Vol.2015-NL-222 No.13
2015/7/16

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

阪本浩太郎[†1†2]　石下円香[†2]　藤田彬[†2]　渋木英潔[†1]
狩野芳伸[†3]　三田村照子[†4†2]　森辰則[†1]　神門典子[†2]

NTCIR12 QA Lab-2 タスクでは現実世界における質問応答システムの実現を目指して世界史の大学入試問題を解くことを目的としている．QA Lab-2 で扱う大学入試の二次試験には，論述問題が存在し，非常にチャレンジングな課題となっている．本論文では，大学入試の論述問題を解く質問応答の評価手法について検討する．論述問題の模範解答を参照要約と捉え，参照要約を用いた ROUGE やピラミッド方式による評価手法を適用するにあたり，参照要約の数による評価値の安定性と参照要約との一致率を計算する単位について考察する．

A Study in Automatic Evaluation of QA Systems for Essay Questions of World History in University Entrance Exams

KOTARO SAKAMOTO[†1†2]　MADOKA ISHIOROSHI[†2]　AKIRA FUJITA[†2]
HIDEYUKI SHIBUKI[†1]　YOSHINOBU KANO[†3]　TERUKO MITAMURA[†4†2]
TATSUNORI MORI[†1]　NORIKO KANDO[†2]

NTCIR12 QALab-2's goal is to investigate the real-world complex Question Answering (QA) technologies using Japanese university entrance exams .In the task, essay question is a big chanllenge.This paper discusses the evaluation method of the QA system for the essay questions. By referring to the gold standards, we apply the evaluation method such as ROUGE and Pyramid method using reference summaries to evaluate the system summary. Then, we investigate the stableness of evalueation score based on the number of the reference summaries and the unit of calculating the accordance rate with the reference summaries.

1. はじめに

我々は，現実世界における質問応答システムの実現を目指して世界史の大学入試問題を解くことを目的とした QA Lab-1 タスク[1]を NTCIR-11 で行い，NTCIR-12 でも QA Lab-2 タスクを行っている．大学入試には，多肢選択問題であるセンター試験と，自由記述問題が主である二次試験があるが，QA Lab では両方を対象としている．特に，二次試験には，「○○について 600 字以内で述べよ」といった論述問題が存在し，非常にチャレンジングな課題となっている．

論述問題を解くことができるシステムの構築も重要な課題であるが，オーガナイザの立場からは，システムが出力した結果をどのように評価するかも重要な課題である．人名や地名といった語句を尋ねる質問と異なり，論述問題の場合，システムの出力が模範解答と表層的に完全一致することはない．また，人間が読んで判断する場合でも，世界史という専門性が高い分野であるため，世界史の知識に乏しい人間には，正誤の判断が困難な解答も多い．したがって，理想的には，システムの出力を世界史の専門家に実際に読んでもらい採点してもらうことが望ましいが，コストの面で厳しいものがある．それゆえ，自動評価，もしくは世界史の知識に乏しい人間でも判断できる程度のサポートで評価可能な方法が望まれる．

以上の背景から，本稿では，大学入試の世界史論述問題における自動評価に向けた調査を行う．2 節で論述問題の評価に対する我々の基本的な考え方と調査すべき内容について述べた後，3 節で具体的な調査方法について説明する．4 節で調査結果とその考察を述べる．5 節は結論である．

2. 論述問題の評価方法と調査内容

論述問題の評価には，模範解答を参照要約とみなせば，ROUGE[2]やピラミッド法[3]といった参照要約を用いた評価手法を適用することができると考えられる．参照要約を用いた評価手法を適用するにあたり，本稿では，以下の 2 点に関する調査を行う．

A) 模範解答（参照要約）の数による評価の安定性
B) 模範解答（参照要約）との一致率を計算する単位

A)に関して，一般に参照要約の数が多いほど，安定した評価値が得られると考えられるが，(A1)論述問題においても同様の傾向がみられるのか，もしみられるならば，(A2)どのくらいの数の模範解答を用意すれば安定するのか，という疑問は明らかではない．本稿では，この 2 点に関して

†1 横浜国立大学
Yokohama National University
†2 国立情報学研究所
National Institute of Informatics
†3 静岡大学
Shizuoka University
†4 カーネギーメロン大学
Carnegie Mellon University

ROUGE を用いた調査を行う.

