

2020年度（令和2年度） 卒業論文

機械学習と統計的検定を利用した  
知見獲得の評価実験

電気情報系学科 卒業論文検印	
学科長	

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

B17T2024C 鎌倉 周平

## 概要

本研究は、教師あり機械学習と統計的検定を用いることにより、毎日新聞のテキストデータから単語対に関わる有益な情報（本稿では素性とする）を取り出す。教師あり機械学習と統計的検定で取り出した素性を分析・評価することを本研究の目的とする。

本研究の成果は2つある。1つ目は、知見獲得の実験で、教師あり機械学習と統計的検定を用いて新聞記事から様々な分野の単語対に関する素性を取り出し、分析することで様々な分野の知見を得ることができたことである。教師あり機械学習は、新聞データに単語対(A, B)が含む文章を機械に学習させ、その文章の分類先をA, Bとして推定し、その手がかりとなった素性を取り出し、分析することで知見獲得を行った。統計的検定では、新聞データに単語対(A, B)が含む文章からA, Bと共起する単語の共起回数を基にA, Bの素性を取り出し、分析することで知見獲得を行った。

2つ目は、教師あり機械学習と統計的検定で収集した単語対の素性を複数の評価実験を通じて、教師あり機械学習と統計的検定の知見獲得の有用性を示したことである。評価実験は3つ行った。1つ目は、Word2vecで収集した単語対の素性との比較である。Word2vecは、単語対と類似性のある素性を収集することができる。そのWord2vecで収集した単語対の素性500個と教師あり機械学習で収集した単語対の素性500個、統計的検定で収集した単語対の素性500個との一致数で評価した。その一致数の平均は教師あり機械学習が30.6個、統計的検定が70.3個であった。2つ目の評価実験は、人手で連想した語句30個とそれぞれの手法で得た単語対の素性500個との一致数で評価した。筆者が連想した単語対の語句との一致数の平均は、教師あり機械学習が7.9個、統計的検定が11.3個であった。3つ目の評価実験は、それぞれの手法で得た単語対の素性100個を有益性の観点から人手で評価した。筆者が行った評価の比率は、教師あり機械学習が0.14、統計的検定が0.07であった。また、2つ目と3つ目の評価実験を筆者以外の被験者3名による被験者実験を行った。連想語句との一致数の平均は、教師あり機械学習が6.7個、統計的検定が10.4個であった。有益性における比率は、教師あり機械学習が0.085、統計的検定が0.059であった。教師あり機械学習は、役に立つ知見につながるような語句の収集に優れていることがわかり、統計的検定は、単語

対と類似度が高い素性や人が連想できる語句など一般的な語句の収集に優れていることがわかった。

# 目次

<b>第1章</b>	<b>はじめに</b>	<b>1</b>
<b>第2章</b>	<b>先行研究</b>	<b>3</b>
2.1	パターンと教師あり機械学習と素性分析を利用したウェブと新聞からの株式相場に関わる知見獲得 . . . . .	3
2.2	機械学習を用いた類義語の使い分けに関する知識獲得 . . . . .	4
2.3	機械学習とデータマイニング (大規模データベースからの知識獲得) . . . . .	4
<b>第3章</b>	<b>問題設定と提案手法</b>	<b>5</b>
3.1	問題設定 . . . . .	5
3.2	知見獲得における提案手法 . . . . .	5
3.2.1	最大エントロピー法による知見獲得 . . . . .	5
3.2.2	符号検定による知見獲得 . . . . .	5
3.3	知見獲得の評価における提案手法 . . . . .	6
3.3.1	Word2vec を用いた評価実験 . . . . .	6
3.3.2	人手による評価実験 . . . . .	6
3.4	最大エントロピー法 . . . . .	7
3.5	正規化 $\alpha$ 値 . . . . .	7
3.6	符号検定 . . . . .	8
3.7	Word2vec . . . . .	9
<b>第4章</b>	<b>知見獲得における実験と考察</b>	<b>10</b>
4.1	実験データ . . . . .	10
4.2	実験方法 . . . . .	10
4.3	実験結果と考察 . . . . .	11

<b>第 5 章</b>	<b>知見獲得の評価における実験と考察</b>	<b>26</b>
5.1	実験方法 . . . . .	26
5.1.1	Word2vec との比較 . . . . .	26
5.1.2	人手評価 . . . . .	26
5.2	実験結果と考察 . . . . .	27
5.2.1	Word2vec を用いた評価実験 . . . . .	27
5.2.2	人手による評価実験 . . . . .	28
5.2.3	被験者による評価実験 . . . . .	32
5.2.4	評価実験に対する考察 . . . . .	36
<b>第 6 章</b>	<b>今後の課題</b>	<b>38</b>
<b>第 7 章</b>	<b>おわりに</b>	<b>39</b>

