様　式　Ｃ－１９、Ｆ－１９－１、Ｚ－１９　（共通）

科学研究費助成事業　　研究成果報告書

令和　５ 年　 ６ 月 １３ 日現在

機関番号：１５３０１
研究種目：基盤研究(C)（一般）
研究期間：2018～2022
課題番号：１８Ｋ１１９８９

研究課題名（和文）学術論文のためのコストセンシティブ情報抽出とサイバーフィジカル閲覧支援

研究課題名（英文）Cost-Sensitive Information Extraction and Cyber-Physical Browsing Support for Academic Papers

研究代表者

太田　学（Ohta, Manabu）

岡山大学・自然科学学域・教授

研究者番号：１０３２６０１９

交付決定額（研究期間全体）：（直接経費）　　3,400,000円

研究成果の概要（和文）：本研究では、学術論文からその参考文献の書誌情報をコストセンシティブに抽出する方法を二つ提案し、実験によりその抽出器の学習データ量と抽出精度の関係を定量的に明らかにするとともに、能動学習や擬似学習データを利用することで、学習データが削減できる見通しを得た。さらに、表の構造解析手法とそれを利用した表の数値データの自動グラフ化を提案するとともに、サイバーフィジカル論文閲覧支援として、論文中の引用箇所の関連情報をサイバー空間から自動集して論文閲覧者に提供するサービスを考案した。

研究成果の学術的意義や社会的意義
本研究で提案した参考文献書誌情報をコストセンシティブに抽出する技術は、電子図書館等において学術論文の書誌情報を整備する際に利用できる非常に有望な技術となっている。また提案した表構造解析手法は、近年提案された手法と比べて遜色のない表構造解析精度を達成している。一方、タブレット端末のカメラを通して紙の学術論文を読む読者へのサイバーフィジカル論文閲覧支援は、ウェアラブル端末を利用した近未来の読書のフィージビリティスタディとなっている。

研究成果の概要（英文）：In this study, we proposed two methods to extract bibliographic information from academic papers' references in a cost-sensitive manner. Through experiments, we quantitatively demonstrated the relationship between the amount of training data for the extractor and the extraction accuracy. We also explored the potential of reducing training data by using active learning and pseudo-training data. Furthermore, we proposed a method to analyze table structures and automatically graph numerical data within tables. Additionally, as a cyber-physical paper browsing support, we devised a service that automatically collects relevant information from the cyberspace on citations in papers and makes it available to paper readers.

研究分野：情報工学

キーワード：電子図書館　学術論文　情報抽出　メタデータ　閲覧支援　サイバーフィジカル　表構造解析

様　式　Ｃ－１９、Ｆ－１９－１、Ｚ－１９（共通）

１．研究開始当初の背景

電子図書館のみならず大学等でも機関リポジトリの構築が進むなど、インターネットアクセス可能な情報アーカイブが分散的かつ組織的に整備されるようになった。利用者が電子文書へ効率よくアクセスするには、書誌情報等の文書のメタ情報の整備とその組織化が不可欠であるが、良質のメタ情報が付与された電子文書の作成技術は成熟したとは言い難い。メタ情報付与のためそれまでに機械学習を利用した文書からの情報抽出器が提案されたが、人手による正解ラベル付き学習データの生成コストならびに抽出誤りに対する事後の修正コストの検討が十分とは言い難かった。

Google Glass のようなウェアラブル端末を使えば、拡張現実の世界で紙の論文を快適に読むことができるか？という素朴な疑問がある。論文を読む際に助けとなる情報、例えば引用文献の情報や専門用語の解説などは、通常サイバー空間にあるが、それをうまく見つけられるかどうかは読者の知識や経験に依存し、また発見には手間もかかる。特に初学者の場合、発見のコストが大きく、初学者ゆえにうまく発見できない可能性もある。そのため、例えば予めこれらの情報をウェブから抽出し、引用箇所や専門用語の理解を助ける補助情報を生成して読者に提供できれば有用な論文閲覧支援となる。

