Vol.2016-NL-227 No.13
2016/7/30

NTCIR-12 QA Lab-2におけるQAシステムの課題 — センター試験の結果を中心として

渋木 英潔[1,a)] 石下 円香[2] 阪本 浩太郎[1,2] 藤田 彬[2] 狩野 芳伸[3] 三田村 照子[4] 森 辰則[1]
神門 典子[2,5]

概要：本稿ではNTCIRワークショップのQA Labタスクに参加したシステムの結果を元にした誤り分析を行う．参加システムの30%以下しか正答できなかった問題を難問と定義し，難問となる質問形式はどのような傾向がみられるのか調査する．また，赤本を用いて人間にとっての難易度と関係があるのか比較する．さらに，誤文と判断するための手がかりの場所と誤文と判断される原因の2通りの観点から正誤判断問題の誤りを類型化し，難問にどのように出現するのか調査する．

キーワード：質問応答システム, 誤り分析, QA Lab, NTCIR

Issues of QA Systems in the NTCIR-12 QA Lab-2 task — With a Central Focus on the Results of the Center Tests

Abstract: Using the Center Test results in the NTCIR QA Lab tasks, error analysis is described in this paper. We defined "difficulty" of questions as the number of systems that could answer correctly. We reported what question formats frequently appeared in the difficult questions and whether the difficulty was related to human beings in comparison with Akahon. For difficult true-or-false questions, we also classified the falsehoods in terms of clue description places and cause types, and reported the occurrence.

Keywords: QA system, error analysis, QA Lab, NTCIR

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2回開催している[1], [2]．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．2015年に行われた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[3]と連携して，全国の大学受験生と共に大手予備校が行う模擬試験に挑戦し，センター模試の世界史Bにおいて76/100点（偏差値66.5）という好成績を収めるシステム[4]も現れた[*1]．これは参加システムがそれぞれに改良を重ねてきた結果といえる．しかしながら，参加システム全体を俯瞰して，現在のQAシステムの多くが不得手とする問題がどのようなものなのか，それらの問題を解くにはどのような知識や技術が必要とされるのか，またどのように類型化されるのか，といった分析はこれまで十分に行われていない．そして，これはタスクオーガナイザである我々が行うべき課題である．

QA Labでは，分野こそ世界史に限定しているものの，センター試験だけでなく，東京大学，京都大学，北海道大

[1] 横浜国立大学
Yokohama National University, Chiyoda, Tokyo 101–0062, Japan
[2] 国立情報学研究所
IPSJ, Chiyoda, Tokyo 101–0062, Japan
[3] 静岡大学
Shizuoka University, Chiyoda, Tokyo 101–0062, Japan
[4] カーネギーメロン大学
IPSJ, Chiyoda, Tokyo 101–0062, Japan
[5] 総合研究大学院大学
SOKENDAI, Chiyoda, Tokyo 101–0062, Japan
[a)] shib@forest.eis.ynu.ac.jp

[*1] 多肢選択形式のセンター模試と比較して，自由記述形式の東大模試の世界史では16/60点（偏差値48.7）が最高点であり，より多くの課題が残されているといえる．

表 1　各 Phase で用いた試験
Table 1　Exams used in each phase

		センター試験	二次試験	模擬試験
QA Lab-1	Phase-1	JA&EN	—	—
	Phase-2	JA&EN	JA&EN	—
QA Lab-2	Phase-1	JA&EN	JA&EN	JA
	Phase-2	—	—	JA
	Phase-3	JA&EN	JA&EN	JA

学，早稲田大学，中央大学の二次試験や大手予備校の模擬試験を用いて幅広い形式の質問を対象としている．また，海外の研究者が参加できるようセンター試験と二次試験に関しては英訳したものを用意し，日本語 (JA) と英語 (EN) の 2 種類のサブタスクを設定している．それぞれのサブタスクは，Phase-1，Phase-2 といった名称で複数の run を行っており，参加者は任意の run に参加することが可能である．さらに，各 Phase においても，センター試験のみに解答する，東京大学の二次試験のみに解答するといった自由を参加者に許可している．表 1 に各 Phase で行われた試験の種類と日英のサブタスクを示す．

