川添　愛（かわぞえ・あい）
九州大学文学部、同大学院ほかにて言語学を専攻し、博士号を取得。津田塾大学女性研究者支援センター特任准教授、国立情報学研究所社会共有知研究センター特任准教授などを経て、言語学や情報科学をテーマに著作活動を行っている。
著書に『聖者のかけら』『数の女王』『働きたくないイタチと言葉がわかるロボット』『白と黒のとびら』などがある。

40字×98行＝3920字

ヒトの言葉　機械の言葉
「人工知能と話す」以前の言語学

川添　愛

2020年11月10日　初版発行

◇◇◇

発行者　青柳昌行
発　行　株式会社KADOKAWA
〒102-8177　東京都千代田区富士見2-13-3
電話　0570-002-301（ナビダイヤル）
装丁者　緒方修一（ラーフイン・ワークショップ）
ロゴデザイン　good design company
オビデザイン　Zapp!　白金正之
印刷所　株式会社暁印刷
製本所　株式会社ビルディング・ブックセンター

角川新書

　ISBN978-4-04-082348-5 C0280

参考文献（注に挙げたもの以外）

柏端達也『コミュニケーションの哲学入門』、慶應義塾大学三田哲学会叢書、2016年
久木田水生、神崎宣次、佐々木拓『ロボットからの倫理学入門』、名古屋大学出版会、2017年
酒井邦嘉『言語の脳科学』、中公新書、2002年
坪井祐太、海野裕也、鈴木潤『深層学習による自然言語処理』、講談社、2017年
中澤敏明「機械翻訳の新しいパラダイム　ニューラル機械翻訳の原理」、『情報管理』2017. 8, Vol. 60, No.5, 299-306.
西山佑司『日本語名詞句の意味論と語用論　―指示的名詞句と非指示的名詞句―』、ひつじ書房、2003年
C. M. ビショップ（著）、元田浩、栗田多喜夫、樋口知之、松本裕治、村田昇（監訳）『パターン認識と機械学習　上』、丸善出版、2012年
福井直樹『新・自然科学としての言語学　生成文法とは何か』、ちくま学芸文庫、2012年
三木那由他『話し手の意味の心理性と公共性　コミュニケーションの哲学へ』、勁草書房、2019年
山本貴光、吉川浩満『脳がわかれば心がわかるか　脳科学リテラシー養成講座』、太田出版、2016年

近年、ＡＩによる意外な間違いの事例がいくつも報告されています。それらの中には、私たち人間から見れば非常に理解しがたい事例が数多く含まれています。読者の皆さんの中にも、道路標識にステッカーを貼る程度の変化を加えるだけで自動運転車が標識を正しく認識しなくなるとか[18]、鹿の画像の中の一点を他の色に置き換えるだけでＡＩがそれを「飛行機」だと誤認識するようになる[19]といった話を聞いたことのある方がいらっしゃるでしょう。つい最近では、画像に写っている物体を高い精度で分類できるＡＩが、必ずしも物体そのものではなく「背景」に依存した分類をしていることが指摘されました[20]。また、先に紹介した言語モデルＢＥＲＴを利用した高性能なＡＩについても、入力文の一部を同義語（意味が同じ別の表現）に置き換えたところ、間違った動作をするようになったという事例が報告されています[21]。

こういった例は、高い性能を誇るＡＩが、必ずしも私たちが想定しているような判断の仕方をしているとは限らないということを示唆しています。この章で見てきたように、機械学習においては、機械は限られた数のデータをお手本にし、その中にある規則性や法則性を見いだそうとします。そうすることで、お手本の中にない入力に対してもできるだけ正解を出せるような関数を求めるわけです。ただし、このときに機械がデータの中のどの

ような側面に目をつけているかは明確ではなく、私たちが認識していないような意外な法則性を利用している可能性があります。もちろん、それがＡＩに求められている仕事にとって本質的な法則性であれば何の問題もありません。しかし、もしそうでない場合は、たとえＡＩがうまく動いているように見えたとしても、それは「目のつけどころは間違っているが、その間違いがまだ露呈していないだけ」ということになります。

