

AI言語モデルを活用した非定型帳票認識技術

Application of AI Language Models to Semi-Structured Document OCR

小野 聡一郎 ONO Soichiro 馮 思萌 FUU Shimou 古畑 彰夫 FURUHATA Akio

人手不足解消や作業効率化の一環として、光学的文字認識（OCR：Optical Character Recognition）の分野では、様々な様式の帳票からデータを自動抽出する非定型帳票認識の精度向上が求められている。近年目覚ましく発展しているAIをOCRに活用する動きがあるが、事前の学習や人による作り込みなどが必要で、手間やコストが掛かる問題があった。

東芝デジタルソリューションズ（株）は、提供している“AI OCR文字認識サービス”にAI言語モデルを組み合わせた非定型帳票認識技術を開発している。文字列認識結果の部分文字列に対する属性推定技術を導入し評価試験を行った結果、事前準備の手間を削減しながら、従来技術と同等の抽出精度が得られた。また、生成AIを活用して、ユーザーがカスタマイズしやすくする技術も開発している。

Various businesses use optical character recognition (OCR) systems, and improving accuracy of extracting information from various business forms, such as semi-structured forms, is increasingly required to deal with labor shortages and to streamline work. The trend of applying the latest artificial intelligence (AI) technologies in OCR systems has accelerated recently, however, the higher development costs for finetuning processes of AI models are serious issues.

Toshiba Digital Solutions Corporation is developing a technology to recognize semi-structured forms by adopting AI language models to its existing AI-OCR service. Our experiments on entity extraction of substrings from the results of recognized strings have confirmed that it achieves extraction accuracy comparable to that of existing methods while saving on advance preparations. In the future, we will consider applying generative AI to allow users to easily customize this technology.

1. まえがき

近年、様々な分野でAI技術が目覚ましい発展を見せている。これまでは、画像認識の分野への活用が主であったが、2022年秋頃から関心が高まっている生成AIに代表される自然言語処理（NLP：Natural Language Processing）の分野にも広がりつつある。OCRの分野は、画像認識と言語処理の技術の融合領域であり、AI技術の製品展開が進んでいる。従来は主に、乱雑な手書きや、取り消し・訂正された記入文字といった難読文字を認識する技術が発展していたが、最近では言語処理技術を活用したサービスが急速に増えている。

その代表例が、非定型帳票認識の分野である。これまでの帳票認識は、帳票内の読み取るべき位置、及びそこに書かれている情報を事前知識として与え、それらに基づいて認識する定型帳票認識が主流であった。この方式は、帳票フォーマットのそれぞれに対してユーザーが事前定義を行う必要があり、特に請求書などは発行元ごとに多様なデザインを持つため、利用前に掛かるコストが高いという問題があった。

非定型帳票認識技術は、これを解決するもので、項目に付与される見出しや書かれ方の一般的な規則を基に、重要な情報を帳票の中から自動的に見つけ出して出力する。これまでの帳票認識と異なり、事前に読み取るべき位置の指定を行う必要がないため、導入時のユーザー作業コストが低いことが特徴である。東芝デジタルソリューションズ（株）は、非定型帳票認識技術を適用したAI OCR文字認識サービスを提供している。しかし、非定型帳票認識は、事前に大量のデータを用意しての学習や人力での作り込みなどを行う必要があり多様な帳票のデザインを網羅するような学習データの準備や抽出ルールの定式化は難しかった。

そこで当社は、NLPの技術に注目し、AI言語モデルを活用した非定型帳票認識技術を開発した。この技術では、AI言語モデルによる属性推定を導入し、それを簡略な抽出ルールと組み合わせることで、表現の多様性に対する受容度を高め、十分に多様なデータがなくても一定の認識精度を達成可能にした。

