青山学院大学 社会情報学研究科　2019年度 修士論文

単語ベクトルを用いた大学レコメンドシステム

学籍番号　38118002

氏名　**北堀 達也**

指導教員　**宮治 裕 教授**

2020年 1月

論文要旨

近年，大学進学という選択肢は高校生にとって一般的な選択肢として受け入れられるようになった．また2020年4月から高等教育の修学支援制度が施行される予定であり，この制度の施行により大学進学率は増加すると考えられる．さらに新設大学，新設学部も年々増加しているなかで，特定の大学に志願者が集中するという問題が提起されている．

本研究では，高校生が自分にあった適切な大学を選択できるように支援するシステムを考案した．ネットの大学に関する記事から大学ごとの単語ベクトルを生成し，大学間の類似度を測る．また，キャンパスの所在地や学部ごとの偏差値といったデータを利用して，大学間の潜在的な関係性の可視化を支援するシステムを構築した．

検証の結果，特定の大学に絞ったシステムを構築し，潜在的な関係性の可視化を実現した．また，より横断的なデータを利用したシステムの精度向上の可能性に関して示した．

謝辞

本研究を進めるにあたって，親身にご指導していただいた指導教官の宮治教授に厚く御礼申し上げます．並びに，多くの助言，ご指摘いただいた研究室の先輩方と後輩の皆様に感謝の意を表します．

また学習に使用したデータとして大学プレスセンターの記事を利用させていただいた，株式会社大学通信の皆様，パスナビのデータを利用させていただいた旺文社の皆様に感謝の気持ちと御礼を申し上げます．最後に家庭内で体調面，精神面で支えて下さった父と母に心から感謝を述べ，謝辞にかえさせていただきます．

目次

論文要旨 i

謝辞 ii

第1章 はじめに 1

1.1 背景 1

1.2 研究目的 5

1.3 関連研究 5

1.3.1 DatingServiceのデータを用いたWord2Vecによる趣味・嗜好の類似度算出 5

1.3.2 アニメの主題歌による類似アニメ検索の検討 5

1.3.3 本研究の新規性 6

1.4 論文構成 6

第2章 本研究で用いた技術 7

2.1 Word2Vec 7

2.1.1 Word2Vecの構造 7

2.1.2 単語ベクトルの次元数 7

2.1.3 反復回数 8

2.1.4 windowサイズ 8

2.1.5 本研究での利用 8

2.2 GloVe 8

2.2.1 概要 8

2.2.2 パラメータ 9

2.3 TF-IDF 9

2.3.1 Term Frequency 9
2.3.2 Inverse Document Frequency 10
2.3.3 本研究での利用 10

第3章 関連大学推薦システム 11
3.1 ユースケース 11
3.2 システム構成 12
3.2.1 概要 12
3.2.2 モデル部 12
3.2.3 大学情報データベース部 13
3.3 提供されるAPI 13
3.3.1 ある大学の単語ベクトルと近い大学の取得 14
3.3.2 大学に単語を加算した大学を取得 14
3.3.3 大学から単語を減算した大学を取得 15
3.3.4 ある大学と他の大学間で共通の似た意味の単語を取得 . . . 15
3.3.5 大学の情報を取得 16

第4章 検証実験 17
4.1 実験の目的 17
4.2 実験の方法 18
4.2.1 単語ベクトルの次元数 18
4.2.2 反復回数 19
4.2.3 windowサイズ 19
4.2.4 x-max 19
4.3 Word2Vecのモデル検証結果 19
4.3.1 WV_A 20
4.3.2 WV_B 21
4.3.3 WV_C 21
4.3.4 WV_D 22
4.3.5 WV_E 23
4.3.6 WV_F 23

4.3.7 WV_G 24
4.3.8 WV_H 24
4.4 Word2Vecモデルの考察 25
4.4.1 青山学院大学に近い大学 25
4.4.2 青山学院大学－キリスト教 26
4.4.3 青山学院大学と明治学院大学で共通の近い単語 26
4.5 GloVeのモデル検証結果 27
4.5.1 GV_A 27
4.5.2 GV_B 28
4.5.3 GV_C 29
4.5.4 GV_D 30
4.5.5 GV_E 30
4.5.6 GV_F 31
4.5.7 GV_G 31
4.5.8 GV_H 32
4.6 GloVeモデルの考察 33
4.6.1 青山学院大学と近い大学 33
4.6.2 青山学院大学－キリスト教 33
4.6.3 青山学院大学と明治学院大学で共通の近い単語 34
4.7 モデル評価のまとめ 34

第5章 おわりに 35
5.1 改善点 35

参考文献 37

第 1 章

はじめに

本論文では，大学に関する記事から文書毎のベクトルを生成することで各大学毎の特色をデータ化し，潜在的な関係性を可視化する研究について記述する．

まず，本研究をおこなう背景となった事柄について述べる．次に，研究目的の詳細について述べ，最後に次章以降の本論文の構成について概略を述べる．

1.1 背景

近年，大学進学という選択肢は高校生にとって，一般的な選択肢として受け入れられるようになった．図 1.1 に過去 10 年間の大学・短大進学率と，大学 (学部) 進学率の推移を文部省公表の資料 [1] から示す．直近 3 年間の推移は比較的横ばい傾向にあるが，高校卒業後の進路に関して 50% 近い学生が大学へ進学している．

更に，2020 年 4 月から高等教育の修学支援制度 [2] が施行される．この新制度では，学生個人に対する要件と，支援対象者の所得に関する要件を満たした場合に授業料などを減免するか，給付型の奨学金を支給する．対象となる学校種は大学，短期大学，高等専門学校，専門学校となる．前提として，少子化が進んでいるために全体の大学進学者数は減少傾向にあるが，この制度の施行により，大学・専門学校進学率は今後増加すると考えられる．

図 1.1: 過去 10 年間の大学進学率推移

一方大学の学校数に関して，日本では 774 校存在している**??**．これは 2019 年 4 月の入学者を募集した大学の数である．それぞれの内訳としては，国立大学が 82 大学，公立大学が 91 大学，私立大学が 592 大学であり，私立大学が全体の約 8 割を占めている．このデータのうち，2019 年度新設大学は 13 大学で，内訳は公立大学 1 校，私立大学 10 校，専門職大学 2 校に上る．また新設学部は国立，私立専門職大学合わせて 61 学部，新設学科は計 118 学科となっている．

志願者数の観点から，大学に関する志願者数の推移を日本私立大学振興・共済事業団の資料 [3] からまとめた．

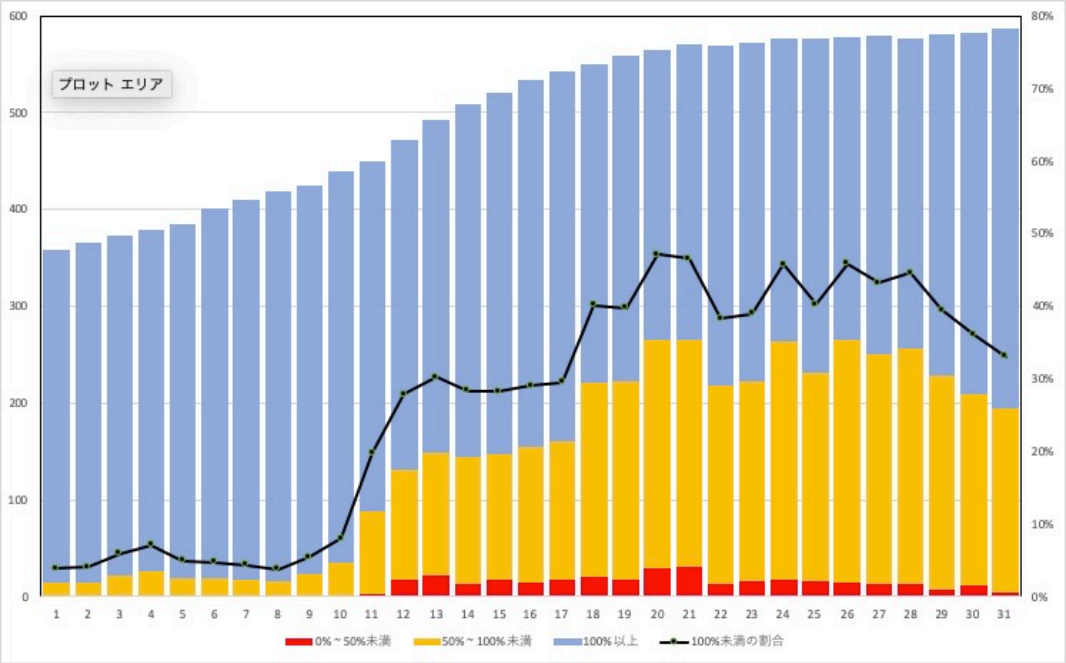

図 1.2: 平成元年〜31 年の私立大学志願者数の推移

大学数は増加傾向にあったものの，ここ 10 年間はほぼ横ばいである．志願者数が募集人数を下回った大学の比率は近年減少傾向にある．これは 2018 年から大学に対して入学者の超過率を厳格に制限 [4] したためであると考えられる．入学者の超過率が厳格に制限されたため，入学者数が絞られる形になり，その分の学生が他の大学に流れた結果定員割れの大学数が減ったと考えられる．更に平成 31 年からは入学定員充足率が 0.9 〜 1.0 倍の場合に入学定員充足率に応じて補助金が増額される [5]．そのため今後，有名大学の定員は減少傾向になると予想される．このような政策を政府が主導して施行した理由を論座の記事 [6] から引用する．

