NTCIR-13 QA Lab-3における大論述問題の事例分析

渋木 英潔 [†1]　阪本 浩太郎 [†1†2]　石下 円香 [†2]
狩野 芳伸 [†3]　三田村 照子 [†4]　森 辰則 [†1]　神門 典子 [†2†5]

[†1]横浜国立大学　[†2]国立情報学研究所　[†4]静岡大学
[†3]カーネギーメロン大学　[†5]総合研究大学院大学

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、NTCIR ワークショップにおいて QA Lab タスクをこれまで 3 回開催している [1, 2, 3]。QA Lab では周囲の文脈理解や高度な推論を要する現実世界の問題の一つとして世界史の大学入試問題を用いたタスクを設定しており、国内外の研究者が多く参加している．大学入試問題には多肢選択問題や語句記述問題など多様な出題形式があるが、中でも論述問題が非常に困難な課題であることが QA Lab-2 の結果[1]からも伺えた。そのため、QA Lab-3 では特に論述問題に焦点を当てたタスクを設定した。

論述問題は、解答の制限字数の観点から数百字前後の大論述問題と百字前後の小論述問題に分類できる。大論述問題の例を図 1 に示す。大論述問題はその長さから複数のサブトピックを含んでおり、それらのサブトピックをどのような構成で記述するかについて考慮する必要がある。QA Lab-2 では論述問題を、問題文のトピックに合わせて教科書などの知識源を要約する query-biased summarization[4] の一種とみなし、要約の評価手法である ROUGE[5] や Pyramid 法 [6, 7] を用いて評価していた[2]。ROUGE や Pyramid 法は参照要約（模範解答）との内容の一致度を測る指標であり、構成的な部分に関する評価は行われない。それゆえ、QA Lab-3 では、QA Lab-2 までの評価に加えて、内容以外の観点からの quality questions による評価を行った。本稿では、QA Lab-3 に提出された大論述問題を対象に quality questions による評価を中心とした事例分析を行い、現状と課題を示す。

近年，13～14 世紀を「モンゴル時代」ととらえる見方が提唱されている。それは，「大航海時代」に先立つこの時代に，モンゴル帝国がユーラシア大陸の大半を統合したことによって，広域にわたる交通・商業ネットワークが形成され，人・モノ・カネ・情報がさかんに行きかうようになったことを重視した考え方である。そのような広域交流は，帝国の領域をこえて南シナ海・インド洋や地中海方面にも広がり，西アジア・北アフリカやヨーロッパまでをも結びつけた。

以上のことを踏まえて，この時代に，東は日本列島から西はヨーロッパにいたる広域において見られた交流の諸相について，経済的および文化的（宗教を含む）側面に焦点を当てて論じなさい。解答は，解答欄（イ）に 20 行以内で記述し，必ず次の 8 つの語句を一度は用いて，その語句に下線を付しなさい。なお，（　）で並記した語句は，どちらを用いてもよい。

ジャムチ	授時暦	染付（染付磁器）
ダウ船	東方貿易	博多
ペスト（黒死病）	モンテ＝コルヴィノ	

図 1: 大論述問題の例

2　NTCIR-13 QA Lab-3

QA Lab-3 では、論述問題の対象として、東京大学の第 2 次学力試験試験問題（世界史）の 2000 年から 2014 年までの 15 年分を用いた。大論述に該当する問題は毎年 1 問出題されるので 15 問存在し、その内、5 問をトレーニングデータ[3]とし、10 問をテストデータ[4]とした。入試問題の記述をそのまま用いた日本語タスクとそれを英訳した英語タスクがあり、日本語の論述問題タスクに参加したチーム数は延べ 8 チーム、チームごとに複数のシステムによる提出を許可したため、提出された解答ファイル数は 23 ファイルであった[5]。提出された解答ファイルの中には大論述の部分を白紙解答にしていたものもあったため、最終的な大論述問題の解答数は 458 となった。本稿では、この 458 解答を対象に分析を行う。

