人工知能学会 インタラクティブ
情報アクセスと可視化マイニング研究会(第12回)
SIG-AM-12-10

レイアウト認識に基づく論文構成要素の抽出

Automatic Role Labeling of OCR Processed Scholarly Papers

岩月憲一 [1*]　加藤恒昭 [2]　山口和紀 [2]
Kenichi Iwatsuki[1], Tsuneaki Kato[2], Kazunori Yamaguchi[2]

[1] 東京大学教養学部
[1] School of Arts and Sciences, The University of Tokyo
[2] 東京大学大学院総合文化研究科
[2] Graduate School of Arts and Sciences, The University of Tokyo

Abstract: Components of scholarly papers bear roles such as title, body, itemization title, or figure. A role label enables advanced searching such as finding papers in which a specified keyword is used in a specified role. In this paper, we propose a fully automatic role labeling method for OCR processed scholarly papers. In the proposed method, we first identify components from the OCR processed images by reconstructing components from incorrectly recognized regions by OCR software. Next, we assign role labels to the components. Our experiment showed that the accuracy of the classification reached 94% in the best case.

1　はじめに

学術論文の紙面は，テキストや図表など，複数の構成要素から成り立っている．そして，それぞれの構成要素が，論文タイトル，著者，見出し，本文といった文書内役割を持っている．

論文検索を行う際に，論文全体を対象とした検索を行うよりも，ある特定の役割を持った構成要素に絞って検索を行いたいことがある．この検索意図に応えるためには，予め，論文から抽出した構成要素に対して，正しく文書内役割を特定しておく必要がある．例えば，節タイトルとそれに対応する本文を特定しておけば，ある特定の節に絞った検索が可能となる [3]．しかし，文書内役割は筆者と読者の共通認識に基づき暗黙的に与えられるものであり，論文データ中に文書内役割が直接記述されることはない．従って，構成要素に対し，明示的に文書内役割を特定することが問題となる．本研究は，学術論文の構成要素に対し，文書内役割を特定することを目的とする．

論文構成要素を得るためには，論文 PDF ファイルを直接 XML 等に変換するか，あるいは論文画像をスキャンし，OCR ソフトウェアで処理をするという方法がある．前者の場合，PDF 作成ソフトウェアの仕様により変換後のデータの性質が異なる場合がある [2]．また，デジタルデータから直接 PDF に変換されておらず，紙媒体をスキャンした PDF については，取り扱うことができない．以上の理由から，本研究では，論文構成要素を得る方法として，論文画像をスキャンし，OCR で処理をするという方法を採る．

OCR ソフトウェアで論文画像を処理すると，各構成要素と大まかに一致する矩形領域を得ることができる．この時，OCR の誤りによって，得られた矩形領域が実際の論文構成要素と一致しない場合があるので，これを修正する．その後，各構成要素に対して，文書内役割を割り当てる．

2　関連研究

先行研究について，(1) 対象とする論文のデータ，(2) 文書内役割特定の手法，(3) 特定する文書内役割の種類の 3 つの観点から述べる．

まず，対象とする論文のデータには，2 種類あり，PDF を変換し XML 等のデータとして得るもの [2, 5, 6, 7]，論文画像をスキャンし OCR で処理したデータを得るものに分かれる．

次に，文書内役割特定の手法には，大きく分けると 2 種類あり，1 つはルールベースによる手法，もう 1 つは機械学習による手法である．

ルールベースによる文書内役割特定には，Klink らが提案している手法がある．そこでは，文書には，あらゆる文書に共通する構造と，例えば学術論文のような特定の領域の文書に特化した構造の 2 種類があると

*連絡先：東京大学教養学部学際科学科
〒 153-8902 東京都目黒区駒場 3-8-1
E-mail: 0191699319@mail.ecc.u-tokyo.ac.jp

しており，構成要素に対しそれぞれルールを適用していく．前者では，ページ上部ならばヘッダ，中黒から始まっているテキストならば箇条書き，といったルールが適用される．後者では，表の上にはキャプションが来る，といったルールが適用される．これらのルールが構成要素に複数適用され，最終的に，最も可能性の高い文書内役割が与えられる [4].

