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　４ 年　 ６ 月 １７ 日現在

機関番号：３２６９２
研究種目：基盤研究(C)（一般）
研究期間：2018～2021
課題番号：１８Ｋ１１４２７

研究課題名（和文）実世界と可能世界が参照可能であるテキストの日本語モダリティ解析

研究課題名（英文）Modality analysis of Japanese texts which are accessible to both of a real world and possible worlds

研究代表者
松吉　俊（MATSUYOSHI, Suguru）

東京工科大学・メディア学部・講師

研究者番号：１０５１２１６３

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、将棋解説テキストに対して、1,622個のモダリティ表現ラベル、5,014個の事象クラスラベル、3,092個の事実性ラベルを付与し、大規模な3階層の日本語モダリティコーパスを構築した。このコーパスの複数種類のラベルをマルチタスク学習するBERTモデルにより実装したモダリティ解析システムは、F値でそれぞれ0.84、0.81、0.83を達成した。また、予測局面データを利用する解析システムは、正解率0.40を達成した。

研究成果の学術的意義や社会的意義
国内外においてモダリティ解析に関する研究は行われているが、非テキスト情報を伴ったテキストを利用した研究は一切行われていない。本研究は、将棋の局面データと予測局面データ(可能世界)を利用してこの難しい課題に世界で初めて挑戦したという学術的意義がある。コーパスの正解ラベルも既存のモダリティ辞書も利用しない解析手法が正解率0.40を達成したことを実験で示し、研究の新しい道を示唆したという学術的意義も本研究にはある。

研究成果の概要（英文）：First, we constructed a corpus of Japanese text annotated with three types of modality labels, which are a modality expression chunk label, an event class label and a factuality label. The corpus contains 1,622, 5,014 and 3,092 manual annotations for the three types, respectively. Then, we developed a Japanese modality analyzer which employs a language model BERT and a multi-task learning framework. For the above three labels, it achieved F-measures of 0.84, 0.81 and 0.83, respectively. We also developed an analyzer which adopts modal logics and uses symbol grounding of event mentions in possible worlds. It achieved an accuracy of 0.40.

研究分野：自然言語処理

キーワード：モダリティ解析　コーパス　日本語モダリティ　将棋解説文　シンボルグラウンディング　様相論理

１．研究開始当初の背景

近年、写真などに写された実世界の物事を自然言語文によって自動的に記述することに注目が集まっているが、人間は目の前に存在する物事についてのみ語るわけではない。人間は、目の前の実世界を認識したのち、推測、仮定、否定などの形式により、存在していないが重要であると判断した物事や命題について語ることができる。推測、仮定、否定などの主観的な態度を表現するために用いられる語句（「ようだ」や「わけがない」、「おそらく」等）を言語学ではモダリティ表現と呼ぶ。自然言語処理において、入力テキストのモダリティ表現を解析し、書き手の主観的な態度を認識するタスクはモダリティ解析と呼ばれ、英語や日本語において活発に研究が行われている。現在はテキストのみを対象とした研究が進められているが、モダリティ表現の認識精度は高くはない。

２．研究の目的

本研究の目的は、実世界と可能世界が参照可能であるテキストの日本語モダリティ解析である。自然言語処理における意味解析の研究を発展させるために、各可能世界における事象の成否を利用する様相論理を取り入れる。具体的には、将棋解説文データに着目して研究を遂行する。将棋解説文データは以下の 3 つの部分からなる。1 つめは、図 1 のような現在の将棋局面（駒の配置や持ち駒など）のデータ（実世界）である。2 つめは、図 1 の下側に示されるような、現在の局面に対する解説テキストである。3 つめは、図 2 に例示されるような、現在の局面に対する先読みアルゴリズムによる予測局面データ（可能世界の集合）である。予測局面データとは、対象局面を根としたゲーム木であり、各局面に対してその良さを表すスコアが付与されている。解説テキストでは、将棋の駒やその動きが淡々と述べられているだけでなく、書き手の主観的な態度も表現される。本研究では、将棋の予測局面という非テキストデータ内に出現しうる、モダリティ表現の対応事象の成否を自動認識し、様相論理を適用することで、高い精度でテキストのモダリティ解析を行うことを目指す。