# 表 目 次

3.1	評価基準 . . . . .	7
4.1	実験データ . . . . .	10
4.2	実験結果 . . . . .	11
4.3	正規化 $\alpha$ 値 (政治の上位 30 個) . . . . .	14
4.4	正規化 $\alpha$ 値 (経済の上位 30 個) . . . . .	15
4.5	正規化 $\alpha$ 値 (輸入の上位 30 個) . . . . .	16
4.6	正規化 $\alpha$ 値 (輸出の上位 30 個) . . . . .	17
4.7	正規化 $\alpha$ 値 (社会主義の上位 30 個) . . . . .	18
4.8	正規化 $\alpha$ 値 (資本主義の上位 30 個) . . . . .	19
4.9	正規化 $\alpha$ 値 (オリンピックの上位 30 個) . . . . .	20
4.10	正規化 $\alpha$ 値 (パラリンピックの上位 30 個) . . . . .	21
4.11	p 値 (政治と経済の下位 30 個) . . . . .	22
4.12	p 値 (輸入と輸出の下位 30 個) . . . . .	23
4.13	p 値 (社会主義と資本主義の下位 30 個) . . . . .	24
4.14	p 値 (オリンピックとパラリンピックの下位 30 個) . . . . .	25
5.1	Word2vec で得られた素性 500 個との一致数 . . . . .	27
5.2	連想語句一致数 . . . . .	28
5.3	評価結果 . . . . .	29
5.4	評価実験の比率 . . . . .	29
5.5	政治 . . . . .	29
5.6	経済 . . . . .	29
5.7	輸入 . . . . .	30
5.8	輸出 . . . . .	30
5.9	社会主義 . . . . .	30
5.10	資本主義 . . . . .	30

5.11	オリンピック	30
5.12	パラリンピック	31
5.13	被験者による連想語句一致数	32
5.14	被験者による連想語句一致数の平均	33
5.15	被験者による評価結果	33
5.16	被験者による評価実験の平均	33
5.17	被験者による評価実験の比率	33
5.18	政治	34
5.19	経済	34
5.20	輸入	34
5.21	輸出	34
5.22	社会主義	34
5.23	資本主義	35
5.24	オリンピック	35
5.25	パラリンピック	35

# 第1章 はじめに

近年、電子テキストが増加し、大量の電子テキストから重要な情報を素早く得ることが求められている。

先行研究 [1] では教師あり機械学習を利用してウェブと新聞から株式相場における知見の収集が行われている。先行研究 [2] では統計的検定を利用した使い分けに関する知見の収集が行われている。しかし、先行研究 [1] では「株価上昇」「株価下降」に対象が限定されておりその他の事象で研究がなされていない。また、それぞれの先行研究は人手での評価実験のみで、その他の手法との比較実験が行われていない。本研究では、機械学習の最大エントロピー法と統計的検定の符号検定を利用して、新聞データから様々な分野の知見を獲得し、それぞれの手法の評価を行う。

本研究の主な主張点を以下に整理する。

- 最大エントロピー法と符号検定を利用することで様々な分野の単語対に関する一般的な情報や有益な情報の収集ができた。例として政治では、「献金」や「天皇」や「トランプ」などがあり、経済では、「アベノミクス」や「リーマン」や「インフレ」などがあった。
- Word2vec で得た単語対の素性 500 個と最大エントロピー法と符号検定で得た単語対の素性 500 個との一致数の平均は、最大エントロピー法が 30.6 個、符号検定が 70.3 個であった。人手で連想した各単語対の語句 30 個とそれぞれの手法で得た単語対の素性 500 個との一致数の平均は、最大エントロピー法が 7.9 個、符号検定が 11.3 個であった。それぞれの手法で得た各単語対の素性 100 個を人手で評価した結果の有益性の比率は、最大エントロピー法が 0.14、符号検定が 0.07 であった。よって、符号検定は、単語対と類似度が高い素性や人が連想できる語句など一般的な語句の収集に優れており、最大エントロピー法は、役に立つ知見につながるような語句の収集に優れている。

本論文の構成は以下の通りである。第 2 章では、本研究に関連する研究としてどのような研究が行われてきたかを記述し、その研究と本研究との関連を説明する。第 3

章では、本研究で扱う問題の設定とそれを解決するために提案した手法について説明を行う。第4章では、本研究で行った知見獲得における実験についての説明と、その実験の結果と考察について記述する。第5章では、本研究で行った知見獲得における評価の実験についての説明と、その実験の結果と考察について記述する。第6章では、今後の課題について記述する。第7章ではまとめを行う。

## 第2章 先行研究

本章では、先行研究について記述する。2.1節では、村田ら [1] が行ったパターンと教師あり機械学習と素性分析を利用したウェブと新聞からの株式相場に関わる知見獲得について記述し、2.2節では、赤江 [2] が行った機械学習を用いた類義語の使い分けに関する知識獲得について記述する。2.3節では、元田ら [3] が行った機械学習とデータマイニングについて記述する。

### 2.1 パターンと教師あり機械学習と素性分析を利用したウェブと新聞からの株式相場に関わる知見獲得

村田ら [1] は、パターンと教師あり機械学習と素性分析を利用して株式相場の知見獲得を行った。データには、ウェブと新聞を利用し、株価の騰落に関わるパターンで収集した。ウェブでは「X による株価上昇」と「X による株価下落」のパターンでウェブから文章を収集し、X に相当する部分を入力テキストデータとする。例えば、「原油価格の混乱による株価下落」の表現から入力テキストとして「原油価格の混乱」を得る。新聞では、「東京株式市場 X 日経平均株価…前日終値比」というパターンで抜き出し、X に相当する部分を入力テキストデータとする。例えば、「10日の東京株式市場は、前日の株価上昇に対し利益確定の売りが広がり日経平均株価は反落、一時、前日終値比306円92銭安の1万6930円85銭まで値を下げた。」という表現から、入力テキストとして、「は、前日の株価上昇に対し利益確定の売りが広がり」を得る。村田ら [1] は入力テキストデータを入力として、株価上昇であるか否かを出力として、教師あり機械学習を行った。素性には、入力テキスト中の全ての1から3個の単語連続を用いて、最大エントロピー法で株価上昇か下落かを調べた。素性分析による知見獲得には、全データを最大エントロピー法で求める  $\alpha$  値を正規化した値 (正規化  $\alpha$ ) を求めた。その値が大きい素性を抽出することで、株式相場に関する重要な知見獲得を行った。

## 2.2 機械学習を用いた類義語の使い分けに関する知識獲得

赤江 [2] は, 教師あり機械学習を用いるとともに統計的検定を用いて, 類義語の使い分けを行い, 類義語の使い分けに関わる知見を得ている.