機械学習による文書内役割特定としては，SVM を用いる手法 [2, 6] と CRF[2, 5, 7] を用いる手法の 2 種類がある．学習・推定に際して用いる素性には大きく分けて 2 種類ある．矩形領域の位置座標や大きさといったレイアウト情報から抽出される素性と，テキスト情報から抽出される素性である．いずれの先行研究も，両方の素性を用いている．増田らが用いている素性には他の研究とは異なり，形態素解析を利用して得たテキスト中の名詞や人名の割合が用いられている [6]．また，

Luong らは，処理を 2 段階に分けており，レイアウト情報とテキスト情報を利用して文書内役割を特定し，なおかつ節タイトル部分についてはテキスト情報のみを利用して節の種類を特定している [5].

最後に，特定する文書内役割の種類について述べる．先行研究には，論文全体を対象としているもの [4, 5, 6] と，一部のみを対象としているもの [2, 7] がある．前者については，タイトル，節タイトル，本文，図，表など 23 種類の役割を特定しているもの [5] から，論文タイトル，著者，本文，ヘッダ，フッタ，ページ番号の 6 種類の役割に留めているもの [6] もある．後者については，論文中の図表とそのキャプションのみを特定しているもの [2]，論文の書誌情報のみ（タイトル，著者，アブストラクト，キーワード）を特定しているもの [7] がある．

本研究では，PDF 作成ソフトに依存せず，かつスキャンされた論文にも対応するため，OCR ソフトウェアで処理したデータセットを用いることにする．

同様に OCR を用いている研究はあるが，いずれも，レイアウト情報の認識誤りを修正する手法については述べていない．しかしながら，Luong らが指摘するように，OCR の誤りによって文書内役割が適切に特定できない場合がある [5] ため，本研究では誤り修正法についても取り組んでいる．

文書内役割特定の手法としては，ルールベースのものは論文の種類によってルールを手動で調節する必要があるため採用せず，機械学習の方法を採ることにした．今回の問題を系列ラベリング問題として捉えることにより，矩形の列の順序を学習・推定のために用いることができるため，CRF を採用することとした．

文書内役割については，情報検索への応用を前提とするため，論文全体にわたる 19 種類を用意した．

3 提案手法

3.1 概要

文書内役割の特定にかかる一連の処理は次の手順で行う（図 1）．OCR で論文画像を処理することによって，論文の構成要素を，複数の矩形領域として得ることができる．この時，構成要素と対応して認識されることが望ましい矩形が，分割されて認識されてしまうという誤りが生じる．

このようにして得られた矩形領域を入力とし，まず矩形領域を一次元列に並べ替える．学術論文は，タイトル→著者→見出し→本文といった一定の順序を持っており，この順序の情報を有効に用いたい．そこで，二次元紙面上に配置された論文の構成要素を，論文を読む際の順序に一致するよう一次元列に並べ替える．そして，この順序の情報を用いて以降の処理を行う．

続いて，誤って分割された矩形領域の統合を行う．

最後に，各矩形領域に対応する文書内役割を与える．ここでは，先に与えた順序を利用して，CRF による文書内役割の学習・推定を行う．OCR ソフトウェアで得られる情報には，論文中のテキストそのもの（テキスト情報）に加え，位置座標，大きさなどのレイアウト情報も含まれている．しかし，テキストそのものを利用しようとすると，文字認識の誤りが大きく影響してしまう可能性がある．従って本研究では，レイアウト情報を主に用いることにする．

認識誤りについて，補足しておく．OCR ソフトウェアは，余白などの視覚情報を元に，まとまったテキストや図を 1 つの矩形領域として設定している．そのため，論文の構成要素と，この矩形領域は必ずしも一致しない．このような不一致には，2 つの種類がある．

2 段組の論文にあっては，1 つの節に属するテキストが 2 つの段にわたって記述されている場合，矩形領域としては 2 つに分割されて手に入ることになる．しかしながら，先述の通り読む順序で並べ替えを行うため，2 つのテキストは系列内で連続する．後の文書内役割特定で，2 つとも本文として認識されれば，2 つのテキストを統合し，1 つの節として抽出することが可能になるから，問題にはならない．

他方，図表中に文字を含む場合，これらの文字を認識してしまい，複数のテキストとして認識するため，1 つの図表が複数の矩形領域に分割されてしまう．このまま文書内役割の特定を行おうとしても，実際の図表と比べて，矩形領域の大きさが極端に小さくなってしまうために，図表として認識されにくくなる．このため，分割されている矩形を文書内役割特定の前に統合しておく必要があるのである．

