



NTT人間情報研究所
上席特別研究員

西田 京介 Kyosuke Nishida

特定分野の専門知識を持った高性能で低消費エネルギーのLLM, tsuzumi

2022年11月に生成AIであるChatGPTが発表され、それ以降大規模言語モデル（LLM：Large Language Models）が世界的に急速に注目されるようになりました。以降次々と新たなLLMが発表される中、高性能という「明」に対して消費エネルギーの増大という「暗」の部分もクローズアップされてきています。2023年11月1日に発表されたNTT版LLM「tsuzumi」は、「LLMの大規模化・一極集中化ではなく、異なる個性を持った多数のAIが連携することにより、ヒトと一緒に社会のWell-beingを実現する未来をめざす」ことをビジョンに開発が進み、「明」の実現と「暗」の解決を同時に成し遂げることが可能になり、2024年3月25日に商用開始が発表されました。NTT人間情報研究所 西田京介上席特別研究員に、tsuzumiのプロジェクト発足経緯・特徴、LLMの動向、そしてレッドオーシャンに向けたチャレンジについての思いを伺いました。



発表から商用開始までを5カ月で達成。NTT版LLM「tsuzumi」

現在、手掛けていらっしゃる研究について教えていただけますでしょうか。

2023年11月に発表しましたtsuzumiをはじめとする、大規模言語モデル（LLM：Large Language Models）の研究をしています。

2009年にNTT入社以降、2013～2015年に事業会社に在籍した期間も含めて、さまざまな研究開発を行ってきました。2017年ごろから取り組み始めた自然言語処理分野の研究では、AI（人工知能）による自然言語テキストの「機械読解」、人が目からテキストを読むようにテキストと併せて視覚情報も融合させて理解させる「視覚的機械読解」等のテーマに取り組み、言語処理学会年次大会（NLP）で2018～2024年に優秀賞（2018年、2021年は最優秀賞）を受賞し、ICDAR（International Conference on Document Analysis and Recognition）という文書解析・認識に関する国際会議における、2021年のVQA（Visual Question Answering）コンペティションでは、2位に相当するrunners-upを受賞しました。

特に、2018年にGoogleからBERT（Bidirectional Encoder Representations from Transformers）という言語モデルが発

表され、NTT版の日本語BERTをつくる話が出てきたころから、言語モデルに関する研究に取り組むようになりました。2022年11月に生成AIであるChatGPTが発表されLLMが世界的に注目されるようになるのと並行して、これまでになかった汎用AIの初期段階の登場とも思えるようなパラダイムシフトが世界で進行していると感じ取り、NTTとしてもLLMに取り組む必要性・重要性を訴求して2023年2月にLLMに関するプロジェクトを立ち上げ、それ以降LLMをメインのテーマとして取り組んでいます。

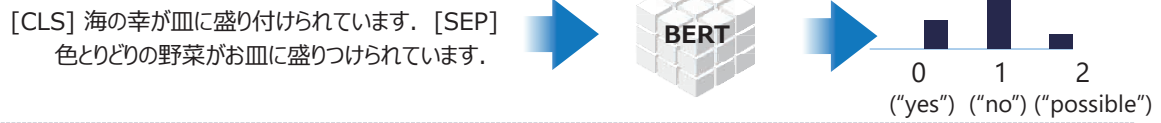
プロジェクトでは、コンピューティングリソースの調達や学習に必要なデータの収集から手掛け、チームメンバーとともにLLMの構築に取り組みました。6月ごろからLLMの事前学習を行い、11月1日にtsuzumiを発表し、2024年3月25日に商用利用開始の発表、という非常に短期間での開発となりました。

LLMとはどのようなものなのでしょうか。

言語モデル（LM：Language Models）は、単語（トークン）列の生成尤度をモデル化し、将来の（あるいは欠落した）トークンの確率を予測することで結果を得るモデルです。LLMは、Transformerと呼ばれるニューラルネットにおける深層学習モデルをベースとしており、高精度な情報検索やプログラムの生成・修正を含むさまざまな言語処理タスクを可能にします。LLMの高

- 事前学習で得た能力をさらに発展させるべく、さまざまなタスクにおける指示と応答を**明示的に学習する**
- このチューニングにより汎用性が高まり、未知のタスクにおいても学習することなく対応できるようになる

従来のタスクデータへのチューニングイメージ



インストラクションチューニングされたモデル

前提文と仮説文の間に含意関係は成り立ちますか？

前提：海の幸が皿に盛り付けられています。
仮説：色とりどりの野菜がお皿に盛りつけられています。
出力は「はい」「いいえ」「どちらでもない」から選んで、
その根拠も述べてください。

