



⑧ なぜ高校生は「意味を理解しない AI」に敗れたか？

—リーディングスキルテストの結果から—

新井紀子（国立情報学研究所） 尾崎幸謙（筑波大学）

なぜ「大学入試」をベンチマークに選んだのか

「ロボットは東大に入れるか」には、大学入試問題をベンチマークとすることで、近未来 AI の知的複合タスクに対する可能性と限界を明らかにすることのほかに、もう1つのミッションがあった。それは、近未来 AI が労働市場へ及ぼす影響を明らかにすることである。

テクノロジーの進歩は、生活を便利にし、生産効率を上げ、長期的に見れば消費者だけでなく労働者にもメリットをもたらす。ただし、それには労働市場の需給調整がうまく行けば、という大前提がある。デジタイゼーションに伴う新しい人材需要に応えられる人材を、教育を通じて供給できなければ、深刻な人手不足と失業が、同時に起こり得る。

20 世紀前半の製造業におけるオートメーションは短期的には大量の失業者を生み出し、保護主義の台頭や第二次世界大戦の遠因になった。一方で、長期的には、機械にはできない高度に知的な仕事が創出され、その担い手を育成するために高度な教育を整備する必要が生じるプラスの側面もあった。つまり、初等中等教育の義務化や高等学校・大学の高い進学率は、「学ぶ・学び続ける」ということが「普遍的に良いこと」だからではなく、むしろ、オートメーション以降の社会が求めたスキルを身につけたホワイトカラー

を効率よく労働市場に送り出すために社会が最適化した結果だといえる。では、そもそも「20 世紀の工場機械と差別化できる能力」としてホワイトカラーに求められてきた標準的な技能や能力とは何か。それは入試問題に端的に現れるに違いない。では、そのうちのどれほどが近未来 AI によって代替され得るか。AI の社会的影響を予測する上で、大学入試問題をベンチマークとして選択した最大の理由はここにある。

本特集のほかの記事から分かるように、東口ボで開発されたプログラムは、Siri や Watson 等の AI プログラム同様に、問題に書かれている文や図を理解しない。理解はしないが、センター模試の偏差値は2年続けて57を超え(表-1)、東大入試の模擬試験の論述(世界史)でも人間の受験生より高い得点をマークした。一体なぜ受験生の75%は東口ボに劣る成績しか残せなかったのか。それを解明するために、リーディングスキルテスト¹⁾を用い、調査を実施した。

リーディングスキルテストとは

リーディングスキルテスト(以下 RST)では、国語の長文読解とは異なり、教科書や新聞から採った50~200字程度の短文を正しく読み解けるかを問う。現状では、問題は択一または複数選択式で、問題ごとの制限時間は設けていない。受験者は与えられた時間の中で、

【成績概要：2016 年度進研模試 総合学力マーク模試・6 月】

	国語 (200)	数学 I A (100)	数学 II B (100)	英語(筆) (200)	英語(リ) (50)	物理 (100)	日本史 B (100)	世界史 B (100)	5 教科合計 (950)
全国平均点	96.8	54.4	46.5	92.9	26.3	45.8	47.3	44.8	437.8
東口ボ得点	96	70	59	95	14	62	52	77	525
東口ボ偏差値	49.7	57.8	55.5	50.5	36.2	59.0	52.9	66.3	57.1

〈注〉5 教科 8 科目文系型(国、数 2 科目、英〈筆記およびリスニング〉、地理歴史 2 科目、理科 1 科目)での受験者 12 万 582 人(受験者総数 26 万 4,604 人)で集計した偏差値。科目名の下のカッコ内は各教科の配点。英語の(筆)は筆記、(リ)はリスニング。

表-1 2016 年度センター模試における東口ボの成績

以下の文を読みなさい。

仏教は東南アジア、東アジアに、キリスト教はヨーロッパ、南北アメリカ、オセアニアに、イスラム教は北アフリカ、西アジア、中央アジア、東南アジアにおもに広がっている。

この文脈において、以下の文中の空欄にあてはまる最も適当なものを選択肢のうちから1つ選びなさい。

オセアニアに広がっているのは () である。

- A. ヒन्दウー教 B. キリスト教 C. イスラム教
D. 仏教

図-1 係り受け認識の問題例 (出典:東京書籍(株)中学校社会教科書『新しい社会地理』p.36)