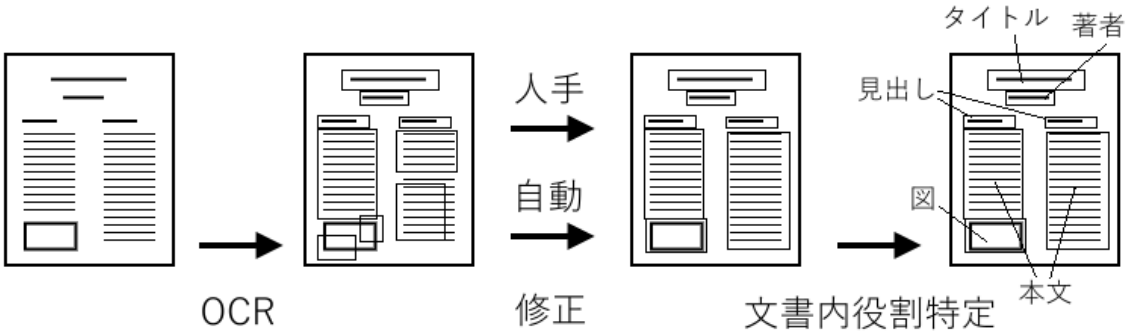


図 1: 処理の流れ

3.2 矩形の順序付け

2 次元平面上に配置された矩形領域を，1 次元の列に並べる手法について述べる．本研究では，論文を読む際の順序に矩形を並べるため，読む順序の特定が必要になる．

読む順序を特定する先行研究としては，Aiello らが次のような手法を提案している [1]．まず，任意の 2 つの矩形領域の座標の情報をもとに，ルールベースで，ありうる読む順序を全て求める．その後，各矩形のテキストに品詞タグを付与し，それぞれの順序ごとに，矩形の最後 2 単語の品詞と次の矩形の最初の 1 単語の品詞が共起する確率を計算し，最も確率の大きい順序に決定するというものである．しかしながら，OCR の文字認識の誤りがあること，および図表が複数の矩形に分割されていることから，各矩形のテキストに対し適切な品詞タグを付与することが困難となるため，本研究で対象としている文書には，この手法を適用することはできない．

本研究では，以下の観察から，ヒューリスティクスに基づいたルールベースで読む順序を特定し，順序付けを行うことにした．

学術論文は，本文部分が 2 段組となっている場合が多い．従って，左上→左下→右上→右下の順に並べ替えることになるが，大きな図表については，2 段にまたがっていることがしばしばある．この場合，図表よりも上の部分と下の部分のそれぞれで，左上→左下→右上→右下の順序に並べ替える必要がある．さらに，図表部分が小さく分割されている可能性があるので，これも考慮する．

3.3 矩形領域の誤り修正

構成要素と対応して 1 つの矩形領域であると認識すべきところ，複数に分割されている矩形を判別し，統合する．

まず，判別の手法であるが，SVM を用いて学習・推定を行った．カーネル関数には，Linear と Radial Basis Function(RBF) をそれぞれ用いた．

矩形の x 座標（横方向），矩形の y 座標（縦方向），矩形の幅，矩形の高さ，矩形内テキストの文字数，矩形の面積に着目し，それぞれ当該矩形の値，1 つ前の矩形との差分，1 つ後の矩形との差分を学習に用いる素性とした．ここでの 1 つ前，1 つ後というのは，前節で並べ替えた一次元列における順序である．合計 18 の素性を用いている．

SVM を用いて統合すべきであると判定された矩形を 1 つの矩形に統合したデータを用意する．統合後の矩形について，位置座標は，最も左上にある矩形のものを採用した．大きさについては，特定された全ての矩形が内部にあるような長方形のうち，最も小さいものを採用した．

3.4 文書内役割の推定

各矩形の文書内役割を学習・推定するために，CRF を用いた．今回特定した文書内役割は，謝辞，謝辞タイトル，著者，本文，数式，図，図キャプション，フッタ，脚注，ヘッダ，箇条書き，注釈，ページ番号，参考文献，参考文献タイトル，節タイトル，表，表キャプション，論文タイトルの 19 種類である．