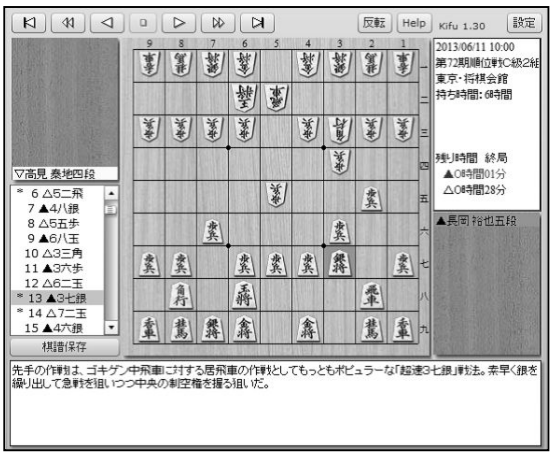

図 1: 将棋局面と解説

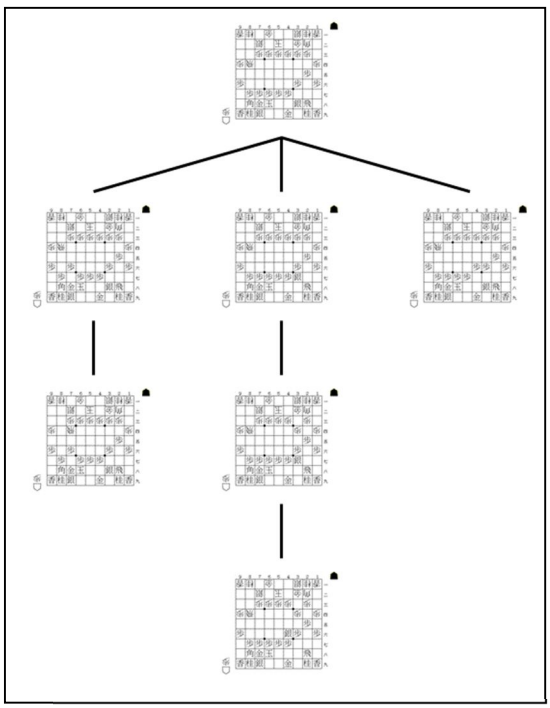

図 2: 予測局面データ(ゲーム木)

３．研究の方法

本研究では、以下の 3 点を中心に研究を進め、最終的に、非テキストデータである局面データを伴うテキストに対して高い精度でモダリティ解析を行うシステムを開発する。

（1）モダリティ情報付与コーパスの構築。既存のモダリティラベルの体系を見直し、機械学習しやすい新しい体系を定義する。将棋解説文を対象として、新しい体系のラベルを付与する。

（2）モダリティ解析システムの開発。深層学習を利用した解析システムを実装する。上記コーパスには将棋に特有の表現に関するラベルも付与されているので、それも含め 4 種類のラベルに関してマルチタスク学習するシステムを実装する。

（3）モダリティ表現の自動獲得。様相論理を利用することにより、未知のモダリティ表現を教師信号なしで自動獲得する。各局面において、将棋に関する事象が成立しているか否かを自動認識し、可能世界集合においてその成否の割合を計算することで、確信度の強さを自動推定する。獲得したモダリティ表現を利用したモダリティ解析システムを実装する。

４．研究成果

（1）新しいモダリティラベルの体系として、「肯定の可能性」、「確実な否定」、「未来」、「仮定」などのモダリティ表現ラベル、「断定」、「希望」、「疑問」、「複合辞」などの事象クラスラベル、「成立を断定」、「高い確信度で成立を推測」、「不成立を断定」などの事実性ラベルからなる 3 階層の体系を定義した。この体系に従い、将棋解説文データに対して、1,622 個のモダリティ表現ラベル、5,014 個の事象クラスラベル、3,092 個の事実性ラベルを人手で付与し、大規模なコーパスを構築した。このコーパスは、研究室のウェブサイトにて無償で提供している。

（2）深層学習に基づく言語モデルである BERT および条件付き確率場 CRF を利用して日本語のモダリティを解析する手法を提案した。将棋固有表現ラベル、モダリティ表現ラベル、事象クラスラベル、事実性ラベルの 4 つの情報をマルチタスク学習することが有効であることを明らかにした。実装した解析システムは、上記 4 つのラベルに関して、F 値でそれぞれ 0.90、0.84、0.81、0.83 を達成した。