正確に、ただしできるだけ多く解くよう指示される。問題には、文の表層的な情報を読み取れているかを測る問題タイプと、文の意味を理解し、正しく推論を実行できるかを測る問題タイプがある。それぞれ(1) 係り受け認識(2) 照応認識(3) 同義文判定の3タイプと、(4) 推論(5) イメージ同定(6) 具体例認識の3つのタイプである。

図-1は(1)にタイプ分けされる、文の係り受け関係を正しく認識しているかを問う問題例であり(正解はB)、現状の構文解析器でも正解できる。一方、図-2は(5)に分類される、文を表象する正しい図を選ぶ問題例であり(正解はA)、よほどフレームを限定しないと機械に解くことは困難だろう。紙幅の制限のため、本稿ではすべての問題タイプを解説できないので、論文等の情報^{1),2)}を参照していただきたいが、大雑把にまとめると、(1)から(3)はすでに人工知能研究のリアルな対象となっているものを、(4)から(6)は不連続かつ劇的なイノベーションがない限りAIでは解決が難しいと考えられる問題を設定した。言い換えると、どのタイプの問題でも正答率が85%を超えるようならば、(人間らしく)意味を理解して文を読んでいるが、そうでない場合は、読解になんらかの課題を抱えており、意味を理解しないまま文を記号列として処理している疑いがあると考えたのである。

まず、本調査を短文の読解力調査として用いることの妥当性を担保するために、項目特性図を用いて結果が不自然な問題を取り除いた。その上で、公立中学生4,586人および進学率がほぼ100%である公立の進学校の高校生1,347人、合計5,933人についての調査結果を報告する。

下記の文の内容を表す図として適当なものを、A~Dのうちからすべて選びなさい。

原点Oと点(1,1)を通る円がx軸と接している。

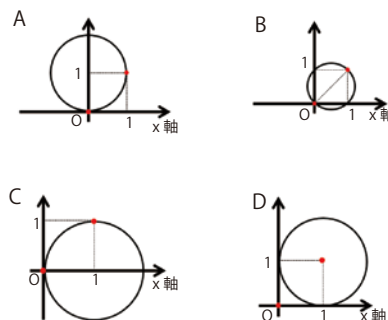


図-2 イメージ同定の問題例

調査結果と分析

本調査はすべての問題が択一または複数選択式なので、ランダムに答えを選んでもある確率で正解し得る。各問題タイプについて、中学1年生から高校3年生までの正答率、および正答率がランダム解答より良いとはいえない生徒(以下、ランダム解答)の割合を統計的に求めたものが表-2である。

すべての問題タイプについて、高い学年ほど正答率も高く、ランダム解答の生徒の割合が低いことが分かる。このことは、中高生は学校および生活の中で各タイプの読解能力を伸ばしていることとともに、RSTのテストとしての妥当性を示している。ただし、中学3年生から高校1年生の差は、各生徒の発達によるものというよりは、読解に課題が多い生徒が高校入試によってスクリーニングされた結果と解釈すべきだろう(より偏差値の低い高校の結果については近日中に公表予定)。

しかしながら、AIが射程に入れている(1)から(3)の問題タイプに比べて、意味理解なしに解くことが難しい(4)から(6)の問題タイプの正答率の方が悪く、ランダム解答の生徒の割合が高いことが分かる。まさに、先に述べたように、AIに代替されやすいような読解力の身につけ方をしていると考えられる。

表-3は、図-1(係り受け認識)と図-2(イメージ同定)で示した問題について、学年ごとに正答率を示したものである。図-1の問題について高校3年生では94%という高い正答率を示しているものの、図-2の問題については45%にとどまっている。図-2の問

8 なぜ高校生は「意味を理解しない AI」に敗れたか？

	係り受け	照応認識	同義文判定	推論	イメージ	具体例
中1	56.4 (44.6)	55.4 (46.5)	67.3 (58.1)	50.6 (68.4)	26.4 (57.7)	23.6 (62.1)
中2	58.4 (40.5)	55.9 (42.6)	71.8 (52.5)	54.0 (58.9)	28.6 (45.4)	25.4 (56.6)
中3	64.5 (29.3)	66.8 (24.4)	74.7 (41.8)	57.9 (49.9)	36.3 (34.5)	33.7 (41.4)
高1	82.0 (6.7)	82.1 (5.8)	87.5 (17.0)	67.9 (32.9)	51.1 (16.8)	49.1 (19.3)
高2	86.4 (3.5)	81.3 (10.4)	88.7 (12.5)	69.7 (29.5)	57.7 (10.4)	50.0 (8.5)
高3	87.5 (3.0)	86.9 (2.2)	90.5 (8.5)	74.9 (19.1)	53.2 (4.4)	51.3 (14.3)

