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　 ７ 日現在

機関番号：１１３０１
研究種目：基盤研究(C)（一般）
研究期間：2019～2022
課題番号：１９Ｋ１２１１２

研究課題名（和文）先行文脈から動的に得る知識と事前学習で得る静的な知識を融合した文章の意味構造解析

研究課題名（英文）Text semantic parsing combining dynamic knowledge obtained from preceding context and static knowledge obtained in pre-training

研究代表者
松林　優一郎（Matsubayashi, Yuichiroh）

東北大学・教育学研究科・准教授

研究者番号：２０５８２９０１

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）：文を「正確に読む」ために必須となる省略解析の技術の性能は、一般的な人間の読解能力と比べて大きな隔たりがあった。本研究課題では、文章の一部分だけを見て解析を行う従来型の意味解析手法を改良し、(1) 先行文脈の文意の蓄積（＝動的知識）に基づいて後方の意味を理解する解析モデルの構築と(2) 推論に必要な常識的知識（＝静的知識）の効果的・効率的な表現方法の確立、(3) これら動的知識と静的知識を融合した自然な推論に基づく意味解析の実現を目指した。本研究の成果として、従来法を大きく上回る解析精度を持った省略解析技術を実現した。最終成果である解析システムはオープンソースソフトウェアとして広く公開した。

研究成果の学術的意義や社会的意義
省略解析は文章の意味を正確に理解するAIの実現に不可欠な要素であり、日本語解析のボトルネックとなっていたこの基盤技術の解析精度向上により、応用技術の発展可能性が増大した意義は大きい。開発したシステムは一般公開し、実世界テキスト解析に適用可能である。精度向上の鍵となったアイデアは、汎用的言語モデルに対して学習の形態を大きく変更することなくシームレスに意味解析の能力を増強するものであり、その他の言語処理技術の性能向上に対しても応用可能性を秘めている。加えて、研究過程で得られた知見から書き手の省略判断分析という新たな研究の方向性を展開し、教育応用等へのシードを生んだ点も学術的意義として挙げられる。

研究成果の概要（英文）：The performance of omission analysis, which is fundamental to "accurate reading" of sentences, had a large gap compared to general human reading comprehension ability. The goal of this research was to improve the conventional semantic parsing technique, which only looks at a few sentences around the target sentence, by (1) constructing a parsing model that understands meaning based on accumulating the meaning of sentences in the preceding context (= dynamic knowledge), (2) establishing an effective and efficient method to express the common sense knowledge required for inference (= static knowledge), and (3) realizing the natural inference based on the combination of such dynamic and static knowledge. As a result, we were successful in implementing an omission analysis system that significantly outperformed previous methods' performance. Our proposed system, developed in this project, has been publicly released as open source software.

研究分野：自然言語処理

キーワード：省略解析　述語項構造　意味解析　文章理解

様　式　Ｃ－１９、Ｆ－１９－１、Ｚ－１９（共通）

１．研究開始当初の背景

機械による文章の精緻な読解を行う上で、当時最大のボトルネックだったのは文章内に明示的に書かれることのない省略された内容の理解であった。通常の言語運用では、人間が読んだときに推論で自明となる内容は記述から省かれるのが一般的である。しかしながら、この省略された内容を正しく読み取るための意味解析技術は、研究開始当初、未だ一般的な人間の読解能力と比べて深刻に低水準な状況にあった。例えば、日本語の省略された主語や目的語を補う処理の精度は、比較的分析手法の研究が進んでいる新聞記事のデータ上でも 50%に満たなかった。我が国の母語である日本語は、主語・目的語の約 40%が省略される言語であり、省略解析の精度が低水準に留まることで生じる分野全体の技術的停滞の深刻さは火を見るより明らかなため、解決が求められる火急の技術的課題であった。

意味解析技術がこのような低い解析精度に留まる現状を打開するため、研究代表者はこれまでに最新のシステムに対し数千文規模でのエラー分析を行い、結果として解析エラーの大部分は、現状の解析器が (1)文脈を考慮する機構と (2)世界知識を保持・利用する機構を持ち合わせていない点に起因することを明らかにしてきた。既存モデルには、先行文脈の文の意味を動的知識として保持しつつ後の文の意味構造解析に利用する機構が不足していた。既存の意味解析モデルの大部分は１～２文単位の処理が中心であり、通常人間が自然に考えるような複数文の意味のつながりや文章全体での焦点の推移などの談話的な要素を考慮していなかった。また、もう一つの課題点は、人間であれば経験的に知っている大量の常識的知識を何らかの方法で保持し、実際の意味計算モデルの中で利用するための適切な表現方法・計算方法を確立することであった。