B)に関して，ROUGE では文字または形態素を単位とした一致率で計算を行っているが，ピラミッド法の SCU[3]や iUnit[4]といった，より正確に内容を把握できる単位での計算を行うことも考えられる．しかしながら，ピラミッド法などのマニュアル評価は，正確である反面，非常にコストがかかるという問題がある.このコストの内訳は大きく，(B1)テキストから適切な単位で切り出すコスト，(B2)単位同士が一致しているか判断するコスト，に分けることができる．我々は，論述内容を把握するのに適切で，(B1)のコストを軽減するのに適した単位として，事態性名詞を考慮に入れた述語項構造を仮定する．世界史の論述問題を評価する場合，名詞句レベルの一致では内容を把握するのに不十分であり，少なくとも「A が B を滅ぼした」のか「B が A を滅ぼした」のかを区別できる必要がある．また，述語項構造解析に関する研究はいくつか行われており[5],まだ実用レベルではないものの比較的自動抽出が可能な単位と考えられる．以上の理由から，本稿では，人手で模範解答を述語項構造に分解し，述語項構造の同一性を判断する上でどのような問題が生じるのかを調査する.

3. 調査方法

3.1 調査対象

東京大学の世界史科目の第一問と京都大学の世界史科目の第一問・第三問は，毎年論述問題が出題されている．それら 3 問の 2004 年～2011 年の 8 年分の計 24 問について，それぞれの模範解答を 5 種類の計 120 解答を用意した.

5 つの模範解答のうち 2 つは，書籍「教学社の大学入試シリーズ」(通称，赤本）と河合塾のウェブサイト「大学入試解答速報」[a]からそれぞれ取得した．残りの 3 つは，模範解答の作製を株式会社アイアール・アルトに依頼し，株式会社アイアール･アルトが雇った次の 3 名の解答である.

【解答作製者1】
東京大学大学院歴史学専攻博士課程在籍
イスラームとヨーロッパ（オスマントルコ史）が専門
現在，大学の非常勤講師
塾講師，家庭教師の経験あり

【解答作製者2】
高校非常勤講師2年目
東京学芸大学史学専攻修士課程修了

【解答作製者3】
東京大学大学院歴史学専攻博士課程在籍
東アジア，中国史が専門
現在，大学の非常勤講師
塾講師，家庭教師の経験あり

これら 120 個の模範解答に対し， ROUGE の値やそれによる順位の安定性評価を行う．また，東京大学の世界史科目の第一問と京都大学の世界史科目の第一問・第三問の 3 問に関して，2005 年，2007 年，2009 年の 3 年度分の解答作製者 3 名の計 27 個の模範解答を対象に人手で述語項構造の分割を行う.

3.2 模範解答（参照要約）の数による評価の安定性

3.2.1 調査の設定

24 問の論述問題に対しそれぞれ 5 種類の模範解答を用意した．この 5 種類をシステム要約と参照要約に分類する．参照要約数が 1 つの場合，2 つの場合,3 つの場合，4 つの場合の 4 通りある．参照要約数が N 個の場合，5 種類の模範解答から N 個を参照要約にする場合が C(5, N)通りあり，残りの 5-N 個から 1 つシステム要約にする場合が 5-N 通りあるため，C(5, N)*(5-N)通りの ROUGE の値が生成される．すなわち，1 問につき計算される ROUGE の値は，

- 参照要約数 1 の場合，ROUGE の値は 20 通り
- 参照要約数 2 の場合，ROUGE の値は 30 通り
- 参照要約数 3 の場合，ROUGE の値は 20 通り
- 参照要約数 4 の場合，ROUGE の値は 5 通り

の計 75 通りである.

