University Entrance Examinations as a Benchmark Resource for NLP-based Problem Solving

Yusuke Miyao
National Institute of Informatics
Graduate University for Advanced Studies
yusuke@nii.ac.jp

Ai Kawazoe
National Institute of Informatics
zoeai@nii.ac.jp

Abstract

This paper describes a corpus comprised of university entrance examinations, which is aimed to promote research on NLP-based problem solving. Since entrance examinations are created for quantifying human ability of problem solving, they are a desirable resource for benchmarking NLP-based problem solving systems. However, as entrance examinations involve a variety of subjects and types of questions, in order to pursue focused research on specific NLP technologies, it is necessary to break down entire examinations into individual NLP subtasks. For this purpose, we provide annotations of question classifications in terms of answer types and knowledge types. In this paper, we also describe research issues by referring to results of question classification, and introduce two international shared tasks that employed our resource for developing their evaluation data sets.

1 Introduction

This paper introduces natural language corpora whose source texts are taken from university entrance examinations. This resource has been developed aiming at benchmarking NLP systems for problem solving. In general, entrance examinations are mostly described in natural language, and the goal of reading the text is clear, viz., to solve questions. Therefore, this is an ideal resource for evaluating end-to-end NLP systems that read natural language text, perform some information processing, and output answers.

University entrance examinations have several desirable features to be used for benchmarking NLP-based problem solving. They are carefully designed for empirically quantifying a certain ability of high-school-level students. Therefore, it is not a trivial task for NLP systems to solve university entrance examinations. On the other hand, it is guaranteed that required knowledge is fairly restricted, and legitimate solutions always exist. Despite such artificial restrictions, investigating the entire process of solving entrance examinations is meaningful, because it is expected to reveal true contributions of current NLP technologies to human-like problem solving tasks. In addition, evaluation results are intuitively understandable, and can be compared directly with human performance. This provides us with empirical evidence for analyzing the relationships between human intelligence and artificial intelligence.

While it is now clear that university entrance examinations are a useful resource for NLP benchmarking, it is also true that they will not be appropriate for focusing on individual NLP tasks, because they involve a variety of subjects and types of questions. It is almost hopeless to invent a single clever algorithm that can solve all types of questions. Therefore, it is necessary to break down entire examinations into NLP subtasks that can be investigated solely. For this aim, we annotate classifications of questions, which allow us to isolate specific NLP subtasks for focused research. An important point here is that, question classification allows us to extract individual NLP tasks, but, at the same time, their contributions to entire performance are always accessible. Therefore, our resource is inherently different from NLP resources that focus on monolithic NLP tasks/applications in nature, such as parallel corpora for machine translation research and evaluation data sets for question answering. Owing to question classifications, subsets of our resources have been adopted in international shared tasks for recognizing textual entailment and reading comprehension, which will also be mentioned in this paper.

Standardized tests for high-school-level stu-

International Joint Conference on Natural Language Processing, pages 1357–1365,
Nagoya, Japan, 14-18 October 2013.

dents are widely accepted in the world; examples include SAT (U.S.), Baccalauréat (France), Suneung (Korea), Gao Kao (China), and Center Test (Japan). In this work, we collect source texts from examinations of Center Test in Japan. Center Test has additional advantages as a NLP resource, because texts are free from copyright issues, and questions are given in a multiple-choice style, which allows for automatic evaluation.

The contributions of this paper are summarized as below:

- Describes details of the design of resources developed from university entrance examinations.
- Classifies questions from the NLP point of view, and discusses research issues involved.
- Introduces present use cases of our resources to show their effectiveness.

The resources introduced in this paper are made available for research purposes. As we will see below, this resource involves a variety of research issues in NLP and related AI technologies, and thus collaborative research based on such open resources is indispensable.

2 Motivation

Current NLP corpora can be classified into two types. One is to focus on specific fundamental NLP technologies, such as Penn Treebank (Marcus et al., 1993) developed for parsing research. The other is application-oriented data sets, meaning that corpora are used for evaluating specific NLP applications, such as machine translation and question answering (Voorhees and Buckland, 2012; Kando et al., 2011). However, despite significant advancement achieved by these resources, it is still unclear how far current NLP technologies have approached human intelligence, in particular, about the ability of generic problem solving. In the current NLP, research topics are inherently determined when corpora are developed, and there is no room for investigating performances of NLP technologies from a holistic view.

