

2022年度（令和4年度） 卒業論文

Word2vecを用いたより格調が高い単語の  
判定

指導教員

村田真樹  
村上仁一

鳥取大学工学部 電気情報系学科

自然言語処理研究室

B18T2057C 汐見 和泰

## 概要

公の場でのコミュニケーションにおいて、その場に適した格調が高い言葉を使用することは重要である。しかし、敬語を正しく使うことに苦手意識を感じている人も多い。そのため、自動で格調が高い言葉遣いを選択することは有意義である。

本研究では、格調が高い単語を判定する手法として、同じ意味を持つ言葉なら格調が高い単語ほうが、より他の格調が高い単語と意味が近くなるという仮説に基づいて実験を行った。いくつかの言い換え単語から、Word2vec を用いて格調が高い単語を判定した。Word2vec は分布仮説に基づき学習を行う。同じ文脈に出現する語彙が似たベクトルを得られることが分かっている。この性質を利用し、分散表現からコサイン類似度を利用して、同義語単語群からどの単語が予め格調が高いとわかっている単語との類似度が高く、格調が高いのかを判定した。

結果は、単語対 277 組中の 192 組を正しく判定でき、判定精度は 0.693 という結果になった。概ね基本的に設定しておく単語対のデータが多ければ多いほどより正答率が上がるという結果になった。予め決めておく単語対を 1 単語対から約 100 単語対まで増やすと、正答率が 0.051 上がった。しかし、約 50 単語対設定から約 100 単語対、約 200 単語対設定での場合で、約 200 単語対設定が最も正答率が高かったものの、正答率にほとんど差は見られなかった。

この手法のみで文の言い換えをすると、実際の言い換えの際は文中の前後との整合性も考慮する必要がある。例えば「姉が妊娠した」という文が「姉が身重した」のように変換され、非文が生成されてしまう。しかし、この手法は大量の文章から格調が高い文を抜き出す場合などに有用だと考えられる。機械学習ではネットなどから大量の文を学習させることが多いが、その大量の文を抜き出す行程が難しい場合もある。そのとき、機械に格調が高い単語を自動判定させることができれば文章収集が容易になるなどの活用がある。

# 目次

|     |                            |    |
|-----|----------------------------|----|
| 第1章 | はじめに                       | 1  |
| 第2章 | 先行手法                       | 3  |
| 第3章 | 提案手法                       | 4  |
| 3.1 | 手法 . . . . .               | 4  |
| 3.2 | Word2vec . . . . .         | 7  |
| 第4章 | 実験                         | 8  |
| 4.1 | 実験方法 . . . . .             | 8  |
| 4.2 | 実験条件 . . . . .             | 9  |
| 4.3 | 実験結果 . . . . .             | 9  |
| 4.4 | 様々な単語対での一対のみでの実験 . . . . . | 21 |
| 第5章 | 考察                         | 22 |
| 第6章 | 今後の展開                      | 24 |
| 第7章 | おわりに                       | 25 |

# 表 目 次

|      |                          |    |
|------|--------------------------|----|
| 4.1  | 5分割クロスバリデーショングループ1 (その1) | 10 |
| 4.2  | 5分割クロスバリデーショングループ1 (その2) | 11 |
| 4.3  | 5分割クロスバリデーショングループ2 (その1) | 12 |
| 4.4  | 5分割クロスバリデーショングループ2 (その2) | 13 |
| 4.5  | 5分割クロスバリデーショングループ3 (その1) | 14 |
| 4.6  | 5分割クロスバリデーショングループ3 (その2) | 15 |
| 4.7  | 5分割クロスバリデーショングループ4 (その1) | 16 |
| 4.8  | 5分割クロスバリデーショングループ4 (その2) | 17 |
| 4.9  | 5分割クロスバリデーショングループ5 (その1) | 18 |
| 4.10 | 5分割クロスバリデーショングループ5 (その2) | 19 |
| 4.11 | 5分割クロスバリデーション結果          | 20 |
| 4.12 | 少ない設定での実験結果              | 20 |
| 4.13 | 様々な単語対での一対のみでの実験         | 21 |
| 5.1  | 良い結果が出た対の例               | 22 |
| 5.2  | 悪い結果が出た対の例               | 22 |

# 図目次

|                          |   |
|--------------------------|---|
| 3.1 仮説概念図 . . . . .      | 5 |
| 3.2 提案手法を説明する図 . . . . . | 6 |

# 第1章 はじめに

公の場でのコミュニケーションにおいて、その場に適した格調が高い言葉を使用することは重要である。例えば、ビジネスでのメールなどで正しく敬語を使えないと困る場面がある。しかし、敬語を正しく使うことに苦手意識を感じている人も多く、また近年読書率の低下で語彙の低下も問題視されている。そのため、自動で格調が高い言葉遣いを選択することができれば様々な場面で役立つ。

格調が高い単語とは、単語のもつ品格などが高いことである。言葉が上品であったり、気高いことである。その言葉を使っているほうが、より優雅な印象を持たせる言葉使いであるとする。例えば、ビジネスなどで使う文としては「行く」に比べて、「伺う」のほうがより礼儀正しいとみなされる。また、「物知り」や「面倒くさい」、「いいかげん」と言うよりも、「博識」や「わずらわしい」、「なおざり」と言う方がより大人らしい。

本研究では、言い換え単語をいくつか取得し、そこから Word2vec を用いて格調が高い単語を判定する方法を提案する。まず、同じ意味を持つ言葉であれば、より格調が高い単語の方が他の格調が高い単語と意味的に近くなるという仮説を立てる。その仮説のもと、Word2vec を用いて単語同士の類似度から格調高さを判定する手法を提案する。Word2vec は、教師なしテキストデータから、単語の文法的、意味的な情報を学習し、数 100 次元のベクトルで表現できる。

