

特集 「人工知能と歴史」—人工知能が歴史を理解する—

# 「ロボットは東大に入れるか」 歴史科目の自動解答

## Solving Examinations of History Subjects in Todai Robot Project

狩野 芳伸  
Yoshinobu Kano

静岡大学情報学部  
Faculty of Informatics, Shizuoka University.  
kano@inf.shizuoka.ac.jp, <http://kanolab.net/kano/>

川添 愛  
Ai Kawazoe

国立情報学研究所社会共有知研究センター  
Research Center for Community Knowledge, National Institute of Informatics.  
zoeai@nii.ac.jp

渋谷 英潔  
Hideyuki Shibuki

横浜国立大学環境情報研究院  
Department of Environment and Information Sciences, Yokohama National University.  
shib@forest.eis.ynu.ac.jp

藤田 彬  
Akira Fujita

国立情報学研究所社会共有知研究センター  
Research Center for Community Knowledge, National Institute of Informatics.  
a-fujita@nii.ac.jp, [http://researchmap.jp/akira\\_fujita/](http://researchmap.jp/akira_fujita/)

**Keywords:** examination solver, question answering, information extraction, natural language processing.

### 1. はじめに

「歴史」という言葉を聞いて最も身近に感じるのは、世界史や日本史といった、学校での歴史系教科ではないだろうか。学校での学習では、試験問題に解答することが求められる。「人工知能と歴史」というテーマに対しては、試験問題の自動解答を通じた、知識源としての教科書の扱い、問題文の処理、人間の解答者との比較など本稿で紹介する「ロボットは東大に入れるか」における研究が大きく関連するといえよう。

「ロボットは東大に入れるか」(以下「東ロボ」)は国立情報学研究所を中心とする長期プロジェクトである。同プロジェクトは、人工知能技術の総合的ベンチマークとして大学入学試験問題に挑戦することを通じて、自然言語処理を含む種々の知的情報処理技術の再統合および新たな課題の発見と解決を目指している。

大学入学試験問題は志願者の知的能力を客観的に測定することを目的として設計されたデータであり、通常ただ1回の試験によって、かつ、受験者間での公平性を担保しながら測定を行うために入念な検討が加えられている。この点で、入学試験問題は言語処理を含む知的情報処理技術の総合的ベンチマークとして恰好の素材であるといえる。

本稿では東ロボで挑戦している科目のうち、世界史および日本史の歴史系科目に焦点を置く。歴史科目は暗記

すれば解けるであろうから、計算機による自動解答は容易だと思われがちだが、必ずしもそうではない。例えば NTCIR-10 RITE2 タスク [Watanabe 12] では、センター試験社会科の問題を素材に含意関係認識のタスクとして精度を競った。含意関係認識タスクとは、例えば「12世紀に日本では鎌倉幕府が開かれた」という文が「鎌倉幕府は1192年に始まったとされていたが、現在では実質的な成立は1185年とする説が支配的だ」という文に含意されるかを自動判定するタスクであるが、4択としての試験成績評価で最良の結果は正答率34.26%と、多くが四択問題で25%がベースラインであることを考えると、それほど良い結果は得られていなかった。

その大部分が選択式問題からなるセンター試験形式のテストは、ごく単純な表層的な手掛かりのみでは正解できないように設計されていると考えられ、現在70~90%の精度に留まっている種々の言語処理技術をより信頼性高く頑健なものへと導くためのガイドラインとして好適である。

また、国立大学の二次試験に代表される記述式試験問題は、論述問題に象徴されるように、要求される情報に適切にアクセスする技術に加えて、それらの情報を高度な知的判断のもとで適切な表現にまとめる技術が求められる。このような点も言語処理技術の総合的なベンチマークとして適している。

