

博士論文 平成25年度 (2013)

動詞連想概念辞書の構築とその応用研究

—連想に基づいた意味理解システム—

慶應義塾大学大学院政策・メディア研究科

寺岡 丈博



## 要旨

# 動詞連想概念辞書の構築とその応用研究 —連想に基づいた意味理解システム—

寺岡 丈博

近年、自然言語処理研究では、コンピュータの言語理解機能を人間に近付けるために様々な言語資源が、人間が持つ言語の背景知識を体系化したものとして構築・応用されてきた。しかし、意味理解・文脈理解に関する精度は十分に向上していないため、このような知識の体系化に対しては従来と異なったアプローチが必要であると考えられる。

そこで本研究は、動詞に関する連想実験のデータから動詞連想概念辞書を構築し、意味理解システムに応用してコンピュータの言語理解機能を向上させることを目的とする。この連想実験では、基本動詞が刺激語として深層格の名称とともに呈示されるため、実験参加者は連想語としてこれらの深層格情報を答える形式となっている。そして、連想実験から得られたデータを用いて、刺激語に対する連想語の連想し易さを単語間距離として定量化し、連想距離と定めた。これらの刺激語、連想語、そして連想距離を用いて動詞連想概念辞書を構築し、知識の体系化を図った。

構築した動詞連想概念辞書を評価するため、まず始めに、既存辞書と共通するデータを比較分析し、連想情報と共起情報の性質の差異を示した。そして、「省略語の推定」と「換喩表現の検出」に関する意味理解システムへ応用し、従来手法を参考にしたベースラインを用いて評価実験を行った。その結果、いずれの応用システムも、従来手法より統計的有意差をもって高い精度を示すことができた。

以上から、動詞連想概念辞書は、共起情報とは異なる連想情報から成り、意味理解において有用であることが、本研究の結論として言える。

キーワード: 連想概念辞書, 連想実験, 動詞, 省略語, 換喩



## Abstract

# Construction of an Associative Concept Dictionary for Verbs and its Applications

—Text Understanding Systems with Associative Information—

Takehiro Teraoka

Recently, on natural language processing research, various language resources have been constructed and used as something systemized with human linguistic background knowledge to improve computer's ability for understanding meanings. However, the accuracy of understanding contextual meanings has not been adequately improved yet; computers need such knowledge which is systemized using different approaches than before.

The purpose of this study is to construct an associative concept dictionary for Japanese verbs from association experiments on verbs, and apply it to text understanding systems to improve the computer ability. In the experiments, basic verbs with deep cases are presented as stimulus words, so experiment participants should answer information based on the deep cases as associated words. In this study, the degree of association of associated words from stimulus ones was quantified by calculating distances between these words with the experiment data, and it was defined as an associated distance. By using these stimulus words, associated ones, and their associated distances, an associative concept dictionary for verbs (hereinafter referred as to Verb-ACD) was constructed to systemize such knowledge.

To evaluate the Verb-ACD, it was compared with existing dictionaries and indicated that association information and co-occurrence information differed in property. It was also applied to text understanding systems of estimating ellipsis words and detecting metonymic expressions, and they were evaluated with baselines based on the conventional methods. Then, the experimental results showed that both the application systems had significantly better accuracy than those of the baselines.

From the results, it is concluded that the Verb-ACD consists of associative information which differs from co-occurrence information and is effective in text understanding.

Keywords: Associative Concept Dictionary, Association Experiment, Verb, Ellipsis, Metonymy

# 目次

<b>第1章</b>	<b>序論</b>	<b>1</b>
1.1	研究の背景	3
1.2	研究の目的と概要	5
1.2.1	動詞連想概念辞書の構築	5
1.2.2	動詞連想概念辞書の応用	6
1.3	研究の特色	6
1.4	本論文の構成	7
<b>第2章</b>	<b>関連研究</b>	<b>9</b>
2.1	言語資源の構築	11
2.1.1	FrameNet	12
2.1.2	日本語フレームネット	13
2.1.3	京都大学格フレーム辞書	14
2.1.4	WordNet	15
2.1.5	名詞連想概念辞書	17
2.2	名詞の連想情報を用いた自然言語処理	18
2.2.1	重要文の抽出	18
2.2.2	多義性の解消	19
2.2.3	比喩理解システム	20
<b>第3章</b>	<b>動詞連想概念辞書の構築</b>	<b>23</b>
3.1	連想実験	25
3.1.1	刺激語の動詞	25
3.1.2	連想課題	27

3.1.3	実験参加者	28
3.2	連想距離の定量化	28
3.3	連想語と連想距離の関係	30
3.4	既存辞書との比較による動詞連想概念辞書の特徴	34
3.5	動詞の汎用性に関する考察	36
3.6	動詞連想概念辞書の拡張	38
<b>第4章</b>	<b>連想情報を用いた省略語の推定</b>	<b>41</b>
4.1	はじめに	43
4.2	提案手法	43
4.2.1	システムの概要	43
4.2.2	名詞連想概念辞書	44
4.2.3	省略語候補の抽出	45
4.2.4	確信度の設定	46
4.3	評価実験	47
4.3.1	ベースラインの作成	47
4.3.2	課題文と正解の設定	49
4.3.3	評価尺度	50
4.3.4	結果と考察	51
4.4	まとめ	54
<b>第5章</b>	<b>単語間連想情報に基づいた換喩表現の検出</b>	<b>57</b>
5.1	はじめに	59
5.2	提案手法	61
5.2.1	方針	61
5.2.2	手法の詳細	62
5.3	評価実験	69
5.3.1	換喩表現	69
5.3.2	学習データ	70



5.3.3	先行研究及び共起情報に基づいたベースライン . . . . .	72
5.3.4	結果 . . . . .	76
5.4	考察 . . . . .	80
5.5	まとめ . . . . .	86
<b>第6章</b>	<b>結論</b>	<b>89</b>
6.1	本研究の成果 . . . . .	91
6.1.1	動詞連想概念辞書の構築 . . . . .	91
6.1.2	動詞連想概念辞書の応用 . . . . .	91
6.2	本研究の課題と展望 . . . . .	92
6.3	おわりに . . . . .	94
	<b>謝辞</b>	<b>97</b>
	<b>参考文献</b>	<b>101</b>
	<b>本研究に関する発表一覧</b>	<b>113</b>
	<b>付録</b>	<b>117</b>



# 目 次

3.1	刺激語「借りる」と「貸す」に対する連想語の関係 . . . . .	32
3.2	刺激語「歩く」から連想距離が比較的近い連想語 . . . . .	33
4.1	省略語推定システムの概要 . . . . .	44
4.2	VNACD と VACD の正解語の最高順位のヒストグラム . . . . .	53
5.1	本手法 (ACD-JWN) の概要 . . . . .	63
5.2	学習データ 1 の全入力文に対する ACD-JWN の決定木学習の結果 . .	81
5.3	学習データ 2 の全入力文に対する ACD-JWN の決定木学習の結果 . .	82
5.4	学習データ 1 に対する提案手法 ACD-JWN とベースライン GT-JWN, CF-JWN における学習曲線 . . . . .	84
5.5	学習データ 2 に対する提案手法 ACD-JWN とベースライン GT-JWN, CF-JWN における学習曲線 . . . . .	85



# 表 目 次

2.1	WordNet における名詞の意味的關係	16
2.2	WordNet における動詞の意味的關係	16
2.3	名詞を刺激語とした連想実験における連想課題	18
2.4	刺激語「辞書」の時の連想語と連想距離の例	18
2.5	多義性解消における評価実験の結果	20
2.6	比喩理解システムの出力例	21
3.1	分類した基礎語の動詞	26
3.2	深層格を基にした連想課題と内容	27
3.3	刺激語「運ぶ」に対する連想語のデータ例	31
3.4	動詞連想概念辞書と格フレーム辞書における共通語の順位相関係数の 平均	35
3.5	動詞連想概念辞書と格フレーム辞書における共通語の順位相関係数の 平均	35
3.6	動詞連想概念辞書の連想情報と格フレーム辞書の共起情報の比較	37
3.7	刺激語「取る」に対する連想語のデータ例	38
3.8	動詞連想概念辞書の規模	39
4.1	名詞連想概念辞書の連想課題と連想語の例	45
4.2	動詞連想概念辞書の連想課題と対応させた格	48
4.3	提案手法と各ベースラインの平均逆順位 (MRR)	52
4.4	提案手法と各ベースラインの N 位正解率	52
4.5	提案手法と各ベースラインの平均精度の平均 (MAP)	54

5.1	空間的隣接性に基づいた換喩表現の関係性 . . . . .	60
5.2	時間的隣接性に基づいた換喩表現の関係性 . . . . .	60
5.3	ACD-JWN で決定木学習に用いた特徴量の内容 . . . . .	67
5.4	ACD-JWN で決定木学習に用いた特徴量の値 . . . . .	68
5.5	学習データ 1 のサンプル文の一部 . . . . .	70
5.6	学習データ 2 のサンプル文の一部 . . . . .	71
5.7	ベースライン GT-GT における決定木学習の特徴量 . . . . .	73
5.8	ベースライン GT-JWN における決定木学習の特徴量 . . . . .	74
5.9	ベースライン CF-GT における決定木学習の特徴量 . . . . .	75
5.10	ベースライン CF-JWN における決定木学習の特徴量 . . . . .	76
5.11	学習データ 1 と学習データ 2 に対するシステム全体の精度 . . . . .	77
5.12	学習データ 1 に対する換喩表現の検出精度 . . . . .	78
5.13	学習データ 2 に対する換喩表現の検出精度 . . . . .	78
5.14	学習データ 1 に対するリテラルな表現の検出精度 . . . . .	79
5.15	学習データ 2 に対するリテラルな表現の検出精度 . . . . .	79
5.16	連想実験の参加者数と ACD-JWN の精度 . . . . .	86

# 第1章

## 序論

現在の自然言語処理研究では、有効な手法として統計的アプローチが主流となっており、基礎的技術から応用的技術まで広範囲に渡って有用な成果が報告されている。しかし、現状ではコンピュータの意味理解や文脈理解の精度向上に限界があると考えられており、コンピュータが人間の言語能力に迫るためには、新しい技術的な展開が期待されている。第1章では、このような背景及び本研究の目的と概要を述べ、本研究の意義を明確にする。





## 1.1 研究の背景

近年、自然言語処理の研究が著しい発展を遂げてきたが、それでもコンピュータが言語の意味理解を人間と同等なレベルで行うまでには至っていない。現在、基礎的な技術である形態素解析や構文解析をはじめ、意味解析の一種である格解析などは、統計的手法の導入により、高い精度を実現している。しかし、意味解析の一種であっても、直喩や隠喩、換喩などの修辞表現を扱う比喩解析や、複数の文の処理すなわち文脈処理を必要とする談話構造解析、照応・省略解析などは、先に述べた形態素解析や構文解析に比べると、精度が高いとは言えない。なぜならば、人間が日常で用いている言語表現の多様性が高いことから、形態素解析で得られる形態素や語順をはじめ、構文解析で得られる係り受け関係などの表層的な情報を高い精度で処理するだけでは不十分だからである。コンピュータが人間と同様に言語の意味を理解するには、単語の多義性や慣用的な表現、修辞的な表現などを解析するために必要な意味情報、すなわち深層的な情報を正確に処理する必要があるが、高精度の処理が可能になっていないのが現状である。そのため、これらの基礎的技術を統合した機械翻訳や対話システムなどの応用システムに対して、人間に匹敵するまでの精度向上を求めることは未だに困難な面がある。

ここで、人間の言語処理について述べると、我々人間は日常生活の中で言葉を話す、或いは文章を書く際に、品詞や文法などの言語学的な情報だけでなく、単語やフレーズなどの意味も含めた言葉の背景にある膨大な情報を一般的な知識として利用している。すなわち、現在の課題となっているコンピュータの意味理解や文脈理解に関する言語理解機能を向上させ、人間の言語能力に近づけるためには、人間が持つ複雑で膨大な言語の背景知識を精緻に体系化し、コンピュータが知識として使用できるようにする必要がある。

これまでに、コーパスやシソーラスなどの言語資源が構築され、形態素解析や構文解析などの基礎的技術から、機械翻訳や情報抽出などの応用的技術まで広範囲に渡って応用されてきた。これらの言語資源は、人手によって作成された電子化辞書 [27, 61] の他に、十数年分の新聞記事や Web 上のテキストデータから自動構築された辞書などもある [36, 44]。特に、Web 上のテキストデータから統計的処理によって

得られる知識は、従来の言語資源よりも大規模であるため、有用な言語資源とされている。しかし、統計的手法によって得られる知識は、言語の表層的な情報が中心になってしまうため、人間が持つ言語の背景知識を十分に表していない。

一方で、従来とは全く異なったアプローチで構築された言語資源として、連想概念辞書 [64] が挙げられる。この連想概念辞書は、名詞を中心に構築されており、WordNet[15] や NTT 日本語語彙大系 [27] などと同様に上位語・下位語・類義語などが体系的にまとめられたシソーラスの一種であるが、他のシソーラスと大きく異なる点は、人間の直感に基づいた大規模な連想実験のデータから構築されていることである。さらに、刺激語<sup>1</sup>に対する連想語<sup>2</sup>の「連想し易さ」を連想距離として定義し、これを定量化することによって単語間距離を計算している。また、連想実験の刺激語は基本的な名詞であるため、日常的な連想語を収集している可能性が高いことから、常識とも言うべき知識を体系化することができていると考えられる。実際に、重要文の抽出 [65] や多義性解消 [66, 67] に応用されており、既存の手法よりも高い精度を実現している。また、自然言語処理研究のみならず、認知科学分野の創造性プロセス分析の研究 [55] にも応用されるなど、連想概念辞書は人間の心理データを表す言語資源として、従来とは一線を画している。このことから、連想概念辞書は従来の言語資源と比べると全体の規模は小さいが、言語の深層的な情報を含んでおり、人間が持つ言語の背景知識に近づいたものとして捉えることができる。しかし、人間の日常の文脈において動作や状態変化を表す動詞が、意味理解や文脈理解で中心的な役割を果たしていることを踏まえると、名詞の連想情報を中心として構築された現在の連想概念辞書だけでは、コンピュータが利用できる知識の質として十分ではないことがいえる。

ゆえに、動詞に関しても連想情報を収集し、体系化したものが必要であると考えられる。そして、動詞の連想情報を用いて体系化された辞書と、名詞を中心とした連想概念辞書を組み合わせて用いることで、コンピュータの言語意味理解能が人間により近づくことが期待できる。

---

<sup>1</sup>連想実験にて実験参加者に呈示される語。

<sup>2</sup>刺激語に対して実験参加者が連想して答えた語。

## 1.2 研究の目的と概要

本研究は、動詞に関する連想実験のデータを用いて動詞連想概念辞書を構築し、それを応用することでコンピュータの意味理解機能を向上させることを目的とする。また、大量のテキストデータから統計的処理を用いて構築された既存の辞書とデータを比較し、動詞連想概念辞書の性質の分析を行う。さらに、動詞連想概念辞書を意味理解システムに応用し、従来手法を用いたシステムよりも高精度になることを確認することで、動詞連想概念辞書の有用性を示す。

以下で、本研究の内容を「動詞連想概念辞書の構築」と「動詞連想概念辞書の応用」に大別し、それぞれ概要を述べる。

### 1.2.1 動詞連想概念辞書の構築

小学校の国語の教科書で用いられている動詞を基本動詞と定め、基本動詞を刺激語として連想実験を行う。この連想実験では、実験参加者が、呈示される刺激語と連想の課題の組み合わせに対して、思いついた語を連想語として答えることになる。さらに、刺激語と連想語の単語間距離を連想距離と定め、実験データと線形計画法を用いて定量化し、刺激語と連想語、そして連想距離を用いて動詞連想概念辞書を構築する。また、自然言語処理研究では、新聞コーパスや Web 上のテキストデータから統計的処理によって抽出された単語の共起情報が用いられることが多いため、このような共起情報と、動詞連想概念辞書に含まれている連想情報を比較分析することで、性質の評価を行う。以下に、評価の方針を記す。

**性質の評価：連想情報と共起情報の比較分析** 京都大学格フレーム辞書は、Web 文書内の述語とその格情報を統計的処理によってまとめたものである [44]。格要素とその頻度が分かるため、Web 文書中の述語と格に入る単語の共起情報から成り立っているといえる。この格フレーム辞書と動詞連想概念辞書はデータ構造が似ている点があるため、ある動詞に対して共起し易い単語と連想し易い単語を比較することで、共起と連想の特徴の差異を分析する。

## 1.2.2 動詞連想概念辞書の応用

動詞連想概念辞書と既存の名詞を中心とした連想概念辞書（以下、名詞連想概念辞書）を意味理解システムに応用することで、コンピュータが人間の連想情報に基づいた意味理解を行えるようになり、性能の向上が期待できる。そのため、従来の手法を用いたベースラインを設定し、精度を比較することで、動詞連想概念辞書の有用性の評価を行う。以下に、評価の方針を記す。

**有用性の評価：意味理解システムへの応用** 動詞連想概念辞書と名詞連想概念辞書を併用することで、コンピュータは人間が動詞と名詞に対して実際に連想した情報を利用できるため、日常で人間が行っている連想に基づいた意味理解に近づけることができる可能性がある。本研究では、「省略語の推定」と「換喩表現の検出」のシステムに応用する。その後、既存の手法を用いてベースラインを作成し、それぞれのシステムにおいて本手法とベースラインの性能を比較し、精度が向上していることを確認することで有効性を示す。

## 1.3 研究の特色

本研究は、1.2節で説明した「動詞連想概念辞書の構築」と「動詞連想概念辞書の応用」のそれぞれについて特色がある。まず、前者について述べると、自然言語処理研究では動詞を中心とした辞書はこれまでに作成されているが、人間を実験参加者とした連想実験のデータを利用して、動詞について構築したものは他に類が無い。また、連想実験の刺激語である基本動詞は、名詞と比べると数が限られているため、既存の名詞連想概念辞書と比べて、質の高い背景知識の体系化が期待できる。さらに、単語の共起情報と連想情報の特徴について十分に分析されてきていないため、本研究を通して共起と連想の差異を示すことができれば、自然言語処理システムのタスクに応じて連想と共起の情報を効果的に使い分けられることが期待できる。

後者の「動詞連想概念辞書の応用」については、自然言語処理研究で精度向上が課題とされてきた意味理解と文脈理解において、連想情報に基づいた新しい手法を提案することで、精度向上の可能性が考えられる。特に省略解析や比喩理解、多義

性解消，文書要約など，表層的な情報だけでなく深層的な情報を適切に抽出する必要がある処理に対して，従来よりも高い精度が期待できる。

最後に，自然言語処理を含む工学の分野以外でも，認知言語学や言語心理学の研究においては，実験材料や言語教育支援ツールの作成などに，本研究の動詞連想概念辞書を利用できる可能性がある。その他にも，言語を扱う研究分野は多岐に渡ることから，研究・教育利用の目的で動詞連想概念辞書のデータを一般公開することで，他分野の研究発展に貢献することが期待できる。

## 1.4 本論文の構成

本論文の構成を以下に記す。まず，第2章では，本研究の関連研究について述べる。ここでは，動詞連想概念辞書の構築と応用の関連研究として，「言語資源の構築」と「名詞の連想情報を用いた自然言語処理」についてそれぞれ述べる。次に，第3章では，動詞連想概念辞書の構築について述べる。はじめに動詞に関する連想実験の内容と連想距離の定量化について詳細を説明した後に，動詞連想概念辞書の連想情報と京都大学格フレーム辞書の共起情報との比較分析について述べる。そして，第4章と第5章では，動詞連想概念辞書の応用研究について述べる。これらの章では，「省略語の推定」と「換喩表現の検出」において，動詞連想概念辞書と名詞連想概念辞書を組み合わせた手法をそれぞれ提案し，既存の手法を用いたベースラインと性能を比較する評価実験の詳細を述べる。最後に，第6章では，本研究の研究成果をまとめ，今後の課題と研究の発展性について述べる。



## 第2章

# 関連研究

第2章では，本研究の関連研究として「言語資源の構築」と「名詞の連想情報を用いた自然言語処理」について述べる．前者については，主に，動詞に関係したコーパスやシソーラスの他に，本研究と同じアプローチの名詞連想概念辞書について触れる．また，後者については，名詞連想概念辞書の応用例として，重要文の抽出やニューラルネットワークと組み合わせた比喩理解システムについて説明する．





## 2.1 言語資源の構築

今日までに、様々なコーパス (corpus) やシソーラス (thesaurus) などの言語資源が開発・構築され、自然言語処理研究に応用されている。ここでは、コーパスとシソーラスについて概説した後に、本研究と関連のある主な言語資源について説明する。

コーパスとは、新聞や Web 上の文書など、実際に書かれた記事や文章を電子テキストとして大量に収集したものを一般的に指す。その他にも、日本語話し言葉コーパス [42] のようなインタビューや講演などで実際に話された内容を電子的にまとめたものも含んでいる。代表的なものとしては、電子的なコーパスの先駆けとして編纂された Brown Corpus (Brown University Standard Corpus of Present-Day American English) [20] やイギリス英語の話し言葉と書き言葉から成る大規模な BNC (British National Corpus) [12]、そして注釈付きコーパスの Penn Treebank [49] などが挙げられる。日本語に関しては、EDR 電子化辞書の中に含まれている EDR 日本語コーパス [61] や、京都大学テキストコーパス [43]、現代日本語書き言葉均衡コーパス [3, 53] などが挙げられる。

一方、シソーラスとは、単語ごとに上位語や下位語、同義語、類義語などをまとめることで体系化した辞書を指す。代表的なものとしては、心理学の知見に基づいて構築された WordNet [15] の他に、日本語では EDR 電子化辞書の中に含まれている EDR 概念体系辞書 [61] や日本語語彙大系 [27]、分類語彙表 [41] などが挙げられる。

上記で触れたコーパスやシソーラスの他に、動詞を中心とした言語資源を挙げると、FrameNet [5] や VerbNet [38]、動詞項構造シソーラス [81] など、英語や日本語で様々構築されている。まず、FrameNet では、フレーム意味論 [19] に基づいて 10,000 語以上に関する語彙意味分析の結果が約 170,000 文のコーパスデータに付与されて蓄積されている。VerbNet は、FrameNet ほどの規模や細かな粒度ではなく、Levin の動詞分類 [48] を基に構築されており、約 3700 語の動詞に対して項構造や意味役割などが記述されている [39]。また、日本語の動詞に関しては、動詞約 4400 語について述語項構造を分類している動詞項構造シソーラスや、FrameNet と連携して構築されている日本語フレームネット [62] などがある。これらは、概して粒度も細かい上に規模も大きく、また付与事例も多く、動詞に関する詳細な情報源となっている。

しかし、人間が動詞に対して連想した情報を中心として成り立ってはいない。そのため、第 1 章で述べたように、言葉の背景知識を体系化したものを必要とする意味理解や文脈理解に関する処理に対しては、上記の言語資源だけでは不十分な可能性がある。

下記の項では、本研究の動詞連想概念辞書の主な関連研究として、動詞を中心に構築されている、または、見出し語に動詞を含んでいるコーパスやシソーラスについて概要を説明する。そして、最後の項で、本研究と同様のアプローチで構築されている名詞連想概念辞書について概要を説明する。

### 2.1.1 FrameNet

FrameNet(FN) は、カリフォルニア州バークレーの International Computer Science Institute(ICSI) において Charles J. Fillmore を中心に 1990 年代後半から構築され始め、2000 年からは Collin Baker がプロジェクト責任者となり、現在まで拡張・管理されてきている。この FrameNet は、話し言葉と書き言葉の両方を含む 1 億語以上の現代イギリス英語コーパス (British National Corpus) とアメリカ英語コーパス (North American Newswire Corpus) からフレーム意味論 (frame semantics)[19] の枠組みで抽出した用例を基に、英語語彙の語義分析を行い、電子的な語彙体系として構築・資源化している [5]。そもそもフレーム意味論とは、人間の言語と経験との関連を重視する経験的意味論であり、そのための分析手法と意味記述の枠組みを示して、その分析結果を表現する記述的枠組みを提示する理論のことである。言語使用における意味は、その概念化において、それが表現主体によってどんな視点で捉えられているか、どんな文化的価値体系を背景にどのようなコンテキストで捉えられているか、などの認知プロセスに支えられており、この認知プロセスの前提になる背景知識が、フレーム意味論におけるフレームにあたる。そして、Fillmore が定義するフレームとは、その中のどれか一部を理解するためには、全体の構造の理解が必要であり、有機的に繋がっている体系的知識構造を指している。

FrameNet では、はじめに対象とする意味フレームとそれを構成するフレーム要素の定義を行い、その意味フレームが関係する語彙項目を洗い出す。次にコーパス

から語彙項目の用例を収集して、オンラインアノテーション支援ツールによって用例文中の語句へタグ付けを行う。このようにフレーム意味論に基づいた分析結果を語彙の情報資源として蓄積する。これまでに定義された意味フレームは1,000以上あり、10,000以上の語彙意味分析の結果が、170,000文以上のコーパスデータに付与されて蓄積されている [28]。また、FrameNetの大きな特徴として、意味タグとしてコーパスに付与されている情報は Agent や Instrument 或いは Object などの古典的な格フレームに基づいた抽象的な役割関係ではなく、意味フレームに即して定義される精度の高いフレーム要素であることが挙げられる。さらに、フレーム間の関係を参照することで、語彙項目同士の関連付けが行われている点も大きな特徴であり、Web上でフレーム間の関係を表示させることが可能である [21]。

なお、ここまでの規模に至る途中の段階において、FrameNetのデータとソフトウェアを自然言語処理の研究者がどのように使用することができるかについての議論 [7] や、VerbNet や WordNet のような他の言語資源との比較分析などが積極的に行われている [4, 6, 13, 17]。また、2000年代から、スペイン語 [78]、ドイツ語 [9]、中国語 [97]、スウェーデン語 [31]、そして日本語などの他言語に関する FrameNet も構築・整備されてきており、これらの規模の拡張も現在進行中である [28, 62]。

### 2.1.2 日本語フレームネット

日本語フレームネット (Japanese FrameNet, JFN) は、2002年に構築され始めた日本語版の FrameNet である。2.1.1 項の FrameNet と同様、コーパスのデータをフレーム意味論の枠組みに基づいて日本語の語彙意味分析を行い、その結果を電子的な語彙体系として構築している。また、英語語彙分析のために FrameNet で定義された意味フレームやフレーム要素が、日本語語彙記述にどこまで適しているかも同時に検討している。他言語資源とのリンクが可能か等の実験を兼ね、質の高いアノテーションを目指すとともに、将来的には理論言語学だけでなく言語教育に貢献できる資源とすることも視野に入れて規模の拡張・整備が行われている [21, 62]。また、自然言語処理研究の分野では、述語項構造における項の意味役割に関する推定 [26, 80] に応用されている他に、多義性分析による曖昧性解消や言い換えについて議

論などがされている [63].

