大学入試「世界史」論述問題解答システムの開発

高田 拓真　　松崎 拓也　　佐藤 理史
名古屋大学大学院 工学研究科 電子情報システム専攻

1　はじめに

現在，質問応答技術に関する研究の一つとして大学入試問題の自動解答への取り組みが進められている [1][2]．大学入試における問題の多様性，様々な知識の利用・応用の要求といった観点から，これらの問題に挑むことで，現在の質問応答技術の達成度を測定し，課題を明瞭化することができる．

本研究は，大学入試問題における世界史論述問題のうち，特に小論述と呼ばれるタイプの問題に対する自動解答を目標とする．世界史論述問題には大きく分けて数百字で答える大論述問題と，数十字で答える小論述問題がある．大論述問題では，解答に含めることを要求される指定語句が 6〜8 程度与えられることが多く，これらは解答の自動生成における大きな手がかりになる．一方，小論述問題では指定語句が与えられない場合が多い．そのため，小論述問題に対する自動解答では，問題文を解析し解答すべき内容を定め，さらに，それを短い字数にまとめることが必要となる．

また，世界史小論述問題は，従来の質問応答で着目されてきたものとは異なるタイプの質問を多く含む．具体的には，従来は「初代江戸幕府征夷大将軍は誰？」といった factoid 型の質問や，「日本が開国したのはなぜですか？」といった why 型の質問などが主な研究課題であった．しかし，世界史小論述問題では例えば「ポリスの形成過程を 60 字以内で答えなさい」といったようなある物事の過程を答える問題や，「北イタリアに結成された都市同盟について 60 字以内で答えなさい」といった主題のどの側面について解答すべきかが表層的には明らかでない問題が出題される．

本稿では，これらの新しい課題への取り組みの第一歩として，教科書を元テキストとする抽出型の要約によって解答を生成する世界史小論述問題自動解答システムを作成し，入試過去問および模試問題によって評価・分析を行った結果を報告する．

システム出力の評価に関しては，ROUGE を用いた自動評価に加え，「ロボットは東大に入れるか」プロジェクトの一環として予備校講師による人手での評価および講評を受ける機会を得た．本稿では，予備校講師による講評で指摘された問題点に対する技術的な検討についても述べる．

キリスト教徒がローマ皇帝に迫害された理由を60字以内で説明しなさい．（2013年度東大）

5世紀におけるフン族の最盛期とその後について，60字以内で説明しなさい．（2012年度東大）

図 1: 世界史小論述問題例

2　世界史小論述自動解答システム

実際の世界史小論述問題例を図 1 に示す．世界史論述問題を観察した結果，知識源となる世界史教科書の記述を抽出しつなぎ合わせることで解答ができる問題が大多数であることがわかった．そこで，問題文との類似度に基づいて教科書の文を抽出し配列することで解答とするシステムを作成した．類似度スコアとしては，服部らが含意関係認識のために提案した表層的類似度に基づくスコア [3] を改良したものを用いた．無論，異なる質問タイプが混在する世界史小論述問題に対し，表層類似度に基づく単一の要約手法のみで対応できるとは考えられない．本稿での狙いは，具体的なシステムの出力の評価と分析を通し，問題文のより深い分析に基づく解答作成へ向けた指針を得ることにある．以下では，作成したシステムの詳細を記述する．

2.1　システム概要

図 2 にシステムブロック図を示す．システムは検索，時間情報付与，圧縮，論述作成の 4 ブロックから成り立つ．本稿の実験では，解答を構成する文の抽出元として表 1 の 4 つの教科書を同時に用いた．また，世界史知識源として世界史イベントオントロジー (EVT) [4] および山川出版の世界史用語集を用いた．

2.1.1　スコアによる検索

検索部では入力の問題文と抽出元である教科書との表層類似度に基づき，教科書の各文に対しスコアリング，ソートしたものを出力とする．以下，検索部での出力を検索文と呼ぶ．スコアの定義を式 (1) に示す．

$$score = \frac{\sum_{x \in \boldsymbol{N^1}} \{\min(f(x,t), f(x,q)) \cdot a(x) + 0.3 \cdot c\}}{\sum_{x \in \boldsymbol{N^1}} f(x,q)} \quad (1)$$

