Overview of Todai Robot Project and Evaluation Framework of its NLP-based Problem Solving

Akira Fujita, Akihiro Kameda, Ai Kawazoe, Yusuke Miyao
National Institute of Informatics,
2-1-2 Hitotsubashi, Chiyoda-ku, Tokyo, Japan
E-mail: {a-fujita, kameda, zoeai, yusuke}@nii.ac.jp

Abstract

We introduce the organization of the Todai Robot Project and discuss its achievements. The Todai Robot Project task focuses on benchmarking NLP systems for problem solving. This task encourages NLP-based systems to solve real high-school examinations. We describe the details of the method to manage question resources and their correct answers, answering tools and participation by researchers in the task. We also analyse the answering accuracy of the developed systems by comparing the systems' answers with answers given by human test-takers.

Keywords: problem solving, document structure annotation, question-type annotation, competition, evaluation framework

1. Introduction

This paper presents the achievements of the 'Todai Robot Project', which focuses on benchmarking NLP systems for problem solving. We report the organization of the Todai Robot Project and the results of its first open evaluation.

The Todai Robot Project aims to develop problem-solving systems that can attain high scores in the Japanese National Center Test for University Admissions by 2016 and pass the University of Tokyo's entrance examination by 2021. For this task, the organizers provide several shared resources for task-takers, which comprise mainly questions in university entrance examinations and textbooks that possess information required to solve problems. Some of the resources were translated into English and are open to all task-takers.

The questions are primarily collected from the Japanese National Center Test for University Admissions (Center Test)[1] and its trial examination held by a prep school. Questions in university entrance examinations are mostly presented in natural language; therefore, it is clear that having a proper understanding of texts is a prerequisite for solving questions. Such questions are carefully designed to empirically quantify the academic skills of high school students. In this regard, these resources are ideal for evaluating end-to-end NLP systems that read natural language text, perform information processing and output answers.

Furthermore, these resources have the following features, which are suitable for solving questions:

- Required knowledge is limited
- Legitimate solutions always exist
- Evaluation of answers by systems is intuitively understandable and can be compared directly with human performance.

These features provide clues regarding the true contributions of current NLP technologies for human-like problem-solving tasks as well as empirical evidence to analyse relationships between human intelligence and artificial intelligence.

However, there are barriers that must be addressed by organizers to enable task-takers to isolate specific NLP subtasks for focused research. It is almost impossible for task-takers to develop smart algorithms that can solve all types of questions because entrance examinations span multiple subjects and question types. It is necessary to clearly indicate the question type to task-takers who can adjust the system to solve questions by applying optimized algorithms. Examination data include several components, such as figures and tables. Therefore, the semantic analysis of the components is not trivial and beyond the scope of current NLP research. Their semantic content requires interpretation into natural language in advance. Schemes are also required to evaluate the systems submitted by task-takers and to provide updates regarding the task status.

We introduce a method to annotate such resources. We also present a method to manage question resources and their correct answers, answering tools and participation of researchers in the Todai Robot Project. Finally, we present the contribution of this task with the results of an open evaluation in which the developed systems solved a test data set in a trial examination of the Center Test.

2. Resources

There are eleven subtasks corresponding to each subject in the Center Test: *Biology, Chemistry, Ethics, English (foreign language), Japanese (native language), Japanese History, Mathematics, Modern Society, Physics, Politics and Economics and World History*. All subjects in the Center Test, excluding Mathematics, consist of multiple choice questions. Mathematics test-takers do not select a choice, but they put numeric characters into a given column in which the number of characters is specified.

[1] Center Test is a nationwide standardized test for university admission in Japan. Nearly all high-school students who aim to enter university take this exam. All questions in Center Test are multiple-choice format, excluding Mathematics.

	dev	test	final-test
Biology	601	541	1085
Chemistry	515	688	1009
Ethics	307	266	495
English	520	589	1111
Japanese	296	263	479
Japanese History	531	505	1056
Mathematics	659	659	1254
Modern Society	822	800	1463
Physics	380	417	799
Politics & Economics	536	480	1036
World History	688	700	1340

Table 1: Total number of questions in each subtask (Center Test).

	dev	final-test
English	48	48
Japanese	35	35
Japanese History	36	36
Mathematics	75	86
Physics	25	24
Politics & Economics	38	36
World History	36	36

Table 2: Total number of questions in each subtask (trial examination).