### 2.1.3 京都大学格フレーム辞書

京都大学格フレーム辞書 [44] は、京都大学の黒橋研究室が Web から大規模格フレームを自動構築したコーパスである。格フレームとは、用言とそれに関する名詞を用言の用法ごとに整理したものであり、この大規模格フレームは Web 上の約 16 億文の日本語テキストから自動的に構築され、約 4 万用言から成っている高次格フレーム辞書である<sup>1</sup>。なお、オンライン上で任意の格フレームについて検索することが、可能である。大規模なコーパスから格フレーム辞書を構築するには用言の多義性に関して大きな問題が伴うのだが、これにも対処して構築されている。以下では多義性の問題に触れつつ、その対処法と高次格フレーム辞書に至るまでの過程を説明する。

大規模コーパスから格フレーム辞書を自動構築する際の最大の問題は、用言の用法の曖昧性、つまり同じ表記の用言でも複数の意味や用法を持つてしまうために取り得る格や用例が異なるということである。例えば、「積む」という用言に対して「トラックに荷物を積む」と「経験を積む」では用法が異なっている。このような用法が異なる格フレームを別々につくるために格フレーム収集の単位を用言とその直前の格要素の組とし、「荷物を積む」と「経験を積む」を単位として格フレームを収集することで問題性を解消している。また格の一致度と用例群間の類似度の積を格フレームの類似度と定義して計算方法を設定し、この類似度に基づいて格フレームをマージするためにクラスタリングを行う。このクラスタリングは、コーパスを構文解析して確信度の高い述語項構造のみを抽出した後に行われ、この過程を経て 1 次格フレーム辞書を得る。そして次に 1 次格フレーム辞書を用いてコーパスを格解析し、新たに分かる確実な情報を抽出し、格フレーム辞書を高度化して 2 次格フレーム辞書を構築する。このように徐々に新たな情報を加えていくことによって高次格フレーム辞書を構築している [34, 36].

格フレームは用言の用法ごとに整理したものであるため、動詞を中心にしたデー

---

<sup>1</sup>京都大学格フレームとも呼ぶ (<http://nlp.ist.i.kyoto-u.ac.jp/>).

タが大部分を占める。動詞連想概念辞書の構築を目的とする本研究とアプローチは異なっているものの、動詞を中心としたデータ構造を確認できる点で共通している。

#### 2.1.4 WordNet

WordNet は、アメリカのプリンストン大学の Cognitive Science Laboratory にて、心理学的な知見に基づいて Gerge A. Miller が中心となって開発したシソーラスである [52]。1985 年から開発され始め、アップデートとリリースを繰り返してきた。昨今では、2005 年に Windows OS 向けの WordNet2.1 が、翌 2006 年に Unix 系 OS 向けの WordNet3.0 がリリースされている。なお、2013 年の現在は WordNet 3.1<sup>2</sup> のバージョンが最新であり、オンラインでのみ利用することが可能となっている [69]。

WordNet の規模については、現在リリースされているバージョンで最も新しい 3.0 では、英単語約 155,000 語が約 118,000 の synset に分類されており、全体で約 207,000 の英単語と synset の組み合わせから成り立っている。また、WordNet は、名詞、動詞、形容詞、副詞を含んでおり、各々の synset の数については、名詞が約 82,000、動詞が約 14,000、形容詞が約 18,000、副詞が約 3,600 となっている [16, 70]。

synset の大部分は、他の synset と意味的關係を持ち、番号によって示される。なお、この意味的關係の種類は、品詞によって異なる。表 2.1 のように、hypernym (上位語)、hyponym (下位語)、coordinate term (同族語)、holonym、meronym が名詞の意味的關係の種類であり、動詞の意味的關係については表 2.2 のように hypernym、troponym、entailment、coordinate term となっている。また、形容詞については、related noun (関係のある名詞) と similar to (似ているもの)、そして participle of verb (動詞の分詞) の 3 種類の意味的關係であり、副詞の意味的關係は root adjectives (原形となる形容詞) の 1 つのみである。synset に含まれる語句は同様な意味を持つ同義語であるために、基本的にこれらの意味的關係は synset 内に適用される。また、同一の synset 内においても単独の語句が、他の語句と antonym (反意語) などの關係を持つこともある。

<sup>2</sup>WordNet Search - 3.1(<http://wordnetweb.princeton.edu/perl/webwn>).

表 2.1: WordNet における名詞の意味的關係

種類	意味
hypernym	X が Y の種類の 1 つならば, Y は X の hypernym である.
hyponym	Y が X の種類の 1 つならば, X は Y の hyponym である.
coordinate term	X と Y が同じ hypernym ならば, Y は X の coordinate term である.
holonym	X が Y の一部であるならば, Y は X の holonym である.
meronym	Y が X の一部であるならば, Y は X を meronym である.

表 2.2: WordNet における動詞の意味的關係

種類	意味
hypernym	X という行動が Y の種類の 1 つならば, 動詞 Y は動詞 X の hypernym である. (例: movement は travel の hypernym)
troponym	Y という行動が X を行う際の様態であるならば, 動詞 Y は動詞 X の troponym である. (例: lisp は talk の troponym)
entailment	X している場合, 必然的に Y しているならば, 動詞 Y は動詞 X に entail されている. (例: snoring と sleeping の関係)
coordinate term	X と Y が同じ hypernym ならば, Y は X の coordinate term である.

さらに, WordNet では, polysemy count という語句の多義性における情報を知ることができる. polysemy count は, 語句が属する synset の数を表しており, ある単語が幾つかの synset に属している場合, その単語はその数だけ意味を持っていると見なすことができる. そして, これらの意味で, ある意味は他の意味と比べて一般的に多く用いられているという関係を知ることができ, このような関係は frequency score で表されている [15].

現在, ドイツ語に関して構築された GermaNet[24, 25] の他に, オランダ語, イタリア語, スペイン語, ドイツ語, フランス語, チェコ語, エストニア語のヨーロッ

パ言語に関する多言語情報から構築された EuroWordNet[1, 71]をはじめ、英語版の WordNet をもとにして構築された日本語 WordNet[10, 29] など、他の言語に関する WordNet が開発・構築されてきている [23].

### 2.1.5 名詞連想概念辞書

名詞連想概念辞書は、慶應義塾大学の石崎研究室が、名詞を中心とした大規模な連想実験の情報から構築したシソーラスである [37, 64]. この連想実験の実験参加者は、刺激語（名詞）について表 2.3 の「上位概念」、「下位概念」、「部分・材料概念」、「属性概念」、「類義概念」、「動作概念」、そして「環境概念」の 7 つの連想課題を対象とし、それぞれあてはまる語を連想する。そして、このように連想された語を連想語としている。なお、この連想実験は、同大学湘南藤沢キャンパス (SFC) の学内ネットワークを用いて行われており、実験参加者は学部生と大学院生が対象となっている。

また、刺激語に対する連想語の連想し易さを連想距離と定義し、実験で得られた連想語のデータを用いて連想距離を定量化している。連想距離は、連想語の頻度と順位の間数として表され、表 2.4 に例を挙げて示すように、連想距離が短い連想語ほど刺激語に対して連想され易い関係を示している。また、連想実験の刺激語の名詞は小学校の国語の教科書で用いられる基本語のため、日常的な連想語を収集しており、刺激語と連想語間を表す連想距離によって、常識とも言うべき言語の背景知識を体系化できていると考えられる。

現在の名詞連想概念辞書は、1 刺激語名詞に対して実験参加者は 50 人で、刺激語約 1,100 語、連想語が延べ約 280,000 語、異なり約 64,000 語の規模となっている。また、2013 年現在、無料で一般公開されている名詞連想概念辞書は、1 刺激語につき試験者が 10 人のデータによる刺激語 1,650 語、連想語が延べ約 130,000 語、異なり約 33,000 語の規模 [37] となっており、重要文の抽出や語の多義性解消モデルなどの自然言語処理研究 [65] の他に、認知科学研究における実験素材 [55] としても用いられている。

表 2.3: 名詞を刺激語とした連想実験における連想課題

連想課題	説明
上位概念	刺激語の特徴を含む広範な、或いは抽象的な概念を表す（上位語）
下位概念	刺激語の特徴を持つ、より細かい概念を表す（下位語）
部分・材料概念	刺激語の部分、或いは材料を表す
属性概念	刺激語が備えている特徴を表す
類義概念	刺激語と意味が同じ、或いは似ている概念を表す
動作概念	刺激語が行う、或いは行われる動作を表す
環境概念	刺激語が動作をする、或いは存在する場所や環境を表す

表 2.4: 刺激語「辞書」の時の連想語と連想距離の例

連想課題	「辞書」の連想語（連想距離）
上位概念	本 (1.17), 書物 (3.33), 書籍 (4.88)
下位概念	英和辞書 (2.31), 国語辞書 (3.57)
部分・材料概念	紙 (1.23), ページ (3.31)
属性概念	重い (2.00), 難しい (5.54)
類義概念	辞典 (1.20), 事典 (5.60)
動作概念	引く (1.63), 調べる (1.86)
環境概念	図書館 (1.66), 本棚 (2.22)

## 2.2 名詞の連想情報を用いた自然言語処理

### 2.2.1 重要文の抽出

我々人間が文書を要約する際、まず始めに文書中から重要な箇所を抜き出し、それらの内容をまとめた上で簡潔に言い換える作業が言語の連想を通して脳内で行われている。そのため、文書中から重要文を高い精度で抽出することが要約の根幹を成しているといえる。岡本らは、名詞連想概念辞書を用いて文書中にある語の重要度を計算して、重要文を抽出する手法を2つ提案している。



まず、1つ目の手法は、従来手法で用いられている語の出現頻度だけでなく、名詞連想概念辞書を用いて連想関係にある語も含めて、語の重要度を計算し、重要文を抽出する手法である。そして、小学校の教科書に記されている8文書に対して実験参加者40人が重要文を抽出した結果を利用して、重要文抽出の一致点数を定義し、評価実験を行っている。その結果、提案手法が40点中23.6点に対し、tfidf法を用いた従来の2つの手法の時はそれぞれ20.5点と21.8点であった。このことから、連想関係によって関連付けられる語も含めて重要度の計算をする方が、人間に近い重要文の抽出ができることが示されている [65]。

2つ目の手法は、名詞連想概念辞書を用いて文書中の語から文脈ネットワーク (Contextual Semantic Network) を構築して利用する手法である。具体的には、ノードの活性値が隣接するノードに伝播する活性拡散の手法を用いて、語の重要度を計算し、重要文を抽出する。この文脈ネットワークは、文書中の語と関連性の高い語を含んでおり、人間の連想機能に近いモデルになっている。この手法に関しても、tfidf法を用いたベースラインを用いて評価実験を行い、その結果、人間の抽出結果に近いことが分かった。

これらの2つの手法の評価実験の結果より、名詞連想概念辞書を用いることで従来手法よりも高い精度を実現しており、人間に近い重要文の抽出が可能になることが示唆されている。

### 2.2.2 多義性の解消

人間は、文脈中に多義語すなわち複数の意味を持つ語が存在する時、前後の文脈の情報を用いて、この語の意味の同定を行っていると考えられる。コンピュータが多義語の意味の同定を正確にできなければ、全く異なる意味内容を出力してしまう処理もあるため、これらの多義性の解消に関する手法は今日までにも様々提案されている。中でも、岡本らは名詞連想概念辞書を用いて多義語を含む入力文に対して文脈ネットワークを作成し、文脈の内容に沿った適切な語の意味の同定手法を以下のように提案している。

まず、名詞連想概念辞書の刺激語と連想語に対応する2つの概念間の連想関係と

表 2.5: 多義性解消における評価実験の結果 [67]

	額 (がく)	額 (ひたい)	精度
WordNet (ベースライン)	8	6	0.58(14/26)
名詞連想概念辞書 (岡本ら)	15	9	0.92(24/26)

連想距離を用いて、概念のネットワークを構築する。そして、入力文中の自立語が入力される度に、この概念のネットワークから自立語に関する情報を抽出して文脈ネットワークを作成する。多義語が自立語として入力された際は、多義に関係する全てのノードをネットワークに追加する。ノードの活性値は、Interactive Activation Model を参考にした式により算出され、入力文内の自立語の情報によって強化、或いは弱体化され、適切な語義ほど活性値が高くなる仕組みになっている。そのため、多義語に対しては、最終的に文脈ネットワーク上で活性値が高いノードから、語義を決定する手法となっている。この手法の評価実験では、「額」という文字を含んでおり、且つ文中の名詞が全て名詞連想概念辞書の刺激語で成り立っている文を Web から評価用課題文 26 文を取得して、各文中の「額」に対して「額 (がく)」か「額 (ひたい)」の判別を行っている。なお、課題文 26 文は、「額 (がく)」を含む課題文が 16 文、「額 (ひたい)」を含む課題文が 10 文となっている。この評価実験の結果、表 2.5 のように、岡本らの提案手法は、WordNet を用いたベースライン手法と比べて高い正解率を得ている [67]。

このように、評価用課題文の数が少ないこと、そして「額」の判別に限定していることを踏まえると、岡本らの手法はカバレッジは低いですが、文中の名詞が全て刺激語から構成されるという一定の条件下では、多義性の解消において名詞連想概念辞書の有用性を示しているといえる。

### 2.2.3 比喩理解システム

坂口らは、名詞連想概念辞書を Integrated-and-Fire モデルによるニューラルネットワーク上に実装して意味ネットワークを構造化し、これを用いて比喩理解システ

表 2.6: 比喩理解システムの出力例 [2]

入力文	比喩理解の候補語	確信度
まるで雪のようなはだだ	白い	95
	冷たい	41
	きれい	11
まるで雪のような心だ	冷たい	95
	白い	39
	きれい	11
まるで鬼のような人間だ	怖い	95
	強い	88
	大きい	78

ムを構築している [72, 73, 74]. 具体的には、「まるで雪のようなはだだ」という直喩表現の入力文の「雪」と「はだ」に対して、意味ネットワーク上から共通概念の「白い」や「冷たい」を確信度付きで得て、最終的には「そのはだは白い」或いは「そのはだは冷たい」というような理解文を出力するシステムである. なお、この確信度はニューロン発火時の膜電位の値を用いて算出されている [73]. 参考として、この比喩理解システムの出力結果を表 2.6 に示す. このように、意味ネットワーク上から共通概念を得る際に、名詞連想概念辞書のデータに含まれている連想距離の値を用いて活性値を計算しており、活性伝播の仕組みを利用することで、人間の直感に合った比喩理解を可能にしている [2].



## 第3章

# 動詞連想概念辞書の構築

第3章では、動詞を刺激語とした連想実験について説明した後に、刺激語と連想語の単語間距離である連想距離の定量化について詳細を述べる。さらに、動詞連想概念辞書の連想語と連想距離の関係について具体的な例を挙げ、最後に既存辞書との比較分析を行い、連想情報と共起情報の性質の差異について言及する。



## 3.1 連想実験

本研究の連想実験は、オンライン上で稼働する実験システムを用いて行った。この実験システムでは、ブラウザの画面中央に刺激語の動詞と連想課題が呈示され、この呈示内容に対して実験参加者はキーボードを介して連想した内容を回答することで、データが記録されることになっている。以下では、この連想実験の刺激語の動詞、連想課題、そして実験参加者について詳細を述べるとともに、実験参加者が呈示内容に対して行った連想の状況を説明する。

### 3.1.1 刺激語の動詞

小学校の国語の教科書で用いられている1,123語の動詞 [32] を基本動詞と見なし、それらの中で基礎語として位置づけられている約200語の動詞 [54] を優先し、他の基本動詞と合わせて計239語を刺激語とした。これらの基礎語は、基礎日本語辞典 [54] の見出し語として取り上げられている全ての動詞の内、刺激語として不適切な語（「ある」、「なる」、「できる」）を除いたものである。また、自動詞「上がる」と他動詞「上げる」を例として挙げられるように、自動詞と他動詞の両方が基本動詞として存在し、且つ、どちらかの動詞が基礎日本語辞典の見出し語すなわち基礎語の場合、これらの動詞が同じ意味合いであれば、他動詞を優先することとした。なぜならば、3.1.2項で述べるように、連想課題は深層格の名称を参考に設定しており、他動詞には「対象」が存在するため、一般的に自動詞よりも深層格が多いからである。さらに、他動詞には受動態があるため、自然言語処理分野への応用を視野に入れた時に、自動詞よりも他動詞を刺激語の動詞として優先することで、受動態に対する処理も可能になるということが理由として挙げられる。

このように、基礎語の動詞を「見出し語」（基礎日本語辞典の見出し語）、「他動詞を優先した語」（自動詞と他動詞の両方が基本動詞であり、そのどちらかが基礎日本語辞典の見出し語）、「不適切な語」（刺激語の動詞としては不適切な基礎日本語辞典の見出し語）の3種類に分類し、連想実験の刺激語として設定した。なお、基礎語の動詞を分類した内容を表3.1に表す。

表 3.1: 分類した基礎語の動詞

分類	基礎語の動詞
見出し語	合う, 会う, 明ける, 開ける, 遊ぶ, 与える, 余る, 洗う, 争う, 表す, 現す, 歩く, 言う, 生きる, 行く, 急ぐ, 要る, 入れる, 浮く, 受ける, 動く, 歌う, 打つ, 撃つ, 討つ, 訴える, 奪う, 売る, 選ぶ, 得る, 追う, 覆う, 置く, 送る, 贈る, 起こす, 興す, 行う, 怒る, 教える, 驚く, 覚える, 思う, 返す, 書く, 囲む, 重ねる, 飾る, 貸す, 勝つ, 通う, 聞く, 着る, 切る, 組む, 比べる, 呉れる (くれる), 暮れる, 越える, 超える, 断る, 困る, 壊す, 指す, 差す, 挿す, 誘う, 去る, 敷く, 従う, しまう, 示す, 調べる, 知る, 過ごす, 住む, 済む, 座る, そびえる, 足す, 助ける, 尋ねる, 訪ねる, 頼む, 食べる, 黙る, 試す, 違う, 使う, 着く, 就く, 造る, 作る, 努める, 勤める, 務める, 詰める, 照らす, 出る, ともなう, 取る, 撮る, 採る, 眺める, 泣く, 鳴く, 無くす, 並べる, 似る, 脱ぐ, 除く, 望む, 伸ばす, 延ばす, 乗る, 入る, 計る, 測る, 量る, 運ぶ, 走る, 働く, 張る, 晴れる, 引く, 弾く, 轆く, 吹く, 拭く, 含む, 触れる, 参る, 増す, 待つ, まねる, 迷う, 認める, 迎える, 結ぶ, 持つ, 基づく, 求める, 破る, 遣る, 譲る, 許す, 寄せる, 呼ぶ, 依る, 喜ぶ, 割る
他動詞を優先した語	上げる, 預ける, 当てる, 集める, 生む, 落とす, 帰る, 掛ける, 欠く, 固める, 傾ける, 構える, 乾かす, 代える, 変える, 替える, 消す, 決める, 込める, 進める, 立てる, 建てる, 足らす, 捕まえる, 付ける, 伝える, 続ける, 通す, 届ける, 止める, 泊める, 留める, 流す, 慣らす, 残す, 載せる, 生やす, 始める, 冷やす, 曲げる, 交ぜる, 混ぜる, 見る, 回す, 向ける, 焼く, 汚す, 渡す
不適切な語	ある, なる, できる



表 3.2: 深層格を基にした連想課題と内容

連想課題	意味内容	助詞の例
動作主	動作を行う主体	～が, ～は, ～と
対象	動作の対象	～を, ～に
始点	動作の始点・起点	～から, ～より
終点	動作の終点・目標	～まで, ～へ, ～に
時点	動作が行われる時刻・時間	～に
場所	動作が行われる場所・空間	～で
手段	動作を行うための道具・材料	～で
様相	動作の様態・様子・程度・頻度	～に, ～で
理由	動作の理由・原因	～によって, ～で
目的	動作の目的	～のために

### 3.1.2 連想課題

刺激語の動詞と一緒に呈示される連想課題は、格文法における深層格 [18] の一部を参考にして、表 3.2 のように設定した。実験参加者が連想課題の意味内容を理解し易いように、いくつかの連想課題については、表 3.2 の「様相」のように「程度」や「頻度」などの複数の深層格を統合した。また、これらの深層格と使用される助詞の一部を例として、実験参加者に説明した内容も表 3.2 に示す。

連想実験では、刺激語の動詞によって連想課題が異なり (e.g., 自動詞の場合は「対象」を連想課題から外す, 移動を表す動詞の場合は「始点」や「終点」を連想課題に加えるなど), 1 刺激語 1 連想課題を 1 問とした。また、「動作主」, 「時点」, 「場所」, 「手段」, そして「様相」の 5 つの連想課題に関しては、どの動詞に対しても対応する深層格が存在すると考えられるため、刺激語の動詞の種類に依らず、実験参加者に呈示されるように設定した。なお、1 回の実験で呈示される刺激語の動詞は全て異なり、実験参加者はランダムに呈示される約 40 問に対して連想語を答える形式になっている。

### 3.1.3 実験参加者

実験参加者は、慶應義塾大学湘南藤沢キャンパス (SFC) に所属する大学生と大学院生から成っており、動詞連想概念辞書を構築する当初は延べ 170 人 [85] であり、段階的に実験参加者数を増やした。また、実験参加者には連想実験を行う前に下記の項目について説明し、呈示された内容に対して短時間で連想するように指示した。

- 1 問 (1 刺激語と 1 連想課題) につき、約 30 秒で連想語を答えること。
- 呈示内容に対して思いついた連想語は、なるべく多く答えること。
- 連想語をしばらく考え込むようであれば、実験を先に進めること。

なお、オンライン上での個人情報の漏洩を防ぐために本実験システムでは氏名や年齢、性別の入力を不要としたため、実験参加者の年齢の平均や分散、男女構成などの属性情報の詳細は不明である。

## 3.2 連想距離の定量化

名詞連想概念辞書の手法 [64] と同様に、刺激語  $x$  に対する連想語  $y$  の連想され易さを単語間距離と見なし、連想距離  $D(x, y)$  を連想頻度の逆数  $F(x, y)$ 、連想順位  $S(x, y)$ 、連想時間  $T(x, y)$  の線形結合で (3.1) 式のように表せると仮定した [90, 92].

$$D(x, y) = \alpha F(x, y) + \beta S(x, y) + \gamma T(x, y), \quad (3.1)$$

$$F(x, y) = \frac{N}{n(x, y) + \delta}, \quad (3.2)$$

$$\delta = \frac{N}{10} - 1 (N \geq 10), \quad (3.3)$$

$$S(x, y) = \frac{1}{n} \sum_{i=1}^n s_i(x, y), \quad (3.4)$$

$$T(x, y) = \frac{1}{n} \sum_{i=1}^n \log t_i(x, y). \quad (3.5)$$

$F(x, y)$  は刺激語  $x$  に対して連想語  $y$  が連想された頻度の逆数を表し、この連想をした人数  $n(x, y)$  に補正值  $\delta$  を加えた値で実験参加者数  $N$  を割った値である。これは

実験参加者数が大幅に増加した時に連想した人数が少ない場合でも  $F(x, y)$  の値が大きくなり過ぎないように配慮することで、連想距離の過度な変動を抑えたものになっている。そして、 $S(x, y)$  は各参加者が刺激語  $x$  に対して連想語  $y$  を連想した順位  $s_i$  の相加平均である。また、刺激語と連想課題が提示されてから連想語を入力し始めるまでの時間を連想時間  $t_i$  としているが、実験参加者間のみならず個人の中においても大幅な変動が生じるために連想時間の対数の相加平均  $T(x, y)$  を用いている。

次に (3.1) 式の係数  $\alpha, \beta, \gamma$  を決定変数として最適値を求めるため、以下のように目的関数  $Z$  と制約条件 (3.7)~(3.9) を設定し、線形計画法を用いて値  $Z$  を最小化する  $\alpha, \beta, \gamma$  の最適解を求める。目的関数の  $d_1, d_2, d_3$  は、連想実験で回答された全ての連想語に対する連想頻度の逆数、連想順位、そして連想時間の相加平均とし  $\alpha, \beta, \gamma$  はこれらの3パラメータによって求められる値  $Z$  を最小化する係数を指している。

$$\text{目的関数 } Z = d_1\alpha + d_2\beta + d_3\gamma, \quad (3.6)$$

$$\text{制約条件 } a_1\alpha + b_1\beta + c_1\gamma \geq 1.0, \quad (3.7)$$

$$a_2\alpha + b_2\beta + c_2\gamma \leq 10.0, \quad (3.8)$$

$$\alpha, \beta, \gamma \geq 0. \quad (3.9)$$

動詞連想概念辞書を構築する際の実験参加者は1刺激語1連想課題につきまず20人で行い、刺激語54語で延べ約170人の規模であった。そのため、(3.7)式の制約条件は、「実験参加者全員 (i.e.,  $N = n(x, y) = 20$ ) が同じ単語を最初に連想し、その単語を入力し始めるまでの時間が十分に短かった」場合の時に得られる  $F(x, y), S(x, y), T(x, y)$  の値を用いて  $(a_1, b_1, c_1) = (20/21, 1.00, 1.00)$  とし、連想距離が最小値1.0になることを表している。その反対に、(3.8)式は「実験参加者のうち1人 (i.e.,  $N = 20, n(x, y) = 1$ ) しか連想せず、連想した順位も低く、入力し始めるまでに十分長い時間を要した」場合の時に得られる上述の3つの値を用いて、 $(a_2, b_2, c_2) = (10.00, 9.00, 5.00)$  とし、この時の連想距離が最大値10.0になることを表している。そして、目的関数においては実験参加者全員の約170人分の実験データから得られた  $(d_1, d_2, d_3) = (7.89, 2.82, 2.64)$  を用いてシンプレックス法によって最適解を求めた。その結果、 $(\alpha, \beta, \gamma) = (7/10, 1/3, 0)$  の解を得た。

そもそも、線形不等式の制約は、幾何学的には実行可能領域と呼ばれる凸多面体を定義し、線形の目的関数の最適解は、その凸多面体の頂点（実行可能領域の境界上）になる。上記の制約条件 (3.7)～(3.9) では、決定変数が 3 つのため、3 次元空間座標内の凸多面体の頂点の時に、目的関数  $Z$  が最小化となる。このような、決定変数の最適解が少なくとも 1 つは 0 になるモデルを用いている理由は、実験データから得られた 3 つのパラメータの内、刺激語から連想語がどれくらい連想され易いかを表す連想距離を定量化する上で、より信頼できるパラメータを選んで利用することを図ったためである。連想時間の係数が 0 になったことから、連想距離は連想頻度及び連想順位で表すことになった。以上より、連想距離  $D(x, y)$  を (3.10) 式で表すこととした。

$$D(x, y) = \frac{7}{10}F(x, y) + \frac{1}{3}S(x, y). \quad (3.10)$$

### 3.3 連想語と連想距離の関係

動詞連想概念辞書のデータの例として、刺激語「運ぶ」に対する連想語の一部を表 3.3 に示す。この表では、各連想課題における連想語を連想距離の降順に記しており、括弧内の数値が「運ぶ」とその連想語間の連想距離を示している。連想距離の値が小さい程、刺激語に対して連想し易いことから、「運ぶ」の各連想課題においては表 3.3 に記載した順に、実験参加者が連想し易かったといえる。

