大学入試の穴埋め問題を解く質問応答システムの検討

松井兵庫[†1]　阪本浩太郎[†1†2]　松永詠介[†1]　神貴久[†1]
渋木英潔[†1]　石下円香[†2]　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
E-mail: {m_hyogo,sakamoto,shin7240,taka_jin,shib,mori}@forest.eis.ynu.ac.jp,
{ishioroshi,kando}@nii.ac.jp

1　はじめに

近年，文書情報に対する情報要求は複雑化，高度化しており，そのような要求を満たす情報アクセス技術として質問応答が注目されている．質問応答とは，利用者の自然言語による質問に対して，情報源となる文書集合から回答そのものを抽出する技術であり，複雑高度な情報要求を自然言語で表現できる点に特徴がある．

しかしながら，従来の質問応答に関する研究では，「バラク・オバマとは誰ですか」といった比較的シンプルな形式の質問を扱うものが多かった．一方，現実世界においては，質問の核心に至るまでの背景や経緯を複数文にわたって説明したり，具体的な名称が思い出せずに「アメリカ初の黒人大統領」といった抽象的な表現を用いたりするなど，従来研究で想定されたのとは異なる質問状況である場合がある．このような質問の背景を説明する記述や，「下線部の政策を行った人物」といった抽象的な記述を含む質問の例として，大学入試問題があげられる．

NTCIR-11[1] の QALab タスク [1] では，世界史 B の大学入試問題（センター試験および二次試験）を対象とした課題が設定されており，我々もこれに参加した．大学入試問題にはセンター試験と二次試験[2] があるが，同じ Factoid 型の質問であっても多肢選択方式なのか自由記述方式なのか，同じ多肢選択式質問であっても内容の正しい文を選ぶのか空欄に当てはまる語句を選ぶのか，といった多様性が存在する．これらの多様性は，質問応答システムの処理に小さくない影響を及ぼすため，ロバストな質問応答システムを実現する上で検討しなくてはならない事項である．本稿では，QALab における Factoid 型質問および Blank 型質問を対象[3] に我々の取り組みと，現実世界における高精度かつロバストな質問応答を実現するための課題を記述する．

2　大学入試問題

2.1　質問形式の分類

世界史 B の大学入試問題について分析を行った．分析対象には，センター試験問題 4 年分（1997，2001，2005，2009 年）と，5 大学（東京大学，京都大学，北海道大学，早稲田大学，中央大学）の二次試験問題 2 年分（2005，2009 年）を用いた．

分析の結果，問題の問われ方や，図や空欄の有無により大きく分けて以下の 6 つの質問形式に分類できた．

1. Factoid 型…設問の答えを求める質問
2. Blank 型…空欄として相応しい語を求める質問
3. YesNo 型…回答候補の正誤を判定する質問
4. Time 型…時代順など選択肢を順番に並べる質問
5. Graph 型…図を読み解き解答する質問
6. Essay 型…数行または数百字で記述する質問

表 1: 大学入試問題（2007）の形式別質問数

質問形式	センター試験		二次試験（5 大学計）	
	自動	人手	自動	人手
Factoid 型	6	5	155	152
Blank 型	5	5	43	43
YesNo 型	23	24	15	15
Time 型	1	1	0	1
Graph 型	1	1	0	1
Essay 型	—	—	18	19

質問形式の分類は，センター試験と各大学の二次試験のいずれにおいても，年度の違いで問題の問われ方に大きな違いが無かったため，ルールベースの手法で行った．

分析時に得られたルールで，2007 年のセンター試験および二次試験の質問形式の分類を行った結果を表 1 に示す．

表における—は，センター試験には Essay 型の質問が一切存在しないことを示している．ルールベースによる分類精度は，センター試験では 97.2%，二次試験では 98.7%であり，本研究を進めるうえで問題はないと考えられる．

本研究では，Factoid 型と Blank 型の質問に焦点を当てた．両質問形式に焦点を当てた理由は，より難度が高いと考えられる二次試験において，質問数の割合が高かったからである．また，YesNo 型質問が，NTCIR-10 の RITE-2 タスク [3] における大学入試サブタスクで既に扱われていたので，まだ，扱われていない質問形式に挑戦しようと考えたからである．