All examination texts prepared in these subtasks are obtained from the Center Test and a trial examination held by a prep school. The Center Test examination texts from 1990 to 2011 were used. Each year, a main examination and its supplementary examination, which is open to students who were unable to take the main examination, are available. These examinations are conducted for each subject. In total, there are 571 examinations in the Center Test. Moreover, the trial examination texts from 2012 and 2013 were used. The contents of these examination texts were also translated into English. Textbooks for Biology, Chemistry, Ethics, Japanese History, Mathematics, Modern Society, Physics, Politics and Economics and World History are also provided to task-takers in XML format data. All textbook data are described in Japanese.

Datasets of each subtask are divided into three sets: *dev* (for development), *test* (for evaluating systems in development progress) and *final-test* (for final validation of system achievement). The 'dev' dataset is provided to tune the systems. To optimize system performance, task-takers repeatedly adjust their system performance using 'dev' and conduct open tests. The 'test' dataset is provided to prove system performance. The 'final-test' dataset analyses how the systems attain scores on a dataset that is not examined by the developers. Consequently, neither the questions nor their correct answers in 'test' and 'final-test' are shown to

`question`	A question region including outer-most question areas and minimal areas. - `@id`: An identifier of the question region. The id starts with 'Q'. - `@minimal`: '`yes`' or '`no`' can be taken. Question regions that do not include other question regions are given the attribute `minimal = 'yes'`, indicating smallest units of questions. - `@answer_style`, `@answer_type`, `@knowledge_type`: (c.f. section 3.4). - `@anscol`: '`@id`'s of '`ansColumn`'s which are included in the question region.
`instruction`	A statement or an instruction for a question.
`label`	A label such as section numbers, question numbers and identifiers of text fragments.
`ansColumn`	An identifier of an answer column. - `@id`: An ID which is referred to in 'correct answer table' (below-mentioned). The id starts with 'A'.
`choices`	A set of choices. - `@anscol`: An identifier of the corresponding answer column. - `@comment`: Explanatory notes of special circumstance which should be noticed.
`choice`	An individual choice. - `@ansnum`: An identifier of the choice which is referred to in 'correct answer table' (below-mentioned). - `@comment`: Explanatory notes of special circumstance which should be noticed.
`cNum`	A label of choice numbers or markers.
`ref`	A symbol that refers to another text fragment or object data. - `@target`: `@id`(s) of the referred text fragment or object data. `Data`, `uText`, `lText`, `note` and `blank` can be referred. - `@comment`: Explanatory notes of special circumstance which should be noticed.
`data`	Data provided to test-takers of reference to help his/her interpretation of the question or to directly express subjects of the question (such as '*What is the name of the building shown in following figure?'*). - `@id`: An ID which is referred by `ref`. The id starts with 'D'. - `@type`: A format of the data, such as text, image, table and graph.
`uText`	An underlined text fragment. - `@id`: An ID which is referred by `ref`. The id starts with 'U'.
`lText`	A labelled text fragment. For example, descriptions which are itemized with some marks in a statement in a question are labelled texts. - `@id`: An ID which is referred by `ref`. The id starts with 'L'.
`note`	A note or footnote which is to help test-takers interpretation. - `@id`: An ID which is referred by `ref`. The id starts with 'N'.
`blank`	A blank which should be filled with the answer by test-takers. - `@id`: An ID which is referred by `ref`. The id starts with 'B'. - `@digits`: A specified character of the blank. (Only in the case that there is a specification)

