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　　元 年　　５ 月 ３１ 日現在

機関番号：８２４０１
研究種目：若手研究(B)
研究期間：2015～2018
課題番号：１５Ｋ１６０４５

研究課題名（和文）超大規模コーパスからの統計的知識獲得と構成論的記号計算の融合による句の類義性認識

研究課題名（英文）Detecting phrase similarity based on compositional symbolic computation and statistical knowledge acquisition using a large-scale corpus.

研究代表者

松林　優一郎（Matsubayashi, Yuichiroh）

国立研究開発法人理化学研究所・革新知能統合研究センター・研究員

研究者番号：２０５８２９０１

交付決定額（研究期間全体）：（直接経費）　　3,100,000円

研究成果の概要（和文）：本研究では、多様な言語表現を頑健に取り扱うことの出来る日本語の意味解析器の確立に向けて、複数の述語の組合せを含むような句、文、複数文の談話的な構造といったより広い範囲の言語構造に焦点を当て、それらの意味計算のモデルを構築することを目指した。具体的には、語の分散表現を計算の主軸として、これを用いて（1）句やイベントの意味構造を解析する技術の構築と改善、(2) 文をまたぐような談話的な意味の繋がりを考慮するための計算モデルの改善、(3) 言語情報と非言語マルチモーダル情報の組み合わせで意味構造をとらえる計算モデル、の３つの課題に取り組み、12本の査読付き論文の成果につなげた。

研究成果の学術的意義や社会的意義
(1)で開発した意味解析器では、現状の日本語解析における重大なボトルネックである「省略された内容を補う」解析の精度を当初の40%強から60%弱まで飛躍的に向上させ実用レベルに近づけた。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。(2)では文の解析に先行文脈の情報を利用するという近年取り組みが減っていたアイデアに再注目、ハイライトした。このアイデアに基づく研究の数は徐々に増えている。(3) では、音楽、絵本の２つの題材を取り上げ、言語の構造とそれを取り巻く情報の構造を相互に考慮することが言語構造の推定に重要な役割を果たすことを示し、分野での先行事例としての役割を果たした。

研究成果の概要（英文）：In this research, in order to establish a semantic parser that can handle various linguistic expressions robustly, we have aimed to build models that can handle semantics of rather longer expressions than words, such as phrases, sentences, and discourse structures of multiple sentences. Specifically, employing distributed word representations as a key technology, we (1) improved semantic structure analysis models for phrases including multiple events, (2) improved semantic models that can handle inter-sentential semantic structures, and (3) proposed computational models that incorporate linguistic and non-linguistic multimodal structures. The contribution in this research project was presented in 26 publications including 11 peer-reviewed papers.

研究分野：自然言語処理

キーワード：自然言語処理　人工知能　項構造解析　談話　知識

様　式　Ｃ－１９、Ｆ－１９－１、Ｚ－１９、ＣＫ－１９（共通）

１．研究開始当初の背景

Web の爆発的な発達と共に、個人の意見や専門家の知見など様々な情報がテキストの形で発信され、人々にとって有用な知識が手近に入手可能となった。その莫大なデータ量は世界知識の縮図とも言える程になり、大量のテキストを解析し、統計的な裏付けを得ることにより、世論の動向や細やかな専門知識をある程度の信頼度を保証し収集することが可能になりつつあった。

一方で、Web 上の情報は玉石混交であるため、記述された事柄を妥当な知識・意見だと認定するためには、同一の意味を持つ言論に対して一定の統計量（根拠）を得る必要がある。しかし、言語の表現は非常に多種多様であり、類似の意味を表現するために用いられる文字列には膨大な類型がある。この類型を正しく認識できないために、結果として、同じ意味を持つ大量の表現一つ一つが、出現頻度が少ない別の表現と捉えられてしまう問題があった。このような事情から、多様な類義表現のロングテール部分をいかに同定し、価値ある情報をすくい取るかということが情報爆発時代の知識処理において欠かすことの出来ない重要課題となっていた。

