ニューラルネットワークを用いた表構造解析手法の検討

Examination of Table Structure Recognition Methods Using a Neural Network

青柳 拓志

Hiroyuki Aoyagi

岡山大学 太田研究室

Ohta Laboratory, Okayama University

概要 学術論文では，実験結果を示すのに表が頻繁に用いられる．しかし，数値の差異や変化を視覚的に読み取るにはグラフの方が適している．そのため，山田らは表からグラフを自動生成するための表構造解析手法を提案した．具体的には，引かれていない罫線である補助罫線を予測するニューラルネットワーク (NN) モデルと，推定した補助罫線などの特徴量を入力としてセルを生成する NN モデルを用いて表構造解析を行った．本稿では，この補助罫線を予測する NN モデルの特徴量に新たな特徴量を追加し，その効果を検証した．

1 はじめに

近年，学術論文データベースの発展により学術論文を容易に入手できるようになった．学術論文では，実験結果を表すのに表やグラフがしばしば用いられる．表は正確な数値を読み取るのに優れており，グラフは数値の差異や変化を視覚的に読み取るのに優れている．そのため，例えば膨大な実験結果を一度に効率よく比較するにはグラフが適している．表からグラフを自動生成するには，表構造を解析することが必要である．山田ら [1] は，トークンの位置関係に基づいて補助罫線の有無を推定するニューラルネットワーク (NN) モデルと，隣接セルの結合を推定する NN モデルを用いた表構造解析手法を提案した．なお，トークンは基本的に表中の単語に対応する．本稿では，この補助罫線推定モデルに入力する特徴量に，補助罫線候補に関係するトークンの分散表現を追加し，その効果を検証した．

2 本稿で対象となる表の構成要素

図 1 に本稿で扱う表の例を示す．図 1 中で，赤い矩形で囲まれたものをトークンと呼び，罫線または補助罫線で囲まれているものをセルと呼ぶ．トークンは，文書 PDF 中の表を pdfalto[1] により XML 化して得られ，表中の単語であることが多い．なお，図 1 で実線が罫線，点線が補助罫線である．さらに，青色でハイライトした列をヘッダ列，黄色でハイライトした行をヘッダ行と呼ぶ．ヘッダ列は各行の名前を表し，ヘッダ行は各列の名前を表す．また，ヘッダ行ではないがデータの種類を区別する際などに使用される行をサブヘッダ行と呼び，それら以外の数値などをデータと呼ぶ．

	Method 1	Method 2	Method 3
Dataset A	0.92	0.83	0.82
Dataset B	0.90	0.87	0.83
Dataset C	0.96	0.92	0.89

図 1: 表と表の構成要素

3 山田らの表構造解析手法

山田ら [1] の表構造解析手法の概要を図 2 に示す．山田ら [1] の表構造手法は大きく 4 つの処理に分けられる．図中の 1 は，文書 PDF 中の表の XML ファイルへの変換と表の罫線検出である．XML ファイルへの変換には pdfalto を用いる．また，罫線の検出には PFMiner[2] と OpenCV[3]を用いる．図中の 2 は，補助罫線推定である．トークンと罫線の特徴を用いて Implicit Ruled Line Identifier(補助罫線推定モデル) により補助罫線を推定する．図中の 3 は，セル結合である．トークンと罫線と補助罫線の特徴を用いて Cell Generator(セル結合推定モデル) により，隣接する 2 トークンを結合すべきかを推定し，セルを生成する．図中の 4 は，後処理である．行や列を結合し，セルの拡張を行い，表構造を決定する．

4 補助罫線推定モデル

4.1 概要

本稿の補助罫線推定モデルは，トークンの特徴とトークンのテキストの分散表現を用いて，表中のトークンの位置関係に基づき補助罫線を推定する．本稿の補助罫線推定モデルを図 3 に示す．黄色でハイライトした箇所は，本研究で山田ら [1] のモデルに新たに追加した部分である．以後これを検討手法と呼ぶ．図 3 に示され

[1] https://github.com/kermitt2/pdfalto

[2] https://github.com/pdfminer/pdfminer.six

[3] https://opencv.org

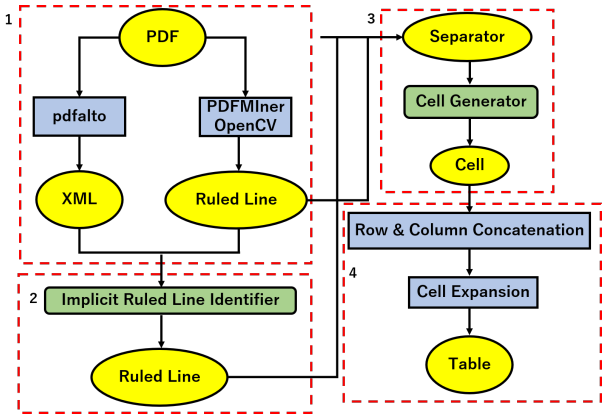

図 2: 山田らの表構造解析手法の概要 [1]

