企業の中に眠っていた宝の山から、新たな価値を効率よく発掘する「データ抽出ソリューション」

事業活動を通じて企業には膨大なデータが日々生まれ、蓄積されています。その中に眠る非構造データを先進の人工知能（AI）技術で構造データに変換し、迅速な経営判断やビジネス変革のために活用するためのソリューションが新登場。社会や企業で進むデータ利活用による新たな価値創出をさらに加速させます。

日の目を見ない「ダークデータ」に光を当てる

IoTの進展やデジタルトランスフォーメーション（DX）への機運の高まりを背景に、企業によるデータ利活用が活発化しています。しかし現在、日々蓄えられる膨大なデータのうち、企業が利活用しているのは、実はごく限られた“氷山の一角”。その水面下では日の目を見ることなく大量のデータが眠っているのが実情です。

活用されずにこうして蓄積されたままになっているデータのことを「ダークデータ」と呼びますが、その中には有用なデータがたくさんあります。それにもかかわらず、それらが放置されてきたのは、その多くが非定型ドキュメントだからです。

近年ではフォーマットが定型・準定型のドキュメントであれば、AI-OCRなどによって高精度なデータの読み取り・抽出もできるようになりました。しかし、発行元ごとに表記や様式がまちまちな請求書や診療明細書といった非定型ドキュメントについては、依然としてその読み取りや抽出が非常に困難だったのです。

そこで日立は、非定型ドキュメントなどに含まれる非構造データも先進のAI技術によって利用可能な構造データに変換できる「データ抽出ソリューション」を開発。従来のAI-OCRでも抽出が困難だった非定型ドキュメントからの効率的なデータ抽出を実現します（図1）。

ダークデータの効率的な活用を実現する2つの技術

日立は、2016年から米国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」に参画しています。また、データサイエンスのトップ人財を集めた「Lumada Data Science Lab.」にはダークデータの専任研究開発チームを設置するなど、最新の研究や技術による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にはこうした先進的な研究成果を基に開発した「データ構造解析」そして、「AI学習」に関する2つの高度な技術が投入されています。

1つ目が、ドキュメント内の表や図、ページ情報、テキストの座標情報といった視覚情報を特徴として認識し、文書を解析する「情報表現構造解析技術」です。従来、情報表現の構造がバラバラなドキュメントから情報を抽出するのは困難でしたが、この技術により非定型ドキュメントからの情報抽出を実現します（図2）。

そして2つ目が、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」です。大量の学習データを基に人手でデータを指定する従来の手動ラベリングには多大なコストを要していました。そこで、データのラベリング作業を自動化することにより、モデル構築を短期間かつ低コストで実現できるほか、

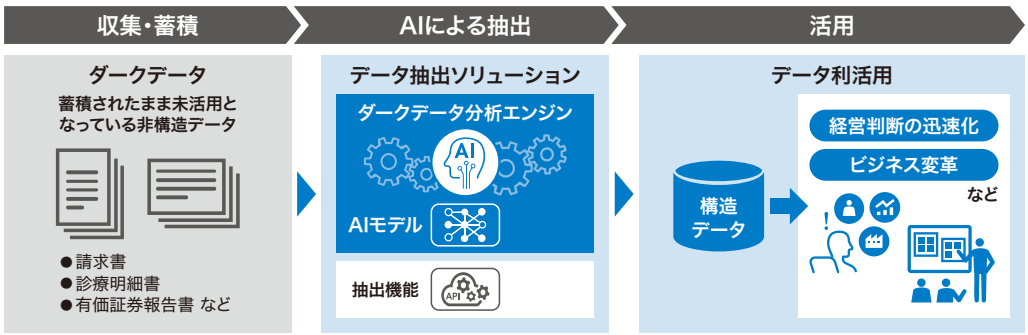

図1「データ抽出ソリューション」の概要

法改正や商品の改定などにともなう追加学習・再学習といったAIモデルの継続的な改善にも柔軟に対応。システム導入時だけでなく、業務稼働後の効率的な運用にも貢献します。

業務を刷新する効率化とコスト削減効果

データ抽出ソリューションでは、日立の専門エンジニアによる業務内容や対象ドキュメントなどに応じたモデル構築や導入・運用コンサルティングを提供。また、他システムとのシームレスなデータ連携を実現するAPIを介して、既存のOCRシステムや業務システムなどとのスムーズな連携も可能です。

