大学入試の論述問題を解く質問応答システムの検討

阪本浩太郎[†1†2]　渋木英潔[†]　石下円香[†2]　森辰則[†1]　神門典子[†2]

[†1]横浜国立大学　[†2]国立情報学研究所
E-mail: {sakamoto,shib,mori}@forest.eis.ynu.ac.jp, {ishioroshi,kando}@nii.ac.jp

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は異なる質問状況である場合がある．こういった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の例として，大学入試問題があげられる．

大学入試問題をコンピュータに解かせる試みとして，「ロボットは東大に入れるか」プロジェクト[1]やNTCIR-11[2]のQA Labタスク [1] などがある．QA Labでは，世界史の大学入試問題（センター試験および二次試験）を対象とした課題が設定されており，我々もこれに参加した．大学入試問題にはセンター試験と二次試験[3]があり，二次試験には数十字から数百字にわたって回答を記述する論述問題が含まれている．論述問題は多肢選択問題と異なる難しさがあり，上述の「ロボットは東大に入れるか」においても世界史の論述問題は対象とされていない．以上の背景から，我々は二次試験の論述問題を解くための質問応答システムを開発している．

論述問題には，回答の際に指定された語句を必ず用いなければならないものと，指定語句がないものの2種類が存在する．指定語句がある場合，文書検索などにおける手がかりが増える一方で回答を生成する際の課題も増えるため，指定語句のない論述問題と比較して必ずしも簡単であるとはいえない．例えば，指定語句は質問内容と直接関係のあるものではないことがあるため，質問内容に合った使われ方をしている記述を見つける必要がある．また，指定語句同士のつながりも希薄であるため，指定語句間のつながりを見つけて記述を補う必要がある．さらに，それらの記述はそれなりの長さをもっているため，字数制限を超えないようにしながら回答を生成しなければならない．それゆえ，本稿では，指定語句がある論述問題を中心[4]に我々の取り組みと，現実世界における高精度かつロバストな質問応答を実現するための課題を記述する．

2　関連研究

近年の複数文書要約では，要約を最大被覆問題として捉える研究が多い [3]．これは，要約対象文書中の概念単位（単語など）を，与えられた要約長を満たす文の集合によって可能な限り被覆することで要約を生成するものである．一見，指定語句のある論述問題と親和性が高いように思われるが，一般に指定語句間のつながりは非常に希薄であることに加え，どれほど他の語句との関係性がなくとも指定された語句は「必ず」用いなくてはならない，という点で最大被覆モデルによる複数文書要約手法をそのまま用いることはできない．したがって，指定語句を起点として，他の語句との希薄な関係性を明瞭にするような文の集合を求める必要がある．

3　大学入試問題

図1に指定語句がある論述問題の例として，２００７年の東京大学の入学試験問題世界史科目の問1を示

[1] http://21robot.org/
[2] http://research.nii.ac.jp/ntcir/ntcir-11/
[3] NTCIR-11 の QA Lab では，2007 年と 2003 年のセンター試験，および，2007 年の，東京大学，京都大学，北海道大学，早稲田大学，中央大学の二次試験を対象としている．
[4] QA Lab に参加したシステムでは，指定語句のない論述問題や穴埋め問題など全ての型の質問に対応している．文献 [2] を参照されたい．

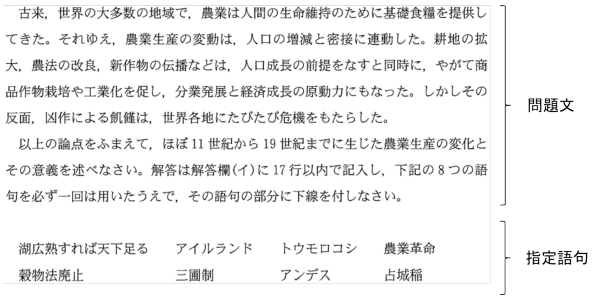

図 1: 二次試験の例

す．東京大学世界史問１では，図１が示すように，問題文と指定語句が与えられる．問題文中には,「１７行以内」といった解答の文字数制限や「ほぼ１１世紀から１９世紀まで」といった解答の時間指定が記述されている．このときの１行は３０字である．例年の傾向では，制限文字数は１７～２０行程度（つまり，５１０～６００字程度），指定語句数は，７，８語である．図１の正解例として「赤本」[5] の解答を図２に示す．