用いた素性は，矩形の位置座標，矩形の大きさ，フォントサイズ，矩形内文字列の 1 文字目の文字種，矩形内文字列の行数である．位置座標と大きさについては，

1 つ前の矩形との差分も利用した．また，位置座標と大きさは 100px 毎に離散化してある．

4 実験と評価

4.1 概要

今回の実験に使用したデータセットについて述べる．言語処理学会年次大会発表論文集のうち，2003 年，2006 年，2009 年，2013 年の論文からそれぞれ 27，23，22，26 本ずつ，合計 98 本の論文を，OCR ソフトウェア [1] で処理し，データセットとした．年度によってスキャンの精度に差異があるため，複数の年度から採用した．

次に，矩形認識の誤りを人手で修正統合し，矩形認識誤りのないデータを作った上で，人手で文書内役割を特定することで，正しく文書内役割が割り振られている正解データを作成した．

今回の実験では，矩形の認識誤りの修正と文書内役割の特定という 2 つの段階があり，合わせて表 1 のように実験を行った．

表 1: 実験の種類

実験番号	誤り修正	文書内役割特定
(1)	人手	CRF
(2)	SVM	CRF
(3)	なし	CRF

認識誤りの修正は，その効果を検証するため，提案した SVM による修正に加え，人手で修正した場合と全く修正しなかった場合も行っている．

文書内役割の特定の結果を正解データと比較することによって，CRF による文書内役割特定の評価を行う．同時に，実験 (1)〜(3) の結果を比較することで，認識誤りの修正が，文書内役割特定に及ぼす影響について確かめる．

4.2 矩形領域の誤り修正

全ての矩形について，統合すべきであるか否かを人手でタグ付けしたデータを用意し，そのデータを用いて SVM[2] で学習・推定を行い，7 分割交差検定をしている．

実験の結果は表 2，3 の通りである．

その後の実験では，RBF カーネルで推定した結果を元に矩形の修正を行ったデータセットを用いている．

4.3 文書内役割の推定の評価法

人手で文書内役割を与えた正解データとの比較をすることで，各実験 (1)〜(3) に対する処理の評価を行うことにする．この場合，それぞれの文書内役割毎に，F 値を計算するのが自然である．

しかしながら，実験 (1)〜(3) の結果は，それぞれ認識誤りの修正の方法が異なるため，矩形列が異なっている．そのため，正解データと，矩形領域が一対一に対応しないために，単純に比較することができない．そこで，以下のような評価法を考える．

まず，正解データでは 1 つの矩形であるところ，実験データでは n 個の矩形に分かれてしまっている場合を考える．この時は，複数の矩形で 1 つの正解データの矩形に対応するので，F 値の計算時には，1/n 個の矩形として扱う（図 2）．

続いて，実験データでは 1 つの矩形となっているところ，正解データでは n 個の矩形が対応している場合を考える．この時は，実験データの矩形に付与されたラベルが，対応する n 個の矩形全てに付与されたものとして扱う（図 3）．

表 2: 矩形領域の誤り修正結果（Linear カーネル）

正解＼推定	分割される	分割されない
分割される	566	784
分割されない	265	5024

表 3: 矩形領域の誤り修正結果（RBF カーネル）

正解＼推定	分割される	分割されない
分割される	558	792
分割されない	250	5039

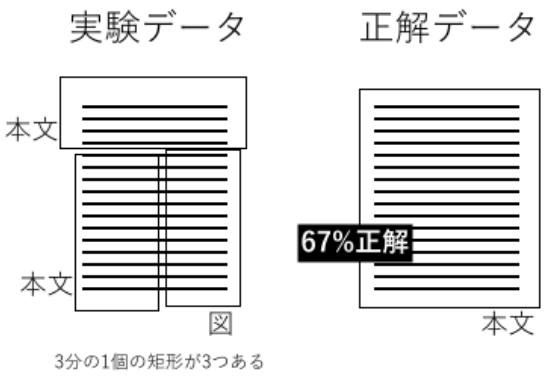


図 2: 矩形が分割された場合の正解率の計算

[1] メディアドライブ社の e.Typist を使用した．
[2] 実装には，libsvm (https://www.csie.ntu.edu.tw/∼cjlin/libsvm/) を用いた．

