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 ２１ 日現在

機関番号：３４３１６
研究種目：基盤研究(C)（一般）
研究期間：2019～2023
課題番号：１９Ｋ１２２４１

研究課題名（和文）自然言語処理技術を用いた快適なWeb利活用支援に関する研究

研究課題名（英文）Research on Comfortable Web Utilization Support Using Natural Language Processing Technology

研究代表者
馬　青（Ma, Qing）

龍谷大学・先端理工学部・教授

研究者番号：３０３５８８８２

交付決定額（研究期間全体）：（直接経費）　　3,300,000円

研究成果の概要（和文）：　言語処理能力が不十分な外国人、高齢者、子供など特定の層を対象に快適なWeb利用を支援する自然言語処理技術の基盤形成を目指し、非構造化文書からの検索用語の抽出によるWeb検索支援、日本語の文法誤りの検出及び誤り文と正しい文の分類・変換を通じた日本語学習者支援、プログラム課題文からの重要箇所の抽出に基づくプログラミング学習者支援、ナレッジコミュニティの活用支援、SNSからの情報獲得の支援など、様々な課題に焦点を当てて研究を推進した。
　これらの研究では、機械学習や深層学習を基軸にしたアプローチを採用し、大規模実データを用いた検証によりその有効性を確認し、さまざまな手法やシステムの開発に成功した。

研究成果の学術的意義や社会的意義
　本研究成果は、言語処理能力が不十分な外国人、高齢者、子供など特定の層の人々に快適なWeb利用を支援する自然言語処理技術の基盤形成に大きく貢献している。複数の科研基盤などで得られた研究成果・技術を融合している点、すなわち、情報検索・情報抽出・意味処理などの単独の応用ではなく、それらの諸自然言語処理技術と、Web関連技術・統計技術・機械学習などを融合的に利用している点が本研究成果の学術的な特色である。また、本研究成果は、他の関連研究の基本要素技術の発展に寄与できる可能性が高く、関連して得られる知見は、人間の言語獲得・理解のメカニズム解明の重要なヒントとなり得ると考える。

研究成果の概要（英文）：　This research aims to construct a foundation for natural language processing technology to support comfortable Web usage targeting specific groups such as foreigners with insufficient language processing abilities, the elderly, and children. The study focuses on various issues, including Web search assistance through the extraction of search terms from unstructured documents, support for Japanese learners through the detection of grammatical errors and the classification and transformation of incorrect and correct sentences, support for programming learners based on the extraction of key points from programming task statements, utilizing knowledge communities, and supporting information acquisition from social media networks (SNS).
　In these studies, approaches based on machine learning and deep learning were adopted, and their effectiveness was confirmed through validation using large-scale real data, leading to the successful development of various methods and systems.

研究分野：知能情報学

キーワード：Web検索支援　日本語学習支援　プログラミング学習支援　情報獲得　用語抽出　系列ラベリング　深層学習　対照学習

様　式　Ｃ－１９、Ｆ－１９－１（共通）

１．研究開始当初の背景

Web の利活用状況に関しては、Web 検索が大きな役割を果たしている。ただし、Web 検索の性能が大幅に向上したとはいえ、それはユーザが適切な検索キーワードを入力できることを前提としており、言語処理能力が不十分な外国人や高齢者、子供にとっては難しい要求である。また、Web 検索エンジンによる機械処理ではなく、人間と人間が Web を介して情報をやり取りするナレッジコミュニティにおいても、外国人や高齢者、子供にとって、完全な意味理解が困難である。さらに、何らかの決定（例: ホテル選定）をするときの Web 上の口コミ情報の利用についても特定の言語（英語や日本語）で書かれているユーザ生成テキストを順に読む必要があり、外国人や高齢者にとって負担が大きい。すなわち、特定の層の人々には、今もなお Web を十分にまたは気軽に享受することが困難であるという障壁が存在する。我々はこの障壁を取り払うために、Web 検索支援に焦点を当てた研究を実施してきた（H25-H28:科研基盤(C)）。

２．研究の目的

提案研究は、先行研究（H25-H28:科研基盤(C)）で確立した手法・知見を中核とし、H20 以降に推進してきた関連の科研費研究で得られた研究成果をベースに、検索支援に限定しない、より広範囲な快適な Web 利活用の支援のための自然言語処理技術の基盤形成を目指した。