> 「私立大学の入学定員管理の厳格化」（以下、定員厳格化）とは、大都市圏の大規模私立大学に学生が集中している状況を改善するため、文科省が2016年度から始めた政策である。
>
> 私立大学の予算には国から交付される助成金が含まれており、その額は大学にもよるが、平均して大学の年間収入額の1割前後にもなる。定員厳格化は、所定の枠を超えて入学させる大学に対してその助成金を交付しないという、いわば「金で大学を縛る」政策なのである。

また論座の記事には，このような定員厳格化の影響で，不本意入学による学生と大学のミスマッチが起こる可能性について論じている．さらにこの状況を示したデータとして，図 1.3 に示した所属大学の選択理由に関する私立大学学生生活白書 2018[7] のアンケート結果が得られた．ここから読み取れるのは，自宅からの通学が可能だったから，自分の力にあっていたから，他に合格した大学がなかったから，大都市にあるから，などの理由が多く回答されているということである．

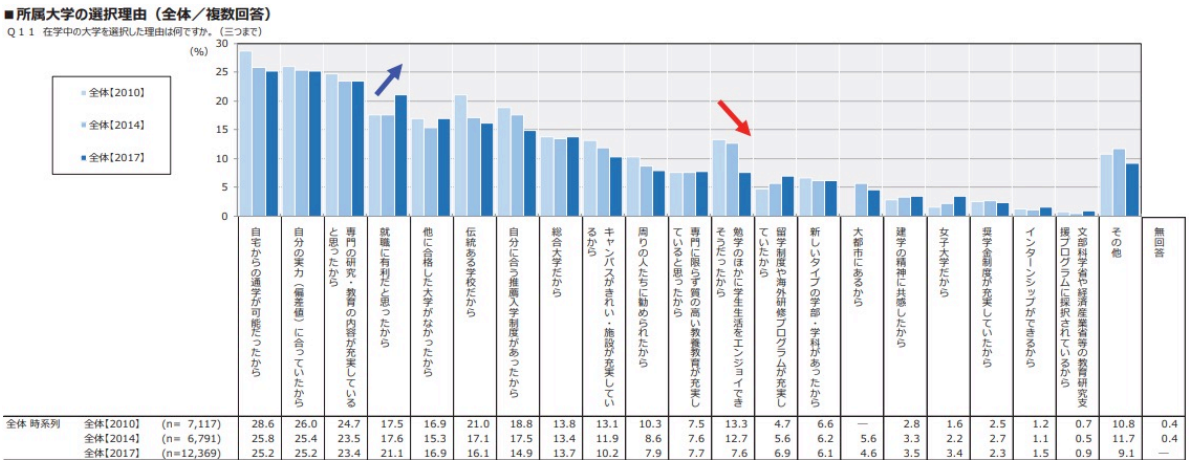

図 1.3: 所属大学選択理由

本来，講義内容や力を入れている研究分野，校風などを加味して自分に合った大学を選ぶべきである．しかし実際には，安定思考や偏差値，立地といった尺度だけで大学を選択する傾向にある．その結果，先に述べた大学定員厳格化の影響と合わせて，志願者の一部が第一志望の学校から第二志望の学校になっただけになる．また，定員を無理に絞ることで偏差値基準で下の大学に学生が流れていくと，最終的に特定の大学に志願者が集中するといった本質的な問題の解決にはならない．

本質的な問題点は，大学を選ぶ際の基準が画一化している点である．そのため大都市の大学，有名大学等に志願者が集中する事態に陥っている．この問題を解決するためには，大学選択において，1 つの尺度で大学を選択するのではく，総合的な観点から関係性を持った大学を提示し，自分にあった大学を選ぶ必要がある．

1.2 研究目的

本研究の大きな目的は，大学を選ぶ際の基準となるような情報を提示するシステムの構築をとする．システムの構築のために，本研究は大学に関する記事から各大学の特徴をベクトルで表現し，分析可能なモデルを生成することで大学間の潜在的な関係性を可視化することを目的とした．具体的には大学に関する記事から単語ベクトルを学習し，偏差値やキャンパス所在地などの情報を補助的に利用して，大学間の関係性を可視化するための支援をするシステムを構築する．

1.3 関連研究

1.3.1 DatingService のデータを用いた Word2Vec による趣味・嗜好の類似度算出

明畠ら [8] の研究では，Word2Vec を用いた趣味嗜好の類似度計算を，DatingService のデータを用いて行なっている．本研究でも Word2Vec と GloVe で生成したそれぞれの単語ベクトルを他の大学の単語ベクトルとのコサイン類似度を計算している．本研究における新規性は，GloVe を用いてモデルを生成し，比較している点である．また，データセットも異なるため，ここでは類似度の計算という点で関連研究として挙げる．

1.3.2 アニメの主題歌による類似アニメ検索の検討

本間ら [9] の研究では，Word2Vec を用いてアニメ主題歌から類似アニメの検索手法を提案している．歌詞内の単語ベクトルと，あらすじ内の単語ベクトルをそれぞれ平均して，1つの文章ベクトルとして扱い，コサイン類似度を比較している．本研究では文章ベクトルは作成しないため，手法が異なるがコサイン類似度を用いてベクトルの類似度を求めている点で関連研究として挙げる．本研究では，大学に関する文書中に出現する大学名の単語ベクトルを用いたシステムの構築が目的であるのに対して，本間らの研究では文章中の全ての単語ベクトルの平均を計算することで文章ベクトルとし比較している点で異なる．

1.3.3 本研究の新規性

本研究は大学名の単語ベクトルに着目した点が新規性として挙げられる．大学に関する文書から大学名の単語ベクトルを生成する先行研究は存在しなかった．また，文書から得られる単語ベクトルだけでなく，最終的なシステムに偏差値や立地などの定量的なデータを補助的に利用する点も先行研究とは違う点として挙げられる．

1.4 論文構成

2章では本研究で使用した詳細な技術について説明する．3章では変研究で提案するシステムの説明と実際に構築したシステムの使用方法を解説する．4章ではシステムの有効性を検証した結果についてまとめる．最後に5章で有効性の考察を考察し，本研究のまとめと今後の課題について述べる．

第 2 章

本研究で用いた技術

本章では本研究で用いた詳細な技術について述べる.

2.1 Word2Vec

本節では Word2Vec の概要とパラメータについて説明する. Word2Vec[10] は, Tomas Mikolov ら [11] によって提案された, 単語をベクトルに変換するためのニューラルネットワークの実装である. 単語をベクトルで表現することで, 単語同士の関連性を定量的に扱うことができる. またベクトルに変換することで, 単語同士でベクトルの距離の足し引きができるようになるため, 単語の演算が可能になる.

2.1.1 Word2Vec の構造

Word2Vec の構造は入力層, 隠れ層, 出力層からなる単純なニューラルネットワークとなっている. 入力層と出力層は学習する単語の数だけ存在する. 隠れ層はあらかじめ指定した次元数 × 単語数 (入力層の数) のベクトルからなる.

入力層で受ける入力は文章を 1-of-K 形式に変換したものとなり, 出力結果が最適になるように隠れ層の単語ベクトルの重みを学習する. 最終的に得られるモデルはこの隠れ層で学習した単語ベクトルになる.