さらに、模試・入試によるシステムの性能測定結果は人間の受験生の正答率や誤りの傾向と直接比較すること

```

<exam source="National Center For University Entrance Examination" subject="SekaishiB(main
exam)" year="2009"><title>2009年度本試験 世界史 B</title>
<question id="Q1" minimal="no">
  <label>【1】</label>
  <instruction> 人類が営む生業と労働は、経済・社会・政治の動きと密接にかかわりながら、大
  きく変容してきた。生業と労働の歴史について述べた次の文章A～Cを読み、以下の問い(問1～
  9)に答えよ。(配点25) </instruction>
  <data id="D0" type="text">
    <label>A</label>清の学者趙翼は、明代の文化人の趨勢を論じて、<uText
  id="U1"><label>(1)</label>唐宋以来、文化・芸術に秀でた者の多くは科挙の合格者であった
  </uText>が、<uText id="U2"><label>(2)</label>明代</uText>になってその担い手は在野の人物に
  移っていったと述べている。明代中期の画家唐寅は、まさにその過渡期の人物と言える。彼は科
  挙で優秀な成績を収めながらも、不運な事件に巻き込まれ、栄達の道を絶たれてからは、蘇州で
  画業をなりわいとしながら自由奔放な生活を送った。明代中期から後期にかけて、在野の芸術家
  や文筆家が続々と現れたのは、<uText id="U3"><label>(3)</label>江南を中心とする商工業の発
  展</uText>によって都市の文化が成熟し、絵画や出版物が広く商品としての価値を持つようにな
  ったからであった。</data>
    <question id="Q2" anscol="A1" answer_style="multipleChoice" answer_type="sentence"
  knowledge_type="KS" minimal="yes">
      <label>問1</label>
      <instruction> 下線部<ref comment="" target="U1">(1)</ref>に関連して、次に挙げる人物は、
      いずれも唐代から宋代にかけての科挙の合格者である。それぞれの人物について述べた文として
      正しいものを、次の①～④のうちから一つ選べ。</instruction>
      <ansColumn id="A1">1</ansColumn>
      <choices anscol="A1" comment="">
        <choice>
          <cNum>①</cNum> 欧陽脩や蘇軾は、唐代を代表する文筆家である。 </choice>
        <choice>
          <cNum>②</cNum> 顔真卿は、宋代を代表する書家である。 </choice>
        <choice comment="1" ra="yes">
          <cNum>③</cNum> 宋の王安石は、新法と呼ばれる改革を行った。 </choice>
        <choice>
          <cNum>④</cNum> 秦檜は、元との関係をめぐり主戦派と対立した。 </choice>
      </choices>
    </question>
  
```

図1 2009年度センター試験世界史BのXMLデータ

が可能である。センター試験は毎年約50万人が受験し、予備校によるセンター試験模試も数千～数万人規模の参加者を集める。このような大規模なサンプルから得られた「普通の人」、「典型的な人」の像とシステムとの比較は、人によるアノテーションに基づく通常の性能測定とは異なった評価指標となっている。

プロジェクトでは、2013年度、2014年度、2015年度の3回にわたり代々木ゼミナールおよびベネッセ主催の全国センター模擬試験(以下「センター模試」)を用いた各科目の解答システムの評価を行い、その結果を公表した。2013年度の各科目の結果については[新井14]を参照されたい。

また、2015年度には駿台予備学校が実施する東京大学二次試験の模擬試験(東京大学実戦模試)の世界史問題を対象として、記述式の試験の解答を行った。

本稿では、自動解答のためにプロジェクトで整備したリソース類(2章)、センター試験の自動解答(3章)、記述式試験の自動解答(4章)、高校生の解答行動(5章)の研究をそれぞれ紹介し、最後に全体をまとめる(6章)。

## 2. 東ロボ歴史科目のリソース

東ロボプロジェクトにおいて、解答器への入力となるのはXML形式の試験問題データである。データの例を

図1に示す。データ中では、問題(question)、指示文(instruction)、問題を解く際に参考にする資料(data)、選択肢群(choices)などの文書構造、下線(uText)やラベル(label)などのテキスト装飾、またラベルや空欄に対するクロスリファレンス(ref、自然言語の参照表現による指示関係以外のもの)をマークアップしている。これは、問題を解くうえで文字認識や文書構造解析などのプロセスを省略し、言語理解に基づく解答に焦点を当てるためである。

歴史科目の解答に際して利用可能なデータとして、教科書その他のリソースが国立情報学研究所から提供されている。教科書データとしては、東京書籍株式会社および山川出版社の世界史B、日本史Bの教科書に、章・節・段落・見出しなどの文書構造をXMLでアノテーションしたものを利用している。このほか、山川出版社の『世界史B用語集』、『日本史B用語集』やWikipediaなどの参照も許可されている。

さらに、世界史科目に関しては、国立情報学研究所が開発・配布しているイベントオントロジーEVT[Kawazoe13]が利用可能である\*1。このオントロジーは、「出来事(イベント)の成立を記述する文」の真偽判断をサポートする目的で作成されたもので、世界史問題の