ここでは、非定型帳票認識分野の技術動向、実用上の課題、及びAI言語モデルを活用した非定型帳票文字認識技術の概要・評価結果について述べる。

2. 非定型帳票認識とその技術動向

非定型帳票認識技術は、事前に読み取るべき文字の位置や属性を与えず、項目に付与される見出しや書かれ方の一般的な規則に基づき、必要な項目を帳票の中から自動的に検出する。抽出する情報の特定には、文字列の言語的特徴や文字列に付与された見出し、項目間の位置や演算関係、例えば“税抜き額+税抜き額×税率=税込み額”などを用いる。文字行の持つ意味（住所や氏名などであり、属性やエンティティーと呼ぶことが多い）の判定には言語的特徴を扱うための深層学習技術であるBERT (Bidirectional Encoder Representations from Transformers⁽¹⁾)などの言語モデルをファインチューニングして用いることが多く、4.1節で述べる技術も同様である。

最近では、文字認識の結果(テキスト)と画像構造をTransformer⁽²⁾構造によって統合して重要な項目の抽出を行う技術であるLayoutLM⁽³⁾など、クロスモーダルな深層学習技術に基づく製品も出てきている。

更に、生成AIとOCRとを組み合わせた非定型帳票認識サービスや、汎用の画像認識サービスなどを併用した非定型帳票認識試作などの報告^{(4), (5)}がある。この方法では、AI自体のファインチューニングを行わず、質問を制御するプロンプトエンジニアリングによって機能を実現する。大規模言語モデルの持つ汎用的な知識を活用して低コストで機能開発できる点に特長があるが、いまだ黎明(れいめい)期にある技術といえる。

3. 非定型帳票認識の種類と課題

ここでは、AI OCR文字認識サービスを例として、非定型帳票認識の主な種類とそれぞれの課題について述べる。

3.1 プリセット型

プリセット型は、サービス提供者が事前に読み取り項目の指定を行うものであり、他社も含めて非定型帳票認識技術の大半はこれに属する。これは、請求書や見積書などあらかじめ定められた種類の帳票に対して、請求元会社名や請求金額などそれぞれ所定の項目に対する記載内容を出力するものである。入力画像と項目抽出結果の例を図1に示す。

AI技術の進展がこの分野に広がる以前、この機能はルールに基づく手法で実現されていた。すなわち、書面内の文字列を全て認識した上で、文字列それぞれに対し以下のような属性情報に沿って条件判定を行い、それに合致するものを所定の項目に対する記載内容として出力するという手法である。

(1) 記載位置

請求書

請求日: 令和5年8月30日

NO. 123456789

〒100-0001 東京都千代田区千代田 1-2-3
ABC株式会社
支店 花子様

振込先 カ/AAA
XYZ銀行 本店 普通 7654321
支払期限 令和5年7月31日
振替コード No.2008-1234

※イベントサービス株式会社
登録番号: T1234567890123

〒100-0001 東京都千代田区千代田 1-2-3
TEL: xxx-xxxx-xxxx
FAX: xxx-xxxx-xxxx

イベントサービス

件名 製品説明会の開催

品名	詳細	数量	単価	金額	税率	課税額
パンフレット印刷費用	A4ページ×100部	100	700	70,000	10%	7,000
会場費	ABCホテルモリタ一室	1	18,000	18,000	10%	1,800
MC出演費	株式会社XYZエンターテインメントセンター	1	18,000	18,000	10%	1,800
材料費	カフェの美味しいお茶	100	150	15,000	0%	0
【税率】	10%対象			158,180	消費税	14,380
10%対象				18,200	消費税	1,300
合計				158,000		15,500
消費税合計						15,500
ご請求金額						174,380

① 請求書番号 ② 請求書発行日 ③ 銀行口座 ④ 本体金額
⑤ 消費税 ⑥ 税込み価格 ⑦ 請求書宛先名 ⑧ 請求元会社名
⑨ 税率別対象額 ⑩ 税率別課税額 ⑪ 登録番号

図1. プリセット型非定型帳票認識の例

特定類型（この場合は請求書）の帳票画像を認識し、開発時に設定した項目に該当する文字列を抽出する。

Example of recognition results of semi-structured document with preset invoice mode