２．研究の目的

本研究の目的は、Web 規模の多様な言語表現を頑健に取り扱うことの出来る日本語の意味解析器の確立に向けて、特に複雑な意味構造を取りやすい、述語の組合せによって成り立つフレーズに焦点を当て、複合的な句の意味をその構成要素の単語の組み合わせを考慮して精密に求める方法を確立することである。研究開始当初は、述語とその項との意味関係を説明する語彙概念構造理論を拡張することで、複合的な句の意味をその構成要素の単語の意味表現から構成的かつ精密に求める方法を確立することを検討していたが、研究の過程で、語や句の性質・意味を実数値ベクトルで表現する分散表現の手法が自然言語処理業界全体で飛躍的に発展し、意味表現･意味計算に関する研究に､ある種のブレイクスルーと劇的な変革をもたらしたことから、当初の研究計画を変更し、語の分散表現を計算の主軸として、これを用いて (1) 句やイベントの意味構造を解析する技術の構築と改善、(2) 文をまたぐような談話的な意味の繋がりを考慮するための計算モデルの改善、(3) 言語情報と非言語マルチモーダル情報の組み合わせで意味構造をとらえる計算モデル、といった、従来中心的に扱われてきた句や文といった単位より、より広範囲の文章表現に対する意味計算を実現するための研究に取り組んだ。

３．研究の方法

(1) の「句やイベントの意味構造を解析する技術の構築と改善」では、これまで申請者が蓄積してきた述語を中心とした意味解析技術の知見を生かして、直接的な記述を伴わない省略などの間接的な意味情報を解析する技術の性能改善を試みた。(2) の「文をまたぐような談話的な意味の繋がりを考慮するための計算モデルの改善」では、特に文章内にある述語間の意味的関係や、段落という文のまとまりごとの意味的関係を考慮することで、文の単位を超えた意味的な関係をモデル化する手法を構築した。(3) の「言語情報と非言語マルチモーダル情報の組み合わせで意味構造をとらえる計算モデル」では、言語と音楽、言語と絵本、の２つの題材を取り上げて、言語の構造とそれを取り巻く情報の中にある構造を相互に考慮することで、言語の構造を考慮するときに補助的かつ重要な役割を果たすことを確かめた。

４．研究成果

(1) 句やイベントの意味構造を解析する技術の構築と改善

①．研究の序盤では、実際の意味解析技術である述語項構造解析と時間関係認識の二つの解析技術に関して、その最先端のシステムにおける現状の問題点を分析し、性能改善のための手がかりを詳細に類型化して報告することで、コミュニティー全体への意識共有を促した。具体的には、項の省略や、述語相当の語が名詞化されている場合に発生する項の機能的曖昧性の解消問題、複数の述語間で項が共有する現象についてそのエラーの原因を詳細に報告し、解析の手がかりとなる新たな情報としてどのような知識を用いるべきかを整理した（松林ら 2015)。また、時間関係認識技術においては、アスペクトやモダリティなどの文法的機能が、句が表す事態の時間的な関係にどのように関係するのかについて、実際の解析モデルを構築しながら計算モデルを分析した(稲田，松林ら 2015)。

②．上記で行った最先端のシステムにおける現状の問題点の分析結果にしたがい、計算モデルの改善を行った。具体的には、述語項構造解析技術において、ニューラルネットワークを用いたモデリングにより、これまで人手により設計されていた特徴量の組み合わせ方法を自動学習する手法と、これまで記号的に表現されていた特徴を数値ベクトルとして表現する手法の導入により、これまでの解析手法と比べ大幅な性能改善を達成し、世界最高精度を実現した（松林，乾 2017; Matsubayashi and Inui 2017)。また、この解析器を事前の構文解析処理を必要としない end-to-end と呼ばれる方式に改良し、文中の複数の述語間の相互作用をモデル化することで、特に日本語に頻出する省略表現の解析精度をさらに飛躍的に向上させることに成功した（図 1）(Matsubayashi and Inui 2018)。その後さらに、このシステムを通常の述語のみならず述語が名詞化されたものの項構造も解析できるよ

