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　６ 年　 ６ 月 １７ 日現在

機関番号：６２６１５
研究種目：挑戦的研究（萌芽）
研究期間：2022～2023
課題番号：２２Ｋ１９８１８

研究課題名（和文）実文書の理解と活用に向けた言語解析手法の深化

研究課題名（英文）Deepening linguistic analysis methods for understanding and utilizing real documents

研究代表者
相澤　彰子（Aizawa, Akiko）

国立情報学研究所・コンテンツ科学研究系・教授

研究者番号：９０２２２４４７

交付決定額（研究期間全体）：（直接経費）　　4,900,000円

研究成果の概要（和文）：本研究では、文書構造や視覚情報を踏まえた自然言語処理手法についての検討に取り組んだ。具体的には、レイアウトされた文書やインタラクティブなウェブ上のフォームなどの文書構造、数式や数字などを含む文書中の非言語要素、テキストの編集的な属性（大文字・小文字の違い）の３つの文書構成要素に注目して、これらの分析やモデル化を提案して有効性を示した。また、自然言語処理分野の国際会議論文を対象として、フォントやレイアウトや図表やインライン数式を含む非言語情報をアノテーションタグの形でテキストに追加したXML形式の文書コーパスを構築した。

研究成果の学術的意義や社会的意義
自然言語処理の分野において、文書から自然言語処理ツールで解析可能な文を抽出する処理は、アドホックで自動化が困難な「前処理」とみなされ、従来はあまり注目されてこなかった。しかしながら、2022年における言語モデルの急速な進展により、当初目指していた本テーマの挑戦性が、訓練データと言語モデルの大規模化によって現実に解決可能な問題となってきた。文書AIが大きな注目を集める中で、本研究で提案したフレームワークや構築した資源は今後の研究に資することが期待される。

研究成果の概要（英文）：This study addressed natural language processing methods based on document structure and visual information. Specifically, we focused on three document components: document structure, such as document layout and forms on the interactive web; nonverbal elements in documents, including mathematical expressions and numbers; and editorial attributes of text (capitalization), and proposed analysis and modeling of these elements to demonstrate their effectiveness. We also constructed a document corpus in XML format for international conference papers in natural language processing, in which nonverbal information, including fonts, layouts, charts, and inline mathematical expressions, were added as annotation tags.

研究分野：研究成果の学術的意義や社会的意義

キーワード：文書理解　　自然言語解析　　言語モデル　文書レイアウト　　視覚的言語理解

様　式　Ｃ－１９、Ｆ－１９－１（共通）

１．研究開始当初の背景

(1) 自然言語処理の分野では、文書から自然言語処理ツールで解析可能な文を抽出する処理は、アドホックで自動化が困難な「前処理」とみなされ、これまであまり注目されてこなかった。しかしながら、不完全な前処理によって文中にノイズが混在し言語解析の性能が低下すること、前処理によって本来は文書理解の助けとなるはずの情報が失われることなどから、実問題への言語処理の適用において、この「前処理」が性能に及ぼす影響は大きい。このため、文書レイアウト構造や非言語要素など文書の視覚的・操作的特徴を踏まえた自然言語処理技術の研究が必要となっていた。

２．研究の目的

(1) 現在の自然言語処理は、入力単位としてトークンの並びである「文」または「文の集合」を想定している。しかし、現実の文書には、レイアウト構造や表示スタイルなどが混然一体となって埋め込まれ、読みを支援する「手がかり」として機能している。人間が文書を読む際には、視覚的な入力を通してこのような手がかりを即座に解読して、「文」の意味理解へとつなげていると考えられる。そこで本研究では、文書中に埋め込まれた非言語情報を抽出・活用するための文書解析手法を検討する。

(2) 具体的には、レイアウトされた文書やインタラクティブなウェブ上のフォームなどの文書構造、数式や数字などを含む文書中の非言語要素、テキストの編集的な属性（大文字・小文字の違い）の３つの文書構成要素に注目する。そして、これらを分析・モデル化することで、言語処理性能を向上する手法を開発する。

３．研究の方法

(1) 構造や視覚特徴を持ち、さらにフォームを経由してユーザとインタラクティブにやりとりをするブラウザ上のウェブ文書に注目して、ブラウザ上の仮想環境で指定のタスクを遂行する言語処理モデルの研究に取り組んだ。このモデルはブラウザに表示される文書画像を読み込んで要求されているタスクを解読し、そのタスクを遂行するためのアクション列を出力する。タスクとしては質問応答やデータベース検索や視覚言語タスクなどを想定する。実験では、先行研究である MiniWoB++のタスクから正解データを人手および疑似的に生成して、正解のアクション列が生成可能であることを示すとともに、現状のモデルの課題について分析を行った。