Our primary motivation to develop a corpus of university entrance examinations is to provide an open data set that encourages research on end-to-end NLP systems for problem solving. By investigating the entire process of solving various types of examinations, we expect to recognize contributions of current NLP technologies and methodologies for integrating them from a holistic view.

For this purpose, university entrance examinations have several advantages as a benchmark, as explained below.

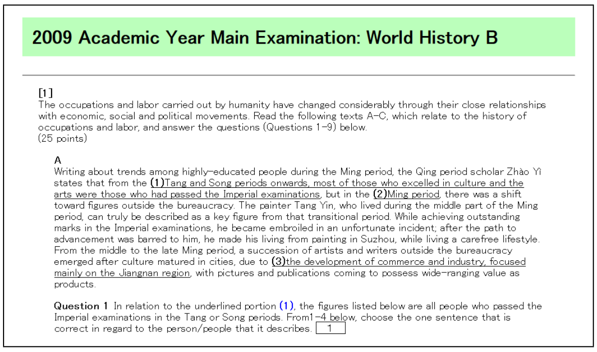
Open but restricted real-world task Since university entrance examinations are developed for empirically quantifying a certain ability of humans, solving them is not a toy task. However, because questions must be solvable by high-school students, this task requires much smaller knowledge space than contemporary NLP applications such as Web-scale question answering. Therefore, we can focus on algorithms of problem solving rather than relying on huge data.

Fair and clear evaluation criteria Intrinsically, standardized tests of university entrance examinations are carefully designed to guarantee fairness. To be more concrete, it is guaranteed that correct answers always exist, and everybody agrees with correct answers. This means that gold standard data is given at almost perfect agreement, which is an ideal feature as a benchmark. In addition, questions of Center Test are given in a multiple-choice style, which allows for automatic evaluation.

Necessity of heterogeneous NLP tasks Since university entrance examinations are aimed at quantifying various aspects of human intelligence, various forms of questions are developed in a variety of subjects. Therefore, to develop an end-to-end system to solve questions, multiple NLP components have to work in a collaborative manner. In some cases, they have to be connected to non-NLP components, such as mathematical solvers and ontology-based inference engines. Thus, investigating entrance examinations promotes interdisciplinary research within NLP, as well as with outside of NLP.

This also indicates the difficulty of focused research on individual NLP technologies. Therefore, we provide annotations for question classification, which enables us to extract a subset of examinations that is relevant to focused NLP tasks (see Section 3 and 4).

Comparison to human performance In our framework, overall performance of NLP systems is quantified as *scores*, which are directly comparable with human performance. We can therefore



2009 Academic Year Main Examination: World History B

[1]
The occupations and labor carried out by humanity have changed considerably through their close relationships with economic, social and political movements. Read the following texts A–C, which relate to the history of occupations and labor, and answer the questions (Questions 1–9) below.
(25 points)

A
Writing about trends among highly-educated people during the Ming period, the Qing period scholar Zhào Yì states that from the (1)Tang and Song periods onwards, most of those who excelled in culture and the arts were those who had passed the Imperial examinations, but in the (2)Ming period, there was a shift toward figures outside the bureaucracy. The painter Tang Yin, who lived during the middle part of the Ming period, can truly be described as a key figure from that transitional period. While achieving outstanding marks in the Imperial examinations, he became embroiled in an unfortunate incident; after the path to advancement was barred to him, he made his living from painting in Suzhou, while living a carefree lifestyle. From the middle to the late Ming period, a succession of artists and writers outside the bureaucracy emerged after culture matured in cities, due to (3)the development of commerce and industry, focused mainly on the Jiangnan region, with pictures and publications coming to possess wide-ranging value as products.

Question 1 In relation to the underlined portion (1), the figures listed below are all people who passed the Imperial examinations in the Tang or Song periods. From1–4 below, choose the one sentence that is correct in regard to the person/people that it describes. 1

Figure 1: A screenshot of university entrance examination (2009 Center Test World History B)

recognize advantages and disadvantages of current NLP technologies compared to human problem solving. We can also empirically investigate relationships between NLP technologies and human ability of language understanding. Although NLP systems do not necessarily imitate human language processing, it is scientifically interesting to explore such relationships.

A possible criticism to our resource would be on its practical value. It is obvious that solving entrance examinations by NLP systems does not have any practical value. However, our intention is on the investigation of the whole process of problem solving. We believe that such holistic analysis of problem solving contributes to better understanding of current NLP technologies.