(2) 書式

(3) 内容の所定単語リストへの合致／非合致

(4) 数値の場合、ほかの項目との算術関係

例えば請求書において、上方にある文字列で、法人名リストに合致するものは請求元又は請求先の会社名である可能性が高い。また、日付や金額は“令和5年6月7日”、“¥123,456”など所定の書式に従っていることが多く、金額の場合であればその値から税抜き価格・税額・税込み価格の候補をそれぞれ絞り込むことが可能である。

近年は、AIに基づく手法が発展している。これは、抽出対象とする項目名に対応する記載内容をあらかじめ正解として教示した帳票データを大量に用意し、これを学習させて記載内容を自動的に推定するものである。

いずれの手法も、開発コストを抑えることが課題である。ルールに基づく手法は人力で絞り込み、ロジックを考案実装する必要があるため、対応する帳票類型や項目が増加するとに開発・維持コストが増大する。一方、AIに基づく手法は大量の学習データが必要となるが、一般的に請求書などの帳票には公表されたデータが少なく、健康診断書や住民票など機微なデータに至っては更に収集が困難である。また、事前に想定した項目以外の記載内容の取得や、特定様式での精度向上などの実用上考えられるニーズへの柔軟な対応も課題であり、これらのコストも抑える必要がある。

3.2 見出し項目サーチ型

プリセット型は高精度な抽出を実現できるが、用途は事前に想定したものに限定される。そこでプリセット型では網羅しきれないユーザーの幅広いニーズに対応するため、見出

し項目サーチという機能を提供している（図2）。

これは、抽出したい項目に対して以下のような情報をユーザーが設定し、当てはまる文字列を記載内容として抽出するものである。

- (1) 見出しとなり得る文字列（複数指定可）
- (2) 記載内容の字種

これにより、既存の類型にない帳票や項目であっても抽出できる。

しかし、記載内容を探すための手掛かりが見出しと字種しかないため、見出しがない項目は抽出できない。また記載位置、書式、算術関係といった情報も使えないことから、対応できる帳票デザインには制約が生じる。例えば図3に示した帳票に対して、見出しとして“銀行”を指定した場合、書式などの付加情報がない限り、右方の“支店”と下方の“東芝”のどちらが見出しに対応する項目かの判別は難しい。

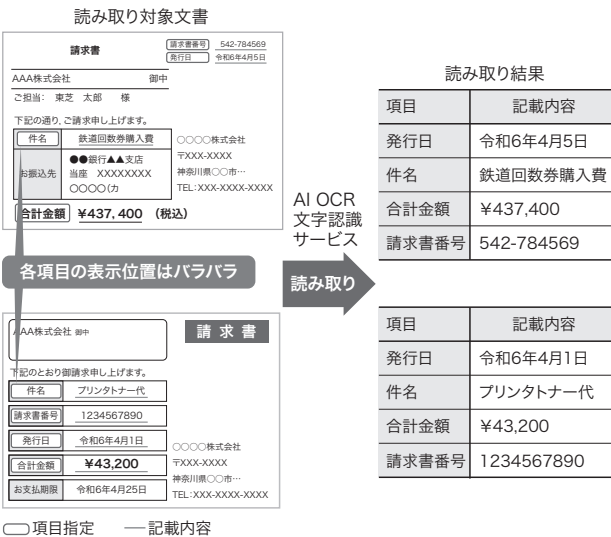


図2. 見出し項目サーチ型非定型帳票認識の例

任意の帳票画像を認識し、見出し文字列などで指定されたユーザー定義項目に対応する記載内容文字列を抽出する。

Examples of recognition results of semi-structured document with header item search mode

銀行	支店
東芝	川崎

図3. 見出し文字列だけでは適切な項目抽出が難しい例

表のレイアウトだけでは、見出しの“銀行”に対応する記載内容が右方と下方のどちらにあるかを判別するのは難しい。

Case in which extraction of appropriate items is difficult by only specified header strings

