大規模言語モデルを用いたマイソク PDF からの情報抽出

本郷 槙一 [1,3]　叶内 晨 [2,3]　齊藤 佑太郎 [3]　岩成 達哉 [3]
[1] 京都大学大学院 情報学研究科
[2]NLPeanuts Inc.　[3] 株式会社 estie
hongo@nlp.ist.i.kyoto-u.ac.jp　shin.kanouchi@nlpeanuts.com
{yutaro.saito, tatsuya.iwanari}@estie.co.jp

概要

本研究では，不動産業界で物件情報の流通に使用される PDF データからの効率的な情報抽出を目指す．実験では，PDF を入力として，あらかじめ指定したカテゴリに対してどの程度情報抽出できるか検証した．実験の結果，OCR による出力後に大規模言語モデルを用いて必要な情報を抽出する 2 ステップの手法の精度が最も高く，Accuracy 0.92 を達成した．また，データ入力業務において本研究の出力結果を活用することで，65%の業務時間削減が可能なことを確認した．

1　はじめに

不動産業界では情報のデジタル化に伴い，情報の収集や管理の効率化が進められている．一方，多くの物件情報はマイソクと呼ばれる PDF フォーマット（図 1 左側．以下「マイソク PDF」と呼ぶ）で提供され，物件情報が 1 枚の PDF に取りまとめられている．しかし，PDF からの情報抽出が依然として難しい問題や，マイソク PDF の記載方法が各社によって異なる問題により，これらの情報を収集・管理する業務が業界で大きな課題となっている．

そこで本研究では，マイソク PDF から必要な情報を抽出する新しいタスクを定義し，検証する．提案手法として，近年急成長を続ける光学文字認識技術（OCR）と大規模言語モデル（LLM）を活用する．既存の商用 OCR サービスでは，レイアウトが複雑な PDF に対しては出力が崩れてしまい，情報を正確に抽出することは依然として難しい．そこで，OCR の出力結果を LLM への入力とし，プロンプトにより出力項目とフォーマットを指定することで，必要な情報の抽出を目指す．また，OCR を利用せずに LLM のみで画像から必要な情報を直接抽出する End-to-End な手法もあわせて検証する．

実験では，不動産業界で実際に使用されているマイソク PDF を評価データとし，あらかじめ定義した 11 カテゴリに対しての情報抽出の精度を評価した．実験の結果，OCR 処理後に GPT-4[1] を用いて情報を抽出する 2 ステップの手法の精度が最大となり，Accuracy 0.92 を達成した．また，本研究の出力結果を活用することで，どの程度データの入力業務を効率化できるか計測した．計測の結果，提案手法による抽出結果を事前入力として利用することにより，手作業で情報を入力する場合に比べて約 65%の時間削減が可能なことを確認した．

2　先行研究

近年，不動産業界におけるデジタル化が進み，自然言語処理や機械学習，深層学習の活用が進展している．不動産取引のための自動価格査定 [2, 3, 4, 5] や，利用者に対する最適な物件の提案 [6, 7]，物件情報の管理や運用における効率化 [8, 9] の研究などが進められている．

GPT-4[1] をはじめとした大規模言語モデル（LLM）は，テキストデータの理解と生成において顕著な能力を示しており，情報抽出や情報の構造化においても有用である [10]．これらのモデルは不動産領域のような専門的な情報に対しても活用可能であると考えられる．

OCR 技術も深層学習の進歩により精度が大幅に向上しており [11]，商用サービスの開発も盛んである [12]．PDF や画像から情報抽出する研究も注目を集め，OCR 処理後に OCR 結果に対して LLM を活用して情報抽出をする研究 [13, 14] や，End-to-End で画像から情報抽出する研究 [15]，画像情報を構造化するベンチマーク [16] などがある．本研究では，OCR 後に LLM を利用する 2 ステップの手法と，LLM のみで画像から直接情報抽出する End-to-End な手法の両手法を検証する．