また、反意語関係にある「借りる」と「貸す」における「動作主」、「対象」、「場所」、「様相」の連想語の一部と連想距離の関係を図 3.1 に示す。図 3.1 の「動作主」、「対象」、「場所」における連想語に関しては、各々の連想距離の大小関係が「借りる」と「貸す」の間で、ある程度似ている傾向があることが分かる。さらに、「対象」と「場所」に関して「本」と「学校」の連想距離がそれぞれ小さい値であることから、学生から成る実験参加者の連想の特徴が表されている。また、「様相」の連想語に注目すると「たくさん」と「しぶしぶ」の連想距離の関係が入れ代わっており、「借りる」と「貸す」の反意語関係が影響する人間の心理的な面が連想距離で表されていることが分かる。このように、連想語と連想距離は、人間の一般的な心理的側面を反映しているが、一部の連想語については学生の連想による特徴が確認できる。

表 3.3: 刺激語「運ぶ」に対する連想語のデータ例

連想課題	「運ぶ」の連想語 (連想距離)
動作主	私 (3.60), 引越し業者 (4.21)
対象	荷物 (1.36), 人 (4.45), 家具 (7.78)
始点	家 (1.45), 学校 (3.81), 部屋 (5.80)
終点	家 (1.92), 駅 (3.73), 目的地 (4.02)
場所	倉庫 (3.73), 工場 (4.50), 会社 (6.22)
手段	車 (1.62), トラック (2.34), 手 (3.47)
様相	必死に (3.17), 一生懸命に (3.36)

そこで、実験参加者が学生から成ることによる動詞連想概念辞書への影響について述べる。上述のように、動詞連想概念辞書は実験参加者すなわち学生の連想から成り立っており、表 3.3 の「始点」の連想語である「学校 (3.81)」や、図 3.1 の「対象」における「本」, 「場所」における「学校」など、連想距離の値が中位の一部の連想語は、明らかに学生による連想の特徴と捉えることができる。そのため、学生と一般の連想の乖離により、動詞連想概念辞書が学生の連想の特徴に偏っている可能性がある。しかし、表 3.3 で連想距離の値が特に小さい、すなわち連想語の上位として「対象」の「荷物 (1.36)」, 「始点」の「家 (1.45)」, 「終点」の「家 (1.92)」, 「手段」の「車 (1.62)」, そして図 3.1 における「対象」の「金 (かね) (1.79)」と「金 (かね) (1.76)」に注目すると、学生による連想の特徴を表すような単語ではなく、どれも一般的な単語といえる。さらに、他の刺激語についても同様の傾向が見られるかを確認するため、刺激語「歩く」に対する各連想課題の連想語 (特に連想距離の値が小さいもの) と連想距離の関係を表すと、図 3.2 のようになった。この図において、「始点」と「終点」の連想語と連想距離に注目すると、それぞれ「学校 (2.31)」と「学校 (1.88)」という学生の連想を特徴付ける連想語が上位にあるが、これら以外の連想語として「始点」では「家 (1.38)」と「駅 (2.48)」, 「終点」も同様に「家 (1.93)」と「駅 (2.17)」が、いずれも上位にあり、且つ、一般的な単語となっている。すなわち、学生特有の連想語は基本的には連想距離の値が中位な関係となっており、学

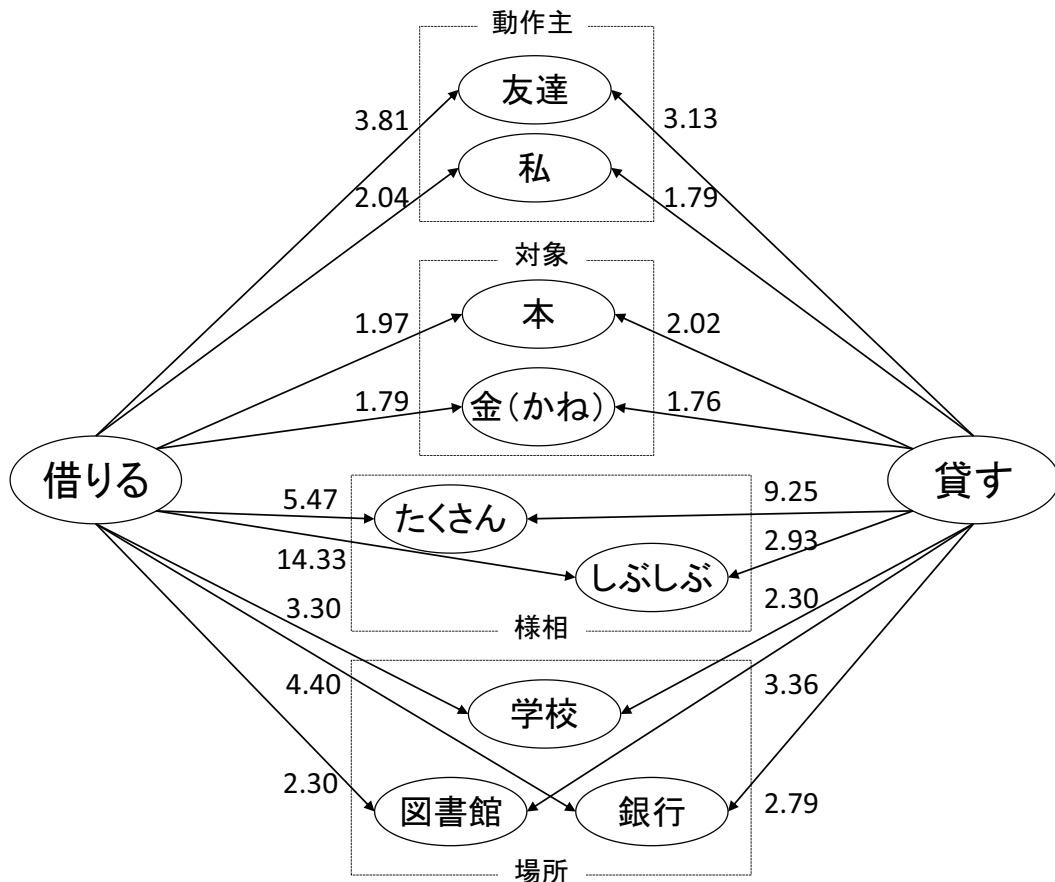


図 3.1: 刺激語「借りる」と「貸す」に対する連想語の関係

生特有の連想語が上位に多少は存在したとしても、連想距離の値が特に小さい上位の連想語は、一般的な単語が占めている。

以上から、動詞連想概念辞書は学生による連想から成り立っているが、連想距離の値が特に小さい連想語、つまり刺激語に対して連想し易い関係にある連想語に関しては、実験参加者が学生であることに依存しておらず、常識とも言うべき一般的な情報が含まれていると考えられる。なお、ここで説明した動詞連想概念辞書は、連想距離を定量化して構築した当初の段階（刺激語 54 語、1 刺激語当たり実験参加者 20 人）よりも拡張されており、刺激語 239 語で 1 刺激語当たり実験参加者が 40 人からの実験データから成っている。これらの刺激語 239 語には、3.1.1 項で説明した基礎語の動詞が全て含まれており、連想語数が延べ約 97,000 語、異なり約 24,000 語の規模となっている [90]。

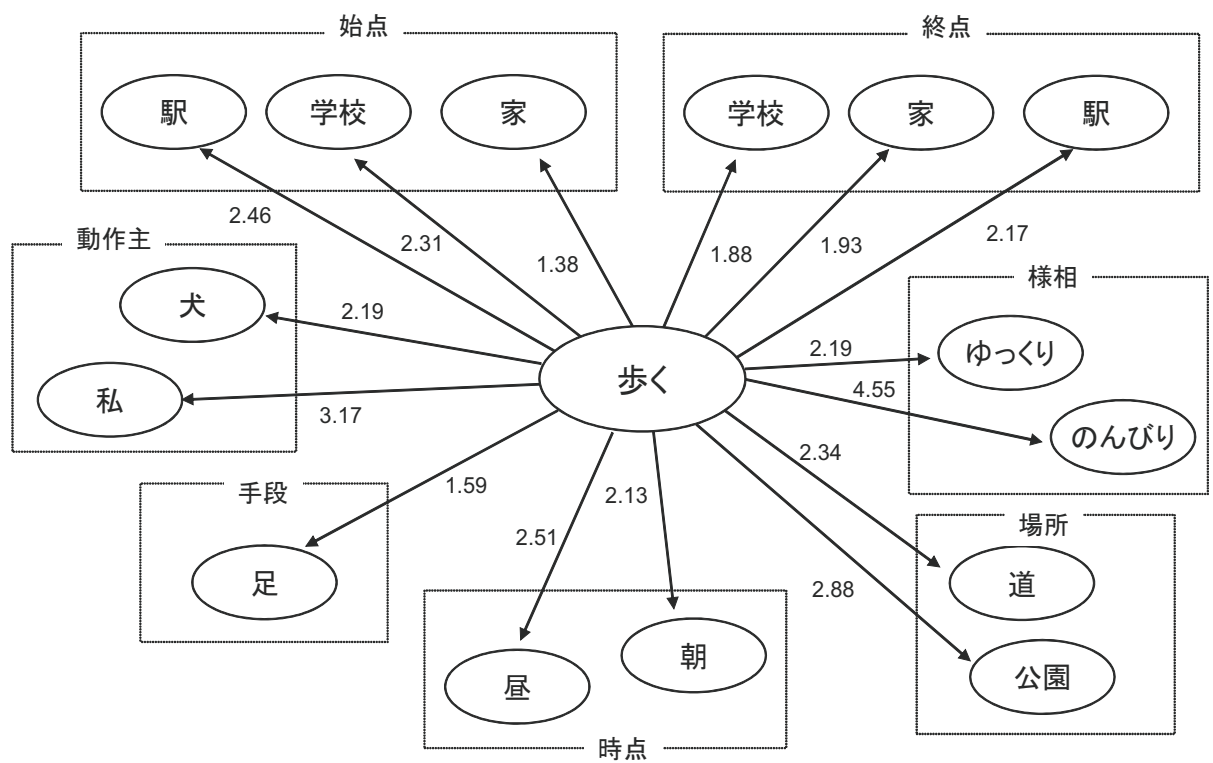


図 3.2: 刺激語「歩く」から連想距離が比較的近い連想語

### 3.4 既存辞書との比較による動詞連想概念辞書の特徴

動詞連想概念辞書を既存の辞書と比較することで、その特徴を分析した。ここでは、動詞連想概念辞書とデータ構造が最も類似していると考えられる京都大学格フレーム辞書 [44] を比較の対象とした。

格フレーム辞書は、Web 上の約 16 億文から自動構築されたものであり、述語とその格フレームの膨大な情報から成り立っている [36]。そもそも格フレームとは、述語とそれが表層の格関係（ガ格、ヲ格、ニ格、デ格など）を持つ単語を整理して記述したものである。述語に関しては、「積む/つむ:動:P1」を例に挙げて説明すると、「積む/つむ」という表記の他に、動詞や形容詞などを区別する述語タイプや、格交替を起こす可能性のある付属語が付いた態タイプの情報に分かれている。「積む/つむ:動:P1」では、「動」は「動詞」を表す述語タイプを、「P」は「れる、られる」を表す態タイプをそれぞれ指している。また、格フレームの格の種類はガ格、ヲ格、ニ格、へ格、カラ格、デ格などの表層格があり、これらの格に入る単語（格要素）とその頻度が記されている。例えば、「積む/つむ:動 1」のヲ格の格要素の情報としては「経験:37342」と「体験:1363」などがあり、ヲ格に入る単語とその単語を含む文の頻度が分かる。そのため、異なる態タイプでも述語の表記が同じ動詞は、ある格に対する格要素が共通する場合に各々の頻度を足し合わせることで、その動詞の表層格に入る単語の頻度（i.e., ある格における動詞と単語の共起頻度）を求めることができる。ゆえに、格フレーム辞書における述語動詞とその格フレームの格（表層格）に入る単語及びその頻度の関係は、動詞連想概念辞書の刺激語動詞と連想課題（深層格）に対する連想語及びその連想距離の関係と対応している。そこで、刺激語動詞の 239 語において動詞連想概念辞書の深層格に基づいた連想課題に対する連想語と対応する格フレームの表層格に入っている単語で共通するものを抽出して両者の比較を行った。なお、動詞連想概念辞書の動詞は全て原形を刺激語としていることから、格フレーム辞書の述語動詞も態タイプが能動態のものを扱った。

表 3.4 と表 3.5 は、動詞連想概念辞書の連想語と格フレームの表層格に入る単語が 5 個以上存在した動詞における連想距離の昇順の順位と格に入る単語の頻度の降順の順位に対して、スピアマンの順位相関係数を求め、連想課題と格の組み合わせご



表 3.4: 動詞連想概念辞書と格フレーム辞書における共通語 ( $n > 5$ ) の順位相関係数の平均

	動作主 - ガ格	対象 - ヲ格	始点 - カラ格
動詞の語数	205	184	26
順位相関係数の平均	0.272	0.185	0.224

表 3.5: 動詞連想概念辞書と格フレーム辞書における共通語 ( $n > 5$ ) の順位相関係数の平均

	終点 - マデ格	場所 - デ格	手段 - デ格
動詞の語数	9	131	141
順位相関係数の平均	0.342	0.183	0.205

とにその平均を表したものである。なお、表中の動詞の単語数は、共通の単語が5個以上ある動詞における単語数である。格フレーム辞書を構成している述語と格関係は Web 文書から得ていることから、比較したデータは Web 文書における共起情報と捉えることができる。そのため、表 3.4 と表 3.5 より、動詞連想概念辞書と格フレーム辞書で同じ動詞に関し、対応する連想課題と格フレームで重複する単語同士の順位の相関がほとんどない、或いは低いことから、同じ動詞に対して連想され易いことと共起し易いことが必ずしも一致していないことがいえる。

そもそも、連想と共起の理論的な違いとして、情報の産出過程が異なることが挙げられる。前者の連想は、本研究では刺激語と連想課題の組み合わせに対して、実験参加者が思いついた内容が反映されている。一方、後者の共起は、既に Web 上で記述された文書や文章に対して、同じ文中に存在する単語同士を整理した内容が反映されている。つまり、連想は動詞と連想課題に対して思いついた内容であり、共起は動詞と同じ文中で記述された内容である。そのため、このような産出過程の違いによって、連想と共起で表された内容が、実際どのように異なるのかを確認する

必要がある。

そこで、連想と共起がどのような点が異なるのかを調べるために、3.3 節で触れた「借りる」と「貸す」に関して、動詞連想概念辞書の連想課題「対象」の連想語と格フレーム辞書の「ヲ格」の格要素で共通している単語の一部をまとめると、表 3.6 になる。この表から、動詞連想概念辞書の方では「借りる」と「貸す」の共通の連想語の並びが同様であることに対して、格フレーム辞書の方ではほとんど異なっていることが分かる。例えば、後者の格フレーム辞書では「力を借りる」と「手を貸す」の頻度が「手を借りる」と「力を貸す」よりも高い。すなわち「借りる」と「貸す」によって単語の共起のし易さが逆転しており、Web 文書の作成者が文章の中で「借りる」と「貸す」の事例で取り上げる傾向を表している。一方、前者の動詞連想概念辞書では「金を借りる」と「金を貸す」、「本を借りる」と「本を貸す」、「手を借りる」と「手を貸す」の順で連想され易いということが分かる。これはすなわち、「金を借りる/金を貸す」や「本を借りる/本を貸す」という動作の担い手と担われ手に関係なくその状況そのものが「借りる」や「貸す」の動作から連想され易いということを示しており、反意語関係が背景となる言葉の対称性を示していると考えられる。このような点で動詞連想概念辞書を構成する連想情報は、共起情報と性質が異なっている。

### 3.5 動詞の汎用性に関する考察

3.3 節と 3.4 節では、「運ぶ」、「歩く」、「借りる」そして「貸す」が刺激語の時の連想語について述べてきたが、これらの動詞が示す動作は、ある程度限定された状況で行われるため、連想課題によっては刺激語の動詞特有の連想語が、連想距離の値が小さい上位にある場合が生じていた (e.g., 「借りる」と「貸す」における「対象」の連想語「金 (かね) (1.79)」と「金 (かね) (1.76)」など)。そのため、これらの刺激語よりも、動詞の汎用性が一般的に高いと考えられる「取る」を刺激語にした場合の連想語について以下で言及する。

表 3.6: 動詞連想概念辞書の連想情報と格フレーム辞書の共起情報の比較

順位	連想語と連想距離 (※昇順)		格要素と共起頻度 (※降順)	
	「借りる」の対象	「貸す」の対象	「借りる」のヲ格	「貸す」のヲ格
1	金 (1.79)	金 (1.76)	力 (18918)	手 (8297)
2	本 (1.97)	本 (2.02)	金 (14864)	耳 (7228)
3	…	…	…	金 (6422)
…	…	…	…	…
6	…	…	本 (7364)	…
…	…	…	…	…
8	手 (6.88)	手 (8.44)	手 (5193)	…
9	…	力 (8.78)	…	本 (608)
…	…	…	…	…
18	力 (9.25)	…	…	…
19	…	耳 (10.58)	…	…
…	…	…	…	…
44	肩 (15.00)	…	肩 (577)	…
…	…	…	…	…
59	…	肩 (16.33)	…	…

表 3.7: 刺激語「取る」に対する連想語のデータ例

連想課題	「取る」の連想語（連想距離）
対象	物 (5.13), 荷物 (5.97), コップ (7.33), 財布 (10.08), 帽子 (10.25)
場所	部屋 (3.88), 教室 (5.88), 店 (5.88), 家 (6.05)
手段	手 (1.62), 箸 (4.50), 棒 (6.05), 足 (7.78), 綱 (8.67)

表 3.7 は、「取る」に対して「対象」、「場所」そして「手段」について、それぞれの連想語を最上位すなわち連想距離の値が最も小さい連想語から順番に 4~5 語を並べたものである。この表の「対象」と「場所」を見ると、「取る」という動詞特有の連想語が表されているとはいえない。たしかに「対象」の「物 (5.13)」と「場所」の「部屋 (3.88)」は、それぞれの連想課題の中では最も連想距離の値が小さい連想語のため、「取る」に対して最も連想し易い語といえる。しかし、連想距離の値そのものは、「借りる」や「貸す」に対する「金 (かね)」の連想距離“1.79”と“1.76”のように、いずれも小さくないため、「取る」に対して特に連想し易い関係にあるわけではないことを示している。

ゆえに、動詞連想概念辞書の中では汎用性のある動詞については、基本的には動詞特有の連想語がない可能性が考えられる。そして、連想課題において最も連想し易い連想語に関しては、その連想距離の値が特に小さくなく、中位の大きさであることが、汎用性のある動詞の特徴として表れていると考えられる。

### 3.6 動詞連想概念辞書の拡張

構築した動詞連想概念辞書は、表 3.8 のように段階を経ながら拡張し、研究に応用してきた。2013 年の現在、動詞連想概念辞書の刺激語の基本動詞は、基礎日本語辞典の項目で取り上げられている動詞の他に、関連語として取り上げられている動詞も全て網羅している。そのため、2013 年の規模を持って、動詞連想概念辞書の第一段階が完成したと見なすことができる。

表 3.8: 動詞連想概念辞書の規模

時期	刺激語数	実験参加者数	連想語延べ数	連想語異なり数	応用研究
2008年	54語	1語 20人	約 12,000語	約 5,400語	[85]
	54語	1語 40人	約 23,000語	約 9,400語	[86]
2009年	232語	1語 20-40人	約 58,000語	約 16,000語	[87, 88, 89]
2011年	239語	1語 40人	約 97,000語	約 24,000語	[90, 92]
	252語	1語 40人	約 101,000語	約 24,000語	[91]
2012年	345語	1語 40人	約 135,000語	約 30,000語	[83, 84]
2013年	440語	1語 40人	約 176,000語	約 36,000語	—



## 第4章

# 連想情報を用いた省略語の推定

第4章では，動詞連想概念辞書の応用研究の1つとして構築した省略語推定システムについて説明する．このシステムは，名詞連想概念辞書と動詞連想概念辞書を併用しており，名詞と動詞の連想情報に基づいた省略語の推定を可能にしている．以下では，この提案手法の詳細を説明した後に，評価実験を通して提案手法の有用性について言及する．





## 4.1 はじめに

日本語は、日常の会話や文章において、主語や目的語が省略される場合が多く [57], 特に読み手や聞き手が簡単に推測できる内容については頻繁に省略される. 機械翻訳や対話システムなどにおいて, コンピュータが会話内容や文章の意味を理解して正しい出力を行うためには, これらの省略語を正確に補完する必要がある. そのため, 自然言語処理の分野で省略語を推定するタスクは, 極めて重要である.

今日までに, 確率モデルによる省略解析の手法 [75] や, 格フレーム辞書を用いてゼロ代名詞を検出する手法 [35] などが提案されている. これらの研究は, 主格或いは目的格に入る語に対して照応解析を行い, 省略語を補完している. 従来の研究では, 主格や目的格などの表層格において, 省略された要素や照応詞が何を指しているのかを示す照応・省略解析が主流であるが, 本研究では動詞と直接意味関係を持つ深層格に入る単語を推定対象として, 人間の直感に近い省略語が推定できることを目指す.

## 4.2 提案手法

本研究では, ある特定の深層格に関して省略語を自動で推定する省略語推定システムを構築した. このシステムは, 本研究で構築した動詞連想概念辞書と既存の名詞連想概念辞書を組み合わせて構成する意味ネットワークを利用し, 省略語を確信度付で出力する. 以下では, 提案手法に基づいた省略語推定システムの概要について触れた後に, 省略語候補の抽出と確信度の計算の処理の詳細について, それぞれ説明する.

### 4.2.1 システムの概要

図 4.1 は, 本手法の省略語推定システムの概要を表している. 図中の四角と矢印に記された内容は, 処理対象の情報, もしくは得られた情報を指している. また, 菱形に記された内容は, 具体的な処理の内容を指している. これらの処理の流れを以

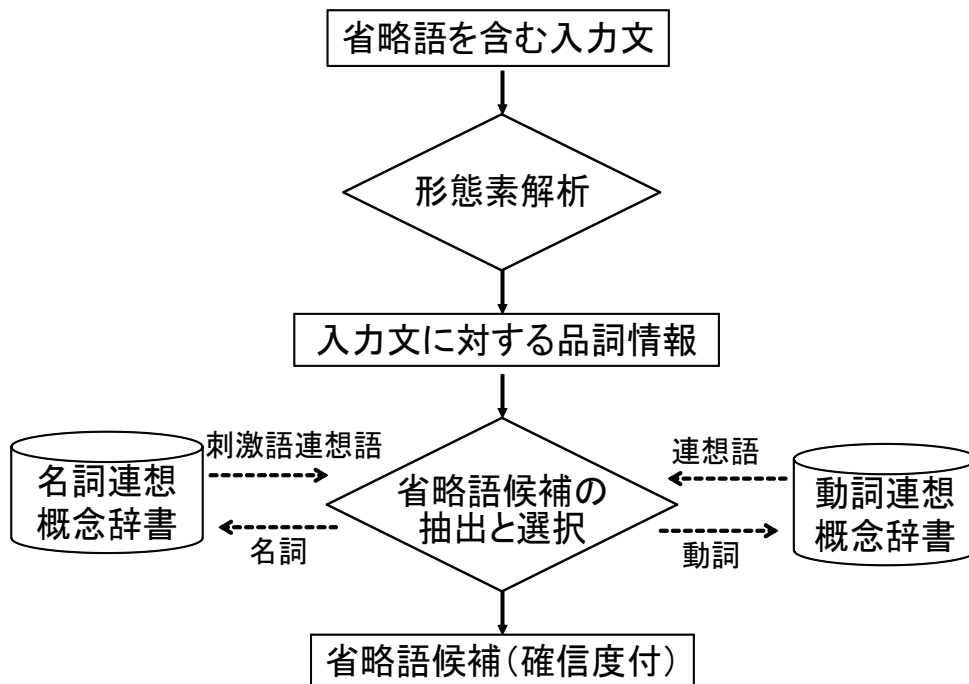


図 4.1: 省略語推定システムの概要

下に概説する．まず，省略語を含む日本語入力文に対して形態素解析を行い，述語動詞に対して推定させる深層格の連想語を動詞連想概念辞書から抽出し，省略語候補とする．次に，名詞連想概念辞書を用いて入力文中の全ての名詞に対して，連想語や刺激語の情報を抽出する．最後に，動詞連想概念辞書と名詞連想概念辞書の連想距離を用いて省略語候補の確信度を求め，この確信度の降順に省略語を出力する [90, 92].