2.1.2 単語ベクトルの次元数

word2vec で単語ベクトルを学習する際に指定する size オプションでは, 隠れ層の単語ベクトルの次元数を指定する. このオプションで指定したサイズ * 全体

の単語数のサイズのベクトルに全ての単語を圧縮し，分散表現を得る．この次元数が大きすぎると効率的な分散表現を学習できないが，次元数が小さすぎると単語の特徴を十分にとらえきれなくなり，学習に時間を要する．

2.1.3 反復回数

iterオプションでは，学習の反復回数を指定する．反復回数が少ないと，最適な分散表現が得られる前に学習が終了してしまう．また，反復回数を増やすと学習に要する時間が増加する．検証実験では，最適な反復回数として，100 1000回を比較対象とする．

2.1.4 window サイズ

ある単語の単語ベクトルを学習する際に，文書に出現した学習対象の単語から指定した単語数まで離れた単語を対象として学習する．windowオプションではこの単語数を指定する．

2.1.5 本研究での利用

本研究で構築するシステムでは，大学間の関連を分析するためにWord2Vecを用いる．具体的には，大学プレスセンター[12]の記事から，あらかじめリストアップした大学に関する記事をスクレイピングにより取得し，それぞれの記事を学習データとしてWord2Vecのモデルを学習した．

2.2 GloVe

本節ではGloVeについて簡易的に解説する．

2.2.1 概要

GloVeはJeffrey Penningtonら[13]によって提案された単語ベクトルを取得する実装である．Word2Vecよりも後に提案された手法であり，C言語で実装されている．GloVeでは，コーパス全体の単語間の共起の数を最小二乗法で学習するモデルとなっている．そのため，Word2Vecと比較して学習時間が短縮できる．ま

た，小さなコーパスでも学習ができ，精度の高さも論文の実験から得られている．Word2Vecでは考慮できない，出現回数が極端に少ない単語に重要度が偏るといった問題も回避できる．これはGloVeでは共起頻度が極端に高い単語と，低い単語を重視しないためである．

2.2.2 パラメータ

基本的なパラメータはWord2Vecと共通であるため割愛する．ここでは本研究で適用したx-maxオプションの値について説明する．x-maxオプションはGloVeの学習の際に指定するパラメータで，共起頻度の閾値を表す．(2.1)式で示すように，重み関数 $f(x)$ は2つの単語i,jの共起頻度がx-maxオプションで指定した値未満の時に $(x/x_{max})^{\alpha}$ となり，それ以上の場合は1となる．α の値は論文から0.75とする．

$$f(x) = \begin{cases} (x/x_{max})^{\alpha} & \text{if } x < x_{max} \\ 1 & \text{otherwise} \end{cases} \tag{2.1}$$

これにより，共起頻度が低い単語は重みが低くなり，共起頻度の閾値以上の単語は重みが1になる．

2.3 TF-IDF

本節ではTF-IDFについて説明する．Term FrequencyとInverse Document Frequencyを掛け合わせたもので，文書中の単語の重要度の計算方法[14]である．

2.3.1 Term Frequency

Term Frequencyは文書内における単語の出現頻度を表す．文書内である単語が出現する頻度が高ければ，その単語は重要であると考えられる．単語 t のTF値の計算方法は，文書 d 内の単語 t の出現回数を文書 d 内の全ての単語の出現回数の総和で割ることで求められる．

$$\mathrm{tf}(t,d) = \frac{n_{t,d}}{\sum_{s \in d} n_{s,d}}$$

2.3.2 Inverse Document Frequency

Inverse Document Frequency は逆文書頻度と呼ばれるもので，ある文書内の単語の，全体の文書における出現頻度の対数になっている．IDF が低いほど，他の文書で出現しないため，重要な単語であると考えられる．単語 t の IDF 値の計算方法は，全ての文書数を単語 t が出現する文書の数で割った自然対数に 1 を足すことで求められる．1 が足されているのは，全ての文書に出現する単語の IDF 値が 0 にならないためである．

$$\mathrm{idf}(t) = \log \frac{N}{df(t)} + 1$$

2.3.3 **本研究での利用**

本研究で構築したシステムでは，大学間で共通している近い意味の単語を学習した単語ベクトルから取得する．その際に普遍的な単語や，TF-IDF 値から重要度が低いと判断できる単語を削除するために TF-IDF を用いた．

第 3 章

関連大学推薦システム

本研究では，生成した単語ベクトルのモデルをWeb API形式で利用できるように構築した．本章では本研究で構築したシステムのユースケースと概要，提供機能について説明する．

3.1 ユースケース

本節では，本研究で構築したシステムのユースケースについて説明する．

まず本システムは，ある大学に関連する大学をユーザに推薦する機能を提供する．具体的には，システム開発者に向けて，Web API形式で関連大学の推薦機能を提供するものである．

想定されるエンドユーザは大学受験を控えた高校生とする．本研究で構築したシステムが提供する機能を組み込んだWebアプリケーションを利用することで，漠然と気になっている大学から関連する大学を検索することができる．類似度の高い大学を検索すると，あらかじめ学習した単語ベクトルから，コサイン類似度が最も近い大学が順番に表示される．表示された大学との共通の近い単語を検索すると，どのような単語を通して検索元の大学と近いのかを確認することができる．

また検索結果の大学から気になる学部を選択すると，偏差値の近い大学の候補が表示される．この時，キャンパスの立地も考慮して検索結果をフィルタリングすることもできる．

さらに，大学名をベクトルで表現しているため，気になっている大学に対して何かしらの単語を足し算したり，引き算したりすることが可能である．

3.2 システム構成

本節では，本研究で構築したシステムの構成について説明する．

3.2.1 概要

本研究では，単語ベクトルと大学に関する情報を用いた大学名の検索機能をWeb API形式で利用できる形で実装した．大学に関する情報は，学部ごとの偏差値とキャンパス名，大学ごとのキャンパスの緯度経度を用いる．Web API形式で実装した理由は，フロントエンド，バックエンド問わずに利用できるためである．また特定のプログラミング言語にとらわれずに利用できることで，様々な利用方法が見出せる．例えばAPIを用いてWebサイトを構築したり，モデルから出力された結果を効率的に可視化することで，一目で関係性を認識できるようにすることなどが挙げられる．

各部は単語ベクトルのモデルを提供するモデル部，学部・偏差値データとキャンパス所在地を提供する大学情報データベース部から構築されている．

3.2.2 モデル部

本システムにおけるモデル部について説明する．モデル部はWord2VecとGloVeを用いて学習した単語ベクトルの機能を提供するものである．大学名と，それに関連する単語をベクトルで表現することで，大学から特定の特徴を加算・減算することができる．また，ベクトル間のコサイン類似度を計算することで，ある大学の単語ベクトルと近いベクトルで表現された大学を取得できる．

本システムで実装した主な機能は，大学名から近い大学を取得する機能，大学に要素を加算する機能，大学から要素を減算する機能，ある大学と他の大学間で共通の近い意味を持った単語の検索機能である．各機能の詳細に関しては次節で説明する．

また本研究では，検索対象の大学を関東近郊の特定の大学に制限した．制限した理由は，単語ベクトルを学習するのに十分なデータを用意することが困難であったためである．そのため，本研究では比較的データが入手できる21校に絞ってシステムを構築した．対象の大学は表3.1に示す．本研究で生成した単

語ベクトルの学習に利用したデータは主に3種類挙げられる．1つは対象の大学それぞれのWikipediaの記事，2つ目はパスナビの各大学のページから沿革，LIFE&STUDY，大学院・研究室，3つ目は大学プレスセンターから各大学名で検索した結果の記事である．

表 3.1: **対象の大学**

青山学院大学	中央大学	立教大学	法政大学	明治大学
早稲田大学	慶應義塾大学	上智大学	国際基督教大学	
日本大学	東洋大学	駒澤大学	専修大学	
成蹊大学	成城大学	明治学院大学	学習院大学	
獨協大学	國學院大学	武蔵大学	東京理科大学	

3.2.3 大学情報データベース部

大学情報データベース部は本システムで対象を絞った21校の大学に関するデータを格納している．データのフォーマットはJSON形式で提供される．データの内容は，大学毎にキャンパスの情報があり，キャンパス情報の中に学部と対応した偏差値とキャンパスの所在地の情報が存在する．

学部と対応した偏差値のデータは，パスナビから取得した偏差値を利用した．偏差値が区間で提供されていた場合は，下限値と上限値の平均を採用した．

所在地はパスナビから取得したキャンパスの住所を元に，キャンパス所在地の緯度と経度を利用した．緯度と経度を採用した理由として，キャンパス間の距離を2次元で比較するためである．