[1]二次試験には 2 チームからの 6 解答しか提出されなかった。
[2]総合的な評価として、人間の専門家による採点も行っている。
[3]2003、2005、2007、2009、2011 の 5 年分
[4]テストランは 2017 年 2 月と 5 月の 2 回のフェイズで行われた。フェイズ 1 は 2000、2004、2008、2012、2013 の 5 年分、フェイズ 2 は 2001、2002、2006、2010、2014 の 5 年分を用いた。
[5]QA Lab-3 のタスクの詳細は文献 [3] を参照されたい。

3 評価指標

3.1 専門家による採点

同一の評価基準[6]を用いて、2 人の専門家により採点をしてもらった。予算の都合により全ての解答を採点することできず、各チームが自分たちの解答につけた優先順位に従って採点する解答を選別した[7]。

3.2 Pyramid 法

Pyramid 法で用いる nugget は 3 名の専門家により個別に作成され、作成された nugget に対して 0（全く無関係）から 3（非常に重要）の整数値で 3 名の専門家に投票してもらった。結果として、各 nugget には、その重要度を表す 1[8]–9 の重みが付いている。各解答のスコアは、解答中に含まれる nugget の重みの総和で計算される。

3.3 quality questions

QA Lab-2 の時の解答を分析した結果を踏まえて、DUC[9]や TAC の Guided Summarization タスク[10]における quality questions を参考とした結果、QA Lab-3 では、以下の 5 項目に対する quality questions を用いた。

- **gramaticality** 助詞の使い方や自動詞/他動詞の区別などの文法的な観点からの品質。
- **non-redundancy** 同じような内容が繰り返されていないかなどの冗長性の観点からの品質。
- **reference clarity** 「いつ、どこで、誰が、何をしたか」が明確かなどの参照先の明瞭性の観点からの品質。
- **fluency** 一文が長すぎる/短すぎるや、句読点が過剰/皆無などの可読性の観点からの品質。
- **coherence and content structure** 因果関係などの意味的なものを含めた一貫性や文章構造の適切性の観点からの品質。評価者が専門知識を得るために教科書や用語集などを利用してもよいこととした。

各項目は、以下の A(excellent)–D(poor) の 4 段階で評価した。

A. 問題なし。スラスラ読める。

B. 比較的容易に理解（修正）可能な誤りが一部にみられる。

C. 多くに誤りがみられる。または、理解するのにかなりの労力を要する誤りがある。

D. 理解不能。まったく読めない。

参加者は上記のガイドラインに従って、他の参加者が提出した解答を評価した。評価の際には、内容の真偽は問わないことを強調し、例え、内容的に虚偽の記述であっても文法的に問題なければ grammaticality に A をつけるよう指示した。

4 事例分析

表 1 と表 2 に、2 名の専門家による採点と Pyramid 法によるスコアの分布をそれぞれ示す。ほとんどの解答が 0 点と採点される結果となった。したがって、まだまだ改善の余地が残されている状態だといえる。

表 3 に、quality question における参加者による評点の分布を示す。未評価の解答もあったため、Total に評価がつけられた解答の総数を示す。また、A を 4 点、B を 3 点、C を 2 点、D を 1 点として Total で割った平均値を Ave. に示す。Ave. の値は、non-redundancy(3.73)、grammaticality(3.64)、fluency(3.46)、coherence and content structure(2.66)、reference clarity(2.63) の順となり、特に reference clarity と coherent and content structure が低かったことが分かる。また、表中の分布から D と評価された数が増加しており、対処の度合いが不十分というよりも、全く対処できていない解答が多かったことを示している。

grammaticality が高い値であったのは、QA Lab-3 に参加したシステムが教科書等からの文抽出を基本としたものであったためと考えられる。grammaticality が D と評価された解答の例を以下に示す。

gramaticality が D の例

国際連盟脱退は日本は、連盟総会で基づく満州撤兵などの勧告案が、42 対 1 で採択されたのを不満として脱退した。

「日本は」以降の記述は国際連盟脱退の説明であるが、それ単独では何についての説明であるか読み手に伝わらない。それゆえ、「国際連盟脱退は」という記述を文頭に追加したと思われるが、結果として非文となって

