Liu, C.-C. et al. (Eds.) (2014). Proceedings of the 22nd International Conference on Computers in Education. Japan: Asia-Pacific Society for Computers in Education

The impact of A.I. on education – Can a robot get into The University of Tokyo?

Noriko H. ARAI* & Takuya MATSUZAKI
Research Center for Community Knowledge, National Institute of Informatics, Japan
*arai@nii.ac.jp

Abstract: The "Todai Robot Project (Can a robot get into The University of Tokyo?)" was initiated by the National Institute of Informatics in 2011 as an AI grand challenge. The goal of the project was to create an AI system that answered real questions on university entrance examinations consisting of two parts, i.e., multiple-choice style national standardized tests and a written test that included short essays. The tasks naturally required the development of ground-breaking underlying technologies in research areas including natural language processing, image processing, speech recognition, automated theorem proving, computer algebra, and computer simulations. It simultaneously required interdisciplinary research synthesis.

Our software took a (digitalized and annotated version of) mock National Center Test for University Admissions (NCTUA), provided by a prep. school, with more than five thousand students. The results revealed that its abilities were still far below the average scores of entrants to The University of Tokyo. However, it was competent enough to pass the entrance exams of 404 out of 744 private universities in Japan. The rapid rise of new AI technologies may negatively affect the labor market in the short term, and we will need to re-design our education systems that were intended to be optimal for the modern industrial society.

Keywords: artificial intelligence, comparative advantage, labor market, automatization

1. Introduction

The question "Can an AI robot pass university entrance examinations?" would have been considered a bad joke ten years ago. However, the tide is turning. IBM Watson defeated a human champion on the American quiz show, "Jeopardy!" in February 2011 (Ferruci et al., 2012), Siri on smartphones answers many spoken questions like "What is the weather like?", and the Russian chatter bot Eugene was claimed to have passed the Turing test in 2013. All of these were considered to be impossible dreams when the Fifth Generation Computer Systems Project ended in failure in the early 1990s.

AI armed with machine learning technologies often surprises us by demonstrating its power in classification problems like medical analysis, and in optimization problems like automated car driving. This therefore raises a natural question. How far can a machine approach human intelligence? Will it replace human workers, especially white-collar workers in the near future? What particular kinds of jobs are being threatened? Will there be any economic returns to higher education when AI is smart enough to "learn" better than most of us? Do we have to set different goals for higher education in the age of AI?

Before jumping to any conclusions, we have to carefully study both the possibilities of and limitations with current AI technologies in comparison to human intelligence especially in the skills, knowledge, and expertise that have traditionally believed to have only been acquired through higher education. That was the motivation when we started the Todai Robot Project ("Can a robot get into The University of Tokyo?") in 2011.

It is of course questionable whether or not university entrance examinations are appropriate to test genuine intelligence; there have always been criticisms that entrance examinations only measure the amount of knowledge and skills in specific fields but not generic skills. However, no one can deny that there is general agreement by the public that we can measure high school students' educational

achievements and developments. These educational attainments are simultaneously known to signal employers of the value of potential employees (Arai, 2013).

University entrance examinations in eastern Asian countries, including Japan, are known to be quite competitive and they cover various skill areas and fields. More than half a million high school graduates take the National Center Test for University Admissions (NCTUA), which is a standardized multiple choice style test, every year, and the top 0.5% students are admitted to The University of Tokyo in Japan. We tried to elucidate in what types of intellectual skills human beings possessed comparative advantages over machines through the Todai Robot Project. It should help us design the education reforms necessary for youth to survive in the AI age.

This paper reports the current status of the Todai Robot Project including the underlying technologies we have developed thus far, and the results we obtained from evaluations. We also discuss problems with the current educational system that need to be addressed to increase human capital in the AI age.