このような意味で、機械学習を利用して開発されるＡＩは、どのように動作するかを完全に予測することができません。つまり、ありとあらゆる入力に対して、つねに１００％正しく動くと保証することができないのです。このことは、ＡＩ全般の品質保証の問題につながります。たとえ商品として出荷する前のテストでは十分な精度が出ていたとしても、お客さんの手元でお客さんが与える入力に対して同じような精度が出るとは限りません。機械学習で開発されたＡＩが社会一般に浸透するには、このような事実が広く認識される必要があるでしょう[22]。

ＡＩにさせたい仕事を定義できるか

ＡＩについてよく尋ねられる質問に、「ＡＩには○○ができますか？」というものがあ

ります。「ＡＩは言葉を理解できますか？」「ＡＩは感情を持つことができますか？」「ＡＩには人間のような思考ができますか？」「ＡＩに哲学はできますか？」……などなど、挙げていけばきりがありません。しかし、こういったことについて考える前に、まずはっきりさせておかなくてはならないことがあります。それは、「その○○は、どんな仕事として定義できるのか？」ということです。

すでに見てきたように、ＡＩの中身は「数（の並び）を入力したら、数（の並び）を出力する関数」です。よって、ＡＩを開発するときには、先に「何を入力として、何を出力するか」、また「入力と出力をどんな数の並びとして表すか」を決めなくてはなりません。つまり、「ＡＩにさせる課題（タスク）を定義しなくてはならない」ということです。深層学習はとても強力な方法なので、入力と出力をきちんと定義することができ、学習に使える良質のデータが大量にそろえば、さまざまな課題を高い精度で行える可能性があります。しかし、ただ「こんなことができるようになってほしいなあ」と思うだけではＡＩは作れません。つまり「言葉を理解できるＡＩ」「感情を持つＡＩ」のような漠然としたイメージを、漠然としたまま実現することはできないわけです。

今すでに世間では「人の言葉が分かるＡＩ」とか「人の心が分かるＡＩ」などといった

ことを謳っているシステムもありますが、それらの実体は「雑談をするＡＩ」だったり、「質問文を入力として受け付け、答えとなる単語を出力するＡＩ」だったり、「文章を入力として、『喜び』『怒り』『悲しみ』などといった感情の種類を出力するＡＩ」であったりします。漠然とした宣伝文句に踊らされないようにするためには、「そのＡＩがする具体的な仕事は、いったいどのように定義されているのか」を見極める必要があるでしょう。

また注意しなくてはならないのは、たとえ人間にとっては似たような仕事であっても、ＡＩにさせる場合は「まったくの別もの」である可能性があるということです。人間の場合は、難しい文章を外国語に翻訳できる人が、文章の要約や日常会話、ましてや言葉の聞き取りもできることなどはほぼ当たり前で、不思議でも何でもありません。よって、機械に対しても、「こんなに高度な翻訳ができるんだから、文章の要約ぐらい簡単だろう」と思いがちです。しかし機械にとっては原則として、翻訳と要約、対話、音声の認識はどれも異なる仕事です。先に述べたように、ＢＥＲＴやＧＰＴ－３といった言語モデルを「何でもできるＡＩ」と見なすのにも無理があります。

また、「機械学習によって作られたＡＩは、人間がすべてプログラムして作ったＡＩよ

り融通が利くはず」という意見を見たこともありますが、機械学習の「融通」は、あくまで「機械学習がうまくいっている場合は、本来の使われ方（＝それが本来するべき仕事）の範囲内で、開発時に使われるデータには存在しないデータに対しても、高い確率で正解を出せる」ということです。これは、必ずしも「本来想定していない使われ方をされても大丈夫」ということを意味しません。こういった点にも注意が必要です。