Another criticism would be on the reason why we focus on *university* entrance examinations, rather than easier questions, such as elementary school tests and TOEFL-like English tests. In this respect, we argue that university entrance examinations are most appropriate as a NLP benchmark. In preliminary investigation, we found that easy tests like elementary school tests rely more on generic knowledge and common sense, probably because the knowledge space and the vocabulary that can be used are too restricted. On the other hand, examinations in more expertized areas, such as medical license tests, are more uniformly developed and involve a less variety of NLP subtasks.

3 Resources

This section describes details of the resources we have developed. As we stated in Section 2, our primary aim is to investigate the entire process of problem solving related to natural language understanding. Therefore, we prepare data sets in which non-linguistic structures are already solved. Figure 1 shows a screenshot of an actual examination, while Figure 2 shows its XML version, where document structures, such as *instruction* and *question* are already given. Relying on these structure annotations, we can easily extract relevant texts for research, such as questions of interest, and their related instructions, etc.

```
<exam subject="World History B (main)" year="2009">
  <title>
    2009 Academic Year Main Examination: World History B
  </title>
  <question id="Q1" minimal="no">
    <label>[1]</label>
    <instruction>
      The occupations and labor carried out by humanity have
      changed considerably through their close relationships
      with economic, social and political movements...
    </instruction>
    <data id="D0" type="text">
      <label>A</label>
      Writing about trends among highly-educated people during
      the Ming period, the Qing period scholar Zhao Yi states
      that from the <uText id="U1"><label>(1)</label>Tang and
      Song periods onwards, most of those who excelled in
      culture and the arts were those who had passed the
      Imperial examinations</uText>, ...
    </data>
    <question id="Q2" minimal="yes" answer_type="sentence"
            knowledge_type="KS">
      <label>Question 1</label>
      <instruction>
        In relation to the underlined portion <ref target="U1">
        (1)</ref>, the figures listed below are all people who
        passed the Imperial examinations in the Tang or Song
        periods.  From 1-4 below, choose the one sentence that
        is correct in regard to the person/people that it
        describes.
      </instruction>
      <ansColumn id="A1">1</ansColumn>
      <choices anscol="A1">
        <choice><cNum>(1)</cNum>Ouyang Xiu and Su Shi are
          writers representative of the Tang period.</choice>
        <choice><cNum>(2)</cNum>Yan Zhenqing is a calligrapher
          representative of the Song period.</choice>
        <choice ra="yes"><cNum>(3)</cNum>Wang Anshi, who lived
          during the Song period, carried out reforms called the
          New Policies (xin fa).</choice>
        <choice><cNum>(4)</cNum>Qin Hui came into conflict with
          the party in favor of war, concerning the relationship
          with the Yuan.</choice>
      </choices>
    </question>
    ...
```

Figure 2: XML data of university entrance examination (2009 Center Test World History B)

	Exam (orig.)	Exam (Eng.)	Textbook
# subjects	11	5	6
# files	571	25	11
# questions	17260	771	N/A
# sentences	79211	5236	35183

Table 1: Statistics of corpora

Basically, all the questions of Center Test are given in a multiple-choice style. The answer data is also given in the XML format. Given answer data, it is almost trivial to compute examination scores, while we also provide tools for automatic evaluation and visualization.

3.1 Corpora

In this work, we collected PDFs and source texts from National Center Test for University Admis-

	Exam (orig.)	Exam (Eng.)	Textbook
Document structure	✓	✓	✓
Question types	✓	✓	
Technical terms	✓		✓
Dependency trees			✓
Coreferences			✓

Table 2: Summary of annotated resources

sion in Japan (a.k.a. Center Test).[1] Center Test is a nation-wide standardized test for university admission in Japan, and almost all high-school students who aim to enter a university in Japan take this exam. Therefore, questions are carefully designed in order to accurately quantify achievement levels of high-school students.

Table 1 shows a summary of source texts. The Center Test corpus includes examination texts from eleven subjects, namely, *World History*, *Japanese History*, *Modern Society*, *Politics & Economics*, *Ethics*, *Physics*, *Chemistry*, *Biology*, *Mathematics*, *Japanese (native language)*, and *English (foreign language)*, used in the years from 1990 to 2011.[2] In each year, a single main examination and a couple of additional examinations are available. In total, we have obtained 571 examinations, each of which contains 30-50 questions. The numbers of questions and sentences are also shown in the table, indicating a comprehensive amount as a corpus for NLP research.

While original texts are in Japanese (except for English tests), a part of examinations of World History, Politics & Economics, Physics, Chemistry, and Biology, are translated into English, in order to allow researchers to work on English NLP as well as cross-lingual NLP.