2. Background and Related Work

The possibility of strong or weak AI has been discussed since Turing's monumental paper (Turing, 1950). More than a half-century of research in AI has failed to produce any firm evidence that a symbol-system can manifest human levels of general intelligence. The body of work on this discussion is too broad to be discussed here (see, e.g., Craine 2003 for an overview). We just need to state that no theory has yet succeeded in bridging the gaps between the ambiguous formulation of contextual knowledge in a powerful language (e.g., a natural language and images) or its sounds, which are reproducible and computational representations in a formal language (e.g., a programming language). Crane stated (2003) "however natural it seems in the case of our own language, words do not have their meaning in and of themselves. … They do not have their meaning 'intrinsically'." These types of research, although philosophically interesting, do not suggest in what particular types of intellectual skills we, as human beings, possess comparative advantages over AI in the labor market.

Much research has recently been done on the impact of computerization on the labor market (Bresnahan, 1999, Brynjolfsson and McAfee, 2011, Frey and Osborne 2013). Frey and Osborne estimated that about 47% of US jobs were being threatened by computerization, and they further provided evidence that wages and educational attainments exhibited a strong negative relationship with an occupation's probability of computerization. Their analysis was based on the modified statistical task model introduced by Autor et al. (2003) that postulated (a) computers were more substitutable for human labor in routine relative to non-routine tasks and (b) a greater intensity of routine inputs increased the marginal productivity of non-routine inputs. However, it is not obvious what kinds of tasks can be determined to be "routine" and "non-routine" since the difficulty of particular types of tasks for AI is not determined by the level of "intelligence". Playing professional-level Shogi (Japanese chess) is apparently more non-routine and requires higher-levels of intelligence than classifying illustrated books, even though the former is much easier than the latter for current state-of-the-art AI.

Several international evaluations and workshops such as the Text REtrieval Conference (TREC)[1], the Cross Language Evaluation Forum (CLEF)[2], and NII Testbeds and Community for Information access Research (NTCIR)[3] have developed varieties of tasks for information retrieval and summarization. Most of them have been designed to evaluate applications for specific functions, and it is impossible to see how well near-term AI is able to perform on more complex tasks.

The Allen Institute for Artificial Intelligence's Project Aristo[4] shared a similar motivation with us, and set a similar goal to build software that would enable computers to "learn" from textbooks, ask questions, and draw tentative conclusions.

[1] http://trec.nist.gov/

[2] http://www.clef-initiative.eu/

[3] http://research.nii.ac.jp/ntcir/index-en.html

[4] http://www.allenai.org/TemplateGeneric.aspx?contentId=8

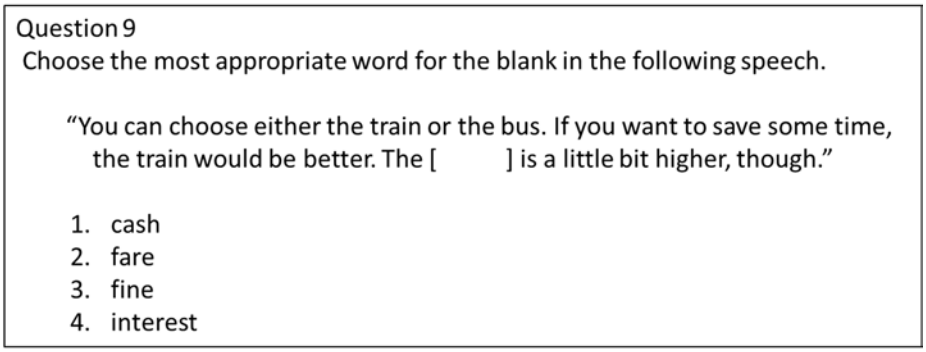
3. Typical Questions Asked

The University of Tokyo has adopted a two-stage selection system for its entrance examination: The first stage involves multiple choice style national standardized tests (NCTUA). One must take seven subjects including Mathematics, English as a Second Language (ESL), Japanese, Social Studies, and Natural Science and achieve a high score with an accuracy of ~80% to take the second written stage prepared by The University of Tokyo to test proficiency in four subjects.

