連想語の自動取得に関する研究

浅見一樹†, 杉本徹‡
†芝浦工業大学大学院　理工学研究科, ‡芝浦工業大学　工学部
{ma17005, sugimoto}@shibaura-it.ac.jp

1. はじめに

連想とは、ある事柄からそれと関連のある事柄を思い浮かべることである。これは人間が日常生活の中で自然に行う働きである。

単語から単語の連想を行う際、人間は多様な関連性をもとに単語を導き出す。例えば「春」であれば、同じく「季節」の単語である「夏、秋、冬」を連想する。しかし関連性は1つのみではなく、「春」の「行事」である「入学､花見」といった単語も同時に連想することができる。

自然言語処理分野において関連のある単語を求める手法はいくつかあり、既存の研究として概念ベース[1]やword2vec[2]を用いる手法がある。しかしこれらの手法で得られる関連のある単語は必ずしも人間が直感的に思い浮かべる単語ではないという問題がある。Deyne[3]らはこの問題に対して心理学的な実験により収集した大規模な単語連想データをもとに連想を実現する方法を提案した。

本研究では被験者から連想語を収集し、N-gram とword2vec を用いた手法を比較することで、人間の直感に近い連想語を自動取得するための手法の検討を行う。

2. 連想語の収集

2. 1. 目的

本研究では、人間が1つの単語（刺激語）から連想する単語を連想語と呼ぶ。連想語を収集する目的は、人間が直感的に連想する単語とはどのような単語であるのかを分析することである。例えば「リンゴ」と「血」という単語は同じ「赤い」という関連を持っている。しかし多くの人は「リンゴ」から「血」を連想することはないだろう。関連ある事柄を想起する連想機能によって得られる単語は、このようにただ関連があれば連想語となるわけではない。そのため実際に人間から連想語を収集し、連想語の性質を分析する必要がある。また人間が連想した連想語と自動取得した関連語の比較を行うことで、手法別に連想に対しての長所と短所を分析することができる。

2. 2. 刺激語の選定

被験者に提示する刺激語を分類語彙表[4]を参考にし、以下の条件で選定した。

・連想しやすい単語である。
・人によって連想結果が偏らない単語である。
・多義性のない単語である。
・分類語彙表内で異なる分類項目に属す。

まず上記の条件を満たす分類項目計 52 項目を選定した。選定した分類項目に属す単語のうち、予備実験において被験者2名がお互い3語以上連想出来た刺激語を合計50語になるように選定した。基本的に1分類項目につき1単語を刺激語として選定し、分類項目内の単語が他の項目より多い場合は最大2単語まで選定した。
選定した分類項目と刺激語50語を表1と表2に示す。

表1 選定した分類項目

類	部門	中項目
体の類	抽象的関係	時間
	人間活動の主体	人間
	人間活動-精神および行為	心、言語、芸術、生活、交わり、事業
	生産物および用具	資材、医療、食料、道具、機械、土地利用
	自然物および自然現象	物質、天地、植物、動物

表2　選定した刺激語

刺激語一覧				
春	通訳	警察官	宴会	天気予報
和風	雪合戦	遠足	農業	洗濯
野球	ケーキ	水着	ドライヤー	ピアノ
テレビ	新幹線	墓場	砂漠	にわとり
蝶	レモン	医者	ダイヤモンド	消しゴム
クリスマス	コンビニ	公園	留学	出張
ラジオ体操	年賀状	クレーム	交通	肥料
弁当	ラーメン	制服	わさび	うちわ
カメラ	スペースシャトル	地震	家畜	川
鏡	おもちゃ	入浴	観察	アイス

2. 3. 連想語の収集

収集する連想語は普通名詞とサ変接続名詞のみとし、被験者数は各刺激語につき 10 人とした。連想実験を行う際に被験者には各刺激語ごとに 40 秒の制限時間内で自由連想を行ってもらった。この制限は長考による関連の弱い単語の出現を抑えるために設けた。また被験者の疲労による連想語の質の低下を防ぐため、1 回の連想につき 30 秒の休憩時間を設けた。