表-2 各問題タイプの正答率（括弧外）およびランダム解答の生徒の割合（括弧内）（単位は%）

	中1	中2	中3	高1	高2	高3
図-1：係り受け	58.2	48.8	64.6	71.8	84.2	94.1
図-2：イメージ	10.7	22.2	25.4	29.0	30.0	45.5

表-3 公立中学高等学校における図-1と図-2の問題の正答率（単位は%）

題に登場する数学用語は中学2年生の全員が学んでおり、しかも二次方程式の解を求めるといような数学的操作は一切求めていない。それにもかかわらず、進学率がほぼ100%の高校3年生であっても、図-2の問題文を読んで理解することができないのだ。

浮かび上がった新たな研究課題

RSTの問題は、どれも（常識を超えるような）知識やスキルは問うてはいない。これらの問題が解けないのであれば、当然のことながら入試問題の提示文や問題文は読めないだろう。では、問題文を読むことができずにどうやって高校生は入試を突破しているのだろうか。あるいは全国学力調査や国際学力調査（PISA）でそれなりの成績をおさめてきたのだろうか。意味を理解できないのに、それなりの精度を達成するAIを研究開発してきた私たちは、その理由をよく知っているはずだ。データ（ドリル）に基づき、単語の共起などのヒューリスティックを利用したりすることで解いている可能性が高い。であれば、データ量と計算量ではるかに彼らを上回る東ロボに敗れるのは当然だったといえよう。

このような人材では、ウィキペディアやMOOC等のオンライン教材を通じて新しい知識を身につけることが期待できないだけでなく、安全管理マニュアルや仕様書、契約書等を正しく読み取れるかどうかも危ぶまれる。

いったいなぜこのような現象が起こったのだろう。過去の結果がないので比較はできないが、この10年で子どもたちの置かれている「読み」に関する環境が激変していることは認識しておく必要がある。幼児や小学生の生活環境から新聞やカレンダーなど字が書かれ

ている紙が激減しているのである。それらはデジタル化され、保護者のパソコンやスマートフォンの中に鍵をかけて仕舞われており、結果として幼い子どもたちが環境としての文字に接する機会は減ってしまった。情報学を担う私たちが、結果として「環境としての文字」を彼らから奪った可能性は否定できない。

一方で、情報学を担う私たちは、この状況を改善するヒントを提供できる可能性もある。今回開発したリーディングスキルテストは言うまでもなく、AIに何ができ、また何ができないかを分析した結果生み出された成果にはほかならない。「意味を理解しないAI」のエラーに日常的に接している私たちだからこそ提案できる教育法もあるかもしれない。その提案が自己満足に陥らないためにも、「半数以上の中学生が、中学校教科書を読めずに卒業している」という状況を直視し、エビデンスに基づく教育システムの再構築が求められる²⁾。

参考文献

- 1) Arai, N. H., Todo, N., Arai, T., Bunji, K., Sugawara, S., Inuzuka, M., Matsuzaki, T. and Ozaki, K.: Reading Skill Test to Diagnose Basic Language Skills in Comparison to Machines, to appear in CogSci 2017.
- 2) 国立情報学研究所：国立情報学研究所プレスリリース、国立情報学研究所（オンライン）、入手先 <<http://www.nii.ac.jp/news/2016/0726/>>（参照 2016-07-26）。
- 3) Lord, F. M. and Novick, M. R.: Statistical Theories of Mental Test Scores, Addison-Wesley (1968).
- 4) van der Linden, W. J. and Glas, C. A. W. (Eds.): Elements of Adaptive Testing, Springer (2010).（2017年3月24日受付）

■新井紀子 arai@nii.ac.jp

イリノイ大学大学院博士課程修了、博士（理学）。広島市立大学情報科学部助手、フィールズ研究所客員研究員等を経て、現在、国立情報学研究所情報社会相関研究系教授、社会共有知研究センター長。専門は数理論理学、人工知能。

■尾崎幸謙 koken@gssm.otsuka.tsukuba.ac.jp

早稲田大学大学院文学研究科心理学専攻博士後期課程修了、博士（文学）、日本学術振興会特別研究員、統計数理研究所等を経て、現在、筑波大学大学院ビジネス科学研究科准教授。専門は心理統計学、教育測定学。