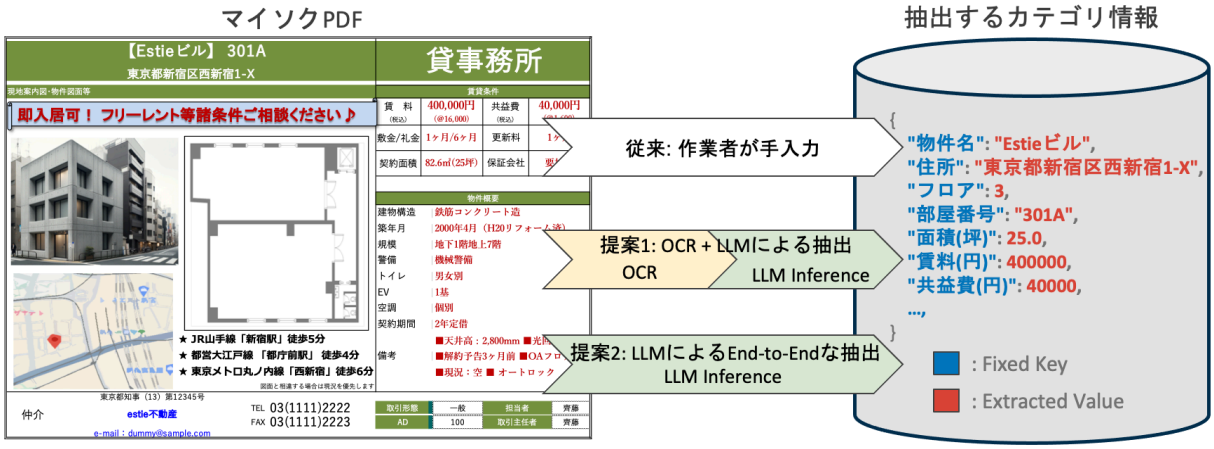

図 1　マイソク PDF からの情報抽出の流れ

表 1　抽出するカテゴリ情報

カテゴリ名	型	概要	Null 比率
物件名	文字列	ビルの名前．ただしフロアや部屋番号は除く（例: Estie ビル）	0 %
住所	文字列	ビルの所在地（例: 東京都新宿区西新宿 1-X）	0 %
フロア	整数	募集物件の位置する階．ただし地下はマイナスで表現 (例: B3F → -3)	0 %
面積 (坪)	小数	募集物件の坪面積．ただし m^2 のみの場合，単位の変換が必要．（例: 25.0）	3 %
部屋番号	文字列	募集物件の部屋番号．ただしフロア全体を借りる場合は Null（例: 301A）	67 %
契約開始日	文字列	契約を始められる日．別名，引渡時期/入居時期（例: 即日, 相談）	6 %
賃料総額 (円)	整数	月ごとに支払う賃料の総額（例: 400000）	2 %
共益費総額 (円)	整数	月ごとに支払う共益費の総額（例: 40000）	3 %
礼金 (ヶ月)	小数	礼金は賃料の何ヶ月分か．ただし記載がない場合は 0 とする（例: 1）	0 %
敷金 (ヶ月)	小数	敷金は賃料の何ヶ月分か．ただし記載がない場合は 0 とする（例: 3）	0 %
契約期間 (年)	小数	定期借家契約の場合の契約年数（例: 2）	5 %

3　タスク概要

不動産の物件情報はさまざまな形式で流通しているが，そのうちの一つとして図 1 左側に示すような「マイソク PDF」と呼ばれるフォーマットがある．このフォーマットでは，物件の概要・間取り図・契約情報などが 1 枚に取りまとめられている．物件の貸主が物件を紹介する際に利用する資料で，不動産業者間で欠かすことのできない情報となっている．

しかし，マイソク PDF は各不動産会社が作成する独自のフォーマットで提供され，これらの情報を効率的に集約し蓄積するには，多大な労力を要する．以下の課題が，処理の自動化を困難にしている．

1. **カテゴリ情報の定義のずれ**: PDF の書き方に正式な決まりがないため，個社ごとに定義が異なり，情報の統合が困難である．
2. **カテゴリ名の省略**: 例えば，図 1 で「Estie ビル」が物件名であることは明示されていない．
3. **カテゴリ情報の混在**: 例えば「Estie ビル 301A」と書いてある場合，物件名が「Estie ビル」，部屋番号が「301A」と推測する必要がある．
4. **例外対応**: 例えば「301」とあればフロアは「3 階」と予想するが，地下 3 階が正解の場合もあり，PDF 全体を考慮して判断する必要がある．

本研究ではこれらの課題を解決するため，マイソク PDF から情報を抽出する新たなタスクを定義する．タスクの流れを図 1 に示す．PDF を入力とし，決められたカテゴリに対する情報抽出を行う．また，抽出する情報はその後データベースに登録されることを考慮して，型を指定した上で抽出する．

各カテゴリについての説明を表 1 に示す．カテゴリは不動産取引において重要視される 11 種類に限定した．また，マイソク PDF には特定のカテゴリ情報が含まれない場合もがあるため，その比率を Null 比率として示す．部屋番号は記載されない場合も多く，Null 比率が 67%となっている．

