プレスリリース

2023 年 12 月 19 日

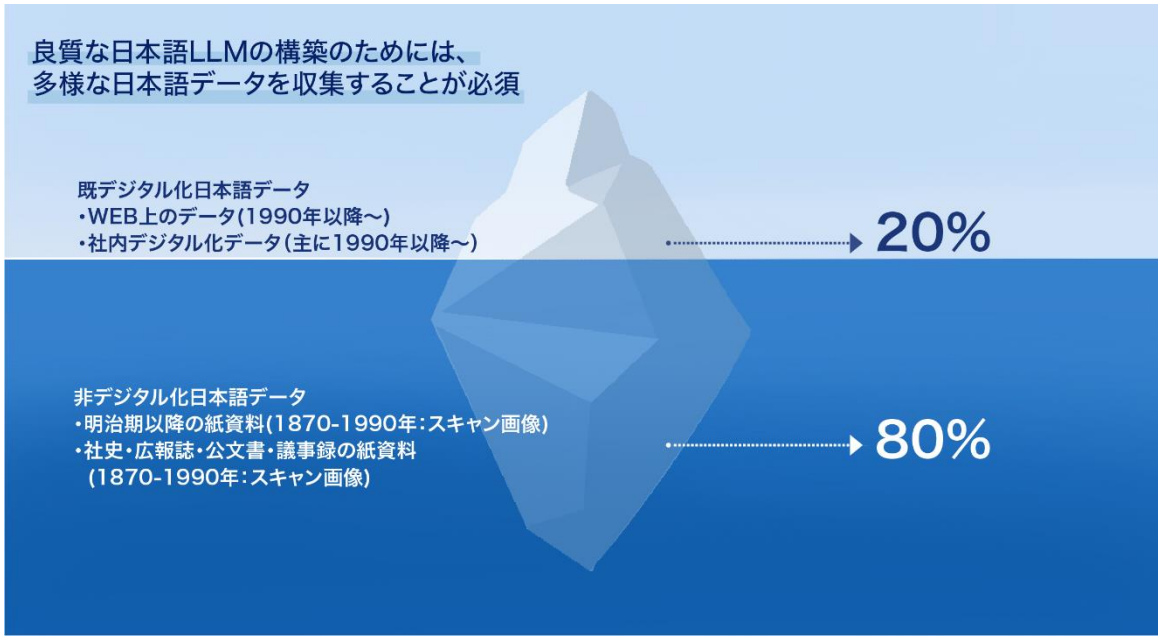
モルフォ AI ソリューションズ、
LLM 向けの日本語データセット生成サービスを提供開始
～文書画像から AI-OCR でテキストデータ作成、良質な日本語 LLM 構築に貢献～

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）出力サービスの提供を 2023 年 12 月 19 日（火）より開始することをお知らせします。

このサービスは、独自 LLM の構築を検討されている組織（企業・官公庁・地方自治体等）や LLM 開発を進める AI 企業・研究機関向けに正確で多様な日本語テキストデータを提供します。

【LLM 構築における日本語学習データの多様性の欠如

良質な日本語 LLM の構築のためには、多様な日本語データを収集することが必須です。しかし、一般的に収集可能な日本語テキストデータは主にインターネット普及後の 1990 年以降のものが中心です。1990 年以前の文書（社史・広報誌・公文書・議事録等の保存文書）の多くはデジタルデータ化されておらず、効率的に収集することができませんでした。そのため、LLM 構築を進める組織の多くは、多様な日本語学習データが用意できず公開された共通のデータセットを活用せざるを得ませんでした。結果として、良質な LLM の構築に制約がかかっています。



【日本語文書に対応した AI-OCR の重要性】

保存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 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帳票ではなく、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＞

国立国会図書館をはじめとして、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/大手新聞社　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頂く事が可能です。

【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、国立国会図書館の 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。

【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

2022 年 11 月 16 日
モルフォ AI ソリューションズ「FROG AI-OCR」を 滋賀県立図書館に提供開始
～郷土資料のデジタルアーカイブ化、視覚障害者向けの書籍テキスト化への活用目指す～
https://www.morphoinc.com/news/20221116-jpr-mais_frog_aiocr

2023 年 2 月 14 日
モルフォ AI ソリューションズ、順天堂大学と「FROG AI-OCR」を活用した武道史研究を開始
～近代史料の計量テキスト分析を実施、新たな武道史研究の視点を提供可能に～
https://www.morphoinc.com/news/20230214-jpr-mais_frog_aiocr

2023 年 7 月 27 日
モルフォ AI ソリューションズ、EU 資金の研究プログラム NONWESTLIT に参画
～伊ボローニャ大学とともに 19 世紀日本語文書の解析を実施～
https://www.morphoinc.com/news/20230727-jpr-mais_frog_aiocr

2023 年 10 月 3 日
モルフォ AI ソリューションズ、 沖縄県豊見城市に「FROG AI-OCR」を導入開始
～自治体 DX として沖縄タイムス系列 Nansei 社と公文書等のデジタル化を支援～
https://www.morphoinc.com/news/20231003-jpr-mais_frog_ai-ocr

【株式会社モルフォ 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 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：〒101-0054 東京都千代田区神田錦町 2－2－1 神田スクエア 11F
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

【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