２．研究の目的

本研究では文章内の意味のつながりと常識的知識のつながりをモデリングする効果的な手法を模索することによって、より優れた意味解析モデルの実現を目指した。本研究課題の核となる研究目標は次の３点である。

目標(1)　「先行文脈（動的知識）を踏まえる意味解析モデルの実現」前方の文章から得る知識を後方の文章の意味解析に役立たせるための優れた表現形式と計算方法の確立
目標(2)　「テンプレート非依存な表現力の高い静的知識ベースの確立」従来法のような予め決められた型（テンプレート）を持たない常識的知識のより柔軟な表現手法の確立
目標(3)　「最先端の意味解析手法における動的知識と静的知識の融合」最新のニューラルネットベースの解析器と親和性の高い知識表現の保持方法、および解析する文章に応じた適切な知識活用方法の確立

３．研究の方法

目標(1) 先行文脈（動的知識）を踏まえる意味解析モデルの実現

文の意味構造解析技術を、これまでの主流であった１～２文単位の解析から、前方の文章で読み取った情報を伝播させながら次の文を解析する手法に拡張する。手法の核となるアイデアは (a) 解析の基本単位である単語の意味を表現するベクトルに、その文で言及された内容を追記する文脈埋め込みモデルと、(b) 得られた文脈付きの単語ベクトルを後の文の解析でどの程度利用するかを決める文脈結合モデルを解析モデルに組み込むことである。これらの文脈情報の埋め込みと結合のモデルは、これまでに申請者が取り組んできた複数の言明間の関係を計算するモデルの自然な拡張として実現を試みる。また、文章の主題の移り変わりや主節・補文などの文構造に応じて各トピックへの着目度を計算することで、意味構造・統語構造・談話構造の３つの観点を組み込んだ自然な読解モデルを実現する。

目標(2)　テンプレート非依存な表現力の高い静的知識ベースの確立

事象間の因果関係などの常識的知識の獲得方法として、Web 上の大規模な文書データからこれらを収集する手法が種々考案されてきたが、多くは（主語, 述語, 目的語）の三つ組のような特定のテンプレートの形で莫大な量の関係を離散的に表現するものであった。このような方法の問題点は、(a) 予め決められたテンプレートに沿った知識しか表現できず、細部が捨象されてしまうことで獲得した一部の知識が人間の直感と矛盾してしまうことと (b) 膨大な言語表現の種類に由来して莫大なメモリ空間を要することにある。
本研究課題ではこの２つの問題を解決するために、近年研究が進んでいる言語表現（フレーズ）全体をベクトル空間に埋め込む技術を応用し、数億ページ規模の超大規模文書データからの教師なし学習によって、決まった型によらず、任意のフレーズと密接な関係のある別のフレーズを予測する高密度かつコンパクトなマトリックス（事象間の因果関係や二項間関係などの常識的

知識の縮図）を獲得することを目指す。この技術によって細部のニュアンスをできるだけ落とさない正確な知識推論の実現を目指す。

目標(3) 最先端の意味解析手法における動的知識と静的知識の融合
目標(2) で確立する知識ベースモデルを目標(1)で開発する意味構造計算モデルに統合するための計算モデルを構築し、実際の解析における常識的知識の有効性を実証する（図２中央)。基本的なアイデアは目標(2)のモデルで前方のフレーズから予測される後方フレーズの意味表現と、現在解析中の文の解析結果との一貫性が保たれるように解析システムを訓練することであり、具体的な計算モデルは目標(1) 目標(2) で構築する計算モデルに応じて柔軟に検討する。最終的に、この前方文脈と常識的知識を同時に利用して意味解析を行うシステムを開発し、リリースするところまでを本研究課題の成果目標とする。

４．研究成果
研究初期においては、省略解析モデルで一般的に入力として用いられている、文中の各単語の性質を数理的に表現しているベクトルを、現在解析を行っている文の文脈に応じて適切に変化させる文脈埋め込みモデルを用いた省略解析技術を構築し、この事によって従来モデルに比べて、特に日本語省略解析の精度を大幅に改善することに成功した。また、省略解析に必要な常識的知識の効果的な学習方法の一つとして、事前に大規模データを用いて学習された言語モデルを使い、教師データ中の単語の一部を別の自然な語で言い換える事により、省略解析のための教師データを増強する手法を提案し、性能向上に有効であることを示した（図１)。加えて、文中の省略箇所について、より推定の簡単な場所から徐々に期待度を上げながら反復的に解析を行うモデルを設計し、その効果を検証した（図２)。

