日本語論文に特化した PDF 文書解析器の構築と性能評価

嘉本名晋 [1]　梅澤悠河 [2]　長尾浩良 [2]　桂井麻里衣 [1]
[1] 同志社大学理工学部　[2] 同志社大学大学院 理工学研究科
{kamoto24，umezawa22，nagao21，katsurai}@mm.doshisha.ac.jp

概要

学術分野の PDF 文書解析のデファクトスタンダードである GROBID は主に英語で訓練されており，日本語論文の解析は困難であった．そこで，本研究では日本語論文を手動でアノテーションし，新たに GROBID モデルを訓練した．実験の結果，我々のモデルは既存モデルに比べて日本語論文に対する識別性能が向上し，特に非レイアウト関連の要素で優れた性能を示した．さらに，解析対象の論文 PDF と同一の収録誌を訓練データに含めることの有用性を明らかにした．また，収録誌ごとに識別性能を算出し，性能差が生じる原因を定性的に分析した．

1　はじめに

学術データベースの普及やオープンサイエンスの推進により，インターネット上の学術論文の量は大幅に増加している [1]．これらの膨大な論文をマイニングすることは，研究動向の把握や，異なる研究分野間の潜在的な関連性の特定に非常に有益である．多くの研究者は，学術データベースに収録されている書誌情報を分析に用いる．しかし，各データベースは，対象とする研究分野の論文情報を網羅しているとは限らず，特に，全文へのアクセスが制限されているデータベースが数多く存在する．知識抽出のための学術情報の網羅性を高めるためには，論文 PDF から直接テキストを抽出する必要がある．

学術論文のレイアウトは出版社や学会によって大きく異なるため，任意の論文 PDF から書誌メタデータやテキストを抽出することは容易ではなく，多くの手法が提案されてきた．中でも GROBID[1] と呼ばれるオープンソースモデルは高い認識性能を示しており [2]，学術分野における PDF 文書解析器のデファクトスタンダードになりつつある．しかし，GROBID は英語と少量のフランス語，ドイツ語の論文でしか訓練されていないため，非ラテン文字で書かれた PDF 文書の解析にはしばしば失敗することが報告されている [3]．

本研究では，日本語論文 PDF に特化した GROBID モデルを新たに構築した．モデルを訓練するために，日本語論文を収集し，それらのレイアウト要素を手動でアノテーションした．実験では，訓練用データセットを拡張した場合の性能を検証した．本研究により得られた知見はその他の非英語言語でのモデル訓練にも役立つと考えられる．

2　データセットの構築

本章では GROBID 訓練のための日本語論文 PDF データセットの構築方法について述べる．はじめに，様々な分野の学術誌から論文 PDF を収集することを考え，著者の研究分野を網羅する目的で学術データベース KAKEN[2] を利用した．具体的には，まず各研究者の科研費採択課題の審査区分を参照し，最多となる大区分をその研究者の専門領域とみなした．次に，11 種ある大区分それぞれから 9 名または 10 名の研究者をランダムに選択した．それらの研究者氏名を検索クエリとして J-Stage Web API[3] を通じて論文 PDF 一覧を取得した．得られた論文データのうち，クエリとした著者氏名が論文内に記載されていること，かつ 2 ページ以上の PDF であることを条件に，一つの研究者氏名につき 1 本の論文 PDF を手動で収集した．このとき，複数の論文を一つの PDF にまとめたものや，学会の議事録に該当するもの，内容が学術的でないものを手動で除外した．最終的に，98 種類の収録誌から 109 本の論文の PDF が収集された．

次に，将来的なデータセットの公開を見据え，J-STAGE 上で CC-BY または CC-BY-SA のライセンスが付与された論文 PDF を検索した．できるだけ多様なレイアウトの論文を収集するために，同じ収

1） https://github.com/kermitt2/grobid

2） https://kaken.nii.ac.jp/ja/

3） https://www.jstage.jst.go.jp/static/pages/JstageServices/TAB3/-char/ja

表 1　注釈の要素と説明

ラベル	説明
front	文書ヘッダー用
note place=“headnote”	ページヘッダー用
note place=“footnote”	ページのフッター及び番号付き注釈用
body	文書の本文用
listBibl	参考文献用
page	ページ番号用
div type=“acknowledgment”	謝辞用
div type=“availability”	データおよびコードの利用可能性声明文用
div type=“annex”	付録やその他の文章用
other	目視では確認できない文字列用