#### 4.2.2 名詞連想概念辞書

省略語候補の抽出の際に，本研究で構築した動詞連想概念辞書の他に名詞連想概念辞書を用いるため，ここで名詞連想概念辞書について簡単に触れておく．2.1.5 項で説明したように，名詞連想概念辞書は，小学校の国語の教科書で用いられる基本的な名詞を刺激語とした連想実験によって表 4.1 の連想課題にあてはまる単語を実験参加者が答えた内容がまとめられている．表 4.1 は，設定されている連想課題ならびに刺激語名詞が「辞書」の時の連想語の一部を例として載せたものであり，こ

表 4.1: 名詞連想概念辞書の連想課題と連想語の例

連想課題	内容	連想語 (刺激語 : 辞書)
上位概念	刺激語の上位語を表す概念	書物, 本
下位概念	刺激語の下位語を表す概念	国語辞書, 英和辞書
部分・材料概念	刺激語の部分や材料を表す概念	表紙, ページ
属性概念	刺激語の特徴を表す概念	厚い, 重い
類義概念	刺激語の同義語や類義語を表す概念	辞典, 事典
動作概念	刺激語に関係する動作を表す概念	調べる, 持つ
環境概念	刺激語が存在する場所を表す概念	図書館, 本棚

ここでは刺激語と連想語の関係を概念として定義している。なお、動詞連想概念辞書と同様に、刺激語の名詞と連想語の連想距離が定量化されている [64]。なお、本手法で用いた名詞連想概念辞書の規模は、約 1,100 語の刺激語名詞に対して連想語数が延べ約 280,000 語、異なり約 64,000 語となっている [66]。

### 4.2.3 省略語候補の抽出

省略語候補を抽出する過程を以下に記すと、(1) 入力文を形態素解析して得た述語動詞に関して、指定した深層格の連想語を全て動詞連想概念辞書から抽出する、(2) 抽出した連想語を省略語候補として連想距離の情報とともに候補リストに一時的に格納する、(3) 名詞連想概念辞書の刺激語名詞と連想語の情報を用いて候補リストの中から省略語候補の選択を行う、という 3 つの処理が挙げられる。例えば、「財布を忘れたので友達から借りた」という文に対して (1) から (3) までの過程を説明すると、何を借りたのかを出力させる場合は動詞連想概念辞書から刺激語動詞「借りる」の連想課題「対象」に関する連想語を全て抽出し ((1) の処理)、連想距離とともに候補リストに格納する ((2) の処理)。一方で、入力文を形態素解析して得た名詞が名詞連想概念辞書に刺激語名詞として登録されている場合は、表 4.1 の「環境概念」と「部分・材料概念」の連想語を抽出する。ここで抽出する連想語は、刺激語が存在

する場所や環境，或いは刺激語を構成する部分や材料を表し，これらの関係を持つ名詞同士のどちらかが文中で省略されている時，文中にある方の名詞からその省略された名詞を理解できる場合が多いため，これらの概念を優先して用いている．また，省略語の候補を幅広く集めるために，刺激語名詞としてではなく連想語として登録されている場合は，その刺激語を抽出する．上記の例文の場合は，刺激語の名詞が「金（かね）」の時の「環境概念」の連想語に「財布」があるため刺激語の「金（かね）」が得られ，刺激語名詞が「友達」の時の「部分・材料概念」の連想語として「手」が得られる．連想語や刺激語に対して上位の関係にある連想語や刺激語の中で連想距離の値が小さい単語に関しても，同様に「環境概念」と「部分・材料概念」の連想語や逆引きから得られる刺激語を抽出する（(3) の処理）．

以上のように，文中の名詞だけでなく上位語に関しても連想語や刺激語の連想情報を抽出することで，文中に無い上位語の情報も追加することができる．また，入力文に固有名詞が含まれている場合は，固有名詞が連想語として登録されている刺激語に関して上記と同様に「環境概念」と「部分・材料概念」にあたる連想語を抽出する．

#### 4.2.4 確信度の設定

省略語候補の確信度は，動詞連想概念辞書の連想距離  $L_V (L_V \geq 1.0)$  と以下に述べる場合に分けて名詞連想概念辞書の連想距離  $L_{N1} (L_{N1} \geq 1.0)$  [64] 或いは  $L_{N1}$  の平均  $L_{N2} (L_{N2} \geq 1.0)$  をそれぞれ用いて表す．まず，動詞連想概念辞書から抽出した省略語候補が名詞連想概念辞書に連想語或いは刺激語として存在する場合は，(4.1) 式のように名詞連想概念辞書の連想距離  $L_{N1}$  を用いて確信度  $e_1$  を求める．この  $L_{N1}$  は，入力文中の名詞と省略語候補の単語との連想距離を指している．

$$e_1 = \frac{1}{L_V L_{N1}} \quad (0 < e_1 \leq 1). \quad (4.1)$$

一方，省略語候補が名詞連想概念辞書に無い場合，その省略語候補の確信度  $e_2$  は，名詞連想概念辞書に存在する省略語候補  $m$  語の連想距離  $L_{N1}$  の相加重平均  $L_{N2}$  を用いて (4.2) 式と (4.3) 式から求める．これは省略語候補が名詞連想概念辞書に存在す

るが、動詞連想概念辞書の連想距離  $L_V$  が非常に大きい単語、つまり動詞に対して省略語候補が連想され難い関係にある単語しかない場合、確信度は全体的に低い値になる。このように、動詞連想概念辞書から抽出した省略語候補の内、省略語として出力するのに適している単語が名詞連想概念辞書内に無いことで確信度が計算できず、最終的に出力されないということを防ぐため、(4.2)式でこれら確信度を近似的に計算している。

$$e_2 = \frac{1}{L_V L_{N2}} \quad (0 < e_2 \leq 1), \quad (4.2)$$

$$L_{N2} = \frac{1}{m} \sum_{z=1}^m L_{N1}. \quad (4.3)$$

4.2.3項で取り上げた「財布を忘れたので友達から借りた」という例文に対して「借りる」の対象に当たる単語を推定したところ、「金（かね）」や「本」、「手」などが出力された。「金（かね）」の確信度 0.274 に比べて「本」は 0.057、「手」は 0.020 のように他の確信度は極端に低い。この例文は、財布を忘れた人が友達から借りた物として「金（かね）」を借りたことが直観的に類推できることから、推定結果の内容の差を踏まえると確信度は省略語として推定される度合いをある程度表現できていると考えられる。

## 4.3 評価実験

### 4.3.1 ベースラインの作成

本手法の有効性を調べるためにベースラインを設定し、Web上のブログ文書中で用いられている文に対して省略語の推定をそれぞれ行った。ベースラインは、動詞連想概念辞書のデータ構造と最も類似していると考えられる格フレーム辞書（京都大学格フレーム Ver1.0）[44]をはじめ、名詞連想概念辞書、そして動詞連想概念辞書のそれぞれから成るシステムを用いて省略語を推定した結果である。本研究の評価では、人間が実際に推測した内容を正解と設定し、これらのベースラインと提案手法を比較して分析することで省略語の推定における手法の有効性を示す。

表 4.2: 動詞連想概念辞書の連想課題と対応させた格

格	連想課題
ガ格	動作主
ヲ格	対象
カラ格, ヨリ格	始点
マデ格, ヘ格	終点
デ格	場所, 手段

### Baseline 1 (CF)

格フレーム辞書 (Case Frame) を用いて省略語を推定する手法を CF とする。2.1.3 項で述べたように、格フレーム辞書は表記の他に「動詞」「形容詞」「名詞+判定詞」を表す述語タイプ、「れる、られる」などの格交替を起こす可能性のある付属語を表す態タイプで述語の種類が分かれていることから、省略語を推定させる課題文に対してどの述語動詞のタイプが適しているか同定する必要がある。そのため課題文を形態素解析した後に文中で述語動詞に係っている名詞を格要素として含む述語動詞の格フレームを探す。その格フレームを持つ述語動詞のタイプの内、その名詞の頻度が最も大きいタイプを選ぶ。この述語動詞のタイプに対して、課題文で推定させる内容に対応させた表 4.2 の格の格要素を省略語候補として全て抽出し、頻度の降順で出力する。以上のようにして、本研究の提案手法と同様に課題文の述語動詞と文中の名詞の両方の情報を考慮した省略語の推定を行う。

### Baseline 2 (NACD)

名詞連想概念辞書 (Associative Concept Dictionary for Nouns) のみを用いて省略語を推定する手法を NACD とする。名詞連想概念辞書内に文中の名詞と表 4.1 に記した「部分・材料概念」或いは「環境概念」の関係を持つ刺激語或いは連想語との間の連想距離の逆数を確信度とし、降順で出力する。すなわち述語動詞に関係なく、文中の名詞から得られる名詞連想概念辞書の情報に基づいて省略語の推定を行う。

### Baseline 3 (VACD)

NACDと異なり、名詞連想概念辞書を使わずに動詞連想概念辞書 (Associative Concept Dictionary for Verbs) のみを用いて省略語を推定する手法を VACD とする。本研究の提案手法で動詞連想概念辞書は省略語候補を抽出することに用いられており、動詞連想概念辞書のみでも省略語候補を求めることは可能だが、提案手法と同様な確信度の計算はできない。そのため、NACD と同様に動詞連想概念辞書内における述語動詞と省略語候補の連想距離の逆数を確信度と見なし、その降順で出力する。つまり文中の名詞はまったく考慮されず、述語動詞に関する動詞連想概念辞書の情報から成っている。

#### 4.3.2 課題文と正解の設定

本評価では、人間が推測する省略語と比べて本研究のシステムがベースラインを基準としてどのように有効なのかを調べることを目的としている。目的格などの必須格の他に一般的には無くても意味が通じる自由格も含め、省略語が頻出するブログから評価用の課題文を抽出し、実験参加者が推測した省略語の内容を正解データとして設定した。以下で、課題文の抽出と正解の設定に至る過程を述べる。

評価用の課題文は、動詞連想概念辞書の刺激語動詞を含む約 41,000 のブログ文書から以下の手順で抽出した。このブログ文書は研究利用を目的にデータセクション (株) から提供を受けたものであり、刺激語の前後 400 文字で抜粋した文章で構成されたデータセットだったため、文が途中で切れる、或いは句点がまったく使われていない場合があった。そのため、句点が文末に存在している約 104,000 文を抽出した。ここから、提案手法とベースラインの全ての手法で省略語の推定が可能な 96 文を選び、各々に対して推定させる深層格を決定した。省略語の推定が可能と判断した基準は、いずれも文中に名詞と動詞を含んでおり、それぞれ名詞連想概念辞書の刺激語名詞と動詞連想概念辞書の刺激語動詞であり、且つ文中の述語動詞に係る名詞をその 1 つの格要素とする格フレームを持つ述語動詞のタイプが、推定内容に対応する格フレームを必ず有するという条件を全て満たしていることとした。また、推定する内容はベースラインの結果と比較できるようにするために表 4.2 の格フレーム

辞書の格と対応している「対象」、「始点」、「終点」、「場所」、「手段」とした。日本語では、一般的に「動作主」を含む主格の省略が目的格のそれと比べて多いが、ブログでは「私」や「自分」などの書き手自身を表す語が多く省略されていたため、「動作主」の推定を行う必要が無いと判断した。以上のように、ブログから該当する文を抽出する条件によって、課題文と推定内容の組み合わせは 126 通りとなった。これらの課題文と推定内容の組み合わせに対して、5 人の実験参加者に課題文から省略語を推測してもらい、答えた人数が過半数の 3 人以上だった単語を評価用の正解と定めた。全部で 126 通りの組み合わせの内、3 人以上が答えた省略語が存在しなかった課題文と推定内容の組み合わせは 10 通りだったため、これらを除外した結果、最終的に 116 通りの組み合わせになった。

### 4.3.3 評価尺度

本研究で提案した省略語推定システム (VNACD) の推定結果を 3 つのベースライン (CF, NACD, VACD) の結果と比較して評価するため、平均逆順位 (Mean Reciprocal Rank; MRR) と N 位正解率、そして平均精度の平均 (Mean Average Precision; MAP) を用いた。(4.4) 式で表される MRR は、評価用の正解単語とシステムが推定した順に省略語を比較して、最初に合致した省略語の推定順位  $r_k$  の逆数 (Reciprocal Rank:  $1/r_k$ ) の平均であり、 $M$  は課題文の総数を指す。ここでは上位に正解の省略語がどれくらい推定されているかを確認するために VNACD を含む 4 種類の結果に対して上位 1 位、上位 5 位、上位 10 位の各 MRR を計算した。MRR と同様に、N 位正解率も上位 1 位、上位 5 位、上位 10 位について求めた。これは、課題文に対して各順位までに正解の省略語が 1 語以上存在する頻度を表している。

$$MRR = \frac{1}{M} \sum_{k=1}^M \frac{1}{r_k}. \quad (4.4)$$

(4.5) 式で表される平均精度の平均 (Mean Average Precision; MAP) は、全課題文に対して (4.6) 式で得られる各課題文の平均精度 (Average Precision; AP) の平均



である．課題文  $l$  における平均精度  $AP_l$  は， $l$  における正解の単語数  $R_l$  に対する上位  $q$  位で正解の単語を含む時の精度  $P_q$  に対して，上位 1 位から  $R_l$  位までの  $P_q$  の平均を表している．MAP は全課題文数  $M$  における  $AP_l$  の平均のため ( $l = 1, \dots, M$ )，本手法と各ベースラインの MAP を求めて比較することでそれぞれの精度を検証した．

$$MAP = \frac{1}{M} \sum_{l=1}^M AP_l, \quad (4.5)$$

$$AP_l = \frac{1}{R_l} \sum_{q=1}^{R_l} P_q. \quad (4.6)$$

#### 4.3.4 結果と考察

VNACD とベースラインの各 MRR を表 4.3 で表し，それぞれの統計的有意差の有無は符号検定を使用して検討した．表 4.3 より，上位 1 位，上位 5 位，上位 10 位のどの場合においても VNACD の MRR が最も高い値であり，その次に VACD が高い値を示している．これらの MRR の値から VNACD と VACD は正解となる省略語を平均して上位で推定していることが分かる．また，VNACD は VACD の間では上位 1 位の MRR のみ統計的な有意差を確認し ( $p < 0.05$ )，他の 2 つのベースラインとの間ではどの場合も統計的な有意差を確認することができた ( $p < 0.01$ )．一方で，CF と NACD は比較的低い順位まで正解の省略語を推定することができなかったことが各 MRR の値から分かる．これらの関係から，動詞連想概念辞書を用いている VNACD と VACD は動詞連想概念辞書を用いていない他の 2 つのベースラインと比べて，より高い順位で正解となる省略語候補を効果的に抽出できており，特に VNACD は VACD よりも平均して高い順位で正解となる省略語を推定できていることが，MRR の値から判断できる．

VNACD と各ベースラインの  $N$  位正解率の結果は，表 4.4 で表している．どれも 1 位正解率より 10 位正解率の値の方が高く，VNACD と VACD は上位 5 位までに正解の単語を少なくとも 1 語は含んでいる確率が 7 割を超えているため，比較的高いことが確認できる．しかし，その一方で CF と NACD のベースラインではどの順位の正解率も比較的低い値となっており，上位 10 位に対しては 5 割にも達していない．そ

表 4.3: 提案手法と各ベースラインの平均逆順位 (MRR)

	CF	NACD	VACD	VNACD
MRR(top1)	0.155	0.086	0.440** ++	0.534** ++ †
MRR(top5)	0.226	0.157	0.554** ++	0.611** ++
MRR(top10)	0.241+	0.164	0.570** ++	0.617** ++

\*, +, † は, それぞれ CF, NACD, VACD に対する統計的な有意を示す. (\* + † $p$ <0.05, \*\* ++ † $p$ <0.01)

表 4.4: 提案手法と各ベースラインの N 位正解率

	CF	NACD	VACD	VNACD
正解率 (top1)	0.155	0.086	0.440** ++	0.534** ++ †
正解率 (top5)	0.353	0.284	0.741** ++	0.733** ++
正解率 (top10)	0.474+	0.345	0.853** ++ ‡	0.741** ++

\*, +, †, ‡ は, それぞれ CF, NACD, VACD, VNACD に対する統計的な有意を示す. (\* + † $p$ <0.05, \*\* ++ † $p$ <0.01)

のため, MRR と N 位正解率の両方の結果から, VNACD と VACD に比べて CF と NACD は省略語を効果的に抽出できておらず, これは動詞連想概念辞書の有無による差が影響していると考えられる. そのため, 省略語の推定という点において動詞連想概念辞書の情報が一定の妥当性を有していることが示唆される. また, VNACD と VACD の関係に注目すると, 1 位正解率は表 4.4 の MRR の上位 1 位と同値であるため, VNACD が VACD よりも高かったが ( $p < 0.05$ ), その逆に 10 位正解率は VACD の方が高い ( $p < 0.05$ ). これは省略語候補を抽出して確信度を計算する際に, VNACD では名詞連想概念辞書を使用したことによる影響を表している.

図 4.2 は VNACD と VACD における正解の省略語の最高順位をグラフに表したものである. 図 4.2 より, VNACD は VACD と比べて正解の省略語を 1 位に多く含んでいるが, 2 位から 4 位にかけては VACD が多いことが分かる. さらに表 4.4 より, 両者の 5 位正解率がほぼ等しいことから, VNACD は名詞連想概念辞書を確信度の計算に用いたことで正解の省略語の順位が VACD と比べて上位に変動したと考えられる. 一方で, 6 位から 10 位まで VACD は正解の省略語を少しずつ含んでいるのに

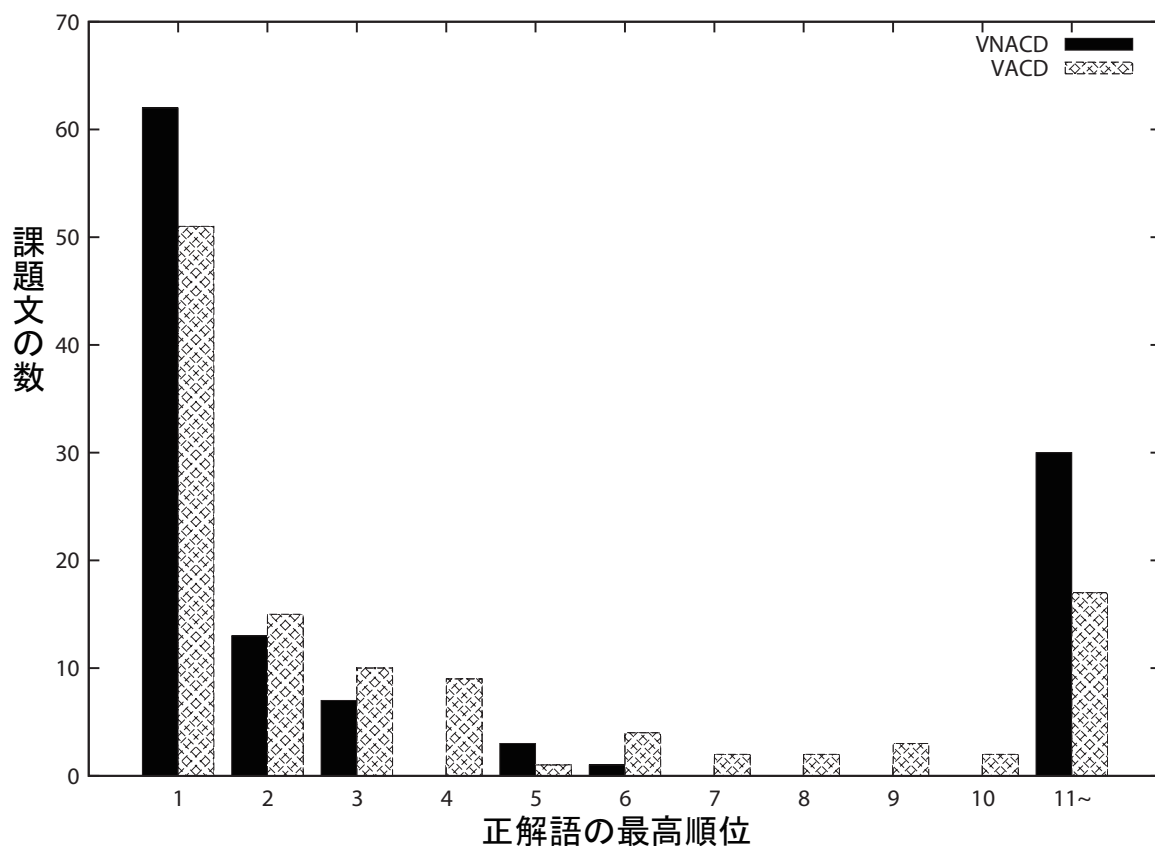


図 4.2: VNACD と VACD の正解語の最高順位のヒストグラム

対して、VNACDには存在しない。その上、11位以降になるとVNACDがVACDよりも多く正解の省略語を含んでいる。これはVNACDでは不正解の省略語が6位から10位の間に順位を上げ、正解の省略語が低い順位で推定されたためである。以上のように、VNACDはVACDに比べて上位、特に1位に正解の省略語がより多く推定されたため、確信度を計算する上で名詞連想概念辞書の一定の有効性を示すことができた。このトレードオフの関係で、6位から10位にかけて省略語の順位が下位へ下がってしまうため、この傾向の軽減については今後取り組む必要がある。

表 4.5 より、VNACD の MAP が各ベースラインの MAP よりも高いことから精度が最も高いことが分かる。そして、MAP や N 位正解率と同様に CF と NACD と比べて VNACD と VACD の MAP の値は十分高く、あらためて動詞連想概念辞書が効果的に利用されていることがいえる。提案手法の VNACD とベースラインの VACD について MAP を比較したところ、動詞連想概念辞書のみを用いている VACD と比べ

表 4.5: 提案手法と各ベースラインの平均精度の平均 (MAP)

	CF	NACD	VACD	VNACD
MAP	0.150	0.092	0.401** ++	0.470** ++ †

\*, +, †, ‡は, それぞれ CF, NACD, VACD, VNACD に対する統計的な有意を示す. (\* + †‡ $p$ <0.05, \*\* ++ †‡‡ $p$ <0.01)

て, 動詞連想概念辞書と名詞連想概念辞書を併用した VNACD は精度が約 7.00% 向上している. 統計的有意差も確認できることから ( $p < 0.05$ ), 文中の省略語の推定において VNACD は VACD よりも統計的には有意に有効な手法であり, 名詞連想概念辞書は省略語の順位付けをする上で必要とした確信度の計算において重要な役割を担っていると考えられる.

本評価の課題文は動詞連想概念辞書と名詞連想概念辞書の各刺激語 (i.e., 小学校の国語の教科書で使用される単語) から成るため, 実験参加者の推測結果が年代によって一般的に大きく変わることは無いと考えられるが, 連想実験の実験参加者と本評価の正解語を推測した実験参加者が同年代であることによって VNACD が最も良い結果になった可能性を考慮する必要もある. そのため, 推定結果において正解語を推測した実験参加者の年代の影響の有無を今後検証していく. その他の課題としては, 先に述べた名詞連想概念辞書を用いることによる正解の省略語の順位が下がることへの対応とともに, 全体的な精度の向上が挙げられる. しかし, 10 位正解率や 11 位以降に正解語が含まれていることを考慮すれば省略語自体の推定はできているが, 確信度が下位のために AP が 0 となり, 結果的に全体の精度である MAP を下げている. そのため, 順位付けに直接関係のある確信度の計算に対して改良を加えることで, システムの更なる精度向上が期待できる.

## 4.4 まとめ

動詞連想概念辞書と名詞連想概念辞書を併用して, 連想情報に基づいた省略語推定システムを構築した. ブログ文書から抽出した評価用の課題文を用いて人間が実

---

際に推測した内容を正解データと定め、格フレーム辞書、既存の名詞連想概念辞書、そして動詞連想概念辞書をそれぞれ単独で用いたベースラインと比較することで、提案手法の評価を行った。その結果、動詞連想概念辞書の妥当性を示すとともに、人間の推測結果により近い省略語の推定を行うことに関して提案手法の有効性を示した。そして、動詞連想概念辞書は省略語候補の抽出において、名詞連想概念辞書は省略語の順位付けにおいて、それぞれ重要な役割を果たしていることが分かった。



## 第5章

# 単語間連想情報に基づいた換喩表現の 検出

第5章では，動詞連想概念辞書の応用研究の2つ目として，比喩表現の一種である換喩の検出について説明する．ここでは，動詞連想概念辞書と名詞連想概念辞書の他に日本語 WordNet の synset を用いることで，単語間の連想情報に基づいた換喩の自動検出を可能にしている．以下では，換喩表現と提案手法の詳細を説明した後，評価実験を通して提案手法の有用性について言及する．





## 5.1 はじめに

換喩（メトニミー）は比喩の一種であるが、喩えるもの（喩詞）と喩えられるもの（被喩詞）が類似性の関係で表される直喩や隠喩とは異なり、隣接性や近接性に基づく関係で表される。換喩を表す関係性は多種に渡るが、一般的に表 5.1 と表 5.2 のように空間的隣接性 [45, 95] と時間的隣接性 [8, 76, 95] に大別することができる [82]。例えば、表 5.1 の「一升瓶を飲み干す」は「容器と中身」の関係で表される換喩表現を含む文であるが、リテラルな（字義通りの）解釈ならば「一升瓶」を丸ごと飲み込んでしまうことになる。これに対して、換喩表現を考慮した一般的な解釈としては、「一升瓶の中に入っている酒」を飲むことになる。これは換喩の機能により、「酒」が空間的な隣接関係のある容器の「一升瓶」に置き換えられているためである。このように換喩表現は、あるものがそれに関係した別のもので表されており、リテラルな解釈と意味的な差異が大きいものにも関わらず、文法的には正しい。そのため、コンピュータが換喩表現を見分けるには形態素解析や構文解析で得られる表層的な情報だけでなく、深層的ともいべき意味的な情報を正確に処理できなければならない。例えば、「マドンナはあの赤シャツを嫌っている」[95] という文に関する処理を例として挙げると、後文に代名詞が含まれている場合、人を指している「赤シャツ」が換喩表現だと分からなければ、その代名詞と「赤シャツ」が照応していることを判断できない。照応解析の他にも、機械翻訳や質問応答システムでは換喩表現かリテラルかの解釈を誤ると全く異なった出力をする可能性がある。よって、換喩解析は自然言語処理の分野で重要な問題である [33, 40]。

換喩解析に関する研究は、主に文中における換喩表現の検出とその言い換えについて行われてきた。前者は文中の名詞に対して換喩表現かリテラルなのかを判断する処理であり、それに対して後者は換喩表現が何を表しているかを同定する処理である。例えば、先で取り上げた「一升瓶を飲み干す」について、前者は「一升瓶」がリテラルではなく換喩表現であると判断すること、後者は「一升瓶」が「酒」を指していると同定することである。

これまでに日本語の換喩解析に関しては、新聞コーパスから収集した名詞の用例と共起関係を統計的なデータとして利用した研究 [56, 93] や、名詞を中心とした連想

表 5.1: 空間的隣接性に基づいた換喩表現の関係性

空間的隣接性	例文 (換喩表現が意味する内容)
容器 - 中身	一升瓶 (→酒) を飲み干す
原料 - 製品	アルコール (→酒) を飲む
手段 - 主体	白バイ (→警官) が逮捕する
作者 - 作品	マーラー (→音楽) を聴く
付属物 - 主体	詰め襟 (→学生) が歩く

表 5.2: 時間的隣接性に基づいた換喩表現の関係性

時間的隣接性	例文 (換喩表現が意味する内容)
	杯を傾ける (→お酒を飲む)
原因 - 結果	筆をとる (→書きはじめる)
	赤面する (→恥ずかしく思う)

の情報を利用した研究 [79] があり、換喩表現の言い換えを中心に議論されている。また、村田らと須賀らは、人手で作成した格フレーム辞書と日本語語彙大系をそれぞれ用いて換喩表現の検出も行っている [56, 79]。いずれも文中の用言に関する構文情報とその対になる単語の情報を使用した検出法を提案している。一方、英語の換喩に関しては、人手で作成された知識やルール、意味ネットワークなどを使用する手法 [11, 14, 30, 50] や、統計的手法を用いたコーパスベースのアプローチなどがある [51]。また、構文情報や意味情報、そして共起情報などを特徴量とし、機械学習による検出の手法も提案されている [59, 60, 68]。中でも、Nastase らは Wikipedia から抽出した概念ネットワークを用いて、換喩表現の可能性のある語に対し確率的な選好性を計算して利用することで、精度の高い換喩理解を実現している [58]。

換喩は隣接性や近接性に基づく比喩の一種であり、言語的・認知的な制約により規制されるものもあるが、一般的には心理的な連想のリンクと密接に関わっているとされる [95, 96]。このような連想の関係性が主張されてきたのにも関わらず、日本

語および英語における換喩表現の検出において、連想の情報がこれまで利用されてきていない。そのため、構文情報や意味情報、共起情報などを用いた既存の手法では、精度が不十分な可能性がある。そこで、本研究では、人間が単語に対して連想した情報に着目し、文中の名詞や動詞の連想関係を用いて、表 5.1 のような空間的隣接性に基づいた換喩表現を自動検出する手法を提案する。また、連想情報の代わりに共起情報を用いた手法も作成し、換喩表現の検出結果を比較することで連想情報と共起情報の相違を明らかにする。なお、時間的隣接性の換喩表現は、事象の時間的な前後関係や因果関係によって表され、文脈中の使われ方で換喩表現かリテラルかが決まるため、複雑なものである。ゆえに、本研究では、換喩表現の判断がより簡易で明確である空間的隣接性で特徴づけられるもののみを対象とし、連想情報を利用して高い性能を実現することを目指す。

