのような知識や仕組みが必要かを考察する．各質問のタイトルは，入試問題の名称_解答欄 ID，質問形式を表しており，括弧内の数字は，(正解のシステム数)/(サブタスクに参加した全システム数) を示している．

2007CT_A11 True/False（2/15）

下線部 (2) の皇帝の事績として正しいものを，次の 1-4 のうちから一つ選べ．
1. 洛陽に遷都し，内政を重視した．
2. 塩・鉄などを非売品とし，また物価の調整と安定に努めた．
3. 文字・度量衡・貨幣を統一した．
4. 郡県制を敷き，三省・六部を設けた．
（正解：2）

下線部 (2) は「漢の武帝」を指しており，正解は 2 番である．しかしながら，参加システムの幾つかは 1 番を誤って出力していた．1 番を行った皇帝は「漢の光武帝」であり，形態素レベルに分解してしまうと「漢の武帝」と混同される可能性が高い．したがって，世界史の用語に対応した固有表現抽出器が必要である．

2007CT_A15 Factoid（2/15）

下線部 (6) に関連して，チンギス＝ハンの時代以後に成立した宗教として正しいものを，次の 1-4 のうちから一つ選べ．
1. 道教　2. キリスト教　3. シク教　4. 仏教
（正解：3）

この質問を回答する上で重要な情報は「チンギス＝ハンの時代以後」の部分であり，これが「13 世紀初頭以後」を指していることを理解しなくてはならない．「建

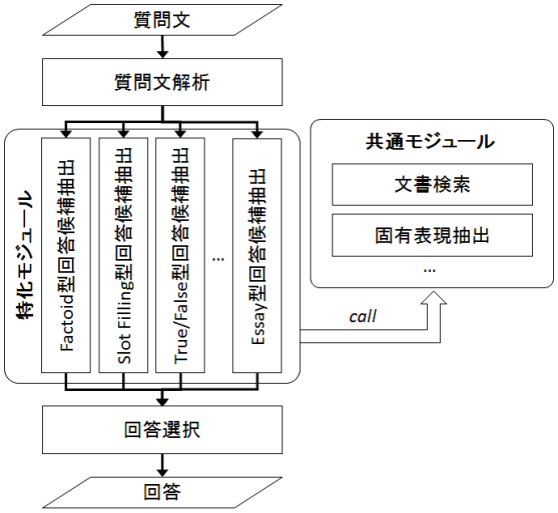

図 7: 垂直統合型質問応答システム

国」や「戦争」といったイベントだけでなく,「チンギス＝ハン」といったエンティティにおいても時間情報が必要である．また,「以後」という時間に関する接辞が重要な意味をもっており，単純な Bag of Words(BoW) では他の単語に埋もれてしまう可能性が高い．したがって，時間情報を適切に扱う必要がある．

2007CT_A12 True/False（2/15）

下線部 (3) に関連して，モンゴル人の支配下における出来事について述べた文として誤っているものを，次の 1-4 のうちから一つ選べ．
1. オゴタイ＝ハンは金を滅ぼし，カラコルムに都を置いた．
2. 元では，イスラーム世界の科学の影響で，授時暦が作成された．
3. 元の支配下では，漢人が重用され，西域出身の色目人は蔑視された．
4. 従来の大運河が補修され，また大都に至る新運河が建設された．
（正解：3）

参加システムの多くは，BoW またはそれに類する指標を用いて処理を行っていた．3 番が誤りなのは,「元の支配下では，漢人が『蔑視』され，西域出身の色目人は『重用』された」からであり，BoW では正しく内容を捉えることはできない．したがって，係り受けなどの関係を正確に把握する必要がある．

2007CT_A18 True/False Combo（1/15）

同じく下線部 (8) に関連して，第二次世界大戦後の植民地の独立について述べた次の文 a と b の正誤の組合せとして正しいものを，下の 1-4 のうちから一つ選べ．
a. イラクが独立した．
b. モザンビークが独立した．
1. a-正 b-正　2. a-正 b-誤
3. a-誤 b-正　4. a-誤 b-誤
（正解：3）

この質問では正誤の判断対象の文が非常に短く，適切な BoW を構築することが難しかったと思われる.「第二次世界大戦後にイラクが独立した」のように適切な文脈を補う必要がある．また，二つ前の質問と同じく，第二次世界大戦「後」という接辞の意味を適切に捉える必要がある．

2003CT_A30 Slot Filling（2/18）

下線部 (8) に関連して，英語の成り立ちについて述べた次の文章中の空欄 a にいれる語として正しいものを，下の 1-4 のうちから一つ選べ．
　ブリテン島には，古代以来，諸民族が繰り返し侵入し，ゲルマン系の言語である英語は，長い年月にわたり様々な言語の影響を受けてきた．現代英語の語彙を語源別に分類してみると，ギリシア語やラテン語のほか，特に a 語の影響が際立っている．a 語が大量に流入したのは，11 世紀後半に a から来た征服者が，イングランド国内に比較的強力な王権を確立した時期以降とされる．
1. ドイツ　2. フランス
3. デンマーク　4. オランダ
（正解：2）

教科書などでは「『ノルマン公国』から来た征服者」と記述される場合が多く，正答の「フランス」という表現と合致しなかったと思われる．こういった表記の違いを処理できる必要がある．

6　まとめ

本稿では，QA Lab における質問形式ごとの正解率をもとに，現在の質問応答システムが不得手とする質問形式が，True/False Combo 型，Graph 型，Time 型であることを明らかにした．これらの質問を解くためには，時間情報や地理情報などに特化した処理を質問応答システムに組み込む必要があるかもしれず，そのような枠組みとして「垂直統合型質問応答システム」の構想を述べた．また，参加システムの多くが解けなかった質問を中心に事例分析を行った結果，世界史の分野に特化した固有表現抽出や表記ゆれ処理，述語項構造や時間関係などを捉えることができる意味表現，適切な文脈への参照といった処理が精度向上のために必要であることが分かった．

謝辞

試験問題データは，株式会社ジェイシー教育研究所から「ロボットは東大にはいれるか」プロジェクトに使用を許諾されたものを，その共同研究の一環として使用した．使用データのうち，東大以外の二次試験データの XML 化，センター試験の一部と二次試験の英訳は QA Lab で行った．

参考文献

[1] 狩野芳伸, 神門典子, 質問応答システムとセンター試験解答フロー: Kachako 対応による標準化・互換化, 2013 年度人工知能学会全国大会（JSAI2013）論文集, 2013.

[2] 石下円香, 狩野芳伸, 神門典子, 質問応答システムでの解答に向けた大学入試問題の分析, 2013 年度人工知能学会全国大会（JSAI2013）論文集, 2013.

[3] H. Shibuki, K. Sakamoto, Y. Kano, T. Mitamura, M. Ishioroshi, K. Y. Itakura, D. Wang, T. Mori, N. Kando, “Overview of the NTCIR-11 QA-Lab Task”, Proceedings of the 11th NTCIR Conference, 2014.