本研究の主な主張点を以下に整理する。

- 同じ意味を持つ言葉であれば、より格調が高い単語の方が他の格調が高い単語と意味的に近くなるという仮説を立てる。
- 上記仮説の信頼性を調べるため実際のデータを用いて実験を行った。正解率は 0.693 だった。
- この手法が発展すれば、格調が高い文章の収集が容易になるなどの活用がある。そのように文書収集が容易になると、他の用途などでも活用ができる。

本論文の構成は以下のとおりである。第2章では、本研究に関連する研究を記述する。第3章では、本研究の提案手法を記述する。第4章では、本研究が行った実験についての説明と、その結果について記述する。第5章では、結果から考察を行う。第6章では、今後の課題について記述する。

## 第2章 先行手法

単語同士の関連性を調べる方法として、共起を用いた手法がある。Turney[1]はコーパスから得られる共起情報を用いて単語を評価した。ある評価表現  $t$  の評価極性値  $SO-score(t)$  は式 (2.1) で算出される。

$$SO-score(t) = PMI(t, "excellent") - PMI(t, "poor") \quad (2.1)$$

ここで、PMI(pointwise mutual information)(Church and Hanks 1989) は2つの語句間の共起の尺度で、任意の語句  $a, b$  の PMI は式 (2.2) で算出される。

$$PMI(a, b) = \log_2 \frac{p(a, b)}{p(a)p(b)} \quad (2.2)$$

つまり、語句  $t$  が肯定極性を示す代表的な語「excellent」と共起しやすければ  $SO-score(t)$  は正の値となり、否定極性を示す代表的な語「poor」と共起しやすいと負の値となる。

また、類似研究として、リフレーミングに関する研究がある。リフレーミングとは、ネガティブな表現をポジティブな枠組みで見ることであり、言い換え表現が生じる。河野ら [2] は、リフレーミングの事例をウェブなどから人手で収集し、リフレーミング生成モデルを学習するためのパラレルコーパスを構築した。そして、作成したパラレルコーパスと収集した対話データを用いて事前学習済みの大規模言語モデルをファインチューニングすることで、リフレーミング生成モデルを学習する研究を行った。

## 第3章 提案手法

### 3.1 手法

手法の概要として、いくつかの言い換え単語群から Word2vec を用いて格調が高い単語を判定するという手法である。

まず、同じ意味を持つ言葉であれば、より格調が高い単語の方が他の格調が高い単語と意味的に近くなるという仮説を立てる。その仮説の概念図を図 3.1 に示す。

Word2vec を利用して得られた単語の分散表現から、コサイン類似度を用いる。コサイン類似度は、ベクトル  $a$ ,  $b$  に対して以下の式で得られる。

$$\cos(a, b) = \frac{a \cdot b}{\|a\| \|b\|} \quad (3.1)$$

Word2vec は分布仮説に従い学習を行うため、似た文脈で出現する語彙との類似度は高くなる。この性質を利用する。

具体的な手法手順は以下の通りである。

1. 格調の高低が予めわかっている、Word2vec 上に存在する単語対をある程度用意する。
2. 格調の高低がわかっていない、Word2vec 上に存在する判定したい単語群と用意した単語対とのそれぞれの類似度を計算する。
3. 格調が高い単語との類似度から格調が低い単語との類似度を引き、その値をその単語の優先度とする。
4. 判定したい単語群の中で、最も優先度が高かった単語が、最も格調が高い単語とする。

ある判定単語  $t$  の優先度は式 (3.2) で算出する。

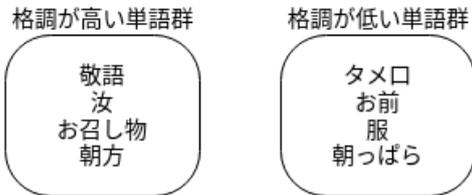
$$\text{優先度}(t) = \sum \{(\cos(t, \text{格調が高い語句})) - (\cos(t, \text{格調が低い語句}))\} \quad (3.2)$$

手法の内容を図示したものを図 3.2 に示す。

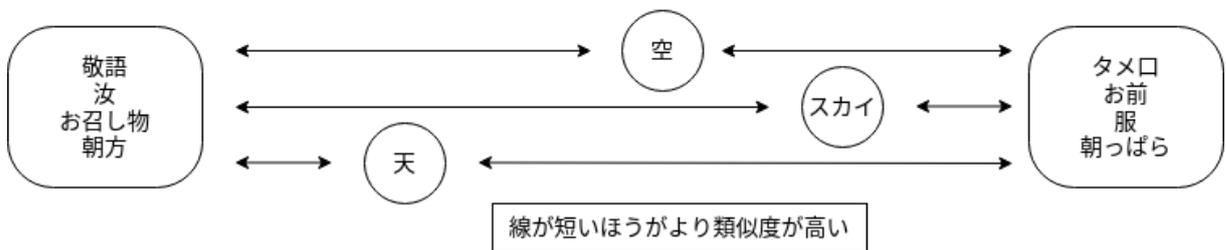


図 3.1: 仮説概念図

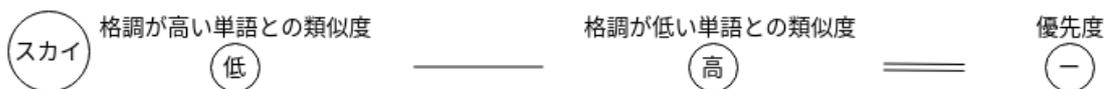
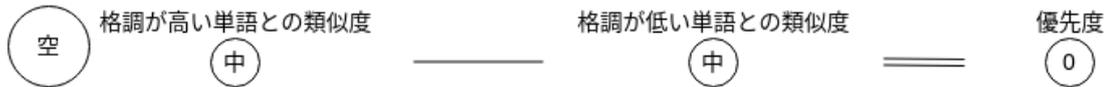
1 予め格調の高低が判明している単語対を用意する