２．研究の目的

本研究は、学術論文の電子文書から様々なメタ情報をコストセンシティブに抽出する方法の開発と、抽出したメタ情報を利用したタブレット端末による新しい論文閲覧スタイルの提案を目的としている。学術論文の参考文献は、引用解析や文書間リンク生成などに利用されるため、その書誌情報を整備することは大変重要である。そこで本研究ではまず、(1)論文の参考文献欄からメタ情報としてその書誌情報をコストセンシティブに抽出する技術を研究開発する。また、学術論文から実験結果などがまとめられた表の数値データを自動抽出できれば、グラフへの自動変換など有用な閲覧支援サービスが提供できる。そのため本研究では、(2)機械学習による表構造解析手法を提案する。さらに本研究では、(3)抽出した論文メタ情報ならびにタブレット端末のカメラと OCR（文字認識技術）を利用して、印刷されたフィジカルな学術論文を、サイバー空間の情報で補いながら読むことができる論文閲覧支援を提案する。

３．研究の方法

(1)参考文献書誌情報抽出

参考文献書誌情報抽出では、多様な雑誌論文の参考文献文字列から低コストで高品質な書誌情報を抽出することが目標である。例えば、出版社名や論文誌名の辞書は参考文献書誌情報抽出において有用であるが、辞書は一般にその整備コストが高くその点が問題である。そのため本研究では、このような高価な特徴量や特徴量エンジニアリングによらない汎用性のある抽出器を開発する。

(2)表構造解析

数値が記載された表は、学術論文では実験結果のような重要な情報がまとめられていることが多く、このような表データを自動解析できれば、その数値の可視化など様々な閲覧支援が行える。一方、表の罫線の引き方やセルの構成など表の構造は著者によって異なる。そのため、機械学習を用いて表の構造を自動解析する方法を研究し、合わせてその解析結果を利用して表中の数値データをグラフ化する方法についても検討する。

(3)サイバーフィジカル論文閲覧支援

タブレット端末が普及しても紙に印刷された論文を読む機会は多く、またその関連情報は論文から抽出したメタ情報を利用してウェブから収集できる。そのため本研究では、タブレット端末のカメラと OCR を用いて、紙媒体の学術論文の閲覧支援を行う論文閲覧支援インタフェースを研究する。また具体的な閲覧支援として、読者が閲覧している論文中の引用箇所に提示するのにふさわしい補助情報の自動生成法を研究する。

４．研究成果

(1)参考文献書誌情報抽出

学術論文の参考文献欄から低コストで書誌情報を抽出するため、ニューラルネットワーク(NN)と Conditional random field (CRF) のハイブリッドな抽出器 (Bi-directional LSTM-CNN-CRF 抽出器)を開発した。この抽出器は整備コストのかかる辞書などを必要としないにもかかわらず、綿密に特徴量設計がされた従来の CRF 抽出器と同様に高精度に書誌情報が抽出できた。例えば、電子情報通信学会の英文論文誌の論文からの参考文献書誌情報抽出の実験では、抽出精度は約93%だった。本研究ではまた、この抽出器の学習データ量と書誌情報抽出精度の関係を実験により明らかにするとともに、書誌情報抽出精度を維持したまま学習データを削減する方法を検討した。実験の結果、書誌情報抽出結果に定義した確信度を利用して、抽出が困難な参考文献文字列を優先的に学習する能動サンプリングを行うことで、少量の学習データで高精度に抽出できることを示した。

本研究ではさらに、この抽出器のモデルを、様々な自然言語処理タスクにおいて成果をあげている BERT に置き換えた抽出器（BERT 抽出器）を提案した。この BERT 抽出器も、Bi-directional LSTM-CNN-CRF 抽出器と同様に煩雑な特徴量設計は不要で、かつ書誌情報抽出精度はそれよりも高くなることを実験により確認した。またこの抽出器による書誌情報抽出実験でも、学習データ量と書誌情報抽出精度の関係を明らかにするとともに、書誌情報抽出精度を維持したまま学習データを削減する方法を検討した。実験の結果、自動生成した擬似学習データを実学習データに加えてデータ量を実学習データの 50 倍にすると、実学習データが 10 件程度と少量でも比較的高い抽出精度となることを確認した。

