

動詞連想概念辞書の構築とその応用 Construction of associative concept dictionary for verbs and its applications

寺岡 文博[†] 岡本 潤[‡] 石崎 俊[†]
Takehiro TERAOKA Jun OKAMOTO Shun ISHIZAKI

1. はじめに

単語の品詞や文法などの言語学的情報だけではコンピュータが人間のように言語を扱うことはできない。なぜなら人間は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からである。即ち、この知識とも言うべき複雑で膨大な情報を体系化したものが必要である。そのアプローチとして言語情報だけでなく意味を表すための概念の収集と構造化が挙げられ、一例として連想概念辞書がある。連想概念辞書とは、人間の直感に基づいた連想実験のデータに基づいて定量化した連想距離を用いて、言葉の背景にある情報を体系化したものであり、現在は名詞を中心とした連想概念辞書が構築されている[3]。連想距離は人間の知識として単語間の関係を表しており、係り受けから単語と単語の近さを表しているモデル[5]と比べてより人間の連想という要素から成り立っている。このような視点から連想概念辞書は、Web などから大量の情報を入手して構築している格フレーム辞書[2]などのコーパスやシソーラスとは異なる面を持つ。ここで人間が日常で使用する文脈の中で動作や変化を表す品詞である動詞は意味理解の重要な役割を担っていることを踏まえると、動詞を中心とした連想概念辞書即ち動詞連想概念辞書の必要性が考えられる。そこで動詞を刺激語とした連想実験から得られたデータを利用し、刺激語と連想語における単語間距離である連想距離を定量化することで動詞連想概念辞書の構築を図った。

本稿では動詞連想概念辞書の構築の過程と考察を述べ、最後に自然言語処理分野への応用について触れる。

2. 動詞連想実験

小学校の国語の教科書で用いられる 1123 語の動詞[1]を基本動詞と見なした上で、他動詞の優先や機能動詞の除外などで効率的に減らして 798 語とした。これらの内の一部の基本動詞に対して、「位置移動動詞」「所有移動動詞」「感情動詞」「知覚動詞」「構築動詞」「破壊動詞」と意味的に分類した。この分類は連想語に及ぼす動詞の種類の影響の有無を明確にして動詞連想概念辞書の今後の拡張に生かすためである。そして後述の連想課題の組み合わせを考慮し、計 54 語の代表的な動詞を刺激語と設定した。また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おり。なぜならば深層格は構文中で動詞と特定の意味関係を持つ格であり、これに基づいた連想実験データの利用により動詞まわりの知識を体系化できる可能性が考えられるからである。幾つかの連想課題については複数の深層格を統合させて被験者の理解の簡易化を図り、表 1 のような 10 種類の連想課題を設定して連想実験を行った。尚、この連想実験には慶應義塾大学湘南藤沢キャンパス (SFC) の学内ネットワークを利用したオンライン上で稼動する実験システムを用いており 1 刺激語 1 連想課題につき被験者

20 人とし、SFC の学部生と大学院生の述べ約 170 人が被験者となっている。

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・程度など
理由	動作の理由・原因
目的	動作の目的

3. 連想距離の定量化

連想距離 D を連想頻度の逆数 F 、連想順位 S 、連想時間を対数計算した値 T' の線形結合で表わす。

$$D = \alpha \times F + \beta \times S + \gamma \times T' \quad (1)$$

F は連想された頻度の逆数を表し、連想した人数に補正値を加えた値で被験者数を割った値であり、補正値を分母に加えることで正規化を行っている。これにより被験者数を大幅に増加させた時に、連想した人数が少ない場合でも F の値が極端に大きくなるのを防ぎ、連想距離 D の極端な変動を抑えることが可能となる。また連想順位 S は各被験者が連想した語の順位の相加平均である。ここで、連想実験システムで刺激語と連想課題を提示された時から被験者が連想語を入力し始めるまでの時間を連想時間と見なし、それを対数計算した値 T' を用いることで被験者毎に生ずる大幅な個人差を解消している。このようにして正規化を行い、式(1)の係数 α 、 β 、 γ を求めるために目的関数 (式(2)) と制約式 (式(3)~(5)) を設定し、線形計画法を用いて目的関数の値 Z を最小にする時の係数の最適解を求める。

$$Z = c_1 \times \alpha + c_2 \times \beta + c_3 \times \gamma \quad (2)$$

$$\begin{cases} a_{11} \times \alpha + a_{12} \times \beta + a_{13} \times \gamma = D_1 & (3) \\ a_{21} \times \alpha + a_{22} \times \beta + a_{23} \times \gamma = D_2 & (4) \\ \alpha, \beta, \gamma \geq 0 & (5) \end{cases}$$

制約式で $(a_{11}, a_{12}, a_{13}, D_1) = (20/21, 1, 1, 1)$ 、 $(a_{21}, a_{22}, a_{23}, D_2) = (10, 9, 5, 10)$ とした時、目的関数において $(c_1, c_2, c_3) = (10, 8, 3)$ とし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 (式(6))。このような連想距離を用いて動詞連想概念辞書を構築した。

$$D = \frac{7}{10} \times F + \frac{1}{3} \times S \quad (6)$$

[†] 慶應義塾大学大学院 政策・メディア研究科
Graduate School of Media and Governance, Keio University
[‡] 慶應義塾大学 SFC 研究所
Keio Research Institute at SFC