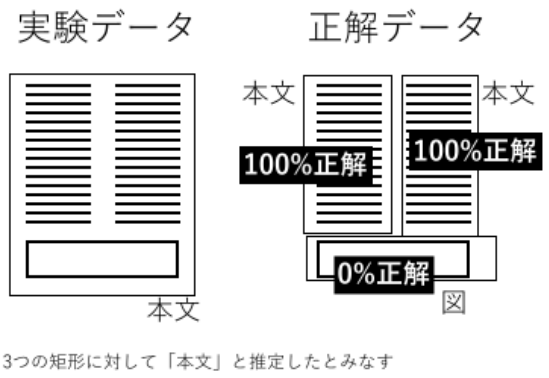


図 3: 矩形が統合された場合の正解率の計算

4.4 文書内役割の推定

CRF[3] を用いて，文書内役割の学習・推定を行った．正解データを用いて学習し，この学習結果を用いて文書内役割の推定を行った．すべて 7 分割交差検定を行っている．4.1 節で説明した 3 種類の実験の結果を表 4 にまとめた．

5 考察

5.1 矩形認識の誤り修正について

5.1.1 分割された矩形の統合が文書内役割特定にもたらす影響

矩形が分割されてしまう誤りを修正するためには，分割されている矩形を特定する処理と，特定した複数の矩形を統合し分割される前の矩形に戻す処理が必要となる．

表 4 によれば，人手で分割された矩形を統合している場合，文書内役割特定の正解率は高いことから，分割されている矩形を統合する必要があることが分かる．しかし，SVM を用いて分割された矩形を特定し統合した場合と，一切修正を加えていない場合の，文書内役割特定の正解率に差がないという結果になっている．この原因について考察を加える．

分割された矩形を統合することによる効果をみるために，分割された矩形から一定の割合でランダムに選ばれた矩形を統合した上で，文書内役割の正解率を計算するという実験を行った．

この実験の結果が表 5 である．全く統合していない場合と，全て統合した場合で，文書内役割特定の正解率に変化は見られない．これは，CRF による文書内役割の特定自体が，分割された矩形に対しても，統合された矩形に対しても，効果的でないことを示している．実際に，元々分割されていた矩形のみに注目した文書内役割の正解率を調べると，統合された矩形の割合に関係なく低い正解率となっている（表 6）．

分割された矩形を統合した後に付与された文書内役割を見ると，箇条書きとなっている矩形が多い．つまり，もともと図表数式であった矩形が分割され，統合されると，図表数式の特徴を失うということである．例えば，図 4 に示すように，元々の矩形と統合後の矩形では，位置と大きさが異なる場合が少なくない．従って，分割された矩形を統合する手法を改善する必要があると言える．

なお，表 5 の結果と，表 4 の SVM+CRF の結果に差異が見られないのは，SVM によって誤って分割されていないのに分割されていると判断された矩形のうち，修正を加えない場合に文書内役割が正解した矩形は 37 個（全体の 0.6%）であり，これらを統合してしまうことによる文書内役割特定の正解率の低下は非常に小さいためである．

図 4: 統合後の矩形

5.1.2 SVM による矩形認識の誤り修正

矩形認識誤りの修正において，レイアウト情報を素性とする SVM がどの定後有効か検討する．

SVM（Linear カーネル）によって求められた分離平面の成分を表 7 にまとめた．これによれば，矩形の幅と，矩形内テキストの文字数が判定に最も大きく寄与していると言える．実際には約 6 割の分割された矩形は特定に至っていない（表 2）．

図 5 は実際に分割された矩形の幅の分布であるが，分割された矩形の半分程度は比較的大きな矩形が占めている．図 6 は全矩形の幅の分布であるが，図 5 と比べると，比較的大きな矩形の大半は分割されていない矩形であり，これらを特定するには，大きさの情報だけでは足りないことが分かる．