これらのデータを元に，偏差値の近い大学や，立地の近い大学を学部毎に比較することができる．

3.3 提供される API

本節では提供するAPIの利用方法について述べる．

3.3.1 ある大学の単語ベクトルと近い大学の取得

ある大学の単語ベクトルと比較した時に，コサイン類似度が高い大学を取得する．リクエストボディにはunivNameで検索元の大学名を指定する．レスポンスは大学名とコサイン類似度のを含む2次元配列で返す．

POST getSimilarUnivs

表 3.2: getSimilarUnivs のリクエストボディ

プロパティ	タイプ	説明
univName	String	検索元の大学名

表 3.3: getSimilarUnivs のレスポンス

プロパティ	タイプ	説明
data	Array	大学名とコサイン類似度を含む 2 次元配列

3.3.2 大学に単語を加算した大学を取得

ある大学に単語を加算することで別の大学を取得する．例えば，明治大学という大学に対して，キリスト教という単語を加算することで，青山学院，立教大学のような明治大学に近くキリスト教の要素を含んだ大学を取得する．

POST addElementsToUniv

表 3.4: addElementsToUniv のリクエストボディ

プロパティ	タイプ	説明
univName	String	検索元の大学名
adds	Array	加算する単語の配列

表 3.5: addElementsToUniv のレスポンス

プロパティ	タイプ	説明
data	Array	大学名とコサイン類似度を含む 2 次元配列

3.3.3 大学から単語を減算した大学を取得

ある大学から単語を減算することで別の大学を取得する．例えば，青山学院大学に対して，キリスト教という単語を減算することで，明治大学，法政大学，中央大学のような青山学院大学からキリスト教の要素を除いた大学を取得する．

POST substractElementsFromUniv

表 3.6: substractElementsFromUniv のリクエストボディ

プロパティ	タイプ	説明
univName	String	検索元の大学名
adds	Array	減算する単語の配列

表 3.7: substractElementsFromUniv のレスポンス

プロパティ	タイプ	説明
data	Array	大学名とコサイン類似度を含む 2 次元配列

3.3.4 ある大学と他の大学間で共通の似た意味の単語を取得

青山学院大学と明治学院大学が近いと考えられる場合，それぞれの大学に近いベクトルを持つ単語が存在すると考えられる．そのような大学間で共通の似た意味の単語を取得する．

POST getCommonTerms

表 3.8: getCommonTerms のリクエストボディ

プロパティ	タイプ	説明
univNameFrom	String	一方の大学名
univNameTo	String	もう一方の大学名

表 3.9: getCommonTerms のレスポンス

プロパティ	タイプ	説明
data	Array	単語とそれぞれの大学名とのコサイン類似度の合計を含む 2 次元配列

3.3.5 大学の情報を取得

本研究で対象となる大学の内，ある大学の学部，偏差値，キャンパス所在地の情報を取得する．

POST getUnivInfo

表 3.10: getUnivInfo のリクエストボディ

プロパティ	タイプ	説明
univNameFrom	String	大学名

表 3.11: getUnivInfo のレスポンス

プロパティ	タイプ	説明
faculty	Object	key にキャンパス名，value に学部名と偏差値を含むオブジェクト
location	Object	key にキャンパス名，value に緯度経度の配列

第 4 章

検証実験

モデルの精度がどの程度妥当かを検証するために，本章ではパラメータを微調整したモデルの出力結果をまとめる．またそれぞれの出力結果に関して評価する．

4.1 実験の目的

Word2Vec と GloVe で単語ベクトルのモデルを生成する場合，いくつかのパラメータを考慮する必要がある．本実験では実際にシステムにモデルを組み込んで，機能を提供する際に最もパフォーマンスの高いモデルを適用するため，いくつかのタスクに分けてモデルを評価する．適用したパラメータの詳細は次節で説明する．

モデルを用いて出力する内容は，青山学院大学に近い大学名，青山学院大学－キリスト教に該当する大学名，青山学院大学と明治学院大学それぞれで共通する意味が近い単語の 3 項目とした．青山学院大学と明治学院大学それぞれで共通する意味が近い単語とは，青山学院大学に近い大学として明治学院大学が得られた場合，青山学院大学と明治学院大学で共通して近い単語を探す．得られた単語からのそれぞれの大学間の距離が近ければ，それぞれの大学が近い要因としての単語を得ることができる．青山学院大学の比較対象に明治学院大学を選んだ理由は，同じミッション系の大学で神学部の統合を通した日本神学校の創立などの歴史的な関係性を持っているためである．

4.2 実験の方法

モデルを評価する際に，それぞれの出力に対して評価軸を定めた．

青山学院大学に近い大学に関する評価項目は，まず偏差値が近いこと，ミッション系の大学であること，キャンパスの立地の近さである．本研究で参考にした偏差値とキャンパスの立地はパスナビ[15]を参考とした．モデルに学部，学科の情報は考慮されていないため，キャンパスの近さは学部を考慮しない．

次に，青山学院大学－キリスト教の評価項目は，非ミッション系の大学であること，偏差値が近いこと，キャンパスの立地が近いこととした．この場合，大学の要素からはキリスト教が消えていることが期待されるため，非ミッション系の大学であることをもっとも重要な評価項目とする．

最後に，青山学院大学と明治学院大学で共通の近い単語の評価項目は，直接的な関係性を示す単語とした．これは，両校がミッション系であることから，キリスト教に関する単語などがあげられる．また，歴史的な背景から日本神学校(現東京神学大学)の創設に両校の神学部が統合したことから，神学部に関する単語も考慮する．さらに専門部の統合から商業学部という単語と，1970年から1987年まで行われた体育会の総合定期戦に関連する単語も挙げられる．

表 4.1: 各出力結果の評価軸

青山学院大学に近い大学	青山学院大学 − キリスト教	青山学院大学と明治学院大学で共通の近い単語
偏差値が近い ミッション系 立地	非ミッション系 偏差値が近い 立地	直接的な関係性を示す単語

4.2.1 単語ベクトルの次元数

Word2VecとGloVeで指定する次元数は，小さすぎると単語の特徴を効率的に学習できず，大きすぎると適切な分散表現が学習できない．一般的に，50 300次元を指定する．本研究で使用したデータセットは比較的サイズが小さいため，単語ベクトルの次元数は50次元と100次元で比較した．

4.2.2 反復回数

Word2VecとGloVeのトレーニングの反復回数を指定する．この数字の大きさに比例して学習に要する時間も大きくなる．また，反復回数が少なすぎると十分に単語の特徴を学習できないため，検証実験では10, 100のパラメータで比較する．

4.2.3 window サイズ

windowサイズは10と1000で比較した．一般的にはwindowサイズは10　20で学習するが，記事の中に出現する大学名の単語ベクトルを学習する際，対象となる単語は記事全体に出現すると考えられるためwindowサイズに1000を適用して比較する．

4.2.4 x-max

GloVeの学習を行う際に，共起頻度の閾値を指定する必要がある．Jeffrey Penningtonらの実験では，100,000,000　600,000,000個のトークンが含まれたコーパスを用いて，x-maxオプションに100を指定した．一方本論文で学習したデータは約2,400,000個のトークンが含まれたデータを用いたため，x-maxオプションに指定する値は10とした．

4.3 Word2Vec のモデル検証結果

Word2Vecのモデル検証結果を示す．今回検証したモデル名と，対応するパラメータの一覧を表4.2に示す．

表 4.2: Word2Vec パラメータの詳細

モデル名	単語ベクトルの次元数	反復回数	window サイズ
WV_ A	50	10	10
WV_ B	100	10	10
WV_ C	50	100	10
WV_ D	100	100	10
WV_ E	50	10	1000
WV_ F	100	10	1000
WV_ G	50	100	1000
WV_ H	100	100	1000

4.3.1 WV_ A

単語ベクトルの次元数が50次元，反復回数とwindowサイズがそれぞれ10で生成したモデルの結果を表4.3に示す．

青山学院大学に2番目に近い大学として明治学院大学が出現したが，それ以外の結果は関連性が不透明である．青山学院大学－キリスト教の結果からは，東京から始まる大学名が頻出するため，学習不足であると考えられる．青山学院大学と明治学院大学で共通の近い単語の出力結果に関しては，明確にお互いの大学に共通する単語が出現していない．