4. 動詞連想概念辞書

4.1 連想語数と異なり語数

構築した動詞連想概念辞書は、刺激語 54 語に対して連想語数が 12174 語、異なり語数が 5461 語となっている。連想語数とは連想された語の全ての合計数であり、異なり語数とは刺激語と連想課題の組み合わせが異なっていたとしても同じ語が連想されている場合は同じ単語として数えた合計数である。表 2 より明らかに「動作主」「時点」「場所」「様相」の連想課題に関しては連想語数と異なり語数の差が顕著であり、異なり語数が連想語と比べてかなり少なくなっている。これは色々な刺激語から多くの同一の語が連想されていることを表している。反対に「始点」「終点」「理由」「目的」の連想課題は連想語数と異なり語数の差が小さい関係であることから、各刺激語に対して連想課題特有の語が連想されていることが分かる。

表 2 連想語数と異なり語数

連想課題	連想語数	異なり語数
動作主	1903	724
対象	1872	1232
始点	462	319
終点	460	338
時点	1686	745
場所	1717	799
手段	1093	731
様相	1824	927
理由	662	560
目的	493	398

4.2 連想語と連想距離

図 1 は反意語関係(双対関係)にある「借りる」と「貸す」について動詞連想概念辞書からデータを抽出して連想語と連想距離を図に表したものである。

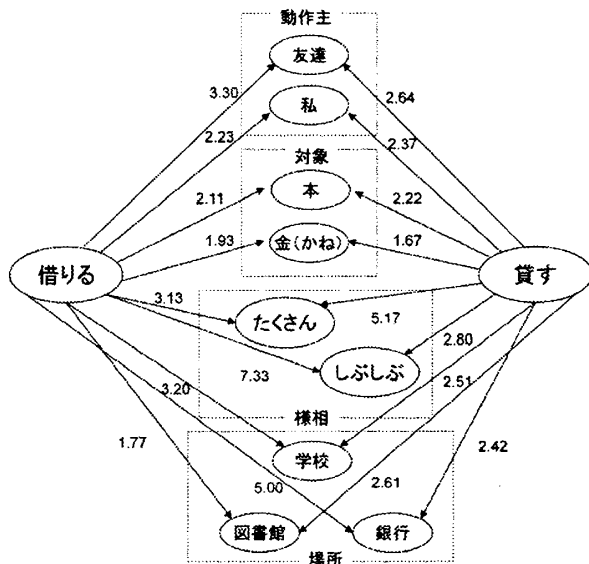


図 1 「借りる」と「貸す」の関係

図 1 から分かるように反意語の関係にある刺激語にも関わらず、「借りる」と「貸す」の「動作主」「対象」「場所」に対して同じ語が連想されており、それらの連想距離の大

小の関係も似ている。つまりこれらの動作を行う「動作主」や動作が行われる「場所」、そして動作の「対象」について被験者が共通して連想したことが示唆されている。また「様相」に注目すると“たくさん”と“しぶしぶ”が連想距離の関係が入れ替わっているのが分かるが、「様相」の内容は動作の様態や様子ということから「借りる」と「貸す」の反意語の関係を反映していると考えられる。ここで各連想課題についてそれぞれ連想距離が短い連想語を用いると「“私”は“図書館”で“本”を“たくさん”借りる」や「“友達”は“学校”で“金(かね)”を“しぶしぶ”貸す」などの文を意味合いが取れた文が作成できる。このように連想語と連想距離の関係によって、意味合いの取れた数々の場面を構成することができる。これは即ち動詞連想概念辞書が動詞と意味的な関係を持つ深層格の情報を表しており、動詞まわりの背景にある知識を反映していると考えられる。

5. 省略語補完への応用

動詞連想概念辞書の応用としてコンピュータが文脈内における省略語を補完するシステムを試作した。概要を述べると、動詞連想概念辞書を用いて補完したい内容について連想語から省略語の候補を抽出した後名詞を中心とした連想概念辞書(名詞連想概念辞書)の情報から省略語の候補の絞り込みを行い、最後に両連想概念辞書の各連想距離を用いて補完順位を示すという流れになっている。これにより、例えば「財布を家に忘れたから借りた」の文に対して「財布を家に忘れたから金(かね)を借りた」というような単純な補完が可能となる。このように連想概念辞書を用いることで文意に沿った内容を補完できる可能性がある。しかし現段階では対象となる文が非常に限られているため、その汎用性を広げるのが今後の課題として挙げられる。

6. おわりに

連想概念辞書は連想という視点で構築されたものとして他に類がなく、人間が言語を連想して文脈内容を把握するといった日常的な言語処理に対してコンピュータが同様に処理するには連想概念辞書が有効であると考えられる。実際に名詞連想概念辞書は、文書の要約に必要な重要文の抽出に関して *tfidf* 法を用いた従来の手法よりも高い精度の手法を実現している[4]。そこで、動詞連想概念辞書との併用で人間の連想に基づく処理により近付けられる可能性があり、更なる精度の向上が期待できる。このような人間の連想に基づいた視点から、コンピュータによる言語意味理解の高精度化を図っていきたい。

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, “格フレーム辞書の漸次的自動構築”, 自然言語処理, Vol.12, No.2 (2005).
- [3] 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).
- [4] 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).
- [5] 工藤拓, 松本裕治, “相対的な係りやすさを考慮した日本語係り受け解析モデル”, 情報処理学会論文誌, Vol.46, No.4(2005).