PDF 文書の階層構造を用いた領域分割の性能向上に関する研究

伊東桂佑 [1]　鶴田直之 [1]　乙武北斗 [1]

概要：住民の地方自治への関心を高めるため，地方自治体の会議資料をテキストにして自然言語処理可能なデータベースの構築を進めている．そこで，PDF 文書で公開されている自治体の議会活動資料から情報を抽出するため，画像処理によって文書の構成要素と配置を抽出し，テキスト化の手がかりとすることを検討している．本研究では先行研究の改良として画像認識用深層学習 Detectron2 を用いることと，文書構成要素の階層構造に関する知識の見直しとを行い性能向上を目指した．実験では，文書構成要素の抽出精度の向上が確認できた．

キーワード：文書画像処理，深層学習，Detectron2，階層構造

A Study on Performance Improvement of Region Segmentation Using Hierarchical Structure of PDF Documents

KEISUKE ITO[†1]　NAOYUKI TSURUA[†1]
HOKUTO OTOTAKE[†1]

Abstract: In order to increase residents' interest in local government, we are constructing a natural language-processable database of local government meeting documents as text. Therefore, in order to extract information from local government meeting activity documents published in PDF documents, we are considering using image processing to extract the components and arrangement of the documents and use them as clues for textualization. In this study, we aimed to improve the performance by using deep learning Detectron2 for image recognition as an improvement of the previous study and by reviewing the knowledge about the hierarchical structure of document components. Experimental results showed that the accuracy of extracting document components was improved.

Keywords: Document image processing, deep learning, Detectron2, hierarchical structure

1. はじめに

住民の地方自治への関心を高めるため，地方自治体の会議資料をテキストにして自然言語処理可能なデータベースの構築を進めている[1][2]．そこで，PDF 文書で公開されている自治体の議会活動資料から情報を抽出するため，画像処理によって文書の構成要素と配置を抽出し，テキスト化の手がかりとすることを検討している．具体的には先行研究[3]において，画像認識深層学習モデルを用いて文書の構成要素を抽出した．その際，文書の構成要素間の階層関係が構成要素抽出の手掛かりとして利用する方法を試みた．本研究では，先行研究の改良として画像認識用深層学習 Detectron2 を用いることと，文書構成要素の階層構造に関する知識の見直しとを行い性能向上を目指した．実験では，文書構成要素の抽出精度の向上が確認できた．

2. 基本技術

先行研究[3]においては，一般物体検出用に提案された深層学習モデルの一つである Single Shot MultiBox Detector(SSD)[4]を用いていた．SSD は，矩形領域での物体検出が可能となる．検出を行う対象が地方自治体の活動資料で，文書の構成要素が四角いものが多いため矩形領域でのアノテーションを行っていた一方，構成要素同士の重なりを考慮することができない一面があった（図 1）．

これに対し本研究では Detectron2 [5]を用いた．Detectron2 は FacebookAI が開発した PyTorch ベースの物体検出ライブラリで，ピクセル単位での領域の Instance Segmentation による物体を検出を行うことができる．また，アノテーションを行う際，アノテーションツール FastLabel[6]を用いた．

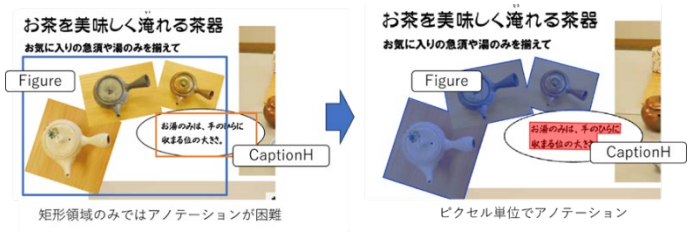

図 1　アノテーション時の方法を変更

1 福岡大学
†1 Fukuoka University

3. 先行研究と課題

3.1 先行研究

地方自治体の議会活動資料に対し文書の構成要素のカテゴリを定義し，これらカテゴリに沿って構成要素の抽出を行った．カテゴリの定義は表 1 の通りである．構成要素同士の包含関係を，文書や図表全体を親（大カテゴリ），文字，図表,キャプションを子（小カテゴリ）とする階層構造として定義する．図 2 親カテゴリと子カテゴリの例を示す．