2.2　回答方式による分類

入試問題には，センター試験のように回答候補を有する多肢選択式質問と，二次試験に多くみられる回答候補を有しない自由記述式質問がある．自由記述式質問では，多肢選択式質問とは異なり，回答候補を探す処理や要求されている回答の数を判定する処理などが必要とされる．本研究で用いた入試問題のデータ構造には，選択肢に関する情報が含まれているため，回答方式の分類は精度 100%で行えている．

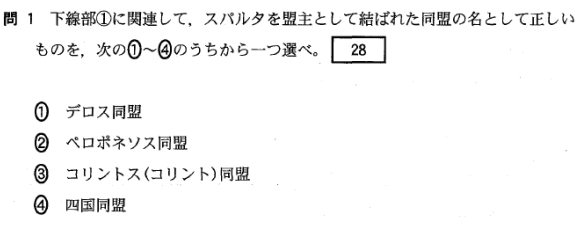
2.3　センター試験の質問形式

センター試験問題は約 38 問程度あり，全てが多肢選択式質問である．質問数の割合は YesNo 型が全体の 3 分の 2 強を占めており，Factoid 型と Balnk 型がその次に高かった．高得点を狙うためには，少なくともこ

[1] http://research.nii.ac.jp/ntcir/ntcir-11/

[2] NTCIR-11 の QALab では，2007 年と 2003 年のセンター試験，および 2007 年の東京大学，京都大学，北海道大学，早稲田大学，中央大学の二次試験を対象としている．

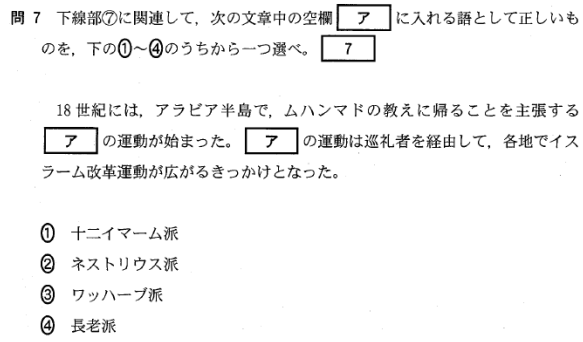
[3] QALab に参加したシステムでは，YesNo 型質問や Essay 型質問など全ての型の質問に対応している．文献 [2] を参照されたい．



問 1　下線部①に関連して，スパルタを盟主として結ばれた同盟の名として正しいものを，次の①～④のうちから一つ選べ。 28

① デロス同盟
② ペロポネソス同盟
③ コリントス(コリント)同盟
④ 四国同盟

図 1: センター試験における Factoid 型質問の例



問 7　下線部⑦に関連して，次の文章中の空欄 ア に入れる語として正しいものを，下の①～④のうちから一つ選べ。 7

18 世紀には，アラビア半島で，ムハンマドの教えに帰ることを主張する ア の運動が始まった。 ア の運動は巡礼者を経由して，各地でイスラーム改革運動が広がるきっかけとなった。

① 十二イマーム派
② ネストリウス派
③ ワッハーブ派
④ 長老派

図 2: センター試験における Blank 型質問の例

れらの 3 つの質問形式に焦点を当てる必要がある．図 1 と図 2 に Factoid 型と Blank 型の質問例を示す．

2.4　二次試験の質問形式

二次試験問題は，多肢選択式の問題も含まれるが，大部分が自由論述式の問題である．また，Essay 型のように文での記述を求められる質問があり，センター試験と比べ，より高度な処理が必要とされる．質問数の割合は Factoid 型の質問が全体の 3 分の 2 弱を占めており，どの大学においても出題されている．次いで Blank 型，Essay 型，YesNo 型の順に割合が高い．図 3 と図 4 に Factoid 型と Blank 型の質問例を示す．

3　基本的な考え方

本研究で扱った Factoid 型と Blank 型の質問は，知識源となるテキスト集合の中から，回答となる語句を探すタスクに帰結できる．また，回答語句を探すタスクは，質問文とテキスト集合との間の類似度に基づき，その類似度は，文（テキスト）中のキーワードのベクトルにより計算される．そのため，質問文のキーワードベクトルと，知識源のテキストのキーワードベクトルを作成する必要がある．

まず，各質問形式における質問文のキーワードベクトルの作成方法について述べる．

Factoid 型質問では，図 5 のように，下線部の文と質問文の 2 文よりキーワードを抽出する．そして，得られたキーワードとその頻度によってキーワードベクトルを作成する．なお，図においてキーワードベクトルの括弧内の数字は，その要素の頻度を示している．この際，1 つだけキーワードベクトルが作成される．