この研究では, いくつかの類義語について実際に使い分けに役立ったと思われる情報を明らかにした. 特に, 類義語の使い分けに関する文献に載っていないような新たな知見が多く得られている. 例えば, 「作成」は「表」「リスト」などを作る時に使われ, 「作製」は「細胞」「遺伝子」などを作る時に使われるなどの素性を得られた. また, 品詞間における類義語の使い分けに関する特徴も得られた. この2つの成果は, 文章を生成する際の類義語の選択, 適切な表現の使い分けの提案に利用できる.

## 2.3 機械学習とデータマイニング (大規模データベースからの知識獲得)

元田ら [3] は大規模データベースからの知識獲得の技術である機械学習の概要とその位置付けをデータマイニングの観点から示した. 機械学習の手法として決定木学習について取り上げた. 表現能力の強力な一階述語論理を基にした帰納論理プログラムによる関係の学習, その中間に位置する命題論理で表現されたグラフパターンの学習について示した. データから公理的な知識を発見する点でデータマイニングの範疇にある科学的知識発見に対する手法を示し, データの測定行為・認知の意味論の重要性を明らかにした.

## 第3章 問題設定と提案手法

### 3.1 問題設定

先行研究 [1] では経済に関わる知見の獲得のみを行っているが、その他の分野での知見の獲得は行われていない。また、先行研究 [1] と先行研究 [2] は知見獲得の評価は被験者実験のみであり、その他の手法との比較を行っていない。そこで、本研究はそれら2点に着目し、機械学習と統計的検定を用いて様々な分野の知見の獲得と知見獲得の評価を目的とする。

### 3.2 知見獲得における提案手法

#### 3.2.1 最大エントロピー法による知見獲得

本研究では、教師あり機械学習に最大エントロピー法を用いる。最大エントロピー法では、どのような語があれば記事中に単語対が出現するかを学習する。例えば、単語対  $A$ ,  $B$  があるとする。  $A$ ,  $B$  を含む文を収集する。収集した文中から  $A$ ,  $B$  を削除し、  $X$  とする。  $X$  とした部分に  $A$ ,  $B$  どちらの語があったのかを学習結果を元に推定する。最大エントロピー法では、学習に役立つ素性に対して分類先への重要度を示す正規化  $\alpha$  値を付与する。正規化  $\alpha$  値が高い素性ほど有用な素性であるので、正規化  $\alpha$  値の高い素性を分析することで単語対に関わる知見獲得を行う。最大エントロピー法の説明を 3.4 節に示す。

#### 3.2.2 符号検定による知見獲得

本研究では、統計的検定に符号検定を用いる。符号検定では、  $A$ ,  $B$  を含む文中から  $A$ ,  $B$  と共起する素性が全データの出現率より偏って多くどちらかの分類先に出現しているかを二項定理に基づく片側検定により求め、有意確率  $p$  値を求める。  $p$  値が

低い素性ほど有用な素性であるので、値の小さい  $p$  値を持つ素性を分析することで単語対に関わる知見獲得を行う。符号検定の説明を 3.6 節に示す。

### 3.3 知見獲得の評価における提案手法

最大エントロピー法と符号検定で得られた素性が有用であることを明らかにするために、以下の評価実験を行う。

#### 3.3.1 Word2vec を用いた評価実験

Word2vec とは、大量の電子テキストを解析し、各単語の意味をベクトル表現することができる手法である。単語をベクトル化することで単語同士の意味の近さを計算することができる。Word2vec を使用して単語対と意味の近い単語を収集し、Word2vec で得られた素性と手法ごとに得られた素性とを比較することで、手法ごとに単語対と関連のある素性が得られたかどうかを評価できる。

#### 3.3.2 人手による評価実験

手法ごとに得られた素性が人にとって役に立つのか、一般的であるかを明らかにするために人手による以下の評価実験を行う。

1. 連想語句による評価

人手で単語対から連想する単語を 30 個書き出す。その単語と手法ごとに得られた素性との一致数を数える。

2. 有益性の評価

最大エントロピー法で得られた素性が符号検定で得られた素性であるかわからない状態で表 3.1 の基準を基に評価する。

以上 2 点の評価により、人が連想する語句をそれぞれの手法が得られているか、知見獲得において有用な手法であるかを評価する。また複数人の被験者で同様の実験を行い、より客観的な評価を行う。

表 3.1: 評価基準

◎	関連語であり役に立つ素性
○	関連語であるが役に立たない素性
×	関係のない素性である

### 3.4 最大エントロピー法

本研究では、教師あり機械学習法に、最大エントロピー法を使用する。最大エントロピー法とは、あらかじめ設定しておいた素性  $f_i (1 \leq j \leq k)$  の集合を  $F$  とするとき、式 (3.1) を満足しながらエントロピーを意味する式 (3.2) を最大にするときの確率分布  $p(a, b)$  を求め、その確率分布にしたがって求まる各分類の確率のうち、もっとも大きい確率値を持つ分類を求める分類とする方法である [1, 4, 5, 6].