表 1: 抽出元とした教科書テキスト

東京書籍出版 世界史 A(2008 年発行)	162743 文字	2850 文
東京書籍出版 世界史 B(2007 年発行)	309943 文字	5329 文
東京書籍出版 新選世界史 B(2007 年発行)	15533 文字	2870 文
山川出版 詳説世界史 B(2010 年発行)	277896 文字	4644 文

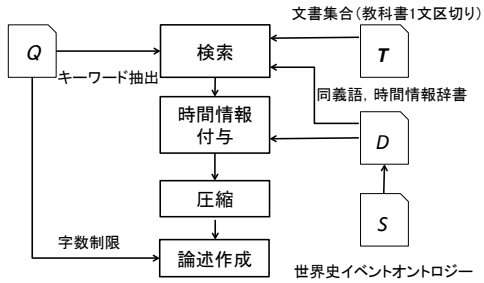

図 2: 解答システムブロック図

ここで t は教科書の一文，q は問題文であり，$f(x,t)$ は集合$\boldsymbol{N^1}$の要素 x が，t 中に出現する回数を表す．本稿における実験では，$\boldsymbol{N^1}$として名詞 1-gram を用いた．また，$a(x)$ は以下に示す重み付けの関数である．

$$a(x) = \begin{cases} 1 & \text{if } x \in \text{世界史用語集の見出し語} \\ 0.7 & \text{else if } x \in \text{固有名詞} \\ 0.5 & \text{else if } x \in \text{wikipedia のページタイトル} \\ 0.1 & \text{otherwise} \end{cases}$$

また，c は q 中で x が何単語目に現れるかを表す変数である．これにより，重要なキーワードであることが多い問題文で後に出現する名詞ほど，スコアに大きく寄与するようにする．また，EVT から世界史用語に関する同義語辞書を作成し，スコア計算の際に同義語辞書に含まれる名詞は辞書中で定めた正規形に統一した．

2.1.2 時間情報付与

問題文で述べられている内容と異なる年代に起きた事柄を含む文は正解となることが少ない．そこで，スコア付けされた教科書の各文に対し，“1800 年-1843 年” といった時間情報を付与し，問題文の内容と年代が合致するか否か判定を行う．時間情報の付与の際には EVT から抽出したイベント・人物などの名前にそれらが生起，存在した時間を対応付けた辞書，もしくは文中にある “13 世紀”，“1745 年” といった表記から時間を特定した．問題文及び検索文の時間情報の関係については以下のように場合分けする．

合致 問題文と検索文の時間情報が合致

不明 問題文と検索文のどちらかに時間情報が付与されていない

合致せず 問題文と検索文の時間情報が合致しない

以上の判定結果を各検索文に付与したものが時間情報付与ブロックの出力となる．

2.1.3 圧縮

論述生成の際，制限字数以内に収まる検索文の中からスコアが一位のものを選択するという手法が考えられるが,「30 字以内で説明しなさい」といった制限字数が少ない問題ではスコアが非常に低いものを選んでしまうという危険性がある．スコアが低い文は上位の文に比べ正解であることが少ないと予想されるため，スコアが上位の文を制限字数以内に収めることで解決を図る．現在のシステムでは，制限字数に収まる検索文の中でスコアが一位の文に対しそのスコアを閾値と比較を行い，閾値を下回る場合，文圧縮を行うことで上記の課題を解決する．以下に手順を示す．

1. 制限字数に収まる検索文の中でスコアが一位の文に対しそのスコアを閾値と比較する
2. 閾値以下の場合，スコアが閾値以上の検索文を読点で分割．分割された各部分を別個の検索文とみなし再度スコアを計算する