（3）本研究では、予測局面データ（可能世界）内に出現しうる事象の成否を利用することで、様相論理の適用により未知のモダリティ表現を自動獲得する予定であった。しかしながら、ある局面に対象の事象が存在するかどうか自動判定するのは非常に難しいこと、および、プロの解説者のように状況に応じて適切にゲーム木の探索空間を絞ることが困難であることにより、研究はうまく進まなかった。その一方で、既知のモダリティ表現が後続する事象を約 5,000 個コーパスから抽出し、確信度の強さを提案手法により自動推定したところ、コーパスの正解ラベルも既存のモダリティ辞書も利用しない手法であるが、正解率 0.40 という結果を得た。このことは、上記 2 つの困難を部分的にも解決することができれば、提案手法は未知のモダリティ表現の自動獲得に有効である可能性を示唆している。今後は、上記 2 つの問題を解決する方法を検討していきたい。また、将棋に特有のモダリティ表現を自動獲得することができれば、将棋の 1 つの局面を入力としてそれを根拠を持って解説するテキストを自動生成するタスクにもその結果を応用していきたい。

５．主な発表論文等

〔雑誌論文〕　計1件（うち査読付論文　1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 亀甲 博貴，松吉 俊，John Richardson，牛久 敦，笹田 鉄郎，村脇 有吾，鶴岡 慶雅，森 信介	４．巻 28
２．論文標題 将棋解説文への固有表現・モダリティ情報アノテーション	５．発行年 2021年
３．雑誌名 自然言語処理	６．最初と最後の頁 847～873
掲載論文のDOI（デジタルオブジェクト識別子） 10.5715/jnlp.28.847	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5件（うち招待講演　0件／うち国際学会　1件）

１．発表者名 嶋田 眞巳，谷村 皓奎，松吉 俊，兼松 祥央，三上浩司
２．発表標題 シナリオのト書きを対象とした主語-述語ペア自動抽出に関する基礎調査
３．学会等名 NICOGRAPH2021 ショートペーパー予稿集
４．発表年 2021年

１．発表者名 亀甲博貴、森信介
２．発表標題 熟練者による解説文内イベントの出現とその根拠のアノテーション
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 友利涼、村脇有吾、松吉俊、亀甲博貴、森信介
２．発表標題 モダリティ表現認識・事象の事実性解析の同時学習
３．学会等名 情報処理学会 第241回自然言語処理研究会
４．発表年 2019年

１．発表者名
亀甲博貴、松吉俊、村脇有吾、森信介

２．発表標題
モンテカルロシミュレーションによる認識的モダリティ表現のグラウンディング手法の検討

３．学会等名
言語処理学会第25回年次大会

４．発表年
2019年

１．発表者名
Suguru Matsuyoshi, Hirotaka Kameko, Yugo Murawaki, and Shinsuke Mori

２．発表標題
Annotating Modality Expressions and Event Factuality for a Japanese Chess Commentary Corpus

３．学会等名
Proceedings of the 11th edition of the Language Resources and Evaluation Conference (LREC2018)（国際学会）

４．発表年
2018年

〔図書〕　計2件

１．著者名 小磯花絵、中俣尚己、木部暢子、小木曽智信、迫田久美子、佐々木藍子、細井陽子、須賀和香子、松吉俊、浅原正幸、窪薗晴夫、有田節子、益岡隆志、野田尚史、原由理枝	４．発行年 2020年
２．出版社 くろしお出版	５．総ページ数 228
３．書名 データに基づく日本語のモダリティ研究	

１．著者名 岡　照晃、伝　康晴、内元清貴、山田　篤、宇津呂武仁、松吉　俊、土屋雅稔、近藤泰弘、 坂野　収、多田知子、岡田純子、山元啓史、荻野綱男、矢澤真人、丸山直子、星野和子、小磯花絵	４．発行年 2019年
２．出版社 朝倉書店	５．総ページ数 224
３．書名 講座日本語コーパス 7 コーパスと辞書	

〔産業財産権〕

〔その他〕

アノテーション付き将棋解説文コーパス http://www.lsta.media.kyoto-u.ac.jp/resource/data/game/

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	森　信介 (MORI Shinsuke) (90456773)	京都大学・学術情報メディアセンター・教授 (14301)	
研究分担者	村脇　有吾 (MURAWAKI Yugo) (70616606)	京都大学・情報学研究科・講師 (14301)	
研究分担者	亀甲　博貴 (KAMEKO Hirotaka) (50827524)	京都大学・学術情報メディアセンター・助教 (14301)	

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