Table 3: Tags and their attributes of document structure.

```
<exam subject="Politics_and_Economics(main exam)" year="2009">
 <title>
   2009 Academic Year Main Examination: Politics &amp;amp; Economics
 </title>
 <question id="Q1" minimal="no">
   <label>[1] </label>
   <instruction>
    Read the following text and answer the questions(Questions 1-10) below.
   </instruction>
   <data id="D0" type="text">
    Today, one of the biggest issues that the international community has to tackle is the <uText
    id="U1"><label>(a)</label>North-South divide</uText>, that is to say, the economic disparity
    between developed and ...
   </data>
  <question anscol="A1" answer_style="multipleChoice" answer_type="symbol-term_location" id="Q2"
    knowledge_type="KS,IC_G" minimal="yes">
    <label>Question 1</label>
    <instruction>
     In regard to the underlined portion <ref comment="" target="U1">(a)</ref>, the diagram below
     shows the relationship between gross national income (GNI) per capita and the infant
     mortality rate in 2005 in Japan, India, ...
    </instruction>
    <ansColumn id="A1">1</ansColumn>
    <data id="D12" type="image">
   <img src="Center-2009--Main-Politics_and_Economics-001.png"/>
    </data>
    <choices anscol="A1" comment="">
      <choice ansnum="1">
      <cNum>(1)</cNum><ref comment="inside of image" target="">A</ref>India<ref comment=" inside
        of image " target="">B</ref>South Korea<ref comment=" inside of image "
        target="">C</ref>Brazil
      </choice>
      <choice ansnum="2">
      <cNum>(2)</cNum><ref comment=" inside of image " target="">A</ref>India<ref comment=" inside
        of image " target="">B</ref>Brazil<ref comment=" inside of image " target="">C</ref>South
        Korea
      </choice>
      <choice ansnum="3">
        ...
    </choices>
 </question>
 ...
```

Figure 1: XML example of document structure (2009 Center Test Politics and Economics)

task-takers. Datasets 'dev' and 'test' both include half of the Center Test examination texts used in odd-numbered years, and the 'final-test' dataset includes the remainder. The trial examination texts from 2012 were assigned to 'dev', and those from 2013 were assigned to 'final-test'. The trial examination data are not assigned to 'test' and do not include Biology, Chemistry, Ethics or Modern Society subtask data. The total number of questions in each subtask is shown in Table 1 and Table 2. We have datasets of correct answers (correct answer tables) for the examination data. These correct answer tables include the allotment of marks for each question. Additional examination texts will become available shortly. For example, the University of Tokyo entrance examination texts are currently being prepared for this task. In future, the total number of questions in our tasks will be increased.

3. Dataset Preparation

We collected PDFs and source texts from the National Center Test for University Admission in Japan and the Yoyogi Seminar prep school. The examination datasets have been annotated with information required for NLP benchmarking. The types of annotation are as follows:

- Document structure
- Formula
- Image
- Question-type

Each annotation type is explained below.

3.1 Document structure annotation

Our examination datasets are provided with human-annotated document structures in XML format. Examination texts should be structured such that users can easily extract questions and relevant text regions because task-takers frequently refer to the questions and collect corresponding information. However, automatic recognition of complicated document structures, such as examination texts, remains a challenge (Schäfer and Weitz, 2012). An examination text comprises several regions: questions and reference data. The question regions include small questions called 'minimal' questions, each of which typically comprise instructions to test-takers and a set of choices. The reference data regions often include images, tables, graphs and texts with extra-linguistic markup, such as underlines. These regions and markup are intricately related to each other in examination texts. An example of annotated examination document structure is shown in Figure 1. The annotation shown in Figure 1 corresponds to the beginning part of an examination description for Center Test Politics and Economy. In the example, there is a part (a non-minimal question) whose ID is Q1. The instructions for this part indicate that the part consists of 10 questions in total. After the instruction, the data which is referred by the 10 questions in this part and the first minimal question are described. The first minimal question has the following attributes: '`@anscol`', '`@answer_style`', '`@answer_type`', '`@id`', '`@knowledge_type`' and '`@minimal`'. Including the definition of these attributes,

```
<answerTable     filename="Center-2009--Main-
  Politics_and_Economics">
  <data>
    <question>1</question>
    <answer_column>1</answer_column>
    <answer>4</answer>
    <score>3</score>
    <answer_type>symbol-
  term_location</answer_type>
    <answer_style>multipleChoice</answer_styl
  e>
    <knowledge_type>KS,IC_G</knowledge_type>
    <question_ID>Q2</question_ID>
    <anscolumn_ID>A1</anscolumn_ID>
  </data>
  <data>
    <question>2</question>
    <answer_column>1</answer_column>
    …
```