3.1 Fill-In-The-Blanks Style Questions/ Factoids

Fill-in-the-blanks and factoids are the most typical styles of questions asked to measure the amount of knowledge in educational testing. Mechanization of these types of question answering tests has been well studied especially over the past two decades along with the rise in Web search technologies and the explosion in the amount of born-digital documents (Ravichandran and Hovy, 2002; Ferruci et al., 2010). The combination of N-gram language models, ranking models for document retrieval, co-occurrence rules, and category filtering explains the process of finding proper keywords for these types of questions in many cases, which was demonstrated in the Jeopardy! Challenge by Watson. There is already a consensus that machines possess absolute advantages over human beings in these types of intelligence.

Figure 1 shows an example of a fill-in-the-blanks style question found in past NCTUA problems. The number of questions falling in these categories is unexpectedly less than 10% of all the problems asked; there is little hope that simple localization of Watson-type factoids will pass university entrance examinations.


Question 9
Choose the most appropriate word for the blank in the following speech.

"You can choose either the train or the bus. If you want to save some time, the train would be better. The [ ] is a little bit higher, though."

1. cash
2. fare
3. fine
4. interest

Figure 1. Example of Fill-in-the-blanks Style Question in NCTUA ESL Test

3.2 True/False Questions

More than 75% of problems asked in NCTUA in the field of social studies, such as world history, are categorized as true/false questions. Figure 2 shows an example.

The intelligence required for human beings to answer factoid and non-factoid styles of questions seems to be the same: the amount of knowledge. However, recent research on textual entailment (TE) has been revealing in that these technologies that are effective for factoids cannot be easily applied to non-factoids (references). TE in natural language processing is a directional relation between text fragments. The relation holds whenever the truth of one text fragment follows from another text. For example, the machine has to recognize that sentence t2 follows from sentence t1.

t1: Yasunari Kawabata won the Nobel Prize in Literature for his novel "Snow Country"

t2: Yasunari Kawabata is the writer of "Snow Country"

Determining truth/falseness based on knowledge written in textbooks is well regarded as a combination of search and TE recognition tasks. It should be noted that t1 and t2 in this example have only four words in common. We cannot tell why t2 follows from t1, despite sentences like "Yasunari Kawabata criticized 'Snow Country'" and "Yasunari Kawabata won the Nobel Prize in Physics for his novel 'Snow Country'" do not only by using the current search technologies.

In Europe north of the Alps, it was not unusual for (4) farmers and those in the lower classes in cities to find it difficult to make a living through a single occupation. In particular, from the (5)latter half of the 15th century to the latter half of the 16th century, a large number of impoverished peasants and poverty-stricken lower classes emerged, due in part to the effects of population increases, and (6)quite a few of them opted to take the path of becoming a mercenary. The lump sum they were paid when joining up and their wages supported their livelihoods. The Thirty Years' War was a war in which such mercenaries played a major role.

Question 4

From 1-4 below, choose the most appropriate sentence that describes rural areas and cities in regard to the underlined portion (4).

1. In the 11th century, in England, enclosure was carried out for the purpose of sheep farming.
2. In the 12th century, the system of domain economies based on compulsory labor spread through the region east of the Elbe River.
3. Guilds were organized in order to guarantee free competition in production and distribution.
4. The Hanseatic League was led by Lübeck.

Figure 2. Example of True/False Question in NCTUA World History Test

Our research team developed a testbed for TE (Miyao et.al, 2012) that utilized the resources taken from the history problems asked in NCTUA, and organized international evaluation tasks at NTCIR-9 and 10. The results from the evaluations revealed that the best system achieved a correct answer ratio of 57%, which is significantly better than a random choice (four choices), but still far below that of human recognition. There is still a long way to go when noting that not only sentence-by-sentence but also paragraph-by-sentence TE recognition combined with searches is required to answer true/false styles of questions.