## 5.2 提案手法

### 5.2.1 方針

「一升瓶を飲む」のように、用言「飲む」に対して係り受け関係のある名詞「一升瓶」が連想されにくい場合、その名詞が、換喩表現として隣接性の関係にある別の語を表している可能性がある。そのため、本研究では、文中の名詞に対して、その名詞を含む文節（名詞文節）ごとに格助詞の情報を用いて、その係り先の用言との連想の関係性を調べ、この関係性が強いほどリテラルであり、逆に弱いほど換喩表現と判断する手法を提案する。本手法では、処理対象を上述の条件を満たす名詞、格助詞、用言の3つ組として、換喩表現かリテラルの判断を行う。具体的には、「一升瓶を飲む」の例で言えば、まず始めに用言（e.g., 「飲む」）とそれに対して連想関係にある語（e.g., 水, ジュース, 酒など）について意味情報を取得し、その意味情報を、処理対象の名詞（e.g., 一升瓶）の意味情報と比較し、共通する概念を特定する。そして、連想関係にある語と当該名詞から、これらの共通する概念までの距離を用いて、連想関係の有無、および、強さを得る。両者の間で共通する概念が無い場合は、処理対象の名詞と用言の間に連想関係が無く、これらの間で意味的差異

が生じているとみなす。すなわち、隣接関係にある別の語を、換喩表現としてこの名詞で置き換えている可能性があると考え。さらに、共通する概念がある場合でも、距離が遠いと連想の関係性が弱いと捉えることができ、隣接関係にある別の語を処理対象の名詞で置き換えているとみなす。一方、用言と連想関係にある語と処理対象の名詞における概念間の距離が近い場合は、当該名詞は用言に対して連想され易い関係であり、意味的差異は生じておらず、換喩の可能性が低いと考える。

なお、本手法では、上記の考え方に則って換喩表現の検出を行うが、概念間距離の近いか遠いかという程度の判断について人手で基準を与えるとアドホックな手法になってしまう可能性があることから、本研究では、データから機械学習の方法により判断基準を獲得するアプローチを取る。学習手法としては、決定木学習を用いる。学習手法として SVM [94] などもあるが、換喩表現の検出を行う上で本研究は最初のステップの位置づけであること、そして学習結果が人間にとって分析し易いことなどの理由から今回は決定木学習を用いることとする [83, 84]。

### 5.2.2 手法の詳細

本手法では、動詞連想概念辞書と名詞連想概念辞書から、刺激語と連想語、そしてそれらの連想距離を用いて、処理対象の名詞との共通の意味情報を調べ、概念間の距離を求める。この概念間の距離を求めるために利用する意味情報は、日本語 WordNet [29]、あるいは、日本語語彙大系 [27] から得る。連想概念辞書と日本語 WordNet を組み合わせて用いる手法を ACD-JWN<sup>1</sup>、連想概念辞書と日本語語彙大系を組み合わせて用いる手法を ACD-GT<sup>2</sup>と呼ぶことにする。両手法とも、用言と連想関係にある語と、処理対象の名詞との概念間距離を特徴量として抽出し、決定木学習を行う処理は同一である。以下では、ACD-JWN における本手法の概要を図 5.1 で表し、特徴量の抽出の過程を (1) から (5) のステップに分けて説明する。なお、図 5.1 では、「データの流れ」を表す矢印の横に「データの値」を記し、菱形が「データの処理」

---

<sup>1</sup>ACD は連想概念辞書 (Associative Concept Dictionary) のことであり、JWN は日本語 WordNet のことである。

<sup>2</sup>GT は日本語語彙大系のことである。

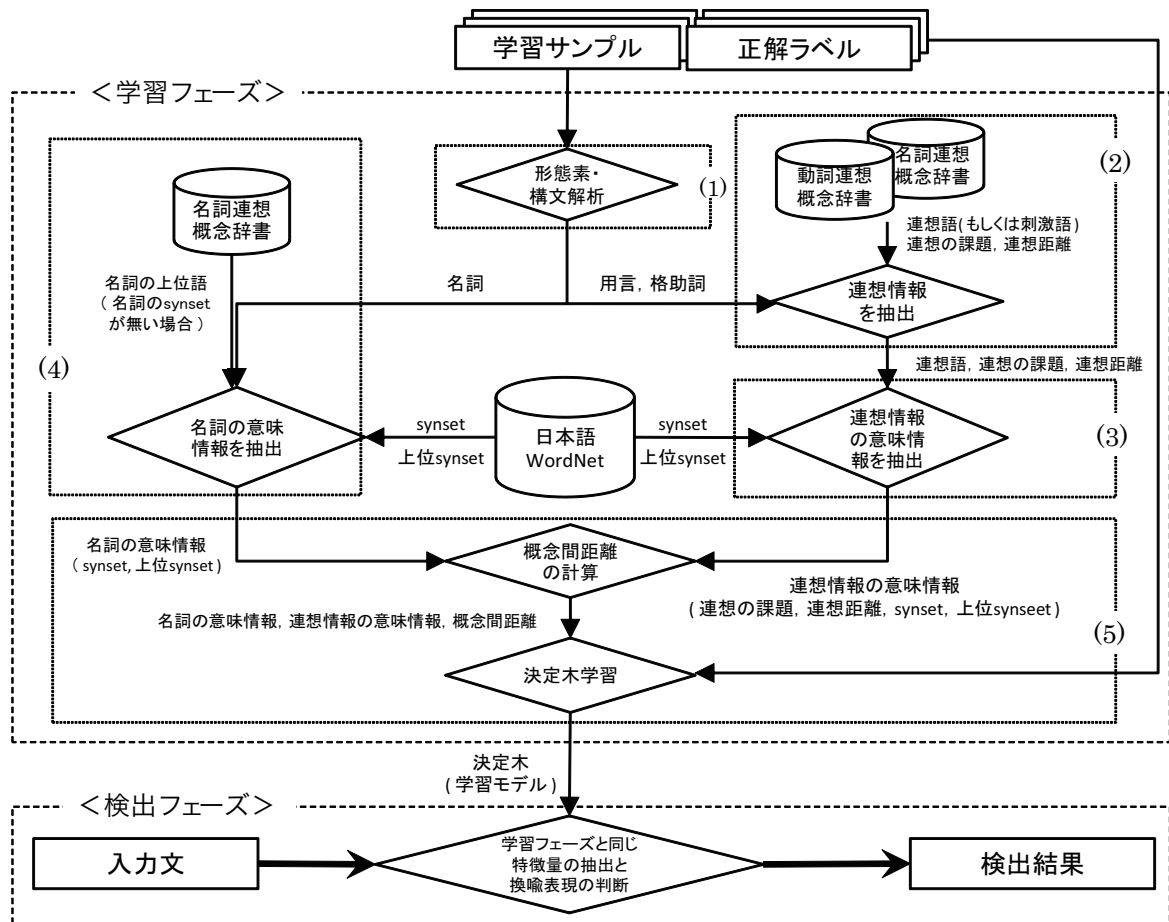


図 5.1: 本手法 (ACD-JWN) の概要. ただし, 図中の (1)~(5) は 5.2.2 項で説明する特徴量の抽出の過程を指す.

を表している。また、ACD-GT については日本語 WordNet の synset の代わりに日本語語彙大系の意味属性を用いるところが異なるのみであるため、説明は割愛する。また、本研究で扱う文は、格助詞を含む名詞文節とその係り先の用言の組み合わせを 1 組以上含んでいることとし、複数組がある場合はそれぞれの組み合わせについて個別に処理を行う。ただし、名詞文節に、(1) の形態素解析で得られる「名詞」が複数含まれる場合、つまり複合名詞の場合は、末尾の名詞のみを用言と係り受けの関係のある名詞として処理を行う。ここで、複数の名詞文節と用言を含む文「道で拾った財布を交番に届けた」の場合に処理する対象を例として挙げると、名詞文節ごとに係り先の用言と名詞との組み合わせに着目するため、名詞「道」と用言「拾った」、ならびに名詞「財布」と用言「届けた」、そして名詞「交番」と「届けた」に対して、それぞれ (2) から (5) の処理を行うことになる<sup>3</sup>。なお、用言文節「拾った」と係り先の名詞文節「財布を」のような用言による連体修飾の関係は、本手法の処理の対象とはしない。

**(1) 文中の係り受け関係の取得** 学習データのサンプル文に対して形態素解析<sup>4</sup>と構文解析<sup>5</sup>を行い、名詞文節とその係り先の用言を取得する。

**(2) 用言の連想情報の抽出** 形態素解析と構文解析の結果から、名詞文節の格助詞とその係り先の用言の情報を用いて、その用言と連想関係にある語を抽出する。具体的には、用言が動詞連想概念辞書の刺激語の場合は、その用言に対して、助詞と対応させた連想の課題 (e.g., 助詞「～が」は「動作主」、助詞「～を」は「対象」など) で連想距離が短い連想語を抽出する。用言が動詞ではなく形容詞の場合は、名詞連想概念辞書の逆引きによって、連想語から刺激語を抽出する。すなわち、その形容詞が表 2.4 で例示する属性概念の連想語である時、連想距離が短い刺激語 (名詞) を抽出する。ここで、抽出する語を 1 語のみに限定すると用言に対する連想情報の多義を許容しないことになるため、最低限の多義に対応することを目的に、用言と助詞に対して連想距離が最も短い語と

---

<sup>3</sup> 「道で」は「拾った」に係り、「届けた」には係らない場合。

<sup>4</sup> MeCab 0.98pre3 を使用。

<sup>5</sup> CaboCha 0.60pre4 を使用。

次に短い語の上位2語を連想情報として得る。なお、2.1.5項と3章のように連想語は表3.3の「引越し業者」のような複合名詞や、人名・地名などの固有名詞となる場合がある。しかし、ここで抽出する連想距離が短い上位2語について言えば、多くの実験参加者にとって連想し易い基本的な語であり、複合名詞や固有名詞が含まれることはほとんどない。また、各連想概念辞書の刺激語は国語の教科書で用いられる名詞と動詞（基本形）のため、逆引きにより得られる刺激語も、基本的な語である。ゆえに、連想情報として抽出する連想語や刺激語は、(1)の形態素解析で単一の形態素として出力される名詞とほぼ単位は変わらないと考えられるため、(1)の形態素の単位における名詞として処理を行う。

**(3) 連想関係にある語の意味情報** 連想情報として抽出した語の意味情報として、日本語 WordNet の synset を用いる。synset とは、同義語の集合であり、WordNet 中では1つの概念に1つの synset が対応している。例えば、「瓶」と同じ synset の語として「ボトル」や「ビン」などが挙げられる。ここでは、用言と連想関係にある語について synset と上位 synset を調べる。なお、あまりに抽象的な概念を指す synset を避けるために、本手法では上位 synset を調べる範囲として語の synset から上位6階層までとした。また、対応する synset が日本語 WordNet に無く、連想関係にある語の意味情報を得ることができない場合は、その語は連想されないということを表していると捉え、意味情報無しとして処理を行う。

**(4) 名詞の意味情報** (3)で抽出した連想関係にある語の意味情報との比較のため、処理対象の名詞に対しても日本語 WordNet の synset とその上位 synset を調べて、名詞の意味情報として抽出する。また、「ニクソン」や「フォード」などの固有名詞が存在する場合は、形態素・構文解析結果でそれぞれ得られる「人名」や「組織」に対応する synset と上位 synset を抽出する。ここで、処理対象の名詞が日本語 WordNet の中に無い場合は、この名詞と関係のある意味情報なるべく得られるように、フォールバックの機構として、名詞連想概念辞書か

ら上位概念の連想語, あるいは連想語の逆引きによる下位概念の刺激語, すなわち上位語にあたる語を抽出し, その上位語の synset とその synset の上位 synset を抽出する. 連想の有無が意味を持つ (3) とは異なり, (4) では処理対象の名詞が示す意味情報を特定することが目的であるため, その名詞が日本語 WordNet に無い場合, 意味情報無しとはせずに, フォールバックを行い, なるべく意味情報を得るようにしている.

**(5) 換喩表現の判断** (2) から (4) の過程で取得したデータを決定木学習の特徴量としてまとめる. 具体的には, 以下の特徴量 F1 から F10 までを抽出する.

**F1, F2, F3:** (2) の連想の課題, 連想距離の上位 2 つの値を順番に特徴量 F1, F2 (連想距離 1), F3 (連想距離 2) とする. F1 は, 連想の課題が動詞連想概念辞書の「動作主」の時は Agent, 「対象」の時は Object, 「始点」の時は Source, 「終点」の時は Goal, 「場所」の時は Location, 「手段」の時は Tool とする. また, 名詞連想概念辞書を用いた時は, NACD とする.

**F4, F5:** (3) で取得した連想関係にある語の synset と上位 synset は, 連想情報の概念の範囲を表しているため, それぞれの異なり数を特徴量 F4 と F5 として求める.

**F6, F7, F8, F9:** (4) で取得した処理対象の名詞の synset と上位 synset の異なり数をそれぞれ特徴量 F6 と F7 とし, 当該名詞が日本語 WordNet に無い場合に名詞連想概念辞書から得る上位語についても synset とその上位 synset の異なり数を特徴量 F8 と F9 として求める.

**F10:** 用言と連想関係にある語と処理対象の名詞のそれぞれの synset から, 共通する synset までのノード間距離の総和を特徴量 F10 (概念間の距離) として抽出する. F10 は, レンジが広いため, 離散的特徴量とする. 具体的には, 表 5.4 のように 6 段階に定めた. None は共通する synset が無い場合を指しており, Near は 2 種類のノード数の平均が 1 未満の場合を, Middle-Near は 1 の場合をそれぞれ表している. さらに, Middle と



表 5.3: ACD-JWN で決定木学習に用いた特徴量. ただし, ※は処理対象の名詞に対応する synset が存在しない場合を指す.

特徴量	説明
F1 連想の課題	処理対象の格助詞に対応させた連想の課題
F2 連想距離 1	用言と助詞に対して連想関係のある語で最も短い連想距離の値
F3 連想距離 2	用言と助詞に対して連想関係のある語で 2 番目に短い連想距離の値
F4 synset 数	連想関係のある語の synset の異なり数
F5 上位 synset 数	連想関係のある語の synset に対する上位 synset の異なり数
F6 名詞の synset 数	処理対象の名詞の synset の異なり数
F7 名詞の上位 synset 数	処理対象の名詞の synset に対する上位 synset の異なり数
F8 上位語の synset 数	処理対象の名詞の上位語の synset の異なり数 (※)
F9 上位語の上位 synset 数	処理対象の名詞の上位語の synset に対する上位 synset の異なり数 (※)
F10 概念間の距離	連想関係のある語と処理対象の名詞の synset から共通する synset までの最小ノード数

Middle-Far はノード数の平均が 1 より大きく 2 以下の場合と 2 より大きく 3 未満の場合をそれぞれ表し, 3 以上の場合は Far と表す.

ここで, 「一升瓶を飲む」を例として具体的な特徴量を挙げる. まず, 名詞文節「一升瓶を」の格助詞「を」に表 3.2 の連想の課題「対象」が対応しているため, F1 は “Object” となる. そして, 用言「飲む」を刺激語と連想の課題「対象」に対して, 連想関係にある語として連想距離が最も短い連想語「水 (1.85)」と 2 番目に短い連想語「ジュース (2.31)」を動詞連想概念辞書から得るため, F2

表 5.4: ACD-JWN で決定木学習に用いた特徴量の値

特徴量	値
F1	Agent, Object, Source, Goal, Location, Tool, NACD のどれか
F2 - F7	数値
F8 - F9	数値 (※処理対象の名詞に対応する synset が存在する場合は 0)
F10	None, Near, Middle-Near, Middle, Middle-Far, Far のどれか

と F3 はそれぞれ “1.85” と “2.31” となる。次に、日本語 WordNet で「水」と「ジュース」を参照し、それらの synset ならびに上位 6 階層までの上位 synset を取得する。これらの synset の異なり数がそれぞれ “6” と “25” のため、これらの値が F4 と F5 となる。一方、「一升瓶」に対しては日本語 WordNet に synset が存在しないため F6 と F7 はともに “0” となる。また、フォールバックの機構により名詞連想概念辞書から表 2.4 の下位概念の連想語として「一升瓶」を持つ刺激語すなわち「一升瓶」の上位語として「瓶」を得る。この時、日本語 WordNet から「瓶」の synset ならびに上位 6 階層までの上位 synset を取得し、F8 と F9 は上位語の synset とその上位 synset の異なり数である “6” と “13” となる。ここまでで取得した「水」と「ジュース」の上位 synset と、「瓶」の上位 synset の間で、共通する synset があり、概念間距離は “6” のため、F10 は “Far” となる。そして最後に、上記の例のように得られる F1 から F10 までの特徴量と正解ラベルを対応させて決定木学習を行い、決定木を得る。なお、学習には C5.0<sup>6</sup>を用いる。未知の入力があれば、得られた決定木を用いて入力文の処理対象となる名詞に対して換喩表現かリテラルかを判別する。

ステップ (1)～(5) で抽出した ACD-JWN の特徴量をまとめると表 5.3 と表 5.4 のようになる。提案手法 ACD-GT は、表 5.3 の synset を日本語語彙大系の意味属性に置き換えたものが特徴量となる。

<sup>6</sup><http://www.rulequest.com/>

## 5.3 評価実験

評価実験用に4つのベースライン (GT-GT, GT-JWN, CF-GT, CF-JWN) を作成し, 提案手法 (ACD-JWN, ACD-GT) と比較した. 決定木学習のための学習データは書籍や先行研究, 新聞コーパスより抽出して作成し, 評価は Leave-One-Out Cross-Validation で行った. 以下では, 学習データ, ベースラインの概要, 決定木学習の特徴量, そして実験結果を述べる.

### 5.3.1 換喩表現

換喩は, 隠喩と共に古くから代表的な修辞表現の1つとして注目を集めてきた. 例えば, 「村は寝静まっている」という表現は, 実際に「寝静まる」のは「(村の) 人」であるが, 「村」という単語で間接的に表すことにより, 読み手の想像を一層掻き立て, 含みを豊にしていると考えられる. しかし, 近年の認知意味論では, 換喩の意味を成立させる人間の認知構造は, 上記のように考えられてきた文章表現だけに収まるものではなく, 一般的な人間の認知メカニズムの反映として, 日常言語に広く分布されていることが分かってきた [47, 82]. Langacker は, 換喩はアクセスしやすいものを参照点 (reference-point) として経由させることで目標 (target) に注意・意識を向けるという, 人間の認知的営み (参照点構造 (reference-point construction)) の現れであると指摘している [46]. 例えば, 「どんぶりを食べる」は, 「どんぶり」という目立った対象を参照点として指示することで, 注意・意識を向けた「ご飯」へのアクセスを容易にしている, という考え方である. このような「どんぶり」 (i.e., 参照点) と, 「ご飯」 (i.e., 目標) との間で, ずれが大きければ, 読み手に深い解釈を求める表現になり, ずれが小さければ, 日常言語の字義的な表現と変わらないような表現になると考えられる.

これまでに換喩に対しては, 前者のような特定の含みを持った修辞表現に限るもの, または後者のような人間の認知メカニズムを反映した日常言語に浸透したものとして広く捉えるものなど, 様々な見解がされてきた [22, 82, 95]. しかし, コンピュータの換喩理解で特に問題になっているのは, 「電話を取る」という文を字義通りに解

表 5.5: 学習データ 1 のサンプル文の一部

換喩表現を含む文	リテラルな文
<u>一升瓶</u> を飲む	酒を飲む
<u>白バイ</u> が違反者を逮捕する	警察官が違反者を逮捕する
<u>顔</u> を剃る	髭を剃る
<u>頭</u> を刈る	髪を刈る
<u>ニクソン</u> はハノイを爆撃した	アメリカ政府はハノイを爆撃した
<u>やかん</u> が沸騰する	水が沸騰する
<u>井戸</u> を汲む	水を汲む
指揮者はその <u>クラリネット</u> を笑った	指揮者はその演奏者を笑った

(下線部は換喩表現を表している.)

積すると、本来の「受話器を取る」のではなく、「電話本体をまるごと取る」ことになってしまうような後者の換喩表現である。ゆえに、本研究の評価実験で扱う換喩表現を「ある対象を参照するのに、隣接性・近接性の関係がある別の対象を用いる言語表現」と広く定義する。

### 5.3.2 学習データ

決定木学習のための学習データについてであるが、2種類のデータセットを設定した。1つは、換喩表現とリテラルな文の出典元が同じデータセット（以降、学習データ1と呼ぶ）で、もう1つは換喩表現とリテラルな文で使用されている単語が同じデータセット（以降、学習データ2と呼ぶ）であり、表5.5と表5.6はそれぞれのサンプル文の一部を示している。両方とも、換喩表現を含むサンプル文は同じであり、主に従来の研究 [56, 79] や書籍 [76, 82, 95] で用いられている文に限定した。これは、換喩表現を恣意的に作成して増やすことによる精度の向上を避けるためである。なお、これらは全て単文であり、複文の扱いについては今後の検討課題である。

一方、リテラルな文は、2種類のデータセットの間で全く異なる。学習データ1の

表 5.6: 学習データ 2 のサンプル文の一部

換喩表現を含む文	リテラルな文
<u>一升瓶</u> を飲む	祭壇に一升瓶を置いた
<u>白バイ</u> が違反者を逮捕する	警察官が犯罪者を逮捕した
<u>顔</u> を剃る	髭を剃る
<u>頭</u> を刈る	田んぼで稲を刈る
<u>ニクソン</u> はハノイを爆撃した	工場を爆撃した
<u>やかん</u> が沸騰する	水が沸騰する
<u>井戸</u> を汲む	水を汲む
指揮者はその <u>クラリネット</u> を笑った	自分を笑った

(下線部は換喩表現を表している.)

リテラルな文は、表 5.5 のように、換喩表現を含む文 (e.g., 「一升瓶を飲む」) それぞれについて、「換喩表現が指す内容として出典元で言及あるいは解説されている語 (e.g., 酒)」と換喩表現の語 (e.g., 一升瓶) を置き換えることで用意した (e.g., 「酒を飲む」)<sup>7</sup>。また、学習データ 2 のリテラルな文は、筆者らが毎日新聞コーパスから抜粋し、換喩表現を含む文と同じような長さの単文を作成した。表 5.6 に示すように、換喩表現を含む文と同程度の文節数であり、且つ、同じ名詞や動詞を含んでいるものを抜粋した。各データセットとも、換喩表現を含む文<sup>8</sup>とリテラルな文は、各 45 文で合計 90 文となっている。

表 5.5 と表 5.6 の換喩表現を含む文にて、下線部が換喩表現に該当するが、本手法は名詞文節ごとに名詞と係り先の用言との関係性を調べるため、入力文 90 文に対して換喩表現かリテラルかの判別部分は、学習データ 1 が 108 箇所、学習データ 2 が

<sup>7</sup>山梨が『比喩と理解』の p.103 で言及している「換喩的に、“理論”、“仮説”などの抽象名詞と“主張する”、“説明する”などの、本来的には人間名詞を要求する動詞との共起関係を保証している」[95] に該当する換喩表現に関しては、“人”に置き換えた。

<sup>8</sup>「仮説が…を説明する」や「理論が…を主張する」は、本来ならば換喩表現と必ずしも見なすことができないが、山梨の『比喩と理解』で用いられていることから、本研究では換喩表現として学習データに加えた。

113箇所ある。その内、換喩表現はそれぞれ 45箇所である。なお、村田ら [56] は山梨の書籍 [95] で取り上げられている換喩表現を含む 23 文の内 17 文に対して実験結果を分析した。これは恣意的に換喩表現を含む文を増やして検出精度を向上させることを避けるため、書籍で取り上げられている文のみで評価を行っていたからである。本研究では、参考にする書籍や先行研究をさらに増やしており、村田らの評価データの規模を考慮すると本研究のデータ量は大きい。従来 [56, 79] は換喩表現のみを評価データにしていたため、ある文が換喩表現を含むかどうかを検出するアルゴリズムを検証することはできなかったが、本研究では 2 種類のデータセットを用意し、両方とも換喩表現を含む文と同数のリテラルな文も揃えることで、換喩表現かリテラルかを判別する評価も可能なデータとなっている。

### 5.3.3 先行研究及び共起情報に基づいたベースライン

ベースラインとして、先行研究 [56] を参考にした 2 つの手法 (GT-GT, GT-JWN) と、その他に提案手法 (ACD-GT, ACD-JWN) で用いた連想情報の代わりに共起情報を用いた 2 つの手法 (CF-GT, CF-JWN) を作成した。なお、いずれのベースラインも、提案手法と同様、用言文節に係る名詞文節における名詞、格助詞と用言文節の用言を処理対象とする。

GT-GT と GT-JWN は、村田らの手法を自動化した手法である。先行研究で、村田らは人手で作成した格フレーム辞書を用いて、文節内の換喩表現を次のように検出していた。入力文「僕が漱石を読む」に対して、まず「読む」の格フレーム“{人}が{本, 新聞, 小説}を[読む]”に照合させる。そして、ガ格の「僕」は“人”で格フレームの意味属性に属すが、ヲ格の「漱石」は“本”, “新聞”, “小説”に属さないため、「僕」はリテラルで「漱石」は換喩表現と判断する。この手法は、述語と助詞の情報から得られる構文情報と意味情報を使用する手法と言える。本評価実験では、このような村田らの手法を日本語語彙大系から構文情報と意味情報を用いて自動化し、それを GT-GT とした。GT-JWN は、GT-GT と同様のアプローチだが、日本語語彙大系の代わりに日本語 WordNet を組み合わせた手法である。

次に、CF-GT と CF-JWN について概略を説明する。これらのベースラインは本

表 5.7: ベースライン GT-GT における決定木学習の特徴量

特徴量	説明	値
格助詞	処理対象の格助詞	ga, ha, wo, de, ni のどれか
意味属性数	構文情報の結合価パターンの 格に入る意味属性数	数値
名詞の意味属性数	処理対象の名詞の意味属性数	数値
名詞の上位意味属性数	処理対象の名詞の意味属性に 対する上位の意味属性数	数値
概念間の距離	結合価パターンの名詞と処理 対象の名詞から共通する意味 属性までの最小ノード数	None, Near, Middle-Near, Middle, Middle-Far, Far のどれか

手法で用いる連想情報の妥当性を調べるために用意されたものであり、提案手法における連想情報の代わりに共起情報を用いる手法である。ここで用いる共起情報は、Web 上の約 16 億文から自動構築された格フレーム辞書 [36, 44] から抽出されたものである。この格フレーム辞書は述語とその格フレーム情報から成り立っており、述語に対して「ガ格」や「ヲ格」などの格に入る語（格要素）とその頻度が記述されている。述語に関しては、動詞や形容詞を区別する述語タイプや、格交替を起こす付属語が付いている態タイプなどで詳細に区別されているが、ここでは述語が基本形で表される情報のみを扱って共起頻度を求めることとした。そして、ある格に対して全ての格要素の総頻度で割った値を、述語と格要素の共起頻度として求める。この共起頻度によって処理対象の名詞と用言の「共起し易さ」が分かるため、本手法で「連想し易さ」を示す連想距離の代わりに用いることができる。そのため、共起頻度が最も高い語と 2 番目に高い語を抽出して、共起情報として使用する。