本稿の実験では閾値を 0.3 とした．

2.1.4 論述作成

ここまでで作成した類似度スコアと「時間情報関係」ラベルつきの検索文集合から，以下の手順で文を選び解答となる論述を作成する．

1. 検索文のうち “合致” が付与された文をスコアが高いものから制限字数に収まるあいだ追加
2. 検索文のうち “不明” が付与された文をスコアが高いものから制限字数に収まるあいだ追加
3. 検索文のうち “合致せず” が付与された文をスコアが高いものから制限字数に収まるあいだ追加

最後に上記の手順で選択された検索文を各文に付与された時間情報に従ってソートし，解答として出力する．

3 評価実験

東京大学の過去問を用いてシステムの評価を行った結果について述べる．また,「ロボットは東大に入れるか」2015 年度の予備校模試を用いた各科目の解答システム評価の一環として，東大模試および慶応大学過去問に対するシステム出力を駿台予備校講師が評価した．この結果についてもまとめる．

3.1 評価データ・評価方法・結果

東京大学過去問の解答では 1995，1998，2000〜2013 年度の東京大学入試試験世界史のうち 30〜120 字で答える形式のもの 50 問を評価データとした．また,「ロボットは東大に入れるか」2015 年度のシステム評価用データには，小論述問題として駿台東大実戦模試 2015 年 8 月の世界史小論述問題 5 問，2015 年度慶応経済の小論述問題 4 問が含まれていた．

東京大学過去問を評価データとした実験では，システムの出力に対し，教学社「東大の世界史 25ヵ年」第 4 版の解答例を参照テキストとして ROUGE スコア

表 2: 東京大学過去問に対する ROUGE スコア

	$ROUGE_1$	$ROUGE_2$	$ROUGE_1$(内容語)
提案手法	0.342	0.080	0.146
重みなし	0.339	0.751	0.135
圧縮なし	0.342	0.080	0.144
時間判定なし	0.334	0.760	0.131

表 3: 駿台東大実戦模試の結果

	得点	配点
問 1(a)	3	6
問 1(b)	0	4
問 2	1	6
問 3(a)	0	4
問 3(b)	0	4
合計	4	24

表 4: 慶応経済 2015 の結果

	得点	配点
問 3(3)	1	1
問 4(1)	2	3
問 12(2)	0	2
問 15	6	6
合計	9	12

[5] を算出した．参照テキストおよびシステム出力の形態素解析には MeCab+UniDic を用いた．また，システムの各要素の有効性を測るため，スコア算出の際の重みなし，圧縮なし，時間判定なしの場合における ROUGE スコアを算出した．表 2 に，全単語を用いた $ROUGE_1$，$ROUGE_2$ スコア，および，内容語（名詞・動詞・形容詞・副詞）のみに対する $ROUGE_1$ スコアを示す．表からいずれの ROUGE スコアも大きな差は見られないが提案手法が最高値を出しており各要素に有効性があることがわかる．論述自動解答システムとしては直接比較可能なベースラインが存在しないため，この ROUGE スコアのみでシステム性能を把握することは難しいが，参考値として一例を挙げれば，阪本らは東京大学過去問における世界史大論述問題を対象とした質問応答システム [6] の性能として 0.22〜0.67 程度の $ROUGE_1$ スコアを報告している．

「ロボットは東大に入れるか」2015 年度のシステム評価では，システムの出力を駿台予備校講師が採点した結果を表 3，4 に示す．問題数は少ないものの慶応過去問に比べ東大の問題はより難しいという傾向が伺われる．駿台模試の受験者の平均点は 6.5 点であり，本システムの得点はそれを下回る結果になった．しかし，得点の差は 2.5 点のみであり，駿台模試の受験者に近いレベルの解答ができていると言える．

4 分析・考察

前節で示した評価結果およびシステム出力に対する予備校講師の講評を分析し，現在のシステムにおける問題点，およびその解決のめに必要な知識，手法について考察した．以下にその詳細を記述する．

4.1 記述の抽象度の違いによる名詞不一致

問題文の記述が抽象的なため問題文の名詞と解答例や教科書に使用される名詞が一致せず，教科書の適切な文のスコアが低いケースが多々あることが判明した．以下に問題例，その解答例，駿台採点基準ポイントを参考に選んだ正解となり得る教科書の文を示す．