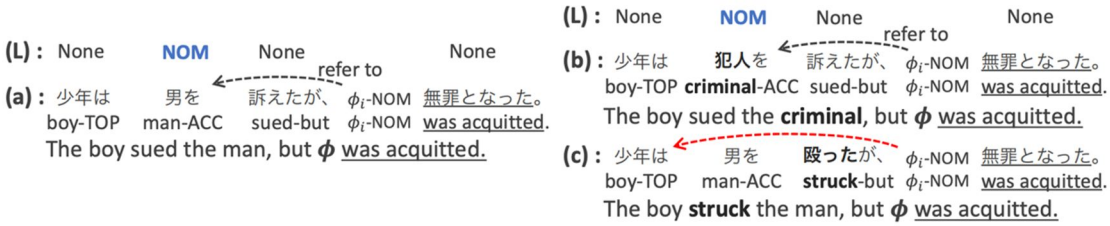

図 1 教師データ中の単語を自然な語で入れ替えることによる省略解析のためのデータ増強

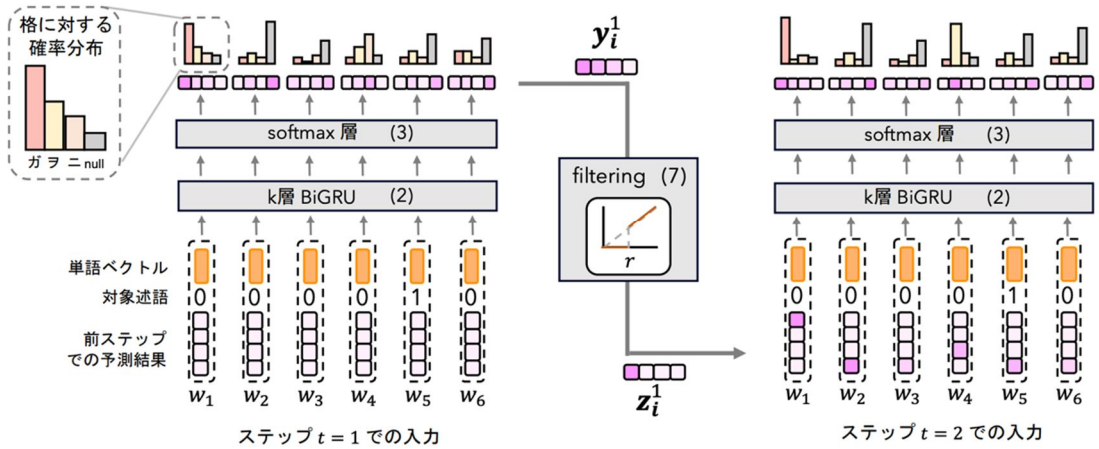

図 2 推定の簡単な場所から反復的に省略解析を行うモデル

次に、前述の成果によって得られた学習手法の知見から、推論に必要な常識的知識 （＝静的知識） の効果的・効率的な表現方法として、大規模データによって学習される言語モデルについて、省略解析のために必要な知識に重点的に学習バイアスをかけながら追加の学習を行うことで実現する方針を取った。前年度作成した目標(1) の学習部分と、 目標(2)で用いる学習手法の双方に対して同一のタスク形式を用いて定式化を行うことにより、最終的に目指す静的知識と動的知識の自然な融合を実現する新たな手法を考え、これを実装した。これらの手法を用いて実験を行い、結果として、日本語の省略解析タスクにおいて昨年度の結果をさらに超えて、これまでの結果を大きく上回る解析精度を実現することに成功した。また、ここで開発した省略解析シ