３．研究の方法

提案研究は快適な Web 利活用の支援のための自然言語処理技術の基盤形成を目指し、各々のテーマについて以下に述べる方法で研究を推進した。

3.1　Web 検索支援：非構造化文書からの検索用語抽出

検索対象を示す適切な検索用語を提示する Web 検索支援として、見出し語（検索用語）と説明文が明確に分かれていない、非構造化文書の「説明テキスト」（例：「偽物のこと。DVD とかに使われる」）からそれを表す用語（例：「海賊版」）の抽出の研究を行った。本研究では、「説明テキスト」で Web 検索を行い、Web 検索結果のテキスト（例：「コピー商品、偽ブランド品等の模倣品をはじめ、DVD、ゲームソフト等の海賊版、コンサートチケット等の偽造品など、」）と「説明テキスト」を入力とする汎用言語モデル BERT（QA タスク仕様）を用いての用語抽出手法を提案した。

3.2　日本語学習者支援：日本語の文法誤り検出、誤り文・正しい文の分類・正しい文への変換

日本語学習支援の研究として、作文の日本語文法の誤り検出に関する研究を行った。文法誤り検出を系列ラベリングタスクとしてとらえ、汎用言語モデル BERT をベースとし、日本語入力を単語分割せずにも解析できる Flair モデルの導入など、深層学習で解く手法を提案した。さらに、追加の事前学習として対照学習と MLM (Masked Language Model)のマルチタスク学習を提案した。

日本語学習支援の研究として、不完全な日本語の文にも対応できる、形態素解析を要求しない深層学習 Character-level CNN (CLCNN)を用い、誤り文・正しい文の分類の研究を行った。誤り文から正しい文への変換も試みた。また、CLCNN の転移学習能力の評価を、作者推定、現代日本語書き言葉均衡コーパスのデータの取得元判定、外国人が書いた日本語作文とその添削文の判定、という三種類の言語処理タスクにより行った。

3.3　プログラミング学習者支援：コーパス構築・プログラム課題文からの重要箇所抽出

小・中学生のプログラミング教育の接続支援に関する研究として、まず、プログラミング課題を記述する文章（課題テキスト）の可読化を目指し、課題テキストとそれを遂行するためのプログラムコードの対応データの収集を行った。課題テキストに対し、入力や与えられた条件、出力などに関する記述箇所（重要箇所）を特定する情報（タグ）を付与し、タグ付きコーパスの構築を行った。構築したコーパスの有用性を検証するために、用語抽出研究で開発した BERT の QA タスク仕様での用語抽出手法での、タグ付きコーパスを用いたファインチューニングと課題テキストのタグ付けの評価実験を行った。

課題テキストの重要箇所の抽出を質問応答としてとらえるアプローチと、系列ラベリングとしてとらえるアプローチを考えた。構築したコーパスを学習データとして汎用言語モデル BERT へのファインチューニングを行い、BERT によるプログラミング課題文からの重要箇所の抽出を行った。系列ラベリングアプローチでは複数の文が連なった課題テキストを対象とし、全体から重要箇所を直接ラベリングする方法と、段階的にラベリングする方法を提案した。

3.4　ナレッジコミュニティの利用支援：トピック抽出と質問分類

ナレッジコミュニティの利用支援の研究として、Yahoo!知恵袋を代表とした Q&A サイトへの投稿のトピック・キーワード抽出と内容分類を行った。手法として、トピック・キーワードの抽

出と教師なし分類を同時に行えるトピックモデルを用いた。トピックモデルの中にももっとも本タスクに適する Online Latent Dirichlet Allocation(LDA)を適用した。 性能向上を図るために、トピックモデルとクラスタリングのハイブリッド手法を提案した。

3.5 SNS からの情報獲得

Web 利活用支援の一環として、SNS から有用な情報の獲得に関する研究を行った。

商品に対する多言語 Amazon レビューテキスト中の日本語レビューテキストに対し、星の数（レーティング）の予測を行った。手法として BERT と RoBERTa (A Robustly Optimized BERT Pretraining Approach)を用いた。これらレビューテキストでファインチューニングした複数の深層学習モデルでのツイトデータ（X のポストデータ）に対するレーティング予測も行った。ツイトデータは話し言葉表現に近いことに着目し、事前学習モデルとして日本語話し言葉 BERT を導入した。また、学習データとしての Amazon レビューテキストとツイトデータの特徴（文字数とテキストスタイル）が異なることから、学習データの要約による短縮化や ChatGPT を利用した話し言葉への変換も検討した。

