大学入試世界史論述問題における 専門知識の有無による評価の一致に関する考察

渋木 英潔[†1]　阪本 浩太郎[†1†2]　石下 円香[†2]　藤田 彬[†2]
狩野 芳伸[†3]　三田村 照子[†4]　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　[†3]静岡大学　[†4]Carnegie Mellon University

1　はじめに

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 において QA Lab タスクを提案している[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世界史の大学入試問題を対象とし，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）に解答するタスクを設定している．特に二次試験には数行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おり，従来の質問応答で想定された状況とは大きく異なるチャレンジングな課題となっている．

図 1 に論述問題の例を示す．核となる質問文の他に質問背景や指定語句などの条件が書かれており，この例ではこれらの条件を満たす記述を 20 行（600 字）以内で解答しなくてはならない．知識源として教科書や Wikipedia などを利用することができるが，600 字という字数は問われている内容と比べて非常に短く，単純な抜粋要約では制限字数内に収めることは困難である．さらに，質問背景には，どのような観点から回答してほしいかが暗黙的に示されており，ただ教科書の内容を要約しただけでは高得点にならない．

このように解答するのが非常に難しいタスクであるが，オーガナイザ側からの別な課題としてはシステムが出力した解答の評価方法がある．世界史という専門知識が要求される解答を正しく評価するためには，世界史の知識をもつ専門家に依頼するのが理想である．しかしながら，全ての解答に対して専門家に依頼することはコストの面で厳しいものがある．そのため，模範解答を参照要約とみなして ROUGE[4]などの自動評価手法を用いることが考えられる．しかしながら，ROUGE は単語 n-gram の表層的な一致に基づいた尺度であるため，内容の真偽を正確に判定しなくてはならない大学入試問題の尺度として適切かは議論の余地がある．単語よりも正確に内容を考慮できる評価単位として，Pyramid 法[5]における nugget や iUnit[6]などの単位がある．

我々は QA Lab-2 において nugget などの単位を用いた評価を行いたいと考えている．しかしながら，大学入試問題という高度な専門知識を要求されるタスクにおいて，専門家でない評価者が解答中の nugget の有無を正しく判定できるのかという懸念がある．それゆえ，本稿では，世界史論述問題を対象に専門家とそうでない評価者との間でどの程度判断が一致するかの実験を行う．また，自動評価に向けて，得点と ROUGE スコアとの間の相関についても考察する．

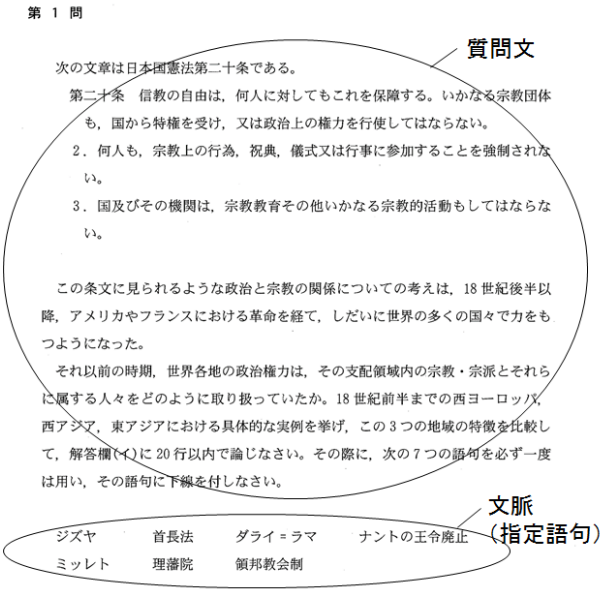

図 1: 論述問題の例

表 1: 各解答の得点（26 点満点）

	解答 1	解答 2	解答 3	解答 4	解答 5
Forst	8	**9**	*9*	*12*	*11*
SML	2	1	3	3	4

※東ロボくんの解答は Forst の解答 2
※Forst の解答 3,4,5 は参考記録（人手による処理あり）

2　対象データ

2015 年 10 月に行われた QA Lab-2 の Phase-2 では，「ロボットは東大に入れるか」プロジェクト（以下，東ロボくん）[7]と連携し，駿台予備学校の協力の下，東大入試実戦模試に挑戦した．東大入試実戦模試では，図 1 に示すような数百字で記述させる大論述論題の他にも，二，三行で解答させる小論述問題，国名や事件名などを解答させる語句問題が存在するが，本稿では大論述のみを対象とする．参加したのは，Forst（横浜国大と NII）と SML（名古屋大）の 2 チームで，両チームとも 5 通りの解答を提出した．採点は駿台予備学校の講師に行っていただき，解答中のどの記述が点数につながったかといった詳細も示していただいた．さらに，模範解答および加点項目や採点に関する留意点についても公開していただいた．加点項目は，

- ルイ 14 世の下で国家体制が整備された

Copyright(C) 2016 The Association for Natural Language Processing.
All Rights Reserved.