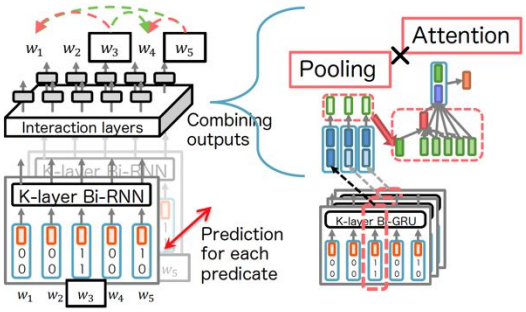

図 1 文中の複数の述語間の相互作用を考慮する述語項構造解析モデル

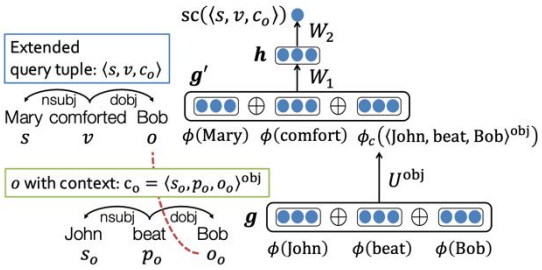

図 2 前方文脈を考慮して述語と項の組み合わせの適切さを評価する計算モデル

う拡張し、そのコードを広く利用可能な形で一般公開している。

③. 日本各地の方言による対話データを書き起こした文章を標準語に翻訳する多言語翻訳の研究を進め、語順の変化がほとんどなく音韻的な変化が中心となる日本語方言においては、文節単位の逐語翻訳が極めて有効な手段であることを突き止めた。また、類似する方言間のデータを共有することで効率的な翻訳が行えることが確かめられた（阿部，松林ら 2018; Abe, Matsubayashi, Okazaki, and Inui 2018)。

(2) 文をまたぐような談話的な意味の繋がりを考慮するための計算モデルの改善

①. 談話的な意味関係を計算するモデルの一例として、述語とその項となる名詞の間の意味的な関係を表現する選択選考モデルを題材として研究を行った。単語の分散表現に基づく句の意味計算モデルを用いて、述語と項の組み合わせに関する妥当性や意味的な類似性を表現する計算モデルを構築した。具体的には、ニューラルネットワークを用いて大規模データによる統計情報を単語横断的に一般化する表現手法により、従来法に比べ述語と項の組み合わせの妥当性をより的確に評価できるモデルを実現した（図 2）（大野，井上，松林ら 2016; Inoue, Matsubayashi, et al. 2016)。

②. 明確な構造を持った文章の談話的な意味計算の一例として、歌詞の文章構造に着目し、歌詞の意味的なつながりをモデル化する研究を進めた。具体的には、A メロ、B メロ、サビといった歌詞のブロックの中にある隠れた意味的トピックの遷移や、歌詞の繰り返しの構造をモデル化することにより、このような構造的な関係をモデル化しない従来手法に比べてより自然な歌詞を生成できることを示した。（図 3）（渡邉，松林ら 2016; Watanabe, Matsubayashi, et al. 2016)

(3) 言語情報と非言語マルチモーダル情報の組み合わせで意味構造をとらえる計算モデル

①. 言語外の情報と文の意味や構造の一貫性を計算する応用研究の一環として、歌詞生成を題材に、入力されたメロディに対してメロディの並びや休符の位置などの音楽的構造と一貫性を持つ歌詞を出力する文章生成モデルを実現した（Watanabe, Matsubayashi, et al. 2018)。このモデルは、従来手法であるメロディーのモーラ数のみを考慮する歌詞生成モデルと比べて、音の構造的な区切れで歌詞の構造も区切れる非常に直感的に歌いやすい歌詞を生成できることがわかり、言語の構造と非言語情報の構造との対応関係を取ることが、相互の構造を説明する際の重要な要素であることを実例的に示した。

②. 日本語のように省略表現が多用される言語において、人間が述語の意味や項構造、文法的振る舞いのルールを獲得するのに必要となる要素を分析するために、絵本に対する項構造アノテーションを実施し、その構築方法とデータの性質を報告した（図 4）（折田，石井，鈴木，松林 2018)。また、このデータを利用して、子供が言語獲得を行う際に、多言語で

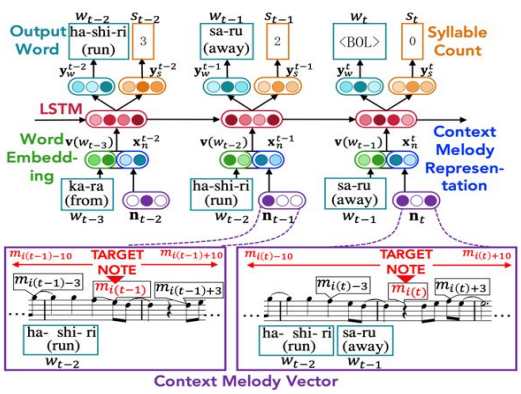

図 3 入力されたメロディの構造を考慮して適切な構造の歌詞を生成するモデル

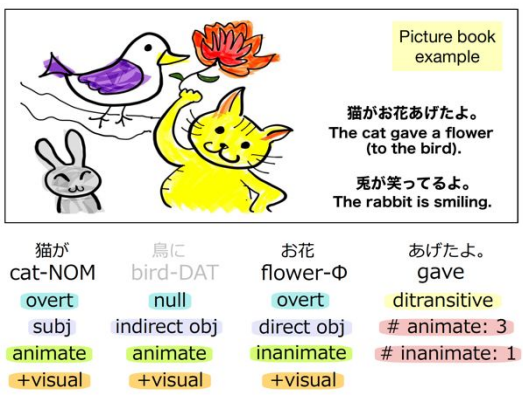

図 4 絵本に対する述語項構造と非言語情報のアノテーション

言われているような文法推論に基づく補助現象が同様に起こりうるかを分析し、項の省略が頻繁に起こっている状況下でも視覚情報（例えば、絵に生き物が出ているかどうか）等の他の情報を共に考慮することで、少なくとも潜在的には同等の文法推論を行えることを客観的に示した (Orita, Suzuki, and Matsubayashi 2019)。