日本最大のコスメ・美容の総合サイト@cosme の利用者支援として、化粧品の個人化推薦手法の開発を行った。推薦に、肌質や年齢といったユーザ情報とブランドなどの商品情報に加え、レビューテキストから美白や保湿などに関する感性評価を機械学習（SVM や深層学習の Stacked Denoising Autoencoder, CNN により抽出し用いた。

機械学習 SVM と深層学習 LSTM を用いての SNS（ツイトデータ）による政党支持率予測の研究を、二つの課題に分けて行った。一つ目の課題は、個々のツイトが政党に対し、肯定・否定・中立の推定を行うことである。二つ目の課題は、ユーザのツイトからユーザが政党を支持・不支持・中立の判定をし、任意の月の政党の支持率の予測を行うことである。

４．研究成果

4.1 Web 検索支援：非構造化文書からの検索用語抽出

Wikipedia のような比較的整った文書に限らない一般の Web 文書から用語の抽出は難しいタスクであること、クエリが整ったものではない場合はさらに難しいことを確認した。複数文書からの用語候補の抽出を行い、頻度に基づいて単純にランキングすることで、1 位に用語を 20〜30%程度の精度で、10 位以内に用語を 40%程度の精度で取得することができた。また、用語候補のランキング方法についても検討を行い、多数決によるランキングが最も有効であることがわかった。類似度に基づく手法は多数決による手法を補完できる可能性があることがわかった。また、提案手法では用語抽出のたびに Web 検索を行って Google スニペットを求めるため, 学習データにない新語・流行語であっても、追加学習なしで抽出できることがわかった。

研究成果は言語処理年次大会論文 1 編と IPSJ 自然言語処理研究会論文 1 編として発表した。

4.2 日本語学習者支援：日本語の文法誤り検出、誤り文・正しい文の分類・正しい文への変換

日本語の文法誤り検出の研究においては、提案手法について、日本語の語学学習のための、誤りを含む文とその訂正文からなる Lang-8 コーパスから抽出・加工した 72 万文のデータを用いて評価した。

提案手法は、適合率（Precision），再現率（Recall），F(0.5) 値のすべてにおいて、先行研究の Bi-LSTM より高かった。また、文字レベルのモデル Flair を導入することにより、適合率がさらに向上し、言語学習者へのより正確なフィードバックが期待できることがわかった。文法誤り検出の性能向上を目的に提案した追加事前学習を行ったモデルと行わなかったモデルとの比較を行い、追加の事前学習の効果を確認した。各単語の意味表現, 文の意味的表現を維持しつつ, 文法誤りを含まない文と文法誤りを含む文を分離できるよう学習ができたと考えられる。

誤り文・正しい文の分類および誤り文から正しい文への変換の研究においては、国立国語研究所が作成・公開している作文対訳データベースから入手した外国人の日本語作文とその添削文の 3 万文を超えるデータを用いて評価した。提案手法は、異なるタスク間の転移学習ができることを確認した。また、三種類の言語処理タスクのうちの 1 つである日本語学習者が書いた文章か添削された文章かの判別タスクにおいては約 70%の正解率が得られた。誤り文から正しい文への変換も試みたが、有効性が確認できなかった。

研究成果は言語処理年次大会論文 2 編として発表した。

4.3 プログラミング学習者支援：コーパス構築・プログラム課題文からの重要箇所抽出

プログラミング学習者支援用に構築したタグ付きコーパスの有用性を、用語抽出研究で開発した BERT の QA タスク仕様での用語抽出手法での、タグ付きコーパスを用いたファインチューニングと課題テキストのタグ付けの評価実験によって確認した。