Blank 型質問では，図 6 のように，空欄を含む文を各選択肢で埋めてから，キーワードを抽出する．そして，得られたキーワードとその頻度によってキーワードベクトルを作成する．この際，選択肢の数だけキーワードベクトルが作成される．

次に，各回答方式における知識源および知識源のテキストのキーワードベクトルの作成方法について述べる．

多肢選択式質問では，知識源として Wikipedia を用いる．Wikipedia における 1 ページを 1 文書とみなして，キーワードの抽出を行い，キーワードと tf-idf 値でキーワードベクトルを作成する．

人類の歴史においては，無数の団体や結社が組織され，慈善・互助・親睦などを目的とする団体と並んで，ときには支配勢力と対立する宗教結社・政治結社・秘密結社もあらわれた。このような団体・結社に関する以下の質問に答えなさい。解答は，解答欄(ハ)を用い，設問ごとに行を改め，冒頭に(1)～(10)の番号を付して記しなさい。

問(1)　18 世紀末の中国では，世界の終末をとなえる弥勒下生信仰に基づく宗教結社が，現世の変革を求めて四川と湖北との境界地区などで蜂起したが，おもに郷勇などの自衛組織に鎮圧された。この宗教結社がおこした乱の名称を記しなさい。

図 3: 二次試験における Factoid 型質問の例

次の文章(A，B)の ［　　］ の中に最も適切な語句を入れ，下線部(1)～(10)について後の問に答えよ。解答はすべて所定の解答欄に記入せよ。

A　中国で古くから精度のかなり高い地図が作られていたことは，1973 年に湖南省長沙の馬王堆(まおうたい)で発掘された紀元前 2 世紀の墓中にあった絹製の地図(1)によって明らかになった。文献には地図に関する記述はけっして多くはない。春秋五覇の筆頭である ［ a ］ に仕えた政治家管仲の著作と伝えられる『管子』には「地図」篇があって地図の軍事的重要性が強調され，また『戦国策』には蘇秦が合

図 4: 二次試験における Blank 型質問の例

自由記述式質問では，知識源として一問一答問題集 [8] を用いる．一問一答問題集の構造は，関連する内容に対し，1 問から数問で一つの固まり（問群）を作っており，問群が複数並んでいる構造になっている．初めに，問群ごとと一問ごとの両単位においてキーワードの抽出を行い，それぞれに対してキーワードと頻度でキーワードベクトルを作成する．

4　回答生成手法

本節では 3 節の考え方をもとに，多肢選択式質問と自由記述式質問のそれぞれにおける，回答生成手法について述べる．多肢選択式質問では，手法 1，手法 2 を提案し，自由記述式質問では，手法 3 を提案する．各手法でキーワードを抽出するために形態素解析器 MeCab[4] を利用した．キーワードには，名詞，複合名詞（アルファベットの連続も含む），形容詞を用いた．いずれの手法においても類似度の計算にはコサイン類似度を用いている．

表 2 に実験に用いる問題の形式別質問数を示す．この表において，手法 1 と手法 2 は多肢選択式質問にしか答えることができないが，手法 3 は自由記述式質問だけでなく，多肢選択式質問でも答えることができる．

4.1　多肢選択式質問

多肢選択式質問では，回答候補を含む類似文書を Wikipedia から検索する．Factoid 型質問の処理の流れを図 7 に示す．

まず，3 節で述べた方法で質問文と Wikipedia それぞれのキーワードベクトルを作成する．そして，作成したキーワードベクトルと回答候補の情報を用いて，各選択肢ごとに選択肢自身を含むページを取得する．回答選択におけるスコアとして，手法 1 では最も類似度の高い文書の類似度を採用し，手法 2 では類似度の高い上位 10 件の文書の類似度の合計を採用している．

