世界史小論述解答システムにおける文圧縮手法の比較・検討

土居 裕典　　松崎 拓也　　佐藤 理史
名古屋大学工学部 電気電子・情報工学科

1　はじめに

現在、「ロボットは東大に入れるか」プロジェクト [1] や NTCIR の QA Lab タスク [2] など、大学入試問題に対する自動解答の研究が行われている。我々は、これらに参加し、世界史記述式試験のうち小論述と呼ばれるタイプの問題の自動解答に取り組んでいる。

世界史小論述問題は、ある歴史的出来事についての問に、30 字から 150 字程度の指定字数以内で解答する問題である。与えられた文の真偽を判定する問題や Factoid 質問型の問題と違い、解答を論述する必要があり、限定的な主題について質問されるため、解答には適切な情報を短くまとめることが求められる。図 1 に小論述問題の例を示す。小論述問題は、教科書などの情報源から解答すべき内容を含む文を抽出したのち、抽出した文から必要な箇所を切り出し、字数に収めるという複合的なタスクと見ることができる。

本稿では、世界史小論述解答システムの最後の段階である解答文生成に焦点を当て、あらかじめ決められたルールに従い圧縮する「ルールベースの圧縮法」と、問題文との関連度を考慮したスコアをもとに圧縮する「クエリ指向要約を用いた圧縮法」の、対照的な二つの手法から得られる解答を比較検討する。この結果から、両手法の欠点を見出し、改善点を明らかにする。

漢の武帝の対匈奴政策と西域政策とのかかわりについて、60字以内で説明しなさい。(2012年度東大)

イタリア戦争はルネサンス期に半世紀以上にわたってくりひろげられた。この戦争の誘因となったイタリアの政治状況について60字以内で説明しなさい。(2003年度東大)

地中海における遠隔地交易を代表する東方交易について，60字以内で説明しなさい。(2009年度東大)

図 1: 世界史小論述問題の例

2　解答文生成アルゴリズム

2.1　ルールベースの圧縮法

本稿で比較した圧縮法のうちの一つ目は、特定のタイプの文節などを削除するいくつかのルールに基づくアルゴリズムとなっている。ルールには、削除する対象が世界史小論述において不必要なものから削除するよう、優先順位をつけている。アルゴリズムの入力は、解答の要素となりうる文を教科書や用語集から抽出し、スコアをつけたものである。具体的な解答システムの構成例は高田らの報告 [3] を参照されたい。

アルゴリズムの大まかな流れを以下に示す。

1. 抽出された文集合をスコアの降順に並べる。
2. スコア順の文リストから解答候補文リストを作る。
3. 文中の単語をより短い同義語で置き換える。
4. 解答候補文リストの先頭の文を、以下に示すルール 4-1 から 4-5 で圧縮する。圧縮の途中で制限字数以下となった時点でそれを解答とする。
5. 全てのルールで圧縮しても制限字数以下にならなければ、解答候補文リストの次の文を圧縮する。

スコアの降順に並んだ文を $s_1, s_2, ..., s_n$ としたとき、$s_1 + s_2, s_1, s_2, ..., s_n$ を解答候補文のリストとした。先頭の $s_1 + s_2$ は、s_1 の後に s_2 を結合したもので、もしも s_1 が字数制限よりも短い文だと s_1 そのものが解答となり、解答に含まれる情報が少なくなってしまうため、s_1 よりも先に解答候補文とした。同義語の置き換えに用いた辞書は、世界史イベントオントロジー [4] から世界史用語の同義語情報を抽出し作成した。

圧縮のためのルールは以下の 5 つである。

4-1. 括弧と文頭の接続詞を削除
4-2. 副詞を削除
4-3. 役割と固有名詞の同格表現のうち、役割の部分を削除: (例)「フランス王シャルル」の場合「フランス王」が役割、「シャルル」が固有名詞
4-4. 時間表現を削除
4-5. 連体修飾句を文末から最も遠い文節から順に削除

上記 5 つのルールで圧縮しても字数制限に収まらなければ、次の解答候補文で同様の圧縮を行う。以上を指定字数に収まる圧縮結果が得られるまですべての解答候補文で繰り返す。最後の文でも圧縮結果が得られない場合は解答は無しとする。

2.2 クエリ指向要約を用いた圧縮法

本稿で比較した圧縮法のうちの二つ目は、森田らの複数ドキュメント要約手法 [5] を簡略化したものである。2.1 のルールベースの手法との大きな違いは、問題文と出力中の語の関連度を考慮している点である。