検証に用いるデータは実際の賃貸オフィス用のマイソク PDF とし，200 件に対して正解情報をアノテーションした．利用する PDF はランダムに抽出するが，1 枚のマイソク PDF に複数の物件情報が記載されているものや，物件が重複しているものなどは事前に目視で確認し除外した．

表 2　マイソク PDF に対する情報抽出精度 (Acc)

モデル名	物件名	住所	フロア	面積	部屋番号	契約開始日	賃料	共益費	礼金	敷金	契約期間	マクロ平均
OCR + GPT-4	0.91	0.99	0.97	0.85	0.95	0.88	0.97	0.95	0.92	0.92	0.99	0.936
OCR + GPT-3.5-turbo	0.87	0.98	0.92	0.81	0.83	0.69	0.91	0.80	0.90	0.69	0.96	0.852
GPT-4V	0.23	0.26	0.82	0.75	0.27	0.21	0.72	0.59	0.33	0.24	0.74	0.469
Gemini Pro Vision	0.26	0.17	0.68	0.36	0.44	0.35	0.73	0.54	0.54	0.15	0.72	0.449

4　提案手法

本研究では，OCR 結果に対して LLM を用いる 2 ステップの手法と，LLM のみで画像から直接情報抽出する End-to-End な手法の 2 つを提案する．

4.1　OCR+LLM による 2 ステップの抽出

マイソク PDF に対して OCR を行いテキスト情報を取得後，さらに LLM を用いて必要なカテゴリ情報を抽出する 2 ステップの処理を提案する．OCR には Azure AI Document Intelligence[1)] を利用する．

LLM には OpenAI の GPT-3.5-turbo（1106）と GPT-4（0613）[1] を用いる．OCR による出力結果とプロンプトを LLM の入力とし，OpenAI の Function calling 機能を利用することで出力形式を指定する．Function calling は OpenAI API の機能であり，JSON Schema を推論時のパラメータとして付与することで，指定した形式通りの JSON を出力することができる．プロンプトには，ドメインが不動産であること，入力テキストは PDF データを OCR した結果であること，Function calling に指定された Key 情報を抽出することなどを指示した．

4.2　LLM による End-to-End な抽出

End-to-End の手法では，OCR を介さずに，LLM を用いて画像情報から直接カテゴリ情報を抽出する．LLM には，GPT-4 Turbo with Vision（GPT-4V）と Gemini Pro Vision [17] を用いた．ただし，モデルへの入力のため，PDF データは一度 PNG に変換した．プロンプトの一部として出力用の JSON Schema の情報を与えることで，出力を制御した．

5　実験

マイソク PDF を入力とし，あらかじめ指定した 11 個のカテゴリに対してどの程度抽出できるか実験を行った．100 件の検証用データでプロンプトチューニングを行い，100 件のテストデータで評価した．評価方法は，カテゴリ毎の完全一致による Accuracy とした．ただし文字列型のカテゴリに関しては表記揺れを吸収する処理として，空白・改行の除去と NFKC 正規化などを行った．

LLM のパラメータとして，*temperature* = 0 に固定した．また，一部モデルにおいて JSON フォーマットが崩れる問題などにより処理に失敗するケースがあった．そのため，エラーが出た場合は機械的に 5 回まで処理を再実行し，それでも出力に失敗する場合は抽出に失敗したものとした．

5.1　実験結果

マイソク PDF に対する各カテゴリの抽出精度を表 2 に示す．OCR+GPT-4 の精度が最大となり，マクロ平均 Accuracy は 0.936 となった．カテゴリ毎のスコアを見ると，住所や契約期間などの正解となりうる候補が少ない項目の精度が最も高くなっている．一方で，面積や契約開始日などは 0.90 を下回る結果となった．

OCR+LLM vs End-to-End　OCR+LLM の 2 ステップの手法に比べ，End-to-End の手法の精度は著しく下回った．特に物件名や住所など，日本語の抽出を必要とするカテゴリにおける差が顕著となった．2 ステップの手法では正解したが End-to-End の手法で間違えてた事例を確認すると，例えば物件名では「カメヤビル」を間違えて「カメヒビル」と抽出していたり，住所では「渋谷区」を間違えて「四谷区」と抽出しており，日本語の認識精度に課題があることを確認した．

