Keywords you should know. 第155回

知っておきたいキーワード

言語系生成AI

延澤志保†

†東京都市大学情報工学部情報科学科

"Generative AI in Natural Language Processing" by Shiho Hoshi Nobesawa (Department of Computer Science, Faculty of Information Technology, Tokyo City University, Tokyo)

キーワード：言語系生成AI，ChatGPT，自然言語処理，深層学習，大規模言語モデル

まえがき

2022年に発表されたChatGPTは世界中に話題を巻き起こしました．人間同士のように言葉で対話しながら欲しい情報を引き出すことができるChatGPTは，各国の政府や高等教育機関などからその使用の是非や注意点などについて次々と声明が出されるなど，社会問題にすらなりました．

本稿では，ChatGPTに代表される「言語系生成AI」について，人間の言葉をコンピュータで扱う自然言語処理の一分野との立場から概説します．

情報検索と質問応答

ChatGPTは，いわゆるAIチャットボットで，質問応答システムと言われる自然言語処理分野技術のひとつに当たります．

情報検索システム（検索エンジン）は1990年代に生まれ，2006年には「ググる」という動詞が流行語候補に選ばれるほど一般的になりました．検索エンジンに与えるキーワード群はクエリと呼ばれます．ユーザは例えば「ChatGPT，生成系，人工知能」などクエリを検索エンジンに提示し，検索エンジンはクエリとして与えられたキーワード群からユーザの検索意図を推定して，ユーザの欲しい情報が載っている可能性が高い文書を提示します．一般に，クエリに関連すると推定される文書は膨大な数に上るため，検索エンジンは文書をランク付けして可能性の高いものから順に出力することでユーザの負担軽減を図ります．ユーザが探す情報が載っている可能性が高いと見做された部分をハイライトする「強調スニペット（図1）」の出現で，情報検索はさらに便利になりました．☞

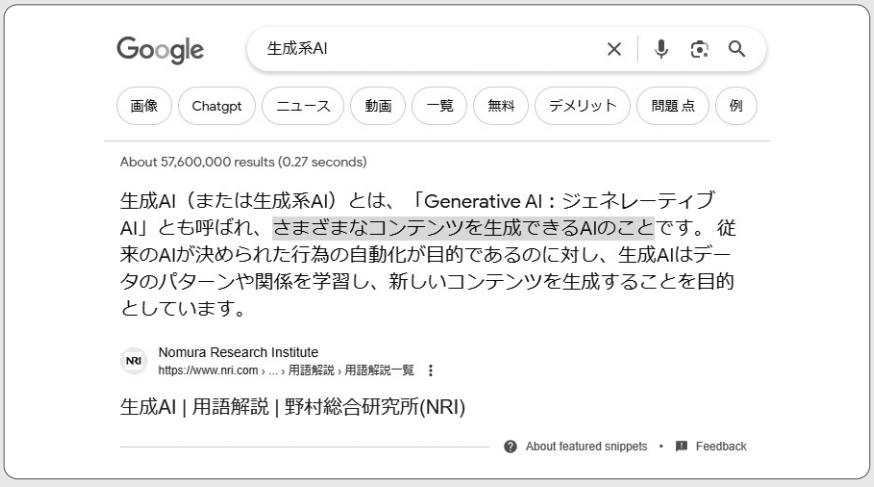

図1　強調スニペットの例

☞ しかし，そもそも情報検索の目的は欲しい情報を得ることにあり，システムが欲しい情報だけを抜き出して提示してくれるのであれば，ユーザは自ら文書を読む必要がなくなります．これが質問応答システムです．2011年2月，IBMの開発した質問応答システム「Watson」がアメリカのクイズ番組Jeopardy!で人間チャンピオン2人を相手に勝利し，人工知能の発展を強く印象付けました．その後Watsonを始め質問応答技術はさまざまな分野に応用され発展を続けています．

言語系生成AI

2022年，アメリカの人工知能開発企業OpenAIがChatGPTを発表すると，その自然な受け答えに世界中が驚きました．ChatGPTは，人間の言葉で入力した質問に対して，人間のように言葉などで回答します．この質問をプロンプトと呼びます．この少し前から，画像処理の分野でも，同様にプロンプトに対して対応する画像を生成するシステムが話題になっていました．ChatGPTの，まるで人間同士のお喋りのようなやりとりは，人工知能や情報科学の専門知識がなくても最先端の人工知能技術を誰でも気軽に使える仕組みを提供するものでした．ChatGPTのような対話的な質問応答システムは言語系生成AIと呼ばれるようになりました．

自然言語処理で用いる知識の源は電子化文書です．従来の自然言語処理技術では個々の語句，言語の構文など，さまざまな要素をそれぞれについて「特徴を掴む」ような学習をする必要がありました．受験に備えて問題をたくさん解いて特徴を覚えるようなイメージです．例えば，図2の英文は2種類の解析が可能な「構文的に曖昧な文」として有名な例です．図2の例では，構文解析を施した上で，語句の出現状況などから「time flies」という複合名詞の出現確率よりも「time」が「flyする」確率の方が高いことを推定することで精度の向上を図るというのが，従来の自然言語処理の基本的な動きでした．

それに対して，言語系生成AIを実現した深層学習では，従来とは比較にならないほどの膨大な量の電子化文書からその言語や語句に「慣れる」学習を行います．留学先で日々その言語にさらされているうちによく聞く言い回しが身に着くようなイメージです．言語系生成AIの出力が自然なのはそのためです．ニューラルネットワークについての研究は古くからありますが，これを大きく発展させ言語系生成AIの実現に貢献したのが深層学習技術による大規模言語モデル（LLM）です．語句をベクトル表現する技術によって，出現頻度の低い語句も含め，膨大なテキストから語句同士の関連などさまざまな情報を自動的に学習し積極的に活用することができるようになりました．また，再帰型ネットワーク技術などによって文脈を考慮した出力が可能になったこと（図3）が，自然な文の生成を実現しました．