表 1　先行研究で用いたカテゴリと階層関係
（H は横書き，V は縦書きを示す）

カテゴリ		定義
大	小	
PTitleV		ページ内に一つだけ存在する大見出し
PTitleH		
PSegment		文章主体の領域
	TitleV	領域内に一つだけ存在するタイトル
	TitleH	
	LeadV	PSegment 内の 1～2 行の文章，リード文
	LeadH	
	ParagraphV	領域内における段落
	ParagraphH	
	Figure	図や画像，挿絵
FSegment		画像主体の領域
	TitleV	図や画像，挿絵
	TitleH	FSegment 内の表全体
	ParagraphV	領域内における段落
	ParagraphH	
	Figure	図や画像，挿絵
	Table	FSegment 内の表全体
	CaptionV	図や表に一つ存在するキャプション
	CaptionH	

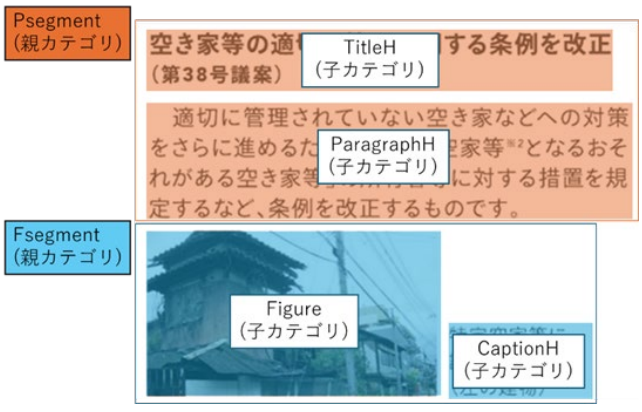

図 2 親カテゴリと子カテゴリの例

そして，この親カテゴリと子カテゴリの関係性を 2 部マッチング問題と捉え，認識・検出精度を向上させる単純な方法として，次の二つの方法と検証していた．

一つは，親カテゴリが高い確率で検出されているにもかかわらず，そこに含まれている子カテゴリが検出されていない場合で，検出時の確率のしきい値を下げ，未検出になっていた子カテゴリ候補を検出とみなすトップダウン方式である．もう一つは，逆に，複数の子カテゴリが高い確率で検出されているにもかかわらず，それらを包含する親カテゴリが検出されていない場合で，検出時の確率のしきい値を下げ，未検出になっていた親カテゴリ候補を検出とみなすボトムアップ方式である．

先行研究では全体的に適合率が低かった図表関連のカテゴリ(FSegment とその子カテゴリ)に対してトップダウン法を採用し，最終的な検出結果は表 2 のようになった．

表 2 ボトムアップ方式を採用した検出結果

	適合率
PTitleV	85.71%
PTitleH	24.21%
PSegment	55.47%
TitleV	78.46%
TitleH	48.45%
LeadV	81.14%
LeadH	68.83%
ParagraphV	83.21%
ParagraphH	70.28%
FSegment	57.22%
Figure	15.11%
Table	45.65%
CaptionV	0.00%
CaptionH	0.00%
平均	50.98%

3.2 先行研究の課題

先行研究の課題として表 2 より，PtitleH ，TitleH と FSegment の子カテゴリである Figure , Table , CaptionV , CaptionH の適合率が 50%以下で, 他のカテゴリと比較して低い数値となっている．特に課題となるのが CaptionV, CaptionH の適合率で，結果が 0%であり検出結果のすべてが誤検出となる．

これらの原因として，それぞれ画像や表の説明文としての役割を持つ Caption が，多くの場合が短い 1 行程度の文となり，リード文の役割持ち同様に短い 1 行程度の文である Lead に類似しているため正確に検出することができていないこと，また Caption に対しての教師データが少ないこと．加えて図表に対してアノテーション時，物体が重なることの 3 つのことが考えられた．更には，各カテゴリに対する定義があいまいでアノテーション時を行う人によって解釈の違いが生じていた．

4. 提案手法

4.1 カテゴリの定義の見直し

本研究では，カテゴリの定義の見直しを行うことで精度向上の手法を提案する．

一例を紹介すると，先行研究において PSegment は文章が主体の領域，FSegment は画像が主体の領域，という定義になっている．図 3 に対してアノテーションによるタグ付けを行う際，全体の領域として PSegment か FSegment，どちらかを選ぶ必要がある．その場合，文章を読むと画像に対しての説明がなく，文章を補足する形で画像が配置されているため PSegment と判断できる．一方，文章を無視した場合，領域全体が画像で占められているので FSegment とも判断できてしまう．