4. AI 言語モデルによる非定型帳票認識技術

4.1 見出し項目サーチのAI 言語モデルによる拡張

対応する帳票類型や、項目、ユーザーごとに異なる幅広い需要などに対応するカスタマイズのために、見出し項目サーチ型の柔軟さとプリセット型の認識精度を両立できる仕組みが必要である。

基本的な考え方として、見出し項目サーチ型で可能な読み取り対象項目の設定に、プリセット型で用いる属性情報を選択的に付加できるようにすれば、任意の対象帳票に対して高い精度での読み取りができるようになる。ただし、あまり複雑な設定をユーザーに要求することは、AI OCR 文字認識サービスが提供する省力化の価値を損なう。このため、できるだけ少ない付加情報から見出し項目を抽出するように設定できるのが望ましい。

このような設定の簡便さと高い読み取り精度の両立のため、当社は、見出し項目サーチ型の非定型帳票認識への、AI 言語モデルによる属性推定の導入（図4）を検討している⁽⁶⁾。

AI 言語モデルによる属性推定は、文字列を単語などの要素に分解し、その中からユーザーが抽出したい情報である蓋然性が高い特定の内容が記載された部分を抽出し、更にそれらに“金額”，“日付”，“法人名”など内容に応じた、抽出したい項目より一段階汎化した分類を付与する技術である。この中から、ユーザーが抽出したい項目に該当する分類を選択することにより、記載内容の候補である文字列をかなり絞り込める。更に、これに対して見出し文字列・記載位置などの少数の設定を付け加えることで、見出し項目サーチに近い手軽さでより高い精度を得ることが期待できる。

見出し項目サーチにAI 言語モデルを導入した手法で、請求書を対象に項目抽出を試行した結果を、プリセット型と比較して図5に示す。汎用を目指した手法であるにもかかわらず、項目によってはプリセット型と同等の抽出精度を実現で

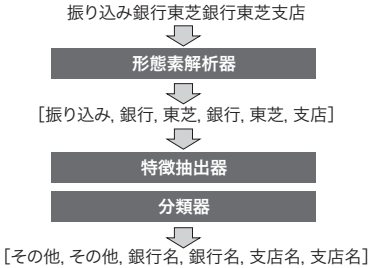


図4. AI 言語モデルによる属性推定

一続きの文字列を単語などの形態素に分解し、それぞれの単語を内容に応じて分類する。

Flow of entity extraction of words using AI language model

きたことが分かる。

この結果はまた、AI 言語モデルで付与された分類を見出し項目抽出条件の指定に用いることで、プリセット型で行う帳票類型の作成や対応項目の新規追加を、より低コスト、短期間で行えることも意味している。

4.2 生成AIの活用

非定型帳票認識にAI 言語モデルを導入することで、従来よりも多様な帳票の認識が可能となるが、対応可能帳票類型や項目の制約がなくなつてはならず、モデル開発コストの問題も残る。これらの課題に対応するため、当社は、大規模言語モデル(Large Language Model : LLM)を用いた生成AIを非定型帳票認識に活用することも検討している。

一般に、LLMを用いた生成AIは、自然言語の入力というユーザーにとってなじみのある手段を使って、広範な問題に対する適切な回答を得られるため、4.1節で述べたようなカスタマイズを容易にすることに、柔軟に対応できる潜在能力を持つ。反面、LLMは汎用用途に学習されている巨大モデルであるため、学習及び運用コストは高く、また必ずしも非定型帳票認識のような特殊な用途に対して最適な出力が得られるとは限らない。このため、生成AIによる非定型帳票認識では、プロンプトエンジニアリングといった小規模な工夫から、モデル自体の追加学習といった比較的大規模な開発まで、いろいろな手段を使って高い認識精度とコスト削減を両立させるための最適解を模索することが必要である。