$$\sum_{a \in A, b \in B} p(a, b) g_j(a, b) = \sum_{a \in A, b \in B} p(a, b) g_j(a, b) \quad (3.1)$$

*for*  $\forall f_j (1 \leq j \leq k)$

$$H(p) = - \sum_{a \in A, b \in B} p(a, b) \log(p(a, b)) \quad (3.2)$$

ただし、 $A, B$  は分類と文脈の集合を意味し、 $g_i(a, b)$  は文脈  $b$  に素性  $f_i$  があつてなおかつ分類が  $a$  の場合 1 となり、それ以外で 0 となる関数を意味する。また、 $(a, b)$  は、既知データでの  $(a, b)$  の出現の割合を意味する。

式 (3.1) は、確率  $p$  と出力と素性の組の出現を意味する関数  $g$  をかけることで出力と素性の組の頻度の期待値を求めることになっており、右辺の既知データにおける期待値と、左辺の求める確率分布に基づいて計算される期待値が等しいことを制約として、エントロピー最大化 (確率分布の平滑化) を行って、出力と文脈の確率分布を求めるものとなっている。

### 3.5 正規化 $\alpha$ 値

正規化  $\alpha$  値とは、最大エントロピー法で求まる  $\alpha$  値を全分類先での合計が 1 となるように正規化した値である。また、素性  $a$  と分類先  $b$  の対によって定まる値であり、素性  $a$

のみが適用される場合に分類先 b となる確率に相当する。各素性の、分類先ごとに与えられた正規化  $\alpha$  値が高いほど、その分類先であることを推定するのに重要な素性であることを意味する。

## 3.6 符号検定

本研究では、統計的検定に、符号検定を使用する。符号検定とは、勝ちと負けや表と裏といった二つの相反するペアに差があるかないかを+と-の符号を付け、その符号の出現率が二項確率に従うかどうかを検定する手法である。以下、符号検定の流れを示す。

### 1. 仮説の設定

帰無仮説 ( $H_0$ ): 「2 群間に差がない」と仮定する。

対立仮説 ( $H_1$ ): 「2 群間に差がある」と仮定する。

2 群間に差がなければ、+と-の符号がでる確率は両方とも  $\frac{1}{2}$  である。もし 2 群間に差があるならば、+と-の符号がでる確率はどちらかに偏る。

### 2. 有意水準の設定

有意水準とは、仮説を棄却するときの基準となる確率のことである。有意水準は 5% や 1% に設定する。

### 3. p 値を求める

$H_0$  が正しいという前提において、極端に偏った検定量が得られる p 値を求める。

### 4. $H_0$ を棄却する

$H_0$  が正しいという前提で求めた p 値が有意水準を下回ることで  $H_0$  の仮定を棄却し、 $H_1$  の仮説が立証することで、2 群間に差があることを証明する。この時初めて「統計的に有意差があった」といえる。

## 3.7 Word2vec

Word2vec は Google 社の Tomas Mikolov ら [7] によって提案されたニューラルネットワーク (Skip-gram) の手法である。Skip-gram は、文脈を利用して与えられた単語と与えられた単語の周辺 に出現する単語を予測できるように、単語ベクトルの学習を行うモデルである。意味的に関連が強い単語はベクトルが近くなる。例えば、「国語」、「数学」、「理科」、「社会」、「英語」は日本の中学校で学ぶ主要な 5 教科で似た単語としてベクトルが近くなる。単語をベクトル変換することで、人手で入力した単語と近いベクトルの単語 (類似した単語) を取得することができる。

## 第4章 知見獲得における実験と考察

本章では、本研究の知見獲得の実験データを4.1節で記述し、本研究で行った知見獲得の実験方法を4.2節で記述し、実験結果と実験の考察について4.3節に記述する。

### 4.1 実験データ

本研究では、実験データに2007年から2018年の毎日新聞のデータを用いる。実験に使用する単語対は「政治」と「経済」、「輸入」と「輸出」、「社会主義」と「資本主義」、「オリンピック」と「パラリンピック」を用いる。それぞれの単語対の実験に使用したデータ数を表4.1に示す。

表 4.1: 実験データ

単語対	データ数
政治と経済	20,000
輸入と輸出	29,938
社会主義と資本主義	2,358
オリンピックとパラリンピック	16,574

### 4.2 実験方法

3.4節と3.5節の方法を使って知見獲得の実験を行う。単語対をA、Bとして、知見の獲得の方法を以下に示す。

**手順1** Aを含む文を収集する。

**手順2** Bを含む文を収集する。

**手順3** 収集したA、B文の数に違いがあればランダムに整列してから同数にする。

手順4 3.4節と3.6節の方法を用いてAとBの素性を抽出する。

手順5 手法ごとに得られたA, Bの素性上位30個を取り出す。

手順6 手順5で得られた素性を分析する。

### 4.3 実験結果と考察

最大エントロピー法では10分割クロスバリデーションを行い、正解率を求めた。その結果を表4.2に示す。表4.2より正解率が8割を超える精度であった。

表 4.2: 実験結果

単語対	正解率
政治と経済	87.17
輸入と輸出	84.85
社会主義と資本主義	83.41
オリンピックとパラリンピック	92.00

最大エントロピー法で得られた単語対に対する素性30個を表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。符号検定で得られた単語対に対する素性30個を表4.11, 表4.12, 表4.13, 表4.14に示す。