[6]評価基準は「○○について書かれていれば+α 点」、「前後のつながりが不明瞭なら-β 点」のように加点や減点のポイントが端的に書かれている。

[7]2 人の専門家が採点した解答集合の関係は他方を完全に包含している関係ではなく一部が重複している関係である。

[8]作成者が 0（全く無関係）を投票することはありえないため最低値は 1 となる。

[9]http://duc.nist.gov/duc2007/tasks.html

[10]http://www.nist.gov/tac/2011/Summarization/Guided-Summ.2011.guidelines.html

表 1: 専門家の採点結果

	0	1–10	11–20	21–30	31–	Total
専門家 A	56 (84.8%)	5 (7.6%)	4 (6.1%)	0 (0.0%)	1 (1.5%)	66
専門家 B	21 (67.7%)	9 (29.0%)	1 (3.2%)	0 (0.0%)	0 (0.0%)	31

表 2: Pyramid スコアの結果

0	1–50	51–100	101–150	151–200	201–	Total
496 (69.6%)	146 (20.5%)	24 (3.4%)	25 (3.5%)	15 (2.1%)	7 (1.0%)	713

いる。文抽出後の加工に関してまだ改善の余地があると思われる。

non-redundancy も高い値であり、サブトピックごとに 1 文を選択するなどの方法で、冗長的な記述を抑制することができていると考えられる。non-redundancy が D と評価された解答の例を以下に示す。

non-redundancy が D の例

（略）カイロの中心は、ナイル川東岸にほど近いタハリール広場である。その北東にはオスマン時代には湖であったところをムハンマド・アリー朝時代に埋め立てて作ったイズベキーヤ公園がある。また、タハリール広場から東に 400m ほどのところにムハンマド・アリー朝の王宮であり、現大統領府であるアブディーン宮殿がある。タハリール広場から南のナイル川沿いはガーデン・シティと呼ばれ、イギリス統治時代にエジプト総督府がおかれ開発が進められたエリアである。このタハリール広場を中心とした地域は新市街と呼ばれ、19 世紀のムハンマド・アリー朝の時代に都市開発がすすめられた地域で、現在でもカイロの中心である。

「タハリール広場」や「ムハンマド・アリー朝」といった重要語句を含む文集合を書いた結果、同じような内容の繰り返しになったものと思われる。しかしながら、それぞれの文には「イズベキーヤ公園」や「アブディーン宮殿」など特有の記述も含まれており、単純に前後の文に包含されているというわけではない。これは、単純な文抽出によるアプローチの限界を示していると考えられ、複数の文を 1 文にまとめるような処理が必要であると考えられる。

fluency が D と評価された解答の例を以下に示す。

fluency が D の例

（略）19 世紀のキリスト教の特徴は、プロテスタントの大部分の福音派の復活であったそして後に現代の聖書奨学金が教会に及ぼす影響。

「復活であった」で終わる文と「そして」で始まる文が連続して書かれたものと思われる。しかしながら、前者の文には句点がなく、後者の文は体言止めであったため、結果として非常に可読性が低い記述となっている。また、fluency が D の場合、他の指標も D となることが多く、この例では、fluency の他に grammaticality と coherence and content structure も D であった。句点の有無や体言止めか否かなど、比較的容易に判断可能かつ細かな点ではあるが、こういった処理を怠ると様々な点で悪影響が表れるといえる。

reference clarity が D と評価された解答の例を以下に示す。

reference clarity が D の例

アヘン戦争の原因をつくりだすことになる。さらにスペインやポルトガルの貿易独占に対抗するため、ブラジルではサトウキビ=プランテーションを経営し、加わった。孫文は、病死した。禁止 (海禁) された。（略）