５．主な発表論文等

〔雑誌論文〕（計 12 件）

1. Naho Orita, Asumi Suzuki, Yuichiro Matsubayashi. Verb arguments in Japanese picture books. In Proceedings of the 41st Annual Meeting of the Cognitive Science Society (CogSci 2019), July 2019. [査読あり]
2. Kaori Abe, Yuichiroh Matsubayashi, Naoaki Okazaki, and Kentaro Inui. Multi-dialect Neural Machine Translation and Dialectometry. In Proceedings of the 32nd Pacific Asia Conference on Language, Information and Computation (PACLIC 32), 10 pages, December 2018. [査読あり]
3. Yuichiroh Matsubayashi and Kentaro Inui. Distance-Free Modeling of Multi-Predicate Interactions in End- to-End Japanese Predicate-Argument Structure Analysis. In Proceedings of the 27th International Conference on Computational Linguistics (COLING), pp. 94-106, 2018. [査読あり]
4. Kento Watanabe, Yuichiroh Matsubayashi, Satoru Fukayama, Masataka Goto, Kentaro Inui and Tomoyasu Nakano. A Melody-conditioned Lyrics Language Model. In Proceedings of the 16th Annual Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (NAACL-HLT), pp. 163-172, 2018. [査読あり]
5. Kento Watanabe, Yuichiroh Matsubayashi, Kentaro Inui, Satoru Fukayama, Tomoyasu Nakano, and Masataka Goto. Modeling Storylines in Lyrics. IEICE Transactions on Information and Systems. Vol.E101-D, No.4, 13 pages, 2017. [査読あり]
6. Yuichiroh Matsubayashi and Kentaro Inui. Revisiting Design Issues of Local Models for Japanese Predicate-Argument Structure Analysis. In Proceedings of the 8th International Joint Conference on Natural Language Processing (IJCNLP), pp. 128-133, 2017. [査読あり]
7. Kento Watanabe, Yuichiroh Matsubayashi, Kentaro Inui, Tomoyasu Nakano, Satoru Fukayama and Masataka Goto. LyriSys: An Interactive SupportSystem for Writing Lyrics Based on Topic Transition. In Proceedings of the 22nd Annual Meeting of the Intelligent User Interfaces Community (IUI 2017), pp.559–563, 2017. [査読あり]
8. Naoya Inoue, Yuichiroh Matsubayashi, Masayuki Ono, Naoaki Okazaki and Kentaro Inui. Modeling Context- sensitive Selectional Preference with Distributed Representations. In Proceedings of the 26th International Conference on Computational Linguistics (COLING 2016), pp. 2829-2838, 2016. [査読あり]
9. Kento Watanabe, Yuichiroh Matsubayashi, Naho Orita, Naoaki Okazaki, Kentaro Inui, Satoru Fukayama, Tomoyasu Nakano, Jordan B. L. Smith and Masataka Goto. Modeling Discourse Segments in Lyrics Using Repeated Patterns. In Proceedings of the 26th International Conference on Computational Linguistics (COLING 2016), pp.1959-1969, 2016. [査読あり]
10. 松林優一郎，中山周，乾健太郎．日本語述語項構造解析タスクにおける項の省略を伴う事例の分析．自然言語処理．22 巻 5 号．pp.433-463，2015．[査読あり]
11. 稲田和明，松林優一郎，乾健太郎．日本語文書内で表現される事象間の時間的な順序関係の推定．情報処理学会論文誌，Vol.56，No.10，pp.2054-2071，2015．[査読あり]
12. 松林優一郎．特集「自然言語処理技術の現状と展望（エラー分析プロジェクトを通して）」-述語項構造解析-．情報処理，57 巻 1 号，pp. 14-15，2015.