[3]実装には，CRF++ (https://taku910.github.io/crfpp/) を用いた．

表 4: 実験結果

文書内役割	(1) 人手+CRF			(2)SVM+CRF			(3) なし+CRF		
	P	R	F	P	R	F	P	R	F
謝辞 (ack)	1.00	0.50	0.67	1.00	0.25	0.40	0.75	0.25	0.38
謝辞タイトル (ackt)	0.17	0.67	0.27	0.75	0.25	0.38	1.00	0.33	0.50
著者 (auth)	0.95	0.93	0.94	0.62	0.93	0.75	0.54	0.93	0.68
本文 (body)	0.88	0.91	0.89	0.59	0.73	0.65	0.57	0.72	0.64
数式 (equ)	0.75	0.77	0.76	0.06	0.01	0.02	0.05	0.01	0.02
図 (fig)	0.74	0.72	0.73	0.44	0.22	0.29	0.40	0.22	0.29
図キャプション (figc)	0.76	0.82	0.79	0.59	0.33	0.43	0.55	0.33	0.41
フッタ (ftr)	0.99	1.00	1.00	0.93	1.00	0.96	0.91	1.00	0.95
脚注 (ftnt)	0.82	0.71	0.76	0.55	0.59	0.57	0.54	0.60	0.57
ヘッダ (head)	0.97	1.00	0.99	0.91	0.86	0.89	0.90	0.87	0.88
箇条書き (list)	0.76	0.72	0.74	0.28	0.69	0.40	0.27	0.68	0.38
注釈 (note)	0.33	0.15	0.21	0.33	0.25	0.29	0.42	0.25	0.31
ページ番号 (page)	0.99	1.00	0.99	0.98	1.00	0.99	0.96	1.00	0.98
参考文献 (ref)	0.88	0.77	0.82	0.84	0.81	0.82	0.80	0.81	0.80
参考文献タイトル (reft)	0.76	0.84	0.80	0.68	0.49	0.57	0.64	0.50	0.56
節タイトル (stitle)	0.95	0.98	0.96	0.76	0.51	0.61	0.73	0.53	0.61
表 (tab)	0.80	0.83	0.82	0.68	0.41	0.51	0.59	0.42	0.49
表キャプション (tabc)	0.80	0.80	0.80	0.66	0.45	0.54	0.59	0.46	0.52
論文タイトル (title)	0.97	0.94	0.96	0.81	0.80	0.81	0.78	0.80	0.79
正解率		0.86			0.61			0.61	

表 5: 分割された矩形の正解率と文書内役割の正解率

分割されていると判定する確率	文書内役割の正解率
0	0.61
0.2	0.61
0.4	0.61
0.6	0.61
0.8	0.61
1	0.61

表 6: 分割された矩形に注目した文書内役割の正解率

分割されていると判定する確率	文書内役割の正解率
0	0.16
0.2	0.15
0.4	0.16
0.6	0.17
0.8	0.17
1	0.13

5.1.3 複数の矩形が統合されてしまう認識誤り

ここで，OCR による矩形認識の誤りについて，改めて述べる．認識誤りには，2 種類ある．1 つは，矩形が複数に分割されてしまうという誤りであり，もう 1 つは，複数の矩形が 1 つの矩形として認識されてしまうという誤りである（図 3）．これまで，前者に着目して修正を行ってきたが，後者の認識誤りの影響について考察する．

今回の実験では，正解データにおける約 3 割の矩形が，複数個で 1 つの矩形として認識されてしまっている．本来異なる文書内役割を持つ複数の矩形が 1 つにまとめられてしまうと，推定の段階で 1 つしか文書内役割を与えられないため，正解率が大きく低下することになる．

表 7: 分離平面の成分

成分	値
x 座標の差分（1 つ前）	0.08
x 座標	0.37
x 座標の差分（1 つ後）	-0.18
y 座標の差分（1 つ前）	0.35
y 座標	-0.15
y 座標の差分（1 つ後）	-0.85
幅の差分（1 つ前）	-0.24
幅	-2.74
幅の差分（1 つ後）	2.36
高さの差分（1 つ前）	0.10
高さ	-0.84
高さの差分（1 つ後）	-0.03
文字数の差分（1 つ前）	1.05
文字数	-4.95
文字数の差分（1 つ後）	2.05
面積	1.05

