業務特化型に活路を見出す日本の生成AI

◆富士通はナレッジグラフとLLMを融合し回答精度を上げる

2024年5月17日、富士通は経済産業省が推進する国内生成AIの開発力を強化するためのプロジェクト「GENIAC（Generative AI Accelerator Challenge)」（詳細は後述）の追加公募に採択され、論理推論を可能とする大規模言語モデル（LLM）の研究開発を開始することを発表した。企業における生成AI活用の課題を解決するため、業務に特化した生成AIの提供を目指す。

生成AIを業務活用する場合の課題の一つが、現行の汎用LLMでは計算量やコスト、精度などがオーバースペックであり、対象業務に最適なものがないというものである。この課題についてはNTTの「tsuzumi」やNECの「cotomi」など多くの日本企業がコンパクトなLLMを提供し始めている。

二つ目の課題は、信頼性の問題である。生成AIは、根拠に基づかないもっともらしい誤りを回答してしまう「ハルシネーション」により、信頼性が求められる業務にLLMの導入が進んでいない。ハルシネーション対策としては、外部データベースから関連する情報を取得し回答を生成するRAG(Retrieval Augmented Generation ： 検索拡張生成)が注目されている。しかし、企業が業務で利用している資料には図表が多く、それを理解して対応できるRAGは少ない。

富士通では、業務での質問をLLMに入力する際、自然言語ではなく知識処理技術の一つであるナレッジグラフ形式で入力すると、より業務知識に基づいた回答ができることに着目した。今回のプロジェクトでは、「自然言語文書をナレッジグラフに変換して形式知にするLLM」と「与えられた質問に対してナレッジグラフ上で関連情報を探索し、論理的に集約し回答するLLM」の二つの特化型LLMを開発し、回答精度を高める。信頼性の問題を解決し、24年度中の業務活用を目指す。

◆経済産業省が進める国内生成AI開発強化プロジェクト「GENIAC」

上記の「GENIAC」は、経済産業省が日本での生成AI開発力を強化していくため、24年2月2日にスタートさせたプロジェクトである。生成AIの基盤モデルを開発する上では、計算資源の確保が大きな課題であり、スパコンなどの確保と利用料金

補助という形で支援する。また、開発者同士のネットワークを広め、知見を共有する環境を提供するとともに、生成AIの利活用を促進するための活動も実施する。

スタート当初の採択事業者として、ABEJA、Preferred Elements、東京大学、Sakana AI、ストックマーク、情報・システム研究機構、Turingの5社、2研究機関が2月に採択されており、5月には、ELYZA、Kotoba Technologies Japan、富士通の3社が追加になった。

◆米IT大手企業は生成AIの開発競争が激化している

24年5月14日、OpenAIは生成AI「GPT」の新たなモデル「GPT-4o（GPT-4omni）」を発表した。テキストはもちろん、音声や画像、映像での入力、音声での応答に対応するマルチモーダル機能で、今回は「スピード」や「使い勝手」を改善した。すなわち、モデルの学習方法を変更して、高速化を実現した。これまで独立した3つの処理モデルを組み合わせていたが、GPT-4oでは一つのモデルで画像や映像などの視覚情報、テキスト、音声などを組み合わせて学習するようにした。この結果、応答遅延時間が平均でGPT-3.5では2.8秒、GPT-4では5.4秒だったが、GPT-4oでは平均0.32秒と大幅に短縮することができた。短時間で変換できることで、感情などの情報を加味した自然な会話が可能になる。GPT-4oを使った英語とスペイン語のリアルタイム翻訳の動画で変換速度の実力が分かる。動画理解能力も向上している。専門家向け45分の講義動画をそのままGPT-4oに入力し、数分で正確に要約する能力を示している（select sampleのLecture summarization参照）。

24年5月15日、Googleは年次イベント「Google I/O」で生成AIモデル「Gemini」に、軽量で高速な「Gemini 1.5 Flash」モデルを新たに追加すると発表した。開発者や企業顧客から、長文のテキスト文章を要約、テキストや図表、音声などマルチモーダルな入力で推論しているなどの活用情報を入手した。そのフィードバックから、アプリケーションによっては、少ない応答時間とコストのニーズが高いことが分かり、新たなモデルを開発した。Gemini 1.5 Flashは、軽量なモデルだが、膨大な量の情報に対するマルチモーダル推論能力が非常に高い。また、要約、チャットアプリケーション、画像やビデオのキャプション作成、長いドキュメントや表からのデータ抽出などに優れている。

生成AIの開発競争が激化する中、日本製生成AIは業務特化型になる。【成田誠】