In addition to examinations, we collected textbooks of World History, Japanese History, Modern Society, Politics & Economics, Ethics, and Biology. Questions in these subjects often ask to recognize *facts*, such as historical facts and biological processes. Because textbooks describe such facts, we can use textbooks as knowledge sources for solving such questions. In fact, these textbooks were adopted as knowledge sources in a shared task on recognizing textual entailment (Section 5).

For these text data, we annotated document structures, question types, technical terms, dependency trees, and coreferences (see Table 2), which are explained in the consecutive sections.

`question`	A question region including outermost question areas and minimal areas. An ID is assigned to each element. Question regions that do not include other question regions are given the attribute `minimal="yes"`, indicating smallest units of questions.
`instruction`	A statement or an instruction for a question.
`data`	Data provided to test-takers of reference, including not only texts but also images, tables, graphs, etc.
`label`	A label such as section numbers, question numbers, identifiers of text fragments, etc.
`ansColumn`	An identifier of an answer column. Each answer column is given an unique ID, which is referred to in answer data.
`choices`	A set of choices.
`choice`	An individual choice. The attribute `ra="yes"` denotes correct choices.
`cNum`	An identifier of a choice.
`ref`	A symbol that refers to another text fragment, such as underlined texts. A referred text fragment is denoted by the attribute `target`.
`uText`	An underlined text fragment. A unique ID is assigned when the text fragment is referred to by `ref`.

Table 3: Document structure tags

3.2 Document Structure Annotation

Examination texts are highly structured, while the automatic recognition of document structures is still a challenging task (Schäfer and Weitz, 2012). Therefore, our resource is provided with human-annotated document structures in the form of XML, as shown in Figure 2. Table 3 shows an excerpt of XML tags used for the annotation.[3] In addition to document structures, texts are also annotated with extra-linguistic markups, such as underlines (`uText`) and references (`ref`).

Owing to the document structure annotations, users can easily extract questions and relevant text regions. For example, a complete list of individual questions can be obtained by extracting elements `<question minimal="yes">`, and their corresponding answer columns and choices can also be extracted easily. Furthermore, text fragments referred to by a label like "`(1)`" can be obtained by following the attribute `target` of `ref` (see the example in Figure 2).

Formulas play crucial roles in examinations of Science and Mathematics. Although understanding of semantics of formulas is indispensable, the

[1] `http://www.dnc.ac.jp/`

[2] Examination corpora of Geography, Geology, and General Science are under construction.

[3] The complete list and the definitions of tags are provided together with annotated corpora.

Answer types	
sentence	Choices are described by sentences.
term	Choices are described by terms (e.g. person names).
image	Choices are represented by images or parts of an image.
formula	Choices are represented by formulas.
combination	Choices are described by a combination of sentences, terms, etc.
Knowledge types	
KS	An external knowledge source (e.g. textbooks) is required.
RT	Reading comprehension of a text given within a question is required.
IC	Image comprehension is necessary.
GK	General knowledge is required.
DM	Domain-specific inference (e.g. laws of dynamics) is required.

Table 4: Top-level categories for question classification

From 1-4 below, choose the one sentence that is correct in regard to the person/people that it describes.
(1) Ouyang Xiu and Su Shi are writers representative of the Tang period.
(2) Yan Zhenqing is a calligrapher representative of the Song period.
(3) Wang Anshi, who lived during the Song period, carried out reforms called the New Policies (xin fa).
(4) Qin Hui came into conflict with the party in favor of war, concerning the relationship with the Yuan.

Figure 3: A true-or-false question

semantic analysis of formulas is not trivial and is beyond the scope of NLP research. Therefore, we marked up all formulas that appear in examination texts with MathML.

3.3 Question Type Annotation

As mentioned in Section 2, university entrance examinations involve a variety of NLP subtasks, which prevents us from focusing on individual NLP tasks. In order to extract questions of interest, we annotate each question with classification categories. By extracting questions assigned specific categories, we can obtain a subset of examinations on which isolated NLP tasks can be studied.

Table 4 shows a subset of top-level categories for question classification. Questions are classified according to two perspectives. The *answer type* specifies the format of answers. For example, if choices are presented with a sentence, it is assigned the category `sentence`, which typically indicates *true-or-false questions* as exemplified in Figure 3. If `term` is assigned, the question is likely to be a *factoid-style question*. These categories are further classified into sub categories; for example, `term` is divided by term categories (e.g. *person names*), while `combination` is further classified with elements of combinations. In total, 25 answer type categories are annotated.

