

言語処理から見たマルチモーダルAI

— LLM のさらなる変革 —

Multimodal AI from the perspectives of Natural Language Processing and Machine Translation



国立研究開発法人産業技術総合研究所 フェロー

辻井 潤一

国立研究開発法人産業技術総合研究所 フェロー、英国マンチェスター大学教授、東大名誉教授、国際計算言語委員会 (ICCL) 委員長、AAMT / Japio 特許翻訳研究会委員長

✉ j-tsuji@aist.go.jp

1 はじめに

ChatGPT が発表されて、まだ3年も経過していないのが意外に感じられるほど、この3年間の大規模言語モデル (Large Language Model、以下では LLM) の技術は急進展している。商用価値の大きさから、巨大 IT 企業、国家、LLM に特化したスタートアップなどが巨大な資金を投入する開発競争になっている。さらにこの動きは、当初 LLM が対象とした言語テキストだけでなく、画像・音声など、別のモダリティの情報も取り込む形のマルチモーダル化も進展している。

LLM を支える計算機構 Transformer の技術が有効なのは言語処理に限らない。また、大きなモデル (基盤モデルと呼ばれる) をまず作って、それを少量の学習データで個別タスク用に作り変える移行学習の考え方も、言語だけでなく他のモダリティの AI 技術にも適用できる。このことから、LLM とは独立に、画像、音声、動画の基盤モデル、さらには、ロボットなどの行動系のための基盤モデルの構築も活発化している。

これら一連のマルチモーダル AI という技術の動向が本年度の YEAR BOOK の特集テーマとなっている。私の所属する産業技術総合研究所でも、この方向での研究に積極的に取り組んでおり、本研究所の佐藤雄隆氏がこれについて寄稿している。

本稿では、私の専門である自然言語処理、とくに、機械翻訳の観点から LLM を論じ、言語理解の観点からのマルチモーダル化を考えることにする。

2 機械翻訳の変貌

私がよく使う LLM の機能に、英文校正と翻訳がある。以前は、改まったメールや手紙、あるいは、論文を書く場合には、懇意にしている人に Native Check を依頼していた。ただ、そう頻繁にチェックを頼むこともできないので、不安は残るが Native Check なしにメールを送ることも多かった。LLM の場合には、そういう気づかいは不要なので、頻繁に校正や翻訳機能を使っている。

私の使っているシステムでは、修正した箇所のリストとその理由も説明してくれるので、英語力の向上にもつながる。先日海外の知人、といっても、友達というほど親しくない人への英文メールを校正させると、随分と立派な英文に仕上げてくれた。ただ、それを読むと立派すぎて、友達とはいえないまでも、一応、良く知った同僚的な人に出すにはあまりにも立派でよそよそしい。システムが、テキストの調子をカジュアルに変えることもできますというので、その代案を頼む。代案は、逆に、相手の教授にこのようななれなれしい言葉遣いはしないだろうという部分があり、結局は、2つの案の折衷案にする作業をすることになった。このような作業が必要ではあるが、一人でいろいろ考えながら、英文メールを書くよりもはるかに早い。しかも、英文ですこし丁寧なメールを書くという、以前であれば気の進まない仕事にも、気楽に取り組むことができる。

ずっと以前、機械翻訳の研究をしていたころ、「原文のない翻訳」という論文を書いた。きちんとした原テキストを用意することなく、相手言語のテキストをユーザ

と対話しながら作っていく、という対話型の翻訳システムの論文である。書きたい人が、粗い原文を用意し、翻訳システムはそれをもとに必要な情報をその人とやりとりしながら、最終的な翻訳テキストを仕上げていくという論文であった。この論文で書いたようなユーザとシステムが会話をしながら、翻訳先の言語でのテキストを作っていくという原テキストなしの翻訳システムが、いま LLM で実現されている。

以前に書いた論文の主旨は、「書き手が伝えたいと思っている情報を相手言語で表現する作業」を共同して行う翻訳システムを考えようとする提案であった。

当時（あるいは、現在でも）、機械翻訳の研究コミュニティには、「整った元言語でのテキストがあり、翻訳は、このテキストが持つ情報を相手言語に移す作業」という考え方が主流であった。これに対して、元言語のテキストが翻訳に必要な情報をすべて含んでいるものではなく、翻訳に必要な情報をテキストの書き手から聞き出しながら翻訳テキストを作っていくシステムを考えるべきではないか、という趣旨であった。