録誌の論文は 20 件までという収集上限を設定し，53 種類の収録誌から 404 本の論文 PDF を収集した．

各 PDF を GROBID を用いて XML 形式に変換し，論文から抽出した文字列に対して表 1 に示す要素ラベルを付与した．作業は同志社大学の大学院生および学部生 8 名によって行われた．アノテータ間の作業のばらつきを防ぐため，全てのデータを第二著者が目視でチェックし，必要に応じて修正を加えた．最終的に合計 148 種類の収録誌から 513 本の論文 PDF データセットを構築した．

3　実験

本章では，2 章の日本語データセットで訓練した GROBID モデルの性能を検証するために，英語論文で訓練された既存モデルとの比較実験を行う．また，データセットに対して様々な条件を設定し，それらがモデルの性能に与える影響を検証する．

GROBID はラテン語で記述された PDF 文書を対象として設計されているため，日本語で書かれた文書に適用するには内部モジュールの改変が必要である．特に，GROBID の既存のトークナイザは日本語の分かち書きに対応していなかった．そこで本研究では代用のトークナイザとして MeCab [4] を採用した．また，トークンが人名か，一般名詞かを判定する内部モジュールを MeCab を用いて日本語に対しても適切に動作するようにした．さらに，論文中のヘッドノートやフットノートの検出に向けて複数回登場する文字列パターンを記録して特徴量に変換する内部モジュールについても，日本語の文字列に対応できるよう書き換えた．

3.1　日本語論文 PDF の解析における既存モデルとの性能比較

2 章で構築したデータセットにおける 5 分割交差検証を通じて提案モデルと既存モデルの性能を評価した．モデルは conditinal random field （CRF）と BidLSTM-CRF [5] の 2 種類を採用した．GROBID は機械学習用に 32 種類の特徴量が設計されている．これには，トークンやトークンの形式（大文字，小文字，数字，イタリック，ボールド），フォントの変化，フォントの大きさの変化，文字のまとまりを表すブロック情報，人名や URL などのパターンと一致するかの辞書情報，文書内やページ内での位置情報，行の長さ，句読点の種類や数，文字列の出現パターンなどが含まれる．BidLSTM-CRF においては，これらの特徴量に加え，トークンを 1 文字ずつに分割したものをそれぞれ埋め込み層を用いてベクトルに変換する．この変換されたベクトルを時系列データとみなし，BidLSTM の入力とすることで，特徴量および単語ベクトルのそれぞれに対して 50 次元のベクトルを得る．これらの 100 次元ベクトルと，事前学習済みの単語埋め込みに基づくトークンの 300 次元ベクトルを結合し，BidLSTM に入力した．BidLSTM-CRF の実装では機械学習ライブラリである DeLFT[4] を利用し，日本語トークンのベクトル算出には日本語で事前学習された fastText [5] を用いた．

要素単位での性能評価結果を表 2 に示す．日本語モデル（CRF）と日本語モデル（BidLSTM-CRF）の比較では全ての要素の F1 スコアにて CRF が優れた性能を示した．BidLSTM-CRF は LSTM 層を含んでいるため，長期依存関係の学習が得意であるが，CRF よりも大量の訓練データが必要となる．そのため，今回構築したデータセットの規模が十分でなかった可能性がある．加えて，論文解析用に手動設計された特徴量が CRF の入力として効果的に機能したものと考えられる．

既存モデルと日本語モデル（CRF）の間で特に注目すべきは acknowledgement，annex，references での性能差である．これらの要素は他の要素，特に本文である body と区別することが難しく，正しく識別するためにはテキスト特徴が必要であると考えられる．例えば acknowledgement ではほとんどの場合「謝辞」またはそれに準じた文字列がセクションタイトルとしてつけられており，annex では「付録」ま

4）https://github.com/kermitt2/delft
5）https://fasttext.cc/docs/en/crawl-vectors.html

