動詞連想情報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—

寺岡 丈博[†]　岡本 潤[‡]　石崎 俊[†]

[†]慶應義塾大学大学院政策・メディア研究科, {teraoka,ishizaki}@sfc.keio.ac.jp

[‡]慶應義塾大学 SFC 研究所，juno@sfc.keio.ac.jp

1. はじめに

自然言語処理技術が著しく発展してきた今日ではあるが，人間に近づくような意味理解を行うまでの精度を求めるのは未だに困難である．もとより，単語の品詞や文法などの言語学的な情報だけではコンピュータが人間のように言語を扱うには不十分である．人間は言葉を話す或いは書く際にそれらの言語学的な情報だけでなく，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．即ち，コンピュータの言語理解機能を向上させ人間に近づけるためには，人間が持つ複雑で膨大な言語関連情報を体系化して使用できるようにする必要がある．

このアプローチの一つとして連想概念辞書[4]が挙げられる．連想概念辞書は人間の直感に基づいた大規模な連想実験のデータに基づいて言葉の背景にある情報を体系化し，連想距離を定量化することによって語間の意味を計算できるようにしている．現在は名詞を中心として構築されており，重要文の抽出[5]や多義性の解消モデル[6]などに応用している．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で中心的な役割を果たしていることから，動詞に関しても言葉の背景の情報を同様に体系化・定量化する必要性が考えられる．

そのため本研究では動詞を刺激語にして深層格情報を抽出する連想実験を行い，動詞連想概念辞書を構築することで動詞における知識の体系化を図っている．そこでは刺激語と連想語の間の連想距離を定量化することにより意味的な距離を計算可能にしている[8]．更にその応用として，ある特定の深層格に関する省略語を推定するシステムを試作した．名詞連想概念辞書と組み合わせて推定した確信度付きの省略語に対し，人間を被験者にして推定された省略語を基準にして評価することでその有効性を確認した．

2. 動詞連想概念辞書

小学校の国語の教科書で扱われている動詞[1]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動詞[3]を基礎動詞と定めて計 200 語の基礎動詞を刺激語とした．刺激語と一緒に提示される連想課題はフレーム意味論における深層格の一部を参考にして表 1 のように設定しており，被験者が理解し易いように幾つかの連想課題については深層格を統合している．

表 1 連想課題の内容

連想課題	意味内容
動作主	動作を行う主体
対象	動作の対象
始点	動作の始点・起点
終点	動作の終点・目標
時点	動作が行われる時刻・時間
場所	動作が行われる場所・空間
手段	動作を行うための道具・材料
様相	動作の様態・様子・程度・頻度
理由	動作の理由・原因
目的	動作の目的

連想距離$D(x,y)$は刺激語xと連想語yの単語間距離であり式(1)のように表わされる[8]．この式は刺激語xに対して連想語yが連想された頻度の逆数$F(x,y)$と，同じく刺激語xに対して連想語yが連想された順位s_iの相加平均$S(x,y)$によって連想距離$D(x,y)$を線形結合で表わし，線形計画法を用いて得られた最適解を利用している．尚，連想実験から得られるパラメータから連想距離$D(x,y)$の最小値が 1.0 程度，最大値が 10.0 程度になるように境界条件を定め，シンプレックス法で計算されている．また$F(x,y)$は補正値δを分母に加えることで正規化を行っており，被験者数Nを大幅に増加させた時に連想した人数nが少ない場合でも$F(x,y)$の値が極端に大きくなるのを防いで連想距離$D(x,y)$の極端な変動を抑えることが可能となっている．

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

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

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

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

このように動詞連想概念辞書は刺激語動詞とそれに対する連想課題毎の連想語，そして各々の連想距離によって成り立っている．現在の規模は，基礎動詞に加えてその他の基本動詞も刺激語に加わり，232 語の刺激語動詞に対して連想語数が 57517 語，異なり語数が 15608 語となっている．

3. 省略語推定システム

3.1 システムの概要

動詞連想概念辞書の応用として文脈内において指定した深層格に対応する省略語を補完するシステムを試作した．その処理の概要について図 1 に沿って述べる．まず入力文に対して形態素解析 [1] を行い，述語動詞に対して動詞連想概念辞書を用いて指定した深層格の内容に対応する連想課題の連想語を抽出して省略語の候補群とする．次に，入力文内の全ての名詞に対して，名詞の連想概念辞書[4]（以下，名詞連想概念辞書）を用いて深層格と対応させた概念に関する連想語や逆引きとして得られる刺激語を抽出し，これらの単語と省略語の候補群で共通する語が省略語の最終的な候補として絞り込まれる．そして最後に省略語の候補に対する動詞連想概念辞書と名詞連想概念辞書のそれぞれの連想距離を用いて求めた確信度の大きい順に出力する．