翻訳のプロでも、質の良い翻訳を作るためには、テキストの書き手から直接情報をとらなければならない。例えば、英語での特許申請書を作る専門家は、整った日本語の特許申請を翻訳するというよりも、申請者からの概略から出発して、特許申請者本人から話を聞いたり、質問したりしながら申請書を仕上げていくという。また、ノーベル賞作家（スペイン語かポルトガル語？）の作品の多くを英語翻訳したマンチェスター大学の教官は、その作家と懇意になり、私的旅行と一緒にするような関係になった、それほど、翻訳の過程で頻繁なやり取りがあった、とのことであった。

翻訳というのは、原文が担う情報を相手の言語に引き移すという作業ではなく、原文を理解し、その理解から相手言語のテキストに必要な情報を取り出して、相手言語のテキストを作り出していく作業である。

こういうと、当たり前のようなのであるが、機械翻訳コミュニティでは理解に基づく翻訳は一種のタブーであった。前の AI ブームの時には、「理解にもとづく翻訳」が提唱されたこともあったが、技術的には不可能であるとされて、忘れられていった。その状況を LLM は大きくかえることとなった。

3 機械翻訳と理解、マルチモーダルの可能性

「文の内容を理解することなく翻訳する」機械翻訳の立場は、ニューラル翻訳、および、それに Attention 機構を入れることで現在の Transformer の基礎となった研究においても、保持されていた。テキストの理解が機械翻訳に影響を持つようになったのは、この Transformer の技術を使って巨大な言語モデル (LLM) が構築されてから以降である。LLM が、理解の能力を示すことが明らかになり、翻訳の機能も、いわば by-product として、実現されている。

元言語の文だけでは情報が不足していて、文の書き手や状況の理解結果から情報を補う必要がある簡単な例を実験したのが、次の例である。私の使っている LLM に日本語の文「テーブルの上のランプを消してください」を翻訳させると、

Please turn off the lamp on the table

という翻訳が出力される。

しかし、日本語の「上」は、英語への翻訳を考える場合には「曖昧」で、「A の上の B」で A の表面に接して B がある場合には on、そうではなくて、接しないで A の上方にある場合には above、と訳出される。これを間違って訳出すると、A と B の位置関係に関して、間違っただけで情報が伝わる。ただ、この情報は原文の日本語には明示されず、正しい翻訳を作るための情報が欠如している。

そこで、ランプは天井からつるされているという状況の記述を付け加えて、

ランプが天井からつるされています。そのとき、次の文を英訳してください。「テーブルの上のランプを消してください」

とすると、LLM は、on を above に変えて翻訳する。翻訳する文ではなく、その前段で状況を説明するテキストを付加すると、その状況記述を理解して、それを翻訳に反映する。

さらに、「なぜ、on ではなく above を使うのか」という質問をすると、上述した on と above の差を一般的に説明したあと、図 1 の説明を付け加える。背景記述

✓ まとめると：

- on the table → テーブルの上に置かれている（接触している）
- above the table → テーブルの上方にある（接触していない）

この違いを踏まえて、あなたの文脈（天井から吊るされたランプ）では
"above the table" が自然で正確な表現になります。

もし他にも似たような前置詞の使い分けで迷っているものがあれば、ぜひ教えてください！

図1 on と above の状況理解による使い分け

の状況から、「天井からつるされているランプは、テーブルの上方にある」という位置関係を理解している。on と above の使い分けの規則を、現在の個別状況に合わせて適用している。

また適切な翻訳は、話し手と聞き手の関係や両者の

置かれている状況でも変化する。以前の翻訳システムでは取り扱えないものであった。状況を背景説明として与えて、翻訳がどのように変化するかを実験したが、図2と図3である。LLM は、おなじ日本語「やめなさい」を、状況によって翻訳を適切に切り替える。

[入力]

3歳の子供が、飼い犬にいたずらをしている状況で、母親が「やめなさい」といった。この「やめなさい」を英語に翻訳してください

[LLM の出力]

この場合の「やめなさい」は、英語では状況や口調によっていくつかの言い方ができますが、母親が子供に注意している場面では、以下のような表現が自然です

- "Stop that."
- "Don't do that."
- "Leave the dog alone." (犬に関する具体的な指示)

例えば、母親が少し厳しく言っているなら

- "Stop it right now."
- "That's enough."

