HITACHI
Inspire the Next

News Release

2021 年 6 月 23 日
株式会社日立製作所

企業内の「ダークデータ」に着目した「データ抽出ソリューション」を提供開始

米国発の AI を活用し、非定型ドキュメントからのデータの効率的な抽出と有効活用を支援

ソリューション概要図

株式会社日立製作所(以下、日立)は、このたび、「ダークデータ」*1 と呼ばれる、日々の企業活動で生成・蓄積されるものの有効活用できていない膨大なデータに光をあて、新たな価値を見いだす「データ抽出ソリューション」を開発し、6 月 23 日より販売を開始します。

本ソリューションは、日立が参画する米国スタンフォード大学の企業参画プログラムで開発された AI を中核としたダークデータ分析エンジンを活用し、請求書や診療明細書といった発行元によって様式や表記が異なる非定型ドキュメントの利活用において、取得したいデータの抽出作業を自動化・高度化するものです。一般的な OCR や AI-OCR*2 では解析が難しい多種多様なドキュメントに対応し、日々蓄積する膨大なダークデータの中から、価値あるデータを導き出し、データ利活用による経営判断の迅速化やビジネスの変革に貢献します。

昨今、IoT の進展により、社会や企業活動から生み出されるデータ量が加速度的に増え続ける中、組織の持つデータのうち、その多くが活用されないまま眠るダークデータであると言われています。日々蓄積されるデータを、意味のある情報として整理し分析することで新しい価値の創出につなげることが、今後のビジネス拡大に重要となっています。

こうした中、近年、ドキュメントデータについては、AI を活用した OCR 技術の進化により、フォーマットが定型または準定型の帳票について高精度な読み取りや情報抽出が可能となってきました。一方で、請求書や診療明細書、有価証券報告書など非定型ドキュメントは、発行元ごとに表記や様式が異なるため、読み取り・抽出の自動化が困難なケースが多く、課題となっています。

日立は、2016 年より、スタンフォード大学工学系研究科が主催するデータサイエンス分野におけるプログラム「Stanford Data Science Initiative」に参画し、ダークデータの効率的な活用に向けた研究を行い、その成果をオープンソースとして公開してきました。また、日立のデータサイエンティ

ストのトップ人財が集まる「Lumada Data Science Lab.」*3 にて、ダークデータに関する専任の研究開発チームを設置するなど、最新の研究やテクノロジーを活用したソリューション開発に取り組んできました。

「データ抽出ソリューション」は、日立が、これまでのダークデータに関する研究成果をもとに開発した、新しい AI ソリューションです。人が文書を読む際に、テキストだけでなく、全体のレイアウトや単語の出現位置など視覚的な情報から文書を捉えるように、AI が、表や図、テキストの座標といったドキュメント内のさまざまな特徴から文書の構造全体を解析し、非定型の多種多様なドキュメントのデータ抽出に対応します。また、少ない教師データから AI モデルを生成できる自動ラベリング機能により、導入時のモデル構築や、追加学習・再学習といったモデルの改修にも柔軟に対応できます。

これにより、企業内で蓄積する膨大なデータの中から、効率的に価値あるデータを見つけ出し、生産性向上や販売力強化、コスト削減といった企業に内在するさまざまな経営課題の解決に向けて迅速なデータ利活用を支援します。

なお、本ソリューションは、日立の専門エンジニアが、お客さまの業務で扱うドキュメントに適したモデルの構築を行うなど、業務内容に応じた最適な導入・運用のコンサルティングを行います。また、他システムとのシームレスなデータ連携を可能にする API により、既存の OCR システムや業務システムとの連携を効率化します。

今後、日立は、画像や映像、音声といった、企業が保有するダークデータ全般に対応するソリューションの実現に向け、AI の抽出機能をさらに強化し、日立の Lumada*4 ソリューションの一つとして、社会や企業におけるデータ利活用による新たな価値の創出や課題解決を支援していきます。

*1 ダークデータ：企業内で日々収集・蓄積されていくデータのうち、活用されていないデータ、または活用されているものの手間がかかり活用効率が悪いデータ。
*2 AI-OCR：AI(人工知能)技術を取り入れた光学文字認識機能(OCR)。定型あるいは準定型の帳票に記載された項目の自動抽出が可能となるため、伝票などの入力作業を効率化でき、生産性向上を支援する。
*3 ニュースリリース(2020 年 3 月 30 日発表)「データサイエンティストのトップ人財を結集した「Lumada Data Science Lab.」を設立」 https://www.hitachi.co.jp/New/cnews/month/2020/03/0330d.html
*4 Lumada：お客さまのデータから価値を創出し、デジタルイノベーションを加速するための、日立の先進的なデジタル技術を活用したソリューション・サービス・テクノロジーの総称。

■本ソリューションの特長

1. 高度な解析技術を活用し、非定型のさまざまなドキュメントからのデータ抽出を効率化

表やページ情報などドキュメント内のさまざまな視覚情報を特徴として捉え、文書を解析する「情報表現構造解析技術」により、対応が難しかった非定型ドキュメントのデータ抽出を可能にします。