3.3 Summarization

Not only NCTUA tests but also the written tests by The University of Tokyo often ask the test takers to summarize given documents from a specific point of view. The summarization problems in Japanese and ESL account for more than 20% at NCTUA, and more than 50% in the written test.

Most state-of-the-art automatic summarization systems produce *extractive summaries* by first extracting important sentences from a (set of) document(s) using surficial clues and then re-arranging the extracted sentences with some modifications to achieve readability of output. This is thus methodologically quite different from an alternative 'deeper' approach that is based on precise understanding of the source document and natural language generation (NLG) from an abstract representation of the main content of the source document. Nencova & McKeown (2011) provided a recent survey of the field.

While targeting extractive summaries is apparently a practical choice given current progress in the generic language understanding technology and NLG, it clearly falls short of solving some exam problems that ask for the recognition or generation of summaries that are designed to test the examinee's true understanding of a text. Our current and future work includes the investigation of students' behaviors in a summarization task, and the evaluation and enhancement of current automatic summarization technologies.

4. Development of Underlying Technologies - Mathematics as Example

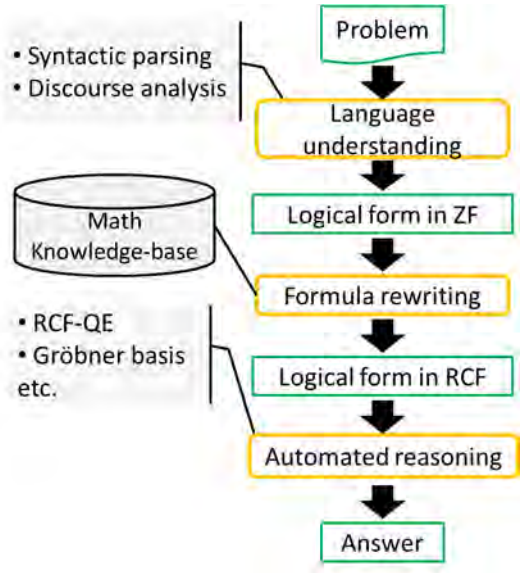


Figure 3. Overview of Math Solver

The current rise of AI has two main origins. The first is, of course, the invention of machine learning. Statistics and optimization deliver their theories to machine learning. The combination of big data and massively parallel computing has enabled machines to "learn" from data existing on the Web, networks, and databases, even though there is only little hope that machine learning technologies will help machines to solve design problems like proving the validity of a given math formula.

The second rather inconspicuous origin is the sophistication of the traditional logical approach. The so-called "knowledge acquisition bottleneck" in the research area of syntactic and semantic parsing of natural language is being relieved by the development of various techniques to extract a massive collection of syntactic/semantic rules from (annotated) text corpora (e.g., Miyao et al., 2004; Hockenmaier & Steedman, 2007; Liang et al., 2011). Similar techniques have also been developed to extract the inference rules used in high-level reasoning tasks including TE recognition (e.g., Lin & Pantel, 2001; Schoenmackers et al., 2010). The virtue of the logical approach is in its ability to express complex input-output relations, such as the mapping from natural language text to its meaning and the logical relation between a premise and its consequences, in a way that a human can understand.

This raises a natural question. How far can a machine approach human intelligence, particularly in its ability of generic problem solving by mashing up these two different types of AI technologies, statistical and logical? Kanayama and Miyao (2012) presented a technique of converting a true/false question to a set of factoid-style questions, that was aimed at overcoming the difficulty of finding direct evidence for logically rejecting a statement as false in an information source. They achieved an accuracy of 65% on the NCTUA data by employing Watson's factoid QA engine as the backend system. Tian et al. (2014) developed a semantic representation framework that allowed efficient inference while capturing various aspects of natural language semantics. They combined a logical inference engine based on their semantic representation with a statistical classifier to enhance the coverage of their methodology beyond simple TE recognition problems that could be handled purely within logical inference.