表 4.3: WV_ A の検証結果

青山学院大学に近い大学		青山学院大学 - キリスト教		青山学院大学と明治学院大学で共通の近い単語	
大学名	類似度	大学名	類似度	単語	類似度の合計
聖心女子大学	0.74	目白大学	0.62	女学院	1.094
明治学院大学	0.722	帝京大学	0.583	英和	0.968
昭和女子大学	0.704	芝浦工業大学	0.552	芸術	0.742
帝京大学	0.703	女子美術大学	0.546	ライン	0.634
清泉女子大学	0.691	東京薬科大学	0.545	山手	0.579
実践女子大学	0.649	東京家政大学	0.544	校友	0.439
白百合女子大学	0.649	東京情報大学	0.542		
津田塾大学	0.638	東京電機大学	0.529		
東京女子体育大学	0.637	東京理科大学	0.525		
女子美術大学	0.636	多摩美術大学	0.524		

4.3.2 WV_B

単語ベクトルの次元数が100次元，反復回数とwindowサイズがそれぞれ10で生成したモデルの結果を4.4に示す．

青山学院大学に近い大学として，ミッション系の大学で偏差値も比較的近いと考えられる上智大学が得られた．また青山学院大学－キリスト教の結果から関西学院大学はミッション系の大学に該当する．さらに，青山学院大学と明治学院大学で共通の近い意味の単語から，キリスト教，神学，定期，合同などのミッション系の大学や歴史的背景を連想させるような単語が得られた．

表 4.4: WV_B の検証結果

青山学院大学に近い大学		青山学院大学 - キリスト教		青山学院大学と明治学院大学で共通の近い単語	
大学名	類似度	大学名	類似度	単語	類似度の合計
上智大学	0.372	成蹊大学	0.317	定期	0.531
東京外国語大学	0.308	駒澤大学	0.293	神学	0.45
駒澤大学	0.308	獨協大学	0.262	芸術	0.415
東京大学	0.295	東京理科大学	0.253	監督	0.37
筑波大学	0.294	中央大学	0.252	キリスト	0.364
早稲田大学	0.289	名古屋大学	0.221	イギリス	0.363
中央大学	0.276	筑波大学	0.204	キリスト教	0.323
オックスフォード大学	0.271	明治大学	0.19	合同	0.304
東京農業大学	0.256	大阪大学	0.167	前期	0.291
成蹊大学	0.256	関西学院大学	0.167	チャペル	0.272

4.3.3 WV_C

単語ベクトルの次元数が50次元，反復回数が100，windowサイズを10で生成したモデルの結果を4.5に示す．

青山学院大学に近い大学で上智大学や明治学院大学などが出現したが，単語ベクトルの類似度の高さでは清泉女子大学と聖心女子大学よりも低い．どちらもミッション系の大学であるが，偏差値などの観点から上智大学や明治学院大学の方が近いと考えられる．

青山学院大学と明治学院大学で共通の近い単語では，神学という単語が出現した．しかし，それぞれの大学との類似度の合計は他の単語に対して高くないため，学習不足であると考えられる．

表 4.5: WV_ C の検証結果

青山学院大学に近い大学		青山学院大学 - キリスト教		青山学院大学と明治学院大学で共通の近い単語	
大学名	類似度	大学名	類似度	単語	類似度の合計
清泉女子大学	0.677	東京薬科大学	0.557	女学院	0.981
聖心女子大学	0.674	帝京大学	0.539	英和	0.893
上智大学	0.633	中央大学	0.525	芸術	0.819
明治学院大学	0.616	目白大学	0.516	ライン	0.625
帝京大学	0.599	東京情報大学	0.498	学位	0.538
大東文化大学	0.585	女子美術大学	0.472	山手	0.538
立教大学	0.58	桜美林大学	0.469	校友	0.515
実践女子大学	0.578	東京電機大学	0.46	神学	0.463
桜美林大学	0.563	成蹊大学	0.458		
昭和女子大学	0.561	工学院大学	0.45		

4.3.4 WV_ D

単語ベクトルの次元数が 100 次元，反復回数が 100，window サイズを 10 で生成したモデルの結果を 4.6 に示す．

青山学院大学に近い大学として，上智大学や中央大学，学習院大学等が得られた．しかし類似度が高い大学で仏教系大学である駒澤大学などが出現した．青山学院大学と明治学院大学で共通の近い単語はキリストや礼拝，教会などミッション系の大学を連想させるような単語が得られた．また，定期，神学といった歴史的背景を連想させる単語も得られた．

表 4.6: WV_ D の検証結果

青山学院大学に近い大学		青山学院大学 - キリスト教		青山学院大学と明治学院大学で共通の近い単語	
大学名	類似度	大学名	類似度	単語	類似度の合計
上智大学	0.393	駒澤大学	0.337	キリスト	0.493
駒澤大学	0.364	中央大学	0.314	定期	0.491
東京外国語大学	0.304	獨協大学	0.261	キリスト教	0.441
中央大学	0.3	関西学院大学	0.244	神学	0.433
東京農業大学	0.288	東洋大学	0.224	礼拝	0.424
関西学院大学	0.283	成蹊大学	0.208	宗教	0.416
早稲田大学	0.254	名古屋大学	0.178	芸術	0.39
筑波大学	0.253	明治大学	0.177	教会	0.358
獨協大学	0.253	筑波大学	0.166	基本	0.34
学習院大学	0.249	法政大学	0.165	チャペル	0.311

4.3.5 WV_E

単語ベクトルの次元数が50次元，反復回数が10，windowサイズを1000で生成したモデルの結果を4.7に示す．

このモデルでは，青山学院大学に近い大学として中央大学，法政大学，立教大学，明治大学が類似度の高い大学としてあげられた．しかし，偏差値や立地を考えると類似度の値が相対的に低い．また，青山学院大学－キリスト教はミッション系の大学が出現せずに法政大学や明治大学を得ることができた．

青山学院大学と明治学院大学で共通の近い単語から，神学やミッション系の大学に関する単語が出現しなくなった．ここから得られる単語では両校の関係性を証明できない．

表 4.7: WV_E の検証結果

青山学院大学に近い大学		青山学院大学 - キリスト教		青山学院大学と明治学院大学で共通の近い単語	
大学名	類似度	大学名	類似度	単語	類似度の合計
東洋大学	0.69	法政大学	0.558	芸術	0.787
日本女子大学	0.651	明治大学	0.527	音楽	0.741
聖心女子大学	0.636	立命館大学	0.49	イギリス	0.732
立命館大学	0.623	東洋大学	0.482	女学院	0.714
昭和女子大学	0.623	昭和女子大学	0.465	バス	0.711
中央大学	0.607	千葉工業大学	0.463	コミュニティ	0.643
法政大学	0.601	横浜市立大学	0.459	相談	0.601
立教大学	0.595	中央大学	0.455		
明治大学	0.584	金沢工業大学	0.455		
東京電機大学	0.569	神奈川大学	0.433		

4.3.6 WV_F

単語ベクトルの次元数が100次元，反復回数が10，windowサイズを1000で生成したモデルの結果を4.8に示す．

青山学院大学に近い大学は近い偏差値の大学やミッション系の大学が多く得られた．このモデルでは北里大学の類似度が立教大学や上智大学などと比べて高く出ている．青山学院大学と明治学院大学で共通の近い単語から，商業という単語が新しく得られた．これは1944年に専門部を閉鎖し，明治学院に合同した際の高等商業学部から関連性があると考えられる．

表 4.8: WV_F の検証結果

青山学院大学に近い大学		青山学院大学 - キリスト教		青山学院大学と明治学院大学で共通の近い単語	
大学名	類似度	大学名	類似度	単語	類似度の合計
明治大学	0.729	明治大学	0.426	心理	0.857
北里大学	0.713	法政大学	0.423	併設	0.85
法政大学	0.707	北里大学	0.385	統合	0.843
明治学院大学	0.66	九州大学	0.362	商業	0.817
短期大学	0.658	立命館大学	0.341	前期	0.794
上智大学	0.654	名古屋大学	0.337	キリスト教	0.791
立教大学	0.65	駒澤大学	0.331	イギリス	0.72
中央大学	0.6	大阪大学	0.321	教会	0.679
早稲田大学	0.567	早稲田大学	0.317	神学	0.612
日本女子大学	0.562	中央大学	0.317	キリスト	0.606