The *knowledge type* describes the types of knowledge that are necessary to solve the question. While Table 4 shows representative top-level categories, they are further divided into fine-grained categories, and, in total, 90 knowledge type categories are annotated. For example, `KS` indicates that to answer the question requires referring to an external knowledge source like textbooks (e.g. Figure 3). This type of questions typically appear in examinations of Social Studies. `RT` indicates a similar type of questions, but necessary information is given as a text within an examination. Therefore, reading comprehension is necessary. `DM` means domain-specific inference is necessary depending on a subject. Individual domains are annotated with finer-grained categories, like *physical mechanics* and *electromagnetics*. For example, to solve questions of physical dynamics, calculation of formulas based on laws of dynamics is required. `GK` indicates any other type of knowledge, such as *typical situations and reactions in a restaurant*. Since the knowledge space is not strictly restricted, we suppose this is the most difficult type of questions for NLP systems.

In Section 4, we will discuss research issues involved in our resource, by observing results of question classification described here.

3.4 Linguistic Annotation

In addition to document structures and question classifications, we have developed resources annotated with technical terms, dependency trees, and coreference relations, in order to support research on fundamental NLP tools.

Technical terms are annotated to examinations and textbooks of World History, Japanese History, Modern Society, Politics & Economics, Ethics, and Biology. These subjects are selected because, as we will see in Section 4, a majority of questions in these subjects are either true-or-false or factoid-style questions, which can be approached by searching textbooks for relevant evidences. In such a scenario, technical terms are crucial keys for accurate search. For example, to solve the question shown in Figure 3, it is necessary to correctly recognize relationships among named entities like *Ouyang Xiu* and *the Tang period*.

	W. Hist.		J. Hist.		M. S.		P. & E.		Ethics		Bio.	
instance	8864	(52.1)	5876	(35.5)	2558	(24.3)	2279	(22.6)	2556	(28.8)	90	(0.6)
class	5592	(32.9)	7808	(47.2)	5084	(48.4)	4779	(47.3)	1237	(14.0)	2382	(15.6)
both	2557	(15.0)	2848	(17.2)	2867	(27.3)	3039	(30.1)	5072	(57.2)	12790	(83.8)
# terms	17013	(100.0)	16532	(100.0)	10509	(100.0)	10097	(100.0)	8865	(100.0)	15262	(100.0)
# sentences	5797		5571		3674		3352		3245		4215	

Table 5: Statistics of technical term annotations

	W. Hist.		J. Hist.		M. S.		P. & E.		Ethics	
True-or-false question	1854	(73.6)	1308	(55.6)	1102	(79.5)	805	(88.6)	656	(81.8)
Factoid question	464	(18.4)	557	(23.7)	192	(13.9)	62	(6.8)	128	(16.0)
Reading comprehension	102	(4.0)	146	(6.2)	43	(3.1)	3	(0.3)	88	(11.0)
General knowledge	1	(0.0)	0	(0.0)	92	(6.6)	8	(0.9)	114	(14.2)
Image comprehension	222	(8.8)	198	(8.4)	111	(8.0)	101	(11.1)	17	(2.1)
# questions	2519	(100.0)	2351	(100.0)	1386	(100.0)	909	(100.0)	802	(100.0)

Table 6: Classification of questions (Social Studies)

We analyzed our corpus of examinations and textbooks, and developed an ontology of technical terms, which involves 72 categories. Their occurrences in examinations and textbooks are annotated manually. Table 5 shows statistics of technical term annotations on examinations.[4] Annotated terms not only include typical named entities (i.e. *instances*) like *person names*, but also include *class concepts* that describe domain-specific abstract terms, such as *genetic trait*. Several categories may include both instance terms and class terms (denoted as "both" in the table); for example, the category *artwork* includes *Isenheim Altarpiece* (instance) and *miniature* (class). Interestingly, the distributions of terms imply characteristics of each subject; for example, World History concerns named entities, while Biology is more focused on abstract concepts.

Dependency trees and coreferences are annotated in order to assess the performance of fundamental NLP tools including dependency parsers and coreference resolution systems. It is expected that these NLP tools work reasonably well on texts of examinations and textbooks, because in general texts in these domains are written in an unambiguous and easy-to-understand way. Currently, we have annotated a subset of a textbook of World History, and will extend the data as necessary.