表 2　日本語論文に対する既存モデルと日本語モデルの性能

モデル	既存モデル			日本語モデル（BidLSTM-CRF）			日本語モデル（CRF）		
評価指標	**Precision**	**Recall**	**F1-Score**	**Precision**	**Recall**	**F1-Score**	**Precision**	**Recall**	**F1-Score**
acknowledgement	0.0	0.0	0.0	52.3	13.6	20.8	**54.2**	**26.7**	**35.8**
annex	25.3	6.6	10.4	37.4	1.8	3.2	**90.3**	**82.4**	**86.1**
availability	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
body	78.3	**98.7**	87.3	87.4	98.6	92.7	**92.9**	97.8	**95.3**
footnote	12.8	0.6	1.1	25.9	1.8	3.2	**64.0**	**22.8**	**33.5**
funding	0.0	0.0	0.0	0.0	0.0	0.0	**27.5**	**6.4**	**8.7**
header	58.7	13.7	22.1	56.4	42.3	48.0	**72.1**	**60.9**	**66.0**
headnote	55.4	9.4	16.0	47.4	30.0	36.6	**73.8**	**46.3**	**56.9**
other	0.0	0.0	0.0	31.2	4.3	7.5	**84.5**	**37.9**	**52.2**
page	77.2	39.5	52.2	63.5	40.2	48.9	**78.4**	**55.4**	**64.9**
references	86.2	18.1	29.8	87.0	52.5	65.3	**90.2**	**83.2**	**86.6**
toc	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
マクロ平均	30.8	14.6	17.1	41.4	24.2	27.7	**61.8**	**44.1**	**49.7**
加重平均	74.6	77.6	71.3	82.6	85.2	82.7	**90.4**	**91.0**	**90.4**

たは「付記」がそれにあたる．しかし，既存モデルでは訓練データに日本語の論文は含まれておらず，これらの関連単語を学習できなかったと考えられる．一方で，本文やページ番号はテキスト情報なしでレイアウト特徴のみから識別できることもあり，高性能ではないが既存モデルでも識別可能であったことがわかる．また，今回 availability や toc は全てのモデルで全く識別できなかった．これらの要素は収集したデータセット内にはほとんど現れず，識別に必要な特徴を学習できなかったと考えられる．

3.2　解析対象と同一の収録誌を訓練データに含めない場合の性能検証

GROBID は訓練データに含まれる学術論文のレイアウトやテキスト情報をもとに学習する．レイアウトは収録誌によって大きく異なるため，訓練データに含まれる収録誌に対する性能と含まれない収録誌に対する性能には差が生じる可能性がある．そこで，モデルを CRF に限定し，解析対象と同一の収録誌を学習したモデルと学習していないモデルの性能を比較した．

まず，データセット内で PDF が 10 件に満たない収録誌のデータ 195 件を第 1 の訓練データとした．次に，第 1 の訓練データに含まれない収録誌のデータ 318 件を収録誌ごとに訓練用・テスト用へと半分ずつに分割し，これに第 1 の訓練データを加えたものを第 2 の訓練データとした．最終的な内訳は，第 1 の訓練データが 195 件，第 2 の訓練データが 348 件，テストデータが 165 件となった．この場合，テストデータからみて同一収録誌のデータを含むのが第 2 の訓練データ，含まないのが第 1 の訓練データとなる．

テストデータにおける各要素認識の F1 スコアを表 3 に示す．訓練データにテストデータと同一の収録誌を含めたモデルがマクロ F1 スコアで 0.17 ポイント上回った．要素ごとでは footnote，headnote，toc が大きく性能を上げている．これらの要素は各収録誌がユニークなレイアウトや文字列を持つことが多く，同一誌からの論文を訓練データに含めることが重要となる．一方で，レイアウトの差に影響を受けにくい references や body では大きな性能の向上は見られなかった．また，今回構築したデータセット内に少数のデータしか含まれなかった funding は識別できなかった．

3.3　収録誌単位での性能検証

学術論文 PDF のレイアウトは収録誌によって様々であり，解析の成功例と失敗例には大きな差が生じうる．そこで，解析対象と同一の収録誌を含めるか否かがどれほど性能に影響が与えるかを収録誌ごとに調査し，レイアウトの性質が与える影響を定性的に分析した．

表 3　評価対象と同一の収録誌を学習済みの CRF モデルと未学習 CRF モデルの F1 スコア

	未学習モデル	学習済みモデル
acknowledgement	61.5	**67.3**
annex	35.5	**41.0**
body	94.4	**95.0**
footnote	0.0	**52.9**
funding	0.0	0.0
header	85.1	**92.8**
headnote	56.4	**83.6**
other	0.0	**0.2**
page	73.1	**83.6**
references	93.1	**94.1**
toc	0.0	**67.8**
マクロ平均	45.4	**61.7**
加重平均	88.1	**90.3**

論文 PDF が 10 件以上の収録誌が 20 種類あり，それぞれの収録誌のデータを訓練用・テスト用へと半分に分割した．対象の収録誌以外の全てのデータを第 1 の訓練データとし，それらに先ほど分割した訓練用のデータを足したものを第 2 の訓練データとみなした．この場合も前節の実験と同様，テストデータからみて同一収録誌のデータを含むのが第 2 の訓練データ，含まないのが第 1 の訓練データとなる．最後に，テスト収録誌ごとに 2 種類の訓練データそれぞれで CRF モデルを構築した．