優しく注意する場合は：

- "Sweetie, please stop bothering the dog."

どの表現を使うかは、母親の口調や子供との関係性、状況の緊急性によって変わります。どんなニュアンスで訳したいか、もう少し教えてもらえますか？

図2 「やめなさい」の翻訳－母親と幼児

[入力]

高校の教室での授業中、後ろのほうに座っている生徒二人が小声で話をしているので、教師が「やめなさい」といった。この時の「やめなさい」を英語に翻訳してください

[LLM の出力]

高校の教室で授業中に教師が生徒に「やめなさい」と注意する場面では、英語では以下のような表現が自然です：

- **"Stop talking."**
- **"That's enough."**
- **"Quiet, please."**
- **"Pay attention."**
- **"Focus on the lesson."**

少し厳しく言う場合は

- **"I need you to stop talking now."**
- **"This is not the time to chat."**

教師らしい口調で、状況に応じて使い分けられます。

図3 「やめなさい」の翻訳—教師と生徒

この例では、「やめてください」の訳のいくつかの可能性を知っているだけでなく、可能な訳の中に Dog、talking、lesson などの背景説明からの補いが現われたり、説明にも教師、生徒、母親、授業中といった単語があらわれたりするなど、LLM が状況を理解し、その理解を翻訳に反映させていることがわかる。また、翻訳の可能性を提示するに際して、相手との関係性、緊急性など、それぞれの訳文のニュアンスについてもコメントしていることも面白い。

これらの実験では、翻訳を行う背景文脈を、言語プロンプトで与えた。このような背景文脈の言語化をしなくても、システム側が自らの視覚や聴覚からの情報を使って、背景文脈を理解し、それを翻訳に反映するマルチモーダルな翻訳器が、マルチモーダル翻訳器である。

現在、商用化されひろくつかわれている携帯型の翻訳器がさらに発展して On-Line 会議や Virtual Office のアバターに翻訳機能を付与したシステムなどが、視覚による状況理解や、音声や表情からの感情認識などを統合し

たマルチモーダルな翻訳システムとして、次の研究課題となる。

実際、本特集で議論されるように、LLM のマルチモーダル化は活発に研究されている。テキストでの状況記述から画像を生成したり、与えられた画像や動画からその状況を説明するキストを生成したりする研究も活発化している。また、音声認識や生成の技術も実用化のレベルに達し、話し言葉の音声から、話者の感情（怒り、喜び、悲しみなど）を認識する研究、あるいは、表情から人の感情状態を推定する研究も進展している。これらの技術が充実することで、言語化された状況説明を必要としないマルチモーダル翻訳器は、夢物語ではなくなりつつある。

4 LLM の理解と機械翻訳

LLM は、論文を要約したり、逆に簡単な箇条書きメモから長いテキストを生成したりと、テキスト操作に関するタスクに優れた能力をしめす。LLM の出現以前に

は、これらのタスクは、知識に基づくテキスト理解の能力を必要とし、一般的に使える技術とするのは困難だと考えられていた。

プログラムの作成とか 数学の問題を解くとか、あるいは、科学の新しい仮説を提案など、さまざまな能力を示す LLM が、人間にとっては、取り立てて特殊な能力とは思えないテキスト理解の能力を示すことは、当然のように見えるかもしれない。

しかしながら、状況に関する膨大な知識をある程度抽象化して持ち、それをテキストに記述される個別な状況に当てはめて理解する能力は、人間知能の根幹を構成するもので、きわめて実現がむづかしいものであった。LLM がこの能力を持ったこと、すなわち、知能を構成する汎用的な能力を持ったことが、LLM の成功を支えている。このような汎用的な知的能力が、プログラム作成・数学の問題解決・科学の新発見といった能力の獲得

をどの程度支えているのかを見る、あるいは、こういう能力を訓練すると、基盤となる汎用の知的能力がさらに向上するのかなどが、現在の研究課題となっている。

本稿では、LLM が示すテキスト理解能力がどのようなものかを翻訳を通して試してみることにする。以前の人工知能研究では、システムのテキスト理解能力をみる典型的な課題の一つとして、代名詞の指示対象同定の課題があった。LLM のこの能力を翻訳タスクで試してみることにする。