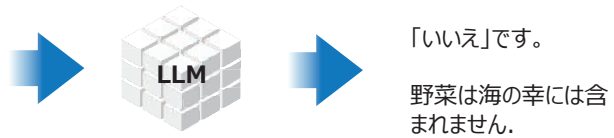


図1 インストラクションチューニング

度な言語処理能力は、主として「事前学習」「インストラクションチューニング」の2つのプロセスを経て実現されます。

事前学習は、言語モデルを事前学習させるプロセスで、Transformerを大量のテキストによる大規模コーパスにより事前学習し、事前学習済言語モデルをつくります。2018年のBERTの登場から、この言語モデルが自然言語処理の基盤として有効であると注目を集めています。2020年に登場したGPT-3では、タスクを定義したプロンプトに続くテキストを生成させることにより、ファインチューニングせずとも任意のタスクをある程度扱えるほど、事前学習済言語モデルの能力が発展しました。

インストラクションチューニングは、事前学習で得た能力をさらに発展させるべく、さまざまなタスクにおける指示と応答を明示的に学習するプロセスです。従来のタスクデータへのチューニングでは、例えば、2つの文章をセパレーターと呼ばれるトークンで渡してBERTに入力することで、0 (“yes”), 1 (“no”), 2 (“possible”)といった3つのラベルがついたスコアを出し、これで入力した2つのトークンの関係性を判定していました。これはデータドリブンな学習で、言語モデル自身はどのようなタスクが与えられているかを全く知らない状態でデータドリブンに学習を行います。一方、インストラクションチューニングでは、これらのタスクを言語で定義していきます。入力テキストに対して、どういうタスクを解いてほしいのか、どういう出力をしてほしいのかということ言語で定義し、LLMがその指示にしたがって答えてくれるようにするチューニングです(図1)。これにより汎用性が高まり、未知のタスクにおいても学習することなく対応できるようになります。

さて、現在のLLMは大規模化する傾向にあり、Scaling Raws*に基づき、モデルサイズ(パラメータ数)の大きな言語モデルが次々とつくられており、パラメータ数が1 T(Trillion) = 1000 B(Billion) = 1兆規模を超えるモデルも開発に入っています(図2)。例えばGPT-4ではパラメータ数が1.76 Tともいわれています。

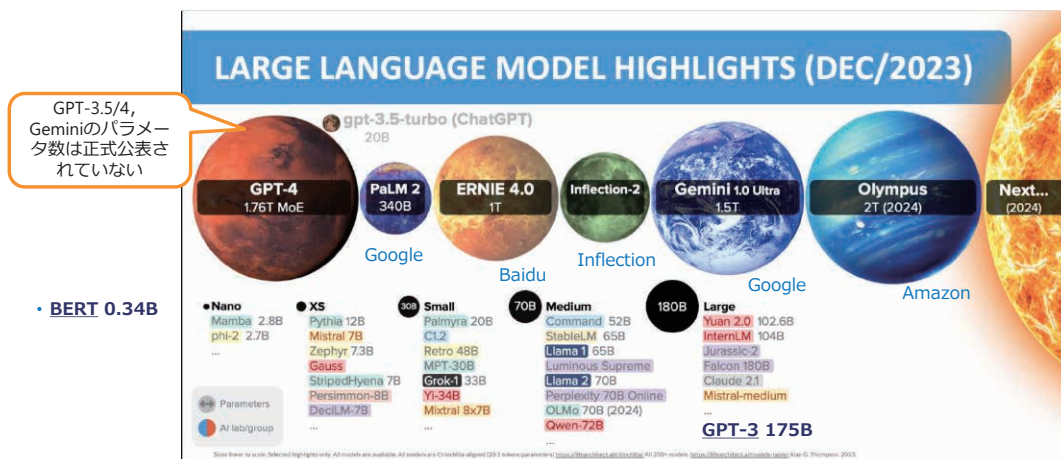
一方で、この傾向が進むにつれて、エネルギー消費も指数関数的に増大し、コストもそれにつれて増大します。175 Bのパラメータを持つといわれているGPT-3でモデルを学習しようとする、1回の学習で原子力発電所1基分に相当する、約1300 MW/hのエネルギーを使うという試算もあります。