問題文 明代の長江流域の農業・工業について，2 行以内で説明しなさい．(駿台実戦模試 問 (3)(b))

解答例 下流域で綿織物など家内制手工業や綿花などの原料栽培が広がり，中流域が穀倉地帯となり「湖広熟すれば天下足る」と称された．

教科書 長江下流域では綿織物や生糸に代表される家内制手工業がさかんになり，原料となる綿花や養蚕に必要な桑の栽培が普及した。

上記の問題は「農業」「工業」といった抽象的な名詞を使用している．これらは，解答例における「綿織物」「家内制手工業」「綿花」「原料栽培」「穀倉地帯」といった内容に対応している．また，教科書の記述にも「農業」や「工業」といった名詞は出てこず，その内容を表す名詞が現れている．そのため，本稿におけるシステムではこれらの文を検索で上位の結果として得ることができなかった．このような例は，特に東大の問題で多く見られた．そのため，この問題が駿台東大実戦模試の点数が慶應経済の点数に比べて低い原因の一つと考えられる．対策方法としては，この例では「綿織物，家内制手工業，綿花，原料栽培，穀倉地帯」が「農業」「工業」の下位語に位置するという知識を用いて検索を行う方法が考えられる．

4.2 駿台模試・慶応過去問に対する予備校講師による講評の分析

「ロボットは東大に入れるか」2015 年度のシステム評価に参加し，自動解答の結果に対して駿台予備校世界史講師からの講評を受ける機会を得た．小論述に関する講評のポイントは以下の 2 点であった．

1. 全体に共通する大前提を無視している．
2. 題意を読み取れていない．

一つ目の点は，本稿のシステムでは大問の導入部は無視し，小問の部分のみを問題文として扱い，解答を行ったことに起因する．駿台模試の問題では，大問の導入部は「世界史上の都市におけるヒトやモノの交流に関する，以下の 3 つの設問に答えなさい。」となっており「ヒトやモノの交流に関する」問題であることを読み取り，これに対応する文から解答する必要がある．

「題意を読み取れていない」ことに関して，駿台模試問 (1)(b)「イスラーム世界の都市の主要な施設について，2 行以内で説明しなさい。」の問題は一般のイスラーム世界に共通する施設について論述の対象にしており，普遍的な事例を書く必要があるのに対し，システムの出力は具体的な事例を答えていると指摘された．原因の一つに論述として何を要求されているかを把握していないことがあげられる．現在のシステムで

表 5: 問題タイプ調査結果

問題タイプ	数	問題例
理由	8	キリスト教徒がローマ皇帝に迫害された理由を 60 字以内で説明しなさい。
過程	13	日本は国際連盟に参加し，後に脱退した。脱退の経緯を 60 字以内で説明しなさい。
結果	2	1948 年 5 月に始まった第一次中東戦争（パレスティナ戦争）がある。この戦争の結果どのようなことが起こったか，60 字以内で説明しなさい。
比較	4	明から清の前期（17 世紀末まで）にかけて，対外貿易と朝貢との関係がどのように変化したかについて，海禁政策に着目しながら，120 字以内で説明しなさい.
説明 (具体的)	14	メロヴィング朝の王は，どのような宗教に改宗したのか，この王の名前とともに，60 字で説明しなさい。
説明 (抽象的)	11	ローマの平和と繁栄を示す都市生活を支えていた公共施設について，60 字以内で説明しなさい。

は「イスラーム世界の都市の主要な施設」の，どの側面について答えればよいか把握しないまま検索を行うため，適切な文の抽出ができていない．解決策としては以下で述べる問題タイプ分類が考えられる．

4.3 問題のタイプ調査・考察

「題意を読み取れていないこと」に対して，解決策として問題タイプ分類を挙げた．その準備段階として東京大学入試試験 50 問の問題文のタイプ調査を行った．ここでは，その結果と考察について述べる．

