

科学研究費助成事業 研究成果報告書

平成 30 年 6 月 27 日現在

機関番号：32692

研究種目：若手研究(B)

研究期間：2015～2017

課題番号：15K16787

研究課題名(和文) 連想インタラクションに基づく日本語語彙学習システムの開発

研究課題名(英文) Development of Japanese Vocabulary Learning System based on Associative Interaction

研究代表者

寺岡 丈博 (TERAOKA, Takehiro)

東京工科大学・メディア学部・助教

研究者番号：30617329

交付決定額(研究期間全体)：(直接経費) 1,900,000円

研究成果の概要(和文)：本研究の目的は、日本語学習者が日本語母語話者による言葉の連想に沿って語彙を学習する支援システムを開発することである。研究期間を通して、大規模な連想実験を実施し、日本語学習用に必要とされる基本動詞を全て満たす規模まで動詞連想概念辞書を拡張した。また、連想情報を換喩解析に適用して精度向上を確認するとともに、品詞間における単語関係の比較を通じて連想情報の特長を示した。そして最後に、これらの特長を応用し、日本語語彙学習システムを試作した。

研究成果の概要(英文)：Our purpose is to develop a support system which learners of Japanese can learn vocabulary according to association of words by Japanese native speakers. During the research period, we conducted large-scale association experiments and extended the Associative Concept Dictionary for Verbs to the scale that satisfied basic verbs required for Japanese learning. We also applied its associative information to metonymy analysis and confirmed the improvement of accuracy. Moreover, we showed the features of associative information through comparing word relations between verbs and nouns. Finally, by applying these features, we tried to develop the Japanese vocabulary learning system.

研究分野：自然言語処理

キーワード：連想 動詞 換喩解析 共起 語彙学習

1. 研究開始当初の背景

海外の日本語学習者は、2012年には約399万人まで増加しており、半数が大学生・大学院生ということからも、第二言語習得の点で日本語学習の需要が高まってきている。そのような中、近年、自然言語処理技術が日本語学習の支援ツール・教材メディアとして応用されてきているが、他の主要言語と比べると言語研究・教育研究の成果に沿った辞書・コーパスの開発は遅れているのが実情である。

従来の日本語学習に関する辞書・コーパスや学習支援システムでは、単語の共起情報を用いたものが多く見られる。これらは、母語話者あるいは学習者によって書かれた文やwebサイト上の文から抽出されたものが多く、そのような意味では書き言葉としてのコンテキストに基づいているといえる。

2. 研究の目的

本研究の目的は、従来使用されてきたような共起情報とは異なる連想情報を用いて、語彙学習システムを提案・開発することである。

近年、研究代表者は基本動詞に関して大規模な連想実験のデータから動詞連想概念辞書を構築し、言語処理システムに適用してきた。この動詞連想概念辞書は、動詞と深層格に対して人間が連想して答えた連想語の頻度や順位の情報を用いて構築されており、日本語母語話者約2,300人分のデータから構成されている。

本研究では、この動詞連想概念辞書の拡張と整理を行うとともに、この辞書を構成する連想情報と、コーパス等から得られる共起情報がどのように異なるかを分析する。そして、これらの分析から得られた知見を利用し、日本語学習者が日本語母語話者の連想に沿って語彙を学習できるような支援システムの開発を図る。

3. 研究の方法

(1) 動詞連想概念辞書の拡張・整理

動詞連想概念辞書の規模(2015年4月)は、動詞519語に対して連想語が延べ約220,000語、異なり約45,000語である。これらの動詞は、日本語基礎辞典の見出し語・索引語として用いられている動詞を全て満たしているが、日本語学習者にとっては不足している。そのため、日本語学習用にまとめられた日本語基本動詞用法辞典に掲載されている動詞728語を満たすまで規模を拡張した上で、表記ゆれ修正など辞書データの整理を目指す。

(2) 連想情報と共起情報の比較分析(応用)

辞書の拡張と並行して、動詞連想概念辞書の連想情報と従来利用されている共起情報を換喩解析システムへの応用を通して比較分析を行う。このシステムは、比喩表現の一つである換喩(メトニミー)を検出し、言い換えるシステムであり、「検出」と「解釈」の2つのフェーズで連想情報と共起情報を用いた手法をそれぞれ比較し、分析を行う。

