PRESS RELEASE

2024 年 4 月 24 日

株式会社 日本経済新聞社

経済情報に特化した大規模言語モデル開発

日本経済新聞社（本社:東京都千代田区、代表取締役社長:長谷部剛）の研究開発部署である日経イノベーション・ラボは、経済情報に特化した大規模言語モデル、NIKKEI Language Model（略称 NiLM、にるむ）を開発しました。インターネット上の公開情報は利用せず約 40 年分の日本経済新聞の記事や日経産業新聞、日経 MJ、日経ヴェリタス、NIKKEI Prime などの専門媒体、日経 BP の各媒体の中から日本経済新聞社グループが著作権や使用権を有する記事のみを使用しました。日本の経済情報に専門特化した最大級の言語モデルとなります。

2024 年 4 月現在、一般公開されているモデルを起点としない独自の事前学習が完了し、最大で 130 億パラメーターのモデルを構築し性能を検証しています。

さらに最大で 700 億パラメーターのモデルをファインチューニング（継続的事前学習や指示学習と呼ばれる手法）した大規模言語モデルの開発も完了し、記事要約や最新ニュースに関する知識など社内の独自タスクにおける性能改善を確認しました。米メタ社の Llama 2 の 700 億パラメーターモデルや、Llama 3 の 80 億パラメーターモデルなどをベースにデータを追加学習させました。

現時点で学習に利用している記事のみでも、Wikipedia や Common Crawl などのデータセットを用いずに、日本語コーパスのトークン量は約 1 兆の規模感に達しました。およそ 150 年に渡って経済情報を世界へ届けてきた日本経済新聞社グループにしかできない最高品質の大規模言語モデルです。

大規模言語モデルは多くのビジネスで応用が始まっており、これまでにない新しい体験を生み出し続けています。しかし汎用的な大規模言語モデルには最新のデータが反映されていないことや、学習データに起因するハルシネーション、様々なメディアのデータが許諾なく使われている可能性が高いという倫理的な課題も含んでいます。日本経済新聞社は報道機関として責任ある立場で生成 AI を利用するために、自らのデータを用い、経済領域への専門特化と、継続的な情報更新が可能なモデルを作成することに焦点を当てた大規模言語モデルの研究開発を続けています。

日経イノベーション・ラボは日本経済新聞の記事を用いて 19 年に BERT モデルを事前学習した後、技術の進展に追従して RoBERTa、GPT-2、T5、DeBERTa といったモデルの事前学習にも取り組んできました。他にも言語モデルの時系列性能劣化や訓練データ抽出、ハルシネーションに関する研究など自然言語処理の分野で多くの実績を残しています。23 年秋に創刊された Minutes by NIKKEI では、編集者が利用するための AI 編集支援ツールを開発しました。AI でアニメーション動画を制作しパリで上映するなど、様々な領域で成果をあげています。

今回開発したモデルは日経イノベーション・ラボが推進しているAIプロダクト群の研究開発などで利用を検討していきます。日本経済新聞社にしか開発することができない経済に専門特化した大規模言語モデルを今後様々な研究開発で利用していく予定です。大規模言語モデルの研究開発は今後も継続し、性能改善や付随する課題に対する検証を進めていきます。

進化を続ける生成AIとどう向き合っていくのか、日本経済新聞社では議論を続けています。ニュースの現場に足を運び、培った経験をもとに分析し、正確な情報を読者に伝えるプロセスを担うのは人間に他なりません。「考え、伝える」メディアとして人の手による責任あるジャーナリズムをこれからも追求していきます。

〈参考〉

石原慧人, 石原祥太郎, 白井穂乃 (2021). BertSumを用いた日本語ニュース記事の抽象型要約手法の検討. 2021年度人工知能学会全国大会（第35回）論文集.
https://www.jstage.jst.go.jp/article/pjsai/JSAI2021/0/JSAI2021_1D4OS3c02/_article/-char/ja
石原祥太郎　(2022). 実践：日本語文章生成　Transformersライブラリで学ぶ実装の守破離, PyCon JP 2022. https://2022.pycon.jp/en/timetable/?id=EEA8FG
Shotaro Ishihara, Hiromu Takahashi, and Hono Shirai (2022). Semantic Shift Stability: Efficient Way to Detect Performance Degradation of Word Embeddings and Pre-trained Language Models. In Proceedings of AACL-IJCNLP 2022. https://aclanthology.org/2022.aacl-main.17/
Shotaro Ishihara (2023). Training Data Extraction From Pre-trained Language Models: A Survey. In Proceedings of Third Workshop on Trustworthy Natural Language Processing.
https://aclanthology.org/2023.trustnlp-1.23/
石原祥太郎 (2023). 事前学習済み言語モデルからの訓練データ抽出：新聞記事の特性を用いた評価セットの構築と分析. 言語処理学会第29回年次大会発表論文集.
https://www.anlp.jp/proceedings/annual_meeting/2023/pdf_dir/Q2-2.pdf
AI編集支援ツールNIKKEI Tailorについて
https://www.nikkei.com/prime/minutes/article/DGXZQOCD1523T0V11C23A1000000

本件に対する問い合わせ
日本経済新聞社 広報室　TEL:(03)3270-0251 (代表)