以下、各ベースラインの処理の流れを説明するとともに、それぞれが用いる特徴量を表 5.7, 表 5.8, 表 5.9, 表 5.10 で示す。

**Baseline1: 先行研究の汎用化手法 (GT-GT)** 入力文に対する形態素解析と構文解

表 5.8: ベースライン GT-JWN における決定木学習の特徴量

特徴量	説明	値
格助詞	処理対象の格助詞	ga, ha, wo, de, ni のどれか
synset 数	結合価パターンの格に入る意味属性 を置き換えた synset の数	数値
名詞の synset 数	処理対象の名詞の synset 数	数値
概念間の距離	結合価パターンの名詞と処理対象 の名詞の synset から共通の synset までの最小ノード数	None, Near, Middle-Near, Middle, Middle-Far, Far のどれか

析の結果から、名詞文節の係り先の用言について、日本語語彙大系の構文体系から結合価パターンを抽出する。例えば、「読む」の結合価パターンの1つとして「N1がN2を読む」を抽出する。ここで、N1とN2には、それぞれ“主体”と“抽象物（精神）、精神”の意味属性が対応する。このように結合価パターンは用言と格要素の意味的な共起関係を表しており、その格要素は名詞の意味属性を指している。この結合価パターンは、日英の翻訳の際の使用頻度に基づいて優先順位が決まっている [77]。GT-GT は、入力文と同じ格を持つ結合価パターンが複数ある場合、最も優先順位が高いパターンを日常の構文情報で使用頻度が高いものとして選ぶ。次に、結合価パターンの格要素に対応する意味属性を得た後、処理対象の名詞の意味属性と比較を行い、共通する意味属性までの階層数を求める。この過程で表 5.7 の特徴量を抽出し、決定木学習で得た決定木を用いて、処理対象の名詞について換喩表現カリテラルかを判断する。

**Baseline2: 日本語 WordNet を用いた手法 (GT-JWN)** GT-JWN は、構文情報で得た名詞の意味属性の代わりに、その意味属性名に対応する全ての synset を用いることで、日本語 WordNet の意味情報を利用する。その他は、GT-GT と同様である。表 5.8 の特徴量を抽出し、決定木学習により得られた決定木を用いて、処理対象の名詞に対して換喩表現カリテラルかを判断する。



表 5.9: ベースライン CF-GT における決定木学習の特徴量

特徴量	説明	値
格助詞	処理対象の格助詞	ga, ha, wo, de, ni のどれか
共起頻度 1	用言と助詞に対して共起頻度が最も高い値	数値
共起頻度 2	用言と助詞に対して共起頻度が 2 番目に高い値	数値
F4 - F10	ACD-GT の F4-F10 と同じ	ACD-GT と同じ値

**Baseline3: 共起情報を用いた手法 (CF-GT)** CF-GT は、連想情報の代わりに共起情報を用いる点以外は、提案手法 ACD-GT と同様の手法である。すなわち、格フレーム辞書から処理対象の用言と格助詞に対して最も共起頻度が高い格要素と、その次に高い格要素をそれぞれ抽出して、処理対象の名詞との概念間距離を日本語語彙大系の意味属性を用いて求める。この過程で表 5.9 の特徴量を抽出し、決定木学習により得られた決定木を用いて、換喩表現かリテラルかを判断する。なお、共起情報を利用する先行研究として Nastase の手法があるが、この手法は、Wikipedia から概念ネットワークを作成し、換喩表現の検出に用いている。この概念ネットワークは、ある動詞に対して主語や目的語に何が当てはまりやすいかを共起情報を用いてスコア化しているため、日本語の格フレーム辞書と同等の情報を持つと考えられる。そのため、本ベースライン CF-GT、および後述の CF-JWN は、Nastase らの手法と同等の特性を有しているものである。

**Baseline4: 共起情報を用いた手法 (CF-JWN)** CF-JWN は、連想情報の代わりに共起情報を用いる点以外は、提案手法 ACD-GT と同様の手法である。CF-GT の時と同様に格フレーム辞書から抽出した格要素に対して、処理対象の名詞との概念間距離を日本語 WordNet の synset を用いて求める。同様に、表 5.10 の

表 5.10: ベースライン CF-JWN における決定木学習の特徴量

特徴量	説明	値
格助詞	処理対象の格助詞	ga, ha, wo, de, ni のどれか
共起頻度 1	用言と助詞に対して共起頻度が最も高い値	数値
共起頻度 2	用言と助詞に対して共起頻度が 2 番目に高い値	数値
F4 - F10	ACD-JWN の F4-F10 と同じ	ACD-JWN と同じ値

特徴量を抽出し，決定木学習により得られた決定木を用いて，換喩表現かリテラルかを判断する．

### 5.3.4 結果

実験結果をまとめると以下の 5 つの表となった．表 5.11 は，学習データ 1 と学習データ 2 において，提案手法（ACD-GT, ACD-JWN）とベースライン（GT-GT, GT-JWN, CF-GT, CF-JWN）が処理対象の名詞を換喩表現かリテラルな表現かを正しく判別したシステム全体の精度を示している．表 5.12 と表 5.13 は，それぞれの学習データにおいて換喩表現の検出に関する適合率と再現率，ならびに，これらの調和平均 F 値をまとめたものである．また，表 5.14 と表 5.15 は，同様に，リテラルの検出に関する結果をまとめたものである．さらに，分母の数が同じで比較可能な場合は，統計的検定を行っている．表 5.11 のように，提案手法 ACD-JWN が各ベースラインよりも高い精度（0.83 と 0.84）を示しており，全てのベースライン（GT-GT, GT-JWN, CF-GT, CF-JWN）に対して統計的な有意差（ $p < 0.01$  或いは  $p < 0.05$ ）を確認することができた．なお，統計的有意差の検出には McNemar 検定を用いた．

学習データ 1 に対する結果に注目すると，表 5.12 と表 5.14 より，ACD-JWN が，

表 5.11: 学習データ 1 と学習データ 2 に対するシステム全体の精度

手法	学習データ 1	学習データ 2
GT-GT	0.65 (70/108)	0.71 (80/113)†
GT-JWN	0.63 (68/108)	0.67 (76/113)
CF-GT	0.61 (66/108)	0.59 (67/113)
CF-JWN	0.64 (69/108)	0.61 (69/113)
ACD-GT	0.73 (79/108)†	0.81 (92/113)*+†††
ACD-JWN	<b>0.83</b> (90/108)**++††††¶	<b>0.84</b> (95/113)*++††††

\*, +, †, ‡, ¶は、ベースライン GT-GT, GT-JWN, CF-GT, CF-JWN, 提案手法 ACD-GT それぞれに対する統計的な有意差を示す. (\* + ††¶p<0.05, \*\* ++ ††††p<0.01)

換喩表現の検出とリテラルな表現の検出の両方で F 値 (0.79 と 0.86) が各ベースラインよりも高く、且つ、適合率 (0.85 と 0.83) と再現率 (0.73 と 0.90) も各ベースラインのそれぞれの値よりも高かった。さらに、換喩表現の検出の再現率 (0.73) は、GT-GT と CF-JWN に対して統計的な有意差があり、リテラルな表現の検出の再現率 (0.90) は全てのベースラインに対して統計的な有意差があった。

また、学習データ 2 に対する結果に関しては、表 5.13 と表 5.15 より、学習データ 1 の時と同様に、ACD-JWN が換喩表現の検出とリテラルな表現の検出の両方で F 値 (0.79 と 0.87) が各ベースラインよりも高く、且つ、適合率 (0.85 と 0.84) と再現率 (0.73 と 0.91) も各ベースラインの値より高かった。さらに、換喩表現の検出の再現率 (0.73) は、全てのベースラインに対して統計的な有意差があり、リテラルな表現の検出は GT-GT 以外の 3 つのベースラインに対して統計的な有意差が確認できた。

以上から、2 種類の学習データに対して、提案手法 ACD-JWN が、全てのベースラインよりも換喩表現の検出とリテラルな表現の検出において、ともに高い F 値と精度を示し、且つ、両方の表現の検出を同時に評価できる全体の精度に関しては全てのベースラインに対して性能の差は有意であることから、提案手法がベースラインよりも有効であることが確認できた。

表 5.12: 学習データ 1 に対する換喩表現の検出精度

手法	適合率	再現率	F 値
GT-GT	0.59 (22/37)	0.49 (22/45)	0.54
GT-JWN	0.55 (26/47)	0.58 (26/45)	0.57
CF-GT	0.53 (25/47)	0.56 (25/45)	0.54
CF-JWN	0.57 (24/42)	0.53 (24/45)	0.55
ACD-GT	0.71 (27/38)	0.60 (27/45)	0.65
ACD-JWN	<b>0.85</b> (33/39)	<b>0.73</b> (33/45)* ‡	<b>0.79</b>

\*, ‡ は, ベースライン GT-GT, CF-JWN それぞれに対する統計的な有意差を示す. (\* ‡p<0.05)

表 5.13: 学習データ 2 に対する換喩表現の検出精度

手法	適合率	再現率	F 値
GT-GT	0.70 (21/30)	0.47 (21/45)	0.56
GT-JWN	0.60 (24/40)	0.53 (24/45)	0.56
CF-GT	0.49 (21/43)	0.47 (21/45)	0.48
CF-JWN	0.51 (21/40)	0.47 (21/45)	0.49
ACD-GT	0.80 (32/40)	0.71 (32/45)** †	0.75
ACD-JWN	<b>0.85</b> (33/39)	<b>0.73</b> (33/45)**++ ††††	<b>0.79</b>

\*, +, †, ‡ は, ベースライン GT-GT, GT-JWN, CF-GT, CF-JWN それぞれに対する統計的な有意差を示す. (†p<0.05, \*\* ++ ††††p<0.01)

表 5.14: 学習データ 1 に対するリテラルな表現の検出精度

手法	適合率	再現率	F 値
GT-GT	0.68 (48/71)	0.76 (48/63)	0.72
GT-JWN	0.69 (42/61)	0.67 (42/63)	0.68
CF-GT	0.67 (41/61)	0.65 (41/63)	0.66
CF-JWN	0.68 (45/66)	0.71 (45/63)	0.70
ACD-GT	0.74 (52/70)	0.83 (52/63)+†	0.78
ACD-JWN	<b>0.83</b> (57/69)	<b>0.90</b> (57/63)*++††‡‡	<b>0.86</b>

\*, +, †, ‡ は, ベースライン GT-GT, GT-JWN, CF-GT, CF-JWN それぞれに対する統計的な有意差を示す. (\* + † $p$ <0.05, ++ ††‡‡ $p$ <0.01)

表 5.15: 学習データ 2 に対するリテラルな表現の検出精度

手法	適合率	再現率	F 値
GT-GT	0.71 (59/83)	0.87 (59/68)††‡	0.78
GT-JWN	0.71 (52/73)	0.76 (52/68)	0.74
CF-GT	0.66 (46/70)	0.68 (46/68)	0.67
CF-JWN	0.67 (48/72)	0.71 (48/68)	0.69
ACD-GT	0.82 (60/73)	0.89 (60/68)††‡	0.85
ACD-JWN	<b>0.84</b> (62/74)	<b>0.91</b> (62/68)+†††‡‡	<b>0.87</b>

+, †, ‡ は, ベースライン GT-JWN, CF-GT, CF-JWN それぞれに対する統計的な有意差を示す. (+ †† $p$ <0.05, †††‡‡ $p$ <0.01)

## 5.4 考察

提案手法 ACD-JWN が高い精度を示した理由としては、用言に対して連想関係にある語と、処理対象の名詞それぞれから、共通の synset までの概念間距離が特に有効であったことが挙げられる。図 5.2 と図 5.3 は、学習データ 1 と学習データ 2 において、それぞれの全入力文を学習データとして用いたときの学習結果（決定木）である。図 5.2 より、学習データ 1 の場合、「名詞の上位 synset 数」が一定数よりも多い場合に、「概念間の距離」によって換喩表現かリテラルかを判断する条件が細かく分岐しているのが分かる。この概念間距離が、None の時は Metonymy すなわち換喩表現と判断し、反対に Near と Middle-Far の時はリテラルと判断している。また、Middle-Near, Middle の時は、それぞれ「名詞の synset 数」、「連想距離 1」の条件で Metonymy あるいは Literal を判断し、Far の時はさらに多くの特徴量を用いて判断している。また、図 5.3 より、学習データ 2 の場合、「概念間の距離」が None と Far の時は Metonymy すなわち換喩表現と判断し、反対に、Near と Middle の時は Literal すなわちリテラルと判断していることが分かる。また、Middle-Near や Middle-Far の時は、「連想距離 2」と「名詞の上位 synset 数」の条件で Metonymy あるいは Literal を判断している。このようなことから、特徴量の中で概念間距離が重要であることがいえる。よって、概念間距離が換喩表現の判断に必要な連想の関係性の強弱を示唆していると考えられ、これが ACD-JWN が優っていた理由の 1 つと考えられる。

換喩表現の検出において、日本語語彙大系あるいは連想概念辞書を用いることで生じた結果の差異を次に考察する。表 5.5 と表 5.6 の「顔を剃る」や「庭を掃く」を例にして説明すると、日本語語彙大系では「剃る」と「掃く」の結合価パターン of ヲ格に換喩表現の「顔」と「庭」が意味属性に含まれていたことから、GT-GT と GT-JWN ではリテラルとして誤って判断された。これに対して、連想概念辞書では「剃る」と「掃く」に対して、それぞれ「髭」や「ゴミ」が抽出されていたため、「顔」と「庭」それぞれから共通の意味属性あるいは synset までの概念間距離を算出することで換喩表現として正しく判断されていた。さらに、表 5.5 と表 5.6 の「指揮者はそのクラリネットを笑った」の文においては、日本語語彙大系では「N1 が N2 を笑う」という結合価パターンにおいて N1 と N2 の意味属性はそれぞれ“人”と“\*”で

```

名詞の上位 synset 数 <= 5: Metonymy
名詞の上位 synset 数 > 5:
: ... 概念間の距離 = None: Metonymy
  概念間の距離 in {Near,Middle-Far}: Literal
  概念間の距離 = Middle-Near:
: ... 名詞の synset 数 <= 11: Literal
:   名詞の synset 数 > 11: Metonymy
  概念間の距離 = Middle:
: ... 連想距離 1 <= 1.54: Metonymy
:   連想距離 1 > 1.54: Literal
  概念間の距離 = Far:
: ... 連想距離 2 > 5.33: Literal
  連想距離 2 <= 5.33:
: ... 上位語の上位 synset 数 > 28: Metonymy
  上位語の上位 synset 数 <= 28:
: ... 上位語の synset 数 > 8: Literal
  上位語の synset 数 <= 8:
: ... 連想距離 1 > 2.97: Metonymy
  連想距離 1 <= 2.97:
: ... 連想の課題 in {Source,Goal,Location,Tool,
                    NACD}: Metonymy
  連想の課題 = Object: Metonymy
  連想の課題 = Agent:
: ... 上位語の上位 synset 数 <= 21: Literal
  上位語の上位 synset 数 > 21: Metonymy

```

図 5.2: 学習データ 1 の全入力文に対する ACD-JWN の決定木学習の結果

```

概念間の距離 in {None,Far}: Metonymy
概念間の距離 in {Near,Middle}: Literal
概念間の距離 = Middle-Near:
... 連想距離 2 <= 2.74: Metonymy
: 連想距離 2 > 2.74: Literal
概念間の距離 = Middle-Far:
... 名詞の上位 synset 数 <= 19: Literal
名詞の上位 synset 数 > 19: Metonymy
    
```

図 5.3: 学習データ 2 の全入力文に対する ACD-JWN の決定木学習の結果

あった。「指揮者」は“人”に属しているためリテラルだと正しく判断されたが、“\*”は全ての意味属性を指していることから「クラリネット」はリテラルだと誤って判断された。これに対して、連想概念辞書では「笑う」から「人」と「TV 番組」がそれぞれ抽出されていた。概念間距離を算出した結果、「指揮者」はリテラルで「クラリネット」は換喩表現と正しく判断された。そもそも「指揮者」が「クラリネット」という楽器自体を笑うことは意味的には間違いではないが、山梨の解釈によれば「指揮者」が「クラリネット」の演奏者を笑う [95]、すなわち「クラリネット」は換喩表現として一般的には捉えられやすい。そのため、連想概念辞書を用いると我々人間の感覚に近い換喩表現の判断ができることに対して、日本語語彙大系を用いるとなるべくリテラルな表現と判断する、すなわち人間の感覚からずれる傾向になることが分かった。

次に、共起情報と連想情報を用いることで生じた結果の差異について考察する。CF-GT と CF-JWN では格フレーム辞書から抽出した用言と格要素の共起情報を用いており、用言によっては提案手法で用いた連想情報と同じ語を利用している場合があった。例えば、「剃る」と「食べる」のヲ格に関して「髭」と「ご飯」がそれぞれ共起頻度が最も高い語として抽出されており、連想情報と同じであったため、「顔を剃る」と「鍋を食べる」というような換喩表現を含む文に関して提案手法と同様



に正しい判断を行っていた。しかし、表 5.12 と表 5.13 から分かるように、実際には誤った判断を行う場合が提案手法よりも多く、その主な理由として共起頻度の高い単語が、換喩表現の検出には不適切となった場合が多かったことが挙げられる。例えば、「煮える」のガ格に対しては「鍋」という換喩表現の語や、「開ける」のヲ格に対しては「幕」という慣用表現の語、そして「走る」のガ格に対しては「痛み」という抽象的な語などを共起情報として利用したために、これらの用言と格助詞を含んだ文は、いずれも換喩表現あるいはリテラルかの判断を誤る結果となった。これは、リソースである Web 文書で用言に対してその格要素が、いかに多く使用されていたかという共起情報の性質を反映しており、CF-GT と CF-JWN は連想情報に含まれる基本的な語をはじめ、換喩表現や慣用表現、そして抽象的な語など幅広い情報を用いたことが、提案手法よりも精度が低かった直接的な原因だと考えられる。ゆえに、共起情報に含まれる語と処理対象の名詞の概念間距離から、換喩表現かリテラルな表現かの判断を行うことが難しいため、本手法においては共起情報よりも連想情報を用いる方が有用であることが分かった。

図 5.4 は、学習データ 1 に対する提案手法 ACD-JWN とベースライン (GT-JWN, CF-GT) における学習曲線を表しており、図 5.5 は学習データ 2 に対する提案手法 ACD-JWN とベースライン (GT-JWN, CF-JWN) における学習曲線を表している。これらの学習曲線は、評価実験で使用した 90 文から成る 108 事例を 28 事例 (テストデータ) と 80 事例 (学習データ) に、同じく 113 事例を 33 事例 (テストデータ) と 80 事例 (学習データ) に、それぞれ無作為に分け、学習データ 80 事例に対して 10 から 80 まで 5 事例の単位で学習データの量を増やし、その都度テストデータについて換喩表現かリテラルの判断を行うという試行を 100 回繰り返すことで求めた。図 5.4 ならびに図 5.5 の縦軸の「全体の精度」は、「学習に用いたサンプル文の数」を 5 事例ずつ増やした各学習データにおいてテストデータに対するシステム全体の精度の平均値を指している。なお、エラーバーは、95%信頼区間を表している。これらの学習曲線から、学習に用いたサンプル文の数が増えるにつれてテストデータに対する精度が向上しており、且つ、ともに 60 事例以降は精度向上の鈍化していることが分かる。ゆえに、本研究で用いた学習データの量は十分ではないかもしれない

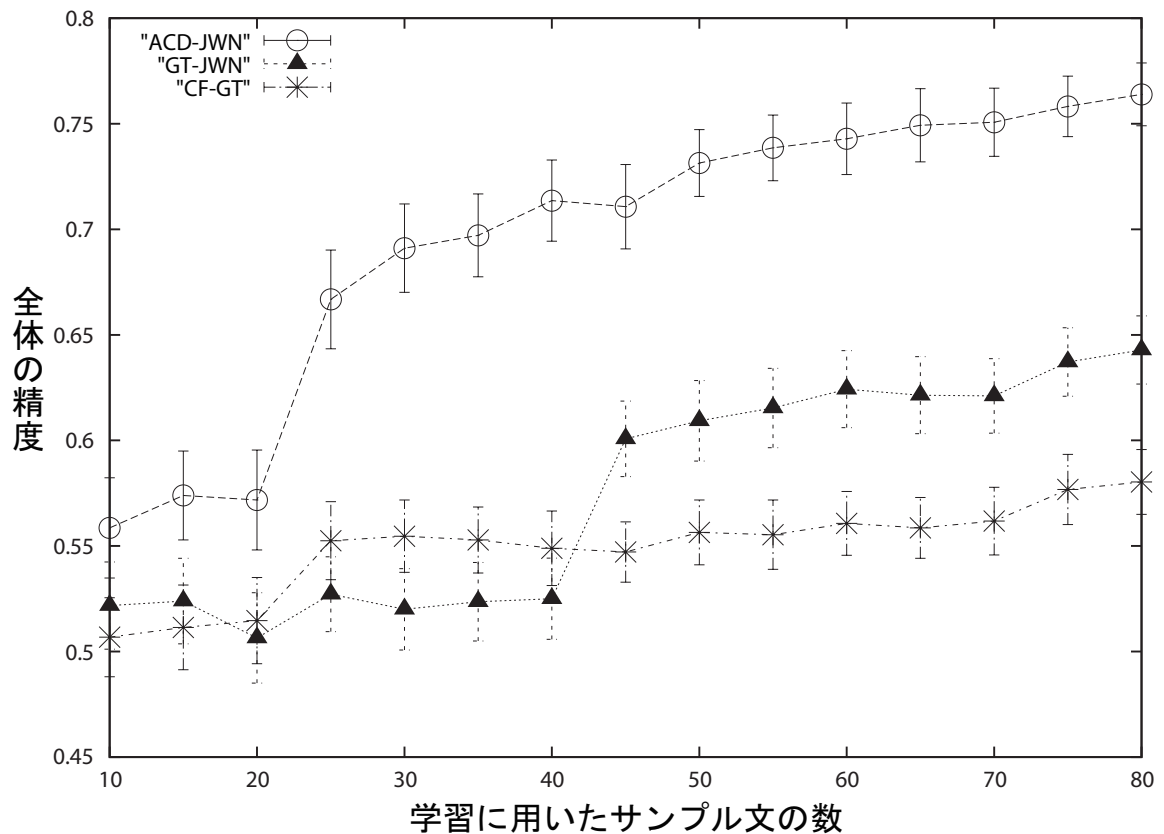


図 5.4: 学習データ 1 に対する提案手法 ACD-JWN とベースライン GT-JWN, CF-JWN における学習曲線

が、目的にかなった学習をしていると考えられる。さらに、各々の学習曲線に注目すると、学習データ 1 と学習データ 2 が 20 事例まではどれも同等の精度であったのが、25 事例以降はベースラインと比べて ACD-JWN の精度が大きく向上していることが確認できる。これは、ACD-JWN では、学習データが増えるにつれて適切に学習ができるような特徴量を用いていることを示唆している。

提案手法の ACD-GT と ACD-JWN で換喩検出ができなかった事例が生じた原因について言及すると、連想概念辞書から抽出した連想情報の中に入力文中の換喩表現が含まれており、概念間距離が短い場合には誤った判別を行っていたことが挙げられる。これは、連想概念辞書が連想実験のデータより構築されているためであり、過去に実験参加者が動詞に対して換喩表現に該当する単語を連想したことに起因する。

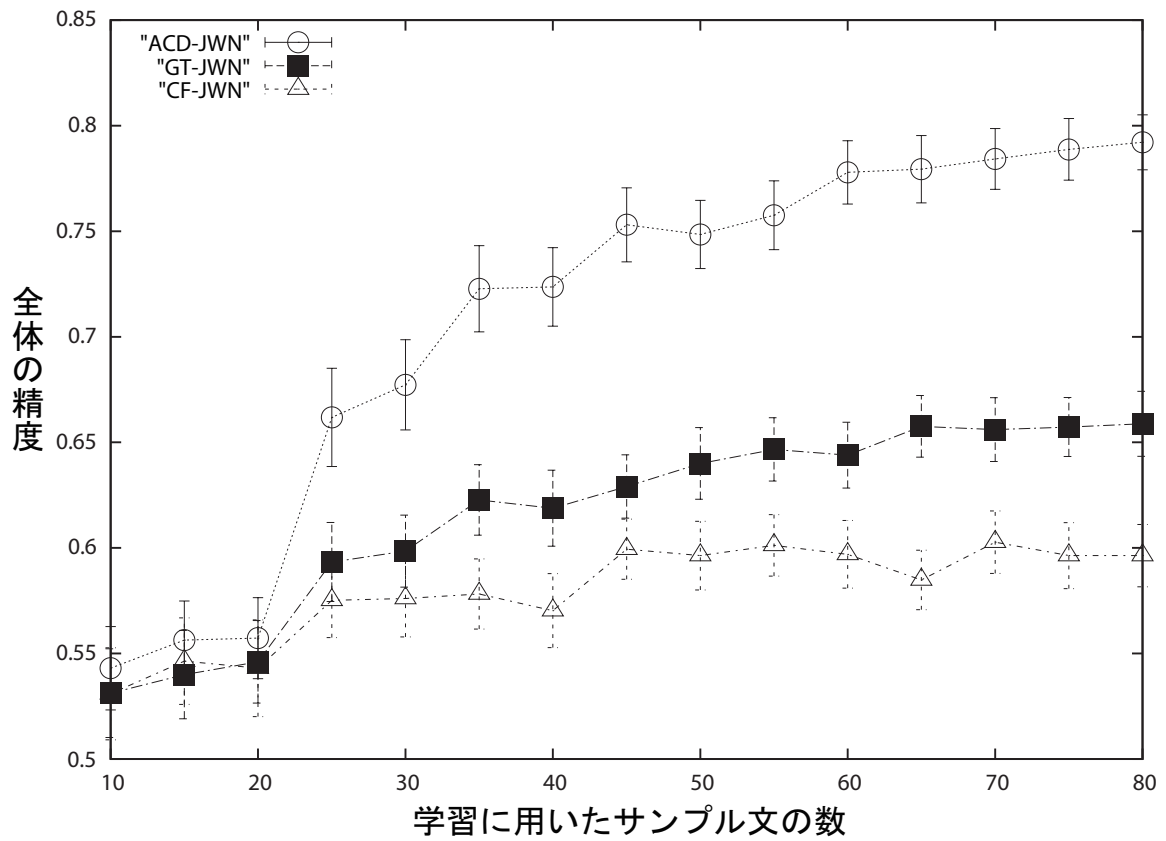


図 5.5: 学習データ 2 に対する提案手法 ACD-JWN とベースライン GT-JWN, CF-JWN における学習曲線

このように、連想情報が換喩表現を含み、短い距離の場合は一様に換喩表現をリテラルだと判断してしまうことが問題ともいえる。この問題で ACD-GT と ACD-JWN は誤った判別になる場合が共通しているが、ACD-JWN の方が換喩表現ならびにリテラルの検出の F 値が高く、特に、学習データ 1 の時の全体の精度は ACD-GT に対して統計的な有意差 ( $p < 0.05$ ) がある。これは、本手法で用いた日本語 WordNet の synset の粒度が日本語語彙大系の意味属性と比べて高いためであり、このことから処理対象の名詞と連想情報に関する概念間距離を求めるには synset が適していると考えられる。