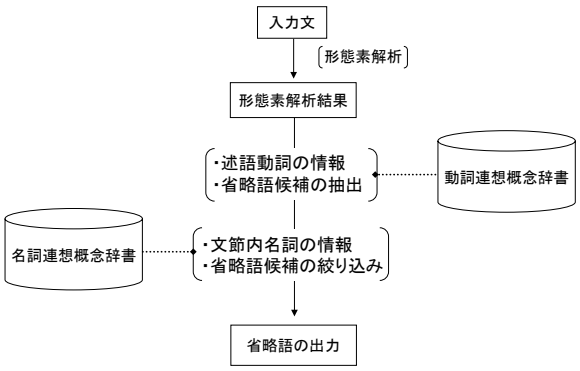

図 1 システムの概要

以上のように動詞連想概念辞書と名詞連想概念辞書を各段階で利用しており，詳細は次節以降で触れる．

3.2 省略語候補の抽出

動詞連想概念辞書の刺激語動詞は基本形で記されている．そのため入力文を形態素解析した結果から得られる述語動詞の基本形を参照し，動詞連想概念辞書からその刺激語動詞に関して実験者が選んだ深層格の連想語を全て抽出する．例えば「財布を忘れたので友達から借りた。」という文に対して何を借りたのかを出力させたい場合，形態素解析して得られる述語動詞の基本形は「借りる」であるため，動詞連想概念辞書からは刺激語動詞「借りる」の連想課題「対象」に関する連想語を全て抽出する．抽出された連想語は省略語の候補として連想距離の情報とともに候補群リストに一時的に格納される．

3.3 省略語候補の絞り込み

省略語の候補群の中から絞り込みを行う時には名詞連想概念辞書の刺激語名詞と連想語の情報を用いる．手順としては，入力文を形態素解析した際に名詞と判別された単語に関して，名詞連想概念辞書に刺激語名詞として登録されているかを調べる．登録されている場合は，省略語の補完内容と関係のある概念 [2]（このシステムでは「環境概念」と「部分・材料概念」を用いる）について前節の候補群リストに格納されている単語について共通する連想語が存在すれば抽出する．しかし文中にある名詞が刺激語名詞として登録されていない場合は，その名詞が前述の概念に対する連想語として登録されている情報の有無を調べ，存在する場合はその時の刺激語名詞が候補群リストに格納されている単語と共通する場合のみ抽出する．また，入力文に固有名詞が含まれている場合は，その固有名詞を連想語として登録されている刺激語（つまり意味的には固有名詞に対応する概念）を探索して，その刺激語の連想語を辿ることで省略語の候補の絞り込みを行えるようにしている．このような処理を文中の全ての名詞に対して行うことで省略語候補が最終的に絞り込まれる．

3.4 確信度による順位付け

最終的な省略語の候補として抽出された単語は，動詞連想概念辞書の連想距離$D_v(D_v \geq 1.0)$と名詞連想概念辞書の連想距離$D_n(D_n \geq 1.0)$[4]から式(2)のように定義した確信度c の大きい順に出力される．

$$\begin{cases} c = \dfrac{1}{D_v D_n} & (D_v D_n \neq 0) \\ 0 < c \leq 1 \end{cases} \tag{2}$$

尚，3.2 節で挙げた例文「財布を忘れたので友達から借りた。」に対しては「金（かね）」，「手」という順で出力されており，確信度cは 0.255 と 0.015 であった．これらの内容の差を踏まえると，確信度c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．

[1] 形態素解析器は MeCab 0.97 を使用．

[2] 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，環境概念の全 7 概念．

4. 省略語推定の評価

4.1 課題文の設定

省略語推定システムの評価を行うために，入力文に用いる省略語推定の課題文と補完させる深層格を表 2 のように 10 文を設定した．これらは客観性を保つために複数の実験者がそれぞれ作成した．尚，課題文中の述語動詞は動詞連想概念辞書の刺激語として，文中の名詞は名詞連想概念辞書の刺激語もしくは連想語として情報があるものとしている．