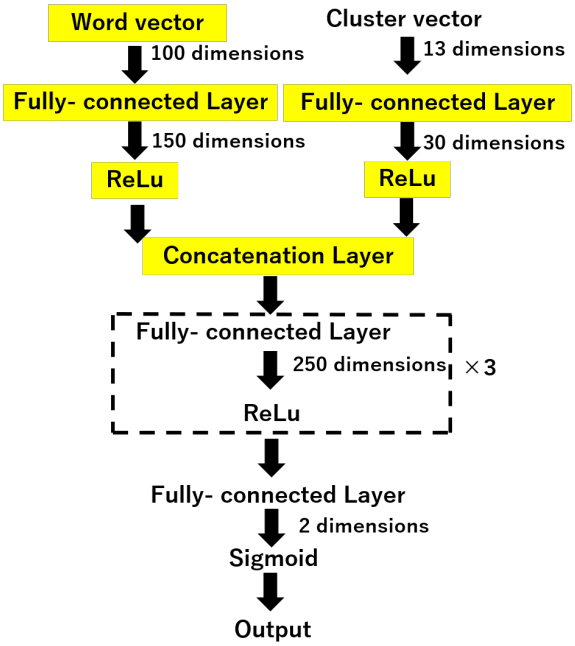

図 3: 検討手法での補助罫線推定モデル

た本稿のモデルへの入力は，補助罫線候補の 13 次元の特徴ベクトルと補助罫線候補中のトークンの 100 次元の分散表現を平均したものである．図 3 では，補助罫線候補の特徴ベクトルは Cluster vector であり，補助罫線候補の分散表現は Word vector である．また，出力は補助罫線か非補助罫線かの 2 値である．中間層の出力次元数は，結合層より前では分散表現を入力とする層は 150，クラスタの特徴量を入力とする層は 30 とし，真ん中の結合層より後では 250 とした．出力層の活性化関数は Sigmoid 関数，それ以外の層は ReLu を用いた．また，損失関数には 2 値クロスエントロピーを用い，最適化関数には Adam，学習率は 0.01 とした．

4.2 補助罫線推定モデルの入力特徴量

山田ら [1] の手法と同様に，補助罫線候補として，垂直方向の補助罫線候補はトークンの左端，右端，水平方向に隣接する 2 トークンの重心の中点の x 座標，水平方向の補助罫線候補はトークンの上端，下端，垂直方向に隣接する 2 トークンの重心の中点の y 座標のそれぞれ 3 種類の点集合を作成し，それぞれの集合の中で重心法を用いてクラスタリングして得られた点集合のクラスタを用いる．補助罫線候補の特徴量は，このクラスタを構成する点の数，表中の水平（垂直）方向に並んだトークン数，補助罫線候補を挟むトークン間の罫線の有無，補助罫線候補の方向，クラスタの種類，補助罫線候補がセル上を通るか，補助罫線候補の位置と表のサイズとする．また，先に述べたクラスタの種類は 6 次元の one-hot ベクトルとする．一方，山田ら [1] の補助罫線推定モデルの入力特徴量に新たに加える補助罫線候補に関するトークンの分散表現は，そのクラスタを構成する各トークンの分散表現を平均したものとする．分散表現は word2vec により得られ，word2vec の学習には Lipzig Copora[4] より English News(2016) と表を含む文章 PDF を pdfalto によりテキスト化したものを用いる．

5 補助罫線の推定実験

5.1 実験の概要

実験では，テストデータとして ICDAR2013 Table Competition にて提供されたテスト用データセットを使用する．このデータセットは，EU，米国政府の発行した様々なドメインの文書 PDF から表を収集したものである．また，学習データとして 81 の文書内の 209 の表から得られた 25,656 のクラスタ (補助罫線候補) を使用する．また，それらのクラスタの中には補助罫線が引かれるものが 21,146，引かれないものが 4,510 であり不均衡であったため，SMOTEENN[2] を用いて補助罫線と非補助罫線の個数を同数にする．

5.2 実験結果

表 1 に山田らの手法による補助罫線推定結果，表 2 に本稿の検討手法による補助罫線推定結果を示す．また，表 3 に山田らの手法による補助罫線推定の再現率，適合率と F 値，表 4 に本稿の検討手法による補助罫線推定のそれらを示す．つづいて図 4 に水平方向のクラスタ (補助罫線候補) を構成する点の数ごとの補助罫線推定結果の F 値，図 5 に垂直方向のクラスタを構成する点の数ごとの補助罫線推定結果の F 値を示す．表 3,