表4.3と表4.4は最大エントロピー法における「政治」と「経済」の素性の収集の結果で、表4.11は符号検定における「政治」と「経済」の素性の収集の結果である。最大エントロピー法による「政治」に関する素性の収集は「討論」や「家」や「疑惑」などの政治に関連する語句の収集が行えている。また、「献金」や「天皇」や「トランプ」などの語句が収集されており、政治に関して知見があまりない人には役立つ素性が得られた。また、「経済」では「成長」や「産業」や「活性」など連語になるような関連語句の収集ができた。「アベノミクス」や「リーマン」「インフレ」などの語句は経済を知る上で重要な語句である。符号検定による「政治」に関する素性の収集は「家」や「選挙」や「献金」などの政治に関連する語句の収集が行えている。最大エントロピー法と比べると関連語句の収集がより多く行えている。しかし、「か」や「う」や「さん」などの政治に関連しない語句が収集されている。また、「経済」では「産業」や「成長」や「市場」などの多くの経済に関連する語句が収集できている。それぞれの素性は経済に関して一般的であるがあまり役に立たない語句である。

表 4.5 と表 4.6 は最大エントロピー法における「輸入」と「輸出」の素性の収集の結果で、表 4.12 は符号検定における「輸入」と「輸出」の素性の収集の結果である。最大エントロピー法による「輸入」に関する素性の収集は「制限」や「頼る」などの輸入に関連する語句の収集が行えている。輸入車の「マーチ」や「BMW」が収集されており輸入に関して知見があまりない人にとって役立つ素性が得られた。しかし「池」や「サウス」や「と共に」など輸入に関連のない語句が得られた。また、「輸出」では「OPEC」や「レアアース」などの輸出に関連する語句の収集できている。「コレラ」や「受精卵」は輸出に一見関連が無さそうであるが、輸出に関しての有用な素性である。符号検定による「輸入」に関する素性の収集は「関税」や「制限」や「産」などの輸入に関する多くの語句の収集が行えている。また「輸出」においても「経済」や「OPEC」や「為替」などの多くの輸出に関する語句の収集ができている。最大エントロピー法と比べるとより多くの関連語句が収集できていることがわかる。

表 4.7 と表 4.8 は最大エントロピー法における「社会主義」と「資本主義」の素性の収集の結果で、表 4.13 は符号検定における「社会主義」と「資本主義」の素性の収集の結果である。最大エントロピー法による「社会主義」に関する素性の収集は「ベトナム」や「キューバ」や「中国」や「ソ連」などの社会主義国の国名の収集が行えている。また、社会主義国の特色である「平等」が収集できており社会主義に関して知見があまりない人に役立つ素性が得られた。また、「資本主義」では資本主義にとって重要である「金融」や資本主義国家である「アメリカ」が収集できた。「恐慌」や「グローバル」は資本主義の特色であるため、資本主義についてあまり知見がない人にとって役立つ素性が得られた。また、社会主義と資本主義との素性を比較することでそれぞれの特徴的な素性が取り出せていることがわかる。符号検定による「社会主義」に関する素性の収集は最大エントロピー法と同様に「キューバ」や「ベトナム」や「ソ連」などの社会主義国の国名の収集が行えている。「左派」は社会主義の特色でありあまり知見がない人にとって役に立つ素性が得られている。しかし、「9」や「1」や「】」などの社会主義に関係のない語句が多く収集されている。また「資本主義」では最大エントロピー法と同様に資本主義にとって重要である「金融」が収集できた。最大エントロピー法では得られなかった「バブル」は資本主義の歴史的な出来事を知る上で重要な語句であり、知見があまりない人にとって役立つ素性だといえる。

表 4.9 と表 4.10 は最大エントロピー法における「オリンピック」と「パラリンピック」の素性の収集の結果で、表 4.14 は符号検定における「オリンピック」と「パラリンピック」の素性の収集の結果である。最大エントロピー法による「オリンピック」に

関する素性の収集は「JOC」や「IOC」や「OCA」などのオリンピックに関連する委員会の略称やロシアオリンピック委員長である「ジューコフ」や水泳競技で活躍している「北島」選手などの人名が収集されている。オリンピックに関して知見を得たい人にとっては役立つ素性である。また、「パラリンピック」ではパラリンピック委員会の略称である「IPC」や「JPC」が収集できた。「アルペンスキー」や「ボッチャ」などはパラリンピックの競技の一つであり、パラリンピックを観戦するきっかけになるような素性である。パラリンピック陸上競技で活躍している「佐藤」の人名が収集できた。符号検定による「オリンピック」に関する素性の収集は最大エントロピー法と同様に「JOC」や「IOC」などのオリンピックに関連する委員会の略称が収集できている。国際オリンピック委員会会長の「バッハ」氏や元国際オリンピック委員会委員の「竹田」氏の名前が収集できた。また、「パラリンピック」ではパラリンピック委員会の略称である「IPC」が収集できた。最大エントロピー法と同様に「アルペンスキー」はパラリンピック競技の一つであり、パラリンピックを観戦するきっかけに繋がる素性である。

表 4.3: 正規化  $\alpha$  値 (政治の上位 30 個)