4 Analysis of Questions

This section discusses research issues involved in solving the university entrance examinations, by analyzing question classification results. Table 6, Table 7, and Table 8 show the number of questions and its ratio (shown in brackets) classified into each category, for examinations of Social Studies, Science, and English/Japanese, respectively.[5] These classifications are obtained from answer type and knowledge type annotations introduced in Section 3, while classification categories are summarized and reinterpreted for readability.

For Social Studies (Table 6), it is obvious that most of the questions can be classified into *true-or-false* and *factoid-style questions*. Low ratios of *reading comprehension* and *general knowledge* indicate that most of the questions can be solved only by referring to external knowledge sources. This is promising, because current question answering methods and/or search-based methods would suffice. As we will see in Section 5, these types of questions have already been tackled in international shared tasks.

For Science subjects (Table 7), Biology looks similar to Social Studies, while results on Physics and Chemistry reveal different characteristics. Almost all questions in these subjects are annotated as *domain-specific inference*, indicating that simply referring to knowledge sources does not suffice, and inference engines, such as formula processing modules and ontology-based reasoning, will be required. In particular, nearly half of the questions in Physics are answered in *formulas*, indicating necessity of formula processing. The integration of NLP components with formula processing should be an interesting frontier.

Results on English and Japanese are totally different. They contain questions at different levels of difficulty. Questions that ask *lexi-*

[4] The statistics for textbooks is omitted for space limitation, but the tendency of the distribution is similar.

[5] The sum of ratios exceeds 100%, because a question might be classified into multiple categories.

	Physics		Chemistry		Biology	
True-or-false question	390	(24.4)	578	(32.5)	938	(52.5)
Factoid question	239	(15.0)	367	(20.7)	564	(31.6)
Formula	683	(42.8)	399	(22.5)	136	(7.6)
Domain-specific inference	1594	(99.9)	1764	(99.3)	522	(29.2)
Reading comprehension	0	(0.0)	3	(0.2)	31	(1.7)
General knowledge	64	(4.0)	0	(0.0)	2	(0.1)
Image comprehension	1105	(69.3)	291	(16.4)	420	(23.8)
# questions	1595	(100.0)	1776	(100.0)	1767	(100.0)

Table 7: Classification of questions (Science)

	English		Japanese	
Lexical knowledge	1085	(44.8)	778	(36.4)
Grammatical knowledge	703	(29.0)	126	(5.9)
Literature knowledge	0	(0.0)	36	(1.7)
Reading comprehension	892	(36.8)	872	(40.8)
Situation comprehension	1213	(50.1)	232	(10.8)
Rhetorical structure	0	(0.0)	173	(8.1)
Translation	0	(0.0)	465	(21.7)
Image comprehension	402	(16.6)	0	(0.0)
# questions	2423	(100.0)	2139	(100.0)

Table 8: Classification of questions (English and Japanese)

cal/grammatical/literature knowledge should be tractable for current NLP systems. However, a significant portion of questions involves *reading comprehension*, which is an outstanding problem in NLP. Research on reading comprehension is recently emerging (Peñas et al., 2011a; Peñas et al., 2011b), while the achievements are still far from satisfactory. Furthermore, English examinations involve a large portion of *situation comprehension* (e.g. selecting an appropriate conversation in a restaurant) and *image comprehension* (e.g. choosing an appropriate description of a given image), which are enormously difficult research issues. In this respect, achieving high scores in English tests can be an ultimate goal of the present effort.

While Mathematics is not shown in the tables, it is totally different from the subjects discussed above. Solving mathematics questions essentially consists of two components. One is natural language understanding, which converts text expressions into mathematical formulas, and the other is mathematical formula processing. Therefore, the primary research issues are to design an interface between the two components, and to increase the accuracy of the two components.

5 Use Cases

In addition to individual studies, two international shared tasks have adopted subsets of our resources for creating their evaluation data sets. Here we briefly introduce these works, which prove the effectiveness of our resources for NLP research.

t: In the period of Emperor Shenzong in the Baisong dynasty, Wang Anshi introduced and promulgated his reform policy (xin fa).

h: Wang Anshi, who lived during the Song period, carried out reforms called the New Policies (xin fa).

Figure 4: A text pair for recognizing textual entailment created from a World History examination.