Figure 2: Example of correct answer table (2009 Center Test Politics and Economics)

```
      <math
    xmlns="http://www.w3.org/1998/Math/MathML">
        <semantics>
          <mrow>
            <msup>
              <mi>a</mi>
              <mn>2</mn>
            </msup>
            <mo>≧</mo>
            <mrow>
              <mn>2</mn>
              <mo></mo>
              <mi>a</mi>
              <mo> + </mo>
              <mn>8</mn>
            </mrow>
          </mrow>
          <annotation-xml encoding="MathML-Content">
            <apply>
              <geq/>
              <apply>
                <power/>
                <ci>a</ci>
                <cn>2</cn>
              </apply>
              <apply>
                <plus/>
                <apply>
```

Figure 3: An example of formula annotation (formula: $a^2 \geq 2a + 8$)

the specification of typical tags that are used for the document structure annotation and their attributes are shown in Table 3. Detailed specifications of the annotation are mentioned in Miyao and Kawazoe (2013).

The correct answer tables are also annotated in XML format (*correct answer table XML*). The following elements for each minimal question are included in the correct answer table XML: answer, score, question ID, answer column ID, answer type, answer style and knowledge type. An example of a correct answer table XML that corresponds to the examination example presented in Figure 1 is shown in Figure 2.

3.2 Formula annotation

Formulas play crucial roles in Physics and Mathematics examinations. Although understanding their semantics is important, the semantic analysis of formulas is complex and beyond the scope of NLP research. We have marked all formulas appearing in examination texts using MathML (World Wide Web Consortium, 2003), which is a common standard format for mathematical expressions. Thus, task-takers can concentrate on NLP research with the assumption that formula semantics are given. An example of the formula annotation is shown in Figure 3.

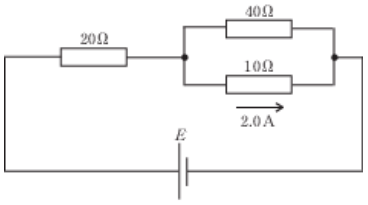


Figure 4: An example of image data (Copyright: Yoyogi seminar)

```
<img    src="YozemiCenter-2013--1-ButsuriI-001.png"
corresponding_obj="X1_1 X1_2 X1_3 X1_4 X1_5" />
  <adata id="X1_1" type="text" field="object"">
    A battery whose electromotive force is E
  </adata>
  <adata id="X1_2" type="text" field="object">
    A resistor whose resistance value is 20Ω
  </adata>
  <adata id="X1_3" type="text" field="object">
    A resistor whose resistance value is 40Ω
  </adata>
  <adata id="X1_4" type="text" field="object">
    A resistor whose resistance value is 10Ω
  </adata>
  <adata id="X1_5" type="text" field="event">
      Circuit diagram. Left end of a registor whose
      resistance is 20Ω is connected to positive
      electrode of a battery whose electromotive
      force is E. Right end of the 20Ω registor is
      connected to left ends of a resistor whose
      resistance value is 40Ω and a resistor whose
      resistance value is 10Ω. The 40Ω and the 10Ω
      resistors are connected in parallel. Right
      end of the 40Ω and the 10Ω resistors are
      connected to negative electrode of the
      battery. 2.0A electrical current is flowing
      in the 10Ω.
  </adata>
```

Figure 5: Image annotation for Figure 4

3.3 Image annotation

Examination data also include images that cannot be analysed with NLP technologies alone, such as pictures, figures, tables and graphs. In some cases, images are important for answering a question and in other cases they are not, and the necessity of the image part is difficult to infer with an NLP system. Therefore, we have annotated their semantic content with natural language sentences or formalized expressions without considering their importance to answering the questions. For example, image data, such as that shown in Figure 4, are annotated, as shown in Figure 5. In the image annotation, illustrated objects and situations are described in natural language with '`adata`' elements. The '`@field`' attributes indicate the types of the described entities. In the case of Figure 4, the resistors and the battery can be regarded as the objects. Since the objects are connected to each other, this connection information is also annotated.