ステムを成熟させ、オープンソースソフトウェアとして広く公開した。

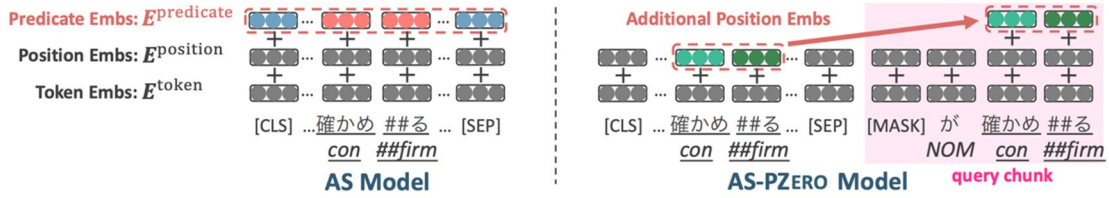

図 3 省略解析に必要な知識にバイアスをかけた言語モデル学習

加えて、この省略解析に関する課題を発展させ、読解時だけでなく、文章記述時にどのような方法で記述内容の省略の是非を決定するかを説明するための計算モデルを試作した。人間が文章内の内容を省略すべきかそうでないかの妥当性を判断したアノテーションデータを収集するため、第一段階ではパイロット版を作成し、その結果を受けて、より高品質で詳細な分析が可能なデータを収集するための収集のプロセスを改善方法を模索し、結果として高品質なデータの作成を達成した。このデータを根拠に、省略の是非を説明するためのいくつかの仮説を検証した。人間の読者間の判断のぶれ、著者と読者間のぶれ、判断の根拠となる因子等の分析を行うとともに、人間の省略判断 を模倣する計算モデルを作成し、その性能について細部の分析を行った。結果として、現在の最先端の日本語言語モデルは、人間の判断をある程度は模倣できるが、同等の能力には至らないことが分かった。作成したデータや研究成果はそれぞれ研究資源、研究論文として公開した。

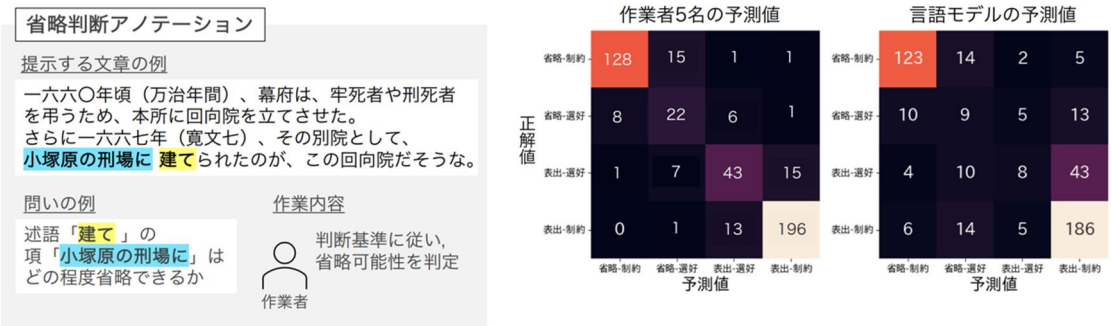

図 4 文章記述時の人間の省略判断に関するデータ収集と分析

５．主な発表論文等

〔雑誌論文〕　計6件（うち査読付論文　6件／うち国際共著　0件／うちオープンアクセス　6件）

１．著者名 Funayama Hiroaki、Sato Tasuku、Matsubayashi Yuichiroh、Mizumoto Tomoya、Suzuki Jun、Inui Kentaro	４．巻 13355
２．論文標題 Balancing Cost and?Quality: An Exploration of?Human-in-the-Loop Frameworks for?Automated Short Answer Scoring	５．発行年 2022年
３．雑誌名 Artificial Intelligence in Education. AIED 2022. Lecture Notes in Computer Science	６．最初と最後の頁 465～476
掲載論文のDOI（デジタルオブジェクト識別子） 10.1007/978-3-031-11644-5_38	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 －

１．著者名 Hiroaki Funayama, Yuya Asazuma, Yuichiroh Matsubayashi, Tomoya Mizumoto and Kentaro Inui	４．巻 --
２．論文標題 Reducing the Cost: Cross-Prompt Pre-Finetuning for Short Answer Scoring	５．発行年 2023年
３．雑誌名 Artificial Intelligence in Education. AIED 2023.	６．最初と最後の頁 12 pages
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 オープンアクセスとしている（また、その予定である）	国際共著 －

１．著者名 Konno Ryuto、Kiyono Shun、Matsubayashi Yuichiroh、Ouchi Hiroki、Inui Kentaro	４．巻 1
２．論文標題 Pseudo Zero Pronoun Resolution Improves Zero Anaphora Resolution	５．発行年 2021年
３．雑誌名 Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing	６．最初と最後の頁 3790-3806
掲載論文のDOI（デジタルオブジェクト識別子） 10.18653/v1/2021.emnlp-main.308	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 －