最後に、提案手法で用いた連想情報は、動詞連想概念辞書の規模と質に大きく左右されると考えられる。そこで、提案手法がベースラインの手法よりも有効に動作

表 5.16: 連想実験の参加者数と ACD-JWN の精度

手法	実験参加者数	学習データ 1	学習データ 2
GT-GT	—————	0.65 (70/108)	0.71 (80/113)
ACD-JWN	1 刺激語 10 人	0.69 (74/108)	0.69 (78/113)
	1 刺激語 20 人	0.77 (83/108)	0.76 (86/113)
	1 刺激語 30 人	0.72 (78/108)	0.71 (80/113)
	1 刺激語 40 人	<b>0.83</b> (90/108)**	<b>0.84</b> (95/113)*

GT-GT は精度が最も高いベースライン。以下の\*は、ベースライン GT-GT に対する統計的な有意差を示す。(\*  $p < 0.05$ , \*\*  $p < 0.01$ )

する条件を確認するために、提案手法で用いた 1 語の刺激語につき 40 人の実験参加者データの他に、10 人、20 人、30 人の実験参加者の連想実験データをサンプリングした後に、それぞれを用いた場合の ACD-JWN の精度を算出し、最も精度が高いベースライン GT-GT と比較した。学習データ 1 と学習データ 2 に対して精度を求めた結果、表 5.16 のように、1 語あたり 40 人の実験参加者のデータを使用した時に統計的な有意差とともにベースラインよりも有効であることが分かった。このことから、動詞連想概念辞書が 1 語あたり 40 人の実験参加者の規模であれば、本手法が有効に動作することが分かった。

## 5.5 まとめ

本研究は、入力文中の名詞が換喩表現かリテラルかを連想の関係性に基づいて判断する手法を提案した。日本語の換喩表現の検出における従来の手法ならびに共起情報を用いた手法をベースラインとして評価実験を行ったところ、提案手法の有効性を確認することができた。本研究の貢献は、先行研究で従来用いられてきた共起情報の代わりに連想情報を用いることの有効性を示し、日本語の換喩表現の検出において最も高い精度を実現したことである。今後は、換喩表現の言い換えに取り組むとともに、テキスト中に換喩表現がどの程度含まれており、それらをどの程度適

切に処理できるかという、言語処理における換喩処理のインパクトについて調べることも検討していきたい。



## 第6章

# 結論

第6章では、本研究の総括を述べる。まず、第3章、第4章、第5章で論じた本研究の成果を確認した上で、総合的な考察を行う。そして、本研究の貢献と限界に触れつつ、今後取り組むべき課題について言及し、本論文の結びとする。





## 6.1 本研究の成果

本研究の内容は、動詞連想概念辞書について「構築」と「応用」に大別できる。以下で、各々について成果をまとめる。

### 6.1.1 動詞連想概念辞書の構築

刺激語を基本的な動詞として、深層格情報を抽出する連想実験を行い、収集した実験データを用いて動詞連想概念辞書を構築した。動詞連想概念辞書は、刺激語と連想語、そして連想距離から成り立っている。刺激語と連想語は単語間距離である連想距離によって関係づけられており、この連想距離は刺激語に対して連想語の連想頻度の逆数と連想順位の線形式で表される。ここでは、実験データのパラメータを目的関数と制約条件に利用した線形計画法の最小化問題に対して、シンプレックス法によって得られた最適解を各々の係数としている。

動詞連想概念辞書の性質を評価するために、既存辞書で最もデータ構造が似ている格フレーム辞書と比較を行った。格フレーム辞書は、Web上の文書から自動構築されていることから、Web上の共起情報と見なすことができるため、これらの辞書の比較は、連想情報と共起情報を比較することになる。そこで、動詞連想概念辞書の連想情報と格フレーム辞書の共起情報を比較分析した結果、ある動詞に対して連想され易いことと共起し易いということが必ずしも一致しないことが分かった。また、反意語関係の動詞については、連想情報では2つの動詞の間における単語の対称性を確認できたことに対して、共起情報ではそのような傾向が見られなかった。これらのことから、連想情報と共起情報は性質が異なっていること、且つ、動詞連想概念辞書は連想情報から成り立っていることを確認することができた。

### 6.1.2 動詞連想概念辞書の応用

動詞連想概念辞書の応用研究としては、省略語の推定と換喩表現の検出を行った。省略語の推定に関しては、動詞連想概念辞書と既存の名詞連想概念辞書を併用して省略語の確信度を計算し、降順で省略語を出力するシステムを構築した。そして、格

フレーム辞書を用いた手法、名詞連想概念辞書のみを用いた手法、動詞連想概念辞書のみを用いた手法をそれぞれベースラインとして設定して評価実験を行った。その結果、提案手法が人間に最も近い省略語の推定が可能になっていることを確認したとともに、動詞連想概念辞書が省略語候補の抽出に、名詞連想概念辞書が確信度による省略語の順位付けにおいて、それぞれ重要な役割を果たしていることを示した。

換喩表現の検出に関しては、動詞連想概念辞書（入力文によっては名詞連想概念辞書）と日本語 WordNet を組み合わせることで、換喩表現の自動検出システムを構築した。ここでは、入力文の述語に対して連想され易い単語と、その述語と係り受けの関係のある文節内の単語の概念間距離を計算して求め、決定木学習で得た学習モデル（決定木）を用いて、換喩表現かリテラルな表現かの判断を行う。この決定木学習は、連想距離をはじめ、連想語や文節内の名詞の日本語 WordNet における synset 数や上位 synset 数の他に、概念間距離の値などを特徴量として抽出して利用した。そして、先行研究の手法を日本語 WordNet と日本語語彙大系をそれぞれ用いて自動化した 2 つの手法の他に、提案手法の動詞連想概念辞書の代わりに格フレーム辞書を用いた 2 つの手法の計 4 つベースラインを設定し、評価実験を行った。その結果、提案手法が最も高い精度であったことを確認したとともに、連想情報の方が共起情報よりも換喩表現の検出に適している理由を明らかにすることができた。

以上のように、省略語の推定と換喩表現の検出を行う意味理解システムに応用して、既存の手法よりも高い精度を実現していることを確認することで、動詞連想概念辞書の有用性を示した。

## 6.2 本研究の課題と展望

本研究の課題として、第一に動詞連想概念辞書の一般公開について述べる。6.1.1 項に記したように、動詞連想概念辞書は、基礎日本語辞典の項目で取り上げられている動詞の他に関連語として取り上げられている動詞も網羅しており、基礎的な動詞に関する連想情報がまとめられた辞書と捉えることができる。そのため、今後は連想語の表記ゆれの修正を行い、基礎動詞編として一般公開することを目指す

もに、動詞連想概念辞書の汎用性をより高くするために基礎的な動詞以外についても実験データを取得して、段階的な拡張を目指したい。

次に、動詞連想概念辞書の拡張性について述べる。表 3.8 からも分かるように、動詞連想概念辞書の拡張性は決して高いとはいえない。これは、連想実験による人間の連想情報の収集に時間的且つ人数的なコストがかかるからである。Web 上の文書や新聞コーパスなどから収集したデータに対して、これまで得てきた連想語に関する特徴量を設定して機械学習によって分類することで、自動拡張できる可能性も考えられる。しかし、本研究で示したように連想情報と共起情報は異なる性質を持っているため、このような大規模な共起情報を含むコーパスと機械学習を利用した自動拡張は、これまでの連想情報とは異なる情報を収集する可能性があり、不適切な面があるといえる。ゆえに、今後も動詞連想概念辞書を拡張するためには、コストはかかってしまうが、従来と同様に実験ベースでデータを収集することが求められる。

さらに、動詞連想概念辞書を構築する上で必要な連想実験の参加者数について述べる。表 3.8 のように、動詞連想概念辞書を構築した当初は、刺激語 1 語あたりの実験参加者数は 20 人の規模であった。そして、連想実験の刺激語を増やしつつ、1 語あたりの実験参加者数を 40 人まで増やして動詞連想概念辞書を拡張してきたが、1 刺激語あたりの実験参加者数が 40 人で十分だという実証はできていない。たしかに、表 5.16 のように実験参加者数をサンプリングした動詞連想概念辞書を用いた結果、1 刺激語 40 人の規模の時に換喩表現の検出に関する提案手法が有効に動作することを示したが、これはあくまでも手法の有効性について言及しただけに過ぎない。表 3.8 より、刺激語が 54 語の時に 1 刺激語あたりの実験参加者数を 20 人から 40 人の規模に拡張すると、連想語の延べ数と異なり数がそれぞれ約 2 倍になっていることが分かる。しかし、実験参加者数を 40 人にした状態で刺激語数を段階別に増やした拡張の際は、連想語の延べ数と比べて異なり数が増加する割合は明らかに小さくなっている傾向が分かる。これは、実験参加者が答える連想語の内容が一部共通しているためであり、常識とも言うべき動詞に関する背景知識を表している可能性が考えられる。このような視点で考えると、刺激語数を固定した状態で実験参加者数を段階別に増やし、連想語の異なり数の増える割合を調べることで、常識と見なせ

るような必要な情報を得るためには、どれくらいまで実験参加者数を増やすことが必要なのかを推測することができる可能性がある。ゆえに、ある刺激語の実験データセットに対して段階的に実験参加者数を増やしていき、連想語の異なり数との関係から、十分と見なせる実験参加者数を検討したいと考えている。

最後に、コンピュータの意味理解について述べる。本研究では、動詞連想概念辞書を用いて省略語の推定や換喩表現の検出を行う意味理解システムを構築し、その有用性を示すことができた。しかし、4.3 節の評価文が少なくなってしまったように、省略語の推定に関する提案手法は、動詞連想概念辞書と名詞連想概念辞書のみを用いたためカバレッジが低く、大規模なテキストデータに対する評価が行えないために、実用レベルでの応用が困難である。一方で、換喩表現の検出に関する提案手法は、動詞連想概念辞書（部分的に名詞連想概念辞書）と日本語 WordNet を連携させることで、先述の問題点がある程度解消できている。これらの点を踏まえると、コンピュータの意味理解機能を様々な応用分野で発展させるには、動詞連想概念辞書や名詞連想概念辞書などの連想情報と既存の言語資源を組み合わせた手法が有効であると考えられる。特に、5.4 節で述べたように、動詞の共起情報には比喩表現や慣用表現、抽象的な表現などが共起され易い単語として入り混じっている性質があることを考慮して、連想情報と使い分ける、或いは併用することが求められる。ゆえに、比喩表現や慣用表現の意味理解システムに、動詞連想概念辞書と既存コーパス・シソーラスを併用することで、精度向上が期待できる。

### 6.3 おわりに

本研究を通して、これまで自然言語処理研究で明確に論じられなかった連想と共起について、性質の違いを示唆することができた。自然言語処理システムのタスクに応じて連想と共起の情報を効果的に使い分けること或いは併用することが、コンピュータの意味理解と文脈理解の精度向上に繋がると考えている。また、本研究で構築した動詞連想概念辞書は、自然言語処理を含む工学の分野以外でも、認知言語学や言語心理学の研究において、実験材料や言語教育支援ツールの作成などに利用できる可能性がある。その他にも、言語を扱う研究分野は多岐に渡ることから、動

詞連想概念辞書を教育・研究目的の利用で一般公開して多くの研究者に利用してもらうことで，他分野の研究発展に貢献することを願ってやまない。



# 謝辞

本論文の謝辞を記すことに当たり、まず始めに、主査として研究の御指導を賜りました慶應義塾大学 石崎俊 名誉教授と慶應義塾大学環境情報学部 今井むつみ 教授の両先生に厚く御礼申し上げます。思い起こせば、私が環境情報学部4年生の時に石崎先生の自然言語処理論を履修したことが、大学院へ進学したきっかけでした。石崎先生には、「コンピュータの言語理解能を人間に近付けるためにはどうすれば良いのか」という自然言語処理研究の命題に対して本質的な課題を見据えた上で、研究のアドバイスや問題提起をしていただきました。修士課程・後期博士課程で研究の御指導をしていただき、本当に感謝致しております。また、今井先生には、後期博士課程の初めの頃から石崎先生が御退職される前までは副査として、そして石崎先生が御退職された後は主査として、御指導していただきました。特に、博士フォーマル発表の際に、認知科学や心理学の観点から本研究に対する疑問点を挙げていただき、自然言語処理の工学的な視点にとらわれがちであった私に研究に対する多角的な視点の必要性を教えて下さいました。誠に感謝致しております。

本論文の副査を引き受けていただきました慶應義塾大学環境情報学部 田中茂範 教授と慶應義塾大学環境情報学部 萩野達也 教授の両先生に厚く御礼申し上げます。田中先生には、本論文における連想情報の性質と換喩表現の種類に関しまして、言語学の観点から貴重な御意見をいただきました。そして、萩野先生には、本論文で取り上げました省略語の推定や換喩表現の検出に関する自然言語処理の内容に関しまして、日本語から英語或いは英語から日本語に翻訳する事例を挙げていただくなど、貴重な御意見をいただきました。

本論文に関しまして、個別にアドバイスしていただきました慶應義塾大学環境情報学部 諏訪正樹 教授、慶應義塾大学環境情報学部 西岡啓二 教授、そして慶應義塾大学環境情報学部 青山敦 専任講師にも御礼申し上げます。諏訪先生には、博士

フォーマル発表後に人工知能分野における推論アルゴリズムを引き合いに動詞連想概念辞書のデータ構造に関しまして貴重な御意見をいただき、且つ、公聴会後には本論文で修正すべき箇所を丁寧に教えていただきました。西岡先生には、修士の時からアドバイスしていただき、動詞連想概念辞書の構築やその応用例につきまして貴重な御意見をいただきました。また、青山先生には、大学院プロジェクトにて公聴会に向けた本論文の発表内容に関しまして、丁寧なアドバイスをしていただきました。

本論文を構成する原著論文の共著者であり、石崎研究室の先輩である嘉悦大学ビジネス創造学部 岡本潤 専任講師と NTT メディアインテリジェンス研究所 東中竜一郎 主任研究員のお二人に深く感謝致します。岡本博士には、動詞連想概念辞書を構築する際に有益なアドバイスをしていただいた他に、応用研究のために名詞連想概念辞書の研究用データを提供して下さいました。東中博士には、研究につきまして様々な面で御指導していただきました。同時に、第一線で御活躍されている研究者としての研究・教育活動や、休日の過ごし方などを間近で見せていただき、とても刺激を受けました。特に、Local Organizer として参加させていただきました YRRSDS 2010 にて、稀有な体験を積みさせていただきましたことは、生涯忘れないことでしょう。お二人に、改めて御礼申し上げます。

慶應義塾インフォメーションテクノロジーセンター (ITC) の皆様に感謝致します。ITC 本部助教として働きながら、本論文に関する研究を続けることができましたのも、皆様の御理解がありましたからこそと思っております。ITC 本部 押見淳 事務長と ITC 本部 金子康樹 課長のお二人には、いつも温かい言葉をかけていただき、元気づけて下さいました。また、湘南藤沢 ITC 関本幸輝 事務長、ITC 本部 浅見健次 主任、日吉 ITC 山根健 主任、湘南藤沢 ITC 山方崇 主任、渡辺陽三 氏、越口直美 氏、元湘南藤沢 ITC の藤井清 氏、鈴木悦子 氏、そして前 ITC 本部助教 直江健介 博士をはじめ、ITC の皆様には、研究発表や論文執筆などに関しまして色々な面で御配慮していただきました。改めて御礼申し上げます。

後期博士課程の研究生活を支えてくれた友人達に感謝致します。慶應義塾大学法学部 佐治伸郎 訪問研究員 (学振) には、ロフトの住人として、東京工科大学の非常



勤講師として、大変お世話になりました。佐治博士には、どんなことでも熱く語れる情熱とどこでもいつでも眠れる適当さで、周りの人を和ませることの心地よさを教わりました。そして、修士の時に石崎研究室で一緒だった玉川大学脳科学研究所 研究員 佐村俊和 博士、ジップインフォブリッジ 西潟一生 氏、NTT コミュニケーションズ 野本済央 氏の三人には、研究以外に楽しむことが研究のために必要であることを教わりました。また、石崎研究室の先輩である湘南工科大学工学部 内山清子 准教授と東京工科大学コンピュータサイエンス学部 渡邊紀文 助教のお二人には、お会いする度に研究について気にかけて下さり、いつも励ましていただきました。後輩である後期博士課程 辰巳奈央 氏には、研究室行事を企画・運営していただき、楽しい時間を過ごさせてもらいました。また、時に研究室の仲間として、時に先輩や後輩として、本論文の研究内容に関する議論に付き合ってくれました石崎研究室の皆様、今井研究室の皆様、そして千葉大学文学部 伝康晴 教授をはじめとする Language and Communication 研究会の皆様にも御礼申し上げます。

最後に、後期博士課程の研究生活を応援してくれた家族に感謝致します。兄 義布史と義姉 正美、甥 義家に深く感謝致します。私が広島の実家に帰省する時は、必ず予定を空けてくれて、家族全員が揃う時間を作ってくれました。いつも会う度に、励ましてくれた兄夫妻、そして構ってくれた甥のおかげで、広島に帰ると精神的に癒されました。昨年、急逝しました祖父 故中村達雄と祖母 中村晴子に深く感謝致します。熊本で会う時や電話で話す時、いつも私の近況を喜んで聞いてくれて、そして学業・研究について励ましてくれました。祖父の生存中に、学位取得の報告ができなかったことが唯一の心残りですが、ここに感謝の意を記します。そして、父 郁人と母 沙津に深く感謝致します。郷里の広島から離れた藤沢の地で、学部から大学院まで自由に過ごすことができましたのも、両親の温かい応援があったからです。二人に心より感謝の意を表したいと思います。本当にありがとうございました。

慶應義塾大学大学院政策・メディア研究科

寺岡 丈博



## 参考文献

- [1] Alonge, A. (1998). “Data on Verb Semantics in the EuroWordNet Database”. In *Proceedings of the 1st International Conference on Language Resources and Evaluation (LREC 1998)*, pp. 385–392.
- [2] 天野真家, 石崎俊, 宇津呂武仁, 成田真澄, 福本淳一 (2007). 自然言語処理. オーム社.
- [3] 浅原正幸, 松本裕治 (2013). “『現代日本語書き言葉均衡コーパス』に対する係り受け・並列構造アノテーション”. 言語処理学会第19回年次大会発表論文集, pp. 66–69.
- [4] Baker, C. F. and C. Fellbaum (2009). “WordNet and FrameNet as Complementary Resources for Annotation”. In *Proceedings of the 3rd Linguistic Annotation Workshop, ACL-IJCNLP 2009*, pp. 125–129.
- [5] Baker, C. F., C. J. Fillmore, and J. B. Lowe (1998). “The Berkeley FrameNet Project”. In *Proceedings of the 36th Annual Meeting of the Association for Computational Linguistics and International Conference on Computational Linguistics*, volume 1, pp. 86–90.
- [6] Baker, C. F. and J. Ruppenhofs (2002). “FrameNet’s Frames vs. Levin’s Verb Classe”. In *Proceedings of the 28th Annual Meeting of the Berkeley Linguistics Society*, pp. 27–38.
- [7] Baker, C. F. and H. Sato (2003). “The FrameNet Data and Software”. In *The Companion Volume to the Proceedings of 41st Annual Meeting of the Association for Computational Linguistics*, pp. 161–164.

- [8] Blank, A. (1999). *Co-presence and Succession: A Cognitive Typology of Metonymy*, pp. 169–191. Amsterdam / Philadelphia: John Benjamins Publishing Company.
- [9] Boas, H. C. (2002). “Bilingual FrameNet Dictionaries for Machine Translation”. In *Proceedings of the 3rd International Conference on Language Resources and Evaluation (LREC 2002)*, pp. 1364–1371.
- [10] Bond, F., K. K. Hitoshi Isahara, and K. Uchimoto (2008). “Boot-strapping a WordNet using Multiple Existing WordNets”. In *Proceedings of the 6th International Conference on Language Resources and Evaluation (LREC 2008)*, pp. 1619–1624.
- [11] Bouaud, J., B. Bachimont, and P. Zweigenbaum (1996). “Processing Metonymy: a Domain-Model Heuristic Graph Traversal Approach”. In *Proceedings of the 16th International Conference on Computational Linguistics*, volume 1, pp. 137–142.
- [12] Burnard, L. and G. Aston (1998). *The BNC Handbook: Exploring the British National Corpus with Sara (Edinburgh Textbooks in Empirical Linguistics)*. Edinburgh University Press.
- [13] Chow, L. C. and J. J. Webster (2007). “Integration of Linguistic Resources for Verb Classification: FrameNet, WordNet and Suggested Upper Merged Ontology”. In *Proceedings of the 8th Conference on Intelligent Text Processing and Computational Linguistics (CICLing 2007)*, pp. 1–11.
- [14] Fass, D. (1991). “met\*: A Method for Discriminating Metonymy and Metaphor by Computer”. *Computer Linguistics*, **17**(1) pp.49–90.
- [15] Fellbaum, C. (1998). *WordNet- An Electronic Lexical Database*. The MIT Press.

- 
- [16] Fellbaum, C. (2006). “WordNet(s)”. In K. Brown, (ed), *Encyclopedia of Language & Linguistics*, volume 13. Oxford: Elsevier, pp. 665–670, Second edition.
- [17] Ferrandez, O., M. Ellsworth, R. Munoz, and C. F. Baker (2010). “Aligning FrameNet and WordNet based on Semantic Neighborhoods”. In *Proceedings of the 7th International Conference on Language Resources and Evaluation (LREC 2010)*, pp. 310–314.
- [18] Fillmore, C. J. (1968). “The Case for Case”. In Bach and Harms, (eds), *Universals in Linguistic Theory*. New York: Rinehart, and Winston, pp. 1–88.
- [19] Fillmore, C. J. (1976). “Frame Semantics and the Nature of Language”. In *Annals of the New York Academy of Sciences*, volume 280, pp. 20–32.
- [20] Francis, W. N. and H. Kucera (1979). “BROWN CORPUS MANUAL”. <http://icame.uib.no/brown/bcm.html>.
- [21] 藤井聖子, 小原京子 (2003). “フレーム意味論とフレームネット”. 英語青年, **149**(6) pp.373–376, 378.
- [22] 深谷昌弘, 田中茂範 (1996). コトバの意味づけ論—日常言語の生の営み—. 紀伊國屋書店.
- [23] Global WordNet Association (2013). “Wordnets in the world”. <http://www.globalwordnet.org/>.
- [24] Hamp, B. and H. Feldweg (1997). “GermaNet - a Lexical-Semantic Net for German”. In *Proceedings of ACL workshop Automatic Information Extraction and Building of Lexical Semantic Resources for NLP Applications*, pp. 9–15.
- [25] Henrich, V., E. Hinrichs, and T. Vodolazova (2012). “WebCAGe - A Web-Harvested Corpus Annotated with GermaNet Senses”. In *Proceedings of the 13th Conference of the European Chapter of the Association for Computational Linguistics (EACL2012)*, pp. 387–396.

- [26] 肥塚真輔, 岡本紘幸, 斎藤博昭, 小原京子 (2007). “日本語フレームネットに基づく意味役割推定”. 自然言語処理, **14**(1) pp.43–66.
- [27] 池原悟, 宮崎正弘, 白井諭, 横尾昭男, 中岩浩巳, 小倉健太郎, 大山芳史, 林良彦 (1999). 日本語語彙大系 CD-ROM 版. 岩波書店.
- [28] International Computer Science Institute (2012). “About - FrameNet”. <http://framenet.icsi.berkeley.edu/fndrupal/about>. 2013年4月25日.
- [29] Isahara, H., F. Bond, K. Uchimoto, M. Utiyama, and K. Kanzaki (2008). “Development of Japanese WordNet”. In *Proceedings of the 6th International Conference on Language Resources and Evaluation (LREC 2008)*, pp. 2420–2422.
- [30] Iverson, E. and S. Helmreich (1992). “Metalle: An Integrated Approach to Non-Literal Phrase Interpretation”. *Computational Intelligence*, **8**(3) pp.477–493.
- [31] Johansson, R., K. F. Heppin, and D. Kokkinakis (2012). “Semantic Role Labeling with the Swedish FrameNet”. In *Proceedings of the 8th International Conference on Language Resources and Evaluation (LREC 2012)*, pp. 3697–3700.
- [32] 甲斐睦朗, 松川利広 (2001). 語彙指導の方法—語彙表編—. 光村図書.
- [33] Kamei, S. and T. Wakao (1992). “Metonymy; Reassessment, Survey of Acceptability, and its Treatment in a Machine Translation System”. In *Proceedings of the 30th Annual Meeting of the Association for Computational Linguistics*, pp. 309–311.
- [34] Kawahara, D. and S. Kurohashi (2006). “Case Frame Compilation from the Web using High-Performance Computing”. In *Proceedings of the 5th International Conference on Language Resources and Evaluation (LREC 2006)*, pp. 1344–1347.
- [35] 河原大輔, 黒橋禎夫 (2004). “自動構築した格フレーム辞書と先行詞の位置選好順序を用いた省略解析”. 自然言語処理, **11**(3) pp.3–19.

- [36] 河原大輔, 黒橋禎夫 (2005). “格フレーム辞書の漸次的自動構築”. 自然言語処理, **12**(2) pp.109–131.
- [37] 慶應義塾大学環境情報学部石崎研究室 (2005). “連想概念辞書 第1版 2005年度版”.
- [38] Kipper, K., H. T. Dang, and M. S. Palmer (2000). “Class-Based Construction of a Verb Lexicon”. In *Proceedings of the 17th National Conference of Artificial Intelligence and 12th Conference on Innovative Applications of Artificial Intelligence*, pp. 691–696.
- [39] Kipper, K., A. Korhonen, N. Ryant, and M. S. Palmer (2008). “A Large-scale Classification of English Verbs”. *Language Resources and Evaluation Journal*, **42**(1) pp.21–40.
- [40] 清田陽司, 黒橋禎夫, 木戸冬子 (2004). “自動抽出した換喩表現を用いた係り受け関係のずれの解消”. 自然言語処理, **11**(4) pp.127–145.
- [41] 国立国語研究所 (2004). 分類語彙表 [増補改訂版]. 大日本図書.
- [42] 国立国語研究所 (2006). “日本語話し言葉コーパスの構築法”. 国立国語研究所報告 124, 国立国語研究所.
- [43] 黒橋禎夫, 長尾眞 (1997). “京都大学テキストコーパス・プロジェクト”. 人工知能学会全国大会論文集, pp. 58–61.
- [44] 京都大学大学院情報学研究科黒橋研究室 (2009). “京都大学格フレーム Ver 1.0”.
- [45] Lakoff, G. and M. Johnson (1980). *Metaphors We Live By*. University of Chicago Press.
- [46] Langacker, R. W. (1993). “Reference-Point Constructions”. *Cognitive Linguistics*, **4**(1) pp.1–38.