QA システムから見たセンター試験と二次試験における最大の違いは，解候補が選択肢として示されている多肢選択形式か，そうではない自由記述形式かという点であり，解候補の抽出をしなくてはならない自由記述形式の方が難しい．QA Lab-2 の Phase-3 では，センター試験に 12 チームから 34 解答の提出があったのに対し，二次試験には 2 チームから 6 解答しか提出されなかった．全体を俯瞰するための十分な量があるとは言い難いことから，本稿では二次試験を調査対象から除外する．また，多肢選択形式の模擬試験についても日本語サブタスクに限定されることから除外する．したがって，QA Lab-1 の Phase-1 と Phase-2，QA Lab-2 の Phase-1 と Phase-3 に提出されたセンター試験の結果に基づいて調査を行う．なお，本稿では，途中経過報告として統計データと基礎的な考察を行う．

2. 関連研究

Project next NLP[*2] では NLP の様々な分野でのエラー分析が行われた．その中で，機械翻訳では，訳出の導出過程を考慮せず出力結果のみに着目するブラックボックス分析が行われた [5]．QA Lab では質問文解析や文書検索の結果などの中間結果の提出は任意 [*3] であるため，タスクオーガナイザの視点から全システムの導出過程を把握することは困難である．それゆえ，本稿ではブラックボックス分析を行うこととする．

QA システムにおけるエラー分析には以下の研究がある．松崎ら [6] は「ロボットは東大に入れるか」プロジェクトにおけるエラー分析を行っており，世界史のエラー分析を行っている．しかしながら，狩野 [7] のシステムに焦点を当てた分析であり，全体を俯瞰する観点では行っていない．Moldovan et al.[8] も，多くのシステムに共通するモジュールを意識して設計されたものではあるが特定のシステムに焦点を当てた分析である．

[*2] https://sites.google.com/site/projectnextnlp/
[*3] QA Lab では標準的な QA システムを想定した中間結果を設定しているが，それに縛られない多様なアプローチを認めている．それゆえ，中間結果の提出は任意とした．

3. 難問の特徴

現在の QA システムの多くが不得手とする問題を調査するために，システムが正答することが困難な問題（以降，難問）を以下のように正答システム率を用いて定義する．質問 q における正答システム率 CSR を以下の式で計算する．

$$CSR(q) = \frac{CSN(q)}{PSN(q)} \quad (1)$$

ここで PSN は問題に解答したシステム数，CSN は正解を解答したシステム数である．本稿では，正答システム率が 30%以下だった問題を難問と定義する．また，難問数を問題数で割った値を難問率と定義する．

表 2 に各 Phase で用いたセンター試験の年度，問題数，参加システム数，難問数，難問率をそれぞれ示す．Phase によって多少のばらつきはあるものの難問率は 4 割を下回ることがなく，回を重ねることで参加システムの質が全体的に向上しているとは言い難い．注意すべき点として，4 割強という難問率は「一部の優秀な」システムに対して「多数の平凡な」システムが足を引っ張った結果では決してない．例えば，模擬試験で好成績を収めたシステム [4] であっても，QA Lab-2 の Phase-3 において 16 の難問中 8 問しか正答できていない [*4]．したがって，難問を分析することは優秀なシステムにとっても必要なことである．

解答方法が多肢選択形式のセンター試験であるが，質問形式は，穴埋め形式や文の正誤判定，時系列順に並び替えるものや地図中の場所を尋ねるものなど多様である．文献 [1], [2] の分類に倣って，本稿では，**語句**，**穴埋め**，**正誤判断**，**時系列**，**図表参照**の 5 種類の質問形式に分類する．表 3 に各 Phase において質問形式に該当する問題数を「全体」の列に，その中で難問に該当する問題数を「難問」の列にそれぞれ示す．括弧内の値は各列の合計を 1 とした時の質問形式が占める割合である．Phase によって質問形式の割合に大きな変化はないことが分かる．難問の半分程度を正誤判断が占めているが，全体においても 6 割以上を正誤判断が占めているため，むしろ正誤判断以外の質問形式の方が不得手とするシステムが多かったと言える．事実，QA Lab に参加したシステムは正誤判断に焦点を当てて設計されているものが多かった．しかしながら，正誤判断に焦点を当てたシステムが多いという事実を考慮すると，難