Blank 型質問の場合もほぼ同様の形で処理できる．

4.2　自由記述式質問

自由記述式質問では，回答候補を求める処理が必要となるが，難しい処理のため，一問一答問題集の構造を

[4] http://code.google.com/p/mecab/

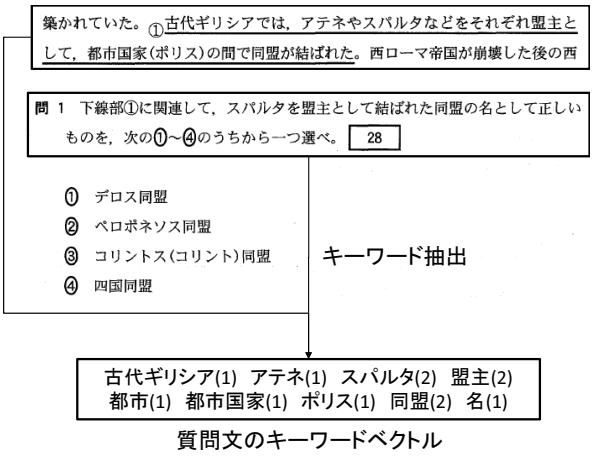


図 5: Factoid 型質問のキーワードベクトルの作り方

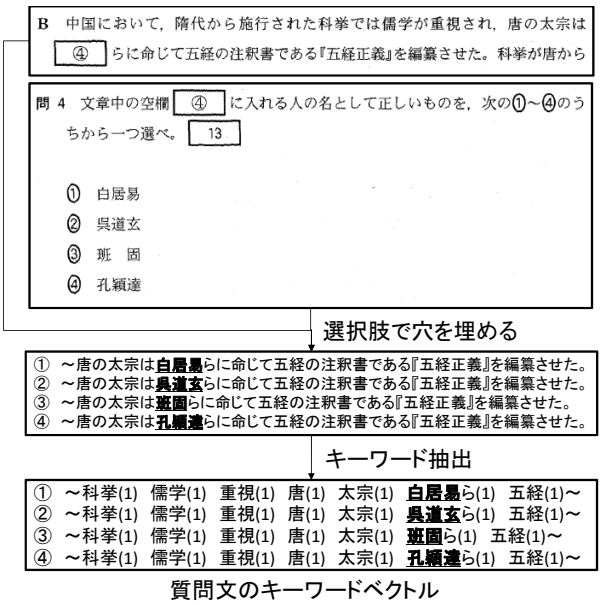


図 6: Blank 型質問のキーワードベクトルの作り方

上手く利用して簡潔に解く方法を用いる．これを手法 3 とする．Factoid 型質問の処理の流れを図 8 に示す．

まず，3 節で述べた方法で質問文と一問一答問題集それぞれのキーワードベクトルを作成する．そして，質問文のキーワードベクトルと，一問一答の各問群のキーワードベクトルを用いて類似度計算を行い，類似度の高い上位 3 つの問群を求める．その後，質問文のキーワードベクトルと，得られた 3 つの問群に含まれる各問題のキーワードベクトルを用いて，再び類似度計算を行い，類似度の最も高い問題を求める．最後に，回答選択として，類似度の最も高い問題の解答をそのまま出力する．

Blank 型質問の場合，通常の多肢選択式質問と同様に，空欄を回答候補（一問一答の解答）で埋めてしまうと，キーワードベクトルの数が膨大なものとなり，一問一答の構造を用いる利点が薄まってしまうため，空欄のある文を Factoid 型質問における質問文として扱い，Factoid 型質問と同様の処理に帰結させている．