鼓と tsuzumi, キーワードは「日本語」「小型軽量」「柔軟なチューニング」「マルチモーダル」

tsuzumiはどのようなLLMでしょうか。

何でもできる汎用的なLLMに向けて大規模化が世界の潮流にある一方で、エネルギー問題が課題となってきています。そこで私たちは、「LLMの大規模化・一極集中化ではなく、異なる個性を持った多数のAIが連携することにより、ヒトと一緒に社会のWell-beingを実現する未来をめざす」ことをビジョンに、モデルサイズで対抗するのではなく、特定分野の専門知識を持った高い性能のLLMを、言語学習データの質と量を向上させるというアプローチでtsuzumiを開発することとしました。ここで、「異なる個性を持った多数のAIが連携する」については、例えば医療分野や教育分野に強いLLMといったような、小型で性能が良い複数のLLMが連携することで、全体として良いLLMを形成する「AIコンステレーション」というコンセプトに基づいています。そして、私たちのLLMの特徴である「日本語に特に強い」「小型・軽量ながらも良い性能を持つ」「チューニングが柔軟に行いやすい」「そしてマルチモーダルへの拡張ができる」という点を楽器の鼓になぞ

* Scaling Raws: 自然言語処理モデルのモデルサイズ(パラメータ数)、データセットのサイズ、トレーニングに使用される計算量が増えるにつれて、損失が「べき乗則」に従って減少するという法則。



出典: <https://lifearchitect.ai/models/>

図2 LLMの大規模化

邦楽器

日本語が特に強い

小さい

小型軽量

調べ緒により調律が容易

柔軟なチューニング

見た目、音、演奏時の所作が美しい

マルチモーダル

図3 鼓とtsuzumi

らえて、tsuzumiと命名しました(図3)。

tsuzumiの評価に関しては、日本の地理・政治・歴史・社会の4カテゴリに関する40問の評価セットにより、2つのモデルの出力を比較して評価を行うRakudaベンチマークでの結果を紹介します。このベンチマークでは、人間による評価の代わりにGPT-4を評価役として用います。図4に示す評価例では、GPT-4がtsuzumiとGPT-3.5の出力を比べて、具体性や詳細度の観点でtsuzumiを高く評価しています。ベンチマーク全体としては、GPT-3.5や他の日本語向けLLMに対して勝率が大きく上回る結果になっており、tsuzumiが高い日本語の理解・生成能力を持っている1つのエビデンスとなっています(図5)。

tsuzumiの事前学習については、専門文書からエンタテインメント系まで幅広いドメインで、合計1000Bを超えるトークン数を用意して事前学習を行いました。今後、さらに学習データの質や量については改善をめざしていきます。また、事前学習中の言語の比率についても重要で、限られた日本語コーパスを基に日本語

に強いLLMを構築するためにどのようなバランスで事前学習すればよいのか、という点はまだ明らかになっていません。最近では、海外製のLLMをベースに、日本語を追加で学習する方式も採用されており、これも非常に良い結果を生み出しています。私たちは、内製でゼロから独自に事前学習を行いました。日英中心にコードや多言語のデータも加えています。特に多言語を加えることの効果など、今後さまざまな観点から検証を行っていきます。

事前学習コーパスの作成においては、文章をトークンに分解するトークナイズや前処理についても注力しています。トークナイズについては日本語の単語を考慮した分割を実施しています。単語の制約を考慮しない場合、学習コーパスの偏りから発生する不自然で冗長なトークンが生まれやすくなってしまいますが、私たちのトークナイズは、長年取り組んできた単語分割に関する研究成果を活用することで、自然な単語の分かち書きに近い分割になります。また、できる限りテキストの質を高めるため、さまざまな方式で、ノイズとなるテキストを除去する作業を行っています。

質問

平成時代から令和時代への移行期に見られた社会的、政治的な変化を挙げ、その意義について説明してください。

tsuzumi-7B の回答

平成時代から令和時代への移行期には、社会的・政治的に多くの変化が見られました。以下にその主なポイントを挙げます。

1. インターネットとデジタル技術の普及:この時期はスマートフォンやパソコンなどの情報通信機器が急速に広がり、人々の生活スタイルにも大きな影響を与えました。特にSNS(ソーシャルネットワーキングサービス)の利用者数が増加し、個人間でのコミュニケーションが容易になったことは特筆すべき点です。また、クラウドコンピューティングやAI(人工知能)の発展も進み、ビジネスシーンでもこれらのテクノロジーが活用されるようになりました。(…)