表 2: 各解答の ROUGE-1 スコア（F 値）

	解答 1	解答 2	解答 3	解答 4	解答 5
Forst	.230	**.245**	.245	.330	.309
SML	.130	.131	.165	.126	.117

※東ロボくんの解答は Forst の解答 2

表 3: 各解答の ROUGE-2 スコア（F 値）

	解答 1	解答 2	解答 3	解答 4	解答 5
Forst	.049	**.049**	.042	.083	.036
SML	.015	.015	.038	.024	.024

※東ロボくんの解答は Forst の解答 2

といった文単位で示され，nugget に相当すると考えられる．加点項目は全部で 41 項目あった．

各解答の得点を表 1 に示す．Forst の解答 3,4,5 は人手による処理が加えられているため参考記録であり，東ロボくんの解答としては自動生成された解答の中での最高得点だった Forst の解答 2 である．しかしながら，本稿では参考記録を含めた全ての解答を対象とした．

3 得点と ROUGE との相関

自動評価の結果として，模範解答を参照要約とした場合の ROUGE-1 と-2 のスコアを表 2 と表 3 にそれぞれ示す．文献[8]と同じく助詞と助動詞を以外の単語を内容語として内容語単位で評価した．表 1 の得点と比較すると，得点の高い解答の方が高い ROUGE スコアを示していることが分かる．Pearson の相関係数は ROUGE-1 との間で.961，ROUGE-2 との間で.825 となり，両者ともに強い相関を示した．特に ROUGE-1 は極めて強い相関であった．このことは世界史の大学入試問題という高度な専門知識を要求されるタスクにおいても，例えば相対的な順位付けといった，ある程度の自動評価が可能となることを示すものと考えられる．

4 専門知識の有無による一致率

4.1 評価方法

専門知識をもたない評価者として，自然言語処理を専攻する大学院生 A~D の 4 名に加点項目の有無を判定してもらった．彼らは大学受験のための世界史を勉強していない．

評価の流れは以下の通りである．最初に問題および加点項目と採点に関する留意点を熟読してもらった．その後，各解答を読み，加点項目ごとに該当すると思われる箇所に下線を引き，該当する加点項目の番号を記入してもらった．ある記述が複数の加点項目に該当すると考えられる場合，複数の下線および番号を許可した．評価者間での相談を禁じたが，内容的に分からない記述があった場合，書籍やインターネット等での調査を許可した．その際，調査した箇所および内容についてメモを残してもらった．その他，判断に迷った箇所や疑問点などについてもメモを残してもらった．

表 4: 模範解答における加点項目の有無

	院生 A	院生 B	院生 C	院生 D
項目 1	○	○	○	○
項目 2	○	○	○	○
項目 3	○	○	○	○
項目 4	○		○	○
項目 5	○	○	○	○
項目 6		○		○
項目 7	○	○	○	○
項目 8	○	○	○	○
項目 9	○	○	○	○
項目 10	○	○	○	○
項目 11	○	○	○	○
項目 12	○	○	○	○
項目 13	○	○	○	○
項目 14	○	○	○	○
項目 15	○	○	○	○
項目 16	○	○	○	○
項目 17	○	○	○	○
項目 18	○	○	○	○
項目 19	○	○	○	○
項目 20	○	○	○	○
項目 21				○
項目 22	○	○	○	○
項目 23	○	○		○
項目 24				
項目 25	○	○	○	
項目 26		○	○	○
項目 27	○	○	○	○
項目 28				
項目 29	○		○	○
項目 30	○	○	○	○
項目 31	○	○	○	○
項目 32				
項目 33	○	○	○	○
項目 34	○	○	○	○
項目 35	○	○	○	○
項目 36	○	○	○	○
項目 37	○	○	○	○
項目 38	○	○	○	○
項目 39	○	○	○	○
項目 40	○	○	○	○
項目 41	○	○	○	○