この連想実験により収集した連想語は2,926単語であり、重複を除くと1,655単語であった。収集した連想語の一部を表3に示す。

本研究では、収集した連想語うち2人以上が共通して連想した単語を対象とし、これらの語を自動取得する手

法の検討を行う。

表3　収集された連想語の例

刺激語	3人以上が共通して連想した連想語
春	桜,花見,夏,入学,秋,冬,入学式,季節,花粉症
警察官	パトカー,逮捕,手錠,公務員,交番,正義,警棒,拳銃
遠足	おやつ,弁当,子供,お弁当,小学生,学校
テレビ	アニメ,ドラマ,番組,ニュース,芸能人,家電,バラエティ
医者	病院,白衣,手術,薬,看護師,病気,注射
ラーメン	味噌,チャーシュー,塩,醤油,とんこつ
カメラ	レンズ,写真,スマホ,スマートフォン,撮影,フィルム
ドライヤー	髪,乾燥,風呂,風,家電,熱風
洗濯	洗剤,洗濯機,水,服,乾燥機,洗濯板
ピアノ	楽器,音楽,鍵盤,楽譜,コンサート,習い事
交通	電車,車,道路,事故,信号機,飛行機,渋滞,自転車

3．　関連語の自動取得

3．1．　自動取得手法

・N-gram

本研究ではGoogle N-gram[5]の2-gramと3-gramを使用した。処理の概要を図1に示す。

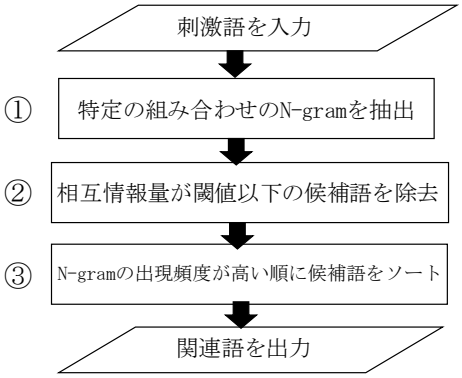

図1　N-gramを用いた関連語の自動取得

①特定の組み合わせのN-gramを抽出

表4に示す組み合わせのN-gramを抽出し、Yを関連語候補とする。

表4 抽出するN-gramの組み合わせ

	2-gram	3-gram
関連語候補Yが普通名詞の場合	刺激語,Y Y,刺激語	刺激語,X,Y Y,X,刺激語
関連語候補Yがサ変接続名詞の場合	刺激語,Y	刺激語,Z,Y
※Xは「の,と,は,や,が」のいずれかの助詞		
※Zは「で,と,は,へ,が,を,の」のいずれかの助詞		

②相互情報量が閾値以下の候補語を除去

抽出した関連語候補語のうち、刺激語との相互情報量が3.0以下の単語を候補語リストから除去する。

③N-gramの出現頻度が高い順に候補語をソート

N-gramの出現頻度順にソートし、出現頻度上位のN-gramに含まれる単語を関連語として出力する。

・2-gramと3-gramの組み合わせ手法

2-gramのみを用いて取得した関連語と3-gramのみを用いて取得した関連語の積集合を求める。その結果を3-gramの出現頻度順にソートし、出現頻度上位の3-gramに含まれる単語を関連語として出力する。

・word2vec

使用コーパスはWikipediaの全文データで、次元数200、文脈窓の大きさは20、学習モデルはskip-gramで学習を行った。学習により獲得した単語分散表現データのうち、Wikipediaコーパス内での出現頻度上位10万語を計算対象とし、各刺激語とのコサイン類似度を計算した。その結果得られる類似度上位の単語のうち、普通名詞、サ変接続名詞のみを出力とした。

以上の手法により出力した関連語の例を、収集した連想語と共に表5に示す。

表5 刺激語「カメラ」の連想語と関連語

連想語	関連語			
	2-gramのみ	3-gramのみ	2-gram&3-gram	word2vec
レンズ	デジタル	価格	撮影	シャッター
写真	好き	電話	レンズ	ファインダー
撮影	ライブ	撮影	映像	レンズ
フィルム	フィルム	レンズ	フィルム	レフ
フラッシュ	防犯	設置	搭載	乾板