たとえば、日付の表記が「発行日」と「診察日」など、発行元によって異なる用語が使われている場合にも、文書の構造から同じ意味をさす単語として認識できるほか、抽出対象が複数ページにまたがるドキュメントでも、対象となる項目を抽出することが可能です。また、一つの区分に対し複数の項目が紐づく 1:N の関係も正しく認識するため、複雑な表のデータ抽出にも適します。

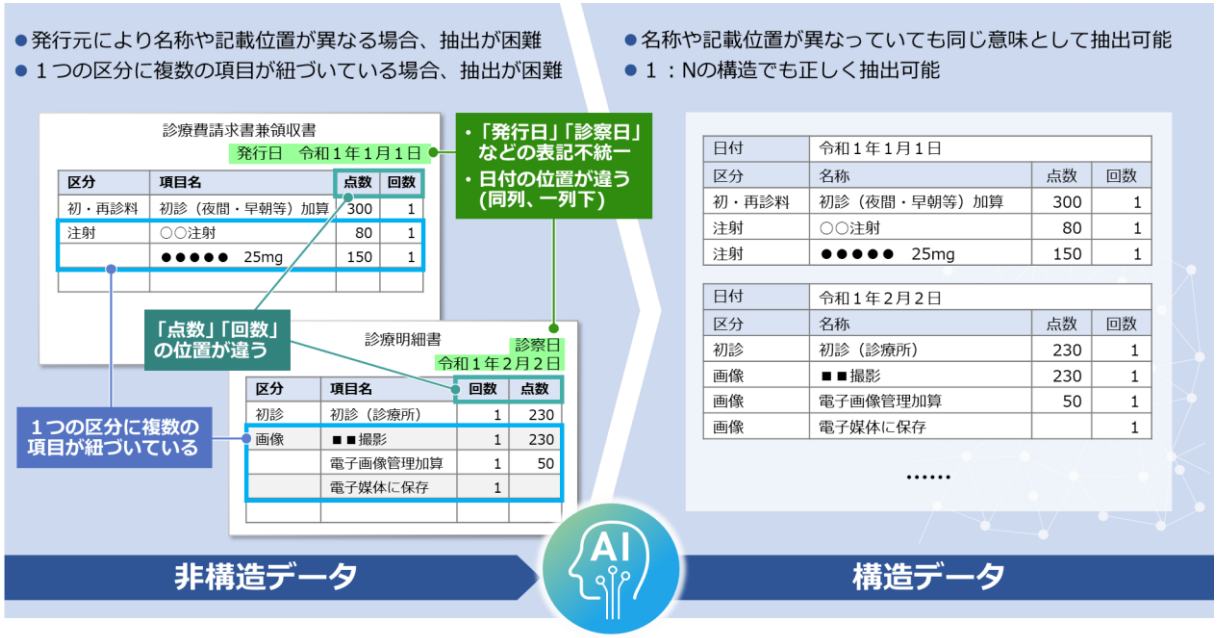
これにより、人手でかかる抽出後のデータ処理時間を削減し、得られたデータを迅速に業務改革に活用するなど、より高度な業務にリソースを充てることが可能となります。

2.少ないデータからAIモデルを構築する技術により、導入時・改定時の作業負荷や期間を削減

従来の AI 技術は、モデル構築にあたって大量の学習データを準備し、人手でデータの指定作業(ラベリング)を行うことが一般的であるため、モデルの構築や精度の維持・運用に多大なコストと時間を要します。本ソリューションでは、少ない学習データで AI モデルを生成できる「弱教師学習技術」により、データのラベリング作業を自動化するため、モデル構築のための期間短縮やコスト削減が可能となるほか、追加学習や再学習といったモデルの継続的な改善にも柔軟に対応できます。

用意する学習用データを削減できることで作業負荷を軽減するため、導入時だけでなく、法改正や商品改定にも迅速に対応でき、運用の効率化に寄与します。

■診療明細書を使ったデータ抽出のイメージ



■本ソリューションの価格および提供開始時期

名称	価格	提供開始日
ダークデータに着目した 「データ抽出ソリューション」	個別見積	6 月 23 日

■「データ抽出ソリューション」に関する Web サイト

https://www.hitachi.co.jp/finance/solutions/application/common/Data-Extraction/

■日立の金融ソリューションに関する Web サイト

https://www.hitachi.co.jp/products/it/finance/

■日立製作所について

日立は、IT(Information Technology)、OT(Operational Technology)およびプロダクトを組み合わせた社会イノベーション事業に注力しています。2020 年度(2021 年 3 月期)の連結売上収益は 8 兆 7,291 億円、2021 年 3 月末時点で連結子会社は 871 社、全世界で約 35 万人の従業員を擁しています。日立の先進的なデジタル技術を活用したソリューション／サービス／テクノロジーである Lumada を通じて、IT、エネルギー、インダストリー、モビリティ、ライフ、オートモティブシステムの 6 分野でお客さまのデータから価値を創出し、デジタルイノベーションを加速することで、社会価値・環境価値・経済価値の 3 つの価値向上に貢献します。

詳しくは、日立のウェブサイト(https://www.hitachi.co.jp/)をご覧ください。

■本件に関するお問い合わせ先

株式会社日立製作所 金融システム営業統括本部 [担当:松浦]
〒100-8220 東京都千代田区丸の内一丁目 6 番 1 号
お問い合わせフォーム:https://www.hitachi.co.jp/finance-inq/

以上