\*1 <http://researchmap.jp/zoelai/event-ontology-EVT/>

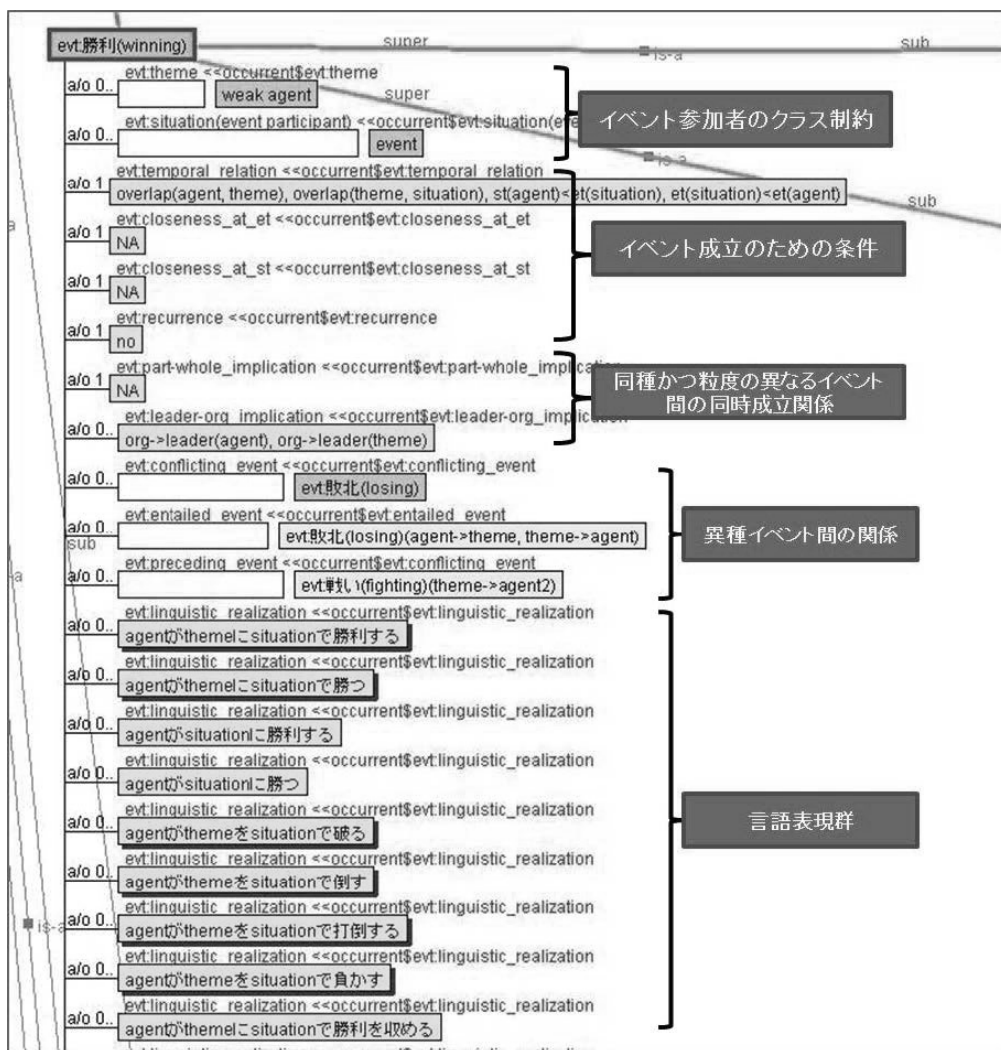


図2 EVTにおけるイベント概念の記述例

選択肢として典型的に現れる「(史実かどうかわからない) 歴史的出来事を記述する文」の正誤判断に利用することができる。上位オントロジーとしてYAMATO [溝口12]をベースとし、内容語、特に動詞によって表されるイベントの成立・不成立の判断に関わる知識・情報を形式的に記述している。イベント概念の記述例(図2)に示されるとおり、そのような知識には以下の5種が含まれる。

1. イベント参加者のクラス制約
  2. イベントが成立するための必要条件: 参加者が満たしているべき時間的・場所的条件, 回数に関する制約など
  3. 同じ種類のイベント間で, 参加者の粒度が異なる場合(例えば北フランス vs. パリ)の同時成立関係
  4. 異なる種類のイベント間の関係(「勝利」と「敗北」との非両立関係, 「放棄」に対する「所有」の先行関係など)
  5. イベントを記述する言語表現群(主に動詞的表現+句構造)
- 現時点では主に高校世界史教科書に現れる720のイベ