GPT-3.5 の回答

平成と令和の移行期には、いくつかの社会的、政治的な変化が起こりました。以下にいくつか挙げて説明します。

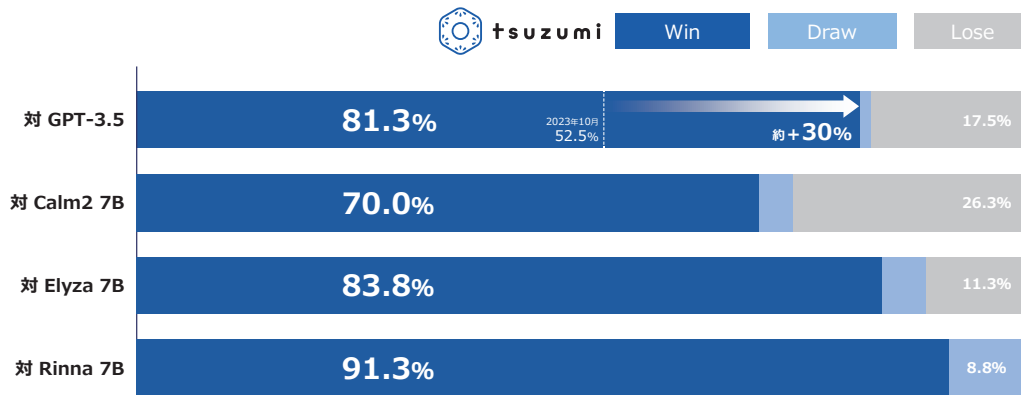
① 政治的变化:平成時代に比べ、令和時代は政治の安定が続いています。また、現在の政権は若い世代の支持を得ています。これは、政治の価値観が世代間で変化してきていることを示しています。また、平成時代に比べ、令和時代は海外からの人口流入が増加しており、外国人材を積極的に受け入れる政策が進んでいます。この流れは、日本の人口減少問題やグローバル化に対応するために重要な変化となっています。(…)

GPT-4の判定:有用性, 関連性, 正確性, 詳細度, 日本語能力

アシスタント1の回答は、平成から令和への移行期に見られた社会的、政治的な変化を具体的に5つのポイントで詳細に説明しています。それぞれのポイントについて、その意義や影響を具体的に述べており、情報の有用性と関連性が高いと言えます。また、日本語の表現も適切で、理解しやすい文章となっています。(…)
両者ともに質問の意図を理解し、適切な回答を提供していますが、より具体的に詳細な情報を提供したアシスタント1の回答の方が優れていると判断します。したがって、私の評価はアシスタント1(※tsuzumi)が高いです。

図4 Rakudaベンチマーク: tsuzumi-7B vs GPT-3.5 判定例

世界トップクラスの日本語性能を向上し、GPT-3.5に対する勝率は約30%アップ



*日本語性能の評価方法
rakudaベンチマーク: <https://yuzuai.jp/benchmark>
日本の地理・政治・歴史・社会に関する40問の質問。GPT-4による2モデルの比較評価(40問×提示順2)で採点

図5 日本語性能比較: Rakudaベンチマーク

インストラクションチューニングについては、幅広いカテゴリの質問や指示でインストラクションチューニングデータを独自に整備して学習に利用するとともに、有益性および安全性の両面から、データの整備を現在も拡大しています。

tsuzumiが商用サービスを開始しましたが、今後どのような展開をお考えでしょうか。

メディカル領域やソフトウェア開発など、専門用語や業界に特有の表現が多く含まれる場合は、従来の生成AIが十分な性能を発揮しないケースも多々あると思います。tsuzumiは、こうした業界に特有のデータに対してもカスタマイズが可能のため、AIを活用できる領域を広げることができます。また、顧客サポート領

域では、お客さまのCX(Customer eXperience)向上のために、図表等のマニュアル類の読解とお客さま情報のアップデートによるパーソナライズが不可欠です。tsuzumiは、世界トップクラスの日本語処理能力とともに図表読解もサポートしていくことで、コンタクトセンタや相談チャットボット等顧客サポート領域における進化を支援します。

特にメディカル領域においては、日本では電子カルテの導入は進んでいますが、同じ症状でもカルテの書き方は病院、医師により異なるため、カルテデータを集積して分析活用することが困難な状況でした。tsuzumiは、超軽量で柔軟に、またセキュアにカルテデータを処理できるため、ドクターが記録した医療データを読解し、共通フォーマットに適切な表現で配置し、分析できる状態にします。

また、NTTドコモのコンタクトセンターでは、お客さまから年間4000万件以上の問い合わせを受けています。現在、コミュニケーションやチャットボットが対応していますが、マニュアル類の種類も多く、新しい情報が追加される頻度も高いため、適切な回答の作成や、チャットボット用のQ & A作成に対して多くの手間がかかっています。tsuzumiを導入することで、通話内容や図表入りのマニュアルを正しく理解してお客さまのご要望を適切に把握し、最適な回答をお返しできるようになると考えています。