英語の代名詞「they」は、指示対象の性別によって「彼ら」「男たち」「彼女ら」「女たち」などと翻訳が変化する。テキストの内容を理解していないと指示対象が正しく決定できない場合があるので、LLM の理解能力を見るわかりやすい例になる。

図 4 が、その例題である。最後の文が (a) か (b) かで、最後の文の they の訳語を変化させる必要がある。最後

次のテキストを日本語に翻訳してください。

Three women, Mary, Jane and Sophia lived in a house far away from the town center. One night, Mary noticed that two men broke in the house. Since they stored guns in the house, they started to shoot them. They also shot back. After a while, they were all killed.

(a) They left the house with money stored in the safe.

(b) They called the local police station.

図 4 テキスト理解と代名詞の指示

最後の文が(a)の場合

以下がご依頼の英語テキストの日本語訳です

メアリー、ジェーン、ソフィアの 3 人の女性は、町の中心から遠く離れた家に住んでいました。ある夜、メアリーは 2 人の男が家に侵入したことに気づきました。家には銃が保管されていたため、彼女たちはその銃で男たちに発砲しました。男たちも撃ち返しました。しばらくして、全員が殺されました。男たちは金庫に保管されていた金を持って家を去りました。←

内容が少しショッキングなので、もしこの文章を使う目的があれば教えてください。必要に応じて調整や補足もできます。

図 5 代名詞の訳 - (a) の場合

最後の文が(b)の場合

以下がご依頼の英語テキストの日本語訳です

メアリー、ジェーン、ソフィアの3人の女性は、町の中心から遠く離れた家に住んでいました。ある夜、メアリーは2人の男が家に侵入したことに気づきました。家には銃が保管されていたため、彼女たちは男たちに向かって発砲しました。男たちも撃ち返しました。しばらくして、全員が殺されました。その後、彼女たちは地元の警察署に通報しました。

この文章には矛盾があるようです。「全員が殺された」とあるのに「彼女たちは警察に通報した」と続いています。もし物語の構成や翻訳の調整が必要であれば、お手伝いできますので教えてください。

図5 代名詞の訳・(b)の場合

の「金をとって去った」のは侵入者、「警察に電話した」のは居住者と考えのが、人間の場合には普通であろう。人間の場合には、「家に押し入ったグループ」が「警察に電話する」のは不自然だとか、「居住者がお金をとって逃走することはなかろう」といった常識から、どのグループが最後の行動をとる可能性が高いかを判断できる。

LLMの翻訳結果を、図5に示す。人間が普通に解釈するように解釈し、それを翻訳に反映している。

さらに、(b)の翻訳のあとには、英文がよくないから、書き直せというコメントがある。最後の文を除いたテキストを読ませ、「誰が殺されたと思うか？」と質問すると、「多分、2つのグループが全員殺された」と答える。確かに、与えた例題がわざとわかりにくくするために、パ

ズルのような悪文になっている。そこで、より明瞭なテキストに書き直すことを依頼した結果が、図6である。

図6のLLMが修正したテキストでは、複数回現れていたtheyが、「the women」、「the intruders」にいずれも正しく置き換えられている。さらには、break-inした男たちをintrudersと、もとのテキストにはなかった単語を使って表現している。

このintruderへの書き換えは、記述された状況と単語intruderの意味とがシステム内部で結び付けられていることを示し、現在のLLMが、人間言語が構成する概念空間をかなりの程度うまく内的に構成していることを示している。

こちらが、内容が自然に伝わるように修正した英文です

Three women—Mary, Jane, and Sophia—lived in a house far away from the town center. One night, Mary noticed that two men had broken into the house. Since they kept guns in the house, the women started shooting at the intruders. The intruders shot back. After a while, the intruders were killed. The women then called the local police station.↵

図6 LLMの校正結果

5 言語とマルチモーダル

さて、intruder の例からをわかるように、言語とそれが表現するもの(とりあえず、概念と呼ぶ)との関係、言語とその周辺文脈との関係は、それほど単純ではない。視覚や聴覚の情報だけで、例えば、授業でのお喋りを認識して、教師の「やめなさい」をお喋りの制止のための発話と推論することは、それほど簡単ではない。