4.3.7 WV_G

単語ベクトルの次元数が50次元，反復回数が100，windowサイズを1000で生成したモデルの結果を4.9に示す．

青山学院大学に近い単語は九州大学や東北大学，名古屋大学など立地的に遠い大学が出現した．更に国立大学が多く見られる．また青山学院大学－キリスト教では，ミッション系の大学である同志社大学が得られたため，モデルの有効性は低いと考えられる．

表 4.9: WV_G の検証結果

青山学院大学に近い大学		青山学院大学 - キリスト教		青山学院大学と明治学院大学で共通の近い単語	
大学名	類似度	大学名	類似度	単語	類似度の合計
九州大学	0.708	九州大学	0.556	イギリス	1.07
東北大学	0.52	名古屋大学	0.424	前期	1.005
明治学院大学	0.503	東北大学	0.392	キリスト教	0.902
上智大学	0.451	東京大学	0.315	基本	0.883
名古屋大学	0.448	大阪大学	0.299	芸術	0.834
短期大学	0.442	同志社大学	0.289	神学	0.824
東京大学	0.44	立命館大学	0.256	併設	0.818
明治大学	0.42	中央大学	0.246	キリスト	0.808
立教大学	0.41	明治大学	0.232	教会	0.774
東京農業大学	0.41	日本女子大学	0.231	山手	0.688

4.3.8 WV_H

単語ベクトルの次元数が100次元，反復回数が100，windowサイズを1000で生成したモデルの結果を4.10に示す．

青山学院大学に近い大学で，偏差値の近い明治大学や中央大学が得られた．また，同じミッション系の大学として，明治学院大学や立教大学，同志社大学，上智大学も得られた．一方で，青山学院大学－キリスト教でミッション系の大学の同志社大学や，オックスフォード大学などの大学が得られた．

青山学院大学と明治学院大学で共通の近い単語では，ミッション系の大学に関する単語や神学という単語とともに，新しく統合という単語が得られた．この単語は神学部や専門部の統合という歴史的な背景から得られたと考えられる．

表 4.10: WV_H の検証結果

青山学院大学に近い大学		青山学院大学 - キリスト教		青山学院大学と明治学院大学で共通の近い単語	
大学名	類似度	大学名	類似度	単語	類似度の合計
明治大学	0.494	明治大学	0.263	キリスト教	0.84
明治学院大学	0.457	北里大学	0.255	イギリス	0.764
立教大学	0.435	東京外国語大学	0.23	キリスト	0.703
同志社大学	0.429	同志社大学	0.225	教会	0.671
北里大学	0.428	学習院大学	0.219	前期	0.661
東京外国語大学	0.424	オックスフォード大学	0.197	神学	0.655
関西学院大学	0.377	東北大学	0.189	統合	0.613
短期大学	0.375	名古屋大学	0.185	併設	0.608
上智大学	0.344	九州大学	0.175	山手	0.606
中央大学	0.337	関西学院大学	0.156	チャペル	0.605

4.4 Word2Vec モデルの考察

3つの出力結果それぞれの観点からモデルの精度を考察する．

4.4.1 青山学院大学に近い大学

最も妥当性が高いと考えられるモデルはWV_F 表4.8である．このモデルは単語ベクトルの次元数が100次元，反復回数が10，windowサイズが1000で生成されたモデルである．このモデルの青山学院大学に近い大学の出力結果には，明治大学，法政大学，立教大学，中央大学が含まれており，同じ偏差値帯からMARCHと分類されている大学が全て得られた．また，同じミッション系である明治学院大学や，ミッション系かつ偏差値が近い上智大学などの大学も得られた．

次に妥当性が高いと考えられるモデルは，WV_H 表4.10である．このモデルは単語ベクトルの次元数が100次元，反復回数が100，windowサイズが1000で生

成されたモデルである．このモデルの青山学院大学に近い大学の出力結果は，明治大学，立教大学，中央大学が含まれている．WV_F と異なる点は，同志社大学と関西学院大学のような地理的に離れた大学が結果に出てきた点が挙げられる．これらの大学は青山学院大学と同じミッション系大学で偏差値も近いが，立地が離れているため次点とした．

これらの結果から，青山学院大学に近い大学の出力に関しては，window サイズは 1000 で反復回数は 10 で十分であると考えられる．

4.4.2 青山学院大学 − キリスト教

このタスクにおいて最も妥当性が高いと考えられるモデルは，WV_E 表 4.7 である．このモデルは単語ベクトルの次元数が 50 次元，反復回数が 10，window サイズが 1000 で生成されたモデルである．出力結果には，明治大学，法政大学，中央大学が含まれており，MARCH と分類されている大学からミッション系の青山学院大学と立教大学が除かれた結果となった．また青山学院大学に近い大学の出力結果から，明治学院大学，上智大学，立教大学が除かれており，各大学からキリスト教という単語の減算が機能していると考えられる．

次に妥当性が高いと考えられるモデルは WV_F 表 4.8 である．このモデルは単語ベクトルの次元数が 100 次元，反復回数が 10，window サイズが 1000 で生成されたモデルである．もっとも妥当性が高いと考えた WV_E と異なる点は，名古屋大学，大阪大学が出力された点である．これらの大学は立地的に離れいるが，一方 WV_E に出現した横浜市立大学は立地的にも近く，偏差値帯も近いため，WV_F の方が妥当性が低いと考えた．

これらの結果から，このタスクにおいては 2 つのモデルの差は小さいが，window サイズは 1000 が妥当で，反復回数は 10 で十分であると考えられる．

4.4.3 青山学院大学と明治学院大学で共通の近い単語

このタスクにおいて最も妥当性が高いと考えられるモデルは，WV_D 表 4.6 である．このモデルは単語ベクトルの次元数が 100 次元，反復回数が 100，window サイズが 10 で生成されたモデルである．出力結果は，キリスト，キリスト教，礼拝，宗教，教会，チャペルといったミッション系の大学を連想する単語と，定期

という過去に18年間開催された定期戦を連想させる単語と，神学という歴史的背景を連想させる単語が出力された．

次に妥当性が高いと考えられるモデルはWV_F表4.8である．このモデルは単語ベクトルの次元数が100次元，反復回数が10，windowサイズが1000で生成されたモデルである．このモデルの出力からは，キリスト教，キリスト，教会というミッション系を連想させる単語と，神学，商業，統合といった歴史的背景を連想させるような単語が出現した．

また，WV_B表**??**とWV_H表4.10もWV_F表**??**と同様に，全体に占める関連性のある単語の出現頻度で同じ結果であった．これらの結果から，このタスクにおいて重要な点は，単語ベクトルの次元数を大きくする点であると考えられる．

4.5 GloVe のモデル検証結果

本節ではGloVeのモデル検証結果をまとめた．GloVeで生成したモデル名と，対応するパラメータの一覧を表4.11に示す．

表 4.11: GloVe **パラメータの詳細**

モデル名	単語ベクトルの次元数	反復回数	window サイズ
GV_A	50	10	10
GV_B	100	10	10
GV_C	50	100	10
GV_D	100	100	10
GV_E	50	10	1000
GV_F	100	10	1000
GV_G	50	100	1000
GV_H	100	100	1000

4.5.1 GV_A

単語ベクトルの次元数が50次元，反復回数が10，windowサイズを10で生成したモデルの結果を4.12に示す．

このモデルからは，青山学院大学に近い大学として中央大学や立教大学，上

智大学，明治大学，法政大学など，偏差値が近い大学とミッション系の大学が取得できた．また，青山学院大学－キリスト教では，上位10校からはミッション系の大学が出現しなかった．しかし東京工業大学や大阪体育大学などは関連性が低いと考えられる．国際連合大学は大学ではないが，大学院は存在し青山学院大学と立地は非常に近いため出現したと考えられる．

青山学院大学と明治学院大学で共通の近い単語に関しては，学習不足であった．

表 4.12: GV_ A の検証結果

青山学院大学に近い大学		青山学院大学 - キリスト教		青山学院大学と明治学院大学で共通の近い単語	
大学名	類似度	大学名	類似度	単語	類似度の合計
中央大学	0.764	中央大学	0.559	学位	0.922
立教大学	0.747	東京工芸大学	0.557		
上智大学	0.71	立教大学	0.537		
駒澤大学	0.71	法政大学	0.506		
明治大学	0.683	国際連合大学	0.499		
学習院大学	0.672	関西大学	0.482		
法政大学	0.671	明治大学	0.482		
実践女子大学	0.623	大阪芸術大学	0.472		
東京理科大学	0.616	大阪体育大学	0.471		
成蹊大学	0.609	金沢医科大学	0.461		

