人間と生成ＡＩの文章理解のちがい

一年四組　高橋　尊斗

現在の生成ＡＩは「記号接地」を全くせずに言語を学んでいるという。[1]人間は知っている言葉が指す対象を知っており、自分自身の体から、概念の様々な特徴を捉え知ることができる（記号接地をして言葉を理解できる）。それに対して生成ＡＩは、身体を持たず、接地した経験がないため、感覚に接地していないまま（真の意味を理解していないまま）記号と記号を結びつけ、置き換えて言葉を理解しているかのように見せている。生成ＡＩを活用するためには、その特性を知らなければならない。人間と生成ＡＩとの文章理解の方法のちがいをふまえ、両者が共存する未来について考えてみたい。

二〇一八年、Ｇｏｏｇｌｅ社によってＢＥＲＴという文章の並び方に確率を割り当てる確率モデルである「言語モデル」が発表された。その言語モデルのうち「計算量」「データ量」「モデルパラメータ数」の三要素を大規模化したものを「大規模言語モデル」（以下、ＬＬＭ）と言い、代表的な例としＯｐｅｎＡＩ社のＧＰＴを挙げることができる。[2]つまり、現在のチャットボット等の生成ＡＩは、与えられたタスクを達成するために確率の高い言葉を次々に出力しているだけであり、真にその言語を「理解」しているわけではない。

では、人間が文章を理解する際、どのような構造に注目しているのだろうか。人間は文章に一貫性を求め、他の知識を照合したり、一度構成したものから新たな問や矛盾を発見し解消したりすることで、「理解」を深めていくことができる。特に一貫性は情報間に矛盾があることに気づき、言葉や世界に関する知識（常識）である既有知識と照合することで、不自然さを吟味し解消する。[3]

「不自然さを解消する」ことは「矛盾の解消」と呼ばれる。[4]情報の整合性がいったん脅かされても適切な事例を説明としてつけ加えることによって、それらの情報を整合性あるものとして理解し直すことは、小学生でも可能であり、幼児期においてもすでに使用していることが示唆されている。[5]

この事例について、大学生を被験者として、与えられた情報の矛盾がどのように修正されるかという実験結果がある。[6]大学生と一対一の面接を行い、例文を一度読み聞かせ、話の中につじつまの合わない所があったか矛盾点を指摘させる。大学生は矛盾を指摘した直後にその情報解釈について「本当にそれで良いか」と念を押す。その後（イ）確かにそう言える

と思うか、（ロ）そう言えるとは限らないと思うが説明できないか、（ハ）そう言えるとは限らない事を説明できる（この場合に限り大学生に説明を求める）かのいずれかひとつを選択させる。最後に例文が全部本当のことだとしたらどう理解すればよいか説明させ、矛盾を解消させる。この結果、人間は最初に与えられる文の表現が変わると、矛盾の解消の際にその文の解釈を変えることが傾向として示された。

この実験と同様のものを、生成ＡＩを使って検証してみる。使用するＬＬＭの種類はＯｐｅｎＡＩ社のＧＰＴ―３・５、ＧＰＴ―４、Ｇｏｏｇｌｅ社のＰａＬＭ２、Ｍｅｔａ社のＬｌａｍａ２の四種類である。なお、ＰａＬＭ２は日本語での出力に応答しなかったため、言語は英語を使用した。具体的な方法として、前項で取り上げた論文に掲載された例文を入力し、人間に対して行ったのと同じように質問を重ねて出力させた。

結果、人間でみられていた各例文の規則性が生成ＡＩでは見られなかった。よってＬＬＭに対して与える文を変えても、矛盾の解消の際にその文の解釈を変えることは難しそうである。これは、冒頭で述べたとおり、生成ＡＩの文章理解の方法が人間と異なるからであると考えられる。また、ＬＬＭの性能が上がるほど矛盾の解消の構成数は大きく増加し、より生成ＡＩ開発の進んでいる英語で行うと構成数は増加する傾向であることも分かった。一方、ヒントを与えることで言及カテゴリー数、回答数ともに増加する傾向があることが分かった。暗黙裏に想定されている条件を意識化させることで、その条件が本当に満たされているのかを疑わせ、その結果、矛盾の解消促進に繋がる傾向があるといえる。

注目すべき点として、生成ＡＩはＬＬＭの種類によって出力結果が変わっていたことがある。回答のあと、本当に妥当なのかと聞くと、最初は矛盾していると答えていたのに、「指摘が間違っていた」や「矛盾している点は実際にはなかった」等、間違った主張に変えてしまう出力パターンが多く確認された。最初は矛盾していないと回答していたＬＬＭが、あとになって矛盾していたと訂正するのである。

以上より、生成ＡＩは本質的な「理解」をしているかのように見せる状態にはほど遠いことがわかる。しかし、一部の側面では人間と比べて、文章理解の能力が同等もしくはそれ以上と見なしうる場合もあった。年々、人間とは異なった仕方で「理解」する、しているように見せる能力が向上しているため、数年後、ＡＩの「理解」が、人間のその能力を上回るようになるかもしれない。

なお、現在開発途中ではあるが、「なぜその答えを出したのか」を説明できるＡＩは、「説

明可能なＡＩ（ExplainableAI）」と呼ばれている。説明可能なＡＩという概念には、ＡＩの出す結論が信頼に足るものなのか最終的に判断し、責任を持つのは人間であるという前提がある。しかし、ＡＩの発展によってＡＩ自身が「なぜその答えを出したのか」を説明し、ＡＩモデルの説明可能性を自身が証明するようになるかもしれない。それが実現可能になれば、人間が介さずとも自身を説明し、自己認識をしているかのように見せることができるということだ。

今後もＡＩは加速度的に発展し続けるだろう。電話交換手のように、新しい道具、技術の発展によって消失する職種は多数存在する。自身が「理解」する能力を超えられた人間は、現時点でも仕事を奪われることは確実だ。現在のＡＩは人間による入力が行われないと動かず出力しないが、自分で自我を持つかのような説明可能なＡＩ、汎用ＡＩが生まれた後には、真のシンギュラリティが待っているかもしれない。

1　今井むつみ・秋田喜美『言語の本質』（二〇二三年・中央公論新書）

2　ＮＲＩ「大規模言語モデル」https://www.nri.com/jp/knowledge/glossary/lst/ta/llm（二〇二三年七月二十二日閲覧）

3　内田伸子「Ⅱ．文章理解と知識」（佐伯胖『認知心理学講座第３巻推論と理解』一九八二年・東京大学出版会）

4・6　鈴木孝子「一見矛盾する情報の理解過程における事例構成」（『教育心理学研究　33巻2号』一九八五年）

5　久保ゆかり「幼児における矛盾する出来事のエピソードの構成による理解」（『教育心理学研究　30巻3号』（一九八二年）