ROUGE は CY Lin の論文[2]に登場する ROUGE-1, ROUGE-2, ROUGE-3, ROUGE-4, ROUGE-5, ROUGE-6, ROUGE-7, ROUEG-8, ROUGE-9, ROUGE-L, ROUGE-W1.2, ROUGE-S*, ROUGE-S4, ROUGE-S9, ROUGE-SU*, ROUGE-SU4, ROUGE-SU9 の 17 種類を出力した．ROUGE は模範解答から抽出された内容語を用いて計算し，内容語を抽出するために形態素解析器 MeCab と辞書 Unidic を用いた.

論述問題 24 問のそれぞれに 75 通りある ROUGE の値を用いて，まず，参照要約数を固定した場合の ROUGE の値の安定性を調査する.次に,参照要約数を変化させた場合,ROUGE の値によるシステム要約の順位の安定性を調査する．さらに,任意の 2 つのシステム要約の順位の安定性を調査する.

3.2.2 参照用約数を固定した場合の ROUGE の値の安定性

参照用約数を 1 つ，2 つ，3 つ，4 つにそれぞれ固定したときの ROUGE の値の分散値を調べる.

3.2.3 参照用約数を変化させた場合のシステム要約の ROUGE による順位の安定性

ROUGE の値によるシステム要約の順位が参照要約数を変

a) 河合塾の大学入試解答速報
http://kaisoku.kawai-juku.ac.jp/nyushi/

化させても同じであるかどうかを調べるために，次の２つの手法で調査する．

ⅰ）Kendall の一致係数 W を用いる方法

ⅱ）任意の 2 つのシステム要約の評価値の比較する方法

ⅰ)は，3 つ以上の順位系列の一致を調べる手法である．これを参照要約数が 1 つの場合，2 つの場合，3 つの場合，4 つの場合のそれぞれのシステム要約の順位の一致を調べるために用いる．1 つのシステム要約に対し，ROUGE の値の数が，参照要約数 1 の場合 4 通り，参照要約数 2 の場合 6 通り，参照要約数 3 の場合 4 通り，参照要約数 4 の場合 1 通りと，参照要約数によって存在する ROUGE の値の数が異なり，単純には比較できない．そこで，参照要約数が等しい ROUGE の値を幾何平均した値に対する，システム要約の順序を調べた．模範解答が 5 種類あるため，システム要約も 5 通りあり，それらの順位は最大 5 位である．参照要約数が 4 通りあるため，順位は 4 通りであり，それらの．順位相関を調べるため Kendall の一致係数 W を計算した．Kendall の一致係数 W は,0 以上 1 以下で与えられ値が高いほど相関が高い．システム要約数 k (= 5)，参照要約数 n (= 4)，システム要約 i の順位の総和 Ri に対し,次の計算式で求める．

$$S = \sum_{i=1}^{k} R_i^2 - \frac{\left(\sum_{i=1}^{k} R_i\right)^2}{k} \quad (1)$$

$$W = \frac{12S}{n^2(k^3 - k)} \quad (2)$$

ⅰ)の手法では,異なる参照要約数ごとの順位相関を調べることができるが，参照要約数が等しい ROUGE の値を平均化することで，結果的に，参照要約を全て 4 つ用いてしまっている．従って，異なる参照要約数ごとの ROUGE の値の特徴が十分に値に反映されているかがわからない．

そこで，ⅱ)の手法では，任意の 2 つのシステム要約 A と B に対し，同じ参照要約を用いた ROUGE の値を比較し，ROUGE(A) > ROUGE(B)の場合の数N_A，ROUGE(A) = ROUGE(B)の場合の数N_E，ROUGE(A) < ROUGE(B)の場合の数N_Bを調査し，ROUGE(A) > ROUGE(B)，もしくは ROUGE(A) < ROUGE(B)に偏っているかを参照要約数ごとに見ることで順位の安定性を調べる．