[*4] 文献 [2] の表 23 を参照．

表 2　センター試験における難問の割合
Table 2　The ratio of difficult questions in the Center Tests

		年度	問題数	システム数	難問数	難問率
QA Lab-1	Phase-1	2007	35*	15	14	0.400
	Phase-2	2003	41	18	19	0.463
QA Lab-2	Phase-1	1999	41	27	17	0.415
	Phase-3	2011	36	34	16	0.444

*実際の出題数は 36 問であるが XML 化の際に不適切なタグが付与された 1 問を評価から除外している

表 3　質問形式の割合
Table 3　The ratio of question formats

	QA Lab-1								QA Lab-2							
	Phase-1（2007 年度）				Phase-2（2003 年度）				Phase-1（1999 年度）				Phase-3（2011 年度）			
	全体		難問		全体		難問		全体		難問		全体		難問	
語句	5	(0.143)	3	(0.214)	1	(0.024)	0	(0.000)	8	(0.195)	4	(0.235)	0	(0.000)	0	(0.000)
穴埋め	4	(0.114)	1	(0.071)	4	(0.098)	2	(0.105)	2	(0.049)	2	(0.118)	3	(0.083)	1	(0.063)
正誤判断	24	(0.686)	8	(0.571)	30	(0.732)	12	(0.632)	28	(0.683)	8	(0.471)	29	(0.806)	11	(0.688)
時系列	1	(0.029)	1	(0.071)	1	(0.024)	1	(0.053)	0	(0.000)	0	(0.000)	3	(0.083)	3	(0.188)
図表参照	1	(0.029)	1	(0.071)	5	(0.122)	4	(0.211)	3	(0.073)	3	(0.176)	1	(0.028)	1	(0.063)
合計	35	(1.000)	14	(1.000)	41	(1.000)	19	(1.000)	41	(1.000)	17	(1.000)	36	(1.000)	16	(1.000)

問の半分程度が正誤判断であるという結果は設計目標を十分に達成できているとは言い難い．5 節では正誤判断問題を対象として誤りの類型化を行う．

4. 人間にとっての難易度との比較

大学入試問題を対象とすることには，
(1) 周辺文脈の理解や推論など比較的高度な知的活動が要求されるので技術の発展を促す．
(2) 模範解答が存在するのでシステムの評価がしやすい．
(3) 参考書の解説などに解答までの道筋が示されているのでアルゴリズムの参考にできる．
といった利点が存在する．本稿では，(3) の利点として，大学入試過去問集である赤本 [*5] を用いた分析を行う．

最初に注意すべき点として，赤本の解説はあくまでも人間を対象としたものであり，システムの処理や評価とは必ずしも一致するわけではない．そのため，システムにとっての難問が人間にとっても難問であるか調査を行う．2008 年度以前のセンター試験には，赤本に受験生のための難易度が示されており，対象データの範囲では，**易**，**標準**，**やや難**が存在している．表 4 に Phase ごとの問題全体と難問における難易度の割合を示す．割合で比較した場合，全体と難問との間に大きな差はない．したがって，システムにとっての難問は，人間にとっての難度とは異なる観点で定義されると考えられる．

5. 誤りの類型化

本稿では，正誤判断の難問を対象に，
(A) 誤文であることを判断するための手がかりが記述されている場所
(B) 誤文と判断される原因
の 2 通りの観点から誤りの類型化を試みる．(A) はシステムが処理しなくてはならない範囲，入力データに関連し，(B) はシステムが行わなくてはならない処理の種類に関連する．

5.1 手がかりの場所

まず，誤文であることを判断するための手がかりが記述されている場所を，**選択肢**，**質問文**，**参照先**の 3 種類に分類する．以下に，各場所に該当する具体例を挙げる．各質問のタイトルは，入試問題の名称_解答欄 ID，質問形式を表しており，括弧内の数字は，正答システム率となる (正解のシステム数)/(サブタスクに参加した全システム数) を示している．

[*5] http://akahon.net/

表 4　赤本による難易度の割合
Table 4　The ratio of the difficulty estimated in Akahon

	QA Lab-1								QA Lab-2*			
	Phase-1（2007 年度）				Phase-2（2003 年度）				Phase-1（1999 年度）			
	全体		難問		全体		難問		全体		難問	
易	9	(0.257)	4	(0.286)	5	(0.122)	2	(0.105)	7	(0.171)	4	(0.235)
標準	20	(0.571)	7	(0.500)	26	(0.634)	12	(0.632)	19	(0.463)	8	(0.471)
やや難	6	(0.171)	3	(0.214)	10	(0.244)	5	(0.263)	15	(0.366)	5	(0.294)
合計	35	(1.000)	14	(1.000)	41	(1.000)	19	(1.000)	41	(1.000)	17	(1.000)