各モデルによるテストデータの識別性能を表 4 に示す．表 4 の資料グループコードとは，J-STAGE で用いられている収録誌の識別コードを指す．ほぼ全ての収録誌で性能が向上した．大きく性能が向上した収録誌に nivr，kokyurinsho がある．これらの収録誌はヘッドノートやページ番号の形式が特徴的であり，訓練データに同一誌のデータを含めたことが有効であったと考えられる．また，高い識別性能を示した jacn，digrajproc，seitai に関しては，一般的な 2 段組のレイアウトであり，訓練データ内に似た論文が大量にあったことが識別の容易さにつながったと考えられる．反対に，論文全体のレイアウトが極めて特徴的な収録誌や，3.1 節で性能の乏しかった要素（annex, footnote, funding）を多く含む収録誌は性能があまり上がらなかった．これは，拡張した訓練データであってもサイズが小さすぎて学習が困難だったためと考えられる．

表 4　テスト論文誌と同一誌からの論文 PDF を訓練データに含まない場合の CRF モデル（未学習モデル）と含む場合の CRF モデル（学習済みモデル）によるマクロ F1 スコア

資料グループコード	未学習モデル	学習済みモデル
sasj	50.0	**61.9**
jacn	79.0	**83.9**
digrajproc	80.1	**81.7**
jptpr	43.5	**50.2**
kokyurinsho	40.4	**71.1**
bmfm	54.6	**77.8**
tetsutohagane	57.1	**62.4**
ktsk	56.2	**67.1**
jrehabilneurosci	55.6	**77.6**
cicommun	59.3	**59.7**
iuws	71.3	**73.7**
jptp	62.2	**69.2**
msjtmsj	54.8	**78.6**
thermoelectrics	60.3	**72.7**
jjccpt	70.6	**79.8**
kinoushi	43.4	**49.7**
jjm	45.7	**52.0**
aaostrans	57.5	**77.2**
seitai	**82.8**	82.3
nivr	26.9	**56.2**

4　おわりに

本研究では手動でアノテーションを施した日本語論文を用いて，新たに GROBID に基づく PDF 文書解析モデルを訓練し，既存のモデルと性能を比較した．実験の結果，既存モデルでは識別が困難であった要素に対する識別性能の向上を示した．これにより，解析対象とする論文と同一の言語でモデルを訓練することの有用性が示唆された．また，訓練データの条件を変えることで GROBID の性能に与える影響を調査した．解析対象の収録誌と同一の収録誌を訓練データに含めることで，識別性能が向上することが明らかになった．さらに，解析対象とする収録誌により大きな差が生じた原因としては，特色のあるレイアウトや出現率の低い要素についての訓練データの少なさが挙げられた．今後は識別が困難であった要素に対しても高い認識性能を持つモデルを構築するために，データセットの拡張や内部モジュールの改良を検討する必要がある．

謝辞

本研究の一部は，2024 年度国立情報学研究所公募型共同研究（24FC05）の助成によって行われた．

参考文献

[1] Jeong-Wook Seo. Changes in the absolute numbers and proportions of open access articles from 2000 to 2021 based on the web of science core collection: a bibliometric study. **Science Editing**, Vol. 10, No. 1, pp. 45–56, 2023.

[2] Norman Meuschke, Apurva Jagdale, Timo Spinde, Jelena Mitrović, and Bela Gipp. A benchmark of pdf information extraction tools using a multi-task and multi-domain evaluation framework for academic documents. In **International Conference on Information**, pp. 383–405. Springer, 2023.

[3] Igor Shapiro, Tarek Saier, and Michael Färber. Sequence labeling for citation field extraction from cyrillic script references. In **SDU 2022: Scientific Document Understanding 2022; Proceedings of the Workshop on Scientific Document Understanding; co-located with 36th AAAI Conference on Artificial Inteligence (AAAI 2022); Remote, March 1, 2022. Ed.: AP Ben Veyseh**, 2022.

[4] Taku Kudo, Kaoru Yamamoto, and Yuji Matsumoto. Applying conditional random fields to Japanese morphological analysis. In Dekang Lin and Dekai Wu, editors, **Proceedings of the 2004 Conference on Empirical Methods in Natural Language Processing**, pp. 230–237, Barcelona, Spain, July 2004. Association for Computational Linguistics.

[5] Guillaume Lample, Miguel Ballesteros, Sandeep Subramanian, Kazuya Kawakami, and Chris Dyer. Neural architectures for named entity recognition. In Kevin Knight, Ani Nenkova, and Owen Rambow, editors, **Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies**, pp. 260–270, San Diego, California, June 2016. Association for Computational Linguistics.