これまで行った評価実験では、比較的簡単な工夫をすることで、非定型帳票認識の精度を高められることを示唆する結果が得られた。このことは、多様なニーズへの対応を更に少ない開発コストで実施できること、またユーザー側でのカスタマイズの柔軟性やそれによる対応可能範囲を更に

拡大できる可能性があることを意味している。

5. あとがき

当社が提供するAI OCR文字認識サービスは機能・性能向上を続けているが、ここでは、非定型帳票認識にAI 言語モデルを適用する取り組みについて述べた。今後は、LLMの活用などをはじめ、最新のAI関連技術をより幅広く導入し、更に柔軟な定義ができる手法や、少量の学習データでも精度向上できる手法などの研究開発を続け、ユーザーの利便性向上に努めていく。

文 献

(1) Devlin, J. et al. "BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding". Proceedings of the 2019 Annual Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (NAACL-HLT) 2019, Volume 1. Minneapolis, MN, 2019-06, Association for Computational Linguistics. 2019, p.4171-4186.

(2) Vaswani, A. et al. "Attention is all you need". Proceedings of the 31st International Conference on Neural Information Processing Systems (NIPS 2017). Long Beach, CA, 2017-12, The Neural Information Processing Systems Foundation. 2017, p.5999-6009.

(3) Xu, Y. et al. "LayoutLM: Pre-training of Text and Layout for Document Image Understanding". Proceedings of the 26th ACM SIGKDD Conference on Knowledge Discovery & Data Mining (KDD 2020). Virtual Event, 2020-08, Association for Computing Machinery. 2020, p.1192-1200.

(4) 石井一志, "さくら情報システム、生成AIとAI-OCRを組み合わせた新サービス「AI Textract (仮称)」を6月にも提供開始". クラウド Watch. <<https://cloud.watch.impress.co.jp/docs/news/1563681.html>>, (参照 2024-01-10).

(5) GMO Internet Group, "請求書OCR自動化 : Document AI+ChatGPT APIで非構造化データをJSONで出力させる". GMO INTERNET GROUPグループ研究開発本部. <<https://recruit.gmo.jp/engineer/jisedai/blog/automated-invoice-ocr-with-chatgpt-and-documentai/>>, (参照 2024-01-10).

(6) 馮 思萌, "言語モデルを用いた帳票文書からの情報抽出の改良". 電子情報通信学会総大会講演論文集 (CD-ROM). オンライン, 2022-03, 2022, D-5-1.

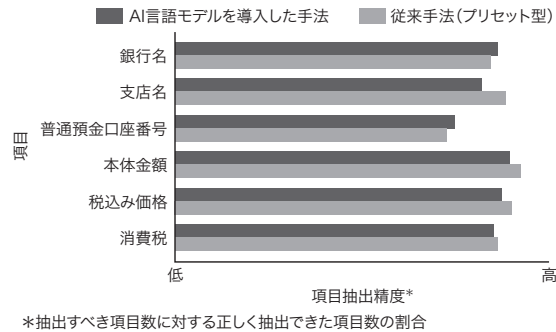


図5. AI 言語モデルを導入した場合と従来のプリセット型による抽出精度の比較

見出し項目サーチにAI 言語モデルを適用すると、事前準備に時間や手間を掛けなくても、プリセット型と同様の抽出精度が得られた。

Comparison of item extraction accuracy of conventional preset-based and AI language model-assisted recognition technologies



小野 聡一郎 ONO Soichiro
東芝デジタルソリューションズ (株)
ソフトウェアシステム技術開発センター コアテクノロジー開発部
情報処理学会会員
Toshiba Digital Solutions Corp.



馮 思萌 FUU Shimou
東芝デジタルソリューションズ (株)
ソフトウェアシステム技術開発センター コアテクノロジー開発部
Toshiba Digital Solutions Corp.



古畑 彰夫 FURUHATA Akio
東芝デジタルソリューションズ (株)
ソフトウェアシステム技術開発センター コアテクノロジー開発部
電子情報通信学会会員
Toshiba Digital Solutions Corp.