3.4 Question-type annotation

Typically, different strategies are employed to solve different types of questions. For example, in the

Mathematics subtask, a participant tries to generate logical formulas from question statements and derive answers by manipulating the formulas (Iwane et al., 2013), while in Physics, a participant uses a physical simulator by providing various parameters from question statements (Yokono and Inamura, 2013). In the social studies subtasks, answers to factoid and true/false questions are addressed by applying textual entailment recognition (Miyao et al., 2012) and question answering (Kanayama et al., 2012). In addition, the selection of resources significantly affects the accuracy of reasoning of truth/falsehood.

Even in a single subtask, questions can contain components that can be regarded as different NLP tasks. For example, it is required to identify the most appropriate solving strategy in Japanese, English and the social studies subtasks. To deal with this situation, we have classified questions by answer style (e.g. multiple-choice), answer types (e.g. terms, sentences, formulas and symbols) and type of knowledge required to solve the question. We have annotated each minimal question with classification information. Knowledge-type information is the most important when selecting strategies to solve questions; therefore, we employ granular classification (90 types in total).

The question-type annotation also provides participants with a function that obtains a subset of examinations in which isolated NLP tasks can be studied. These question classifications are treated as attributes (i.e. '`@answer_style`', '`@answer_type`' and '`@knowledge_type`' of the `question` tag.

4. Framework for the competition

This task aims to encourage competition among task-takers and measures real contributions of current NLP studies and technologies when solving questions in each subtask. To this end, we prepared an automatic scoring framework and some baseline systems to define the performance of task-taker systems.

4.1 Automatic Evaluation

Task-takers can submit their answers to questions on our website. The answers must be in XML format (*answer XML*). The format of the answer XML is the same as the correct answer table. In answer XML, the indispensable tags in each answer that correspond to each minimal question are '`answer`' and '`anscolumn_ID`'. Logs in a solving process (*process log*) can be optionally described in the answer XML.

The submitted answer XML is automatically scored by our server. After scoring, four values representing the scoring results are sent to the task-takers and shown on a webpage (*result page*). The four values are as follows:

- Score (integer ≥ 0)
- Percentage of correctly answered questions (real number ≥ 0 and ≤ 1)
- Number of correctly answered questions
- Number of answered questions in submitted examination data

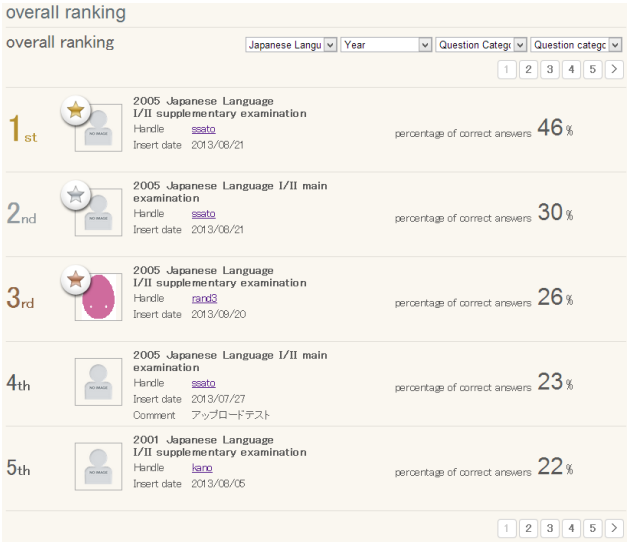


Figure 6: Screenshot of the ranking page

Subtask	Number of submitted answer XML
English	50
Japanese	55
World History	25
Physics	3

Table 4: Number of submitted answers according to task-taker system.

On this webpage, task-takers can refer to these values and also the correct answers, question statements, knowledge type and process log for each minimal question in the submitted examination data, as is shown in Figure 7.

Task-takers can also add comments for each question in the *comment region* provided in the answer XML. These comments appear in each question region in the result page so that task-takers can use them for error analysis.

4.2 Baseline system

We prepared three baseline systems to provide a target for comparison with task-taker results. All baseline systems are implementations of random algorithms.

The score submitted by task-takers and the baseline systems are listed in our *Ranking Page*[2], as is shown in Figure 6. This page is accessible by the public.