4.5.2 GV_ B

単語ベクトルの次元数が100次元，反復回数が10，windowサイズを10で生成したモデルの結果を4.13に示す．

このモデルでは学習不足により青山学院大学と明治学院大学で共通の近い単語が得られなかった．青山学院大学－キリスト教の結果から，中央大学を除いて関連性の低い大学が出現した．

表 4.13: GV_ B の検証結果

青山学院大学に近い大学		青山学院大学 - キリスト教		青山学院大学と明治学院大学で共通の近い単語	
大学名	類似度	大学名	類似度	単語	類似度の合計
中央大学	0.639	東京工芸大学	0.52		
立教大学	0.588	中央大学	0.471		
学習院大学	0.562	大阪体育大学	0.46		
大東文化大学	0.551	大東文化大学	0.449		
明治大学	0.545	大阪芸術大学	0.435		
上智大学	0.537	神戸市外国語大学	0.415		
愛知淑徳大学	0.532	金沢医科大学	0.412		
大阪体育大学	0.53	京都女子大学	0.408		
芝浦工業大学	0.513	名城大学	0.407		
駒澤大学	0.496	会津大学	0.396		

4.5.3 GV_ C

単語ベクトルの次元数が50次元，反復回数が100，windowサイズを10で生成したモデルの結果を4.14に示す．

青山学院大学に近い大学として得られた大学は偏差値，立地の観点から妥当なものと考えられる．しかし青山学院大学－キリスト教から，上智大学と立教大学が得られた．これらの大学はミッション系であるため，モデルの妥当性は低いと考えられる．また，このモデルでは学習不足により青山学院大学と明治学院大学で共通の近い単語が得られなかった．

表 4.14: GV_ C の検証結果

青山学院大学に近い大学		青山学院大学 - キリスト教		青山学院大学と明治学院大学で共通の近い単語	
大学名	類似度	大学名	類似度	単語	類似度の合計
立教大学	0.691	中央大学	0.476		
中央大学	0.667	明治大学	0.466		
上智大学	0.632	成蹊大学	0.437		
明治大学	0.616	法政大学	0.426		
駒澤大学	0.612	武蔵大学	0.407		
法政大学	0.601	上智大学	0.404		
学習院大学	0.563	駒澤大学	0.391		
早稲田大学	0.497	立教大学	0.366		
成蹊大学	0.479	福岡大学	0.354		
東京理科大学	0.476	成城大学	0.347		

4.5.4 GV_D

単語ベクトルの次元数が100次元，反復回数が100，windowサイズを10で生成したモデルの結果を4.15に示す．

青山学院大学に近い大学として新しく大東文化大学が得られた．しかし評価項目を考慮すると，法政大学や上智大学よりも類似度が高い結果は妥当性が低いと考えられる．また，このモデルでは学習不足により青山学院大学と明治学院大学で共通の近い単語が得られなかった．

表 4.15: GV_D の検証結果

青山学院大学に近い大学		青山学院大学 - キリスト教		青山学院大学と明治学院大学で共通の近い単語
大学名	類似度	大学名	類似度	単語　類似度の合計
中央大学	0.642	中央大学	0.55	
明治大学	0.549	法政大学	0.412	
立教大学	0.542	明治大学	0.377	
学習院大学	0.539	関西大学	0.375	
大東文化大学	0.52	淑徳大学	0.37	
上智大学	0.52	武蔵大学	0.359	
成城大学	0.509	東洋大学	0.355	
駒澤大学	0.486	東京理科大学	0.349	
法政大学	0.474	駒澤大学	0.348	
芝浦工業大学	0.455	東京工科大学	0.348	

4.5.5 GV_E

単語ベクトルの次元数が50次元，反復回数が10，windowサイズを1000で生成したモデルの結果を4.16に示す．

青山学院大学に近い大学からは，女子美術大学がもっとも類似度が高い大学として得られた．青山学院大学との関連性はキャンパス所在地が相模原市という点があげられる．青山学院大学－キリスト教の結果からは相模女子大学が得られた．キャンパスの立地が近いための結果であると考えられる．

青山学院大学と明治学院大学で共通の近い単語は，ミッション系の大学を連想する単語が出現したものの，類似度の合計値が高い単語は関連性が不明確なものであった．

表 4.16: GV_ E の検証結果

青山学院大学に近い大学		青山学院大学 - キリスト教		青山学院大学と明治学院大学で共通の近い単語	
大学名	類似度	大学名	類似度	単語	類似度の合計
女子美術大学	0.647	相模女子大学	0.442	芸術	1.024
立教大学	0.631	神奈川工科大学	0.396	コミュニティ	0.93
聖心女子大学	0.622	芝浦工業大学	0.391	音楽	0.925
東京工業大学	0.615	北海道教育大学	0.389	学位	0.874
津田塾大学	0.613	女子美術大学	0.373	礼拝	0.854
上智大学	0.611	東京農業大学	0.372	心理	0.854
関西学院大学	0.61	国際連合大学	0.362	合同	0.825
明治学院大学	0.605	目白大学	0.362		
帝京大学	0.602	創価大学	0.359		
東京農業大学	0.599	帝京大学	0.351		

4.5.6 GV_ F

単語ベクトルの次元数が 100 次元，反復回数が 10，window サイズを 1000 で生成したモデルの結果を 4.17 に示す．

青山学院大学に近い大学からは，関西学院大学や聖心女子大学，明治学院大学などのミッション系の大学が得られた．関西学院大学は立地を除いて青山学院大学に近い大学であると考えられる．また，青山学院大学と明治学院大学で共通の近い単語からは，関係性のある単語が取得できなかった．

表 4.17: GV_ F の検証結果

青山学院大学に近い大学		青山学院大学 - キリスト教		青山学院大学と明治学院大学で共通の近い単語	
大学名	類似度	大学名	類似度	単語	類似度の合計
関西学院大学	0.541	神奈川工科大学	0.344	女学院	0.753
聖心女子大学	0.525	相模女子大学	0.334	英和	0.742
早稲田大学	0.514	北海道教育大学	0.333	芸術	0.731
明治学院大学	0.513	目白大学	0.315	前期	0.705
実践女子大学	0.503	国際連合大学	0.315	音楽	0.676
東京工業大学	0.495	関西学院大学	0.299	コミュニティ	0.674
女子美術大学	0.492	関東学院大学	0.293	貢献	0.631
津田塾大学	0.489	会津大学	0.285		
帝京大学	0.485	大妻女子大学	0.283		
東京大学	0.474	聖心女子大学	0.279		

4.5.7 GV_ G

単語ベクトルの次元数が 50 次元，反復回数が 100，window サイズを 1000 で生成したモデルの結果を 4.18 に示す．

このモデルでも青山学院大学に近い大学として関西学院大学が得られた．また，青山学院大学と明治学院大学で共通の近い単語からは，関係性のある単語が取得できなかった．

表 4.18: GV_ G の検証結果

青山学院大学に近い大学		青山学院大学 - キリスト教		青山学院大学と明治学院大学で共通の近い単語	
大学名	類似度	大学名	類似度	単語	類似度の合計
関西学院大学	0.505	神奈川工科大学	0.41	女学院	0.63
女子美術大学	0.497	東京農業大学	0.38	芸術	0.612
聖心女子大学	0.47	明治大学	0.377	英和	0.572
立教大学	0.459	相模女子大学	0.372	学位	0.485
国際大学	0.456	東京情報大学	0.369		
相模女子大学	0.445	帝京大学	0.355		
同志社大学	0.445	神奈川大学	0.35		
津田塾大学	0.425	帝京平成大学	0.347		
愛知淑徳大学	0.421	東京農工大学	0.333		
東京情報大学	0.411	福島県立医科大学	0.328		

4.5.8 GV_ H

単語ベクトルの次元数が 100 次元，反復回数が 100, window サイズを 1000 で生成したモデルの結果を 4.19 に示す．

このモデルでは，青山学院大学にもっとも近い大学として，関西学院大学が得られた．青山学院大学 – キリスト教の結果からは，もっとも近い大学として関西学院大学が得られたが，ミッション系の大学であるため，モデルの精度は妥当ではないと考えられる．青山学院大学と明治学院大学で共通の近い単語からは，ミッション系の大学を連想する単語と，統合という歴史的な背景を連想する単語が得られた．また定期という単語からは，1970 年から 1987 年まで行われていた総合定期戦からの結果であると考えられる．

表 4.19: GV_ H の検証結果

青山学院大学に近い大学		青山学院大学 - キリスト教		青山学院大学と明治学院大学で共通の近い単語	
大学名	類似度	大学名	類似度	単語	類似度の合計
関西学院大学	0.353	関西学院大学	0.315	聖書	0.471
東京外国語大学	0.322	東京農業大学	0.272	定期	0.458
東京農業大学	0.271	神戸大学	0.214	統合	0.455
静岡大学	0.27	東京理科大学	0.213	女学院	0.451
青山学院女子短期大学	0.257	横浜市立大学	0.213	英和	0.426
オックスフォード大学	0.242	静岡大学	0.18	チャペル	0.369
神戸大学	0.235	昭和女子大学	0.179	前期	0.365
東京工業大学	0.229	北海道大学	0.179	キリスト	0.359
九州大学	0.229	明治大学	0.175	キリスト教	0.357
筑波大学	0.221	九州大学	0.175	申し出	0.351

4.6 GloVe モデルの考察

GloVe で生成したモデルで出力した3つの結果それぞれの観点からモデルの精度を考察する.