駿台予備学校から示された採点の詳細には，解答中のどの記述が加点対象であるかが示されていたが，加点項目の何番目に該当するかは示されていなかった．それゆえ，第一著者が加点対象の記述からどの項目に対応するかを推測した．

解答の質による影響を調査するため，評価者には各チームの解答と一緒に模範解答についても，模範解答と知らせずに評価してもらった．一致率の尺度として Fleiss の κ 値を用いた．

Copyright(C) 2016 The Association for Natural Language Processing.
All Rights Reserved.

表 5: Forst の解答 2 における加点項目の有無

	駿台	院生 A	院生 B	院生 C	院生 D
項目 1					
項目 2					
項目 3					
項目 4					
項目 5					
項目 6					
項目 7		○			○
項目 8					
項目 9					
項目 10	○				
項目 11					
項目 12					
項目 13					
項目 14					
項目 15					
項目 16		○		○	○
項目 17					
項目 18	○	○	○	○	○
項目 19					
項目 20					
項目 21					
項目 22					
項目 23					
項目 24					
項目 25					
項目 26					
項目 27	○	○	○		○
項目 28					
項目 29	○	○	○	○	○
項目 30		○	○	○	
項目 31	○				
項目 32					
項目 33			○	○	
項目 34					
項目 35					
項目 36					
項目 37	○	○	○	○	○
項目 38	○	○	○	○	○
項目 39	○	○	○	○	○
項目 40	○	○	○	○	○
項目 41					

4.2 結果と考察

表 4 と表 5 に，模範解答と Forst の解答 2 における加点項目の有無をそれぞれ示す．評価者が「有」と判断した項目に「○」をつけている．模範解答における院生 4 名の一致率は.615 であり，両チーム合わせた 10 の解答における院生と駿台予備学校の 5 名の一致率は.753 であった．どちらも高い一致率といえる．

表 6: 解答ごとの一致度（Fleiss）

	解答 1	解答 2	解答 3	解答 4	解答 5
Forst	.763	**.752**	.752	.950	.622
SML	.350	.283	.932	.557	.568

※東ロボくんの解答は Forst の解答 2

表 7: 二者間の一致率（Cohen）

	院生 A	院生 B	院生 C	院生 D
駿台	**.704**	**.731**	**.620**	**.758**
院生 A		.777	.796	.783
院生 B			.791	.715
院生 C				.777

※駿台と院生間の平均　**.703**
二人の院生間の平均　.773

解答ごとの一致率を表 6 に示す．解答ごとのばらつきが大きいことが分かるが，SML の解答 1，2 を除き，中程度以上の一致率となっている．SML の解答 1，2 の一致率が低いのは，表 1 に示すように得点が低く，項目を「無」とすることによる偶然の一致率が高く見積もられてしまったことによると思われる．表 5 に示すように比較的「有」が存在する場合には全体的に一致する傾向があった．

ここで問題となるのは，大学入試の正解は多数決ではないため，一致率が高いことが正解とはならないことである．そこで，Cohen の κ 値を用いて二者間の一致率を求めた．両チーム合計の 10 の解答における結果を表 7 に示す．いずれの組み合わせにおいても高い一致率であるが，表中で下線を引いた駿台予備学校との一致率の平均は.703 であり，それ以外の院生間の一致率の平均は.773 であった．このことは専門家とそうでない評価者との間に判断のずれがあることを示している．例えば，表 5 において項目 10 と 31 は駿台予備学校のみが「有」と判断しているのに対し，項目 16 と項目 30 は院生 3 名のみが「有」と判断している．

専門家とずれがあった加点項目を以下に示す．専門家のみが「有」と判断したのは以下の 2 項目であった．

- 項目 10.　啓蒙思想における旧制度批判が 11.[1] の背景
- 項目 31.　要職に満州人と漢人を同数配置する（or 懐柔策として）満漢併用制がとられた

一方，専門家が「無」と判断し，3 人以上の院生が「有」と判断したのは以下の 5 項目であった．

- 項目 2.　ルイ 14 世の下で国家体制が整備された
- 項目 3.　フランスでは絶対王政（絶対主義）が確立した

[1] 項目 11.は「フランス革命が勃発した」

Copyright(C) 2016 The Association for Natural Language Processing.
All Rights Reserved.

