CyberAgent が開発した国産 LLM（大規模言語モデル）について

2023/6/27

著：藤本

（この PDF は回覧自由です。社内やお知り合いに転送してご利用ください）

様々な番組を提供するサービス ABEMA やウマ娘 プリティダービーなどの大ヒットゲームを多くリリースしている CyberAgent が 2023/5/17 に LLM*、「OpenCALM」を一般公開しました。

*LLM（Large Language Model、大規模言語モデル）とは、膨大な量のテキストデータを学習させることで機械翻訳、質問応答、文章生成、要約、感情分析などを可能にしたアルゴリズムのことです。

有名な LLM として Open AI の GPT-3.5、GPT-4 や Google が開発した PaLM 2 があります。LLM を比較する指標としてパラメータ数があります。現状、このパラメータが大きいほどより複雑なタスクをこなすことができます。GPT-3.5 は 1750 億パラメータですが、CyberAgent が開発した OpenCALM は最大 68 億パラメータです。GPT-3.5 と比べると少ないように思えますが、日本国内ではかなり注目されています。

OpenAI の GPT-3.5 や GPT-4、Google の PaLM 2 などの既存の LLM は英語のテキストデータを中心に学習しているため、英語圏の視点、価値観、事実認識などに偏る出力をする可能性があります。また、GPT-4 Technical Report によると、入力する文章が英語である場合に比べて日本語の場合は正確さに欠けてしまうという結果が出ています。こうした背景から、日本語、日本文化に強い LLM の開発が期待されてきました。

OpenCALM は Wikipedia や Common Crawl*といったオープンな日本語データを学習しています。これにより、日本語、日本文化に特化した LLM が出来上がりました。商用利用も可能になっています。今後より OpenCALM が進歩し、普及していくことで、学術とビジネスの両面で日本の自然言語処理技術*が発展していくことが期待されています。今後の国産 LLM の発展に目が離せません。

* Common Crawl とはデータセットを提供する非営利団体

*自然言語処理技術とは、人間が日常的に使用する言語（自然言語）をコンピュータが理解、解析、生成するための技術です。

参考
https://www.cyberagent.co.jp/news/detail/id=28817

GPT-4 Technical Report（OpenAI） https://cdn.openai.com/papers/gpt-4.pdf

---

ChatGPT ビジネスレポート（picture academy）