構築したコーパス（課題文 2,172 件、計 12,471 文）を用いてプログラム課題文からの重要箇所抽出手法の評価を行った。質問応答手法ではその有効性が確認できなかった。一方、直接系列ラベリング方法では適合率・再現率・F(0.5)でその有効性が確認できた。段階的ラベリング方法では適合率・再現率・F(0.5)がそれぞれ 0 さらに向上した。特に出現頻度が低いラベルに対して大きな改善が得られることがわかった。

研究成果は言語処理年次大会論文 1 編として発表した。

4.4　ナレッジコミュニティの利用支援：トピック抽出と質問分類

1000 万ユーザを有するといわれる Q&A サイトの OKWave から収集した 25000 を超える質問文（これらは 19 カテゴリ 200 サブカテゴリ 1000 サブサブカテゴリに属する）を用いて提案手法の評価を行った。評価の結果、教師なし質問分類タスクにおいてはトピックモデルの性能（Purity）がクラスタリングよりよく、ハイブリッド手法の性能がトピックモデルより優れていることが確認できた。また、質問文のベクトル化に TF-IDF の導入が性能向上に寄与することがわかった。トピックモデルのトピック・キーワード抽出における有効性を topic coherence, topic difference 及び人手評価で確認した。

研究成果は国際会議論文 1 編として発表した。

4.5　SNS からの情報獲得

レーティング予測の研究については、Amazon レビューテキスト中の約 20 万日本語レビューテキスト及び人手で収集した 300 弱程度のツイトデータを用いて評価を行った。

レーティング予測を順序付きの多値分類問題とみなし、順序付きの多値分類の性能評価に QWK を用いた。Amazon レビューテキストに対するレーティング予測においては、深層学習モデル RoBERTa の QWK が一番高かった。一方，日本語ツイトデータに対するレーティング予測においては，日本語話し言葉 BERT の QWK が一番高かった。日本語話し言葉 BERT の導入が有効であった。一方、学習データとしての Amazon レビューテキストとツイトデータの特徴（文字数とテキストスタイル）が異なることから、短縮した学習データと話し言葉に変換したデータを用いた実験も行ったが効果は見られなかった。

個人化推薦の研究については、@cosme に投稿されている化粧水カテゴリのレビューテキストを収集し、人手で５つの感性項目に対し４段階の評価値のラベル付けを行い 1000 件のデータのタグ付きコーパスを構築して利用し、提案した感情情報推定手法の評価を行った。評価の結果、機械学習が有効であることが確認できた。用いる SVM と SdA と CNN に感性推定精度に差がほとんどなかったが、処理時間は SVM がもっとも短かった。また、レビューテキストのベクトル化前に係り受け解析を導入しているがそこに文節の繋ぎなおしと文の分割を取り入れたことにより感性推定の精度がわらに向上した。

SNS による政党支持率の予測に関する研究においては、さまざまな方法の組み合わせを考案し評価実験を行った。有効な方法の組み合わせによる政党支持率の予測は結果として、マスメディアの出した政党支持率と同様の変化が見られ、ある程度社会の流れをつかめた予測を行うことができた。政党支持率自体が不確かなため、数値的に正当性を示すことはできないが、 ツイッターなどの SNS でも政党支持率の予測や変化を見ることができることがわかった。各マスメディアの世論調査の電話調査では膨大な人手が必要なところ、本研究では機械学習を用いることで大幅な人員・コスト削減を可能にした。

研究成果は国際会議論文 1 編、招待講演 1 件、言語処理年次大会論文 6 編として発表した。

５．主な発表論文等

〔雑誌論文〕　計2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　1件）

１．著者名 Binggang Zhuo, Masaki Murata, Qing Ma	４．巻 E106-D
２．論文標題 Auxiliary Loss for BERT-Based Paragraph Segmentation	５．発行年 2023年
３．雑誌名 IEICE Transactions on Information and Systems	６．最初と最後の頁 58-67
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 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 Masaki Murata, Kensuke Okazaki, Qing Ma	４．巻 28
２．論文標題 Improved Method for Organizing Information Contained in Multiple Documents into a Table	５．発行年 2021年
３．雑誌名 Journal of Natural Language Processing	６．最初と最後の頁 802～823
掲載論文のDOI（デジタルオブジェクト識別子） 10.5715/jnlp.28.802	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17件（うち招待講演　1件／うち国際学会　2件）