4 知識源

NTCIR-11 QA Lab タスクで配布された次の教科書のテキストデータを教科書データとして使用した．

- 株式会社山川出版社・詳説世界史 B（世 B 016）
- 東京書籍株式会社･世界史 A（平成２０年度発行）
- 東京書籍株式会社･世界史 B（平成１９年度発行）
- 東京書籍株式会社・新選世界史 B（平成１９年度発行）

5 提案手法

Sakamoto et al.[2] が指定語句ありの論述問題を解く際に使用した方法を本稿の提案手法１とする．さらに，提案手法１の問題点を説明し，手法２を提案する．

提案手法１と提案手法２は，図４のように知識源から指定語句を含む文を検索し，得られた文の集合から解答に含む文を選択し，文を時間順に並び替え，並び替えられた文を結合して，１つのテキストとして出力するというおおまかな流れが共通している．しかし，提案手法１と提案手法２では解候補の適切性の判断基準が異なるため，得られた文の集合から解答に含む文を選択する処理が異なる．

提案手法１と提案手法２で共通する処理の流れは次である．

[5]教学社が出版する大学入試問題集

中世西ヨーロッパでは、11世紀頃から普及した三圃制農法が犁・水車の改良とあいまって農業生産力を増大させ、人口増加とそれに伴う都市の発展や東方植民の原因となった。同時期の中国では、宋代にベトナムから日照りに強い占城稲が導入され、長江下流の水稲地帯で集約的農法が発展したが、明代になると長江下流では家内制手工業が盛んとなり、原料である綿花や桑の栽培が普及したため、米の主産地は「湖広熟すれば天下足る」といわれるように長江中流域に移動した。16世紀以降にはアメリカ大陸のトウモロコシが家畜の飼料として世界各地に普及し、アンデス原産のジャガイモは主にヨーロッパの寒冷地に拡大してその人口増を支えたがアイルランドでは19世紀半ばの「ジャガイモ飢饉」によって多くの餓死者とアメリカへの移民を生むことになった、一方、18世紀のイギリスではノーフォーク農法や第2次囲い込みに代表される農業革命によって農業の資本主義化が進み、この結果土地を失った農民は都市に流入して産業革命を支える労働力となった。また産業革命によって台頭した産業資本家の要求により、地主保護のため制定されていた穀物法廃止が実現し、イギリスでは自由貿易体制が確立した。

図 2: ２００７年度の問題に対する「赤本」の正解例

(1) 問題から，問題文，指定語句，制限文字数，指定された時間情報（例えば,「１１世紀から１９世紀まで」）を抽出する．

(2) 教科書から指定語句を含む文を検索する．

(3) 検索された文を指定語句ごとにグループ化する．

(4) 検索された文ごとにその文に記述された内容が関連する時間を推定する．

(5) 検索された文の集合から問題文で指定された時代に該当しない文を削除する．

(6) 検索された文内において不要な記述を削除する．

(7) 検索された文の集合から解答に含む文を選択する．ここは，提案手法１と提案手法２で異なる．

(8) 取り出された文の組み合わせから，文を時間順に並び替える．

(9) 並び替えられた文を結合して，１つのテキストとして出力する．

5.1 提案手法１

提案手法１は，指定語句，制限文字数，および時間情報を手掛かりに解答を作成する．

同手法ではまず教科書データから指定語句が含まれる文を取り出す．指定語句が複数含まれる文がほとんど存在しなかったため，おおむね，指定語句ごとに文を取り出すことになる．指定語句ごとに取り出された文の内，最小文字数の文を選択し，それらを組み合わせることで得られる文字数は制限文字数以内に収まっていることが多い．しかし，最小文字数の文は指定語句以外の情報をほとんど含んでいない場合が多いため，それらを組み合わせても解候補の適切性は低くなると考えた．そこで，解候補の適切性を解答の総文字数の