4.3 Participation

We accept applications for participation in this task on our web site. Contract forms for the use of resources are available on the website, and task-takers can download resources from the site at the conclusion of the contract. Table 4 shows the status of answer submissions as of March 2014. The number of submissions is shown for each subtask. All task-takers used datasets in Japanese. We also prepared datasets in English.

[2] http://21robot.org/task/ranking/

Scores	English	Japanese	Japanese History	Math IA	Math IIB	Physics	World History
Task-takers' system	52 (/200)	62 (/150*)	56 (/100)	57 (/100)	41 (/100)	39 (/100)	58 (/100)
Average of Students	88.3	72.2	45.6	52.0	47.6	42.0	46.6

* Full marks in the case excepting Chinese classics (full marks is 200)

Table 5: Results of the open evaluation of task-takers' systems
(The numbers in parentheses are the full marks of corresponding subjects).

5. Open Evaluation

This section describes the analyses of answering accuracy for task-taker systems. We performed open evaluation of the task-taker systems. The datasets for trial examinations in 2013 by Yoyogi Seminar were used to evaluate the task-taker systems as an independent task. This trial examination was also conducted with students who will write university entrance examinations this year.

The evaluation was performed with some subtasks, i.e. English (foreign language), Japanese (native language), Japanese History, Mathematics, Physics and World History. Mathematics was separated into two subtasks, Math IA and Math IIB, according to the field of questions.

In this paper, we report the results of the systems that obtained the highest score for each subtask. All subtasks consist of multiple-choice questions, excluding Mathematics. In some subtasks, some specific questions were not solved by the system. The Japanese (native language) subject consists of four different sub-subjects: Essay, Novel, Classics and Chinese classics. The system for the Japanese subtask did not solve Chinese classics questions. This exception is due to the difficulty in parsing the Chinese classics texts. Each system solved one subtask for all subtasks, with the exception of Japanese History and World History[3]. This does not mean that a system for a specific subtask applied only a single algorithm to solve all questions in the subtask. Most of the systems applied different algorithms according to the type of question.

Results of the evaluation are shown in Table 5, including the average scores of the students. In addition, the data with percentages of correct answers by the students for each question are accessible by the task-takers. We describe the features of task-taker systems on some subtasks as follows.

In the social studies subtasks, Japanese History and World History, there were two approaches. One system applied a module that is an application of the recognizing textual entailment (RTE) technique (Miyao et al., 2012). The other system applied a QA engine (Ishioroshi et al., 2013). The system that applied the QA engine obtained a better score than the system that applied the RTE engine in both Japanese History and World History.

The system for English applied an inference module that can recognize a textual entailment. However, to solve questions in the English subtask, common sense must be applied in many cases. For example, there are some fill-in-the-blank-type questions that question the appropriateness of an utterance for a specified dialogue. The interpretation of inconsistency of the conversation must be implemented in the system to solve such questions. Humans can mostly recognize the strangeness of incorrect choices that are suitable to the context but unnatural as a dialogue. The inference architecture in the system could not perform sufficiently to solve these types of questions. The judgement to determine whether a text is natural, such as the detection of strangeness in dialogue, is a significant NLP task.

The system for Mathematics derives first-order predicate logic formulas from higher-order predicate logic formulas, which are equivalent to the question statement, and solves a problem using a formula manipulation algorithm. Most questions that are given as formula with fewer instructions written in natural language were solved by this formula manipulation. However, if the whole question statement was written in natural language, it is difficult to establish formula from the question because there are many proposed parsing results for such questions.

Similar to Mathematics, the interpretation of a question statement is a task to be solved in Physics. A physical simulator is used in the task-taker system for providing parameters from question statements. Since Physics question statements are qualitative descriptions, elements that are not essential to the question (such as atmospheric temperature in an experiment to measure gravitational acceleration) are not described in the statement. However, these elements must be derived from the questions or be set to arbitrary values after being assessed as optional information because the simulator requires parameters of elements to be conceivable as its input.

These above problems that emerged in our open evaluation are relevant to recent and significant NLP tasks.