表 2 省略語推定の課題文

課題文	補完深層格
1.駅のホームで待つ。	対象
2.説明のために、黒板に書く。	対象
3.チューリップを植える。	場所
4.友人と弁当を食べる。	場所
5.老人に席を譲った。	場所
6.駅で少年は降りた。	始点
7.富士山を登る。	終点
8.運転手は選手達を試合会場まで運んだ。	手段
9.音楽室でノクターンを弾く。	手段
10.台所で野菜を切る。	手段

4.2 評価基準の作成

上記で設定した課題文に対する省略語推定システムの出力結果を人間が推定した結果と比較するために，被験者 10 人に対して実験を行った．実験の刺激は表 2 の課題文と述語動詞に関して補完する深層格（連想課題）とし，被験者には省略語として何が推定できるかを答えてもらった．この実験では，被験者が推定した単語の頻度と各々の推定した順位のパラメータを得ることができるため，式(1)と同様の式を用いて単語間の距離D_sを算出した．この距離D_sが小さいほど，その単語が推定され易いということを表しており，これらを基準データと定めた．

4.3 推定結果の評価

そもそも文脈上で省略語を決定するにはその前後関係から割り出す必要があり，本研究で設定した課題文は単文であることから省略語をただ 1 つの正解として決定付けることは難しい．しかし，今回の比較は省略語推定システムが人間と同じような推定がどの程度まで可能かどうかを調べるのが目的である．そのため，課題文に対する省略語推定システムの結果について基準データを正解として，正解率と順位一致率を求めた．表 3 は課題文 10 に関する出力結果を基準データとともにまとめたもので，基準データの単語数に比べて省略語推定システムが出力した単語数は少なくなっている．この傾向はどの課題文に対しても見られたため，ここで扱う正解率は省略語推定システムが出力した単語の内，基準データに含まれている単語が占める割合を指している．

表 3 課題文 10 に対する推定結果

順位	基準データ		システムの出力	
	単語	D_s	単語	c
1	包丁	1.21	包丁	0.122
2	ナイフ	2.00	ハサミ	0.105
3	ハサミ	3.11	手	0.029
4	手刀	4.33	ナイフ	0.026
5	スライサー	7.67	水	0.015
6	カッター	8.00		
7	手	8.33		

また，順位一致率pは省略語推定システムで出力された単語の順序δと基準データの中でn個の共通する単語（つまり正解した単語）の順序τについて単語iのそれぞれの順位$r_{\delta i}$と$r_{\tau i}$を用いた Spearman's Footrule 距離$d_{(\delta,\tau)}$とその最大値d_{max}で式(3)のように表される．

$$\begin{cases} p = \dfrac{d_{max} - d_{(\delta,\tau)}}{d_{max}} \\ d_{(\delta,\tau)} = \sum_{i=1}^{n} |r_{\delta i} - r_{\tau i}| \end{cases} \tag{3}$$

表 4 のように正解率が 100.0%の課題文もあるが，10 個の課題文の全体で正解率が 54.9%と低めなのに対して順位一致率が 70.3%と比較的に高めになっている．このことから，共通する単語の順位の大小関係が似ていることが分かる．つまり共通している単語は人間が推定する順位と似ていることが言える．その反面，正解率が半分強であることから省略語推定システムが出力した単語の半分弱は人間には推定されない単語であり，その点で基準データとかけ離れている可能性が考えられる．そこで実際に正解とされた単語に関して確信度に用いられた 2 つの連想距離D_vとD_nの積の平均を求めて，その逆数から得られた値を確信度の閾値(c_t= 0.025)を設定した．実際に確信度cが閾値c_tより大きい単語だけを出力させると，課題文 3 と課題文 5 は正解率に変化が見られなかったが，残りの課題文に対しては総じて正解率が上昇している．その結果，閾値を設けた場合は全体で正解率が 76.3%となった．順位一致率も 75.0%と上昇したが，閾値により出力された単語数が減少したため課題文 1 の順位一致率は 0.0%になった．これは出力された単語が 6 語から 3 語に，基準データと共通する語が 3 語から 2 語に減ったため，順位一致率が 0.0%か 100.0%のどちらかの値しか取らないからである．また課題文 2 と課題文 8 の順位一致率の値が無いのは，閾値の設定により出