5 評価実験

4 節で述べた各回答生成手法の効果を調べるために評価実験を行った．

表 2: 実験に用いる問題の形式別質問数

質問形式	センター試験（多肢選択）	二次試験 多肢選択	二次試験 自由記述
Factoid 型	18	25	127
Blank 型	27	0	43

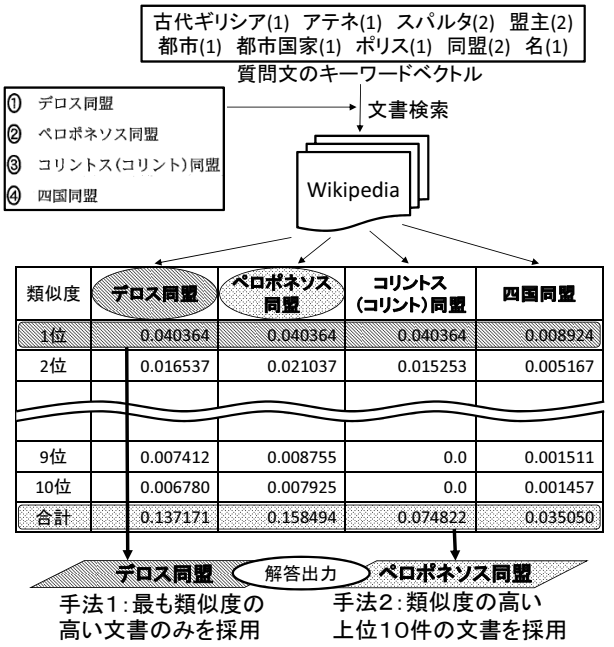


図 7: 多肢選択式 Factoid 型質問の解答生成手法

実験には，センター試験（6 年分）と二次試験（2007 年の 5 大学）の問題（図 2）を用いた．評価指標には正答率を用いた．

$$\text{正答率} = \frac{\text{システムの正解数}}{\text{対象とした問題数}}$$

システムの正解は，多肢選択式質問については，実際の解答と完全に一致した場合のみを正解とし，自由論述式質問については，人手によって表記ゆれを考慮して正解を判定した．実験の結果を表 3 に示す．表における—は，該当手法では理論的に解けないことを示している．

センター試験における Factoid 型の手法 1 と手法 3 は，基本的にセンター試験が 4 択であることを考えると良い精度とは言えない．

多肢選択式 Factoid 型質問においてセンター試験では手法 2 の方が良い結果であったが，二次試験では手法 1 の方が良い結果であった．このことより，上位の文書を重視しつつも，上位数件の文書についても考慮する必要があると考えられる．そのため，手法 2 のように上位 10 件の文書の類似度を単純に足し合わせるのではなく，上位の文書ほど重要度をあげるような重みづけをして，類似度を足し合わせる方が良いと考えられる．

以下，正しく解答できなかった問題とその原因について述べる．提案した手法では，時間情報を扱っていないため，「チンギス=ハンの時代以後に成立した宗教として正しいものを選べ」といった，時間情報が無ければ解くことが難しいような問題には対応できていなかった．このことより，時間情報を扱えるような仕組みが必要であると考えられる．

また，単語の有無のみで問題を解いているため，一文で複数の解答を求めてくる問題や，一文内に複数の空欄があるような問題には，どちらにも同じ解答を返してしまうことがあり，正しく対応できていなかった．

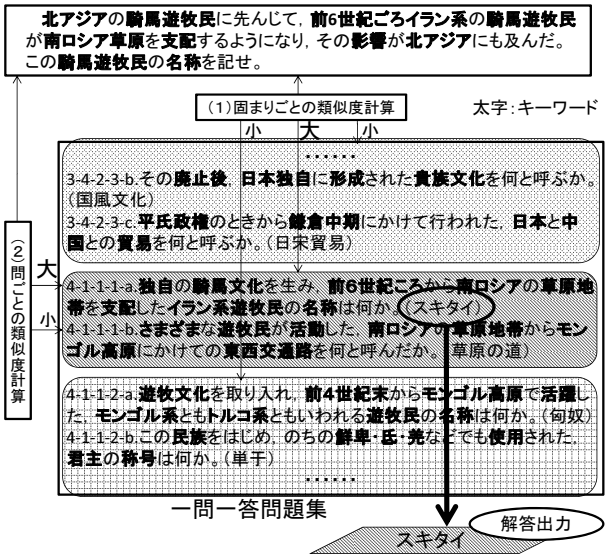


図 8: 自由記述式 Factoid 型質問の解答生成手法

表 3: 各入試問題の形式別正答率

質問形式	手法	センター試験（多肢選択）	二次試験 多肢選択	二次試験 自由記述
Factoid 型	手法 1	0.28	0.52	—
	手法 2	0.50	0.36	—
	手法 3	0.28	0.36	0.27
Blank 型	手法 1	0.48	(問なし)	—
	手法 2	0.48	(問なし)	—
	手法 3	0.0	(問なし)	0.16

これらは，係り受け構造の利用で対応できると考えられる．

他に，反乱の名称を聞かれているのに，人名を出力したり，思想家の名を聞かれているのに協定の名を出力したりと，聞かれているものと違う型の回答を出力している場合が多くあった．これらは，質問型や質問の焦点を利用することで対応できると考えられる．

5.1 QALab の結果

1 節で述べた QALab について参加した結果を簡単に述べる．QALab における Factoid 型と Blank 型の結果を図 9 と図 10 に示す．本チームの結果はグラフの横軸（「チーム名-言語-番号」）において「Forst-JA-01」で示してある．グラフの縦軸は，各形式の質問に対する正答率を足し合わせたものである．正答率は 0.0 から 1.0 までの値であり，図では Factoid 型と Blank 型に関して示してあるので，縦軸の最大値は 2.0 となっている．