１．著者名 Konno Ryuto、Matsubayashi Yuichiroh、Kiyono Shun、Ouchi Hiroki、Takahashi Ryo、Inui Kentaro	４．巻 1
２．論文標題 An Empirical Study of Contextual Data Augmentation for Japanese Zero Anaphora Resolution	５．発行年 2020年
３．雑誌名 Proceedings of the 28th International Conference on Computational Linguistics	６．最初と最後の頁 4956-4968
掲載論文のDOI（デジタルオブジェクト識別子） 10.18653/v1/2020.coling-main.435	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 －

１．著者名 Funayama Hiroaki、Sasaki Shota、Matsubayashi Yuichiroh、Mizumoto Tomoya、Suzuki Jun、Mita Masato、Inui Kentaro	４．巻 1
２．論文標題 Preventing Critical Scoring Errors in Short Answer Scoring with Confidence Estimation	５．発行年 2020年
３．雑誌名 Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics: Student Research Workshop	６．最初と最後の頁 237-243
掲載論文のDOI（デジタルオブジェクト識別子） 10.18653/v1/2020.acl-srw.32	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 －

１．著者名 Abe Kaori、Matsubayashi Yuichiroh、Okazaki Naoaki、Inui Kentaro	４．巻 27
２．論文標題 Multi-dialect Neural Machine Translation for 48 Low-resource Japanese Dialects	５．発行年 2020年
３．雑誌名 Journal of Natural Language Processing	６．最初と最後の頁 781～800
掲載論文のDOI（デジタルオブジェクト識別子） 10.5715/jnlp.27.781	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11件（うち招待講演　0件／うち国際学会　0件）

１．発表者名 石月由紀子，栗林樹生，松林優一郎，大関洋平
２．発表標題 情報量に基づく日本語項省略の分析
３．学会等名 言語処理学会第28回年次大会
４．発表年 2022年

１．発表者名 舟山弘晃，佐藤汰亮，松林優一郎，水本智也，鈴木潤，乾健太郎
２．発表標題 記述式答案自動採点における確信度推定とその役割
３．学会等名 言語処理学会第28回年次大会
４．発表年 2022年

１．発表者名
今野颯人，松林優一郎，清野舜，大内啓樹，乾健太郎

２．発表標題
事前学習とfinetuningの類似性に基づくゼロ照応解析

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
菊地正弥，尾中大介，舟山弘晃，松林優一郎，乾健太郎

２．発表標題
項目採点技術に基づいた和文英訳答案の自動採点

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
舟山弘晃，王天奇，松林優一郎，水本智也，佐藤汰亮，鈴木潤，乾健太郎

２．発表標題
実用的な自動採点のための確信度推定と根拠事例の提供

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
今野颯人，松林優一郎，清野舜，大内啓樹，高橋諒，乾健太郎

２．発表標題
マスク言語モデルを利用したデータ拡張に基づく日本語文内ゼロ照応解析

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
宮脇峻平，清野舜，松林優一郎，今野颯人，高橋諒，大内啓樹，乾健太郎

２．発表標題
反復改良法を用いた日本語述語項構造解析

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
舟山弘晃，佐々木翔大，水本智也，三田雅人，鈴木潤，松林優一郎，乾健太郎

２．発表標題
記述式答案自動採点のための確信度推定手法の検討

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
石月由紀子，栗林樹生，松林優一郎，笹野遼平，乾健太郎

２．発表標題
日本語話者の項省略判断に関するアノテーションとモデリング

３．学会等名
言語処理学会第29回年次大会

４．発表年
2023年

１．発表者名
岩瀬裕哉，舟山弘晃，松林優一郎，乾健太郎

２．発表標題
文章構造グラフを用いた国語記述式答案への自動フィードバック生成

３．学会等名
言語処理学会第29回年次大会

４．発表年
2023年

１．発表者名 Hiroaki Funayama, Yuya Asazuma, Yuichiroh Matsubayashi, Tomoya Mizumoto, Kentaro Inui.
２．発表標題 What can Short Answer Scoring Models Learn from Cross-prompt Training Data?
３．学会等名 言語処理学会第29回年次大会
４．発表年 2023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

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