ブラックボックス問題

ここまでお話ししてきたように、機械学習を使ってＡＩを開発するときになされることは、私たち人間が行う知的な仕事を「何が入力で、何が出力か」という観点から定義し直し、機械がそれを上手に真似するための「関数」を求めることです。そこには、私たち人間がそういった仕事をどのように行っているかという視点は必ずしも入っていません。よって、ＡＩによる近似がいくらうまくいっていたとしても、ＡＩが人間と同じようにその仕事を行っているという保証はありません。

また、先にお話ししたとおり、現在の「言葉を扱うＡＩ」のほとんどは、深層学習を用いて開発されたニューラルネットワークです。つまり、言葉を扱うＡＩの中身は膨大なパ

ラメータを持つ関数であるわけです。言葉を扱うＡＩの中身がそういった関数であるということは、「今の機械の言葉」についての一つの重要な示唆を持っています。それは、「ＡＩの中身を見ても、なぜＡＩが言葉をそのように扱ったのかがよく分からない」ということです。

たとえば、「This is a pen.」という英語の文を入力として受け付けて、「これはペンです。」という日本語文を出力する機械翻訳システムを考えてみましょう。このＡＩにとっては、入力の英語の文も、出力の日本語の文も「数の並び」です。たとえば機械翻訳システムを深層学習で開発する場合、「英語の原文」と「日本語の訳文」のペアを訓練データとして使い、ニューラルネットワークの中のパラメータを調整していきます。調整の結果、そのネットワークが十分な数の英語の文に対して正しい日本語訳を出力できるようになれば、開発は成功したことになります。

しかし、うまく開発できたネットワークを見ても、いったい「This is a pen.」という原文の何がどうなって「これはペンです。」に訳されるのかは分かりません。私たちに見ることができるのは、膨大な数のパラメータを持つ巨大なニューラルネットワーク、つまり非常に複雑な関数です。もちろん、「This is a pen.」に相当する数の並びが、ネットワー

クの中の各部でどのように計算されるか、またその計算の結果が出力に向かってどのように伝わっていくかは分かります。しかし、そのネットワークが「これはペンです。」に相当する数の並びを出力するまでの過程は、あくまで「数の計算」でしかなく、原文のどの部分が訳文のどの部分にどのように反映されているのかを直接見ることはできません。つまり、私たち人間に理解できる形で「ＡＩの振る舞い」を捉えることはきわめて難しいのです。

こういった、「ＡＩがなぜそのような振る舞いをしたのかが理解できない」という問題は、ＡＩの「ブラックボックス問題」と呼ばれています。これは、言葉を扱うＡＩに限らず、深層学習を利用して開発したＡＩに共通して見られる問題です。「分からなくたって、うまく動くなら問題ないじゃないか」と思う人もいるかもしれませんが、先に指摘したように、今のＡＩは私たちが想定していないような間違いをします。たとえば平昌オリンピックの際には、ノルウェー選手団の料理人が機械翻訳を使って卵を１５００個注文したところ、訳文ではなぜか1桁増えて「１５０００個」となっており、注文の十倍の卵が届いて困ったという話もありました。[23]

中身がよく分からないということは、ＡＩが間違いを起こした場合に原因を突き止めることが難しく、改善点を見つけるのも難しいということを意味します。こういったことは、人々の健康や安全、経済的・社会的な利益に関わるような部分にＡＩ技術を応用するときに大きなハードルになります。

このような問題を受け、近年では、ＡＩの振る舞いを人間に理解可能な形で説明する技術に関心が高まっています。アメリカの国防高等研究計画局（ＤＡＲＰＡ）では、「説明可能ＡＩ（Explainable AI: XAI）」の開発が推進されています。[24] ただし、一部の研究者からは、そういった技術が本当にブラックボックス問題を解決できるのかを疑問視する声も上がっています。[25]

今のＡＩは言葉を理解している？

以上、この章では機械の言葉の現状についてお話ししました。今のＡＩは、以前よりも格段にうまく言葉を扱えるようになっています。しかし、私たち人間と同じように言葉を理解しているかというと、そうとは言えない面が多々あります。

もちろん、今のＡＩの中身で起こっていることの大部分が私たちに理解できない以上、