任意の 2 つのシステム要約の ROUGE に対し，同じ参照要約を持つ場合は 7 通りあり，参照要約数が n 個の場合の数 Dn は，D1 = 3，D2 = 3，D3 = 1 である．システム要約を 2 つ使用するため，参照要約数が 4 つの場合はこの方法では調査できない．参照要約数を変化させたときに順位の安定性を比較するために，$|N_A - N_B|$を求め，正規化するために Dn で割り，これを総和幾何平均した値

$$average(\frac{|N_B - N_A|}{D_n}) \quad (3)$$

を参照要約数 n における順位の安定性の評価値とする．

参照要約数が 3 つの場合は，ROUGE(A) = ROUGE(B)という特殊な場合を除いて全て$|N_B - N_A| = 1$である．ROUGE(A)も ROUGE(B)も 0 である場合を除いて ROUGE(A) = ROUGE(B)となることは非常に稀であるため，参照要約数が 3 つの場合の (3)は順位の安定性の評価値を表しているとは言いがたい．従って，手法ⅱ)に関しては，参照要約数が 1 つの場合と 2 つの場合の順位の安定性の比較を行う．

3.3 模範解答から人手で分割された述語項構造の同一性

株式会社アイアール・アルトの解答作製者 3 名が自身が作製した解答を述語を中心に分割した．それらを我々はさらに次のように分割する．

まず，事態性名詞を考慮に入れた述語項構造になるように分割する．例えば,「EEC の発足により、西欧の経済復興が進んだ」は,「発足」が事態性名詞なので,「EEC が発足した」と「西欧の経済復興が進んだ」に分割する．さらに，並列構造が含まれる場合は分割する．例えば,「戦後の国際秩序と安全保障構築の試みがあった」の場合は,「国際秩序」と「安全保障」が並列に書かれているため,「戦後の国際秩序構築の試みがあった」と「戦後の安全保障構築の試みがあった」に分割する．ただし,「トウモロコシやジャガイモなどは大航海時代にヨーロッパへと伝播した」の「トウモロコシ」と「ジャガイモ」のように例示で書かれている場合は分割しない．例示はひとつの概念を外延的に記述していると考えられるため，仮に「トウモロコシ」と「ジャガイモ」以外に「トマト」や「サツマイモ」が含まれていても，同じ内容であるとみなせるからである．

以上のように，人手で模範解答を述語項構造に分解する上でどのような問題が生じるか，また，分解された述語項構造の同一性を判断する上でどのような問題が生じるかを調査する．

4. 結果と考察

4.1 複数の模範解答に対する ROUGE の安定性の調査結果

表 1 のように，ROUGE-N の N が 3 以上のものには ROUGE の値が 0 となるものが含まれていた．これはシステム要約と参照要約の中の内容語がひとつも一致しないことを示す．同じ論述問題の模範解答の内容が全く一致しないということは考えづらいため，ROUGE-N の N が 3 以上のものは評価に用いるのは適さないと考えられる．

表 1　ROUGE の種類ごとの評価値 0 の個数

ROUGE の種類	ROUGE の種類ごとに得られる 1800 個の評価値の内，完全不一致を示す 0 となったものの個数
ROUGE-1	0
ROUGE-2	0
ROUGE-3	18
ROUGE-4	238
ROUGE-5	460
ROUGE-6	932
ROUGE-7	1222
ROUGE-8	1398
ROUGE-9	1612
ROUGE-L	0
ROUGE-W1.2	0
ROUGE-S*	0
ROUGE-S4	0
ROUGE-S9	0
ROUGE-SU*	0
ROUGE-SU4	0
ROUGE-SU9	0