6. Related Work

NLP research that develops benchmark data from questions originally designed to evaluate human performance has been emerging. For example, the Halo project (Angele et al., 2003) targeted Chemistry tests, while IBM's Deep QA (Ferrucci, 2012) employed factoid-style quizzes. However, their benchmark data sets are not open to the public. Collaborative research is indispensable for our purpose because entrance examinations involve various NLP subtasks. Therefore, it is necessary to develop open

[3] For Japanese History and World History, three systems with different strategies solved questions.

resources, as described in this paper.

7. Conclusion

In this study, we introduced the achievements of the 'Todai Robot Project' task, which focuses on benchmarking NLP systems for problem solving.

We described the details of the method to manage the resources of questions and their correct answers, answering tools and participation by researchers. In the proposed method, we focus on optimizing specifications of annotation for the resources. Consequently, these studies enabled NLP-based systems to solve real examinations that are normally taken by high-school students. Furthermore, we analysed the answering accuracy of the developed systems by comparing the systems' answers with answers given by human test-takers.

To encourage collaborative and interdisciplinary research, the tasks described in this paper can be viewed on our website. In addition, the resources that are translated into English are available.

8. Acknowledgements

We would like to express our gratitude to the JC Educational Institute, National Center for University Entrance Examination and Yoyogi Seminar for their support. This research was supported by the Todai Robot Project at the National Institute of Informatics.

9. References

Angele, J., Moench, E., Oppermann, H., Staab, S., and Wenke, D. (2003). Ontology-based query and answering in chemistry: Ontonova @ project halo. *In Proceedings of the Second International Semantic Web Conference*, pp. 913–928.

Ferrucci, D. A. (2012). Introduction to "this is Watson". *IBM Journal of Research and Development*, 56(3.4), pp. 1:1–1:15.

Ishioroshi, M., Kano, Y., and Kando, N. (2013). An analysis of the questions of the university entrance examination to answer using the question answering system. *The 27th Annual Conference of the Japanese Society for Artificial Intelligence*, 2A4-6.

Iwane, H., Matsuzaki, T., Anai, H., and Arai, N. (2013). A symbolic approach to university entrance examination problems in mathematics and issues combining with national language processing. *The 27th Annual Conference of the Japanese Society for Artificial Intelligence*, 2A4-2.

Kanayama, H., Miyao, Y., and Prager, J. (2012). Answering yes/no questions via question inversion. *In Proceedings of COLING 2012*, pp. 1377–1392.

Miyao, Y. and Kawazoe, A. (2013). University entrance examinations as a benchmark resource for NLP-based problem solving. *In Proceedings of IJCNLP 2013*.

Miyao, Y., Shima, H., Kanayama, H., and Mitamura, T. (2012). Evaluating textual entailment recognition for university entrance examinations. *ACM Transactions on Asian Language Information Processing*, 11(4), p. 13.

Schäfer, U. and Weitz, B. (2012). Combining OCR outputs for logical document structure markup: technical background to the ACL 2012 contributed task. *In Proceedings of the ACL2012 Special Workshop on Rediscovering 50 Years of Discoveries*, pp. 104–109.

World Wide Web Consortium. (2003). Mathematical markup language (mathml) version 2.0 (second edition), www.w3.org/TR/MathML2.

Yokono, H. and Inamura, T. (2013). Physics problem solving based on physics simulation and event recognition from time-series data. *The 27th Annual Conference of the Japanese Society for Artificial Intelligence*, 2A4-3.

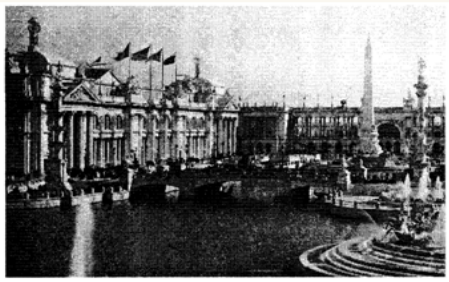
questions

Center-2005--Main-SekaishiA

【1】

歴史上の出来事を記念した建造物や祝祭・儀礼は，民衆の心のよりどころとされるとともに，政治的な意味を持つこともあった。各地の建造物や祝祭・儀礼について述べた次の文章A～Cを読み，以下の問い(問1～9)に答えよ。
(配点 25)