5.1 Recognizing Textual Entailment

The RITE task at the NTCIR conference is a shared task on recognizing textual entailment (Watanabe et al., 2013). RITE consisted of several subtasks, one of which adopted a subset of our resource as an evaluation data set. As described in Section 4, a significant portion of Social Studies consists of true-or-false questions, which can be solved by recognizing textual entailment relations. For example, Figure 3 shows a typical true-or-false question. Test-takers are required to find relevant facts from their knowledge, and judge whether each sentence is true or false. For NLP systems, this corresponds to finding an evidential text from a knowledge source like a textbook or Wikipedia, and judge whether a text fragment in the knowledge source *entails* each sentence. In fact, by extracting a relevant text from Wikipedia, we can create a text pair as shown in Figure 4, which can be used as evaluation data for recognizing textual entailment.

In the RITE task, true-of-false questions are extracted from four subjects, namely, World History, Japanese History, Modern Society, and Politics & Economies, while evidential texts are provided from Wikipedia and textbooks. In total, 510 text pairs are provided as a training set, and 448 pairs as a test set.

While the RITE task reveals that recognizing textual entailment can be applied directly to true-or-false questions, this is not the only solution for this type of questions. Actually, Kanayama et al. (2012) demonstrated a method for applying a factoid-style question answering system to solve true-or-false questions, and evaluated their system using a World History portion of our resource. This reveals that a variety of approaches can be attempted to achieve the same goal, i.e., solving examinations.

5.2 Reading Comprehension

Shared tasks called Question Answering for Machine Reading Evaluation (QA4MRE) at the CLEF conferences (Peñas et al., 2011a; Peñas et al., 2011b) have been focusing on NLP technologies for reading comprehension tasks. In the task setting of QA4MRE, a short document is given, and systems are required to answer multiple-choice questions by reading the given document. Because given texts are small, methods that rely on huge texts as in typical question answering systems cannot be applied, while accurate and deep analysis of given texts is necessary. Original evaluation data sets for QA4MRE have been developed from scratch, focused on several topics like "Aids" and "Climate Change."

In QA4MRE at CLEF 2013,[6] a pilot task that uses reading comprehension questions from English tests of our resource has been organized. The novelty of this pilot task is that questions are originally developed for assessing human English ability, rather than specifically developed for NLP system evaluation. Therefore, it is expected that various aspects of human natural language understanding appear in solving such questions.

6 Related Work

Recent advancement of empirical NLP owes much to language resources, such as annotated corpora and lexicons. Language resources to date have been developed specifically for focused NLP tasks, such as syntactic/semantic parsing, coreference resolution, and word sense disambiguation (Marcus et al., 1993; Kingsbury and Palmer, 2002; Hovy et al., 2006; Ide et al., 2010; Tateisi et al., 2005; Kawahara et al., 2002; Iida et al., 2007). Another type of corpora has been developed for evaluating NLP applications, such as machine translation and question answering, which are often provided in application-oriented evaluation campaigns (Voorhees and Buckland, 2012; Kando et al., 2011; Catarci et al., 2012). In other words, the development of language resources is initiated by the demand for NLP tasks/applications. However, the resources presented in this paper are motivated in an opposite way. We start from texts that involve problem solving by humans, i.e., university entrance examinations, and by analyzing them we can identify NLP tasks that we have to tackle with. It can be said that the framework and the resources described in this paper provide another direction of NLP research.

NLP research that develops benchmark data from questions originally designed for evaluating human performance has also been emerging. For example, the Halo project (Angele et al., 2003) targeted Chemical tests, while IBM's Deep QA (Ferrucci, 2012) employed factoid-style quizzes. However, their benchmark data sets are not open, and therefore collaborative research based on shared standard data cannot be pursued. Collaborative research is indispensable for our purpose, because entrance examinations involve a variety of NLP subtasks, and a single research group cannot solve the entire problem. Therefore, it is necessary to develop open resources as described in this paper.

7 Conclusion

We have introduced an NLP resource that is developed from university entrance examinations, aiming at the development and the evaluation of end-to-end NLP systems for problem solving. In total 571 examinations are collected from 11 subjects, involving 17260 individual questions, revealing a comprehensive resource for NLP benchmarking.

While the ultimate goal is to develop an integrated NLP system that can solve a wide range of questions, this also means it is difficult to focus on individual NLP subtasks. Therefore, we annotated question classifications so that users can extract fragments of the resource that are relevant to a focused NLP subtask. In fact, subsets of our resources have already been adopted by two international shared tasks, namely, NTCIR RITE for recognizing textual entailment, and CLEF 2013 QA4MRE, for reading comprehension.