１．発表者名 岡本昇也，南條浩輝，馬青
２．発表標題 文法誤り検出BERTのためのマルチタスク追加事前学習
３．学会等名 言語処理学会第30回年次大会
４．発表年 2024年

１．発表者名 森廣勇樹，南條浩輝，馬青
２．発表標題 Xのポストデータに対するレーティング予測
３．学会等名 言語処理学会第30回年次大会
４．発表年 2024年

１．発表者名
門井仁弥，南條浩輝，馬青

２．発表標題
プログラミング課題文からの重要箇所抽出

３．学会等名
言語処理学会第30回年次大会

４．発表年
2024年

１．発表者名
岡本昇也，南條浩輝，馬青

２．発表標題
BERT による系列ラベリングを用いた文法誤り検出

３．学会等名
言語処理学会第29回年次大会発表論文集（ 2023年3月） pp．1607-1611

４．発表年
2023年

１．発表者名
森廣勇樹，南條浩輝，馬青

２．発表標題
日本語レビューに対するレーティング予測の精度比較

３．学会等名
言語処理学会第29回年次大会発表論文集（ 2023年3月） pp．1686-1689

４．発表年
2023年

１．発表者名
池内省吾， 南條浩輝， 馬青

２．発表標題
BERTを用いたWeb文書からの用語検索

３．学会等名
IPSJ研究報告自然言語処理（NL） pp．1-6

４．発表年
2021年

１．発表者名
董卜睿，村田真樹，馬青

２．発表標題
機械学習と統計的検定を利用した知見獲得とその評価

３．学会等名
言語処理学会 第28回年次大会 発表論文集（2022年3月）pp. 903-908

４．発表年
2022年

１．発表者名
池内省吾，南條浩輝，馬青

２．発表標題
Web 文書からの用語検索における用語候補のランキングの検討

３．学会等名
言語処理学会 第28回年次大会 発表論文集（2022年3月）pp. 1284-1288

４．発表年
2022年

１．発表者名
三木謙志，村田真樹，馬青

２．発表標題
賛成を得やすい文章の機械学習を利用した収集と分析

３．学会等名
言語処理学会 第28回年次大会 発表論文集（2022年3月）pp. 1541-1545

４．発表年
2022年

１．発表者名
Q. Ma, M. Tsukagoshi and M. Murata

２．発表標題
Estimating Evaluation of Cosmetics Reviews with Machine Learning Methods

３．学会等名
2020 International Conference on Asian Language Processing (IALP)（国際学会）

４．発表年
2020年

１．発表者名
本田涼太，村田真樹， 馬青

２．発表標題
単語制約を用いた概念ネットワークの改良

３．学会等名
言語処理学会 第27回年次大会

４．発表年
2021年

１．発表者名
符 家俊，村田真樹，馬青

２．発表標題
単語クラスタリングによって文書情報を整理する手法の改良

３．学会等名
言語処理学会 第27回年次大会

４．発表年
2021年

１．発表者名
Qing Ma and Masaki Murata

２．発表標題
Topic Extraction and Classification for Questions Posted in Community-Based Question Answering Services

３．学会等名
2019 International Conference on Computational Science and Computational Intelligence (CSCI)（国際学会）

４．発表年
2019年

１．発表者名
Qing Ma

２．発表標題
Estimation of Twitter Opinion Trends of Approval of Political Parties Using SVM and LSTM

３．学会等名
The 19th China-Japan Joint Conference on Natural Language Processing (CJNLP2019)（招待講演）

４．発表年
2019年

１．発表者名
塚越三蘭，馬青，村田真樹

２．発表標題
化粧品レビューテキストに基づく個人化推薦システム

３．学会等名
言語処理学会第26回年次大会

４．発表年
2020年

１．発表者名
村田真樹，中原裕人，馬青

２．発表標題
機械学習と言語処理による株価予測と知識獲得

３．学会等名
言語処理学会第26回年次大会

４．発表年
2020年

１．発表者名
花岡見帆，馬青，村田真樹

２．発表標題
ツイートデータに基づく政党支持率の推定

３．学会等名
言語処理学会第26回年次大会

４．発表年
2020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
研究分担者	南條　浩輝 (Nanjo Hiroaki) (50388162)	滋賀大学・データサイエンス学系・教授 (14201)	

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