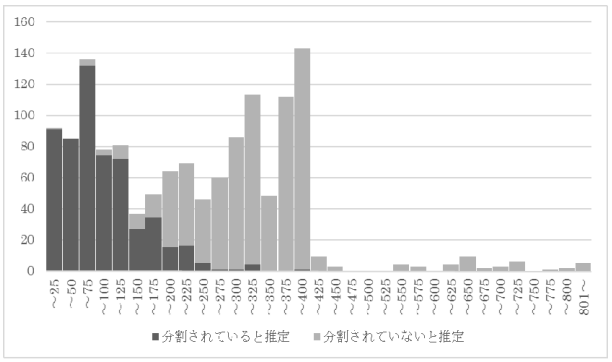
5.2 文書内役割の推定について

5.2.1 CRF を利用した文書内役割の推定

実験 (1) における，CRF による文書内役割特定における混同行列を表 9 に示した．

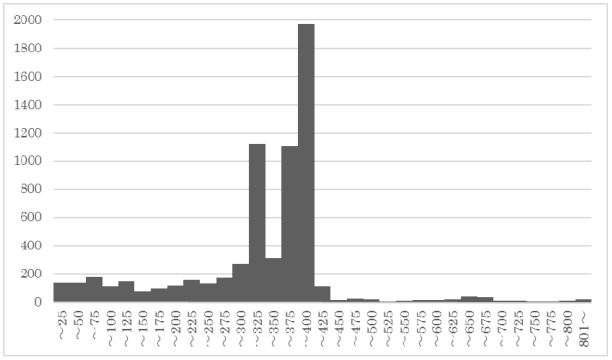
謝辞，謝辞タイトルの F 値（表 4）が低いのは，そもそも謝辞を掲載している論文数がデータセット中に少なく，十分な学習ができなかったためである．

最も目立つのは，本文と箇条書きの混同である．特に本文は最も矩形数が多いため，F 値 0.89 と他の矩形と比較しても目立って低いわけではないにも拘わらず，全体の正解率を下げている．この原因を考察する．



横軸は幅 [px]，縦軸は矩形数

図 5: 分割された矩形の幅の分布



横軸は幅 [px]，縦軸は矩形数

図 6: 全矩形の幅の分布

当該矩形の文書内役割が「本文」「箇条書き」となる素性関数の中で，重みが大きいもの上位 10 位のうち共通するものが 4 含まれている．これは，本文と箇条書きには，類似点が多いということを意味する．そのため，ある程度の混同が生じると思われる．

そもそも，箇条書きは本文部分に記述されるものであるから，位置座標の点では，本文と大差ない．また，大きさについても，複数の項目を列挙する箇条書きでは，ある程度の行数と幅を持ち，本文と大差ない．大きく異なるであろう部分は，箇条書きの場合，最初の文字が「・」や「(」などの記号になり得るという点である．実際に，重み 5 位に「最初の文字が記号」が来ている．逆に言えば，箇条書きと本文を区別するためには，レイアウト情報では足りず，行頭の記号だけでなく，それに続く文の長さ，体言止めかどうか，などのテキスト情報が必要になると思われる．

続いて，図，表，数式の混同を減らす方法について検討する．矩形認識の誤りがない場合であっても，F 値は高くない．また，混同行列を見ると，この 3 つはお互いに混同されていることが分かる．いずれも，レイアウト情報を考えると，ある程度の大きさを持った矩形になっていると考えられる．テキスト情報については，図の場合はグラフなどの文字列が認識される場合があり，表の場合は各セルの文字列が，数式はアルファベット・数字が認識される場合がある．これらの文字列の区別がつきにくい場合があると想像される．3 つの文書内役割を特定する際に用いられる素性関数の重みを調べたところ，大きいもの上位 1 位に来るのは，いずれも「フォントサイズが不明」である．文字列を認識しなかった際にもこの値が設定されるためである．他に上位に来ているものとしては，図・表の場合にはキャプションの情報がある．表については，表キャプションは通常表の上部に来るため，読む順序に並べ替えた後，表キャプション→表の順に矩形が並ぶので，表キャプションさえ特定できれば，表の特定は難しくない．実際に，表を特定する素性関数の 3 位には,「1 つ前の矩形が表キャプション」が来ている．ところが，図表は縦に連続して複数紙面に現れることがしばしばある．この場合，表キャプションと表が，あるいは図と図キャプションが交互に出現するため，図表とキャプションの上下関係が逆転されて学習される場合がある．実際に，図の特定にも，重み上位 3 位の素性関数に「1 つ前の矩形が図キャプション」を素性に持つものが来ている．図表数式の正解率を上げるためには，レイアウト情報，テキスト情報以外の情報，例えば一定の長さの曲線や直線の有無，数学記号の有無といった画像処理によって得られる情報などを用いる必要があると考えられる．

以上で本文と箇条書き，図表数式の混同について述べてきたが，検索意図と文書内役割という観点から，これらの混同の重要度について述べる．箇条書きは通常本文の一部と考えることができるため，これらを区別しない状況も十分にあり得る．また，図表に含まれる文字列や数値の情報のうち，重要な部分は本文中でも言及されることがあり，図表を区別して検索する必要がない場合も考えられる．これらを考慮し，箇条書きを本文として，図表数式を図として（表キャプションも図キャプションとする）文書内役割の特定を行った場合，正解率は表 8 の通りとなり，一定の上昇がみられる．