問題が要求する内容を含まない句や節などを不要と判断して削除することで文内要約を行った結果、その文単独では何を意味しているか不明瞭になったものと思われる。したがって、問題の要求だけではなく、通常の文として成立するかどうかも考慮して削除を行う必要があると考えられる。

coherent and content structure が D と評価された解答の例を以下に示す。

表 3: quality question の結果

	A		B		C		D		Total	Ave.
grammaticality	587	(83.1%)	39	(5.5%)	26	(3.7%)	54	(7.6%)	706	3.64
non-redundancy	576	(81.7%)	89	(12.6%)	18	(2.6%)	22	(3.1%)	705	3.73
reference clarity	277	(39.3%)	102	(14.5%)	116	(16.5%)	210	(29.8%)	705	2.63
fluency	515	(73.0%)	73	(10.4%)	44	(6.2%)	73	(10.4%)	705	3.46
coherence etc.	268	(38.2%)	150	(21.4%)	57	(8.1%)	226	(32.2%)	701	2.66

coherent and content structure が D の例

その後、カール 5 世によって、ネーデルラント 17 州がハプスブルク家の支配下に統合される。ニューヨークやボストンの港は奴隷貿易港として栄えるようになった。南アフリカ戦争、ブール戦争ともいう。ヨーロッパ統合は経済・市場統合にとどまらず，さらに共通の外交・防衛政策を採用して政治統合をめざすヨーロッパ連合条約（マーストリヒト条約）が１９９３年に発効され，ヨーロッパ連合（ＥＵ）が成立した。（略）

抽出された文を時間順に並べているが、文の前後で場所的にも内容的にもつながりがないため論旨の一貫性が感じられない。このことは抽出された文を時間順に並べるだけでは、適切な論述構成とはいえないことを示している。

5 まとめ

本稿では、NTCIR-13 QA Lab-3 の論述問題タスクに提出された 458 解答を用いた事例分析を行った。QA Lab-3 では内容一致の観点に加えて、grammaticality、non-redundancy、reference clarity、fluency、coherence and content structure の 5 つの観点からの quality questions による評価を行った。その結果、現在の解答は reference clarity と coherence and content structure の品質に問題があるものが多かったことが分かった。

本稿で述べた課題は世界史の大学入試問題に特有のものではなく、自動要約や長文生成などにも適用される一般的なものであり、課題を解決することでそういった技術の発展につなげていきたいと考えている。今後、さらに分析を進める予定である。

謝辞

本研究で用いた専門家による評価の一部を提供していただいた大学入試センター石岡恒憲教授に深く感謝いたします。

参考文献

[1] Shibuki, H., Sakamoto, H., Kano, Y., Mitamura, T., Ishioroshi, M., Itakura, K.Y., Wang, D., Mori, T. and Kando, N.: Overview of the NTCIR-11 QA-Lab Task, *Proc. the NTCIR-11 Conference*, pp.518–529 (2014).

[2] Shibuki, H., Sakamoto, H., Ishioroshi, M., Fujita, A., Kano, Y., Mitamura, T., Mori, T. and Kando, N.: Overview of the NTCIR-12 QA Lab-2 Task, *Proc. the NTCIR-12 Conference*, pp.392–408 (2016).

[3] Shibuki, H., Sakamoto, H., Ishioroshi, M., Kano, Y., Mitamura, T., Mori, T. and Kando, N.: Overview of the NTCIR-13 QA Lab-3 Task, *Proc. the NTCIR-13 Conference*, pp.112–128 (2017).

[4] Tombros, A. and Sanderson, M.: Advantages of Query Biased Summaries in Information Retrieval, *Proc. of the 21st Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, 2-10, (1998).

[5] Lin, C.Y.: ROUGE: A Package for Automatic Evaluation of Summaries, *Proc. of Workshop on Text Summarization Branches Out*, 74–81, (2004).

[6] Nenkova, A. and Passonneau, R.J.: Evaluating Content Selection in Summarization: The Pyramid Method, *Proc. of the Human Language Technology Conference of the North American Chapter of the Association for Computational Linguistics*, 145–152, (2004).

[7] Passonneau, R.J., Chen, E., Guo, W. and Perin, D.: Automated Pyramid Scoring of Summaries using Distributional Semantics, *Proc. of the 51st Annual Meeting of the Association for Computational Linguistics*, 143–147, (2013).