素性	$\alpha$ 値	
	政治	経済
家	0.968443	0.031557
献金	0.963250	0.036750
カネ	0.962053	0.037947
釈放	0.917942	0.082058
犯	0.916634	0.083366
空白	0.914494	0.085506
決断	0.913377	0.086623
トランプ	0.913341	0.086659
疑惑	0.899198	0.100802
勝負	0.883225	0.116775
自主	0.878872	0.121128
天皇	0.873095	0.126905
介入	0.862312	0.137688
宗教	0.858922	0.141078
決着	0.856707	0.143293
不信	0.855970	0.144030
討論	0.855303	0.144697
働きかけ	0.853082	0.146918
真剣	0.850821	0.149179
参画	0.848286	0.151714
ムシャラフ	0.846637	0.153363
スタイル	0.845880	0.154120
駆け引き	0.845597	0.154403
歴訪	0.844527	0.155473
亡命	0.842529	0.157471
アサド	0.841522	0.158478
学術	0.838719	0.161281
信認	0.837837	0.162163
興味	0.836123	0.163877
吉永	0.835802	0.164198

表 4.4: 正規化  $\alpha$  値 (経済の上位 30 個)

素性	$\alpha$ 値	
	政治	経済
成長	0.012212	0.987788
産業	0.018125	0.981875
制裁	0.027451	0.972549
アベノミクス	0.036924	0.963076
損失	0.043601	0.956399
界	0.055139	0.944861
発展	0.063159	0.936841
評論	0.067460	0.932540
官庁	0.067678	0.932322
再生	0.070035	0.929965
活性	0.078331	0.921669
指標	0.082160	0.917840
封鎖	0.082611	0.917389
同友会	0.090558	0.909442
学部	0.103990	0.896010
疲弊	0.110713	0.889287
研究所	0.111938	0.888062
対策	0.113250	0.886750
リーマン	0.114126	0.885874
効果	0.114788	0.885212
水域	0.119452	0.880548
実体	0.124294	0.875706
先行き	0.125276	0.874724
好転	0.128663	0.871337
閉鎖	0.128932	0.871068
統計	0.129921	0.870079
インフレ	0.133191	0.866809
資本	0.135495	0.864505
関西	0.136014	0.863986
実業	0.136340	0.863660

表 4.5: 正規化  $\alpha$  値 (輸入の上位 30 個)

素性	$\alpha$ 値	
	輸入	輸出
頼っ	0.981405	0.018595
マーチ	0.956909	0.043091
逆	0.944326	0.055674
エチオピア	0.942129	0.057871
双	0.930359	0.069641
頼る	0.919903	0.080097
原材料	0.916640	0.083360
取締	0.906719	0.093281
敷設	0.901834	0.098166
自給	0.897627	0.102373
盤	0.887150	0.112850
慢性	0.880772	0.119228
博覧	0.879825	0.120175
付与	0.878230	0.121770
制限	0.877568	0.122432
集会	0.868799	0.131201
エレベーター	0.866319	0.133681
池	0.863840	0.136160
サウス	0.858697	0.141303
届け	0.858044	0.141956
切り替え	0.857908	0.14209
雑貨	0.856869	0.14313
通知	0.856817	0.143183
袋	0.855989	0.144011
BMW	0.855760	0.144240
武内	0.853711	0.146289
生態	0.851165	0.148835
由宇	0.850588	0.149412
と共に	0.850514	0.149486
化石	0.849602	0.150398

表 4.6: 正規化  $\alpha$  値 (輸出の上位 30 個)

素性	$\alpha$ 値	
	輸入	輸出
入	0.001529	0.998471
還付	0.011393	0.988607
O P E C	0.028486	0.971514
レアアース	0.047835	0.952165
武器	0.066961	0.933039
出稼ぎ	0.098927	0.901073
F D A	0.101283	0.898717
全農	0.112053	0.887947
サバ	0.114059	0.885941
コレラ	0.114122	0.885878
潤う	0.115400	0.884600
工芸	0.118654	0.881346
促進	0.121396	0.878604
狙う	0.123958	0.876042
水銀	0.124667	0.875333
分離	0.128149	0.871851
受精卵	0.128396	0.871604
満たす	0.129889	0.870111
タス通信	0.130390	0.869610
ウェイ	0.131391	0.868609
減産	0.132236	0.867764
順守	0.132457	0.867543
フィット	0.134445	0.865555
コシヒカリ	0.137360	0.862640
柿	0.138680	0.861320
インフラ	0.141487	0.858513
大田	0.142674	0.857326
精液	0.143074	0.856926
拿捕	0.143124	0.856876
遮断	0.143186	0.856814

表 4.7: 正規化  $\alpha$  値 (社会主義の上位 30 個)

素性	$\alpha$ 値	
	社会主義	資本主義
ベトナム	0.877773	0.122227
キューバ	0.868781	0.131219
特色	0.864560	0.135440
建設	0.825644	0.174356
ソ連	0.796728	0.203272
市場	0.794500	0.205500
実現	0.792832	0.207168
事件	0.786228	0.213772
全	0.785367	0.214633
モンゴル	0.783590	0.216410
旧	0.777432	0.222568
思想	0.775422	0.224578
民主党	0.771840	0.228160
間違い	0.770140	0.229860
政党	0.769802	0.230198
演説	0.769059	0.230941
回	0.759170	0.240830
強調	0.751139	0.248861
者	0.748307	0.251693
中国	0.746626	0.253374
左派	0.744638	0.255362
支持	0.739194	0.260806
若者	0.736897	0.263103
サンダース	0.735624	0.264376
計画	0.733819	0.266181
政権	0.730150	0.269850
強国	0.728835	0.271165
宣言	0.728673	0.271327
人民日報	0.727313	0.272687
平等	0.723119	0.276881