*赤本には QA Lab-2 Phase-3（2011 年度）の難易度は示されていない．

5.1.1　選択肢

2003CT_A6 正誤判断（1/14）

下線部 (6) について述べた文として誤っているものを，次の 1〜4 のうちから一つ選べ．
1. 国際連合は，パレスティナを分割する案を採択した．(解答システム数：11)
2. パレスティナ国家の建設を目指して，パレスティナ解放機構 (PLO) が活動した．(解答システム数：1)
3. ナセル大統領の下で，エジプトはイスラエルと平和条約を結んだ．(解答システム数：1)
4. パレスティナ解放機構 (PLO) とイスラエルは，パレスティナ暫定自治協定に調印した．(解答システム数：1)
＜正解：3＞

3 番が誤りなのは，エジプト-イスラエル平和条約を結んだ時のエジプト大統領が「ナセル」ではなく「サダト」だからである．そして，この情報は 3 番の選択肢だけから得ることができる．

5.1.2　質問文

2007CT_A18 正誤判断（1/15）

同じく下線部 (8) に関連して，第二次世界大戦後の植民地の独立について述べた次の文 a と b の正誤の組合せとして正しいものを，下の 1〜4 のうちから一つ選べ．
a. イラクが独立した．（正答システム数：5，誤答システム数：10）
b. モザンビークが独立した．（正答システム数：10，誤答システム数：5）
1. a-正 b-正　2. a-正 b-誤
3. a-誤 b-正　4. a-誤 b-誤
＜正解：3＞

文 a が誤りなのは，イラクが独立したのが 1932 年の第二次世界大戦以前だからである．文 a の「イラクが独立した」という記述自体に誤りはなく，質問文中の「第二次世界大戦後」という条件を考慮してはじめて誤文と判断できる．

5.1.3　参照先

2007CT_A11 正誤判断（2/15）

下線部 (2) の皇帝の事績として正しいものを，次の 1〜4 のうちから一つ選べ．
1. 洛陽に遷都し，内政を重視した．(解答システム数：5)
2. 塩・鉄などを非売品とし，また物価の調整と安定に努めた．(解答システム数：2)
3. 文字・度量衡・貨幣を統一した．(解答システム数：7)
4. 郡県制を敷き，三省・六部を設けた．(解答システム数：1)
＜正解：2＞

下線部 (2) は「漢の武帝」を指しており，漢の武帝の事績として正しいのは 2 番である．解答システム数が多かった 1 番と 3 番はそれぞれ「漢の光武帝」と「秦の始皇帝」の事績であり，質問文の「皇帝の事績として正しいもの」に該当する．したがって，正しく判断するためには下線部 (2) の参照先が必要である．

5.2　誤りの原因

次に，誤文と判断される原因を，**時間情報**，**場所情報**，**同意語**，**反意**，**交換**の 5 種類に分類する．以下に，各原因に該当する具体例を挙げる．

5.2.1 時間情報

2011CT_A5 正誤判断（10/34）

下線部 (5) に関連して，伝達の技術や道具について述べた次の文 L4 と L5 の正誤の組合せとして正しいものを，下の 1〜4 のうちから一つ選べ．
a. グーテンベルクが 13 世紀に，活版印刷術を改良・実用化した．(正答システム数：19，誤答システム数：15)
b. 秦の蔡倫が，製紙法を改良した．（正答システム数：15，誤答システム数：19）
1. a-正 b-正　2. a-正 b-誤
3. a-誤 b-正　4. a-誤 b-誤
＜正解：4＞

文 a の誤りは，グーテンベルグが活版印刷術を改良・実用化したのが「13 世紀」ではなく「15 世紀半ば」であるという時間情報のずれが原因である．また，文 b の誤りも，蔡倫が製紙法を改良したのが「秦」の時代ではなく「後漢」の時代であることが原因である．本稿では，文 b のように，時代を特定できる「秦」や「後漢」といった語句が誤り原因である場合も「時間情報」に分類することとした．

