お知らせ

2024 年 7 月 10 日

モルフォ AI ソリューションズ、
国立情報学研究所から学術論文用の AI-OCR 開発を受託
～新規コーパス開拓･整備を通じて国産の大規模言語モデル（LLM）構築に貢献～

モルフォグループにおいて AI の事業化を担う、株式会社モルフォ AI ソリューションズ（所在地：東京都千代田区、代表取締役：神田武、以下 モルフォ AIS）は、日本語 LLM（Large Language Model：大規模言語モデル）の学習データを生成するための、AI-OCR（Optical Character Recognition：光学文字認識）出力サービスを 2023 年から提供しています。

このたび、大学共同利用機関法人 情報･システム研究機構 国立情報学研究所（所在地：東京都千代田区、所長：黒橋 禎夫、以下 国立情報学研究所）より、日本語学術論文に特化した AI-OCR の開発を受託しましたのでご報告します。当該事業を通じて、国立情報学研究所が推進する日本語に強い国産 LLM の開発に貢献していきます。

■開発の概要

国立情報学研究所は、2024 年 4 月 1 日、文部科学省の「生成 AI モデルの透明性･信頼性の確保に向けた研究開発拠点形成」事業を実施する拠点として、新たに LLM の研究開発を行う「大規模言語モデル研究開発センター（以下、LLM 研究開発センター）」を開設しました（※1）。LLM 研究開発センターは、1750 億パラメータ規模の国産 LLM 構築に向けて、コーパス整備、計算環境整備、評価用ベンチマーク作成などを行うとともに研究開発用の LLM 構築を進めています。

LLM 研究開発センターでは、日本語学術論文 PDF からのテキストデータ抽出を進めています。学術論文 PDF からの本文抽出は、レイアウト（テキストフロー）解析、構造解析（本文領域推定）などの前処理を要します。これらの機能を備えた各種ツールは英語論文を前提にチューニングされているものが

多く、特定の論文誌に限定されない汎用かつ実用的に日本語論文の本文抽出が可能なものを用意する必要がありました。

モルフォ AIS は、LLM 研究開発センターからの委託事業として、日本語学術論文に特有のレイアウトの認識や、本文領域のテキスト抽出を可能とする AI-OCR の機能開発を行います。これにより、国産 LLM 構築のために必要となる良質かつ大量の日本語のテキストデータの生成に貢献していきます。

■モルフォ AIS が提供する OCR テキスト出力サービス

画像として保存された文書のデジタル化のためには OCR が必要となりますが、市販 OCR の多くは請求書や領収書といった「帳票向け」に開発されたものです。日本語の文書は多様なレイアウト（縦書き、横書き、多段組等）、多様な文字種が混在するため、市販の OCR では読み順を含めた正確な日本語の抽出が難しいという課題があります。

モルフォ AIS の提供する AI-OCR 出力サービスは、上記の市販 OCR が苦手としている文章の読み順まで含めた高精度のテキスト生成を行います。これによって、組織が保有するスキャン画像データから多様かつ正確な日本語を生成することで、日本語 LLM の学習データの作成を支援します。

■サービス内容、特徴、実績

＜サービス内容＞

既存文書（社史・広報誌・公文書・議事録等）のデジタル化と LLM 学習データへの変換

＜特徴＞

①文書に対応した AI-OCR

- LLM に入力する際に重要な読み順まで再現
- 文字種は約 7000 種類で、複雑な漢字も読み取り可能

②画像（JPEG,PDF,PNG 等）が含まれている雑多な文書を、テキスト（様々なフォーマット）で出力可能

＜実績＞

様々な機関向けにテキスト生成を実施済み

（沖縄県豊見城市/ボローニャ大学/順天堂大学/滋賀県立図書館 等多数）

■お申込み・問い合わせ窓口

https://frog-ai-ocr.morphoai.com/

こちらより無償トライアルしていただくことが可能です

■FROG AI-OCR 紹介

FROG AI-OCR は、お手軽に OCR 適用業務が行えるよう NDLOCR の高精度な OCR 処理に加えて、校正・テキスト出力機能も 1 つのパッケージとしてご提供しております。機能は全てクラウドで利用可能で、出力テキストの確認・修正作業を効率良く行うことが可能となります。FROG AI-OCR は、国立国会図書館が CC BY のライセンスで公開している NDLOCR（https://github.com/ndl-lab/ndlocr_cli）をコアエンジンとして利用しています。

■注釈

※1：国立情報学研究所に「大規模言語モデル研究開発センター」新設

～国産 LLM を構築し、生成 AI モデルの透明性・信頼性を確保する研究開発を加速～

https://www.nii.ac.jp/news/release/2024/0401.html

■関連プレスリリース

2022 年 6 月 14 日

世界初（注 1）近代書籍対応の市販 AI-OCR ソフト「FROG AI-OCR」新発売

～デジタルアーカイブ事業・読書バリアフリー法対応に最適、図書館 OCR の決定版～

https://www.morphoinc.com/news/20220614-jpr-mais_frog_aiocr

2023 年 12 月 19 日
モルフォ AI ソリューションズ、LLM 向けの日本語データセット生成サービスを提供開始
～文書画像から AI-OCR でテキストデータ作成、良質な日本語 LLM 構築に貢献～
https://www.morphoinc.com/news/20231219-jpr-mais_ocr

【モルフォ AI ソリューションズについて】
モルフォ AI ソリューションズは、AI（人工知能）の事業化に取り組む企業です。
行政、電力、交通、製造といった社会インフラの領域で、AI-OCR、AI-カメラをはじめとする最先端の AI 技術の導入と実運用を推進しております。

所在地：東京都千代田区神田錦町 2-2-1 KANDA SQUARE 1 0 階
代表者：代表取締役　神田　武
設立：2019 年 12 月
事業内容：AI コンサルティング、システムインテグレーション、SW・HW 販売など
ホームページ：https://www.morphoai.com
X（旧 Twitter）：https://twitter.com/MorphoAIS
FROG AI-OCR：https://frog-ai-ocr.morphoai.com/

■お問合せ先
株式会社モルフォ AI ソリューションズ　神田
メール：contact@morphoai.com