プロンプトの言語で表現された状況

**子犬にいたずらをしている子供に母親が注意
授業中のお喋りを教師が注意**

をマルチモーダルなセンシングデータから認識するためには、LLM が持つ一般的な状況に関する知識にセンシングデータをうまく結びつける処理が必要となる。

このような言語と認識の相互関係に関係する課題を全体として解くのは容易ではない。当面は、状況を限定した応用場面でのマルチモーダル LLM の応用が進むことになる。

典型的な例は、ロボットとの対話に出てくる具象名詞の記号接地 (Symbol Grounding) と直示 (Deictic Reference) の問題がある。これは、現在、多くの研究グループが取り組んでいる課題である。

たとえば、

「テーブルの上のリンゴをとって、棚におけ」

という依頼をロボットが実行するためには、リンゴという単語を対応する概念（一般的なリンゴの形状、色、重さなど）に接地し、その情報を使って、ロボットのおかれている状況の中での特定のリンゴを同定するという直示の参照先決定を行わなければならない。

マルチモーダル LLM が「リンゴ」という概念とセンサー情報との関係を保持していれば、この直示的な参照課題を解くことができる。

この課題は、多方面への展開があり得る。直示的な側面を強調すると、人の指差行動を理解し、複数のリングの中の一つを同定するとか、物を把持する行為も物の形状、重さ、重心の位置などによって異なるので、物によって実際の把持行為を変化させるといった課題もある。後

者の問題は、Vision-Language-Action の AI モデルという、視覚・言語・行動を結びつけるという大きな研究分野に発展しつつある。

また、より言語的な応用としては、コールセンター業務への LLM 応用がある。現在、LLM は、自動応答やオペレータの負担軽減に使われ、有効との報告がされている。ただ、文字によるユーザの問い合わせだけでは、ユーザの状況、例えば、急いでいて苛立っている、強い不満を持っている、余裕をもって回答を求めているなどが把握できず、個別のユーザの状態に合わせたサービスができていない。

文字テキストでは同じに見えるメッセージであっても、人間の場合、ユーザのおかれた状態・声の調子・表情・身振りなどから、まったく違ったメッセージを受け取り、それに応じて適切な応答を行う。コールセンターだけではなく、ユーザとの対話によってサービスを提供するシステムは、数多く研究され実用化も進んでいる。LLM とマルチモーダル技術の融合により、個別ユーザのおかれた状況に応じて、より適切なサービス提供が可能になろう。マルチモーダル AI と LLM の融合は、人との対話で自律的にタスクを実行するエージェント AI での必須の技術となる。

6 おわりに

言語は、ある状況に置かれた人間がその状況を認識したのちに人間内部の情報処理機構が表出したもの、あるいは、外部の状況から切り離された状態で思考した結果が外部に表出されたもの、である。人間の解釈・思考が間にはいることで、実際の物理世界からは切り離された情報媒体である。

これに対して、視覚、聴覚、触覚といった五感からの情報は、外部の物理世界から人間内部の情報処理機構に送り込まれるものである。これらが人間内部の情報処理に結びつけられ、思考・推論を経て、再び言語として表出される、あるいは、外部世界に働きかける行動として、再び、外部世界に戻される。

膨大なテキスト集合から学習された LLM という内的情報処理に視覚、聴覚という外部からの結び付け、その結果として、新たな言語の表出や行動を行うという模式は、AI の究極の形であろう。LLM という言語的思考の技術の次なる技術として、マルチモーダル LLM の研究

の活発化は、人工知能技術の自然な流れであろう。

本稿では、人間知能をなぞる技術としての人工知能という見方から、LLMのマルチモーダル化を議論した。しかしながら、マルチモーダル化を視覚、聴覚といった人間の感覚器からのデータに限定する必要はない。本稿の議論は、対象世界をデータ化したものを思考・推論の処理に結び付けるAI一般、例えば、自動走行車の技術にも当てはまる。あるいは、株価、為替、貿易取引データなど、金融という対象世界のデータと思考に結びつける投資の意思決定を支援する人工知能システムといった、より幅の広い応用分野の技術にも当てはまる。外界からの情報と、言語的な定性的な情報を統合して判断に結びつける枠組み一般の議論である。

我々が手にしたLLMという極めて強力な技術革新が、マルチモーダル化を経て、次のさらに大きな革新へとつながることを期待したい。