(3) 連想情報と共起情報の比較分析(性質)

辞書の拡張が終わり次第、動詞連想概念辞書の動詞に対する連想語と新聞コーパスの動詞と共起する単語を抽出して比較する。動詞連想概念辞書の連想語は、動詞とその深層格の名称(e.g.,「対象」「手段」など)に対して実験参加者が連想したものである。これらと対応する格助詞(e.g.,「対象」と格助詞「を」,「手段」と格助詞「で」など)と同じ文節で使用される単語の頻度を抽出する。そして、ある動詞に対して、連想しやすい語と共起しやすい語がどのように異なるか示し、両情報の差異性について分析を行う。

(4) 日本語語彙学習システムの開発

(2)と(3)の分析結果を踏まえて、動詞連想概念辞書の連想情報を用いた日本語語彙学習システムの開発を目指す。ただし、このシステムの評価については、日本語学習者に一定期間利用してもらわないとできないため、本研究期間内で実施することは予定していない。

4. 研究成果

(1) 動詞連想概念辞書の拡張・整理

日本語基本動詞用法辞典の動詞728語と動詞連想概念辞書の動詞519語で重複しているものを調べ、不足分の動詞254語について連想実験を実施した。この連想実験は、web上で稼働する実験システムを用いており、実験参加者は画面上に呈示される刺激語と連想の課題(深層格の名称)に対して連想した語(連想語)を入力する。3年間を通して実験参加者は延べ約700人であり、これにより、動詞連想概念辞書を基本動詞773語、連想語が延べ約300,000語、異なり約58,000語の規模まで拡張した。

(2) 連想情報と共起情報の比較分析(応用)

動詞と名詞に関する連想概念辞書と日本語WordNetそして日本語語彙大系を用いて、文章中の名詞と述語動詞における単語間の最大概念間距離を求め、換喩表現を検出した後にその解釈となる単語を連想情報から抽出する手法を提案した。そして、新聞コーパスにおける単語の共起情報に基づいた従来の統計的手法による結果と比較し、連想情報の有用性を示した(表1・表2)。換喩解析における連想情報と共起情報の特長をそれぞれ明らかにした。

表1:「検出」における全体の精度とF値

	全体の精度	検出のF値
従来手法1	0.61	0.54
従来手法2	0.64	0.55
提案手法	0.85	0.81

表2:「解釈」の1位正解率と5位正解率

	1位正解率	5位正解率
従来手法1	0.28	0.44
従来手法2	0.21	0.42
提案手法	0.74	0.86

(3) 連想情報と共起情報の比較分析(性質)

動詞連想概念辞書と新聞コーパスを用いて、動詞における連想しやすい語(名詞)と共起しやすい語(名詞)について比較分析を行い、「動詞-名詞」の関係について両者が異なることを示した。また、連想実験の約3,000人分の生データと5年分の新聞コーパスに対して前処理を行い、word2vecを用いて作成した単語の分散表現を比較した。「動詞-動詞」の類似度関係について分析した結果、連想情報では動詞(図1「食べる」)が表す動作に対して「状況が似た動作」(「飲む」)や「前後の動作」(「残す」、「たいらげる」)、「一部の動作」(「味わう」、「噛む」)を表す動詞の類似度が高かった。一方で、共起情報では「複合動詞」にあたる動詞の類似度が高い関係であることがわかった。

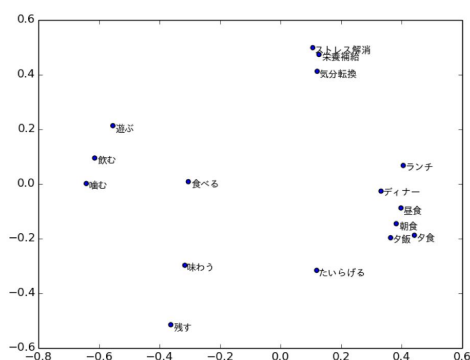


図1: 主成分分析の結果(「食べる」)

(2)とこれらの結果より、本研究で開発を目指す連想に基づいた語彙学習システムが、従来の辞書・コーパスや学習システムとは、特長や性質が異なったものになることが示唆された。