図 3　PSegment か FSegment どちらか判別しにくい例

そのため，以下のような再定義を行う．PSegment はタイトル，リード文，文章を持つかたまり．FSegment は画像，表，その説明文をもつかたまり．この再定義した 2 つのカテゴリを図 3 に対して適用すると図 4 のようになる．

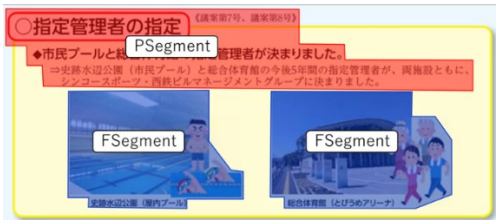

図 4　再定義した FSegment と PSegment を適用した例

また，PSegment と FSegment 双方に内包していた Title と Paragraph を PSegment のみに属し， Figure を FSegment のみ属するように再定義した．以下に新たに再定義したカテゴリと，文書の構成要素間の階層関係（図 5）を示す．

カテゴリの定義は以下の通りである．

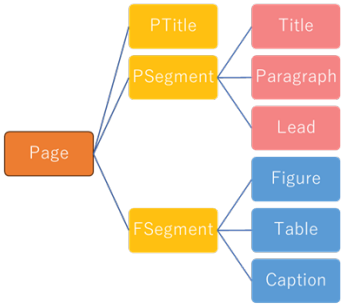

図 5　文書の構成要素間の階層関係

先行研究から変更した点は 3 つある．1 つ目は PTitle の横書き縦書きの統一である．PTitle の特徴は 1 ページを代表するタイトルで，縦書き，横書きで記載されている場合もあれば，イラストと共に書かれていたり，文字を立体的にイラストのように描かれることもある．横書き縦書きかどうかの判断がつかない場合を加味して再定義を行った．

2 つ目は箇条書きであるものを Paragraph として定義していたことである．ページ内に箇条書きで記載されている箇所のアノテーションが Paragraph と Table であることがどちらも見受けられた．そのためアノテーション時に区別がつくよう定義を付け加えた．これに従って Table は箇条書きではない．

3 つ目は Table の定義を，罫線で囲まれた文字や数字とした．2 つ目の変更点と同様に Table である表に対する定義を明確にした．

4.2 学習データの準備

学習データは 215 枚の画像を用い，アノテーションは具体例とともにフローチャートを用いたマニュアルを作成し，アノテーションを事業として行っている業者に依頼した．

5. 実験

5.1 実験の目的と方法

以下の 2 つの目的で実験を行った．

- 実験 1 : 再定義したカテゴリを元にした学習データを用いて Detectron2 の物体検出によって文書の構成要素を抽出した結果を評価する．
- 実験 2 : 実験 1 で得られた結果の図表関連のカテゴリに対してトップダウン法を適用し，評価する．

なお，実験環境は，GoogleColaboratory 及び NVIDIA A100 Tensor Core GPU 80GB 搭載の PC で行った

実験 1 では，再定義したカテゴリを元にした学習データを用いて Detectron2 による学習を行い，推論し，文書の構成要素を抽出した結果を評価した．評価方法は，確率があるしきい値以上のものを検出とみなし，検出した結果に対して適合率，再現率，F 値を算出した．

実験 2 では，先行研究と同様に，実験 1 で得られた結果の図表関連のカテゴリに対してトップダウン法を適用し，推論を行い評価した．評価方法は，実験 1 と同様に，確率があるしきい値以上のものを検出とみなし，検出した結果に対して適合率，再現率，F 値を算出する．

5.2 実験結果

5.2.1 実験 1

再定義したカテゴリを元に，Detectron2 の物体検出によって文書の構成要素を抽出した結果を示す（表 3）．実験で

は, 確率が 70%以上のものを検出とみなした. 結果として, 全体では, 平均適合率が 90.82%, 平均再現率が 88.85%, 平均 F 値は 89.64%であった. 先行研究と比較すると性能が大幅に向上しており, 一意なカテゴリの定義と文書の構成要素同士の重なりを避けた結果が精度向上へ繋がったと考えられる. カテゴリごとの検出精度を見ると, 先行研究で課題となっていた Caption の適合率が向上し, 検出可能となった.