While the design of the semantic representation was in itself a key research issue in Tian et al.'s approach to solving true/false problems in social studies, we already have a well-established system for representing the meaning of mathematical propositions and problems, viz., Zermelo–Fraenkel (ZF) set theory. Thus, the main technical challenges in developing a math problem solver are 1) to accurately translate a natural language math problem into its semantic representation based on ZF set theory, and 2) to mechanically solve a problem expressed as a formula in ZF set theory (Figure 3).

The translation from a natural language text to its formal representation occurs in two steps. The first is the derivation of the meaning of each sentence through syntactic parsing using combinatory categorial grammar (CCG) (Steedman, 2001). The meaning of a sentence is composed in the process of CCG parsing by combining the formal semantic representations of the words in the sentence. The second step is the derivation of the meaning of a problem through the composition of the meanings of the sentences in the problem. We extended the discourse representation structure (DRS) (Kamp and Reyle, 1993) for the theoretical basis of the second step to express the complex inter-sentence semantic composition mechanism that we often encounter in math problems (Matsuzaki et al., 2013).

Table 1: Evaluation results in National Center Test for University Admissions mock tests.

Score	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)
Avg. of students	88.3	72.2	45.6	52.0	47.6	42.0	46.6
T-scores of systems	41.0	45.9	56.1	51.9	47.2	48.3	55.2

Table 2: Evaluation results on University of Tokyo entrance exam mock tests (Math).

Score	Math (Humanities course)	Math (Science course)
Test-takers' system	40(/80)	40(/120)
T-scores of systems	59.4	61.2

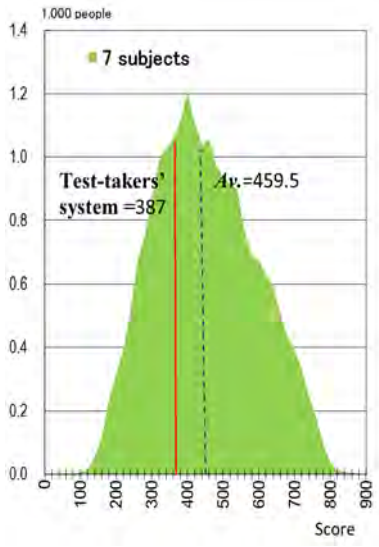


Figure 4. Distribution of total scores in 7 subjects

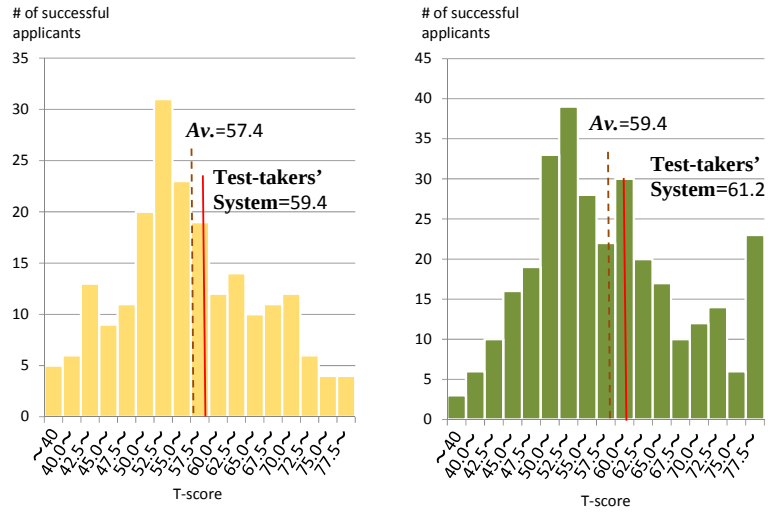


Figure 5. Distribution of T-scores on Univ. Tokyo mock math test achieved by successful applicants in Univ. of Tokyo 2013 exam (Left: Humanities course & Right: Science course)