力されて共通する単語が 1 語のみとなったため順位の比較ができないためである．

表 4 正解率と順位一致率

課題番号	正解率(%)		順位一致率(%)	
	$c > 0$	$c > c_t$	$c > 0$	$c > c_t$
1	50.0	66.7	50.0	0.0
2	42.9	100.0	50.0	--
3	50.0	50.0	77.8	100.0
4	61.5	80.0	75.0	50.0
5	50.0	50.0	100.0	100.0
6	36.4	50.0	100.0	100.0
7	50.0	66.7	100.0	100.0
8	28.6	100.0	0.0	--
9	100.0	100.0	100.0	100.0
10	80.0	100.0	50.0	50.0
計	54.9	76.3	70.3	75.0

5. 考察

これまでの省略・照応解析の研究分野で考えると，確率モデルによる省略解析の手法[7]や格フレーム辞書を用いてゼロ代名詞を高い精度で検出可能にした手法[2]と比べて，連想情報を用いる省略語推定システムは単純でオリジナルな手法ではあるが，このシステムの精度を十分に示すには至っていない．なぜならば本研究の評価に用いた課題文は単文であり，文の数が少ないことなどが理由として挙げられる．しかし，表 4 から分かるように省略語を推定する内容が「対象」「場所」「始点」「終点」「手段」と幅広く対応するという特長がある．「手段」に関しては特に正解率が高く，表 3 を例にして理由を次に述べる．省略語として上位に推定された「包丁」や「ハサミ」については，名詞連想概念辞書の情報を用いて絞り込みを行う際に文中の「台所」は刺激語として登録されており，その「部分・材料概念」[3] の連想語として「包丁」や「ハサミ」が存在したため最終的な候補として出力されている．これは文中の名詞である「台所」は「省略語」にとって「部分として存在する場所」である．つまり省略語の候補を絞り込む上で，文中の名詞と推定する省略語が名詞連想概念辞書の概念によって明確に対応している関係であるため，「手段」を推定する正解率が高いと考えられる．

4.3 節を通じて，省略語推定システムは確信度に閾値を設定して出力させる単語を制限することで，無駄な単語を省いて正解率を大幅に上げられたことから，人間が省略語を推定した結果の内容に近づいたと言えるため，一定の評価を与えることができる．そして，これは省略語の候補を抽出する段階で述語動詞と意味的な関係を持つ深層格情報がきちんと取り出せているためであり，最初の候補を抽出する段階で人間が推定する内容を確実に含んでいることが前提になっている．そのため，省略語の候補を抽出する際に用いる動詞連想概念辞書には，人間が言葉を扱う上で用いている動詞の背景にある知識が含まれていると考えられる．ゆえに応用において動詞連想概念辞書の有効性を示すことができたと言える．また同時に，省略語候補の絞り込みで用いる名詞連想概念辞書についても同様に当てはまるであろう．

[3] 部分・材料概念は刺激語に対して連想語が部分や材料を表し，ここでは環境概念の双対になっている．

6. おわりに

本研究では，動詞連想概念辞書と名詞連想概念辞書を併用して省略語推定システムを作成し，課題文に対して確信度付きで省略語を推定させた結果と人間が推定した内容の比較を行った．そして閾値を設定することで，最終的には正解率 76.3%，順位一致率 75.0%を得ることができ，その有効性を示すことができた．今後の課題としては動詞連想概念辞書のデータの拡充することの他に名詞連想概念辞書と表記を合わせることが挙げられる．また，今回の省略語推定の課題文は人間の推定結果との近さを確認するためだったが，いずれは新聞コーパスや Web 上の文書を対象にして省略語推定システム自体の精度を確認していきたい．

参考文献

[1]甲斐睦朗, 松川利広, “語彙指導の方法 - 語彙表編 - ”, 光村図書(2001).

[2]河原大輔, 黒橋禎夫, “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, Vol.11, No.3(2004).

[3]森田良行, “基礎日本語辞典”, 角川学芸出版(1989).

[4]岡本潤, 石崎俊, “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, Vol.8, No.4(2001).

[5]岡本潤, 石崎俊, “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, Vol.10, No. 5(2003).

[6]Okamoto, J., Uchiyama, K. and Ishizaki, S., “A Contextual Dynamic Network Model for WSD Using Associative Concept Dictionary”, LREC (2008).

[7]関和弘, 藤井敦, 石川徹也, “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, Vol.9, No.3(2002).

[8]寺岡丈博, 岡本潤, 石崎俊, “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(2008).