GPT-4 vs GPT-3.5-turbo　OCR+LLM の 2 ステップの手法において，GPT-4 を用いた手法が GPT-3.5-turbo を用いた手法に比べて 8.4 ポイント高い結果となった．敷金や共益費について大幅な精度の低下が見られたためミス事例を個別に目視で確認した結果，GPT-3.5-turbo では関連性の高いカテゴリがある場合に値の入れ替えや格納ミスが多いことを確認した．例えば「敷金と礼金」や「共益費と賃料」のペアは情報が似ており，それぞれの値を入れ替えてしまったり，片方の値に引っ張られるケースが散見さ

1）https://azure.microsoft.com/ja-jp/products/ai-services/ai-document-intelligence

表 3 編集距離ごとの抽出精度 (Acc)

編集距離	物件名	住所	部屋番号	契約開始日	マクロ平均
1.0	0.91	0.99	0.95	0.88	0.933
0.8	0.93	0.99	0.95	0.89	0.940
0.6	0.95	0.99	0.95	0.90	0.948

表 4 抽出に失敗した事例の分類

エラーの種類	件数	(割合)
抽出位置に関するミス	21	(30%)
抽出後に推論が必要	20	(29%)
抽出漏れ	12	(17%)
勝手な補完	7	(10%)
その他	10	(14%)

れた．GPT-4 ではこれらの誤りが少なく，モデル間の処理能力の差が顕著に表れていた．

編集距離と抽出精度　表 3 に，文字列型のカテゴリに関して編集距離が 0.6，0.8，1.0 の閾値以上の場合は正解とした際の OCR+GPT-4 の精度を示す．編集距離の閾値を下げることで表記揺れなどによる誤りが減ったため，各カテゴリで Accuracy が改善し，特に物件名に関して，4 ポイント上昇した．編集距離の閾値を下げたことで正解となった事例を確認すると，正解が「2023 年 11 月上旬」に対して抽出結果が「西暦 2023 年 11 月上旬」のように，意味的に同義な事例を正解に含めることができるようになった．一方で，情報が 2 行に分かれているにもかかわらず 1 行目しか抽出できていない事例や，前後にノイズを含んでしまっている事例も誤って正解としてしまうケースも確認された．

推論が必要なケースへの対応　ほとんどの場合は PDF の情報をそのまま抽出することで正解となるが，抽出後に推論が必要なケースがある．例えば敷金において，一般的には賃料の n ヶ月分と表現されるが，円で表記されることがある．賃料が 98,000 円の物件で敷金が 196,000 円と記載されているとき，正解は $196,000/98,000 = 2$ という計算に基づいた月単位の値である．GPT-3.5 モデルではこの種の推論に基づく応答が難しい一方で，GPT-4 モデルは正解している事例を複数確認した．この事例は，GPT-4 が単なるテキストからの情報抽出に加えて，推論を行った結果を出力できることを示している．

同様に，面積 (坪) を抽出する際にも推論が必要となることがあった．PDF に m^2 単位の記載しかない場合，その値を 3.3 で割って坪単位に変換した値が正解となる．こちらも GPT4 モデルでは数件正解していたが，敷金の場合と比べると計算結果の誤りや m^2 単位のまま記載する失敗が目立った．

抽出結果の誤り事例の分類　OCR+GPT-4 の手法による出力のうち，テストデータに対する誤り事例全 70 件を分類した結果を表 4 に示す．最も多い誤りは抽出位置に関するもので，OCR によってテキストの位置情報を失ったために生じた値の入れ違いや，前後にノイズを含んでしまう事例が該当する．次に誤りの多い抽出後に推論が必要なケースは，前述の敷金や面積などにおいて情報の変換が必要な事例である．残りは正解が空にもかかわらず勝手に値を補完するケースと，予測結果が空である抽出漏れの誤りである．その他は，PDF データの欠陥により判別がつかないケースなどがある．

6　業務効率化の測定

実際にデータ入力業務を行っているオペレータに協力を依頼し，どの程度業務を効率化できるか検証した．一定時間データの入力作業をしていただき，1 時間あたり何件のマイソク PDF を処理できたかを計測した．比較対象は以下とする．

1. PDF 情報の事前入力なしに手入力した場合
2. OCR+GPT4 による 2 ステップの手法による出力結果を事前入力として利用した場合

計測の結果，(1) の事前入力がない場合が平均 20 件/ h だったのに対して，(2) の提案手法による事前入力ありの場合は平均 57 件/ h となり，約 65%の時間削減が可能なことを確認した．(1) の場合は，すべての情報をタイピングする必要があるが，(2) の場合は事前入力された情報を確認し，修正するのみであるため，大幅に速度が向上した．オペレーターに対するヒアリングの結果, タイピングの負担が大幅に軽減されてより情報の内容に集中できたという意見が得られた．特に住所については文字数が多く正解率の高いカテゴリであるため，好評であった．

