ホンヤク出版社のソリューション

動かない AI を生み出さないため、不揃いなデータをデータ整形（データクレンジング）で統一しませんか？

「伝える」をデザインする

株式会社ホンヤク出版社

動かない AI を生み出さないため、不揃いなデータを**データ整形**（データクレンジング）**で統一しませんか？**

弊社の翻訳業務で培ったデータ前処理技術を AI データクレンジングに利用すれば、 不揃いなデータを効率的に統一できます。その結果、AI にかける前のデータの前処理作業に費やしていた膨大な時間を大幅に短縮し、本来の分析に専念できます。

課題

翻訳作業の前には、必ず、翻訳前処理作業を行います。翻訳前の原文は、形式もバラバラで、使われている用語も不統一な場合があります。

場合によっては、オリジナルのアプリケーションファイルがなく、PDF だけしかなかったり、PDF になっていたとしても、文字データを抽出できない画像化されたデータだったりと状態は様々です。

また、データ変換が必要な場合でも、PDF をテキスト変換すると不要な改行や空白が入ってしまったりして、それを取り除く作業が発生します。

しかし、この前処理作業を怠ると、翻訳作業の後工程に悪影響が出て、結局、二度手間ややり直しによるムダな作業が発生し、納期遅延やコスト増大につながってしまいます。

逆に前処理作業を行った後では、スムーズに業務が進み、納期短縮、コスト削減、クオリティ維持につながります。

AI のデータ活用でも状況は同様です。

AI にかける前のデータ形式がテキストデータ、エクセル（Excel）、PDF、Word、XML、JSON などバラバラな場合、形式の統一、標準化などの AI 学習用のデータの加工は自動化が難しいため、どうしても多くの時間をとられてしまいます。この前工程作業が本来の分析作業の足かせとなっています。

解決策

AI 学習用データの前処理工程をできる限り自動化することで本来の業務に専念できます。

その１：データ形式が不ぞろいな場合のデータ形式統一作業の自動化

例：
- Word→テキストへの変換
- PDF→テキストへの変換
- スキャニングされたデータ→テキストへの変換
- その他、様々な変換に対応しています。

その２：データ変換時に不要な改行や空白などを取り除く作業の自動化

その３：大量文書（紙または画像化された PDF しかない場合）の最適な OCR 処理とチェック

その４：非構造化データ（Word、PDF 等）の構造化データ（XML）への変換

まとめ

AI 学習用データの前処理工程をできる限り自動化することで、自動変換による時間短縮とコスト削減、統一・標準化によるクオリティ維持をはかれます。

その結果、間違った翻訳、動かない AI の発生など、起こり得る不具合を未然に防ぐ、予防対策につながります。

前処理工程は、一番大切な翻訳の精度、AI の精度の根源をカバーする重要な役割を果たします。弊社には、この前処理工程でのお客様のお悩みを解決するためのノウハウがあります。

（2019 年 7 月 1 日）

ホンヤク出版社のソリューション一覧

- Web を簡単に多言語化する方法
- Word だって自動組版できるんです
- 改訂履歴の見られる電子マニュアルはいかがですか？
- Illustrator 多言語版の時短テクニック
- 動かない AI を生み出さないため、
  不揃いなデータをデータ整形（データクレンジング）で統一しませんか？
- 翻訳前に「改行」や「空白」などの制御文字を事前に取り除く方法とは？
- 用語集を自動作成! 同時に表記揺れも検出!
- 翻訳資産管理
- マニュアル診断から、その先へ
- 英訳する日本語原稿は何に気をつけて書く？
- 翻訳費用削減事例
- Microsoft Access を使ったデータ管理

上記に類似したお悩みごとはございませんか？
弊社がおカになれるかもしれません。
各種ソリューションは弊社の Web サイトからご覧いただけます。
是非ご確認くださいませ！

QR コード読み取りまたは
「ホンヤク出版社」で検索

ホンヤク出版社　検索

ページ上方のメニューから
「ソリューション」にアクセス

「ソリューション」ページ URL：https://www.honyaku-shuppan-sha.co.jp/solution.html

お問い合わせ

03-3355-4411 (9:00～17:30)
Web サイトからもお問い合わせいただけます。
トップページの「お問い合わせ」ボタンからフォームに入力してご送信ください。