5.2.2 場所情報

1999CT_A30 正誤判断（8/27）

下線部 (10) に関連して，18 世紀のプロイセンの官僚制度を支えた階層とその役割について述べた文として正しいものを，次の 1〜4 のうちから一つ選べ．
1. ブルジョワジーが上級官職を独占し，国政を担った．(解答システム数：5)
2. ユンカーを中心に，軍事色の濃い国家が築かれた．(解答システム数：8)
3. 地方行政を行う上で，ジェントリが重要な役割を果たした．(解答システム数：10)
4. 治安維持と領土の拡大を行う上で，コサックが大きな役割を果たした．(解答システム数：3)
（未解答システム数：1）
＜正解：2＞

1 番の「ブルジョワジーが国政を担った」のは 18 世紀の「プロイセン」ではなく，19 世紀の「フランス」などである．同様に，3 番のジェントリは「イギリス」，4 番のコサックは「ロシア」である．「場所情報」と「時間情報」の両方にずれが生じている場合，どちらに分類するか難しい問題であるが，社会的階層など（質問の焦点となっている時間情報を含んで）比較的長い期間にわたって存在していたものに関しては「場所情報」に分類することとした．

5.2.3 同位語

2011CT_A10 正誤判断（10/34）

下線部 (1) について述べた次の文 a と b の正誤の組合せとして正しいものを，下の 1〜4 のうちから一つ選べ．
a.『アヴェスター』は，マニ教の経典である．(正答システム数：17，誤答システム数：17)
b. ウラディミル 1 世は，ギリシア正教を国教とした．(正答システム数：24，誤答システム数：10)
1. a-正 b-正　2. a-正 b-誤
3. a-誤 b-正　4. a-誤 b-誤
＜正解：3＞

文 a が誤りなのは，『アヴェスター』は「マニ教」ではなく「ゾロアスター教」の経典だからである．このように共通のカテゴリー（この例では「宗教」）に属する別の語に置き換わることで誤文となるものを**同位語**と定義する．ただ，「同位語」と「時間情報」や「場所情報」との線引きは難しい．例えば，5.2.2 の「場所情報」の例で「プロイセン」が「国名」という共通カテゴリーに属する「フランス」に置き換わっているとみなすこともできるからである．本稿では，共通カテゴリーが時間や場所を象徴するものではない場合に「同位語」に分類することとした．

5.2.4 反意

2011CT_A24 正誤判断（4/34）

下線部 (5) に関連して，19・20 世紀の農業や農民について述べた文として誤っているものを，次の 1〜4 のうちから一つ選べ．
1. プロイセンで，シュタイン・ハルデンベルクらの改革によって農民解放 (農奴解放) が行われた．(解答システム数：8)
2. ロシアで，ストルイピンがミール (農村共同体) を保護しようとした．(解答システム数：4)
3. アメリカ合衆国で，ホームステッド法 (自営農地法) が制定され，西部開拓が促された．(解答システム数：11)
4. イギリスでは，第 3 回選挙法改正によって，農業労働者に選挙権が与えられた．(解答システム数：11)
＜正解：2＞

2 番が誤りなのは，ストルイピンはミールを「保護しようとした」のではなく「解体を図った」ためである．このように述部が反対の意味になることで誤文となるものを**反意**と定義する．

「反意」が誤り原因の文をシステムが判断することを考える．「時間情報」，「場所情報」，「同位語」では固有表現などの特徴語を認識するだけでも比較的対処可能であると

表 5　正誤判断の難問における手がかりの場所
Table 5　Clue descriptions in the difficult True-or-False questions

	QA Lab-1				QA Lab-2			
	Phase-1		Phase-2		Phase-1		Phase-3	
選択肢	9	(0.500)	16	(0.615)	11	(0.550)	17	(0.773)
質問文	4	(0.222)	0	(0.000)	4	(0.200)	2	(0.091)
参照先	5	(0.278)	10	(0.385)	5	(0.250)	3	(0.136)
合計	18	(1.000)	26	(1.000)	20	(1.000)	22	(1.000)

表 6　正誤判断の難問における誤りの原因
Table 6　Wrong types in the difficult True-or-False questions