A
(1)1890年代初め，アメリカ合衆国のシカゴで万国博覧会が開催された。「白い都市」と命名された中心会場の建物群に，来場者は驚嘆した。当時のシカゴでは，新しい世代の建築家たちが，(2)19世紀後半に実用化が進んだ鋼鉄材による建築を発展させ，近代的な高層ビルを建設して注目を集めていたのに対し，「白い都市」は，古代ローマなどの旧い時代の建築様式を模していたからである。この「白い都市」の様式は大人気を博した。その理由の一つは，(3)海外へ勢力を拡大しつつあったアメリカ合衆国のイメージが，ローマ帝国の栄光と重ねられたためと考えられる。また，心のよりどころとしての過去がヨーロッパに求められたという面もあった。



「白い都市」。左側の建物には，古代ローマの様式が取り入れられている。右側へ伸びるコロネード(柱廊)も，古代ギリシアやローマでよく使われたものである。

Introduction of the question

問1
下線部(1)の博覧会は，ある歴史上の出来事の400周年を記念して計画されたものである。その出来事について述べた文として最も適当なものを，次の①～④のうちから一つ選べ。
1

The question

RandomAnswerAE
① コロンブスが，西インド諸島に到達した。(正答) <- Correct answer
② イギリスが，北アメリカで最初の植民地を建設した。(×:システム誤答) <- The answer by the system
③ 北アメリカの13の植民地が，イギリスからの独立を宣言した。
④ アメリカ合衆国憲法が制定された。

Result of checking the answer

Process_Log(RandomAnswerAE)

【選択肢1】[keyword:1890年代, count:1, score:2.0] [keyword:アメリカ合衆国, count:17, score:0.12] [keyword:シカゴ, count:2, score:1.0] [keyword:万国博覧会, count:1, score:2.0] [keyword:博覧会, count:2, score:1.0] [keyword:ある, count:15, score:0.14] [keyword:出来事, count:5, score:0.4] [keyword:計画, count:1, score:2.0] [keyword:ある, count:15, score:0.14] [keyword:出来事, count:5, score:0.4] [keyword:コロンブス, count:1, score:2.0] [keyword:西インド諸島, count:1, score:2.0] 【選択肢2】[keyword:1890年代, count:1, score:2.0] [keyword:アメリカ合衆国, count:17, score:0.12] [keyword:シカゴ, count:2, score:1.0] [keyword:万国博覧会, count:1, score:2.0] [keyword:博覧会, count:2, score:1.0] [keyword:ある, count:15, score:0.14] [keyword:出来事, count:5, score:0.4] [keyword:計画, count:1, score:2.0] [keyword:ある, count:15, score:0.14] [keyword:出来事, count:5, score:0.4] [keyword:北アメリカ, count:3, score:0.66] [keyword:最初, count:1, score:2.0] [keyword:植民地, count:4, score:0.5] [keyword:建設, count:7, score:0.28] 【選択肢3】[keyword:1890年代, count:1, score:2.0] [keyword:アメリカ合衆国, count:17, score:0.12] [keyword:シカゴ, count:2, score:1.0] [keyword:万国博覧会, count:1, score:2.0] [keyword:博覧会, count:2, score:1.0] [keyword:ある, count:15, score:0.14] [keyword:出来事, count:5, score:0.4] [keyword:計画, count:1, score:2.0] [keyword:ある, count:15, score:0.14] [keyword:出来事, count:5, score:0.4] [keyword:北アメリカ, count:3, score:0.66] [keyword:植民地, count:4, score:0.5] 【選択肢4】[keyword:1890年代, count:1, score:2.0] [keyword:アメリカ合衆国, count:17, score:0.12] [keyword:シカゴ, count:2, score:1.0] [keyword:万国博覧会, count:1, score:2.0] [keyword:博覧会, count:2, score:1.0] [keyword:ある, count:15, score:0.14] [keyword:出来事, count:5, score:0.4] [keyword:計画, count:1, score:2.0] [keyword:ある, count:15, score:0.14] [keyword:出来事, count:5, score:0.4] [keyword:アメリカ合衆国憲法, count:2, score:1.0]

Process log by the system

Answer_Type:sentence Knowledge_Type:KS

Question type of the question

Figure 7: Example of results page with auxiliary information of submitted answer