7　おわりに

本研究では，不動産業界における物件情報の情報抽出に関する新しいタスクを提案し，その有効性を実証した．OCR と GPT-4 を組み合わせた方法は，マイソク PDF の処理において高い精度を達成し，実業務における効率化を実現した．今後は，さらなる精度向上と自動化の範囲拡大を目指したい．また，この技術は他の業界における類似の課題にも応用可能であり，多方面への展開が期待される．

参考文献

[1] Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altenschmidt, Sam Altman, Shyamal Anadkat, et al. GPT-4 technical report. **arXiv:2303.08774**, 2023.

[2] Gordon S. Linoff and Michael J. A. Berry. Data mining techniques: For marketing, sales, and customer relationship management. **Wiley Publishing**, 2011.

[3] 清水千弘. 市場分析のための統計学入門. 朝倉書店, 2016.

[4] Hamza Usman, Mohd Lizam, and Muhammad Usman Adekunle. Property price modelling, market segmentation and submarket classifications: A review. **Real Estate Management and Valuation**, Vol. 28, No. 3, pp. 24–35, 2020.

[5] Vilius Kontrimas and Antanas Verikas. The mass appraisal of the real estate by computational intelligence. **Applied Soft Computing**, Vol. 11, No. 1, pp. 443–448, 2011.

[6] 大知正直, 関喜文, 川上登福, 小野木大二, 野村眞平, 吉永恵一, 松尾豊. 推薦そのものがユーザに与える影響を考慮した情報推薦. 人工知能学会全国大会論文集, Vol. JSAI2014, pp. 2M32–2M32, 2014.

[7] Alireza Gharahighehi, Konstantinos Pliakos, and Celine Vens. Recommender systems in the real estate market—a survey. **Applied Sciences**, Vol. 11, No. 16, 2021.

[8] Danielle Sanderson and Dustin Read. Recognizing and realizing the value of customer-focused property management. **Property Management**, Vol. ahead-of-print, , 07 2020.

[9] Taran Kaur and Priya Solomon. A study on automated property management in commercial real estate: a case of india. **Property Management**, Vol. 40, pp. 247–264, 09 2021.

[10] Alexander Dunn, John Dagdelen, Nicholas Walker, Sanghoon Lee, Andrew S. Rosen, Gerbrand Ceder, Kristin Persson, and Anubhav Jain. Structured information extraction from complex scientific text with fine-tuned large language models. **arXiv:2212.05238**, 2022.

[11] Nishant Subramani, Alexandre Matton, Malcolm Greaves, and Adrian Lam. A survey of deep learning approaches for ocr and document understanding. **arXiv:2011.13534**, 2020.

[12] Mathew Salvaris, Danielle Dean, and Wee Hyong Tok. Deep learning with azure. **Building and Deploying Artificial Intelligence Solutions on Microsoft AI Platform, Apress**, 2018.

[13] Vincent Perot, Kai Kang, Florian Luisier, Guolong Su, Xiaoyu Sun, Ramya Sree Boppana, Zilong Wang, Jiaqi Mu, Hao Zhang, and Nan Hua. Lmdx: Language model-based document information extraction and localization. **arXiv:2309.10952**, 2023.

[14] 佐藤栄作, 木村泰知. Topix100 の有価証券報告書の表を対象とした chatgpt による pdf から json への自動変換の試み. 第 22 回情報科学技術フォーラム（FIT2023）, 9 2023.

[15] Geewook Kim, Teakgyu Hong, Moonbin Yim, JeongYeon Nam, Jinyoung Park, Jinyeong Yim, Wonseok Hwang, Sangdoo Yun, Dongyoon Han, and Seunghyun Park. Ocr-free document understanding transformer. In **European Conference on Computer Vision**, pp. 498–517. Springer, 2022.

[16] Zilong Wang, Yichao Zhou, Wei Wei, Chen-Yu Lee, and Sandeep Tata. VRDU: A benchmark for visually-rich document understanding. In **ACM SIGKDD Conference on Knowledge Discovery and Data Mining**, pp. 5184–5193, 2023.

[17] Gemini Team, Rohan Anil, Sebastian Borgeaud, Yonghui Wu, Jean-Baptiste Alayrac, Jiahui Yu, Radu Soricut, Johan Schalkwyk, Andrew M Dai, Anja Hauth, et al. Gemini: a family of highly capable multimodal models. **arXiv:2312.11805**, 2023.