ント概念, 2000超の言語表現をカバーしている。この言語表現数は, 1990~2011年度の大学入試センター試験世界史科目の正誤問題の選択肢に現れる動詞出現数の8割程度をカバーしている。また, EVTに対応するインスタンスデータでは, 国家, 人物, 場所, 社会制度などを含む20のカテゴリーで4700個のインスタンスを記述している。

### 3. センター試験歴史科目の自動解答

センター試験歴史科目の問題は, 入力 of 正誤を判定する問題として一般化できる。具体的には, 図1にあるような「四つの選択肢から正しいもの(あるいは誤ったもの)を選べ」というタイプの問題が過半を占め, ほかに「正誤の組合せのうち正しいものを選べ」というような正誤の判定を行うタイプの問題や, 年代順が正しいものを答えよというタイプの問題がある。なお, ここでは画像の認識が必要な問題は対象にしない。

センター試験歴史科目の問題は, 東ロボ模試タスクだけでなく, 東ロボプロジェクトとの連携により

NTCIR RITE タスクシリーズ [Matsuyoshi 14, Shima 11, Watanabe 13] や, NTCIR QALab タスクシリーズ [Shibuki 14, Shibuki 16] でも自動解答を試みるべく用いられた. 本稿では東ロボ歴史科目の模試チャレンジで高得点を上げたシステム [狩野 14, Kobayashi 16] に加え, これらの関連タスクも含めた全体を概説する.

### 3.1 知識源の利用

正誤を判定する問題は, 質問応答や検索のタスクに近いものがあるが, 誤りを検出できなければならないという点で大きく異なる.

正誤の判定を行うシステムとして, 機械学習による手法が考えられる. 高性能な質問応答システムの多くは, 内部で機械学習を用いている. しかし, 機械学習により高性能を発揮するためには十分な学習データが必要である. センター試験においては, 基本的に過去に出題された問題と同じ問題は出題されないから, 過去の問題を教師として学習するのは現実的ではない. 知識源との比較による, 情報抽出的な正誤の判定が必要と考えられる.

知識源として利用可能な文書には, 教科書データに加え Wikipedia があげられる. Wikipedia における歴史関連事項の記述は網羅性が高いが, 記述の粒度がまちまちである. 教科書データのほうが記述の粒度や表現がそろっているが, それでも常に一問が一文に対応するわけではない. 「誤った選択肢」の場合はさらに問題の解決は複雑であり, 単に知識源の量を増やせば得点が向上するとは限らない.

### 3.2 キーワードの抽出

どのシステムでも, キーワードの抽出が必要であることは共通している. 単語分割という点では, 辞書の充実が重要であり, Wikipedia の見出し語や用語集のデータを導入することでかなりの網羅性が得られる.

加えて, キーワードのカテゴリーも正答率の向上に貢献し得る. 世界史については, 前章で述べた世界史オントロジーがカテゴリーとそのインスタンスを記述している. 世界史オントロジーでは関係性をも記述しているが, 現在の解答器システムではその記述を十分に生かしているものではなく, 単語とカテゴリーの利用に留まっている.

### 3.3 スコアリングと解答選択

多くの解答器は, 問題文と知識源の間のある種の類似度を, なにがしかの基準でスコアとして算出し最も高いスコアの選択肢を解答としている. スコアリングにおいては, 研究グループによってさまざまな手法が試みられている. 係り受け, 論理関係, 格解析などの手法は今のところ正答率の向上には寄与しておらず, 知識源における単語の分布を用いた手法が最も正答率が高い.

表 1 に模試チャレンジ採点結果の各年度最高点と偏差値を示す. 表には含まれていないが, 過去の問題の採点

表 1 各年度の東ロボ模試チャレンジにおけるセンター試験の最高点 (括弧内は偏差値)

科目	2013年度	2014年度	2015年度
世界史 B	58 (55.2)	52 (56.1)	76 (66.5)
日本史 B	56 (56.1)	44 (48.2)	55 (54.8)

結果によると, 年度による変動も大きい点に注意が必要である.

### 3.4 今後の展望

現在の解答器システムは, 模試チャレンジにおいて一定の成果を収めているが, より高度な自然言語処理という観点と, 東京大学に合格し得る点数という観点からは, 正答できていない残り数割の問題が大きな課題である.

これらの問題の解答には, 単語レベルの処理を超えたさまざまな自然言語処理技術が必要と考えられるが, 前節で述べたように係り受け, 論理関係, 格解析などより深い解析手法を取り入れた研究でも必ずしも得点の向上につながっていない. 各年度で最も高い性能を上げていたのは単語レベルの処理によるもので, 年後ごとの性能向上も単語のカテゴリー利用と分布の計算の工夫が貢献している. この状況からすると, 4 択問題に正答するには要素技術の性能がまだまだ不十分なのではないだろうか.

