生成AIは2024年どのような進化をするか

◆NTTデータは文章検索・回答生成システムと生成AI「tsuzumi」を連携させる

2024年1月16日NTTデータは、自社の文章検索・回答生成システム「LITRON Generative Assistant(リトロン ジェネラティブアシスタント)」に、NTTが開発した大規模言語モデル（LLM)「tsuzumi」を連携した新サービスを4月から提供開始すると発表した。従来から提供している関連文章検索機能に業務データを学習した「tsuzumi」を組み合わせることで、より業務に特化した日本語の回答文章を生成する（図.1)。また、閉域環境で利用できるため会社の重要情報を扱うユースケースにも適用でき、生成AIの利用を拡大することができる。「tsuzumi」は追加学習する場合においても、大規模なマシンリソースが不要となるため、従来に比べ1/20以下にコストを低減することができる。

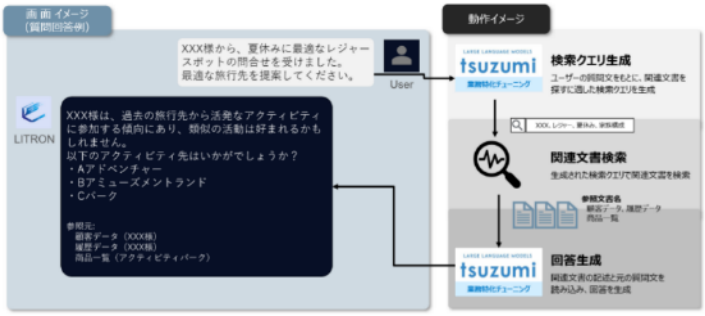

図.1 LITRON Generative Assistant動作イメージ
出典：NTTデータ

◆軽量ながら世界トップレベルの日本語処理性能を持つLLMを開発

NTTが開発した「tsuzumi」は、パラメータ数が60～70億個（OpenAIのChatGPT3.5はパラメータ数が1750億個）と軽量であるため、クラウド提供型LLMの課題である学習やチューニングに必要となるコストを低減できる。また、日本語と英語に対応する「tsuzumi」は、1GPUやCPUでの推論動作を可能にする。日本語性能比較ではChatGPT-3.5と同等の性能を示した。言語や視覚、聴覚のマルチモーダルに対応し、特定の業界や企業組織に特化したチューニングが可能である。

ChatGPTでは1回あたりの学習に要する電力（1300MWh）は原発1基1時間分の発電量以上に相当すると言われている。また、ChatGPTの運用には大規模なGPUクラスタを必要とし、さまざまな業界に特化するためのチューニングや推論にかかるコストが膨大であるため、環境負荷および企業が学習環境を準備するための経済的負担面で課題があった。

言語＋視覚のモーダル拡張により、言語による質問だけでなく、文書画像を提示しながらの質問への回答が可能になる。例では、見積書の画像を入力として与え、それに対し、10%の消費税を抜いた合計金額はいくらと質問すると、9,500円と答える。

図.2 tsuzumi活用事例（言語＋視覚）

出典：NTT

◆Googleは高性能AIモデル「Gemini」を発表

23年12月7日、Googleは、高性能AIモデル「Gemini」を発表した。マルチモーダルとしてゼロから構築された「Gemini」は、テキスト、画像、音声、動画、コードなど、さまざまな種類の情報を同時に受け取り、シームレスかつ自然に処理を行うことができる。「Gemini」はこれまでで最も柔軟なモデルでもあり、データセンターからモバイルデバイスまで、あらゆる場所で効率的に動作する。すなわち、3つのサイズに最適化した「Ultra」「Pro」「Nano」のモデルを提供している。「Gemini Ultra」は、数学、法律、医学など57の科目を組み合わせて知識と問題解決能力をテストするMMLU（大規模マルチタスク言語理解）で90.0%のスコアを実現し、人間の専門家（89.8%）を上回るパフォーマンスを示した。この時点で報告されていたGPT-4のスコアは86.4%であった。テキスト、画像、音声などを同時に認識して理解できるようにトレーニングされているため、ニュアンスを含んだ情報をより理解し、複雑な質問に答えることができる。

「Gemini Nano」はエッジデバイス用の効率的なモデルで、Googleのスマホ、Google Pixel 8 Proで提供される。AI内蔵スマホはGoogle Tensor G3を使用し、レコーダーアプリのサマライズやGboardのスマートリプライで使用可能となる。

レコーダーアプリでは、ネットワーク接続がオフラインであっても、録音された会話、インタビュー、講義、プレゼンテーションなどの要約ができる。Gboardのスマートリプライでは、チャットアプリの使用中で返信候補を提案してくれる機能で、「Gemini Nano」が受信メッセージに基づいて応答を提案してくれるので、入力がとても楽になる。また、スマートフォン上でデータが処理されるので、プライバシーが保護される。

24年の生成AIは、マルチモーダルとエッジデバイスで進化する。　【成田誠】