Even though the language of ZF set theory is powerful enough to express almost all mathematical concepts, no current automatic reasoning technology is capable of making inferences directly based on ZF set theory. We thus need to first rewrite the semantic representation based on ZF set theory into a form that is more amenable to automatic reasoning (AR) such as the theory of real closed fields (RCF) and Presburger Arithmetic. We selected RCF as our first target since elementary geometry and calculus are known to be contained in RCF. However, direct and word-by-word automatic translation often results in very long and redundant formulas beyond the theoretical capacity of an AR algorithm called quantifier elimination for RCF. We succeeded in overcoming many difficulties (Matsuzaki et al., 2014; Iwane et al., 2014) by developing new algorithms. Our empirical study indicated that about half the entrance examination problems expressible in RCF could be solved automatically.

5. Evaluation

We organized an open evaluation task for the NCTUA in 2013. The Yoyogi Seminar, which is one of the major preparatory schools in Japan, provided their mock tests as the source. The fields of the questions included Mathematics (IA and IIB), Physics, World History, Japanese History, Japanese, ESL, and the datasets were provided with human-annotated document structures in XML format (Fujita et al., 2014a). The reason we used mock tests instead of the real tests was that the National Center for

University Entrance Examinations only offers the average scores of the test takers, but not their distributions.

We were able to compare our systems and average senior high school students by administering the mock tests. Table 1 summarizes the results. The average for human test takers was 459.5 out of 900 points. Our systems marked 387, and their T-score was 45.0. Figure 4 plots the status of our systems compared to the human test takers. The capability of our current systems is far below that of the entrants to any of the prestigious universities, including The University of Tokyo. Closer error analysis tells us that our systems made mistakes on problems that required deep reading, situational understanding, world knowledge, common sense, and modeling. Their shallow "literacy" was not good enough to solve these problems. However, our systems did relatively well in some subjects such as world and Japanese histories and mathematics. The Yoyogi Seminar concluded by analyzing the results that our system was capable of passing the entrance exams of 404 out of 744 private universities in Japan.

We evaluated our math problem solving system not only with a multiple choice style test but also with a written test specially designed for those who planned to take the entrance examination of The University of Tokyo. Figure 5 plots the status of our systems compared to the human test takers.

6. Human-Machine Collaboration

Much research has been done on the relation between computerization and recent trends in labor market polarization, with growing employment in high-income cognitive jobs and low-income manual occupations, accompanied by a hollowing-out of middle-income routine jobs (see, e.g., Frey and Osborne 2013 for an overview). As machines get smarter, human beings are expected to become even smarter to be able to survive in the labor market. This trend will be accelerated even more by the emergence of smarter machines. Digital abundance has already decreased the significance of simple memorization and calculation skills, while it has increased that of skills that make use of the new technologies to augment one's talents and abilities. These skills are sometimes referred as a part of information literacy (AASL/AECT, 1998) and as a part of key competency (Rychen and Salganik, 2003). Public education is expected to play the main role in assisting students to acquire these generic skills that seem to provide a comparative advantage to human beings.

We investigated the factors affecting the problem solving skills of third-year high school students, taking search and editing tasks as an example (Fujita et. al., 2014b). Search technologies are the most typical and powerful AI technologies currently available. It is simultaneously difficult for current machines to determine the truth/falseness of found documents, as was explained earlier in this paper. Students educated for 12 years in schools are expected to be intelligent enough to evaluate the appropriateness of found documents to achieve the goals of searches. We were particularly eager to know whether or not academic histories in specific areas would have positive effects on the accomplishments of tasks.

The participants in our experiment were 70 students from two public high schools. The students had to pass an entrance examination to enroll in the schools and both were considered to be in the top 7% in the prefecture. All the participants were motivated to go on to high ranking universities in Japan.