またエラー分析の試み [松崎 16] から, 特に正答できていない問題群に関しては複合的な要因を見て取ることができる. 単純に, 例えば係り受けの性能が向上すればよい, というものではないため, 今後の正答率向上にはより高度な要素技術の研究が必要であり, 一見小さな得点の向上を果たすにもかなりの研究期間が必要なのである.

## 4. 記述式試験歴史科目の自動解答

記述式試験歴史科目の自動解答においても, 基本的な内容はセンター試験の場合と共通している. 教科書などの有用な知識源の利用やキーワードの抽出とそれに基づくスコアリングなどである. 一方で, センター試験には存在していた選択肢が存在しないなどのいくつかの違いもある. 本稿では, センター試験における自動解答との差異に焦点を当てて述べる.

記述式試験歴史科目の問題は, 「○○を行った人物の名を記しなさい」といった語句記述問題と, 「○○について 20 行以内で論じなさい」といった論述問題に大別できる. これはファクトイド型質問応答 [磯崎 09] とクエリ指向型文書要約 [Tombros 98] に相当する大きな違いであり, 解答までの流れに違いが生じる. 以下, 語句記述問題と論述問題に分けて概説する.

### 4.1 語句記述問題

語句記述問題の自動解答の流れは, 一般的なファクト

表2 2015年度の東ロボ模試チャレンジにおける記述式試験（世界史B）の点数（括弧内は偏差値）

大問1 大論述	大問2 小論述	大問3 語句記述	全体
9/26 (61.8)	4/24 (44.4)	8/10 (57.0)	21/60 (54.1)

イド型質問応答と同じである。すなわち、質問文解析、文書検索、解候補抽出、解候補選択の流れである。

センター試験における自動解答と比較した場合、解答タイプの重要性が大きい。センター試験の場合、選択肢の解答タイプは同一であることが多く、例えば、選択肢として「ロンドン」、「パリ」、「ニューヨーク」が示されていれば、仮に解答タイプが〈地名〉と判断できなくとも解答への影響は小さい。一方、語句記述問題ではこのような選択肢が示されないため、解答タイプを正しく認識しなければならない。また、解答タイプには固有表現抽出の結果が用いられるが、歴史科目を対象とするには一般的な固有表現抽出の分類では不十分であることが報告されている[石下13]。精度を上げるためには、〈宗派〉や〈部族〉といった歴史問題特有の解答タイプに対応する必要がある。

表2は2015年度の模試チャレンジにおける記述式試験の採点結果と偏差値を示す。センター試験と異なり受験した科目は世界史Bだけである。三つの大問のうち、語句記述問題に相当するのは大問3である。表2に示すように、歴史問題特有の解答タイプに対応することで、語句記述問題は10点満点中の8点（偏差値57.0）を達成しており比較的好成績を収めている。

#### 4.2 論述問題

論述問題の自動解答には、質問応答だけではなく自動要約の技術も必要である。論述問題の解答が知識源にそのまま記述されていることはないため、知識源中にばらばらに存在するテキストをつなぎ合わせて解答を生成する必要がある。それゆえ、知識源から解答の一部となるテキストを抽出する処理と、それらのテキストを適切に加工して文章を生成する処理との複合となる。結果として、教科書などの知識源を要約対象としたクエリ志向型の複数文書要約に近い処理といえる。

論述問題には20行（600字）程度の解答を要求する大論述問題と、2～3行（100字以内）の解答を要求する小論述問題がある。両者には、抽出すべきテキストの数や種類、解答を生成する際の要約率といった点でいくつか違いがあるが、処理の流れは基本的には共通である。質問文解析、文書検索、重要文抽出、文圧縮、時間順整列、解候補選択という流れである。

大論述問題の場合、異なる時代や地域のテキストを抽出する必要があるため、複数のクエリを用いた文書検索を行う。クエリには、解答に含めるよう指定された語句を用い、質問文中のキーワードにより重要文を抽出する。

一般に重要文の合計字数は問題で指定される制限字数を超えるため、不要な節を削除するといった文圧縮を行う。重要文を時間順に並べ替えて解候補を作成し、全体的な内容の網羅性や制限字数を考慮して解候補を選択する。

