LINE、NAVERと共同で、世界初、日本語に特化した超巨大言語モデルを開発　新規開発不要で、対話や翻訳などさまざまな日本語AIの生成を可能に

2020.11.25　AI関連サービス

従来の特化型言語モデルとは異なる、汎用型言語モデルを実現予定。
処理インフラには世界でも有数の、700ペタフロップス以上の高性能スーパーコンピュータを活用

LINE株式会社（所在地：東京都新宿区、代表取締役社長：出澤剛）はNAVERと共同で、世界でも初めての、日本語に特化した超巨大言語モデル開発と、その処理に必要なインフラ構築についての取り組みを発表いたします。

超巨大言語モデル（膨大なデータから生成された汎用言語モデル）は、AIによる、より自然な言語処理・言語表現を可能にするものです。日本語に特化した超巨大言語モデル開発は、世界でも初めての試みとなります。

従来の言語モデルは、各ユースケース（Q&A、対話、等）に対して、自然言語処理エンジニアが個別に学習する必要がありました（特化型言語モデル）。

一方、汎用言語モデルとは、OpenAIが開発した「GPT」※1や、Google の「T5」※2に代表される言語モデルです。
新聞記事や百科事典、小説、コーディングなどといった膨大な言語データを学習させた言語モデルを構築し、その上でコンテキスト設定を行うためのFew-Shot learning＊1を実行するだけで、さまざまな言語処理（対話、翻訳、入力補完、文書生成、プログラミングコード等）を行うことが可能となり、個々のユースケースを簡単に実現できることが期待されます。
＊1：ブログの書き出しや、プログラミングコードの一部などを与えること。それをもとに、最もそれらしいと判断した文字列を生成します。たとえば、与えた言葉（「おはよう」）に対して、これまで学習した中から最もそれらしいと判断した文字列（「おはようございます」等）を返すといったことが考えられます

今回、日本語に特化した汎用言語モデルを開発するにあたり、1750億以上のパラメーターと、100億ページ以上の日本語データを学習データとして利用予定です。これは現在世界に存在する日本語をベースにした言語モデルのパラメーター量と学習量を大きく超えるものとなります。パラメーター量と学習量については、今後も拡大してまいります。

本取り組みにより、日本語におけるAIの水準が格段に向上し、日本語AIの可能性が大きく広がることが予想されます。

現在、超巨大言語モデルは世界でも英語のみが存在・商用化＊2しており、他言語の開発についても、ごく少数の取り組みが発表されているのみとなります。その理由の一つとして、高度なインフラ環境の必要性があげられます。
超巨大言語モデルの処理には数百ギガバイトものメモリーが必要と考えられており、世界でも指折りの性能を持つスーパーコンピュータなど、高度なインフラ環境が必要です。
＊2 OpenAIが開発し、Microsoftがライセンスを保有する「GPT-3」

今回LINEはNAVERと共同で、当モデルを迅速かつ安全に処理できる700ペタフロップス以上の性能を備えた世界でも有数のスーパーコンピュータを活用し、超巨大言語モデルの土台となるインフラの整備を年内に実現予定です。

英語にて実現している精度に匹敵する、またはそれ以上の、日本語の超巨大言語モデルを創出してまいります。
開発された超巨大言語モデルは、新しい対話AIの開発や検索サービスの品質向上など、AIテクノロジーブランド「LINE CLOVA」をはじめとするLINE社のサービスへの活用のほか、第三者との共同開発や、APIの外部提供についても検討予定です。

※1：OpenAI「GPT(Generative Pre-trained Transformer)」
米国の技術開発会社OpenAIが2019年2月に発表した、文章生成に強い能力を持つ汎用型言語モデルに関する論文。
その後2019年11月には15億のパラメーターをもつ汎用型言語モデル「GPT-2」をリリース。2020年5月に1750億のパラメータを持つ「GPT-3」の構想が発表され、 翌月にベータ版を公開、8月には商用化した。「GPT-3」は「GPT-2」と比較して圧倒的なデータ量を持つことにより、長文の文章生成能力が飛躍的に向上（キーワードからメール文生成や、話し言葉の質問から流暢な回答文を生成する、など）し、世界的に注目を浴びている。

※2：Google 「T5（Text-to-Text Transfer Transformer）」
GPTと同じくトランスフォーマーと呼ばれる自然言語処理技術を用いるが、文章生成よりも翻訳、質疑応答、分類、要約などの文書変換処理を目的とした構成を採用している。入力（タスク）と出力（回答）の両方をテキストのフォーマットに統一して、転移学習を行うことで、全てのタスクを同じモデルで解く。学習データを変更することで、同じモデルでさまざまなタスクが解けるとされる。

【LINE CLOVAとは】
AIテクノロジーブランド「LINE CLOVA」は、「CLOVA Chatbot」「LINE AiCall」などのAI技術やサービスを通して、生活やビジネスに潜む煩わしさを解消すること、社会機能や生活の質を向上させることで、より便利で豊かな世界をもたらしたいと考えています。「ひとにやさしいAI」が自然なカタチで生活やビジネスの一部となるような、「これからのあたりまえ」を創出するべく、引き続きAI技術のさらなる向上や、ビジネスの連携を進めてまいります。