多さと捉え，各指定語句ごとにその指定語句を含む文の文字数の総和が制限文字数を超えない範囲で可能な限り文字数を増やした．

図 3 に示すように，同手法の文選択の処理の流れは次である．

(1) 検索された文の集合から指定語句を可能な限り全て含み，かつ制限文字数にもっとも近い組み合わせを得る．

(2) (1) の過程で複数の指定語句が含まれている文を発見した場合はその文を残し，含まれなかった指定語句について (1) と同様の処理を行う．

5.2 提案手法 2

提案手法 1 の問題点を説明し，その問題を解決するために本研究で用いた尺度とそれを導入したシステムの処理の流れを説明する．

提案手法 1 では，解候補の適切性を各指定語句ごとにその指定語句を含む文の文字数を増やすときの解答の総文字数の多さとしていた．解候補の適切性を，問題文と解答の関連性，文の内容の結束性や質問の型に応じた記述スタイルを満たす度合で測ることでよりよい解答が生成されることが考えられる．

目標とするシステムの処理の流れは次である．

(1) 教科書データから指定語句を含む文の検索

(2) 検索された文の問題文と解答の関連性の計測

(3) 指定語句を全て含むような組合せを全て作成し，文の内容の結束性の計測

(4) 時間や地域，結束性などによる文の並び替え

(5) 文に適切な接続詞を加えることで，文章全体の一貫性の向上（未実装）

(6) 質問の型に応じた記述スタイルを満たす度合の計測（未実装）

(7) 計測されたスコアから解の選択

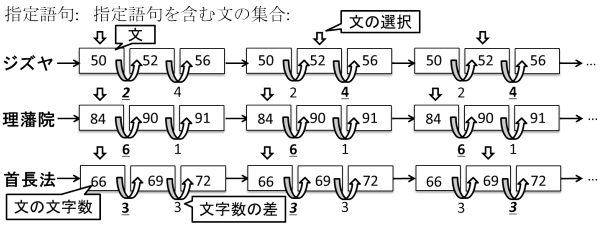

図 3: 提案手法 1 の文選択

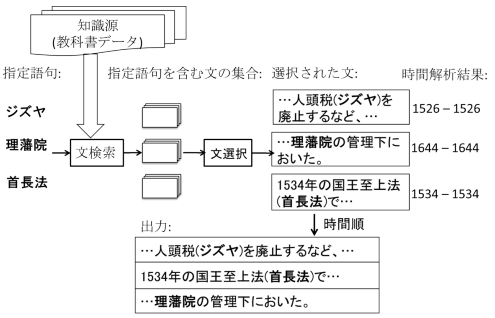

図 4: 提案手法の処理の流れ

本稿の提案手法 2 を用いたシステムでは，(5) と (6) は未実装である．

提案手法 2 を用いたシステムでの解候補の適切性は，

【尺度 1】問題文と解答の関連性

【尺度 2】解答に含む文の内容の結束性

の組み合わせで見積もる．

【尺度 1】は問題文と解答に含む文の類似度の総和で計算する．また,【尺度 2】は解答に含む文の間の類似度の総和で計算する．類似度は，内容語（動詞，形容詞，名詞）の頻度ベクトルのコサイン類似度で計算する．