	QA Lab-1				QA Lab-2			
	Phase-1		Phase-2		Phase-1		Phase-3	
時間情報	11	(0.611)	8	(0.308)	4	(0.200)	9	(0.409)
場所情報	1	(0.056)	3	(0.115)	5	(0.250)	1	(0.045)
同位語	5	(0.278)	7	(0.269)	4	(0.200)	7	(0.318)
反意	0	(0.000)	6	(0.231)	7	(0.350)	3	(0.136)
交換	1	(0.056)	2	(0.077)	0	(0.000)	2	(0.091)
合計	18	(1.000)	26	(1.000)	20	(1.000)	22	(1.000)

考えられる．しかしながら，「反意」では「保護する」や「解体を図る」といった一般的な語句にも着目する必要がある．また，「保護する」vs.「保護しない」といった機能語による単純な反意表現ではないことから語句の意味を正確に理解する必要がある．こういった点が「反意」における課題であると考えられる．

5.2.5　交換

2007CT_A12 正誤判断（2/15）

下線部 (3) に関連して，モンゴル人の支配下における出来事について述べた文として誤っているものを，次の 1-4 のうちから一つ選べ．
1. オゴタイ＝ハンは金を滅ぼし，カラコルムに都を置いた．(解答システム数：3)
2. 元では，イスラーム世界の科学の影響で，授時暦が作成された．(解答システム数：9)
3. 元の支配下では，漢人が重用され，西域出身の色目人は蔑視された．(解答システム数：2)
4. 従来の大運河が補修され，また大都に至る新運河が建設された．(解答システム数：1)
＜正解：3＞

3 番が誤りなのは，「元の支配下では，漢人が『蔑視』され，西域出身の色目人は『重用』された」からである．このように文中の語句に誤りはないが，その組合せが異なることで誤文となるものを**交換**と定義する．また，以下の例の 2 番のように交換された 2 文の片方しか記述されていないものも交換に分類することとした．

2011CT_A31 正誤判断（9/34）

下線部 (4) に関連して，20 世紀のスペインの歴史について述べた文として最も適当なものを，次の～のうちから一つ選べ．
1. フランス軍が，スペインでゲリラ戦に苦しんだ．(解答システム数：6)
2. スペイン内戦では，ドイツとイタリアは不干渉政策を貫いた．(解答システム数：10)
3. フランコが，独裁体制を確立した．(解答システム数：9)
4. アメリカ合衆国に，フィリピンを奪われた．(解答システム数：9)
＜正解：3＞

スペイン内戦は政府軍とフランコ軍の戦いであり，「ドイツとイタリアは『フランコ側を支援し』，イギリスとフランスは『不干渉政策を貫いた』」が正しい内容である．したがって，2 番の記述は交換した 2 文の片方に相当する．

「交換」が誤り原因の文をシステムが判断することを考える．Bag of Words などの単純な語の重なりで解決することは理論的に不可能であり，依存関係などの文の構造を正確に把握する必要がある．また，「反意」の課題とも関係するが，「フランコ側を支援する」や「不干渉政策を貫く」といった単純な動詞を超えたレベルで述部の意味を理解する必要がある．こういった点が「交換」における課題であると考えられる．

5.3　統計データ

表 5 と表 6 に Phase ごとの手がかり場所と誤りの原因をそれぞれ示す．この値は，誤りを含む選択肢ごとに集計し

表 7　手がかり場所と誤り原因の共起頻度
Table 7　Cooccurrence between the clue descriptions and the wrong types

	時間情報		場所情報		同位語		反意		交換		合計	
選択肢	15	(0.174)	4	(0.047)	21	(0.244)	9	(0.105)	4	(0.047)	53	(0.616)
質問文	6	(0.70)	4	(0.047)	0	(0.000)	0	(0.000)	0	(0.000)	10	(0.116)
参照先	11	(0.128)	2	(0.023)	2	(0.023)	7	(0.081)	1	(0.012)	23	(0.267)
合計	32	(0.372)	10	(0.116)	23	(0.267)	16	(0.186)	5	(0.058)	86	(1.000)