4.1.1　参照用約数を固定した場合の ROUGE の値の安定性の調査結果

図 1 は ROUGE の種類ごと参照要約数ごとの分散値の総和幾何平均である．ROUGE-N の N が高まるにつれ分散値が小さくなっている．ROUGE-1 を除いて，すべての ROUGE で参照要約数が増えると分散値が減少していることがわかる．従って，ROUGE-1 以外の ROUGE は参照要約数が増えると値が安定する．参照要約数が増えるごとに減少幅が減っていて，参照要約数が 3 つのときと 4 つのときの分散値がほぼ等しいため，参照要約数が 3 つのときがほぼ収束値であると考え，参照要約は 3 つ以上あると安定することがわかった．

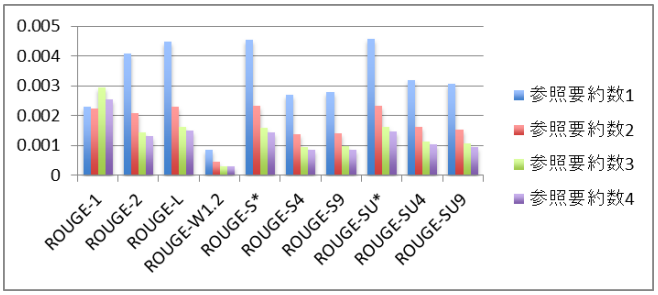

図 1　参照要約数ごとの分散値

4.1.2 手法ⅰ)による順位の安定性の調査結果

表 2 のように，完全に順位が一致したものが大半であった．また完全一致しなかったものも Kendall の一致係数 W が 0.95 もしくは 0.9625 であり，ほぼ一致していた．もし平均化の影響が無いとすれば，参照要約数 1 つで順位が安定することを示す．

表 2　ROUGE の種類ごとの順位の完全一致数

ROUGE の種類	ROUGE の種類ごとに得られる 24 個の評価値の内，完全一致を示す 1 となったものの個数
ROUGE-1	23
ROUGE-2	23
ROUGE-L	24
ROUGE-W1.2	23
ROUGE-S*	24
ROUGE-S4	24
ROUGE-S9	23
ROUGE-SU*	24
ROUGE-SU4	23
ROUGE-SU9	24

4.1.3 手法ⅱ)による順位の安定性の調査結果

図 2 は参照要約数が 1 つの場合と 2 つの場合において，順位がどの程度安定しているかを示す．参照要約数が 3 つの場合は，
図 2 から，ROUGE の種類に関係なく，参照要約数が増えれば順位が安定することがわかった．従って，同時に 4.1.2 の結果からは順位の安定性はわからないことがわかった．

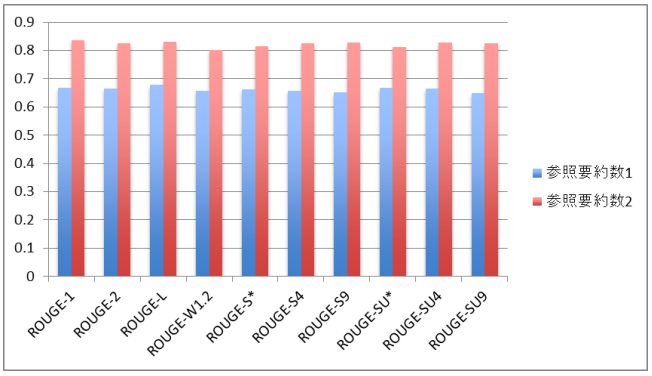

図 2　参照要約数ごとの順位の安定性評価の分散値

4.2　模範解答から人手で分割された述語項構造の同一性判断の調査

本項では，人手で分割された述語項構造の同一性判断をする上でどのような問題が生じるのか事例分析する．

一般知識による推論が必要な例：
1-1)「独仏対立への反省があった」
1-2)「独仏間の対立解消した」
この例では，「対立への反省があった」のであれば「対立解消した」と推論する必要がある．

2-1)「集約的農法が発展した」
2-2)「農業生産力が増大した」
この例では，「農法が発展した」のであれば「農業生産力が増大した」と推論する必要がある．

専門知識による言い換えが必要な例：
3-1)「日本は奉天事件を起こした」
3-2)「日本は張作霖爆殺事件を起こした」
この例は，「奉天事件」と「張作霖爆殺事件」の言い換えが必要である．