(2)表構造解析

論文中の表の構造を解析する手法を考案して、その解析結果を利用して表中の数値データのグラフを自動生成することを検討した。数値が記載された表は学術論文にしばしば現れるが、その活用にはまず表構造を解析する必要がある。本研究では、表中の不可分なデータが記載される領域であるセルに着目し、表の罫線と補助罫線、また表中のトークンを使ってセルを推定することで表構造を決定する。なお、セルを分割する線分のうち、実際に表に書かれているものが罫線で、書かれていないがセルの分割に必要なものが補助罫線である。本研究ではこの補助罫線を、表中のトークンの配置などに基づいて推定する。本研究では、表領域が指定された PDF 文書の表に対して補助罫線の推定やセルの推定を行うため、合わせて四つの NN モジュールからなる表構造解析手法を提案した。図 1 は、そのうちの一つのセル生成モジュールが、表中の隣接トークンを結合していくイメージを示している。図 1 で、赤の線で囲まれているのがトークンで、このモジュールは、結合できるトークンがなくなるまで隣接する 2 トークンを水平方向と垂直方向に交互に結合して、セルを生成する。

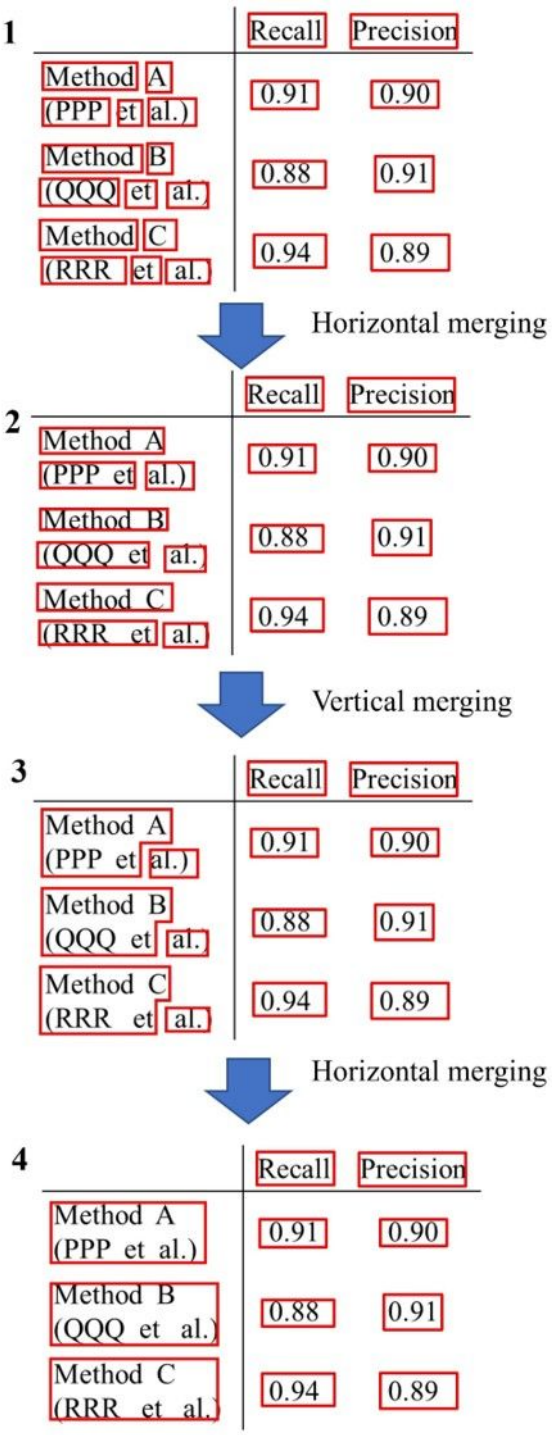

図 1 トークンの結合によるセルの生成

実験では、文書解析の著名な国際会議である ICDAR2013 の表構造解析タスクにおいて提供された 156 の表の構造解析を行い、同コンペティションにおいて定義されたセルの隣接関係の再現性を示す評価指標を算出して評価した。実験の結果、提案手法はその評価指標である再現率、適合率、F 値がそれぞれ、0.967、0.977、0.972 となった。この F 値は、同コンペティションにおいて最高の結果の F 値 0.946 を 2.6 ポイント上回る優れたものだった。

