The impact of AI—can a robot get into the University of Tokyo?

Noriko H. Arai

The ‘Todai Robot Project (Can a robot get into the University of Tokyo?)’ was initiated by the National Institute of Informatics (NII), the Japan’s only general academic research institution of informatics, in 2011 as an AI grand challenge. The goal of the project is to create an AI system that answers real questions on university entrance examinations. This paper reports the current status of the project including the underlying technologies we have developed thus far, and the results we obtained from evaluations.

There have been various types of AI challenges in the past: the chess and shogi matches against professional players, the quiz show challenge against human champions. Todai Robot Project is unique mainly in the following two senses. First, it targets the integrated interdisciplinary research synthesis. Second, it enables us to compare the current status of the project including the underlying technologies we have developed thus far, and the results we obtained from evaluations.

The tasks require not only the development of ground-breaking underlying technologies in various AI research areas but also their interdisciplinary research synthesis. Second, it enables us to compare the performance of the software and numerous well-educated students. 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 National Center Test for University Admissions (NCTUA), standardized multiple choice style tests, every year in Japan, and less than top 3% students are allowed to take the second written test specially designed to select entrants of the University of Tokyo.

Our research team developed a testbed [1, 2] that utilized the resources taken from the history problems asked in NCTUA, and organized international evaluation tasks at NTCIR-9, 10 and 11. The participants pursued mainly two approaches to solving history questions. One is traditional statistical factoid approach. Kanayama and Miyao [3] manually converted original true/false world history questions to a set of factoid-style questions and achieved an accuracy of 65% on the NCTUA data by employing Watson’s factoid QA engine as the backend system. Kano [4] developed domain-independent and language-independent keyword-based system, and achieved an accuracy of 51% on the same data. The other approach is to combine a logical inference engine based on their semantic representation with a statistical classifier. Considering that the accuracy rate of Watson at Jeopardy! Challenge was 69%, it was most likely some deeper language analysis would be required to achieve the accuracy rate of 80%. Tian et al. [5] developed a semantic representation framework that allowed efficient inference while capturing various aspects of natural language semantics. Okita and Liu [6] built a QA system in which they used a semantic parser to acquire meaning representation from history textbooks using common-sense knowledge, and they achieved an accuracy of 68% on the NCTUA data.

Higashinaka et al. [7] applied dialogue and machine translation technologies to English exams; taking into account both user intentions and sentiment analysis results yielded significant gains in the dialogue completion task and using language models constructed from large text corpora led to an accuracy of around 80% in the sentence completion task.

It has been a long dream to solve mathematical problems (expressed in natural language and images) by machine since Turing’s age. Though the possibility of complete automatization of mathematics was denied by the undecidable results, a very important theory of mathematics was proved to be decidable: the theory of real closed field (RCF). Furthermore, RCF admits the quantifier elimination. It suggests that the major part of Euclidean geometry, algebra and calculus especially that is asked in the entrance exams is automatizable. We first targeted the problems falling in this category, and developed a prototype system to solve them [8–10].
Table 1. Evaluation results in NCTUA mock tests.

| Score   | English | Japanese | World history | Japanese history | Math IA | Math IIB | Physics |
|---------|---------|----------|---------------|------------------|---------|----------|---------|
| Task-takers’ system | 95 (/200) | 49 (/100) | 52 (/100) | 44 (/100) | 40 (/100) | 55 (/100) | 31 (/100) |
| Students’ ave.         | 93.1     | 45.9     | 40.8       | 47.2       | 47.1     | 50.4     | 32.7     |
| T-scores              | 50.5     | 51.9     | 56.1       | 48.2       | 46.9     | 51.9     | 49.0     |

The system receives an annotated problem text as the input. The sentences in the problems are then translated to logical forms through grammar-based parsing, using the annotations on the text as the constraints in the parsing process. The sentence-level logical forms are then combined into a problem-level semantic representation according to the intersentence logical relations annotated on the problem. The output is expressed in the most powerful language in mathematics, the higher order formula in Zermelo–Fraenkel’s (ZF) set theory, which is, of course, undecidable. The obtained ZF formula is iteratively rewritten by applying several kinds of equivalence-preserving transformation rules until we find that the rewritten formula is directly reinterpretable in RCF. When it succeeds, we invoke the quantifier elimination algorithm to find the answer. Fig. 1 illustrates the process. Though the linguistic part still requires the human error correction, it solved 23 problems out of 47 taken from the past entrance exams of seven major national universities including the University of Tokyo. We plan to apply our framework to non-RCF problems as well as to physics.

Our research team took a (digitalized and annotated version of) mock National Center Test in 2013 and 2014 provided by a preparatory 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 beyond the median of the human test takers. It was competent enough to pass the entrance exams of 472 out of 581 private universities in Japan (Table 1).

Our results suggest that it is getting harder and harder for the students to get economic returns from investing higher education since AI is ‘smarter’ than the average students especially in building and manipulating given knowledge.

Noriko H. Arai
National Institute of Informatics, Japan
E-mail: arai@nii.ac.jp

REFERENCES
1. Miyao, Y., Shima, H. and Kanayama, H. et al. ACM Transactions on Asian Language Information Processing (TALIP), 2012, Vol. 11: 13.
2. Shibuki, H., Sakamoto, K. and Kano, Y. et al. Proceedings of NTCIR-11. Tokyo: NTCIR 2014, 518–29.
3. Kanayama, H., Miyao, Y. and Prager, J. Proceedings of COLING 2012. Mumbai: The COLING 2012 Organizing Committee, 2012, 1377–92.
4. Kano, Y. Proceedings of NTCIR-11. Tokyo: NTCIR 2014, 530–31.
5. Tian, R., Miyao, Y. and Matsuzaki, T. Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics. Baltimore, Maryland: Association for Computational Linguistics 2014, 79–89.
6. Okita, T. and Liu, Q. Proceedings of NTCIR-11. Tokyo: NTCIR 2014, 570–77.
7. Higashinaka, R., Sugiyama, H. and Isozaki, H. et al. Proceedings of the 21st Annual Meeting of the Association for Natural Language Processing (NLP2015) 2015, (to appear; in Japanese).
8. Matsuzaki, T., Iwane, H. and Anai, H. et al. Proceedings of the 28th Conference on Artificial Intelligence. Palo Alto, California: The AAAI Press 2014, 1098–104.
9. Arai, NH, Matsuzaki, T. and Iwane, H. et al. Proceedings of the 39th International Symposium on Symbolic and Algebraic Computation. New York, New York: ACM 2014, 1–8.
10. Iwane, H., Matsuzaki, T. and Arai, N. et al. Proceedings of the 10th International Workshop on Automated Deduction in Geometry. 2014, 75–84.