表 4.8: 正規化  $\alpha$  値 (資本主義の上位 30 個)

素性	$\alpha$ 値	
	社会主義	資本主義
金融	0.103321	0.896679
共産	0.143487	0.856513
グローバル	0.174572	0.825428
型	0.178802	0.821198
産業	0.197320	0.802680
世界	0.201226	0.798774
◆	0.217065	0.782935
会社	0.217912	0.782088
恐慌	0.226821	0.773179
暴走	0.235940	0.764060
東側	0.235940	0.764060
ん	0.236874	0.763126
著書	0.253963	0.746037
矛盾	0.259704	0.740296
利益	0.262056	0.737944
なき	0.263382	0.736618
—	0.265868	0.734132
アメリカ	0.269077	0.730923
香港	0.270501	0.729499
発言	0.270657	0.729343
問う	0.272996	0.727004
社会	0.274950	0.725050
新しい	0.275457	0.724543
躍進	0.276419	0.723581
維持	0.277681	0.722319
自然	0.278464	0.721536
人間	0.278773	0.721227
見える	0.280219	0.719781
ます	0.282854	0.717146
介入	0.284961	0.715039

表 4.9: 正規化 $\alpha$ 値 (オリンピックの上位 30 個)

素性	$\alpha$ 値	
	オリンピック	パラリンピック
JOC	0.983756	0.016244
IOC	0.983522	0.016478
スペシャル	0.972607	0.027393
ジュエコフ	0.912187	0.087813
OCA	0.910274	0.089726
委員	0.904504	0.095496
青少年	0.897462	0.102538
ス	0.892998	0.107002
アクアティクスセンター	0.891972	0.108028
北信越	0.883448	0.116552
大陸	0.876135	0.123865
ムーブメント	0.867295	0.132705
駒沢	0.860937	0.139063
古代	0.859589	0.140411
数学	0.847427	0.152573
ジュニア	0.844121	0.155879
恒和	0.842250	0.157750
興奮	0.838999	0.161001
北島	0.834524	0.165476
評議	0.827506	0.172494
体育	0.827031	0.172969
井沢	0.826670	0.173330
少年	0.826658	0.173342
高	0.826185	0.173815
昭和	0.825482	0.174518
演技	0.820819	0.179181
立候補	0.817465	0.182535
制	0.816870	0.183130
運	0.816084	0.183916
石原	0.810517	0.189483

表 4.10: 正規化 $\alpha$ 値 (パラリンピックの上位 30 個)

素性	$\alpha$ 値	
	オリンピック	パラリンピック
I P C	0.001154	0.998846
車いす	0.007046	0.992954
障害	0.016862	0.983138
J P C	0.018508	0.981492
パラ	0.057064	0.942936
中立	0.057200	0.942800
組織	0.059341	0.940659
車椅子	0.070936	0.929064
バリアフリー	0.072179	0.927821
ウクライナ	0.082331	0.917669
成田	0.083028	0.916972
国枝	0.086735	0.913265
・	0.088193	0.911807
障	0.091210	0.908790
アルペンスキー	0.095014	0.904986
クレーブ	0.105431	0.894569
ボール	0.117578	0.882422
純一	0.122034	0.877966
義足	0.129251	0.870749
喜朗	0.141125	0.858875
主	0.142435	0.857565
邦男	0.147205	0.852795
視覚	0.147908	0.852092
切断	0.148608	0.851392
マスケット	0.148699	0.851301
鳥	0.156418	0.843582
ボッチャ	0.156456	0.843544
健全	0.162136	0.837864
佐藤	0.162202	0.837798
中森	0.162862	0.837138

表 4.11: p 値 (政治と経済の下位 30 個)

政治		経済	
素性	p 値	素性	p 値
家	0.00E+00	産業	0.00E+00
選挙	2.72E-149	成長	0.00E-01
小沢	2.81E-131	景気	4.38E-163
規正	2.60E-113	金融	2.20E-161
議員	1.67E-111	市場	2.42E-156
献金	8.84E-104	財政	5.64E-141
資金	8.30E-103	省	2.80E-134
氏	2.90E-102	%	6.09E-131
団体	3.76E-91	投資	5.41E-116
政党	6.69E-90	生産	1.60E-108
民主党	6.03E-88	対策	1.98E-90
カネ	8.37E-84	世界	4.05E-90
国会	1.07E-77	回復	5.49E-90
選	6.25E-71	高	9.18E-86
自民党	6.70E-71	企業	1.80E-84
衆院	8.73E-67	消費	1.91E-83
党	6.94E-65	中国	2.64E-82
記載	3.53E-64	効果	2.78E-76
国民	5.72E-63	減速	1.11E-75
代表	1.24E-54	上昇	3.11E-74
ない	7.47E-54	輸出	5.58E-74
か	4.21E-53	発表	1.81E-69
私	1.40E-48	相	4.35E-68
一郎	9.10E-47	年度	9.65E-68
う	5.40E-46	安	2.72E-67
法	7.90E-46	比	1.12E-66
事務所	7.55E-45	価格	1.89E-62
憲法	6.09E-43	再生	2.60E-62
さん	3.49E-42	緩和	6.14E-62
民主	1.32E-41	G D P	9.04E-62