The participants were provided with an interface where they could edit answer sentences after searching a document, which was a Japanese history textbook. The participants could copy and paste statements derived in search form to answer form and edit them in answer form.

Two problems about Japanese history were asked in the experiment. Both of them had to be answered within 15 minutes. If a participant selected a certain page (correct page) from the document in Problem 1 (P1) and extracted a certain set of consecutive sentences (correct part), he/she could completely answer the problem. Problem 2 (P2) required a slightly more complex cognitive procedure. The participants had to select two particular pages, extract certain consecutive statements from each page, and summarize the statements to meet the requirements. If the participants in P2 only used the exact extracted text from the correct pages as the answer text, they would exceed the word count limitations.

The percentage of criteria participants met was about 70% in P1 (average score was 1.39 and SD = 0.60) and about 48 % in P2 (average score was 3.36 and SD = 1.55). They naturally performed more poorly when more complex cognitive processing was required. Most participants found the

correct pages in P2, but they could not meet nearly half of the criteria. This indicated that participants had trouble unifying and editing information.

Out of the 70 participants, 43 took a class in Japanese history in their second year of high school, while 37 did not. We unexpectedly found that academic history had no significant effects on the number of correct pages selected or the test scores for either problem (Figure 6). These findings suggest that learning in a traditional setting (i.e., the combination of lectures, textbooks, and standardized tests) did not contribute to developing abilities to evaluate the appropriateness of found documents and summarize them properly to meet the requirements.

Further cognitive and pedagogic analysis is necessary to conclude whether or not they would be able to perform better if they had learned in a different setting, such as active learning and computer supported collaborative learning. We left this for future work.

7. Conclusion

We introduced the aims and the state of progress in the "Todai Robot Project". The purpose of the project was to open up new horizons by reintegrating the subfields of AI that have come about since the 1980s, and encourage researchers in these areas to share progress and findings over the past thirty years. We simultaneously tried to elucidate the abilities of the current AI in comparing it to human students using university entrance examinations as a testbed. It should help us to understand the possibilities of and limitations with near-term AI technologies.

We have thus far developed AI systems that were capable of passing the entrance examinations of more than half the universities in Japan. The progress with our math problem solving system especially suggested that human beings cannot maintain their comparative advantage over machines in problem solving in the most important areas of geometry and calculus. We have to wait for future research to conclude whether or not this will be the same case for TE recognition and information summarization.

We analyzed the performance and the effects of the academic histories of high school students in solving Japanese history problems with a search-and-edit interface. Unfortunately, the results revealed that their academic histories had no significant effects on performance. This suggests that it is necessary to re-design the education system by taking into consideration pressures from the labor market.

Acknowledgements

We would like to express our gratitude to the National Center for University Entrance Examinations, the Yoyogi Seminar, Tokyo Shoseki Co., Ltd and, Yamakawa Shuppansha Ltd. for their support. The Todai Robot Project was financially supported by the National Institute of Informatics.

References

Anderson, R. E. (1992). Social impacts of computing: Codes of professional ethics. *Social Science Computing Review*, *10*(2), 453–469.

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

Arai, K. (2013). The Economics of Education: An Analysis of College-Going Behavior, Springer-Verlag.

Turing, A. M. (1950). Computing Machinery and Intelligence. *Mind*, *49*, pp. 433–460.

Crane, T. (2003). The mechanical mind: A philosophical introduction to minds, machines and mental representation. Presbyterian Publishing Corp.

Bresnahan, T.F. (1999). Computerisation and wage dispersion: an analytical reinterpretation. *The Economic Journal*, vol. 109, no. 456, pp. 390–415.

Brynjolfsson, E. and McAfee, A. (2011). Race against the machine: How the revolution is accelerating innovation, driving productivity, and irreversibly transforming employment and the economy. Digital Frontier Press Lexington, MA.