2 判定したい単語群のそれぞれの単語との類似度を計算する



3 格調が高い単語との類似度から格調が低い単語との類似度を引く、それを優先度とする



4 優先度が最も高かった単語が最も格調が高い単語となる



図 3.2: 提案手法を説明する図

## 3.2 Word2vec

Word2vec は Google 社の Tomas Mikolov ら [3] によって提案されたニューラルネットワーク (Skip-gram) の手法で、コーパスから単語の分散表現を得る手法である。分散表現とは、単語の意味を固定長のベクトルで表現することである。単語をベクトル空間内に埋め込むことで、ベクトル間の距離や演算を利用することができる。Word2vec は、分布仮説という、「意味の似ている単語は似た文脈で出現する」という考えに基づき、単語の分散表現を学習する。この時、周辺語彙からある単語を推測する方法で学習するモデルを CBoW (continuous bag-of-words)、ある単語から周辺語彙を推測するモデルを skip-gram と言い、どちらも 2 層のニューラルネットワークである。Word2vec で学習した分散表現は、似た意味を持つ単語のベクトルは似たベクトルになり、 $king - man + woman = queen$  といった加法性があることも知られている。これらの性質から、単語の意味の類似性や関連性を測定したり、加法構成性の評価タスクなどに利用される。

# 第4章 実験

## 4.1 実験方法

提案手法がどの程度信頼できるのかを調べるため、予め格調の高低が判明している語彙の対を使って実験を行い、元のデータと上記手法の結果の一致率を調べる。

格調が高い語彙と格調が低い語彙の対は書籍「大人の言い換えハンドブック」[4]より作成した。今回の実験では言い換えの書籍より格調の高低が判明している単語を入手するため、言い換えの単語群から判定する提案手法と異なり、判定に用いる語彙も2単語の対から判定することになる。提案手法と実験では条件が異なるが、提案手法は単語群が2単語しかない場合も想定され、また、2単語対で正しく判定可能ならば、それよりも単語数が多い場合でも判定が可能と予想される。そのため、今回の実験は手法の精度を調べる上で有効である。提案手法より、Word2vec は分布仮説に従い学習を行うため、似た文脈で出現する語彙との類似度は高くなる性質を利用し、書籍と判定結果が一致するかを調査する。

実験には5分割クロスバリデーションを用いる。具体的な実験手順は、以下の通りである。

1. 入手した単語対を5つのグループにわけるとする。
2. そのうちの1グループを判定する単語対とする。
3. 残りの4グループ全てを予めわかっている単語対に設定する。
4. 単語対の両方に対して判定を行い、優先度を算出する。
5. 優先度より、どちらがより格調が高いかを比較する。具体的には、単語対の格調が高い方から格調が低い方の優先度の値を引き、正の値なら書籍と結果が一致、負の値なら書籍と結果が不一致となる。
6. グループを変えて、これを5回行う。

## 4.2 実験条件

本研究で使用する Word2vec のモデルは、Python のオープンソースライブラリである gensim[5] を利用して作成する。学習データは日本語 Wikipedia より作成した。モデルは skip-gram であり、vector\_size は 100 とし、window は 10 として、学習を行った。

書籍より入手した語彙の対の数は 734 対であったが、その中から両方の語彙が Word2vec 内に存在する対のみに絞った結果、277 対となった。それをランダムにシャッフルし、5 分割した。それにより、1 グループ内の単語対の数は 55, 56 単語対の約 50 単語対。4 グループ使用した場合は 221, 222 単語対の約 200 単語対となった。

## 4.3 実験結果

5 分割クロスバリデーションで行った実験結果を表 4.1～表 4.10 に示す。1 グループ目を残りの 4 グループで判定した結果を表 4.1, 表 4.2 に示す。同様に、2, 3, 4, 5 グループを判定した結果についてもそれぞれ表 4.3, 表 4.4, 表 4.5, 表 4.6, 表 4.7, 表 4.8, 表 4.9, 表 4.10 に示す。

表 1 列目の格調が低い単語とは、書籍で言い換えられる前の単語であり、表 2 列目の格調が高い単語とは書籍にて言い換えられた後の単語である。表 3 列目の優先度差とは、単語対の格調が高い単語の優先度の合計に格調が低い単語の優先度の合計を引いた差である。表 4 列目の評価とは、判定結果が書籍と一致したか否かについて示す。○なら一致、×なら不一致とする。

表 4.1: 5分割クロスバリデーション グループ1 (その1)