2024年3月25日にtsuzumiのサービス開始をニュースリリースしましたが、研究のフェーズとしては基盤になるモデルをつくったところであり、それをNTTグループ全体のエンジニアとともにお客さまに価値を提供していくためにブラッシュアップしています。私はそれをサポートするとともに、AIコンステレーションを含めてtsuzumiを汎用化し、人とAIが自然に共生する世界をめざして研究を進めていきたいと思っています。LLMは目覚ましい発展を遂げていますが、社会実装を進め、AIがあらゆる環境で人を支援するためにはまだ不足している点が多いです。特に、AIが人と同等の入出力インタフェースを持ち、マルチモーダルなタスク、身体性を必要とするタスクに取り組めるようにすることが重要です。現在、私たちは視覚と言語の融合理解をメインに取り組んでいますが、今後、聴覚、力覚、触覚、さらには脳波などの生体信号を言語モデルとどのように結び付けていくかについて取り組んでいきたいと考えています。

今重要なこと、今自然にやるべきことをめざすという意識でチャレンジ

研究者として心掛けていることを教えてください。

私は研究を進めていくうえで、「アカデミックとビジネスの両方に関与することをしたい、その時々で皆が自然にやるべきだと思える重要な研究をしたい」という思いを持っており、チームのメンバーにもよく話しています。2017年ごろは、表データを人のように読む「表の読解」をテーマとして取り組み、それがコンタクトセンターのソリューションに利用され、その後2018年ごろに、自然言語を理解・生成する機械読解・要約に組み込み、COTOHA[®]というNTTコミュニケーションズのサービスに利用されました。そこから視覚的な文章読解というような、より人間に近い条件下で人間の言葉を理解させる方向性に進んできており、その時々技術レベルに応じて価値を創出できるテーマに対してアプローチすることを実践してきました。そして、現在はLLMに取り組んでいます。LLMは競争の激しい分野であり私にとっても非常に大きなチャレンジだったのですが、このタイミングでNTTも絶対に取り組むべきだという思いで、プロジェクトを立ち上げました。

さて、機械読解を研究テーマとしているとき、初めてMS MARCOというMicrosoft主催の機械読解を競い合う国際的なコンペティションにチャレンジしてみました。それまで、たくさん

の研究者が参加する主要なテーマで戦うという経験はありませんでしたが、私たちのモデルが当初1位になり、リーダーボードの最上位にNTTの文字が載りました。自分自身もすごく嬉しかったのですが、周りの人もとても喜んでくれたことが印象に残っています。このときの経験から自分たちも世界で戦っていけると思い、その後もいくつかのコンペティションやリーダーボードに参加して、上位の成績を取ることができました。LLMは世界中が注目しており、その進化も日進月歩なので、LLMに取り組むことには勇気が必要でしたが、MS MARCOの成功体験が私の背中を押してくれたと感じます。LLMはビジネスからもとても期待が大きいですし、アカデミックなテーマとしても、LLMのメカニズムの分析から新しいモデルの提案まで可能性を多く秘めているので、大きなやりがいを持って取り組んでいます。

後進の研究者へのメッセージをお願いします。

メインストリームにある分野の研究テーマは、多くの研究者が切磋琢磨しているレッドオーシャンであり、なかなか成果が出にくい環境であるため避ける方もいると思います。一方で、今一番重要なテーマであるからこそ、たくさんの研究者が取り組んでいるともいえます。実際に自分たちでやってみると新しい発見や手ごたえがあり、十分に戦っていける部分が多いと思いますので、若い研究者の方にもぜひチャレンジしていただきたいです。一方、メインストリームではないが新しいテーマに取り組むときは、それをやることで他の研究者が後ろからついてくるような価値のあるテーマになっているかを意識して取り組むのが良いと思います。例えば、若いころはややインクリメンタルであっても重要と思えるテーマに取り組み、より実力が付いたらメインストリーム、レッドオーシャンのテーマであっても覚悟を持って飛び込んでいくということもあるのではないのでしょうか。いずれにしても重要で価値のあるテーマにチャレンジすることが大切と考えます。

LLMについては多くの研究者が取り組む、まさにレッドオーシャンではありますが、一方で歴史の浅く、変化が激しい分野でもあるため、全く追いつけないということはなく、最前線で戦っていける可能性が十分にあるテーマだと考えています。たくさんの方に協力いただきながら進めており、これから良い成果が出せると期待しています。

さて、研究分野やテーマを変えることは勇気があることではあると思いますが、自然言語処理以外にもさまざまな分野・テーマに取り組んできた自分の経験からは、柔軟にいろいろなことにチャレンジしてみて良かったと思います。今重要なこと、今自然にやるべきことをめざすという意識でチャレンジすれば、分野・テーマを変えることに対して、比較的抵抗が少ないのではないかと思います。