このアルゴリズムは、ソース文書中のすべての部分木の中で、文字数あたりの目的関数の増分が最も大きいものを順次要約文に加えていく。目的関数はソース文書中の各単語に与えられたクエリとの関連度スコアに基づいており、要約 S に対し、以下の式で与えられる。

$$f(S) = \sum_{w} \left\{ qsb(w) \sum_{i=0}^{count_s(w)} d^i \right\} + \gamma(c(S) - n(S))$$

ここで $d < 1$ は減衰率、$count_s(w)$ は要約 S 中で単語 w が含まれる文の数、$c(S)$ は要約の文字数、$n(S)$ は要約に含まれる文の数、$qsb(w)$ は単語 w の関連度スコアである。第一項は冗長性のペナルティを含んだクエリとの関連度スコアを表し、第二項は断片化した要約を避けるためのペナルティを表している。

各単語に与えられるクエリとの関連度スコアは Morita らによる Query Snowball スコア [6] を用いた。問題文中のキーワード q の集合を Q とし、Q をクエリとみなす。教科書データ中で、キーワード $q \in Q$ と同一文内で共起したキーワード $r1$ の集合を $R1$、$r1$ と同一文中で共起したキーワード $r2$ の集合を $R2$ とする。$s_b(w)$ を単語 w の idf、$freq(q, r1)$ を q と $r1$ の共起回数、$dist(x, y)$ をキーワード x と y が共起した文の文節係り受け木上での x と y の最短文節間距離、$sum_Q = \sum_{q \in Q} s_b(q)$ とするとき、$r1$ の関連度スコアは以下の式で与えられる。

$$s_r(r1) = s_b(r1) \sum_{q \in Q} \left(\frac{s_b(q)}{sum_Q} \right) \left(\frac{freq(q, r1)}{dist(q, r1) + 1.0} \right)$$

同様に $sum_{R1} = \sum_{r1 \in R1} s_r(r1)$ のとき、$r2$ の関連度スコアは以下の式で与えられる。

$$s_r(r2) = s_b(r2) \sum_{r1 \in R1} \left(\frac{s_r(r1)}{sum_{R1}} \right) \left(\frac{freq(r1, r2)}{dist(r1, r2) + 1.0} \right)$$

$R1$ と $R2$ 以外のキーワードのスコアは 0 とする。

以上の定義のようにこの手法では要約に含まれるキーワードと問題文の関連度に基づいた圧縮が行われるため、問題文を考慮しないルールベースの手法よりも問題文に合った情報が解答に残りやすいと考えられる。また、森田らは、述語とその必須格および直前格の文節が分断されないようにするため、述語とこれら格要素の文節を係り受け木上の 1 ノードとして表すようにしているが、本稿ではこれを用いない。

表 1: 抽出元とした教科書

東京書籍 世界史 A(2008 年発行)	162743 文字	2850 文
東京書籍 世界史 B(2007 年発行)	309943 文字	5329 文
東京書籍 新選世界史 (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

3 実験

3.1 実験設定

東京大学の過去問を用いて、2 つの圧縮手法による解答を比較・検討した。問題は東京大学過去問の 1995 年度および 2000〜2012 年度の小論述問題のうち 39 問を用いた。本稿では、解答システムのうち文抽出の部分は無視して、圧縮の部分のみを評価する。そのため、解答に必要な情報が含まれた文集合を教科書 (表 1) から人手で抽出し、その文集合をアルゴリズムで圧縮し解答を生成した。1 問あたりの抽出文の平均数は 5.7 文であった。抽出した文の形態素解析には JUMAN を、係り受け解析には KNP を用いた。スコア計算に用いる idf の計算およびクエリとキーワードの共起情報の取得には、山川出版「詳説世界史 B」(2010 年発行) を用いた。idf の計算に用いた文書の単位は、小節の中にある小見出しで区切られた部分とした。目的関数のパラメータ γ と d の値はともに 0.2 とした。

解答の評価は以下の 2 点に基づき行った。

(a) 問題で尋ねられた情報が含まれているか
(b) 文として自然か否か

(a) は東大世界史に関する解説書 [7] を参考に、解答に加点ポイントが含まれているか否かを人手で判断し、含まれる加点ポイントの数で評価した。解説書 [7] では、例えば「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。」という問題に対して、加点ポイントとして
(1) 皇帝や有力者が、市民の支持を得るためインフラを整備した。
(2) 水道、道路などのライフライン。
(3) 広場・公衆浴場・円形闘技場など政治・娯楽施設。
の計 3 個が列挙されている。例えば、解答が「都市には浴場・凱旋門・闘技場が建設され、道路や水道橋もつくられた。」の場合、加点ポイント (2) と (3) が解答に含まれるとした。解説書で与えられている一問あたりの加点ポイントの平均数は 5.5 個であった。