4-1)「日本国憲法は非軍事化政策に基づいた」
4-2)「日本国憲法で平和主義が掲げられた」
この例は，「非軍事化政策」と「平和主義」の言い換えが必要である．

専門知識による代表格の推定が必要な例：
5-1)「ドイツでは，英米仏の占領地区とソ連占領地区との分断が進む」
5-2)「ドイツは，西側陣営の西ドイツと，東側陣営の東ドイツに分断された」
この例では「英米仏」が「西側陣営」の代表格であり「ソ連」が「東側陣営」の代表格であることがわかる必要がある．

6-1)「11 世紀頃，西欧では三圃制などが普及した」
6-2)「11 世紀のヨーロッパで新に三圃制が導入された」
この例では「西欧」は「ヨーロッパ」の代表格であることがわかる必要がある．

冗長な表記の例：
7-1)「ユダヤ人がイスラエルを建国した」
7-2)「イスラエルを建国した」
この例では，「ユダヤ人が」が書かれているか否かが異なるが，イスラエルを建国したのがユダヤ人以外に候補が無ければ，同一とみなせる．

問題文の指示により解答中で語を省略している例：
8)「戦後の安全保障構築の試みがあった」
この例の「戦後」は，論述問題の問題文中に「第二次世界大戦中に生じた出来事が､いかなる形で 1950 年代までの世界のありかたに影響を与えたのか」とあるため，「第二次世界大戦後」を指しており，その省略表現である．

5. まとめ

論述問題の評価に模範解答を参照要約とみなせば，参照要約を用いた評価手法を適用することができると考えられる．参照要約を用いた評価手法を適用するにあたり，本稿では，以下の 2 点に関する調査を行った．
A) 模範解答（参照要約）の数による評価の安定性
B) 模範解答（参照要約）との一致率を計算する単位

A)に関しては，ROUGE-1 を除く ROUGE の評価値は，参照要約数を増やした方が安定し，ROUGE の評価値による順位は，ROUGE-1 も含め，参照要約数を増やした方が安定することがわかった．ROUGE-1 を除く ROUGE の評価値は，参照要約は 3 つ以上あると安定することがわかった．ROUGE-N の N が 3 以上のものは評価に用いるのは適さないことがわかった．

B)に関しては，述語項構造の同一性判断をする上で，一般知識や専門知識による言い換え，代表格の推定，冗長性の判定，問題文の文脈を見なければ判断できないものがあることがわかった．

謝辞 模範解答の作製に関して，株式会社アイアール・アルトの渡部恵理子氏と解答作製者 3 名の皆様に大変お世話になりました．感謝致します．

参考文献

1) Hideyuki Shibuki, Kotaro Sakamoto, Yoshionobu Kano, Teruko Mitamura, Madoka Ishioroshi, Kelly Y. Itakura, Di Wang , Tatsunori Mori, Noriko Kando, Overview of the NTCIR-11 QA-Lab Task, Proceedings of the 11th NTCIR Conference (2014)
2) CY Lin , Rouge: A package for automatic evaluation of summaries, Text summarization branches out: Proceedings of the ACL-04 workshop 8 (2004)
3) Nenkova, A., Passonneau, R., and McKeown, K, The Pyramid Method: Incorporating Human Content Selection Variation in Summarization Evaluation, ACM Transactions on Speech and Language Processing, Vol. 4, Issue 2 (2007)
4) Makoto P. Kato, Matthew Ekstrand-Abueg , Virgil Pavlu , Tetsuya Sakai, Takehiro Yamamoto, Mayu Iwata, Overview of the NTCIR-11 MobileClick Task, Proceedings of the 11th NTCIR Conference (2014)
5) 林部祐太，小町守，松本裕治，述語と項の位置関係ごとの候補比較による日本語述語項構造解析，自然言語処理, Vol.21, No.1, pp.3-26. (2014)