| 格調が低い単語  | 格調が高い単語  | 優先度差   | 評価 |
|----------|----------|--------|----|
| 利口       | 聡明       | 7.60   | ○  |
| 楽しみ      | 醍醐味      | 5.92   | ○  |
| とんでもない   | 以ての外     | 9.05   | ○  |
| バランス     | 兼ね合い     | 11.02  | ○  |
| 早とちり     | 早合点      | 1.32   | ○  |
| みなさん     | 皆様       | 7.60   | ○  |
| あさる      | 探す       | -1.49  | ×  |
| だいたい     | 総じて      | 8.01   | ○  |
| 付け込む     | 乗じる      | 9.65   | ○  |
| 猿も木から落ちる | 弘法にも筆の誤り | 2.08   | ○  |
| 屁理屈      | 理屈       | 4.45   | ○  |
| 消極的      | 慎ましい     | -10.71 | ×  |
| 生意気      | 小憎らしい    | 9.00   | ○  |
| 終わり      | お開き      | -4.33  | ×  |
| 面汚し      | 名折れ      | 2.72   | ○  |
| 心得       | たしなみ     | -1.39  | ×  |
| 軽率       | 不用意      | -0.89  | ×  |
| 悪妻       | 恐妻       | -2.65  | ×  |
| 適当       | 適切       | 3.75   | ○  |
| 飲んべえ     | 酒豪       | 2.39   | ○  |
| 贈る       | 差し上げる    | -3.98  | ×  |
| 控えめ      | しおらしい    | -7.28  | ×  |
| コネ       | つて       | 2.99   | ○  |
| ほじくる     | えぐる      | 5.94   | ○  |
| 育ち       | 生い立ち     | -3.57  | ×  |
| 塗りたい     | 塗りつける    | -0.01  | ×  |
| 朝っぼら     | 朝方       | 13.03  | ○  |
| 恥        | 名折れ      | 5.14   | ○  |
| まるで      | あたかも     | 10.27  | ○  |
| 食い物      | 食べ物      | -2.21  | ×  |
| 嫌う       | 煙たがる     | 1.92   | ○  |
| 得意       | お手の物     | -4.91  | ×  |
| かわいがる    | いつくしむ    | 10.60  | ○  |
| なまじっか    | なまじ      | -2.01  | ×  |
| 妊娠       | おめでた     | -3.52  | ×  |

表 4.2: 5分割クロスバリデーション グループ1 (その2)

| 格調が低い単語 | 格調が高い単語 | 優先度差   | 評価 |
|---------|---------|--------|----|
| つらい     | やるせない   | 5.27   | ○  |
| 棺桶      | 棺       | 7.72   | ○  |
| すごい     | すばらしい   | 8.09   | ○  |
| 最中      | たけなわ    | 6.22   | ○  |
| 火葬場     | 斎場      | 3.48   | ○  |
| とにかく    | ともあれ    | 17.06  | ○  |
| 失敗      | 不首尾     | 7.22   | ○  |
| 臨月      | 産み月     | 1.91   | ○  |
| 中途半端    | 不徹底     | 13.27  | ○  |
| プライベート  | 私事      | 9.56   | ○  |
| 最高      | 理想的     | -4.10  | ×  |
| とっちめる   | 懲らしめる   | -1.56  | ×  |
| チップ     | 心付け     | -0.84  | ×  |
| へばる     | 疲れる     | -10.65 | ×  |
| 猿真似     | 物真似     | -4.31  | ×  |
| 隠す      | 伏せる     | 6.69   | ○  |
| とんずら    | 逃亡      | 4.03   | ○  |
| 誤り      | 誤謬      | 0.84   | ○  |
| まさか     | よもや     | 11.87  | ○  |
| 天気      | 雲行き     | -0.20  | ×  |
| 様子      | たたずまい   | 11.49  | ○  |

表 4.3: 5分割クロスバリデーション グループ2 (その1)

| 格調が低い単語 | 格調が高い単語 | 優先度差  | 評価 |
|---------|---------|-------|----|
| だまし討ち   | 不意打ち    | -4.68 | ×  |
| 敵       | 好敵手     | 8.01  | ○  |
| もっばら    | ひとえに    | -3.00 | ×  |
| 体つき     | 背格好     | -5.01 | ×  |
| 出発      | お立ち     | -0.81 | ×  |
| 取り繕う    | 弥縫策     | 12.40 | ○  |
| 仮に      | よしんば    | -1.07 | ×  |
| 臆病      | 慎重      | 11.05 | ○  |
| 札びら     | お札      | 4.90  | ○  |
| 理屈っぽい   | 理詰め     | 9.77  | ○  |
| 変わり者    | 変わり種    | 5.64  | ○  |
| 落第      | 不合格     | -1.16 | ×  |
| そっくり    | 生き写し    | 7.54  | ○  |
| のろい     | 緩慢      | 6.72  | ○  |
| 静寂      | しじま     | -4.01 | ×  |
| 照れる     | 気恥ずかしい  | -4.47 | ×  |
| わたし     | わたくし    | 6.36  | ○  |
| 物知り     | 博識      | 9.54  | ○  |
| 下手      | つたない    | -8.05 | ×  |
| 匂う      | 香る      | 2.45  | ○  |
| 損害      | 不利益     | 3.14  | ○  |
| 大げさ     | 仰々しい    | 6.96  | ○  |
| で       | にて      | 3.70  | ○  |
| 食う      | いただく    | 4.49  | ○  |
| 気持ち     | 微意      | 19.57 | ○  |

表 4.4: 5分割クロスバリデーション グループ 2 (その 2)

| 格調が低い単語 | 格調が高い単語 | 優先度差  | 評価 |
|---------|---------|-------|----|
| いつまでも   | 幾久しく    | 4.97  | ○  |
| 共稼ぎ     | 共働き     | -4.53 | ×  |
| 出身地     | お国      | 4.10  | ○  |
| 日没      | 日の入り    | 1.79  | ○  |
| 延期      | 日延べ     | 4.36  | ○  |
| ぴかー     | 白眉      | 9.17  | ○  |
| 末期      | 終期      | 2.76  | ○  |
| 醜い      | 見苦しい    | 3.35  | ○  |
| 配偶者     | 連れ合い    | -3.92 | ×  |
| だんだんに   | 次第に     | 2.85  | ○  |
| 上品      | たおやか    | -0.35 | ×  |
| 腹這い     | うつ伏せ    | -2.19 | ×  |
| 無能      | 非力      | 2.27  | ○  |
| 馬鹿馬鹿しい  | 片腹痛い    | 3.51  | ○  |
| 氏素性     | 身元      | -3.54 | ×  |
| のっぽ     | 長身      | 6.90  | ○  |
| 露骨      | あからさま   | -2.67 | ×  |
| ひょろ長い   | 細長い     | 6.47  | ○  |
| 計画      | 腹づもり    | 0.96  | ○  |
| 手引き     | 手ほどき    | -4.48 | ×  |
| へっぴり腰   | 及び腰     | 8.94  | ○  |
| 叱る      | 諭す      | 5.52  | ○  |
| しこたま    | 多い      | 2.27  | ○  |
| 墓場      | 霊園      | 9.48  | ○  |
| 盗み見る    | かいま見る   | 7.47  | ○  |
| 甘ったれる   | 甘える     | -2.36 | ×  |
| せわしない   | 慌ただしい   | 0.52  | ○  |
| へこむ     | 落ち込む    | -0.62 | ×  |
| お受け取り   | お納め     | 9.50  | ○  |
| 皮算用     | 目算      | 2.71  | ○  |
| へこたれる   | 屈する     | 12.57 | ○  |