表 2: 評価結果

	ルールベース	クエリ指向要約
加点ポイント被覆率	23%	22%
文は自然か	43%	23%

表 3: 不自然な解答の分類

	ルールベース	クエリ指向要約
連体修飾句の欠如	18	8
格要素の欠如	0	17
代名詞の照応関係の欠如	1	2
先行文のない接続詞	3	4
時間表現の不完全な削除	7	1
その他	0	4

(b) は解答が自然か不自然かを人手で評価した。具体的には、以下の基準のいずれかひとつでも当てはまる場合に不自然であると判断した。

- 述語に対し省略不可能な格要素が削除されている
- 連体修飾句のうち必須のものが削除されている
- 代名詞の照応関係が分からない
- 接続詞が正しく使われていない

3.2 実験結果

評価結果を表 2 に示す。全問題の加点ポイントのうち解答に含むことができたものの割合 (被覆率) と、解答が自然な文であった問題の割合を両手法で求めた。両手法で加点ポイントの被覆率はほぼ同じであり、両者の解答内容の質には大きな差がないと言える。一方、自然な文で解答できたかどうかには差が現れた。

不自然な解答の分類を表 3 に示す。表の数字は、それぞれの分類項目に当てはまる不自然な解答の数を表す。複数の項目に分類される解答に対しては、それぞれの項目についてカウントした。連体修飾句の欠如には「(遊牧民エフタルの) 侵入をうけたが」、「(北イタリアの) 諸都市は」、「(世宗の) 時代には」、「(諸国の独立を認めた) ため」などの句において括弧内に示した修飾句が削除された解答があった。述語の格要素の欠如には、「移した」「置いた」「して」などの述語で必須の目的語が削除されている例があった。代名詞の照応関係の欠如の例としては、解答が「彼は」や「この時代」といった代名詞から始まり、先行詞が解答中に存在しないものがあった。時間情報の不完全な削除には、時間表現を削除した結果、「後半」「ころ」「かけて」などの時間表現のあとに続く語句のみが残った解答があった。その他には、「(同一の民族と) しての」「(宗教的権威に) よって」といった文節の直前格が削除され、「しての」「よって」などの表現が残った解答があった。

問題：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。

抽出文

文1: 一方フン人は，5世紀前半にアッティラ王(1)がパンノニアを中心に大帝国をたてた(2)。(37字)

文2: ドナウ川中流域のパンノニア平原を拠点(2)としたフン族はアッティラ(1)に率いられて各地に攻勢をかけ，5世紀半ばには大勢力となったが，カタラウヌムの戦いでローマとゲルマン人の連合軍にやぶれ(3)，やがて内紛におちいり瓦解した(5)。(104字)

文3: しかし，最盛期をきずいたアッティラの死後，王国は分裂を重ねて弱体化し(5)，フン族は，アヴァール人やブルガール族などに吸収されていった。(65字)

加点ポイント

(1) アッティラ
(2) パンノニア平原を拠点に大帝国を建設
(3) カタラウムの戦いで、西ローマ・ゲルマン連合軍に敗北した
(4) イタリアに侵入したが、ローマ教皇レオ1世の説得により撤退した
(5) アッティラの死後、帝国は崩壊した

図 2: 問題文・抽出文・加点ポイントの例

4 検討

いずれの手法でも加点ポイントのうち 20%程度のみを含む解答しか得られなかった要因のひとつは、情報源の教科書だけではすべての加点ポイントを網羅しておらず、教科書から抽出された文だけでは入力文として不十分であったことが考えられる。実際に、教科書から抽出された入力文中には約 7 割ほどの加点ポイントしか含まれていなかった。また、多くの加点ポイントを含む解答を生成するには、複数の文から短い単位で適切な箇所を抽出しつなぎ合わせることが必要である。そのことを示す値として、入力文 1 文あたりの加点ポイント数の平均は 1.6 個であった。図 2 に入力文と加点ポイントの例を示した。図では、入力文中の加点ポイント該当箇所に下線を引いている。ルールベースの圧縮法は 2 文もしくは 1 文から解答を作るアルゴリズムになっており、クエリ指向要約の圧縮法は断片化を避けるペナルティを考慮した目的関数となっている。その結果、要約に実際に用いられた抽出文の平均数はルールベースの圧縮法では 1.4 文、クエリ指向要約の手法では 1.8 文と少なくなった。そのため、両手法とも多くの加点ポイントを解答に含むことができなかったと考えられる。以下では、解答が不自然な文になった要因を圧縮法ごとに考える。