Frey, C. B. and Osborne, M. A. (2013). The future of employment: how susceptible are jobs to computerisation? Oxford University.

Autor, D., Levy, F., and Murnane, R.J. (2003). The skill content of recent technological change: An empirical exploration. The Quarterly Journal of Economics, vol. 118, no. 4, pp. 1279–1333.

Ravichandran, D. & Hovy, E. (2002, July). Learning surface text patterns for a question answering system. In *Proceedings of the 40th Annual Meeting on Association for Computational Linguistics* (pp. 41–47). Association for Computational Linguistics.

Ferrucci, D., Brown, E., Chu-Carroll, J., Fan, J., Gondek, D., Kalyanpur, A. A., Lally, A., Murdock, J. W., Nyberg, E., Prager, J., Schlaefer, N., and Welty, C. (2010). Building Watson: An overview of the DeepQA project, *AI magazine*, *31*(3), pp. 59–79.

Nenkova, A., & McKeown, K. (2011). Automatic Summarization. Foundations and Trends in Information Retrieval, 5, (2–3), 103–233.

Miyao, Y., Ninomiya, T., & Tsujii, J. (2004). Corpus-oriented grammar development for acquiring a head-driven phrase structure grammar from the Penn treebank. In Proceedings of the First International Joint Conference on Natural Language Processing, pp. 684–693.

Hockenmaier, J. & Steedman, M. (2007). CCGbank: A Corpus of CCG Derivations and Dependency Structures Extracted from the Penn Treebank. Computational Linguistics, 33, (3), pp. 355–396.

Liang, P., Jordan, M. I., & Dan Klein, D. (2011). Learning dependency-based compositional semantics. In Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies, pp. 590–599.

Lin, D. & Pantel, P. (2001). Discovery of inference rules for question answering. Natural Language Engineering, 7, (4), pp. 343–360.

Schoenmackers, S., Etzioni, O., Weld, D. S., & Davis, J. (2010). Learning First-order Horn Clauses from Web Text. In Proceedings of the 2010 Conference on Empirical Methods in Natural Language Processing, pp. 1088–1098.

Tian, R., & Miyao, Y., & Matsuzaki, T. Logical Inference on Dependency-based Compositional Semantics. In Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics, pp. 79–89.

Kanayama, H., & Miyao, Y., & Prager, J. Answering Yes/No Questions via Question Inversion. In Proceedings of COLING 2012, pp. 1377–1392.

Matsuzaki, T., Iwane, H., Anai, H., & Arai, N. (2013). The Complexity of Math Problems – Linguistic, or Computational? In Proceedings of the 6th International Joint Conference on Natural Language Processing, pp. 73–81.

Matsuzaki, T., Iwane, H., Anai, H., & Arai, N. H. (2014). The Most Uncreative Examinee: A First Step toward Wide Coverage Natural Language Math Problem Solving. In Proceedings of 28th Conference on Artificial Intellegence, pp. 1098–1104.

Iwane, H., Matsuzaki, T., Arai, N., & Anai, H. (2014). Automated Natural Language Geometry Math Problem Solving by Real Quantier Elimination. In Proceedings of the 10th International Workshop on Automated Deduction in Geometry, pp. 75–84.

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

Fujita, A., Kameda, A., Kawazoe, A., & Miyao, Y. (2014) Overview of Todai Robot Project and Evaluation Framework of its NLP-based Problem Solving. In Proceedings of the 9th edition of the Language Resources and Evaluation Conference, pp. 2590-2597.

Ryche Rychen, D. S. E., & Salganik, L. H. E. (2003). *Key competencies for a successful life and a well-functioning society*. Hogrefe & Huber Publishers.

Fujita, A., Suzuki, M., & Arai, N. H. (2014) Cognitive Model of Generic Skill: Cognitive Processes in Search and Editing, In Proceedings of the 36th annual meeting of the Cognitive Science Society, pp. 2234-2239.