表 4.5: 5分割クロスバリデーション グループ3 (その1)

| 格調が低い単語 | 格調が高い単語 | 優先度差  | 評価 |
|---------|---------|-------|----|
| 豚に真珠    | 猫に小判    | -2.70 | ×  |
| どれほど    | いかほど    | 0.86  | ○  |
| 甘ったるい   | 甘い      | -2.76 | ×  |
| 考える     | 鑑みる     | 18.22 | ○  |
| 言いつけ    | 仰せ      | 11.22 | ○  |
| 裸足      | 素足      | 2.24  | ○  |
| 直接      | 直々      | 7.91  | ○  |
| 言いぐさ    | 言い分     | 3.25  | ○  |
| 泥臭い     | 土臭い     | 3.83  | ○  |
| 間違い     | 行き違い    | 10.30 | ○  |
| 目つき     | まなざし    | 3.21  | ○  |
| 足       | おみ足     | 1.34  | ○  |
| あのとき    | 先般      | 8.96  | ○  |
| デマ      | 風説      | 12.73 | ○  |
| 荒唐無稽    | 夢物語     | -3.98 | ×  |
| 朝晩      | 朝夕      | 2.64  | ○  |
| どす黒い    | 浅黒い     | -0.56 | ×  |
| そして     | 次に      | 1.59  | ○  |
| 雪景色     | 銀世界     | -3.55 | ×  |
| これから    | 今後      | 10.98 | ○  |
| 性格      | 気立て     | -2.35 | ×  |
| 健康      | すこやか    | -0.79 | ×  |
| 誘う      | いざなう    | 6.03  | ○  |
| 無駄骨     | 徒労      | 2.87  | ○  |
| 妊娠      | 身重      | 4.97  | ○  |
| 目立つ     | 水際立つ    | 4.60  | ○  |
| やけくそ    | 自暴自棄    | 3.48  | ○  |
| 真っ昼間    | 昼日中     | 9.45  | ○  |
| がつつく    | 食べる     | -3.01 | ×  |
| へま      | ポカ      | 0.17  | ○  |
| 金繰り     | 資金繰り    | 5.03  | ○  |

表 4.6: 5分割クロスバリデーション グループ3 (その2)

| 格調が低い単語 | 格調が高い単語   | 優先度差  | 評価 |
|---------|-----------|-------|----|
| ご苦労さま   | お世話になりました | -1.91 | ×  |
| 口出し     | 口添え       | 8.32  | ○  |
| どける     | のける       | -2.02 | ×  |
| だいたい    | おおむね      | 7.19  | ○  |
| 一日中     | ひねもす      | -5.38 | ×  |
| なくなる    | 尽きる       | -2.51 | ×  |
| 借金      | 借財        | 9.09  | ○  |
| しかし     | しかしながら    | -0.82 | ×  |
| 腕っこき    | 腕利き       | 0.33  | ○  |
| ちゃち     | 貧弱        | 9.47  | ○  |
| 祝う      | 寿ぐ        | 4.09  | ○  |
| かわいそう   | 痛ましい      | 11.67 | ○  |
| うるさ型    | 論客        | 1.22  | ○  |
| うろつく    | ぶらつく      | -0.24 | ×  |
| おっかない   | 恐ろしい      | 3.41  | ○  |
| いかさま    | 不正        | 7.82  | ○  |
| 役立てる    | 資する       | 4.99  | ○  |
| 細君      | 奥様        | -7.94 | ×  |
| えこひいき   | 身びいき      | 6.52  | ○  |
| まもなく    | ほどなく      | 4.71  | ○  |
| 奥さん     | 奥様        | 5.72  | ○  |
| その      | 当該        | 8.94  | ○  |
| 必ずしも    | あながち      | -2.82 | ×  |
| 顔見知り    | 顔なじみ      | 0.78  | ○  |

表 4.7: 5分割クロスバリデーション グループ4 (その1)