(2) 自然言語処理分野の国際会議論文を対象として、フォントやレイアウトや図表やインライン数式を含む非言語情報をアノテーションタグの形でテキストに追加した XML 形式の文書コーパスを構築した。非言語要素として、抽象的な数量概念と対応付けられながらも、文脈的な解釈を必要とする文中の数字（例：「220cm の男性→背が高い」）を対象として、k 近傍言語モデルを用いた数字の埋め込み表現を提案し、構築した論文コーパスを含むことなるデータに適応して有効性を評価した。また、複雑な談話構造をわかりやすく提示するインタフェースにかかわるものとして、数学定理証明の可読性向上に関する研究に取り組んだ。

(3) 英語における大文字と小文字表記の違いに注目して、固有表現抽出タスクにおける影響を分析して性能改善に取り組んだ。

４．研究成果

(1) ブラウザ上で表示されるウェブ文書を対象として、画像、テキスト、ユーザアクションを統一的に扱うためのツールを実装して、データ収集やモデル構築を行った（Iki et al., 2023)。

さらに、ブラウザ上の仮想環境で指定のタスクを遂行するアクション列を出力するタスクであるMiniWoB++に対して、大規模言語モデルのエージェントを用いて、新たなタスクおよび正解アクション列を自動生成するフレームワークを提案・実装した。検証により、エージェントのフレームワークを用いて、新たなMiniWoB++タスクインスタンスを正解のアクション列とともに自動生成可能であることを示した。

(2) 自然言語処理分野の主要な国際会議論文をCreative CommonsライセンスのもとでアーカイブしているACL Anthologyサイトから全PDF論文をクロールして解析、各単語ボックスに、フォント、ページ内位置情報、レイアウトボックス、セクション、セクションタイプ、後続スペース、文書 ID（文書のメタ情報）などの情報を対応付けるとともに、論文全体をブラウザで閲覧可能なxhtml形式に変換したコーパスを構築した。また、k近傍言語モデルを用いた数字の埋め込み表現については、構築したデータセット用いて有効性を評価し、その結果を国際会議で発表した（Sakamoto et al., 2023)。さらに、数学定理証明の可読性向上については、自然言語による定理証明の記述構造と、定理証明器Leanの記述構造を比較するとともに(Tsurusaki et al., 2023)、Leanのコメント文として入力した自然言語テキストをCCGパーザーで構文解析する枠組みを実装した。

(3) 英語における大文字と小文字表記の違いについては、トランスフォーマー型の言語モデルにおいて、大文字・小文字の違いを保持するモデルとしないモデルの２通りを用いて、固有表現抽出タスクにおける分野間転移などの影響を詳細に分析し、得られた知見に基づきデータ増強による性能改善に取り組んで国際学会で発表した（Dao et al., 2022)。

(4) 2022年における言語モデルの急速な進展により、当初目指していた本テーマの挑戦性が、訓練データと言語モデルの大規模化によって現実に解決可能な問題となってきた。文書AIが大きな注目を集める中で、本研究で提案したフレームワークや構築した資源は今後の研究に資することが期待される。

５．主な発表論文等

〔雑誌論文〕　計4件（うち査読付論文　2件／うち国際共著　0件／うちオープンアクセス　2件）

１．著者名 壹岐太一，増本雄斗，相澤彰子	４．巻 -
２．論文標題 ディスプレイ操作記録ツールの提案と有効性の検証	５．発行年 2023年
３．雑誌名 言語処理学会第29回年次大会(NLP2023)	６．最初と最後の頁 -
掲載論文のDOI（デジタルオブジェクト識別子） なし	査読の有無 無
オープンアクセス オープンアクセスとしている（また、その予定である）	国際共著 －

１．著者名 Tuan An Dao, Akiko Aizawa	４．巻 -
２．論文標題 Effect of Letter Case on Named Entity Recognition Performance	５．発行年 2023年
３．雑誌名 Proceedings of The 28th International Conference on Natural Language and Information Systems	６．最初と最後の頁 -
掲載論文のDOI（デジタルオブジェクト識別子） 10.1007/978-3-031-35320-8_45	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 津留崎堅章，相澤彰子	４．巻 -
２．論文標題 文芸的プログラミングによる形式的証明の可読性向上	５．発行年 2023年
３．雑誌名 情報アクセシビリティをめぐる諸問題に関する研究集会	６．最初と最後の頁 -
掲載論文のDOI（デジタルオブジェクト識別子） なし	査読の有無 無
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 Taku Sakamoto, Akiko Aizawa	４．巻 -
２．論文標題 Predicting Numerals in Text Using Nearest Neighbor Language Models	５．発行年 2023年
３．雑誌名 The 61st Annual Meeting of the Association for Computational Linguistics (ACL)	６．最初と最後の頁 4795-4809
掲載論文のDOI（デジタルオブジェクト識別子） 10.18653/v1/2023.findings-acl.295	査読の有無 有
オープンアクセス オープンアクセスとしている（また、その予定である）	国際共著 －

〔学会発表〕　計0件

〔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

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