- 項目 16. （スンナ派の）イスラーム法（シャリーア）に基づく統治
- 項目 30. 諸制度の具体例として科挙（or 儒教 or 朱子学）に基づく皇帝専制体制
- 項目 38. 人口増加と開墾による環境破壊（or 土地不足）が進み，社会不安が増大した or 中央集権が弱まった

院生が評価した際のメモを見ると，前者の専門家のみが「有」と判断した項目に関しては，「旧制度批判について明記されていない」や「フランス革命の背景としては書かれていない」といったメモが残されており，有無を悩んだ末に「無」と判断しているものが多かった. 一方, 後者の 3 人以上の院生が「有」と判断した項目に関しては, メモが残されておらず, 悩むことなく「有」と判断しているものが多かった.

これらの項目について院生にインタビューを行った結果，前者に関しては，加点項目に書かれている条件を厳密に適用しようとした結果，加点項目に関して述べている記述であっても該当しないと判断したことが分かった．駿台予備学校の採点詳細と合わせると,「該当しないことを明確に示す記述がない限りは該当するとみなす」と考えた方が良い結果になる可能性がある．この点について今後さらに分析を進めていきたい. また, 後者に関して,「ルイ 14 世」,「絶対王政」,「イスラーム法」,「科挙」といった, その項目に特有の語句が含まれている記述に対して, あまり深く考えずに「有」と判断したことが分かった．3 節で述べたように，単語単位での表層的な一致度を測る ROUGE スコアが文章全体に対する得点と強い相関がある一方で，後者の項目のように，単語単位の有無だけでは，文章よりも小さな nugget 単位の有無を正確に判断できない事実もあることから，両者の違いが何に由来するのか，さらに分析を進めていきたいと考えている.

5 まとめ

本稿では，世界史の東大入試実戦模試の結果を用いて，論述問題の自動評価に向けて，専門知識の有無による nugget レベルでの加点項目の一致率を考察した．模範解答を参照要約とした ROUGE-1 スコアと得点との間には.961 という極めて高い相関があり，専門知識を必要とする世界史論述問題においても，表層的な一致に基づいた自動評価に妥当性があることが分かった．評価者の専門知識の有無に関係なく，nugget の有無に関する判断は κ 値で.735 という高い一致率となった．一方で，専門知識の有無により nugget 有無の判断が異なる項目も存在し，条件的な記述を厳密に適用することにより誤った判断をしてしまう場合があることが分かった．今後，専門家でなくとも自動評価できるよう，さらに分析を進めていきたいと考えている.

謝辞

本研究の実施にあたって, 各種データの提供や各解答の詳細な採点をしていただいた駿台予備学校の皆様, プロジェクト「ロボットは東大に入れるか」を推進している新井紀子教授をはじめとする国立情報学研究所の皆様に深く感謝いたします.

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", In Proc. of the 11th NTCIR Conference, 2014.

[4] C.-Y. Lin, "Rouge:A package for automatic evaluation of summaries", In Proc. ACL workshop on Text Summarization Branches Out, pp.74–81, 2004.

[5] J. Lin, and D. Demner-Fushman, "Will pyramids built of nuggets topple over?", In Proc. of the main conference on Human Language Technology Conference of the North American Chapter of the Association of Computational Linguistics. 2006.

[6] M. P. Kato, M. Ekstrand-Abueg, V. Pavlu, T. Sakai, T. Yamamoto, M. Iwata, "Overview of the NTCIR-10 1CLICK-2 Task", In Proc. of the 10th NTCIR Conference, pp.182–211, 2013.

[7] 新井紀子, 松崎拓也. ロボットは東大に入れるか?-国立情報学研究所「人工頭脳」プロジェクト. 人工知能学会論文誌, 9 2012.

[8] 阪本 浩太郎, 石下 円香, 藤田 彬, 渋木 英潔, 狩野 芳伸, 三田村 照子, 森 辰則, 神門 典子..大学入試の世界史論述問題における質問応答システムの自動評価に関する一考察. 第 222 回情報処理学会自然言語処理研究会(SIG-NL), 2015.

Copyright(C) 2016 The Association for Natural Language Processing.
All Rights Reserved.