| 格調が低い単語 | 格調が高い単語 | 優先度差  | 評価 |
|---------|---------|-------|----|
| 飯       | ご飯      | -6.86 | ×  |
| 地べた     | 地面      | 1.88  | ○  |
| 敵       | ライバル    | -1.02 | ×  |
| そのうえ    | あまつさえ   | -5.43 | ×  |
| いまいち    | 今ひとつ    | 6.09  | ○  |
| 情けない    | 嘆かわしい   | 5.90  | ○  |
| 忙しい     | 多忙      | 6.95  | ○  |
| 焼き直し    | 改作      | 4.65  | ○  |
| 鈍い      | マイペース   | -6.67 | ×  |
| 夕方      | たそがれ    | -6.65 | ×  |
| 悪臭      | 異臭      | -1.20 | ×  |
| 過保護     | 温室育ち    | 2.21  | ○  |
| 夜       | 夜分      | 10.29 | ○  |
| さえ      | すら      | 3.83  | ○  |
| 身持ち     | 素行      | 5.28  | ○  |
| 老人      | 高齢者     | 6.89  | ○  |
| おべっか    | お世辞     | -2.64 | ×  |
| 愚痴      | 繰り言     | 16.04 | ○  |
| 外人      | 外国人     | 3.50  | ○  |
| 案       | 叩き台     | 1.54  | ○  |
| やっぱり    | やはり     | 7.67  | ○  |
| やかましい   | かまびすしい  | 16.32 | ○  |
| 自分勝手    | 恣意的     | 14.20 | ○  |
| 迷う      | ためらう    | -0.53 | ×  |
| 出費      | 物入り     | -4.55 | ×  |
| 風変わり    | 型破り     | 1.37  | ○  |
| 哀れ      | 不憫      | 0.46  | ○  |
| 野暮ったい   | 無粋      | -2.25 | ×  |
| いっそう    | ひとしお    | -6.60 | ×  |
| 手紙      | 封書      | 9.27  | ○  |
| 厚ぼったい   | 厚手      | 4.39  | ○  |
| おぼこい    | うぶ      | 4.36  | ○  |
| 反対      | 不賛成     | -2.21 | ×  |
| 服       | お召し物    | 7.76  | ○  |

表 4.8: 5分割クロスバリデーション グループ 4 (その 2)

| 格調が低い単語 | 格調が高い単語 | 優先度差  | 評価 |
|---------|---------|-------|----|
| 生まれつき   | 生まれながら  | 1.67  | ○  |
| とか      | やら      | 1.26  | ○  |
| 死亡      | 逝去      | 11.17 | ○  |
| 不満      | 不本意     | 4.04  | ○  |
| ばらまく    | ふりまく    | -4.95 | ×  |
| 新米      | 新進      | 10.63 | ○  |
| 地肌      | 素肌      | -4.58 | ×  |
| かっぱらう   | 掠める     | 6.66  | ○  |
| ひよろひよろ  | ほっそり    | 2.11  | ○  |
| 漬物      | 香の物     | -0.48 | ×  |
| 裏道      | 裏通り     | -7.53 | ×  |
| いいかげん   | なおざり    | 13.11 | ○  |
| お釣り     | お返し     | -6.96 | ×  |
| 無愛想     | 無口      | 1.19  | ○  |
| 書く      | 認める     | 8.02  | ○  |
| ダメ      | 不適當     | 24.25 | ○  |
| ひどく     | 完膚なきまでに | 5.91  | ○  |
| ほっぺ     | 頬       | 9.05  | ○  |
| 体       | 御身      | 9.53  | ○  |
| 古株      | ベテラン    | -6.81 | ×  |
| 目糞鼻糞を笑う | 五十歩百歩   | 0.62  | ○  |

表 4.9: 5分割クロスバリデーション グループ5 (その1)

| 格調が低い単語        | 格調が高い単語 | 優先度差  | 評価 |
|----------------|---------|-------|----|
| 会う             | お目もじ    | 5.51  | ○  |
| ひじょうに          | 甚だ      | 7.57  | ○  |
| ちゃんちゃらおか<br>しい | 笑止千万    | -2.52 | ×  |
| やる             | 行う      | 12.02 | ○  |
| 中年             | 壮年      | 12.55 | ○  |
| なんとか           | かろうじて   | 8.58  | ○  |
| もちろん           | もとより    | 15.31 | ○  |
| 騒音             | ざわめき    | -6.48 | ×  |
| 目指す            | 期する     | 13.83 | ○  |
| 話す             | 語らう     | 9.45  | ○  |
| 乗っかる           | 乗る      | 2.71  | ○  |
| ちょっと           | 少々      | 9.94  | ○  |
| うれしい           | 喜ばしい    | 8.98  | ○  |
| 怪しい            | 不可解     | 6.48  | ○  |
| 馬鹿笑い           | 大笑い     | -3.89 | ×  |
| とぼっちり          | しわ寄せ    | 2.47  | ○  |
| 考え             | 所存      | 12.72 | ○  |
| 足代             | 交通費     | -2.03 | ×  |
| さま             | 様       | -3.50 | ×  |
| 考える            | 勘案      | 20.40 | ○  |
| 読む             | 繙く      | 12.42 | ○  |
| 膝小僧            | 膝頭      | 1.41  | ○  |
| 思いつき           | ひらめき    | -3.11 | ×  |
| 愚痴る            | こぼす     | -0.11 | ×  |
| きのう            | 昨日      | -2.78 | ×  |
| 素人             | アマチュア   | 4.27  | ○  |
| ときどき           | 時として    | 8.50  | ○  |
| 確かめる           | 見定める    | 4.39  | ○  |

表 4.10: 5分割クロスバリデーション グループ5 (その2)

| 格調が低い単語  | 格調が高い単語 | 優先度差  | 評価 |
|----------|---------|-------|----|
| くだもの     | 水菓子     | 4.01  | ○  |
| すごく      | いたく     | 8.09  | ○  |
| せびる      | 無心する    | 2.56  | ○  |
| 何度も      | たびたび    | 1.84  | ○  |
| 薄っぺら     | 薄手      | 9.97  | ○  |
| 面倒くさい    | わずらわしい  | 7.64  | ○  |
| 下品       | はしたない   | 2.66  | ○  |
| こうした     | かかる     | -3.03 | ×  |
| いつか      | いずれ     | 6.86  | ○  |
| のような     | ごとき     | -5.58 | ×  |
| どっこいどっこい | 互角      | 9.70  | ○  |
| 口喧嘩      | 口論      | 4.21  | ○  |
| 若い       | 若々しい    | 10.29 | ○  |
| 思いがけなく   | たまさか    | 5.53  | ○  |
| 日常生活     | 起き伏し    | -2.56 | ×  |
| 髪        | 御髪      | 8.39  | ○  |
| 自腹       | ポケットマネー | 0.19  | ○  |
| いちゃもん    | 難癖      | -0.89 | ×  |
| 初心者      | ビギナー    | 0.01  | ○  |
| 万歳       | 万々歳     | -1.67 | ×  |
| しかめっ面    | 渋面      | 5.16  | ○  |
| また       | ならびに    | 7.84  | ○  |
| 古狐       | 古狸      | 3.16  | ○  |
| いじる      | もてあそぶ   | 1.21  | ○  |
| 左うちわ     | 悠々自適    | 3.74  | ○  |
| 人だかり     | 人波      | 3.96  | ○  |
| 大雑把      | 大まか     | 2.62  | ○  |