3．2．　評価と考察

各手法において上位5単語に占める2人以上が連想した単語の割合を一致率とし、比較評価を行う。この際連想人数に偏りがあった単語や、刺激語が形態素解析により複数の形態素に分解される単語(天気予報、ラジオ体操)は除去した。評価結果を図2に示す。

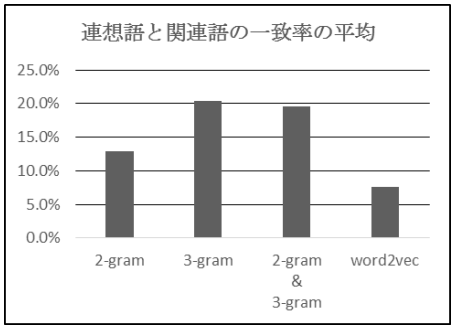

図2 連想語と関連語の一致率

・3-gramのみを用いた手法

図2よりword2vecや2-gramを用いた手法よりも、3-gramを用いた手法の方が人間が実際に連想する単語を多く出力できていることがわかる。表3で示した連想語の多くが「ピアノ、は、楽器」、「カメラ、で、撮影」といった表4に示した組み合わせで使用されることが3-gramを用いた手法の一致率が高かった理由として考えられる。

一方で、表6のように有名な映画や本のタイトルに使用されている刺激語の一致率が低くなる傾向があった。

このような単語は日常の文脈中で共起することは少ないが、Web 上で出現頻度が高い共起関係であったため、関連語として出力されたと考えられる。

表 6 3-gram を用いた手法において一致率の低かった刺激語およびその連想語と関連語

砂漠		鏡		ピアノ	
連想語	関連語	連想語	関連語	連想語	関連語
砂	月	反射	塔	楽器	種類
オアシス	王女	姿見	女王	音楽	森
ラクダ	オアシス	自分	法則	鍵盤	ため
乾燥	死	光	中	楽譜	音
サソリ	真ん中	顔	国	コンサート	先生
※下線の単語は映画や本などのタイトルに使用されている単語					

・2-gram と 3-gram の組み合わせ手法

次に一致率の高かった手法が 2-gram と 3-gram を組み合わせた手法であった。この手法は 3-gram のみを用いた手法と比較して 0.8%一致率が低かった。2-gram と 3-gram を組み合わせた手法と表 6 の関連語の比較を表 7 に示す。

表 7 2-gram と 3-gram を組み合わせた手法と 3-gram のみの手法の関連語比較

砂漠		鏡		ピアノ	
3-gramのみ	2-gram & 3-gram	3-gramのみ	2gram & 3-gram	3-gramのみ	2-gram & 3-gram
月	オアシス	塔	法則	種類	練習
王女	砂	女王	池	森	発表
オアシス	緑化	法則	挿入	ため	ヴァイオリン
死	遺跡	中	水面	音	演奏
真ん中	横断	国	観察	先生	名曲
※下線の単語は映画や本などのタイトルに使用されている単語					

表 7 より 2-gram と 3-gram を組み合わせることで、一般的に文脈中で共起することが少ない単語を取り除くことが出来ていることがわかる。しかし一方で、2-gram のみの手法の一致率がもともと低いこともあり、結果として不適切な語の出力を増やすこととなった。

・2-gram のみを用いた手法

2-gram のみの手法では 3-gram のみの手法と比べて助詞の指定による共起語の制限が行えないため、不適切な語が多く含まれる結果となった。しかし一部の単語では他の手法より一致率が高い結果を得られた。表 8 に一致率が高かった単語を示す。

表 8 2-gram を用いた手法において一致率が高かった刺激語およびその連想語と関連語

テレビ		ラーメン		ケーキ	
連想語	関連語	連想語	関連語	連想語	関連語
アニメ	番組	味噌	屋	誕生日	チーズ
ドラマ	ゲーム	チャーシュー	店	生クリーム	屋
番組	液晶	塩	醤油	スポンジ	シフォン
ニュース	ドラマ	醤油	味噌	チョコレート	チョコレート
芸能人	ケーブル	とんこつ	塩	クリスマス	クリスマス