本研究ではまた、表構造解析の前処理として、文書中の表領域検出についても検討した。これは、通常文書の中には本文や図など表以外の構成要素が含まれており、表構造解析を自動化するにはまず文書中の表領域を特定する必要があるためである。

表構造解析結果を利用した数値データのグラフ化については、棒グラフを自動生成する手法を検討した。

(3)サイバーフィジカル論文閲覧支援

タブレット端末のカメラと OCR を用いて、紙媒体の学術論文の閲覧支援を行う論文閲覧支援インタフェースでは、ユーザがカメラ機能を用いて紙媒体の論文を撮影すると、撮影画像中のテキストを OCR で文字認識して重要語を抽出し、Wikipedia などのウェブ上の関連情報と合わせて表示する。本研究では、このインタフェースに組み込むためにユーザの使用履歴を利用した備忘録を設計した。この備忘録では、論文閲覧画面と備忘録画面の切り替えや、備忘録に表示する重要文、検索語句の履歴、メモ機能などを検討した。

本研究ではまた、読者が閲覧している論文中の引用箇所に提示するのにふさわしい補助情報の自動生成手法を提案した。具体的には、論文中の引用箇所における著者の引用意図を推定し、その引用意図に基づいて引用論文以外のウェブの情報源（Wikipedia、Papers with Code）も活用して、とりわけ初学者の読者の助けになる補助情報を自動生成する。本研究では、読者の閲覧支援の観点から、手法やアルゴリズムを引用する Method、結論を引用する Conclusion、主張の根拠とする内容を引用する Basis、実験やそこで用いるデータを引用する Data の四つの引用意図を定めた。実験では、その引用意図ごとに定めた方法で引用箇所の補助情報を生成し、被験者にその有用性を評価させた。その結果、生成した補助情報は、引用意図、補助情報の情報源、生成方法によってその有用性は異なったが、例えば引用意図 Method では引用論文や Papers with Code の記事から生成した補助情報が有用であることなどを確認した。

５．主な発表論文等

〔雑誌論文〕　計5件（うち査読付論文　5件／うち国際共著　0件／うちオープンアクセス　0件）

１．著者名 Manabu Ohta, Ryoya Yamada, Teruhito Kanazawa, Atsuhiro Takasu	４．巻 -
２．論文標題 Table-structure recognition method using neural networks for implicit ruled line estimation and cell estimation	５．発行年 2021年
３．雑誌名 Proc. 21st ACM Symposium on Document Engineering (DocEng 2021)	６．最初と最後の頁 1-7
掲載論文のDOI（デジタルオブジェクト識別子） 10.1145/3469096.3469870	査読の有無 有
オープンアクセス オープンアクセスではない、又はオープンアクセスが困難	国際共著 －

１．著者名 Hiroyuki Aoyagi, Teruhito Kanazawa, Atsuhiro Takasu, Fumito Uwano, Manabu Ohta	４．巻 -
２．論文標題 Table-structure Recognition Method Consisting of Plural Neural Network Modules	５．発行年 2022年
３．雑誌名 Proc. 11th International Conference on Pattern Recognition Applications and Methods (ICPRAM 2022)	６．最初と最後の頁 542-549
掲載論文のDOI（デジタルオブジェクト識別子） 10.5220/0010817700003122	査読の有無 有
オープンアクセス オープンアクセスではない、又はオープンアクセスが困難	国際共著 －

１．著者名 Manabu Ohta, Ryoya Yamada, Teruhito Kanazawa, Atsuhiro Takasu	４．巻 -
２．論文標題 A Cell-detection-based Table-structure Recognition Method	５．発行年 2019年
３．雑誌名 Proc. 19th ACM Symposium on Document Engineering (DocEng 2019)	６．最初と最後の頁 1-4
掲載論文のDOI（デジタルオブジェクト識別子） 10.1145/3342558.3345412	査読の有無 有
オープンアクセス オープンアクセスではない、又はオープンアクセスが困難	国際共著 －