比較実験として、予め決めて置く語句を4グループから半分の2グループの110, 111単語対の約100単語対に減らした実験。設定する単語対を「敬語」と「タメ口」の一对のみとした実験も行った。

約200単語対を設定した場合も含め、実験の正解数を表4.11に示す。

表 4.11: 5分割クロスバリデーション結果

|                 | 1  | 2  | 3    | 4    | 5  | 正解数計 | 正解率   |
|-----------------|----|----|------|------|----|------|-------|
| 約200単語対設定       | 36 | 38 | 39   | 37   | 42 | 192  | 0.693 |
| 約100単語対設定<br>平均 | 39 | 38 | 38.5 | 34.5 | 40 | 190  | 0.686 |
| 敬語, タメ口のみ       | 37 | 33 | 31   | 37   | 38 | 176  | 0.635 |
| 全体の対の数          | 56 | 56 | 55   | 55   | 55 | 277  |       |

表4.11の、列の1~5はクロスバリデーションで分割したグループごとの結果である。約100単語対設定は別のデータで二回行い、平均を取っている。

さらに少ない条件での実験として、5分割した1グループのみの約50単語対のみを予め設定しておく単語対として実験した。また、277単語対を10分割し、そのうちの一つのグループである27, 28単語対の約25単語対のみを予め設定しておく単語対とした場合の実験も行った。結果を表4.12に示す。

表 4.12: 少ない設定での実験結果

|             | 正解数計  | 正解率   |
|-------------|-------|-------|
| 約50単語対設定の平均 | 191.5 | 0.691 |
| 約25単語対設定の平均 | 165.7 | 0.664 |

表4.11, 表4.12より、約50単語対を設定した場合の正解率が約100単語対を設定した場合よりも高かったが、基本的に設定しておく単語対のデータが多ければ多いほどより正解率が上がるという結果になった。単語対を一つから約100単語対まで増やすと、正答率が0.051上がり、約25単語対設定から約50単語対設定に増やすと正答率が0.027上がった。しかし、約50単語対設定から約100単語対, 約200単語対設定の場合で、約200単語対設定が最も正答率が高かったものの、正答率にほとんど差は見られなかった。

## 4.4 様々な単語対での一対のみでの実験

今度は様々な単語対一対のみを設定する場合においても実験を行った。結果を表 4.13 に示す。

表 4.13: 様々な単語対での一対のみでの実験

|                | 正解数計 | 正解率   |
|----------------|------|-------|
| 敬語, タメ口のみ      | 176  | 0.635 |
| 聡明, 利口のみ       | 118  | 0.534 |
| 醍醐味, 楽しみのみ     | 111  | 0.502 |
| 以ての外, とんでもないのみ | 140  | 0.633 |
| 兼ね合い, バランスのみ   | 113  | 0.511 |
| 早合点, 早とちりのみ    | 136  | 0.615 |
| 皆様, みなさんのみ     | 97   | 0.439 |
| 全体の対の数         | 277  |       |

単語対によって正解率は様々であるが、いずれの場合においても「敬語」「タメ口」のみを設定した場合よりも低い正解率となり、約 25 単語対を使用した際の平均正解率よりも低い値となった。

## 第5章 考察

考察のため、実験結果より良い結果が出た中から例を表 5.1 に、悪い結果が出た例を表 5.2 に示す。

表 5.1: 良い結果が出た対の例

| 格調が低い単語 | 格調が高い単語 | 優先度差  | 評価 |
|---------|---------|-------|----|
| 利口      | 聡明      | 7.60  | ○  |
| すごい     | すばらしい   | 8.09  | ○  |
| だんだんに   | 次第に     | 2.85  | ○  |
| のっぼ     | 長身      | 6.90  | ○  |
| へこたれる   | 屈する     | 12.57 | ○  |
| デマ      | 風説      | 12.73 | ○  |
| これから    | 今後      | 10.98 | ○  |
| まもなく    | ほどなく    | 4.71  | ○  |

表 5.2: 悪い結果が出た対の例

| 格調が低い単語 | 格調が高い単語 | 優先度差   | 評価 |
|---------|---------|--------|----|
| 消極的     | 慎ましい    | -10.71 | ×  |
| 落第      | 不合格     | -1.16  | ×  |
| 静寂      | しじま     | -4.01  | ×  |
| 露骨      | あからさま   | -2.67  | ×  |
| 手引き     | 手ほどき    | -4.48  | ×  |
| 古株      | ベテラン    | -6.81  | ×  |
| 鈍い      | マイペース   | -6.67  | ×  |
| 豚に真珠    | 猫に小判    | -2.70  | ×  |

悪い結果の例より、Word2vec による判定では「ベテラン」や、「マイペース」、「ライバル」など、横文字を格調が低く判定する傾向があった。また、書籍の言い換えの傾向としては「落第」のようにはっきりとした物言いをするより、「不合格」と言った方が相手に与える印象が柔らかいという観点から「不合格」の方が格調が高いとなっているが、これは人によっては二字熟語を使うほうが格調が高いという意見も出ると考