In order to encourage collaborative and interdisciplinary research, the resources described in this paper are made available for research purposes.[7]

[6] http://celct.fbk.eu/QA4MRE/

[7] The resource is available at http://21robot.org/.

Acknowledgments

We gratefully acknowledge the National Center for University Entrance Examination, JC Educational Institute and Tokyo Shoseki Co., Ltd. for original Center Test data and textbook data.

References

J. Angele, E. Moench, H. Oppermann, S. Staab, and D. Wenke. 2003. Ontology-based query and answering in chemistry: OntoNova @ Project Halo. In *Proceedings of the Second International Semantic Web Conference*.

T. Catarci, P. Forner, D. Hiemstra, A. Penas, and G. Santucci, editors. 2012. *Information Access Evaluation. Multilinguality, Multimodality, and Visual Analytics*. Springer.

D. A. Ferrucci. 2012. Introduction to "this is watson". *IBM Journal of Research and Development*, 56(3.4).

Eduard Hovy, Mitchell Marcus, Martha Palmer, Lance Ramshaw, and Ralph Weischedel. 2006. OntoNotes: The 90% solution. In *Proceedings of HLT/NAACL 2006*.

Nancy Ide, Christiane Fellbaum, Collin Baker, and Rebecca Passonneau. 2010. The manually annotated sub-corpus: a community resource for and by the people. In *Proceedings of the ACL 2010 Conference Short Papers*, pages 68–73.

Ryu Iida, Mamoru Komachi, Kentaro Inui, and Yuji Matsumoto. 2007. Annotating a Japanese text corpus with predicate-argument and coreference relations. In *Proceedings of Linguistic Annotation Workshop*, pages 132–139.

Hiroshi Kanayama, Yusuke Miyao, and John M. Prager. 2012. Answering yes/no questions via question inversion. In *Proceedings of COLING 2012*.

Noriko Kando, Daisuke Ishikawa, and Miho Sugimoto, editors. 2011. *Proceedings of the 9th NTCIR Workshop Meeting on Evaluation of Information Access Technologies: Information Retrieval, Question Answering and Cross-Lingual Information Access*.

D. Kawahara, T. Kurohashi, and K. Hasida. 2002. Construction of a Japanese relevance-tagged corpus. In *Proceedings of the 8th Annual Meeting of the Association for Natural Language Processing*, pages 495–498. (In Japanese).

P. Kingsbury and M. Palmer. 2002. From Treebank to PropBank. In *Proceedings of LREC 2002*.

M. P. Marcus, B. Santorini, and M. A. Marcinkiewicz. 1993. Building a large annotated corpus of English: The Penn Treebank. *Computational Linguistics*, 19(2):313–330.

Anselmo Peñas, Eduard Hovy, Pamela Forner, Àlvaro Rodrigo, Richard Sutcliffe, Corina Forascu, and Caroline Sporleder. 2011a. Overview of QA4MRE at CLEF 2011: Question answering for machine reading evaluation. In *CLEF 2011 Labs and Workshop Notebook Papers*, pages 19–22.

Anselmo Peñas, Eduard Hovy, Pamela Forner, Àlvaro Rodrigo, Richard Sutcliffe, Caroline Sporleder, Corina Forascu, Yassine Benajiba, , and Petya Osenova. 2011b. Overview of QA4MRE at CLEF 2012: Question answering for machine reading evaluation. In *CLEF 2011 Labs and Workshop Notebook Papers*.

Ulrich Schäfer and Benjamin Weitz. 2012. Combining OCR outputs for logical document structure markup: technical background to the ACL 2012 contributed task. In *Proceedings of the ACL-2012 Special Workshop on Rediscovering 50 Years of Discoveries*, pages 104–109.

Y. Tateisi, A. Yakushiji, T. Ohta, and J. Tsujii. 2005. Syntax annotation for the GENIA corpus. In *Proceedings of IJCNLP 2005 Companion Volume*.

E. M. Voorhees and Lori P. Buckland, editors. 2012. *Proceedings of the Twenty-First Text REtrieval Conference (TREC 2012)*.

Yotaro Watanabe, Yusuke Miyao, Junta Mizuno, Tomohide Shibata, Hiroshi Kanayama, C.-W. Lee, C.-J. Lin, Shuming Shi, Teruko Mitamura, Noriko Kando, Hideki Shima, and Kohichi Takeda. 2013. Overview of the Recognizing Inference in Text (RITE-2) at NTCIR-10. In *Proceedings of the 10th NTCIR Conference*.