たものであることに注意されたい．例えば，5.1.3 の例では誤りの選択肢が 3 つあるため，それぞれに対して分析をしている．1 番の選択肢の場合，参照先の「漢の武帝」と同じ「皇帝」カテゴリーの「漢の光武帝」に置き換わっていることが原因なので，手がかり場所は「参照先」，誤り原因は「同位語」と判断した．一方，4 番の選択肢に対しては，選択肢中の「郡県制」の時代（秦）と「三省・六部」の時代（唐）が異なっていることが原因なので，手がかり場所は「選択肢」，誤り原因は「時間情報」と判断した．また，表 7 に，手がかり場所と誤り原因との間の共起頻度を示す．この表 7 に限って，割合（括弧内の値）は 4Phase 分の合計を 1 として計算している．

表 5 から，平均して 6 割程度が「選択肢」の情報だけで解答することが可能な問題であることが分かる．次いで多かったのが「参照先」であり，「質問文」の情報までを必要とするものは多くても 2 割程度にとどまった．表 6 から，「時間情報」が誤り原因である問題が一番多かったことが分かる．世界史のという分野を考えると，時間情報を正しく理解できているかが問われる問題が多いのは自然なことであると思われる．一方，GeoTime としてセットで扱われることの多い「場所情報」が誤り原因である問題は「時間情報」の 1/3 程度と少なく，順位にして 4 番目であった．日本史のように特定の地域ではなく世界全体を対象としていることを考えると少し意外な結果であった．2 番目に多かった誤り原因は「同位語」であり，3 番目が「反意」，最も少なかったのは「交換」であった．表 7 を見ると，「同位語」の手がかり場所はほとんどが「選択肢」であることが分かる．また，「質問文」が手がかり場所である場合は「時間情報」か「場所情報」のどちらかが誤り原因であることが分かる．今回の調査では，時間的制約から難問のみに絞ったこともあり調査対象となる問題数が少なく十分な結論を出すことが難しい．しかしながら，現時点の分析でも上で述べた傾向がみられており，今後さらに調査を続けていきたい．

6. おわりに

本稿では，NTCIR ワークショップの QA Lab タスクに参加したシステムの結果に基づいて，現在の QA システムが不得手とする問題を分析した．参加システムの 30%以下しか正答できなかった問題を難問と定義し，センター試験の難問を中心に分析を行った結果，質問形式の占める割合には大きな差がなかったこと，システムにとっての難問と赤本に書かれた人間にとっての難易度との間には関係がなかったことが確認できた．誤文と判断するための手がかりの場所と誤文と判断される原因の 2 通りの観点から正誤判断問題の誤りを類型化した結果，「時間情報」と比較して「場所情報」が誤り原因であることは少なかったこと，「質問文」が手がかり場所である場合，誤り原因は「時間情報」か「場所情報」のどちらかであったこと，「同位語」が誤り原因である場合，そのほとんどの手がかり場所が「選択肢」であったことが，難問に限定した小規模な調査であるが確認できた．今後，更に調査対象を広げて分析していきたい．

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] 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—, 人工知能学会誌, Vol.27, No.5, pp.463–469 (2012).
[4] Kobayashi, M., Miyashita, H., Ishii, A. and Hoshino, C.: NUL System at QA Lab-2 Task, *Proc. the NTCIR-12 Conference*, pp.413–420 (2016).
[5] 赤部晃一, Graham Neubig, 工藤拓, John Richardson, 中澤敏明, 星野翔: Project Next における機械翻訳の誤り分析, エラー分析ワークショップ (Project Next) (2015).
[6] 松崎拓也, 横野光, 宮尾祐介, 川添愛, 狩野芳伸, 加納隼人, 佐藤理史, 東中竜一郎, 杉山弘晃, 磯崎秀樹, 菊井玄一郎, 堂坂浩二, 平博順, 南泰浩, 新井紀子: 「ロボットは東大に入れるか」プロジェクト： 代ゼミセンター模試タスクにおけるエラーの分析, 自然言語処理, Vol.23, No.1, pp.119–159 (2016).
[7] 狩野芳伸: 大学入試センター試験歴史科目の自動解答, 第 28 回人工知能学会全国大会論文集, pp.1–4 (2014).
[8] Moldovan, D., Pasca, M., Harabagiu, S., and Surdeanu, M.: Performance Issues and Error Analysis in an Open-Domain Question Answering System, *Proc. the 40th Annual Meeting of the Association for Computational Linguistics*, pp.33–40 (2002).