- [47] Langacker, R. W. (1995). “Raising and Transparency”. *Language*, **71**(1) pp.1–62.
- [48] Levin, B. (1993). *English Verb Classes and Alternation: A Preliminary Investigation*. The University of Chicago Press.
- [49] Marcus, M. P., B. Santorini, and M. A. Marcinkiewicz (1993). “Building a Large Annotated Corpus of English: The Penn Treebank”. *Computational Linguistics*, **19**(2) pp.313–330.
- [50] Markert, K. and U. Hahn (2002). “Understanding metonymies in discourse”. *Artificial Intelligence*, **135**(1) pp.145–198.
- [51] Markert, K. and M. Nissim (2003). “Corpus-Based Metonymy Analysis”. *Metaphor and Symbol*, **18**(3) pp.175–188.
- [52] Miller, G. A. (1995). “WordNet: A Lexical Database for English”. *Communications of ACM*, **38**(11) pp.39–41.
- [53] 文部科学省科学研究費特定領域研究「日本語コーパス」総括班 (2008). 『現代日本語書き言葉均衡コーパス』完成記念講演会予稿集.
- [54] 森田良行 (1989). 基礎日本語辞典. 角川学芸出版.
- [55] 森田純哉, 永井由佳里, 田浦俊春, 岡田亮士 (2008). “概念合成によるコンセプトのデザインと連想: 概念の連想数と動作概念の役割”. *認知科学*, **15**(4) pp.599–614.
- [56] 村田真樹, 山本専, 黒橋禎夫, 井佐原均, 長尾真 (2000). “名詞句「A の B」「AB」の用例を利用した換喩解析”. *人工知能学会誌*, **15**(3) pp.503–510.
- [57] 成山重子 (2009). 日本語の省略がわかる本—誰が? 誰に? 何を? How can we know who did what to whom in Japanese? [The Grammar of Omission: Less is More]. 明治書院.



- 
- [58] Nastase, V., A. Judea, K. Markert, and M. Strube (2012). “Local and Global Context for Supervised and Unsupervised Metonymy Resolution”. In *Proceedings of the 2012 Conference on Empirical Methods in Natural Language Processing*, pp. 183–193.
- [59] Nastase, V. and M. Strube (2009). “Combining Collocations, Lexical and Encyclopedic Knowledge for Metonymy Resolution”. In *Proceedings of the 2009 Conference on Empirical Methods in Natural Language Processing*, pp. 910–918.
- [60] Nicolae, C., G. Nicolae, and S. Harabagiu (2007). “UTD-HLT-CG: Semantic Architecture for Metonymy Resolution and Classification of Nominal Relations”. In *Proceedings of the 4th International Workshop on Semantic Evaluations (SemEval-2007)*, pp. 454–459.
- [61] 日本電子化辞書研究所 (1993). EDR 電子化辞書.
- [62] 小原京子 (2006). “フレーム意味論と日本語フレームネット”. *日本語学*, **25**(6) pp.40–52.
- [63] 小原京子 (2012). “日本語フレームネットにおけるコーパスデータの多義性分析”. 言語処理学会第18回年次大会発表論文集.
- [64] 岡本潤, 石崎俊 (2001). “概念間距離の定式化と既存電子化辞書との比較”. *自然言語処理*, **8**(4) pp.37–54.
- [65] 岡本潤, 石崎俊 (2003). “連想概念辞書の距離情報を用いた重要文の抽出”. *自然言語処理*, **10**(5) pp.139–151.
- [66] Okamoto, J. and S. Ishizaki (2010). “Homographic Ideogram Understanding Using Contextual Dynamic Network”. In *Proceedings of the 7th International Conference on Language Resources and Evaluation (LREC 2010)*, pp. 1180–1186.
- [67] 岡本潤, 石崎俊 (2012). “文脈ネットワークを用いた語の多義性解消”. *KEIO SFC JOURNAL*, **12**(1) pp.97–111.

- [68] Poibeau, T. (2007). “UP13: Knowledge-poor Methods (Sometimes) Perform Poorly”. In *Proceedings of the 4th International Workshop on Semantic Evaluations (SemEval-2007)*, pp. 418–421.
- [69] Princeton University (2010a). “About - WordNet”. <http://wordnet.princeton.edu>.
- [70] Princeton University (2010b). “wnstats - WordNet 3.0 database statistics”. <http://wordnet.princeton.edu/wordnet/man/wnstats.7WN.html>.
- [71] Rodriguez, H., S. Climent, P. Vossen, L. Bloksma, W. Peters, A. Alonge, F. Bertagna, and A. Roventini (1998). “The Top-Down Strategy for Building EuroWordNet: Vocabulary Coverage, Base Concepts and Top Ontology”. *Computers and the Humanities*, **32** pp.117–152.
- [72] 坂口琢哉 (2010). “連想概念辞書のニューラルネットワークへの符号化と比喻理解システムへの応用”. *安田女子大学紀要*, **38** pp.169–179.
- [73] Sakaguchi, T. and S. Ishizaki (2001). “Simile Understanding System Using Associative Concept Dictionary and Pulsed Neural Network”. In *Proceedings of the 7th Pacific Association for Computational Linguistics Conference(PACLING)*, pp. 229–235.
- [74] 坂口琢哉, 石崎俊 (2003). “連想概念辞書の実装による意味ネットワークと比喻理解システムへの応用”. *情報処理学会研究報告*, **91** pp.21–24.
- [75] 関和弘, 藤井敦, 石川徹也 (2002). “確率モデルを用いた日本語ゼロ代名詞の照応解析”. *自然言語処理*, **9(3)** pp.63–85.
- [76] 瀬戸賢一 (2002). *日本語のレトリック*. 岩波書店.
- [77] 白井諭, 横尾昭男, 中岩浩巳, 渡邊いづみ, 高橋直美, 関嘉美, 池原悟, 宮崎正弘 (1998). “構文意味辞書と構文体系”. *言語処理学会第4回年次大会発表論文集*, pp. 194–197.

- [78] Subirats, C. and H. Sato (2004). “Spanish FrameNet and FrameSQL”. In *Proceedings of the 4th International Conference on Language Resources and Evaluation (LREC 2004). Workshop on Building Lexical Resources from Semantically Annotated Corpora*, pp. 13–16.
- [79] 須賀智秋, 石崎俊 (2006). “連想概念辞書を用いたメトニミー意味解析システムの構築”. 言語処理学会第12回年次大会発表論文集, pp. 817–820.
- [80] Tagami, H., S. Hizuka, and H. Saito (2009). “Automatic Semantic Role Labeling based on Japanese FrameNet - A Progress Report -”. In *Proceedings of the 11th Pacific Association for Computational Linguistics Conference (PACLING)*, pp. 181–186.
- [81] 竹内孔一, 乾健太郎, 竹内奈央, 藤田篤 (2008). “意味の包含関係に基づく動詞項構造の細分類”. 言語処理学会第14回年次大会発表論文集, pp. 1037–1040.
- [82] 谷口一美 (2003). 認知意味論の新展開. 研究社.
- [83] 寺岡丈博, 東中竜一郎, 岡本潤, 石崎俊 (2013). “単語間連想関係を用いた換喩表現の自動検出”. 人工知能学会論文誌, **28**(3) pp.335–346.
- [84] Teraoka, T., R. Higashinaka, J. Okamoto, and S. Ishizaki (2012). “Automatic Detection of Metonymies using Associative Relations between Words”. In *Proceedings of the 34th Annual Conference of the Cognitive Science Society (CogSci)*, pp. 2417–2422.
- [85] 寺岡丈博, 岡本潤, 石崎俊 (2008a). “動詞連想概念辞書の構築とその応用”. 第7回情報科学技術フォーラム一般講演論文集第2分冊, pp. 229–230.
- [86] 寺岡丈博, 岡本潤, 石崎俊 (2008b). “動詞連想概念辞書の構築とその特徴”. 言語処理学会若手の会第3回シンポジウム (YANS).
- [87] Teraoka, T., J. Okamoto, and S. Ishizaki (2009). “Estimating Elliptical Words

- with Associative Information from Verbs”. In *Proceedings of the 11th Pacific Association for Computational Linguistics Conference(PACLING)*, pp. 60–65.
- [88] 寺岡丈博, 岡本潤, 石崎俊 (2009a). “動詞連想概念辞書を用いる省略語の推定法と評価—動詞連想概念辞書の構築と応用—”. 言語処理学会第15回年次大会発表論文集, pp. 845–848.
- [89] 寺岡丈博, 岡本潤, 石崎俊 (2009b). “連想情報に基づいた省略語推定システム”. 日本認知科学会第26回大会発表論文集, pp. 246–247.
- [90] Teraoka, T., J. Okamoto, and S. Ishizaki (2010). “An Associative Concept Dictionary for Verbs and its Application to Elliptical Word Estimation”. In *Proceedings of the 7th International Conference on Language Resources and Evaluation (LREC 2010)*, pp. 3851–3856.
- [91] Teraoka, T., J. Okamoto, and S. Ishizaki (2011). “Detecting Metonymic Expressions with Associative Information from Words”. *Computational Linguistics and Related Fields (Procedia-Social and Behavioral Sciences)*, **27** pp.105–112. Elsevier.
- [92] 寺岡丈博, 岡本潤, 石崎俊 (2011). “連想情報を用いた省略語の推定法と評価—動詞連想概念辞書の構築とその応用—”. *KEIO SFC JOURNAL*, **11**(1) pp.185–197.
- [93] 内山将夫, 村田真樹, 馬青, 内元清貴, 井佐原均 (2000). “統計的手法による換喩の解釈”. *自然言語処理*, **7**(2) pp.91–116.
- [94] Vapnik, V. N. (1995). *The Nature of Statistical Learning Theory*. Springer.
- [95] 山梨正明 (1988). 比喩と理解. 東京大学出版会.
- [96] Yamanashi, M. (1987). “Metonymic interpretation and associative processes in natural language”. In M. Nagao, (ed), *Language and Artificial Intelligence*. Elsevier Science Publishers B.V. (North-Holland), pp. 77–86.

- [97] You, L. and K. Liu (2005). “Building Chinese FrameNet database”. In *Proceedings of the 2005 12th IEEE International Conference on Natural Language Processing and Knowledge Engineering (IEEE NLP-KE '05)*, pp. 301–306.



# 本研究に関する発表一覧

## 原著論文

1. 寺岡文博, 東中竜一郎, 岡本潤, 石崎俊 (2013). “単語間連想関係を用いた換喩表現の自動検出”. 人工知能学会論文誌, **28**(3) pp.335-346. (ISSN: 1346-0714)
2. 寺岡文博, 岡本潤, 石崎俊 (2011). “連想情報を用いた省略語の推定法と評価—動詞連想概念辞書の構築とその応用—”. *KEIO SFC JOURNAL*, **11**(1) pp.185-197. (ISSN: 1347-2828)

## 国際会議

1. Teraoka, T., R. Higashinaka, J. Okamoto, and S. Ishizaki (2012). “Automatic Detection of Metonymies using Associative Relations between Words”, In *Proceedings of the 34th Annual Conference of the Cognitive Science Society (CogSci)*, pp.2417-2422.
2. Teraoka, T., J. Okamoto, and S. Ishizaki (2011). “Detecting Metonymic Expressions with Associative Information from Words”. In *Proceedings of the 12th Pacific Association for Computational Linguistics Conference (PA-CLING)*, #13.
3. Teraoka, T., J. Okamoto, and S. Ishizaki (2010). “An Associative Concept Dictionary for Verbs and its Application to Elliptical Word Estimation”. In *Proceedings of the 7th International Conference on Language Resources and Evaluation (LREC 2010)*, pp.3851-3856.

4. Teraoka, T., J. Okamoto, and S. Ishizaki (2009). “Estimating Elliptical Words with Noun and Verb Semantic Network -Application of Associative Concept Dictionaries-”. In *Proceedings of the 13th Conference of the International Conference on the Processing of East Asian Languages (ICPEAL)*, p.49.
5. Teraoka, T., J. Okamoto, and S. Ishizaki (2009). “Estimating Elliptical Words with Associative Information from Verbs”. In *Proceedings of the Conference of the Pacific Association for Computational Linguistics (PACLING)*, pp.60-65.

## その他

### 学術雑誌

1. Teraoka, T., J. Okamoto, and S. Ishizaki (2011). “Detecting Metonymic Expressions with Associative Information from Words”. *Computational Linguistics and Related Fields (Procedia-Social and Behavioral Sciences)*, Elsevier, **27** pp.105-112. 2011. (ISSN: 1877-0428)

### 国内会議・シンポジウム等

1. 寺岡文博 (2012). “連想の関係性を用いた換喩表現の自動検出”. 第37回 LC (Language and Communication) 研究会.
2. 寺岡文博, 岡本潤, 石崎俊 (2011). “動詞と名詞の連想による換喩検出法の提案”. 言語処理学会第17回年次大会発表論文集, pp.540-543.
3. 寺岡文博 (2010). “動詞連想概念辞書の構築とその応用～連想に基づいた意味理解の研究～”. 第20回 LC (Language and Communication) 研究会.



4. 寺岡文博, 岡本潤, 石崎俊 (2009). “連想情報に基づいた省略語推定システム”. 日本認知科学会第 26 回大会発表論文集, pp.246-247.
5. 寺岡文博, 岡本潤, 石崎俊 (2009). “動詞連想情報を用いる省略語の推定と評価—動詞連想概念辞書の構築と応用—”. 言語処理学会第 15 回年次大会発表論文集, pp.845-848.
6. 寺岡文博, 岡本潤, 石崎俊 (2008). “動詞連想概念辞書の構築とその特徴”. 言語処理学会若手の会第 3 回シンポジウム.
7. 寺岡文博, 岡本潤, 石崎俊 (2008). “動詞連想概念辞書の構築とその応用”. 第 7 回情報科学技術フォーラム一般講演論文集第 2 分冊, pp.229 -230.

## 招待講演

1. 寺岡文博 (2009). “連想概念辞書—実験材料ツール・知識ベースのモデルとしての可能性—”. 日本心理学会ディスコース研究部会第 2 回定例研究会, 法政大学.



# 付録

## 動詞連想概念辞書のデータ

### 1: 3.3節で使用した刺激語「運ぶ」の連想情報

連想課題	「運ぶ」に対する連想語と連想距離
動作主	私 (3.60), 引越し業者 (4.21), 運送屋 (4.28), 運転手 (4.45), 自分 (7.33), 友達 (7.78), 先生 (8.33), 父 (8.44), あなた (9.25), 人間 (9.42), 大工 (9.42), 彼 (9.42), 子供 (9.58), ラクダ (9.92), 自動車 (9.92), 車 (10.25), 兄 (10.42), 電車 (10.58), 船 (10.75), …
対象	荷物 (1.36), 人 (4.45), 物 (5.07), 足 (6.72), 段ボール (7.56), 箱 (7.67), 家具 (7.78), 机 (8.22), 本 (9.08), 危険物 (9.25), 麻薬 (9.25), ゴミ (9.42), 料理 (9.42), ダンボール箱 (9.75), テント (9.92), 食事 (9.92), 道具 (10.08), 仕事 (10.75), テレビ (10.92), …
始点	家 (1.45), 学校 (3.81), ここ (5.80), 部屋 (5.80), 倉庫 (6.38), 駅 (7.56), 店 (7.89), 日本 (7.89), 教室 (8.00), 自宅 (9.42), 実家 (9.92), 工場 (10.25), 部室 (10.25), 海 (11.58), …

終点	家 (1.92), 学校 (3.43), 駅 (3.73), 目的地 (4.02), 店 (4.64), 部屋 (4.72), アメリカ (4.77), 車 (5.56), 港 (6.53), 会社 (6.80), 日本 (7.05), あそこ (7.44), 引越し先 (7.56), 自宅 (7.67), 荷物 (9.08), 倉庫 (9.08), 友人宅 (9.42), 海 (9.56), 病院 (9.58), 会場 (9.92), 東京 (9.92), 教室 (10.25), ロンドン (10.42), 公園 (10.42), 空港 (10.58), 大学 (10.58), コンビニ (10.92), イギリス (11.08), ドイツ (11.42), 山 (12.08), …
時点	朝 (2.71), 昼 (2.94), 夜 (3.17), 昼間 (4.50), 深夜 (5.60), 夜中 (5.88), 午前中 (7.33), 午後 (7.56), 夕方 (7.56), 早朝 (9.08), 明日 (9.92), …
場所	家 (2.48), 学校 (3.27), 倉庫 (3.73), 工場 (4.50), 部屋 (5.53), 階段 (6.13), 会社 (6.22), 教室 (8.33), 電車 (9.08), 道 (9.08), 外 (9.42), 車 (9.42), 庭 (9.42), グランド (9.75), 道路 (9.75), 病院 (10.58), アメリカ (10.92), …
手段	車 (1.62), トラック (2.34), 手 (3.47), 電車 (3.57), 飛行機 (4.1), 自転車 (4.43), 船 (4.88), 台車 (4.95), カート (6.63), 足 (8.44), 人力車 (9.42), 滑車 (9.75), バイク (10.08), …
様相	必死に (3.17), 一生懸命 (3.36), ゆっくり (3.43), 急いで (3.61), 軽々 (6.13), 懸命に (7.67), 重そうに (7.89), 毎日 (8.11), どんどん (9.08), せっせと (9.25), さっさと (9.58), 早く (9.58), 丁寧に (9.58), 静かに (9.75), …

(複数の実験参加者が答えた連想語を主に記載している.)

## 2: 3.3節で使用した刺激語「借りる」の連想情報

連想課題	「借りる」に対する連想語と連想距離
動作主	私 (2.04), 友達 (3.81), 客 (5.07), 彼 (5.47), 母 (6.05), 彼女 (6.47), 自分 (7.33), 後輩 (7.56), 学生 (7.89), 父 (8.00), 弟 (8.33), 人 (9.08), あなた (9.42), 姉 (10.08), 従兄弟 (10.08), …
対象	金 (かね) (1.79), 本 (1.97), CD(3.76), 鉛筆 (4.45), 消しゴム (5.8), ノート (6.30), 自転車 (6.80), 手 (6.88), 家 (8.00), 教科書 (8.00), 車 (8.00), ビデオ (8.11), 知恵 (8.11), 服 (8.22), パソコン (8.33), DVD(8.56), 筆記用具 (9.11), 力 (9.25), 傘 (9.42), 道具 (9.42), 物 (9.42), 人手 (9.75), PC(10.42), 助け (10.58), 靴 (11.58), …
様相	たくさん (5.47), 頻繁に (5.88), 申し訳なさそうに (6.05), 何度も (7.33), 丁寧に (7.33), すぐに (7.67), 少し (8.11), 簡単に (9.08), 毎日 (9.08), いつも (9.58), 申し訳なく (9.75), 必死に (9.75), …, しぶしぶ (14.33), …
場所	図書館 (2.30), 学校 (3.30), 家 (3.77), 銀行 (4.40), 部屋 (7.44), レンタルショップ (7.56), 店 (8.11), サラ金 (9.08), 不動産屋 (9.25), オフィス (9.75), 友達の家 (9.92), …

(複数の実験参加者が答えた連想語を主に記載している.)

### 3: 3.3節で使用した刺激語「貸す」の連想情報

連想課題	「貸す」に対する連想語と連想距離
動作主	私 (1.79), 銀行 (2.62), 友達 (3.13), 親 (3.70), 友人 (4.18), 先生 (4.69), 人 (5.80), 子供 (6.13), 母 (7.05), 自分 (8.22), 彼 (8.33), 高利貸し (8.44), 教授 (8.56), 図書館 (8.56), 会社 (8.67), 父 (8.78), 質屋 (9.58), 家主 (9.75), 金融会社 (9.75), 兄 (9.75), 彼女 (9.92), ビデオ屋 (10.08), 生徒 (10.08), レンタルビデオショップ (10.25), 企業 (10.42), 国 (10.42), 金持ち (10.92), …
対象	金 (かね) (1.76), 本 (2.02), CD(4.41), 友達 (5.00), 服 (6.55), ノート (6.72), 自転車 (7.89), 手 (8.44), 力 (8.78), ゲーム (9.42), ビデオ (9.42), 物 (9.42), 教科書 (9.75), 家 (10.08), 漫画 (10.08), 消しゴム (10.42), 耳 (10.58), 部屋 (11.08), …
様相	しぶしぶ (2.93), 快く (5.88), 嫌々 (5.88), 少し (7.89), 気軽に (9.08), 頻繁に (9.08), たくさん (9.25), すんなり (9.42), 仕方なく (9.42), 喜んで (9.58), 毎日 (9.58), たまに (10.42), …
場所	学校 (2.30), 銀行 (2.79), 図書館 (3.36), 家 (3.81), 店 (4.39), 教室 (5.80), 部屋 (7.44), 会社 (7.89), ツタヤ (9.08), ビデオショップ (9.08), 金融機関 (9.58), 自宅 (9.58), 居酒屋 (9.75), 大学 (9.75), 公園 (9.92), ビデオ屋 (10.08), レンタルショップ (10.75), …

(複数の実験参加者が答えた連想語を主に記載している.)

## 4: 3.3節で使用した刺激語「歩く」の連想情報

連想課題	「歩く」に対する連想語と連想距離
動作主	犬 (2.39), 私 (3.17), 人 (3.28), 子供 (4.72), 猫 (5.22), 動物 (5.73), 人間 (5.88), 彼 (6.13), 赤ちゃん (7.44), 老人 (7.89), あなた (8.00), 猿 (8.00), おじいちゃん (9.08), 僕 (9.08), おばあさん (9.42), わたし (9.42), ロボット (9.58), おじいさん (9.75), 君 (10.08), 昆虫 (10.08), 彼女 (10.42), アヒル (10.92), …
始点	家 (1.38), 学校 (2.31), 駅 (2.46), 公園 (5.47), 自宅 (6.55), バス停 (8.33), 店 (9.44), 教室 (9.58), 海 (9.67), ここ (9.75), 会社 (9.92), 図書館 (10.58), 塾 (10.92), 駐車場 (11.92), 銀行 (12.25), …
終点	学校 (1.88), 家 (1.93), 駅 (2.17), ゴール (6.38), 店 (6.38), 目的地 (8.11), バス停 (8.33), 自宅 (9.75), 職場 (9.75), 疲れる (9.75), 友達の家 (9.75), 公園 (10.08), 山頂 (10.08), …
時点	朝 (2.13), 昼 (2.51), 夜 (3.87), 午前中 (4.14), 午後 (5.17), 夕方 (5.40), 早朝 (6.38), 毎朝 (7.33), 明日 (8.11), 夜中 (8.56), 今 (9.08), 真夜中 (9.42), いつも (9.58), 食後 (9.58), …
場所	道 (2.34), 歩道 (2.37), 公園 (2.88), 廊下 (4.95), 道路 (6.30), 通り (7.56), 学校 (7.67), 歩道橋 (7.67), キャンパス (8.22), 山 (9.25), 横断歩道 (9.42), 遊歩道 (9.58), 街 (9.75), 浜辺 (9.92), 墓地 (9.92), 床 (10.08), 並木道 (10.75), …
手段	足 (1.59), 靴 (3.77), 杖 (4.02), 松葉杖 (5.27), 竹馬 (5.53), 義足 (6.30), 徒歩 (7.33), 手 (7.89), 裸足 (8.00), サングル (8.33), 歩行器 (9.25), スニーカー (9.42), 自分の足 (9.58), スリッパ (9.75), ブーツ (9.75), 靴下 (9.92), 片足 (9.92), 機械 (10.08), 素足 (11.42), …

---

様相	ゆっくり (2.19), 必死に (4.31), のんびり (4.55), てくてく (6.13), 急いで (6.38), はやく (7.67), とぼとぼ (7.89), 悲しそうに (8.56), 軽やかに (9.08), 楽しそうに (9.11), 一生懸命 (9.25), 速く (9.25), 毎日 (9.25), だらだら (9.75), 毎晩 (9.92), せかせか (10.08), …
----	--

---

(複数の実験参加者が答えた連想語を主に記載している.)



## 5: 3.5節で使用した刺激語「取る」の連想情報

連想課題	「取る」に対する連想語と連想距離
動作主	私 (2.76), 子供 (4.50), 人 (5.00), 自分 (5.13), 先生 (6.20), 彼女 (6.22), 犬 (7.67), 彼 (7.67), 猿 (7.78), 俺 (9.08), サル (9.25), 泥棒 (9.25), あなた (9.58), 動物 (9.58), 友人 (9.58), 友達 (9.75), 猫 (9.92), 父 (10.08), 後輩 (10.42), 母 (10.58), 妹 (10.58), 兄 (11.08), …
対象	物 (5.13), 荷物 (5.97), コップ (7.33), 単位 (7.44), 金 (7.56), 塩 (7.67), 資格 (7.67), お金 (7.78), ボール (7.78), 手 (7.89), ペン (9.08), りんご (9.08), 本 (9.08), 商品 (9.25), ティッシュ (9.42), 消しゴム (9.42), リモコン (9.58), 舵 (9.58), 歳 (9.75), ゴミ (10.08), 魚 (10.08), 財布 (10.08), 帽子 (10.25), …
始点	机 (3.47), 家 (4.12), 皿 (7.44), 棚 (7.67), 本 (8.22), かご (9.08), 本棚 (9.08), 学校 (9.25), ケース (9.92), 引き出し (9.92), 人 (9.92), 手 (10.08), 箱 (10.25), 冷蔵庫 (11.08), …
時点	朝 (2.58), 昼 (2.79), 夜 (3.06), 今 (4.07), 午前中 (4.93), 午後 (6.06), 明日 (6.13), 最後 (6.30), 夕方 (7.05), さっき (7.78), 今日 (7.78), 昨日 (7.89), 早朝 (8.89), 深夜 (9.33), 夜中 (9.89), 秋 (10.08), 夏 (10.25), 空き時間 (11.08), …
場所	部屋 (4.28), 教室 (5.88), 店 (5.88), 家 (6.05), 公園 (7.44), 学校 (7.56), 席 (9.08), テーブル (9.25), 街中 (9.42), 図書館 (9.42), 棚 (9.42), スーパー (9.58), 畑 (9.75), 机 (9.92), 道 (10.25), …
手段	手 (1.62), 箸 (4.50), 棒 (6.05), フォーク (6.27), トング (6.63), スプーン (6.80), アーム (7.44), ペンチ (7.56), 足 (7.78), 網 (8.67), 包丁 (10.42), はさみ (11.08), …

---

様相	必死に (3.13), 素早く (3.61), ゆっくり (6.00), そっと (8.00), さつと (9.08), すつと (9.08), 急いで (9.25), しっかり (9.42), 適当に (9.58), 簡単に (9.75), 多く (9.75), がむしやらに (9.92), 丁寧に (10.25), 慎重に (10.42), …
----	---

---

(複数の実験参加者が答えた連想語を主に記載している.)