4.6.1 青山学院大学と近い大学

このタスクにおいて最も妥当性が高いと考えられるモデルは, GV_ C 表 4.14 である. 出力結果から, 立教大学, 中央大学, 明治大学, 法政大学が取得できた. また, ミッション系の大学で上智大学, 偏差値が近いと考えられる東京理科大学や成蹊大学, 学習院大学が得られた.

GV_A 表 4.12 と GV_ B 表 4.13, GV_ D 表 4.15 も出力結果に大きな差はなかった. これらの結果から, モデルの window サイズは 10 が妥当であることが分かる. おそらく window サイズを 1000 に設定すると, 学習が収束する前に終わってしまうため, 適切な関係性を学習しきれなかったと考えられる.

4.6.2 青山学院大学 – キリスト教

このタスクにおいて最も妥当性が高いと考えられるモデルは, WV_ D 表 4.15 である. 理由としてはまず出力結果にミッション系の大学が含まれていない点が第 1 に挙げられる. 第 2 に中央大学, 法政大学, 明治大学が高い類似度で示されている点が挙げられる.

このほかでは GV_ C 表 4.14 が挙げられるが, 上智大学と立教大学が出力され

ている点でキリスト教の減算がうまく機能していないと考えられる．これは単語ベクトルのサイズの違いが影響していると考えられるが，これらの結果から，単語ベクトルのサイズは100次元が妥当であり，windowサイズは10で十分であると言える．

4.6.3 青山学院大学と明治学院大学で共通の近い単語

GloVeで生成したモデルでは，windowサイズが小さい場合，このタスクの結果が得られなかった．最も妥当性の高いと考えられるモデルはGV_H 表 4.19である．出力から，聖書，チャペル，キリスト，キリスト教などのミッション系の大学を連想させる単語と，統合や定期などの歴史的背景を連想させる単語が得られた．他のモデルでは十分な出力を得られなかったため，このモデルが今回の検証実験で最も妥当性が高いと言える．

4.7 モデル評価のまとめ

Word2Vecで生成したモデルに関しては，総合的に評価して，WV_ F 表 4.8が精度が高いと言えた．

またGloVeで生成したモデルに関しては，適切なモデルがタスクによって異なった．青山学院大学に近い大学を出力するタスクと，青山学院大学－キリスト教のタスクでは，GV_ D 表 **??**が総合的に評価して妥当性が高かった．青山学院大学と明治学院大学で共通の近い単語を出力するタスクでは，よりwindwoサイズが大きなGV_ H 表 4.19の精度が高く，windowサイズを小さくすると結果が得られなかった．

Word2Vecで生成したモデルと，GloVeで生成したモデルを比較すると，全てのタスクに対して安定した出力が期待できるモデルは，Word2Vecで生成したモデルであった．しかし青山学院大学に近い大学を出力するタスクと，青山学院大学－キリスト教のタスクではGloVeで生成したモデルの出力の方が，偏差値，ミッション系，立地の観点から妥当性の高い結果となった．

第5章

おわりに

本研究では，大学に関するWikipediaの記事と大学プレスセンターの記事をデータセットとして，単語ベクトルを学習しモデルを生成した．また，パスナビの大学のページから取得した学部ごとのキャンパスや偏差値の情報と，キャンパス所在地の緯度経度といった情報を補助的に用いて，大学間の関係性を可視化するための支援をするシステムを考案した．

検証実験から，特定のタスクに対して精度の高いモデルを評価して，タスクに応じたモデルを利用することで出力結果の妥当性を向上させた．データセットの都合上，対象の大学は関東近郊の21校に絞ったが，Web API形式で実装することで様々な環境で単語ベクトルのモデルを利用できるシステムを構築した．

5.1 改善点

まずデータセットに偏りがあることが改善点として挙げられる．今回使用したデータは，Wikipedia, パスナビ，大学プレスセンターの記事をマージしたものであるが，大学プレスセンターの記事がデータセットにおけるほとんどの割合を占めている．そこで大学プレスセンターの記事に偏りがあると，モデルの精度に影響が出る．さらに記事の数も大学によって差があるため，安定したデータセットを収集する必要がある．しかし日本全国の全ての大学を考慮に入れる場合，有名大学と新設大学では利用できるデータの量に大きな差がある．

また暗黙知のように我々が知っている情報がインターネットから取得できない点も挙げられる．例えば大学によっておしゃれな印象であったり，お金持ちが多い大学などといった情報は収集するのが難しい．このような情報はTwitter

や5ちゃんねるなどから取得できる可能性があるが，ノイズが非常に多いことが予想される．それゆえ，様々な媒体から横断的に大量のデータを収集する必要がある．

参考文献

[1] 文部科学省. 平成30年度画稿基本調査(確定値)の好評について. https://www.mext.go.jp/component/b_menu/other/__icsFiles/afieldfile/2018/12/25/1407449_1.pdf.

[2] 文部科学省. 高等教育の修学支援新制度. https://www.mext.go.jp/a_menu/koutou/hutankeigen/index.htm.

[3] 日本私立学校振興・共済事業団. 平成31(2019)年度私立大学・短期大学等入学志願動向. https://www.shigaku.go.jp/files/shigandoukouH31.pdf.

[4] 文部科学省. 平成28年度以降の定員管理に係る私立大学等経常補助金の取り扱いについて(通知). https://www.mext.go.jp/a_menu/koutou/shinkou/07021403/002/002/__icsFiles/afieldfile/2015/07/13/1360007_2.pdf.

[5] 文部科学省. 平成31年度以降の定員管理に関わる私立大学等経常補助金の取り扱いについて(通知). https://www.mext.go.jp/a_menu/koutou/shinkou/07021403/002/002/__icsFiles/afieldfile/2018/09/19/1409177.pdf.

[6] 篠原秀雄. 次第定員厳格化がひきおこした大学受験の大混乱. https://webronza.asahi.com/science/articles/2019112600002.html?page=2.

[7] 北篠英勝. 私立大学学生生活白書2018. Technical report, 一般社団法人日本私立大学連盟, 2018.

[8] 明畠利樹, 中西健太郎, 岩本拓也. Datingserviceのデータを用いたword2vecによる趣味・嗜好の類似度計算, 2017.

[9] 本間直人, 北原鉄朗ほか. アニメの主題歌による類似アニメ検索の検討. 研究報告音声言語情報処理(SLP), Vol. 2018, No. 42, pp. 1–2, 2018.

[10] 西尾泰和. word2vecによる自然言語処理. 株式会社オライリー・ジャパン, 2014.

[11] Tomas Mikolov, Kai Chen, Greg Corrado, and Jeffrey Dean. Efficient estimation of word

representations in vector space. Technical report, Google Inc. Mountain View, CA, 9 2013.

[12] 大学プレスセンター. https://www.u-presscenter.jp/.

[13] Jeffrey Pennington, Richard Socher, and Christopher D.Manning. Glove: Global vectors for word representation. Technical report, Computer Science Department, Stanford University, Stanford, CA 94305, 2014.

[14] 天野真家, 石崎俊, 宇津呂武仁, 成田真澄, 福本淳一. 自然言語処理システム. 自然言語処理, p. 138. 2007.

[15] 大学受験パスナビ. https://passnavi.evidus.com/.