(4) 日本語語彙学習システムの開発

最終年度に、日本語語彙学習システムを試作した。このシステムは、語彙学習において(2)と(3)を含めた連想情報の特長を応用している。

具体的には、動詞連想概念辞書を用いて、穴埋め形式問題を次の手順で自動生成する。まず、動詞連想概念辞書の動詞を述語とする単文を生成する。この単文には、動作主や対象に関して連想しやすい語と格助詞を組み合わせた文節の他に、括弧と格助詞からなる文節が含まれている。また、この単文と一緒に4つの単語が括弧に入る選択肢として呈示される。これらの選択肢も動詞連想概念辞書の情報を用いて自動で生成される。そして、学習者が単語を選択すると正解語のみが画面に残り、どの単語が最も連想しやすい語なのかが分かるようになっている。このような仕組みで、学習者が日本語母語話者の連想に沿って語彙を学習することが期待できる。

今後は、試作した語彙学習システムのUIや携帯端末における使用などについても検討し、完成を目指す。その後、日本語学習者

を対象にしてシステムの評価を行いたいと考えている。

5. 主な発表論文等

(研究代表者、研究分担者及び連携研究者には下線)

[雑誌論文](計1件)

1. 安本匡佑, 寺岡文博, VISTouch: 複数の携帯端末間における動的な立体連携, 査読有, 情報処理学会論文誌, Vol. 59, No. 2, 2018, pp. 690-702

[学会発表](計11件)

1. 石本祐一, 寺岡文博, 榎本美香, 言語情報と韻律情報に基づく自発発話終了位置の統計的予測モデルの構築, 査読無, 日本音響学会 2018 年春季研究発表会, 2018
2. Yuichi Ishimoto, Takehiro Teraoka, and Mika Enomoto, "End-of-Utterance Prediction by Prosodic Features and Phrase-Dependency Structure in Spontaneous Japanese Speech", 査読有, INTERSPEECH 2017, 2017, pp. 1681-1685
3. 石本祐一, 寺岡文博, 榎本美香, 統語情報と韻律情報を用いた発話順からの漸進的発話未予測の検討, 査読無, 日本音響学会 2017 年秋季研究発表会, 2017
4. 寺岡文博, 石崎俊, 動詞における連想と共起の差異性, 査読無, 情報処理学会第79回全国大会, 2017
5. 石本祐一, 寺岡文博, 榎本美香, 韻律情報と文節係り受け構造を用いた発話未予測モデルの構築, 査読無, 日本音響学会 2017 年春季研究発表会, 2017
6. Yuichi Ishimoto, Takehiro Teraoka, and Mika Enomoto, "A study on prediction of end-of-utterance by prosodic features and phrase-dependency structure in spontaneous speech", 査読有, The 5th Joint Meeting of the Acoustical Society of America and the Acoustical Society of Japan, 2016, p. 3393
7. 石本祐一, 寺岡文博, 榎本美香, 韻律情報と文節係り受け構造を用いた発話未予測モデルの検討, 査読無, 日本音響学会 2016 年秋季研究発表会, 2016
8. Masasuke Yasumoto and Takehiro Teraoka, "Game Applications of Dynamic and Spatial Connections between Multiple Mobile Devices", 査読無, Replaying Japan 2016, 2016, pp. 51-52
9. Takehiro Teraoka, "Metonymy Analysis Using Associative Relations between Words", 査読有, The 10th International Conference on Language Resources and Evaluation (LREC), 2016, pp. 4614-4620
10. 寺岡文博, 石崎俊, 単語間連想関係を考慮した換喩表現の解析, 査読無, 情報処理学会第78回全国大会, 2016

11. Masasuke Yasumoto and Takehiro Teraoka,
“ VISTouch ” ,査読有,ACM SIGGRAPH 2015,
2015

6 . 研究組織

(1)研究代表者

寺岡 文博 (TERAOKA, Takehiro)
東京工科大学・メディア学部・助教
研究者番号 : 30617329

(4)研究協力者

石崎 俊 (ISHIZAKI, Shun)
慶應義塾大学・名誉教授
研究者番号 : 00245614

寺岡 杏子 (TERAOKA, Kyoko)
研究者番号 : 20713241