論述問題への取組みは、センター試験や語句記述問題と比較して日が浅く、既存の技術をそのまま応用したという感が強い。表2に示すように、偏差値としては語句記述問題と同程度の結果であるが、素点で見るとまだまだ改善すべき点は多いと思われる。

#### 5. 高校生の記述式問題での解答行動

記述式問題での解答行動は、解答に必要な情報を得た後の要約過程における認知的な負荷の大きさの別に、1) 情報源から重要箇所を抽出する手法（重要箇所抽出）、2) 情報源内の概念の重みを考慮して要約する手法（重み付け要約）、3) 抽象的な概念について特定のクエリに沿ってまとめる手法（query-biased summarization）[Tombros 98]、の3種に分類できる。このうち、重要箇所抽出と重み付け要約は、一定水準の性能をもった自動手法が実現されている。一方、query-biased summarization を行う技能は、自動化が実現されておらず、人間がもつ技能として特にユニークであり得る。この種の高度な要約を行う技能の育成は、学校教育においても目標とされてきた。

しかしながら、著者らの行った調査によると、実際はquery-biased summarization を完遂する人間は多くないと見積られる。入試偏差値が60を超える高校の3年生70名に、日本史の教科書内を検索する機能をもった端末を用いてquery-biased summarization の問題「8世紀から10世紀前半に、日本の政府が動員する軍事力の構成や性格はどのように変化したか」に解答する課題を提示した。この課題については、大学受験予備校における日本史論述問題の評価指針に沿って定められた、答案に意味内容として含まれるべき事柄（評価項）が9項ある。専門家が、評価項に書かれた事柄が答案に記述されているかをそれぞれ判断したところ、1項以上の評価項を解答に含めることができた高校生は約20%であり、最も多いケースでも解答に含まれる評価項は4項であった。

学校教育では、query-biased summarization の技能を習得する足場掛けとして、重要箇所抽出、重み付け要約の手法が教材に用いられてきた。現在の高校生がどの程度の正確性をもってこれらの要約操作を行うか、また正確な要約が行われるときどのような方策がとられる傾向にあるか、調査を行った。

#### 5.1 調査設定

高校3年生（2014年度）303名が調査に参加した。対象者らが高校を受験した2011年度の各校の入試偏差

値は、60～70以上であった。

調査対象者は資料（東京書籍の日本史教科書『日本史B』）を検索しながら解答を編集する機能をもつインタフェースを用いて問題に解答する。資料は、歴史上特筆されるイベントや時代を特徴付ける社会背景を取り扱うトピックという単位の文書に分かれる。インタフェースでは、トピックを検索対象に完全一致検索ができる（AND検索可）。検索を実行すると、検索語を含むトピックへのハイパーリンクとトピック内で検索語がヒットした付近の文字列の組合せのリスト（スニペット）が表示される。ハイパーリンクをクリックするとトピックが表示される。表示されたトピックからは文字列選択およびコピーが可能である。インタフェースでのキー操作およびマウス操作は、時刻とともに記録される（ログ）。出題した問題は下記のとおりである。

- 問1：平安時代の政治体制では、藤原詮子や藤原彰子など、天皇の生母に当たる女性が重要な役割を占めていた。その理由を、当時の政治体制の在り方とともに、110字から130字でまとめよ。（解答時間：10分）
- 問2：鎌倉時代、幕府により諸国に地頭が置かれたが、この地頭がどのような名目で設置され、実際にはどのような職務を担っていたか。またその後、承久の乱を経てどのように変化していったのかを、140字から160字でまとめよ。（解答時間：15分）

問1は、教科書上のある一つのトピックを探し出し、連続した箇所を抜き出すことで、過不足なく答えられる（重要語抽出）。問2は、二つの離れた箇所に記述されたトピックを探し出し、それぞれからある連続した箇所を1か所ずつ抜き出し、内容を要約することで、過不足なく答えられる（重み付け要約）。

問1の評価項は2項、問2は7項ある。評価項が答案に含意されるか否かを専門家が判定した。1件の答案を、異なる2名の専門家が評価した。結果が異なる際は合議のうえで判定を確定した。

## 5.2 結果・考察

「ログ」および「答案に含意される評価項の数（含意項数）」の間の関係を問題別に分析する。解答中に使用端末にトラブルがあった場合、当該調査対象者のログおよび答案を無効とした。ログ、答案ともに有効であった対象者は、問1で291名、問2で256名であった。問題別の含意項数の分布を表3に示す。