高度な先進技術の活用に加え、こうした優位性でも企業のさらなるデータ利活用を促進する本ソリューションを導入することで、業務効率の改善と運用コストの削減が期待できます。運用中の業務に適用することで処理時間を短縮でき、ドキュメント読み取り・確認作業のスループットも向上するほか、情報を抽出する手作業の削減により人件費の抑制にも貢献します。また、効率的な業務運用による顧客体験（カスタマーエクスペリエンス）の向上も期待できます。

残されたダークデータの利活用でさらなる価値創出へ

さらに今後は、抽出処理に関するさらなる機能強化を図りながら、抽出されたデータの分析を担うAIや各種ソリューションとの連携によって、データ利活用のより一層の効果性・有用性をめざす考えです。

データ抽出ソリューションが対象とするドキュメント以外にも、企業の内部には画像や映像、音声など、さまざまな非構造データが蓄積されています。そしてそれらの大部分は利活用されないまま放置されたダークデータです。日立はAIのさらなる強化などによって、こうした膨大なダークデータからの価値ある情報の抽出を促進し、これからも社会や企業の新たな価値創出と課題解決を支援していきます。

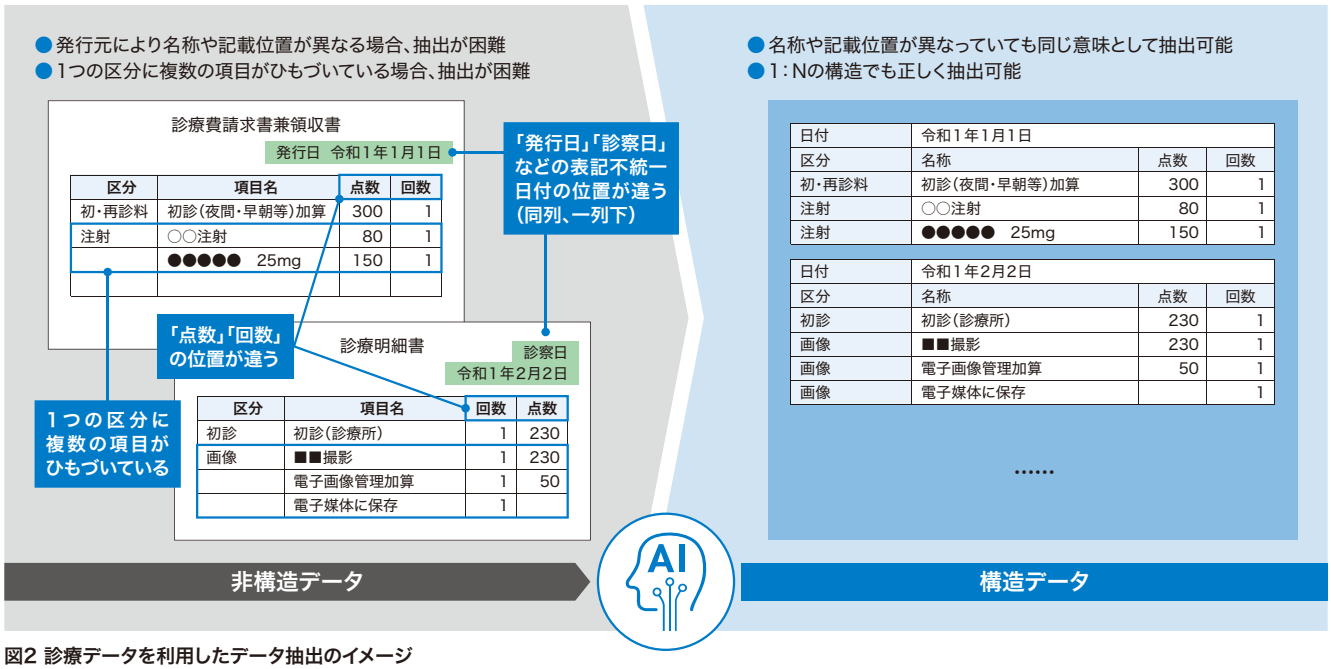

図2 診療データを利用したデータ抽出のイメージ

お問い合わせ先・情報提供サイト

(株) 日立製作所 金融システム営業統括本部
https://www.hitachi.co.jp/Data-Extraction/