調査結果を表 5 に示す.「理由」「過程」「結果」「比較」タイプはそれぞれ問題文中の事例の「理由」「過程」「結果」「比較」について説明するよう要求される問題である．また,「説明 (具体的)」は「モンロー宣言の内容を述べよ」など，問題文で具体的な要求が示されるタイプの問題であり,「説明 (抽象的)」は具体的な要求が示されないタイプの問題である． これらのタイプのうち「理由」「過程」「結果」「比較」「説明 (具体的)」は，問題文中に物事のどの側面について記述すべきか直接明示されるため要求点の判別が容易だが,「説明 (抽象的)」はその判別が難しい．判別が容易な問題タイプの一つである「理由」の問題例を下記に示す．

問題文 キリスト教徒がローマ皇帝に迫害された理由を 60 字で説明しなさい。(2013 年度東大)

この問題では「ローマ皇帝によるキリスト教徒の迫害」の理由を述べた文を要求されていることが読み取れる．解答手法としては教科書の談話構造解析結果を用いて理由にあたる文を抽出する方法が考えられる．

一方，判別が難しい「説明 (抽象的)」には，下記のような問題がある．

問題文 北イタリアに結成された都市同盟について，60 字で説明しなさい。(2009 年度東大)

解答例 イタリア政策を進める神聖ローマ皇帝に対してミラノを中心にロンバルディア同盟が結成され，皇帝を破って自治権を確認させた。

上記の問題の解答には,「北イタリアに結成された都市同盟」の説明が求められる．参考書では同盟の結成された「理由」とその「結果」の部分が解答になっている．しかし，これらの情報を問題文上から読み取ることは困難である．これらの問題に対しては対象となる事例のタイプから解答すべき内容を決定する方法が考えられる．この例では対象は「同盟」であり,「同盟」は属性として [構成員，目的，結果] などを含む．この情報から「目的，結果を要求されることが多い」といったヒューリスティックで，求められていると予想される情報を選択，抽出することで解答が作成できる．

5 まとめ

大学入試「世界史」論述問題のうち，特に小論述に対する自動解答システムを開発，評価を行った．結果として，名詞の上位下位関係の把握，問題タイプ分類が必要なことなどがわかった．

謝辞

本研究で用いた教科書テキスト，知識源および予備校講師による評価の機会を提供していただいた NTCIR12 QALab-2 タスクオーガナイザーの皆様と「ロボットは東大に入れるか」プロジェクトに深く感謝いたします．

参考文献

[1] 新井紀子, 松崎拓也. ロボットは東大に入れるか? 人工知能学会誌, Vol. 27, No. 5, pp. 463–469, 2012.

[2] Hideyuki Shibuki, Kotaro Sakamoto, Yoshionobu Kano, Teruko Mitamura, Madoka Ishioroshi, Kelly Y Itakura, Di Wang, Tatsunori Mori, and Noriko Kando. Overview of the NTCIR-11 QA-Lab Task. In *Proc. NTCIR-11 Conference*, 2014.

[3] 服部昇平, 佐藤理史, 駒谷和範. 表層類似度に基づく日本語テキスト含意認識. 人工知能学会論文誌, Vol. 29, No. 4, pp. 416–426, 2014.

[4] Ai Kawazoe, Yusuke Miyao, Takuya Matsuzaki, Hikaru Yokono, and Noriko Arai. World history ontology for reasoning truth/falsehood of sentences: Event classification to fill in the gaps between knowledge resources and natural language texts. In *New Frontiers in Artificial Intelligence*, pp. 42–50. Springer, 2014.

[5] Chin-Yew Lin. Rouge: A package for automatic evaluation of summaries. In *Text summarization branches out: Proc. ACL-04 workshop*, Vol. 8, 2004.

[6] 阪本浩太郎, 渋木英潔, 石下円香, 森辰則, 神門典子. 大学入試の論述問題を解く質問応答システムの検討. 言語処理学会第 21 回年次大会発表論文集, pp. 180–182, 2015.