4.1 ルールベースの圧縮法の検討

ルールベースの圧縮法による解答で多く見られたタイプの不自然な圧縮結果を含む解答例を以下に示す。

解答例 ササン朝は後半、侵入をうけたが、時代に、突

厥と結んでエフタルをほろぼし、また戦いも優勢にすすめ、和平を結んだ。

この例の「(5 世紀の) 後半」の部分でも見られるように、時間表現の削除のルールにない表現、例えば「〇〇世紀になると」などは、ルールにある「〇〇世紀に」の部分だけが削除され、「なると」のみが解答に残る。これは時間表現の削除のルールが不十分であったためである。同様に、接続詞の削除においてもルールが不十分であったため、ルールになかった「一方」や「ついで」といった文頭の接続詞が削除されず解答に残った。また、上の例の「(エフタルの) 侵入」、「(ホスロー 1 世の) 時代」の部分のように、実際には省略不可能な連体修飾句が削除される例が多く見られた。同様の例としては他にも「一部を手に入れ」、「影響をうけて」などの連帯修飾句が削除されたために、意味がわからない場合があった。この問題に対しては、必須の連体修飾句をもつ名詞の辞書を作成し、辞書中の名詞の連体修飾句を削除しないようルールを精緻化することが考えられる。

4.2 クエリ指向要約の圧縮法の検討

クエリ指向要約を用いた圧縮法の解答で多く見られた、目的語の欠如を含む解答例を以下に示す。

解答例 ササン朝は 5 世紀の後半、中央アジアの遊牧民エフタルの侵入をうけたが、ほろぼし、またビザンツ帝国との戦いもすすめ、結んだ。

上の例の「ほろぼし」、「結んだ」において必須の格要素「エフタルの」「和平を」が削除されてしまっている。また、この圧縮法で解答に加えられる部分木は、必ず係り受け木の根を含む部分木である。そのため、制限字数に少しでも余裕があれば根の文節のみの部分木が解答に加えられ、不自然な解答になる。以下はその例である。

解答例 大月氏と同盟して攻撃するため、張騫を派遣したことをきっかけに、タリム盆地のオアシス諸都市にまで支配をひろげた。支配した。

これもまた、「支配した」の目的語を削除可能としたため起こった問題であるとも言える。本実験では、森田ら [5] が提案したように、述語とその必須格および直前格の文節が分断されないようにするために、係り受け木上の 1 ノードとして表すことは行っていないが、これを導入すれば必要な目的語が削除される問題の大部分は解決できると予想される。

5 おわりに

世界史小論述解答システムにおける文圧縮手法について、「ルールベースの圧縮法」と「クエリ指向要約を用いた圧縮法」から得られた解答を比較検討した。両手法による解答は予想される得点に大きな差は見られなかったが、文の不自然さには傾向の違いが見られた。今後の課題としては、ルールベースの手法では現在の不完全なルールを不要な部分を正確に削除できるルールを精緻化すること、連体修飾句をもつことが必須である名詞の辞書を作成して利用すること等が挙げられる。クエリ指向要約の手法では述語に対する必須格および直前格を削除しないことや、接続詞や副詞をあらかじめルールで削除することが可能な改良として挙げられる。

謝辞 教科書データを提供いただいた東京書籍および山川出版、また「ロボットは東大に入れるか」プロジェクトに感謝します。

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, Yoshinobu 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*, 2014.

[3] 髙田拓真, 土居裕典, 松崎拓也, 佐藤理史. 主題と焦点の同定に基づく『世界史』小論述問題の自動解答. 言語処理学会第 23 回年次大会, to appear, 2017.

[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 *JSAI International Symposium on Artificial Intelligence*, pp. 42–50. Springer, 2013.

[5] 森田一, 笹野遼平, 高村大也, 奥村学. 劣モジュラ最大化アルゴリズムを用いた文抽出と文圧縮に基づくクエリ指向要約. 言語処理学会第 19 回年次大会, pp. 500–503, 2013.

[6] Hajime Morita, Tetsuya Sakai, and Manabu Okumura. Query snowball: a co-occurrence-based approach to multi-document summarization for question answering. In *Proc. ACL-HLT,Volume 2*, pp. 223–229, 2011.

[7] 渡辺幹雄, 茂木誠. 24 カ年徹底分析　テーマ別東大世界史論述問題集. 駿台文庫, 2013.