各問には評価項と別に、問1に二例、問2に一例の模

表3 含意項数の分布

	0項	1項	2項	3項	4項	5項	6項	7項
問1	55	141	95	-	-	-	-	-
問2	7	34	39	75	78	22	1	0

範解答が用意されている。この模範解答は、評価基準を設定した専門家が、教科書の記述内容に沿って、教科書での出現順どりに評価項全項が記述される答案を作成したものである。この模範解答と各答案の間で文字レベルでの差を測る編集距離 Levenshtein distance を計算し、群間で差を検定した。Wilcoxon rank sum test を実施したところ、問1、問2とも高得点群、低得点群の間で編集距離に有意な差が認められた ( $p < 0.01$ )。模範解答と編集距離に近いものほど含意項数が多く、高得点群の答案は低得点群の答案に比べ、使用語および記述順序が教科書の記述に類似する傾向があることがわかる。

この原因を明らかにするため、解答に必要な情報が記載された正解トピックを表示した後の調査対象者の行動を分析する。1回目に正解トピックを表示した直後に行った行動を a) 正解部抜出、b) 文字入力、c) それ以外の3種に、人手で分類した。高得点群と低得点群をさらに正解トピック表示後の方策によって群に分けたうえで、Fisher's exact test を実施したところ、問1、問2の両問について、「正解部抜出の方策を取るか否か」と「どちらの得点群に属するか」という二つの要因の独立性が成立しないことが認められた（両側検定、 $p < 0.05$ ）。また、問2について「文字入力の方策を取るか否か」と「どちらの得点群に属するか」という二つの要因の独立性が成立しないことが認められた（両側検定、 $p < 0.05$ ）。

概して正解部抜出を行った者のほうが含意項数が多く、解答上重要な情報を認識したうえで自らの言葉で記述することを試みた者のほうが含意項数が少ない傾向にあることがわかる。情報検索ツールを直接的に利用して問題解決を図るほうが正確であるように報酬系が働き、そのとおりに行動した者が十分な解答を得たと考えられる。Query-biased summarization は読み取った情報を再構成する能力がメリットとして働くタスクであるが、その技能取得の足場掛けとされる重要箇所抽出、重み付け要約の課題を解決する際に、読み取った情報を再構成する方策を取り、かつ適切な答えをつくる高校生は少ないことがわかった。重要箇所抽出、重み付け要約を query-biased summarization の技能習得の足場掛けとすることの妥当性は再考の余地があると考えられる。

## 6. おわりに

本稿では、「ロボットは東大に入れるか」プロジェクトにおける大学入試歴史科目の取組みについて、リソース整備、センター試験、記述式試験、高校生の解答行動の各点から紹介した。自動解答による得点は年々向上しているが、向上するほどにより難しく、かつ本質的な問題の解決を迫られる。人間の解答行動を観察した知見も併せ、現在の技術で何ができ何が足りないかというベンチマークを提供するとともに、人間の知的処理との差異を明らかにする手掛かりとなる有用な試みであり、今後

さらに研究を発展させていきたい。

## 謝 辞

国立情報学研究所の新井紀子教授はじめ、東ロボプロジェクトメンバに深謝申し上げる。大学入試センター試験の問題および解答データの一部については、株式会社ジェイシー教育研究所が販売する「大学入試センター試験問題データベース センター Ten 2011 通常版 全教科セット」を利用した。また、教科書・用語集データをご提供いただいた東京書籍株式会社、株式会社山川出版社、および模擬試験データをご提供くださった学校法人高宮学園、ベネッセコーポレーション、駿台予備学校に感謝を申し上げます。