表 8: 文書内役割をまとめた場合の正解率

実験種別	正解率
人手+CRF	0.94
SVM+CRF	0.69
なし+CRF	0.70

表 9: 混同行列 (実験 (1))

		推定された文書内役割																		
		ack	ackt	auth	body	equ	fig	figc	ftr	ftnt	head	list	note	page	ref	reft	stitle	tab	tabc	title
正解	ack	6			6															
	ackt		8													3	1			
	auth			192	5							1				1	5			2
	body	1		1	1823			18		2		136			9	2	7	3	7	
	equ					151	36	1										7		
	fig					34	210	2										41	4	
	figc				8		3	221				10	1	3	1	2	1	1	19	
	ftr								103											
	ftnt				7			2		80		17		2			4		1	
	head										75									
	list		2		185			10		13		676	2	1	3	3	32		16	
	note				3			2				6	3	1			1	1	3	
	page													299						
	ref				7							20			96	1	1			
	reft				3			3				3			3	85	4			
	stitle			1	3			3		3		12		1			1083	1	1	
	tab					11	36											224		
	tabc			1	5		1	35				8		1		2	1		220	
	title			4							2									96

6 おわりに

学術論文の構成要素抽出においては，PDF データを直接用いるよりも，論文画像を OCR ソフトウェアで認識した方が，スキャンデータしかない論文も利用でき，PDF 作成ソフトウェアの使用にも左右されないという利点がある．しかしながら，OCR による認識誤りの修正をしなければ，十分に実用的であるとは言えない．

認識誤りには，論文構成要素が複数に分割されてしまうものと，複数の論文構成要素が 1 つにまとめられてしまうものが存在する．本研究では，SVM を用いて分割された矩形の統合を行った．矩形のレイアウト情報から，分割されているか否かを学習・推定したが，レイアウト情報のみでは判別が難しいことが分かった．また，複数の構成要素が 1 つにまとめられる誤りの修正を行わなければ，文書内役割の推定の正解率は，一切の認識誤りがない場合と比べて，大きく低下することが分かった．

矩形認識の誤りがなければ，レイアウト情報を主な素性として，CRF によって文書内役割を学習・推定する手法は，正解率 0.86 とおおむね良好な結果を出すことができた．本研究では，図，表，数式の混同ならびに本文，箇条書きの混同がみられたが，この混同を減らすためには，レイアウト情報に加え，テキスト情報や図形的情報を用いる必要があると言える．

参考文献

[1] Aiello, M., Monz, C., Todoran, L., Worring, M.: Document understanding for a broad class of documents, *International Journal on Document Analysis and Recognition*, Vol. 5, No. 1, pp. 1–16 (2003)

[2] 樫本達矢, 太田学, 高須淳宏: 学術論文からの構成要素抽出手法の改良, In *Proc. of The 7th Forum on Data Engineering and Information Management* (2015)

[3] 加藤恒昭, 岩月憲一, 山口和紀: 文書構造に基づく対話的情報アクセスにむけて, 人工知能学会 第 10 回インタラクティブ情報アクセスと可視化マイニング研究会, pp. 1–8 (2015)

[4] Klink, S., Dengel, A., Kieninger, T.: Document structure analysis based on layout and textual features, In *Proc. of International Workshop on Document Analysis Systems*, pp. 99–111 (2000)

[5] Luong, M., Nguyen, T.D., Kan, M.: Logical Structure Recovery in Scholarly Articles with Rich Document Features, *International Journal of Digital Library Systems*, Vol. 1. No. 4, pp. 1–23 (2012)

[6] 増田勝也, 丹治信, 植松すみれ, 美馬秀樹: 研究動向分析のための論文のデジタルテキスト化とマイニングシステム, 言語処理学会第 *20* 回年次大会発表論文集, pp. 792–795 (2014)

[7] Ohta, M., Inoue, R., Takasu, A.: Empirical evaluation of CRF-based bibliography extraction from research papers, In *Proc. of the IADIS International Conference on Information Systems*, Vol. 7, No. 2, pp. 18–26 (2012)