表 3 先行研究 （再掲，左）実験 1 の検出結果（右）

カテゴリ	適合率(%)
PTitleV	85.71
PTitleH	24.21
PSegment	55.47
TitleV	78.46
TitleH	48.45
LeadV	81.14
LeadH	68.83
ParagraphV	83.21
ParagraphH	70.28
FSegment	57.22
Figure	15.11
Table	45.65
CaptionV	0
CaptionH	0
平均	50.98

カテゴリ	適合率(%)	再現率(%)	F値(%)
PTitle	89.84	92.88	91.33
PSegment	86.43	84.66	85.54
TitleV	89.79	80.60	84.94
TitleH	76.49	92.05	83.55
LeadV	89.25	91.56	90.39
LeadH	92.12	72.84	81.35
ParagraphV	97.80	97.99	97.90
ParagraphH	94.33	92.13	93.22
FSegment	86.83	90.38	88.57
Figure	95.83	95.61	95.72
Table	93.60	89.62	91.57
CaptionV	94.44	83.22	88.48
CaptionH	93.85	91.58	92.70
平均	90.82	88.85	89.64

5.2.2 実験 2

実験 1 で得られた結果を図表関連のカテゴリに対してトップダウン法を適用した後，評価した結果を示す（表 4）. 結果として図表関連のカテゴリが全体的に適合率が減少し，一方で再現率は向上した. 総合的な評価として F 値を参照すると，Table,CaptiopV が向上し，Figure は減少，CaptiopH の F 値は変わらなかった. 全体の平均では 0.11%向上した. このことから未検出以上に再検出を行ってしまい，誤検出が増えることで適用率が減少し，その一方で再現率の大きな向上がないことから，全体的な精度の向上は見られなかったと考えられる.

表 4 トップダウン法を適用した検出結果

	適合率(%)	再現率(%)	F値(%)
PTitle	89.84	92.88	91.33
PSegment	86.43	84.66	85.54
TitleV	89.79	80.60	84.94
TitleH	76.49	92.05	83.55
LeadV	89.25	91.56	90.39
LeadH	92.12	72.84	81.35
ParagraphV	97.80	97.99	97.90
ParagraphH	94.33	92.13	93.22
FSegment	86.83	90.38	88.57
Figure	95.51	95.90	95.70
Table	93.24	91.04	92.12
CaptionV	93.85	85.31	89.38
CaptionH	92.55	92.84	92.70
平均	90.62	89.24	89.75

6. おわりに

本研究では先行研究において課題であった文書の各構成要素の検出精度を向上を目的としカテゴリの見直し，再定義や階層構造を用いた精度向上手法の適用を行った.
実験 1 では再定義したカテゴリを元に，構成要素の重なりを避けたアノテーションを行うことで精度向上し，先行研究の課題であった，図表関連のカテゴリに対して精度が向上するという結果を得られた.
一方で，先行研究で使用された階層関係を利用する制度向上方法では，大きな効果は得られなかった.

謝辞

学習データの作成にご協力いただきました(株)正興電機製作所に感謝申し上げます. なお，本研究は JSPS 科研費 JP22K12740 の助成を受けたものです.

参考文献

[1]Yasutomo Kimura, Keiichi Takamaru, Takuma Tanaka, Akio Kobayashi, Hiroki Sakaji, Yuzu, Uchida, Hokuto Ototake and Shigeru Masuyama: Creating Japanese Political Corpus from Local Assembly Minutes of 47 Prefectures, ALR12, The COLING 2016 Organizing Committee, pp.78-85, 2016.
[2]乙武北斗，内田ゆず，高丸圭一，木村泰知: 構造化データ作成を 目的とした PDF 地方議会資料のテキスト抽出に関する分析， 第 37 回ファジィシステムシンポジウム講演論文集，pp.431- 436，2021.
[3]林侑生「Single Shot MultiBox Detector を用いた PDF 文書の領域分割に関する研究」福岡大学工学部卒業論文, 2024.3
[4]Liu, W. et al. (2016). SSD: Single Shot MultiBox Detector. In: Leibe, B., Matas, J., Sebe, N., Welling, M. (eds) Computer Vision – ECCV 2016. ECCV 2016. Lecture Notes in Computer Science(), vol 9905. Springer, Cham. https://doi.org/10.1007/978-3-319-46448-0_2
[5]https://github.com/facebookresearch/detectron2
[6]https://fastlabel.ai/