〔学会発表〕（計 13 件）

1. 今野颯人，松林優一郎，大内啓樹，清野舜，乾健太郎．前方文脈の埋め込みを利用した日本語述語項構造解析．言語処理学会第 25 回年次大会，pp.53-56，March 2019.
2. 松林優一郎，乾健太郎．複数の述語間の関係を考慮した End-to-End 日本語述語項構造解析．言語処理学会第 24 回年次大会，pp.101-104，2018.
3. 阿部香央莉，松林優一郎，岡崎直観，乾健太郎．ニューラルネットを用いた多方言の翻訳と類型分析．言語処理学会第 24 回年次大会，pp.304-307，2018.
4. 折田奈甫，石井啓太，鈴木あすみ，松林優一郎．絵本述語項構造コーパスの設計と構築．言語処理学会第 24 回年次大会，pp.432-435，2018.
5. 渡邉研斗，松林優一郎，深山覚，乾健太郎，後藤真孝，中野倫靖．メロディ条件付き歌詞言語モデル．言語処理学会第 24 回年次大会，pp.1296-1299，2018.
6. 浅野広樹，水本智也，松林優一郎，乾健太郎．文法誤り訂正の文単位評価におけるリフ

ァレンスレス手法の評価性能．情報処理学会第 234 回自然言語処理研究会，pp.1-8，2017．
7. 菊地凜，松林優一郎，乾健太郎．日本語メタファー表現に対する分散表現の評価と考察．第 12 回 NLP 若手の会 シンポジウム（YANS），2017．
8. 高橋諒，松林優一郎，中山周，田然，乾健太郎．Dependency-based Compositional Semantics Vector に基づく文脈付き選択選好モデル．第 12 回 NLP 若手の会 シンポジウム（YANS），2017．
9. 渡邉研斗、松林優一郎、深山覚、中野倫靖、後藤真孝、乾健太郎．メロディと歌詞の相関に基づく自動歌詞生成．情報処理学会第 231 回 自然言語処理研究会，pp.1-12，2017．
10. 松林優一郎，乾健太郎．ニューラルネットワークによる日本語述語項構造解析の素性の汎化．言語処理学会第 23 回年次大会，pp.394-397，2017．
11. 渡邉研斗，松林優一郎，乾健太郎，深山覚，中野倫靖，後藤真孝．ストーリー展開と一貫性を同時に考慮した歌詞生成モデル．人工知能学会第 30 回全国大会，4 pages，2016．
12. 大野雅之，井之上直也，松林優一郎，岡崎直観，乾健太郎．分散表現による文脈情報を用いた選択選好モデル.言語処理学会第 22 回年次大会，pp.885-888，2016．
13. 大野雅之，井之上直也，松林優一郎，岡崎直観，乾健太郎．分散表現に基づく選択選好モデルの文脈化．情報処理学会研究報告．自然言語処理研究会報告，Vol.2016-NL-225，No.1，pp.1-9，2016．

〔図書〕（計 1 件）
1. 福原裕一，松林優一郎，乾健太郎．自然言語処理における意味・談話情報のコーパスアノテーション．小川芳樹，長野明子，菊地朗（編），コーパスからわかる言語変化・変異と言語理論，Part V，pp．423-442，開拓社，2016．

〔その他〕

受賞（計 5 件）
1. 情報処理学会 自然言語処理研究会 学生奨励賞 2017 年 5 月 16 日. 渡邉研斗, 松林優一郎, 深山覚, 中野倫靖, 後藤真孝, 乾健太郎.
2. 言語処理学会第 23 回年次大会 優秀賞, 2017 年 3 月 16 日. 松林優一郎, 乾健太郎.
3. 言語処理学会第 22 回年次大会 最優秀賞, 2016 年 3 月 10 日. 大野雅之, 井之上直也, 松林優一郎, 岡崎直観, 乾健太郎.
4. 情報処理学会 自然言語処理研究会 優秀研究賞 2016 年 1 月 22 日. 大野雅之, 井之上直也, 松林優一郎, 岡崎直観, 乾健太郎.
5. NLP 若手の会 第 10 回シンポジウム デモ賞 2015 年 9 月 5 日. 渡邉研斗, 松林優一郎, 乾健太郎, 深山覚, 中野倫靖, 後藤真孝.

招待講演（計 1 件）
1. 乾健太郎，松林優一郎．言語コーパスへの重層的意味情報付与 ～自然言語処理から見たコーパス分析～．言語変化・変異研究ユニット第二回ワークショップ「コーパスからわかる言語の可変性と普遍性」，2015．

ホームページ等
1. 公開ソフトウェア Showcase: Japanese Predicate-Argument Structure Analyzer. http://www.cl.ecei.tohoku.ac.jp/showcase/

６．研究組織

(1)研究分担者　　該当なし
(2)研究協力者　　該当なし