１．著者名 Shunsuke Tanabe, Manabu Ohta, Atsuhiro Takasu, Jun Adachi	４．巻 -
２．論文標題 An Approach to Estimating Cited Sentences in Academic Papers Using Doc2vec	５．発行年 2018年
３．雑誌名 Proc. 10th International Conference on Management of Digital EcoSystems (MEDES'18)	６．最初と最後の頁 118-125
掲載論文のDOI（デジタルオブジェクト識別子） 10.1145/3281375.3281391	査読の有無 有
オープンアクセス オープンアクセスではない、又はオープンアクセスが困難	国際共著 －

1．著者名 Ryoya Yamada, Manabu Ohta, Atsuhiro Takasu	4．巻 -
2．論文標題 An Automatic Graph Generation Method for Scholarly Papers Based on Table Structure Analysis	5．発行年 2018年
3．雑誌名 Proc. 10th International Conference on Management of Digital EcoSystems (MEDES'18)	6．最初と最後の頁 132-140
掲載論文のDOI（デジタルオブジェクト識別子） 10.1145/3281375.3281389	査読の有無 有
オープンアクセス オープンアクセスではない、又はオープンアクセスが困難	国際共著 －

〔学会発表〕　計23件（うち招待講演　0件／うち国際学会　0件）

1．発表者名 西海真祥，金澤輝一，上野史，太田学
2．発表標題 引用意図を利用した初学者向け学術論文閲覧支援方法の検討
3．学会等名 第21回情報科学技術フォーラム（FIT2022）
4．発表年 2022年

1．発表者名 青柳拓志，金澤輝一，高須淳宏，上野史，太田学
2．発表標題 表検出を含むエンドツーエンド表構造解析手法の評価
3．学会等名 ARG 第18回Webインテリジェンスとインタラクション研究会
4．発表年 2022年

1．発表者名 西海真祥，金澤輝一，上野史，太田学
2．発表標題 文の類似度と Extractive QA による被引用文特定の一手法
3．学会等名 第15回データ工学と情報マネジメントに関するフォーラム（DEIM2023）
4．発表年 2023年

１．発表者名
中山竣平，金澤輝一，高須淳宏，上野史，太田学

２．発表標題
BERTによる参考文献書誌情報抽出の誤り検出の評価

３．学会等名
第15回データ工学と情報マネジメントに関するフォーラム（DEIM2023）

４．発表年
2023年

１．発表者名
青柳拓志，金澤輝一，高須淳宏，上野史，太田学

２．発表標題
グラフニューラルネットワークを用いたエンドツーエンド表構造解析手法の提案

３．学会等名
第15回データ工学と情報マネジメントに関するフォーラム（DEIM2023）

４．発表年
2023年

１．発表者名
細谷亮太，金澤輝一，上野史，太田学

２．発表標題
ニューラルネットワークによる日本語を含む表の構造解析の一手法

３．学会等名
第15回データ工学と情報マネジメントに関するフォーラム（DEIM2023）

４．発表年
2023年

１．発表者名
高橋春成，金澤輝一，上野史，太田学

２．発表標題
初学者の論文閲覧支援のための日本語論文からの専門用語抽出の一手法

３．学会等名
第15回データ工学と情報マネジメントに関するフォーラム（DEIM2023）

４．発表年
2023年

１．発表者名
荒川瞭平，金澤輝一，高須淳宏，上野史，太田学

２．発表標題
BERTによる参考文献書誌情報抽出における擬似学習データの有効性評価

３．学会等名
ARG 第17回Webインテリジェンスとインタラクション研究会

４．発表年
2021年

１．発表者名
高橋春成，金澤輝一，高須淳宏，上野史，太田学

２．発表標題
BERTによる和文の参考文献文字列からの書誌情報抽出の評価

３．学会等名
第14回データ工学と情報マネジメントに関するフォーラム (DEIM2022)