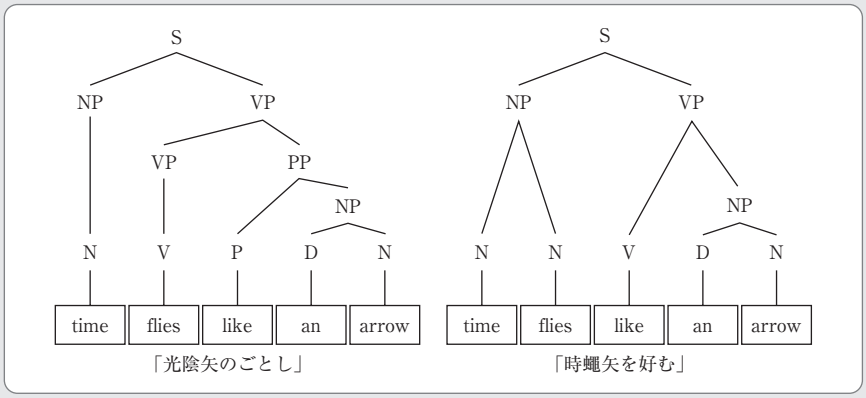

図2　従来の自然言語処理

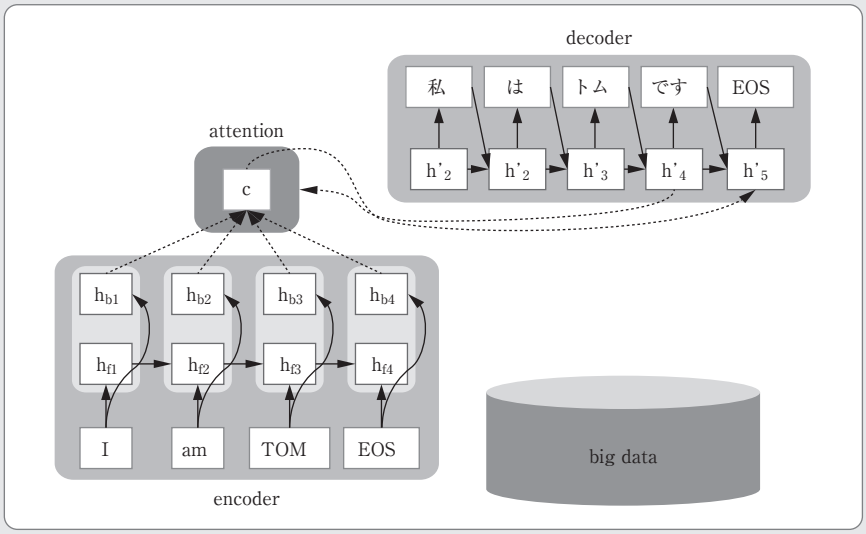

図3　ディープラーニングとアテンション

言語系生成AIとの共存

現在の言語系生成AIは，膨大な文書から得た言語モデルを基に出現可能性の高い言い回しを並べて出力しているようなレベルに過ぎず，流暢さ，自然さを主な目標としている段階で，回答の正確さや適切さについては今後の検討が必要です．現時点ではまだ，言語系生成AI自身が内容を理解し妥当性を判断しているわけではありません．そのため，場合によってはさらりと嘘をつく可能性（ハルシネーション）もあります．これは，質問応答システムとして活用するには，致命的な課題です．またヘイトスピーチに代表されるような不適切発言の抑制は，現時点では知識ベースの管理に頼っていますが，これは本来，各企業に任せてよい問題ではありません．現時点では話題性が先行しており，この部分は今後早急に対応が進められることになります．言語によってベースとする電子化文書の量も質も大きく異なるため，他言語から得た知識を活用する方法などの研究も進められています．

むすび

これまで，自然言語処理の分野では，情報検索，質問応答，意図推定，文生成，対話処理など，人工知能の実現に向けてさまざまな課題について技術を磨き上げてきました．今，人工知能は自ら学ぶ方法を得，自然言語処理の基本技術になろうとしています．人工知能が研究者から手法や教師データを与えられて処理を行う時代はそろそろ終わりを迎え，人工知能自身が自ら学び，研究者はその健やかな成長を手助けする立場へと変化する段階にきたようです．現時点の人工知能はまだまだ発展途上ですが，言語系生成AIの誕生は，人工知能の分野のひとつの歴史的なポイントと考えてよいように思います．

（2023年9月26日受付）

参 考 文 献

1）乾健太郎：“ChatGPTの出現は自然言語処理の専門家に何を問いかけているか”，自然言語処理，30，2（2023）

延澤 志保（のべさわ しほ） 慶應義塾大学大学院理工学研究科計算機科学専攻後期博士課程修了．東京理科大学理工学部情報科学科助手を経て，現在，武蔵工業大学知識工学部（現，東京都市大学情報工学部）情報科学科講師．博士（工学）．

キーワード募集中

この企画で解説して欲しいキーワードを会員の皆様から募集します．ホームページ（https://www.ite.or.jp/）の会員の声より入力可能です．また電子メール（ite@ite.or.jp），FAX（03-3432-4675）等でも受け付けますので，是非，編集部までお寄せください．

（編集委員会）