スコアは，0 以上 1 以下のパラメタ α を用いて次の式で計算した．

$$Score = \alpha【尺度1】 + (1-\alpha)【尺度2】 \quad (1)$$

5.3 提案手法 2 の処理の流れ

図 5 に示すように，提案手法 2 の文選択の処理の流れは次である．

(1) 検索された文の集合から指定語句を可能な限り全て含む文の組み合わせをすべて作成する．

(2) 文の組み合わせが,【尺度 1】と【尺度 2】をどれだけ満たしているのかスコアを計算する．

(3) 文の組み合わせの集合の中でスコアが最大となった組み合わせを得る．

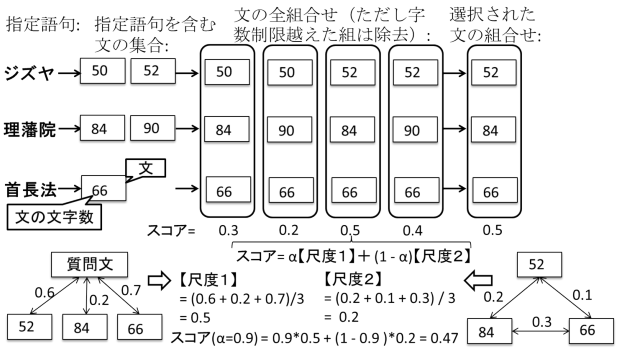

図 5: 提案手法 2 の文選択

6　評価実験

２００５，２００７，２００９年度の東京大学前期試験世界史問１をシステムの入力とした．また，知識源として４冊の世界史の教科書を用いた．

本稿において調査対象として扱った東京大学世界史問１の採点基準は公表されていない．そして，予備校の先生のように世界史の論述問題の採点に長けている方で採点をお願いできる方が見つからなかった．そのため，本稿では「赤本」での正解例を参照要約として用いることで，要約問題で一般的に使用される ROUGE-1, ROUGE-2, および ROUGE-L[4] のスコアを計測し，解答のよさを調査した．

提案手法１，および提案手法２のパラメタ α を 0.0, 0.1, ... , 1.0 として与えた場合について，それぞれ ROUGE1, ROUGE2, および ROUGE-L のスコアを計測する．

7　結果と考察

２００５年度の問題に対する結果を表 1，２００７年度の問題に対する結果を表 2，２００９年度の問題に対する結果を表 3 に示す．α の値が変わっても出力される文章が同じだったものは $\alpha = 0.1$〜0.9 のようにまとめた．なお，$\alpha = 1.0$ は，問題文との関連性のみ，$\alpha = 0.0$ は，文の内容の結束性のみがスコアに影響している．

提案手法２の α 値による出力結果を見ると，$\alpha = 0.0$ のときとそうじゃないときの出力が異なるが，$0 < \alpha$ のときはほとんど同じテキストが出力されている．これは α 値が効きすぎていると考える．今回の実験で使用した【尺度１】の計算方法だと，質問文が複数の文を含むため【尺度１】は複数文と単文の類似度を測る．しかし,【尺度２】は単文と単文の類似度を測るため【尺度１】より小さい値になった．複数文と単文の粒度の違いを均すため,【尺度１】で使用する質問文としての複数文を複数の単文に分割することで，単文と単文の類似度の平均値により【尺度１】を求め,【尺度２】と同じ粒度での類似度を計算することができる．　２．

表 1: ２００５年度の問題に対する ROUGE の値

	ROUGE-1	ROUGE-2	ROUGE-L
提案手法１	0.22	0.16	0.22
提案手法２			
$\alpha = 0.0$	0.30	0.22	0.30
$\alpha = 0.1$〜0.4	0.32	0.24	0.32
$\alpha = 0.5$〜1.0	0.13	0.00	0.13

表 2: ２００７年度の問題に対する ROUGE の値

	ROUGE-1	ROUGE-2	ROUGE-L
提案手法１	0.67	0.31	0.53
提案手法２			
$\alpha = 0.0$	0.56	0.10	0.56
$\alpha = 0.1$〜1.0	0.20	0.00	0.20

表 3: ２００９年度の問題に対する ROUGE の値

	ROUGE-1	ROUGE-2	ROUGE-L
提案手法１	0.44	0.00	0.44
提案手法２			
$\alpha = 0.0$	0.60	0.25	0.60
$\alpha = 0.1$〜1.0	0.60	0.25	0.60

8　まとめ

【尺度１】の計算方法を【尺度２】と同じ粒度の単語と単語の類似度で計算する必要がある．ROUGE による自動評価ではなく，マニュアル評価により本当に解答として適切かどうか評価する必要がある．今回行った実験で，指定語句を教科書データで検索しても一件も見つからないことが１問あたり１，２語存在した．その対策として，同義語により検索できるようにすることや指定語句を見つけられるように検索対象文書を増やすことが今後の課題である．

参考文献

[1] 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.

[2] K. Sakamoto, H. Matsui, E. Matsunaga, T. Jin, H. Shibuki, T. Mori, M. Ishioroshi, N. Kando, “Forst: Question Answering System Using Basic Element at NTCIR-11 QA-Lab Task,” Proceedings of the 11th NTCIR Conference, 2014.

[3] 西川仁, 平尾努, 牧野俊朗, 松尾義博, 松本裕治, 冗長性制約付きナップサック問題に基づく複数文書要約モデル, 自然言語処理, Vol.20, No.4, pp.585–612, 2013.

[4] Lin, C.Y., “ROUGE: A Package for Automatic Evaluation of Summaries,” Proceedings of the Workshop on Text Summarization Branches Out, 2004.