４．発表年
2022年

１．発表者名
青柳拓志，金澤輝一，高須淳宏，上野史，太田学

２．発表標題
ニューラルネットワークを用いた表構造解析の一手法

３．学会等名
第13回データ工学と情報マネジメントに関するフォーラム (DEIM2021)

４．発表年
2021年

１．発表者名
西海真祥，金澤輝一，高須淳宏，上野史，太田学

２．発表標題
引用意図を利用した学術論文閲覧支援情報生成の一手法

３．学会等名
第13回データ工学と情報マネジメントに関するフォーラム (DEIM2021)

４．発表年
2021年

１．発表者名
岩本拓実，金澤輝一，上野史，太田学

２．発表標題
ユーザの興味を利用した学術論文閲覧支援の一手法

３．学会等名
情報処理学会第83回全国大会

４．発表年
2021年

１．発表者名
荒川瞭平，金澤輝一，高須淳宏，上野史，太田学

２．発表標題
BERTによる参考文献書誌情報抽出の精度向上

３．学会等名
情報処理学会第83回全国大会

４．発表年
2021年

１．発表者名
岩本拓実，高須淳宏，太田学

２．発表標題
学術論文閲覧支援のための備忘録の設計

３．学会等名
第18回情報科学技術フォーラム（FIT2019）

４．発表年
2019年

１．発表者名
荒川瞭平，太田学，金澤 輝一，高須淳宏

２．発表標題
少量学習データとBi-directional LSTM-CNN-CRFによる参考文献書誌情報抽出

３．学会等名
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（DEIM2020）

４．発表年
2020年

１．発表者名
山田凌也，太田学，金澤 輝一，高須淳宏

２．発表標題
機械学習を用いた表構造解析の一手法

３．学会等名
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 (DEIM2020)

４．発表年
2020年

１．発表者名
八田谷翔太，太田学

２．発表標題
能動学習を用いた実験情報抽出の一手法

３．学会等名
電子情報通信学会2020年総合大会 情報・システムソサイエティ特別企画 学生ポスターセッション

４．発表年
2020年

１．発表者名
田邊俊介，太田学

２．発表標題
学術論文の被引用文章生成の一手法

３．学会等名
電子情報通信学会2020年総合大会 情報・システムソサイエティ特別企画 学生ポスターセッション

４．発表年
2020年

１．発表者名
浪越大貴，太田学，高須淳宏，安達淳

２．発表標題
Bi-directional LSTM-CNN-CRFによる参考文献書誌情報抽出

３．学会等名
電子情報通信学会データ工学研究会，情報処理学会データベースシステム研究会

４．発表年
2018年

１．発表者名
木下諒，太田学，高須淳宏

２．発表標題
転移学習を用いた少量学習データによる参考文献書誌情報抽出

３．学会等名
第11回データ工学と情報マネジメントに関するフォーラム（DEIM2019）

４．発表年
2019年

１．発表者名
田邊俊介，太田学，高須淳宏

２．発表標題
引用文脈の分散表現を利用した学術論文の被引用文章要約の一手法

３．学会等名
第11回データ工学と情報マネジメントに関するフォーラム（DEIM2019）

４．発表年
2019年

１．発表者名
岩本拓実，太田学，高須淳宏

２．発表標題
学術論文閲覧支援インタフェースにおける備忘録の自動生成の一手法

３．学会等名
第11回データ工学と情報マネジメントに関するフォーラム（DEIM2019）

４．発表年
2019年

１．発表者名
山田凌也，太田学，高須淳宏

２．発表標題
グラフの自動生成のための表の構造解析の一手法

３．学会等名
第11回データ工学と情報マネジメントに関するフォーラム（DEIM2019）

４．発表年
2019年

〔図書〕　計0件

〔産業財産権〕

〔その他〕

―

６．研究組織

	氏名 （ローマ字氏名） （研究者番号）	所属研究機関・部局・職 （機関番号）	備考
研究協力者	高須　淳宏 (Takasu Atsuhiro)		

７．科研費を使用して開催した国際研究集会

〔国際研究集会〕　計0件

８．本研究に関連して実施した国際共同研究の実施状況

共同研究相手国	相手方研究機関