## ◇ 参 考 文 献 ◇

- [新井 14] 新井紀子：ロボットは東大に入れるか，イースト・プレス（2014）
- [石下 13] 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（第27回）（2013）
- [磯崎 09] 磯崎秀樹，東中竜一郎，永田昌明，加藤恒昭：質問応答システム，コロナ社（2009）
- [狩野 14] 狩野芳伸：大学入試センター試験歴史科目の自動解答，平成26年度人工知能学会全国大会論文集，Vol. 28，pp. 1-4（2014）
- [Kawazoe 13] Kawazoe, A., Miyao, Y., Matsuzaki, T., Yokono, H. and Arai, N.: World history ontology for reasoning truth/falsehood of sentences: Event classification to fill in the gaps between knowledge resources and natural language texts, *Proc. LENLS 10*, Submission 11（2013）
- [Kobayashi 16] Kobayashi, M., Miyashita, H., Ishii, A. and Hoshino, C.: NUL System at QA Lab-2 task, *NTCIR-12 Workshop*, pp. 413-420, Tokyo, Japan（2016）
- [Matsuyoshi 14] Matsuyoshi, S., Miyao, Y., Shibata, T., Lin, C.-J., Shih, C.-W., Watanabe, Y. and Mitamura, T.: Overview of the NTCIR-11 recognizing inference in TExt and validation (RITE-VAL) task, *11th NTCIR (NII Testbeds and Community for Information Access Research) Workshop*, pp. 223-232, *inproceedings*（2014）
- [松崎 16] 松崎拓也，横野光，宮尾祐介，川添愛，狩野芳伸，加納隼人，佐藤理史，東中竜一郎，杉山弘晃，磯崎秀樹，菊井玄一郎，堂坂浩二，平博順，南泰浩，新井紀子：「ロボットは東大に入れるか」プロジェクト：代ゼミセンター模試タスクにおけるエラーの分析，自然言語処理，Vol. 23, No. 1, pp. 119-159（2016）
- [溝口 12] 溝口理一郎：オントロジー工学の理論と実践，オーム社（2012）
- [Shibuki 14] Shibuki, H., Sakamoto, K., Kano, Y., Mitamura, T., Ishioroshi, M., Itakura, K. Y., Wang, D., Mori, T. and Kando, N.: Overview of the NTCIR-11 QA-Lab task, *NTCIR-11 Workshop*, pp. 518-529, Tokyo, Japan（2014）
- [Shibuki 16] Shibuki, H., Sakamoto, K., Ishioroshi, M., Fujita, A., Kano, Y., Mitamura, T., Mori, T. and Kando, N.: Overview of the NTCIR-12 QA Lab-2 task, *NTCIR-12 Workshop*, pp. 392-408, Tokyo, Japan（2016）

- [Shima 11] Shima, H., Kanayama, H., Lee, C., Lin, C., Mitamura, T., Miyao, Y., Shi, S. and Takeda, K.: Overview of NTCIR-9 RITE: Recognizing inference in TExt, *NTCIR-9 Workshop*, pp. 291-301（2011）
- [Tombros 98] Tombros, A. and Sanderson, M.: Advantages of query biased summaries in information retrieval, *Proc. 21st Annual Int. ACM SIGIR Conf. on Research and Development in Information Retrieval*, pp. 2-10, New York, USA（1998）
- [Watanabe 13] Watanabe, Y., Miyao, Y., Mizuno, J., Shibata, T., Kanayama, H., Lee, C.-W., Lin, C.-J., Shi, S., Mitamura, T., Kando, N., Shima, H. and Takeda, K.: Overview of the recognizing inference in text (RITE-2) at NTCIR-10, *NTCIR-10 Workshop*, pp. 385-404, Tokyo, Japan（2013）

2016年9月25日 受理

## 著 者 紹 介



狩野 芳伸（正会員）

静岡大学情報学部行動情報学科准教授。博士（情報理工学）。2001年東京大学理学部物理学科卒業，2007年同大学院情報理工学系研究科博士課程単位取得退学。東京大学大学院情報理工学系研究科特任研究員，科学技術振興機構さきがけ研究者などを経て，2014年より現職。質問応答，電子カルテの処理，論文のテキストマイニングなど自然言語処理技術の応用と，より自然な対話システムの構築に興味をもつ。



川添 愛（正会員）

国立情報学研究所社会共有知研究センター特任准教授。博士（文学）。1996年九州大学文学部文学科卒業（言語学専攻）。2005年同大学院文学研究科博士課程修了。2002年より国立情報学研究所にて研究員として自然言語処理，知識表現の研究に従事。津田塾大学女性研究者支援センター特任准教授を経て，2012年より現職。



渋木 英潔

横浜国立大学環境情報研究院科学研究費研究員。博士（工学，経営学）。1997年小樽商科大学商学部商業教員養成課程卒業。2002年北海道大学大学院工学研究科博士後期課程修了。2006年北海学園大学大学院経営学研究科博士後期課程修了。自然言語処理に関する研究に従事。



藤田 彬

国立情報学研究所社会共有知研究センター特任助教。博士（情報学）。2008年横浜国立大学教育人間科学部卒業，2012年同大学院環境情報学府博士後期課程修了。横浜国立大学産学連携研究員，国立情報学研究所コンテンツ科学研究系特任研究員を経て，2015年より現職。読解の認知プロセスと，小論文・作文の自動評価など自然言語処理技術の応用に興味をもつ。