[4] http://wortschatz.uni-leipzig.de/en/download/

表 1: 山田らの手法の補助罫線推定結果

		推定	
		非補助罫線	補助罫線
正解	非補助罫線	4,708	465
	補助罫線	1,699	18,949

表 2: 検討手法の補助罫線推定結果

		推定	
		非補助罫線	補助罫線
正解	非補助罫線	4,203	970
	補助罫線	1,112	19,536

表 3: 山田らの補助罫線推定の再現率，適合率と F 値

	再現率	適合率	F 値
非補助罫線	0.91	0.73	0.81
補助罫線	0.92	0.98	0.95

表 4: 検討手法の補助罫線推定の再現率，適合率と F 値

	再現率	適合率	F 値
非補助罫線	0.81	0.79	0.80
補助罫線	0.95	0.95	0.95

表 4 より，本稿の検討手法による補助罫線の再現率と非補助罫線の適合率がそれぞれ山田らの手法の結果を上回ったが，非補助罫線の再現率と補助罫線の適合率は下回った．なお，再現率，適合率は下記の式で算出した．

$$\text{(非) 補助罫線の再現率} = \frac{\text{推定結果の正しい (非) 補助罫線数}}{\text{正解データの (非) 補助罫線数}}$$

$$\text{(非) 補助罫線の適合率} = \frac{\text{推定結果の正しい (非) 補助罫線数}}{\text{推定した (非) 補助罫線数}}$$

5.3 考察

検討したモデルは，山田らの手法と比較して補助罫線の再現率が高くなったこと，非補助罫線の再現率が低くなったことから，相対的に補助罫線候補を補助罫線とみなしやすいモデルである．また，図 4 では水平方向の補助罫線候補の点の数が多くなるにつれて非補助罫線の F 値が高くなるのに対して，図 5 では垂直方向の補助罫線候補の点の数が多くても非補助罫線の F 値が高くならないことが分かる．これは，点の数が比較的多く補助罫線ではない垂直方向の補助罫線候補が存在したことが原因である．例えば，図 1 中の “Dataset” の右端から生成される補助罫線候補は補助罫線ではなく，点の数が 3 である．この点の数は，図 1 の表における垂直方向の補助罫線候補の点の数が，最大で 4 であることを考慮すると比較的多い点の数である．このような点の数が比較的多く補助罫線でない垂直方向の補助罫線候補が存在するのは，おおよそ単語がトークンに対応するためである．

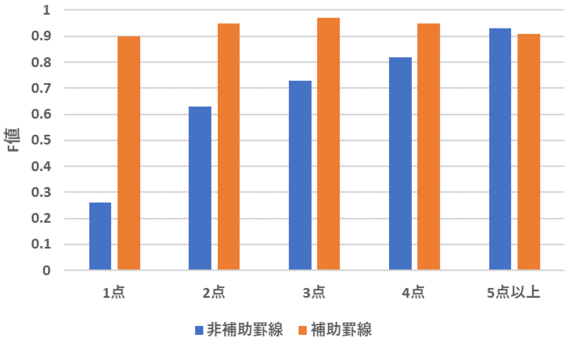

図 4: クラスタの点の数ごとの水平方向の補助罫線推定結果の F 値

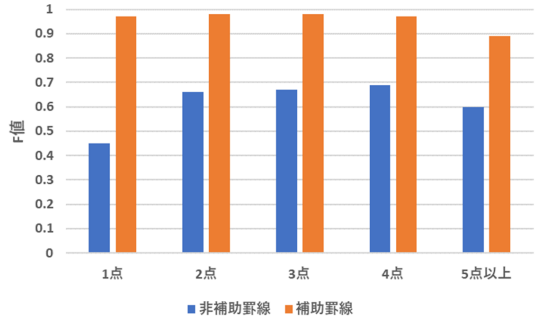

図 5: クラスタの点の数ごとの垂直方向の補助罫線推定結果の F 値

6 まとめ

本稿で検討した補助罫線推定モデルは，山田らの手法と比較して相対的に補助罫線候補を補助罫線とみなしやすいモデルとなった．また今後の展望として，複数の単語から構成される語を事前に一つのトークンにまとめることを考えている．

参考文献

[1] 山田凌也，太田学，金澤輝一，高須淳宏，“機械学習を用いた表構造解析の一手法,” 第 12 回データ工学と情報マネジメントに関するフォーラム，E6-4, 2020.

[2] G. Batista, R. Prati, and M. Monard, “A study of the behavior of several methods for balancing machine learning training data,” SIGKDD Explorations, Vol. 6, pp. 20-29, 2004.