まず表 8 より、修飾または被修飾の関係の単語が多い刺激語の場合、人間が連想する単語は修飾または被修飾の関係の単語になる傾向があることがわかる。2-gram の性質上、隣接する単語を共起語として収集するため、このような性質をもった刺激語に対して他の手法より一致率が高くなると考えられる。

・word2vec を用いた手法

N-gram を用いた手法と比べて word2vec を用いた手法の一致率が低い結果となった。先行研究[6]により word2vec を用いた手法は、刺激語とその関連語を被験者に提示した際、被験者が連想可能であると判断する単語を多く含むことがわかっている。しかし比較評価の際に最も一致率が低い結果となった。word2vec を用いた手法によって得られる関連語の例を表 9 に示す。

表 9 word2vec を用いた手法の関連語の例

ラーメン	消しゴム	ダイヤモンド	年賀状	地震
麺	ボールペン	サファイア	ダイレクトメール	余震
焼きそば	シャープペンシル	ルビー	消印	津波
うどん	文房具	宝石	礼状	本震
お好み焼き	色鉛筆	ゴールド	返信	強震
チャーシュー	鉛筆	エメラルド	メール	噴火

表 9 より、関連語として各刺激語のもつ性質と似た性質を持った単語が取得できていることがわかる。しかし表 3 より、実際は人間が連想を行う際に似たような性質の単語を連想する頻度は少ないことがわかる。word2vec の手法が連想可能な単語を多く含むにも関わらず、一致率が低かったのはこの差異のためであると考えられる。

また各手法で共通して一致率が低かった刺激語の特徴と例を表 10 に示す。

表 10 一致率の低かった刺激語の特徴と単語例

単語の特徴	単語例
抽象的な単語	春, 和風
サ変接続名詞	出張, 留学, クレーム
類義語がある単語	警察官, 墓場

抽象的な単語は多様な文脈中で出現するため、不適切な関連語も多くなったと考えられる。

サ変接続名詞は人間の連想した単語の数が、他の刺激語より少ない傾向にあった。そのため、サ変接続名詞からは連想が行いづらいと考えられる。

「警察官、警官、警察」といった類義語がある単語は、コーパス中でさまざまな表記をされるため、共起語が少なくなり、不適切な単語が増える結果になったと考えられる。

4. まとめ

本研究では人間が直感的に連想する単語の自動取得手法を検討するため、まず自由連想形式で被験者から連想語を収集した。次に N-gram と word2vec を用いて関連語

を自動取得する方法を試みて、得られた関連語を収集した連想語データと比較した。比較評価の結果 3-gram を用いた手法により人間の直感に最も近い連想語が得られることがわかった。

今後の課題としてより複雑な句の構造を指定できる 4-gram、5-gram の利用や、N-gram と word2vec の併用を試みることがあげられる。

参考文献

[1] 小島一秀, 渡辺広一, 河岡司, “連想システムのための概念ベース構成法—語間の論理的関係を用いた属性拡張” 自然言語処理 Vol.11 No.3, 2004.

[2] T. Mikolov, K. Chen, G. Corrado, and J. Dean , “ Efficient Estimation of Word Representations in Vecter Space” In ICLR Workshop. 2013.

[3] Simon De Deyne, Amy Perfors, and Daniel J.Navarro, “Predicting Human Similarity Judgments with Distributional Models: The Value of Word Associations, ” COLING, 2016.

[4] 国立国語研究所: 分類語彙表-増補改訂版, 大日本図書, 2004

[5] 工藤拓,賀沢秀人著 : Web 日本語 N グラム第 1 版, 言語資源協会発行, 2007

[6] 浅見一樹, 杉本徹, “雑談対話システムにおいて話題転換に用いる連想語の取得に関する研究”信学技報 Vol.117, No.82, NLC2017-11, pp.59-64, 2017.