えられる。さらに、「豚に真珠」を「猫に小判」に言い換えるなどの例も人によって意見がわかれると考えられる。

一方、良い結果には例にも示したように「すごい」、「だんだんに」、「デマ」を「すばらしい」、「次第に」、「風説」に言い換えるなどの人間が判断しても明確に言い換え後の語彙の方が格調が高く感じられるものが多く見られた。

Word2vec による判定の正解率があまり上がらなかった原因として、設定した単語対の格調高さの種類がちぐはぐであったことも考えられる。例えば、「どっこいどっこい」を「互角」に言い換えるなど、「格調の低い言葉」を「普通の言葉」にする言い換えがあり、他には「妊娠」を「身重」にするような「普通の言葉」を「格調が高い言葉」にするような言い換えもあった。そうすると、予め設定しておく格調の低い単語群にも、格調が高い単語群にも「普通の言葉」が入ってしまい、Word2vec による判定の妨げになると考えられる。よって、言い換えの種類が正確に統一されていなかったことも正解率が上がらなかった原因であると考えられる。

他にも、今回の実験では Word2vec のモデルの学習データを日本語 Wikipedia より作成したが、Wikipedia は百科事典であるためある程度文体が固定されることが考えられる。そのため、更に格調の高低が明確に分かれるような様々な文体を含む文を学習データに用いて Word2vec のモデルを学習させることでも正解率を上げることが出来ると予想できる。

そのため、格調が低い単語から格調が高い単語へ言い換える対のデータを増やすことはもちろん、対のデータ、学習データも整備して、判定の種類を増やすことで精度の向上が見込められると思われる。

## 第6章 今後の展開

本研究では、言い換えの書籍から格調の高低が判明している単語対を入手し、Word2vecを用いて格調高さの判定を行ったが、このままでは文中の言い換えには使えないという問題が残っている。この手法のみで文の言い換えをすると、実際の言い換えの際は文中の前後との整合性も考慮する必要がある。例えば「姉が妊娠した」という文が「姉が身重した」のように変換され、非文が生成されてしまう。

しかし、この手法は大量の文章から格調が高い文を抜き出す場合などに有用だと考えられる。機械学習ではネットなどから大量の文を学習させることが多いが、その大量の文を抜き出す行程が難しい場合もある。そのとき、機械に格調が高い単語を自動判定させることができれば文章収集が容易になるなどの活用がある。また、今回は格調高さのみに着目したが、それ以外にも例えば「口語」と「文語」などで文書を整理する際にも有効に活用できると予想できる。

## 第7章 おわりに

本研究では、格調が高い単語を判定する手法として、同じ意味を持つ言葉なら格調が高い単語ほうが、より他の格調が高い単語と意味が近くなるという仮説に基づいて実験を行った。いくつかの言い換え単語から、Word2vec を用いて格調が高い単語を判定した。Word2vec は分布仮説に基づき学習を行う。同じ文脈に出現する語彙が似たベクトルを得られることがわかっている。この性質を利用し、分散表現からコサイン類似度を利用して、同義語単語群からどの単語が予め格調が高いとわかっている単語との類似度が高く、格調が高いのかを判定した。

結果は、単語対 277 組中の 192 組を正しく判定でき、判定精度は 0.693 という結果になった。概ね基本的に設定しておく単語対のデータが多ければ多いほどより正答率が上がるという結果になった。予め決めておく単語対を 1 単語対から約 100 単語対まで増やすと、正答率が 0.051 上がった。しかし、約 50 単語対設定から約 100 単語対、約 200 単語対設定での場合で、約 200 単語対設定が最も正答率が高かったものの、正答率にほとんど差は見られなかった。

また、この手法のみで文の言い換えをすると、実際の言い換えの際は文中の前後との整合性も考慮する必要がある。例えば「姉が妊娠した」という文が「姉が身重した」のように変換され、非文が生成されてしまう。しかし、この手法は大量の文章から格調が高い文を抜き出す場合などに有用だと考えられる。機械学習ではネットなどから大量の文を学習させることが多いが、その大量の文を抜き出す行程が難しい場合もある。そのとき、機械に格調が高い単語を自動判定させることができれば文章収集が容易になるなどの活用がある。

# 謝辞

また，研究の進め方や本論文の書き方など，細部にわたる御指導を頂きました，鳥取大学工学部知能情報工学科自然言語処理研究室の村田真樹教授に心から御礼申し上げます．また，本研究を進めるにあたり，御指導，御助言を頂きました，村上仁一准教授に心から御礼申し上げます．その他様々な場面で御助言を頂いた自然言語処理研究室の皆様に感謝の意を表します．

## 参考文献

- [1] Peter D. Turney. : Thumbs Up or Thumbs Down? Semantic Orientation Applied to Unsupervised Classification of Reviews. In Proc. of ACL-2002, pp. 417-424, 2002.
- [2] 河野誠也, 湯口彰重, 吉野幸一郎: 大規模言語モデルを用いたリフレーミング表現の自動生成とその評価. 情報処理学会研究報告, Vol.2022-NL-252, No.9, pp.1-6, 2022.
- [3] Tomas Mikolov, Ilya Sutskever, Kai Chen, Greg S Corrado, and Jeff Dean.: Distributed representations of words and phrases and their compositionality. Advances in Neural Information Processing Systems 26, pp.3111-3119, 2013.
- [4] 話題の達人倶楽部: **大人の言い換えハンドブック**. 青春出版社, 2018.
- [5] gensim: Topic modelling for humans,  
<https://radimrehurek.com/gensim/#>