他のチームの処理を確認してみると，DCUMT[4]，FLL[5]，nnlp[6]，CMUQA[7] などの多くのチームで時間情報を利用しており，結果に良い影響を与えていたので，その有用性が分かる．また，DCUMT や CMUQA などのチームでは意味解析に関する処理や技術を利用しており，必要な処理の一つであると思われる．

6 まとめ

大学入試の穴埋め問題を解くための質問応答システムにおける取り組みについて述べた．穴埋め問題を解くために 3 つの手法を提案し，各手法の評価実験を行った結果，正答率は 3 割から 5 割程度であった．

より正答率を上げるためには，時間情報や係り受け構造などの利用が必要であることが分かった．また，質問型を用いなかったことによる誤りが多かったので，今後は世界史 B に特化した質問型を定義し，利用することを検討していきたい．

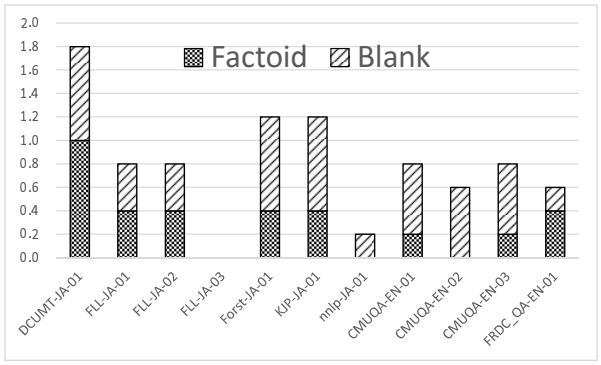


図 9: QALab の結果（センター試験 2007・Phase1）

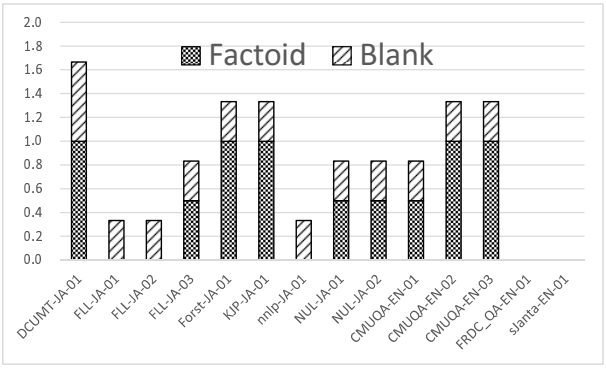


図 10: QALab の結果（センター試験 2003・Phase2）

参考文献

[1] 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”, Proceedings of the 11th NTCIR Conference, (2014)

[2] K. Sakamoto, H. Matsui, E. Matsunaga, T. Jin, H. Shibuki, T. Mori, M. Ishioroshi, N. Kando, “Forst: Question Answering System Using Basic Element at NTCIR-11 QA-Lab Task”, Proceedings of the 11th NTCIR Conference, (2014)

[3] Y. Watanabe, Y. Miyao, J. Mizuno, T. Shibata, H. Kanayama, C. Lee, C. Lin, S. Shi, T. Mitamura, N. Kando, H. Shima, K. Takeda, “Overview of the Recognizing Inference in Text (RITE-2) at NTCIR-10”, Proceedings of the 10th NTCIR Conference, (2013)

[4] T. Okita, Q. Liu, “The Question Answering System of DCUMT in NTCIR-11 QALab”, Proceedings of the 11th NTCIR Conference, (2014)

[5] T. Makino, S. Okura, S. Okajima, S. Song, H. Suzuki, “FLL: Answering World History Exams by Utilizing Search Results and Virtual Examples”, Proceedings of the 11th NTCIR Conference, (2014)

[6] Y. Kimura, F. Ashihara, A. Jordan, K. Takamaru, Y. Uchida, H. Ototake, H. Shibuki, M. Ptaszynski, R. Rzepka, F. Masui, K. Araki, “Using Time Periods Comparison for Eliminating Chronological Discrepancies between Question and Answer Candidates at QALab NTCIR11 Task”, Proceedings of the 11th NTCIR Conference, (2014)

[7] D. Wang, L. Boytsov, J. Araki, A. Patel, J. Gee, Z. Liu, E. Nyberg, T. Mitamura, “CMU Multiple-choice Question Answering System at NTCIR-11 QA-Lab”, Proceedings of the 11th NTCIR Conference, (2014)

[8] 今泉 博, 一問一答世界史 B 用語問題集, 山川出版社, (2009/01)