表 4.12: p 値 (輸入と輸出の下位 30 個)

輸入		輸出	
素性	p 値	素性	p 値
産	1.34E-197	景気	5.00E-229
関税	5.88E-174	経済	1.04E-181
制限	6.81E-164	成長	1.56E-179
発動	7.93E-138	企業	7.07E-176
食品	2.40E-134	武器	4.99E-174
価格	6.30E-123	入	9.79E-157
販売	2.59E-121	回復	5.64E-153
牛肉	3.05E-117	向け	1.43E-142
国産	2.37E-106	減速	6.17E-141
品	2.91E-98	関連	1.06E-127
ランプ	3.91E-87	株	2.75E-125
値上げ	3.02E-83	市場	6.02E-122
法	1.22E-76	投資	8.78E-119
厚生	3.49E-75	業績	2.71E-116
から	1.28E-73	生産	2.92E-108
剤	2.35E-70	O P E C	6.19E-107
検出	9.97E-68	為替	3.42E-104
アルミニウム	1.91E-67	高	4.05E-104
課す	2.84E-67	金融	6.80E-101
冷凍	3.44E-65	アジア	2.61E-92
業者	3.77E-65	電機	5.08E-91
労働省	9.77E-64	株価	1.53E-88
牛	7.35E-62	相場	9.89E-85
取締	5.08E-61	修正	2.59E-83
トン	2.85E-58	中心	1.61E-80
都	1.30E-56	機構	1.84E-79
ガード	7.11E-56	インフラ	1.26E-78
区	9.66E-55	G D P	4.28E-77
セーフ	9.45E-54	悪化	3.41E-74
会社	1.06E-53	期	2.55E-70

表 4.13: p 値 (社会主義と資本主義の下位 30 個)

社会主義		資本主義	
素性	p 値	素性	p 値
党	1.54E-35	金融	1.14E-29
政権	1.42E-22	グローバル	1.90E-22
9	4.03E-20	社会	4.23E-18
キューバ	8.70E-20	世界	2.03E-16
年	2.75E-19	か	6.70E-15
左派	1.01E-18	ない	1.94E-12
1	2.65E-18	』	3.63E-12
日	4.53E-18	型	3.69E-12
旧	9.89E-17	人間	5.24E-12
連立	4.44E-16	『	5.39E-12
ギリシャ	6.40E-16	その	9.44E-12
ソ連	8.80E-16	会社	1.54E-11
運動	1.04E-15	産業	1.90E-11
指導	1.43E-15	資本	2.58E-11
7	4.88E-15	それ	1.19E-10
2	1.30E-14	バブル	3.96E-09
大会	5.95E-14	近代	6.97E-09
】	6.65E-14	成長	8.49E-09
【	6.65E-14	という	9.36E-09
5	8.13E-14	日本	1.07E-08
4	9.05E-14	渋沢	1.49E-08
ベトナム	1.32E-13	消費	1.84E-08
体制	2.22E-13	著書	1.84E-08
特色	3.06E-13	学者	2.86E-08
大統領	4.54E-13	原発	3.47E-08
全	8.90E-13	な	4.75E-08
民主	1.31E-12	分析	5.04E-08
建設	1.58E-12	利益	5.04E-08
6	2.04E-12	水野	5.96E-08
軍	6.64E-12	里山	5.96E-08

表 4.14: p 値 (オリンピックとパラリンピックの下位 30 個)

オリンピック		パラリンピック	
素性	p 値	素性	p 値
JOC	0.00E+00	障害	0.00E-01
IOC	0.00E+00	・	6.81E-294
国際	0.00E+00	車いす	2.77E-194
委員	0.00E+00	東京	1.10E-134
会	0.00E+00	視覚	4.44E-91
総会	4.37E-171	大会	1.42E-82
会長	2.80E-109	I P C	3.61E-78
招致	1.32E-105	リオデジャネイロ	2.14E-63
理事	4.62E-101	0	1.74E-62
竹田	4.30E-96	者	7.87E-58
恒和	4.52E-95	2	3.13E-57
)	5.92E-95	年	1.77E-52
(	7.45E-95	向け	3.90E-52
立候補	2.02E-84	切断	1.72E-45
夏季	9.01E-66	T	1.36E-41
バツハ	1.01E-59	国枝	2.56E-41
連盟	9.00E-49	出場	2.77E-41
都市	1.25E-48	義足	3.05E-41
ローザンヌ	7.35E-45	位	3.08E-41
スイス	1.94E-40	バリアフリー	5.09E-41
共同	2.36E-38	組織	9.46E-41
評議	3.12E-38	主	1.78E-40
ロゲ	7.52E-37	整備	4.31E-40
処分	7.40E-35	会場	8.55E-39
ブエノスアイレス	1.49E-32	座	4.37E-38
O C A	4.31E-32	リオ	7.29E-38
承認	9.02E-30	成田	1.08E-37
全日本	7.31E-29	アルペンスキー	1.02E-36
マドリード	8.78E-29	場	2.88E-36

# 第5章 知見獲得の評価における実験と考察

本章では、本研究の知見獲得の評価実験の方法を5.1節で記述し、実験結果を5.2節に記述する。5.3節で実験の考察について記述する。

## 5.1 実験方法

4.2節で収集した単語対の手法ごとの素性を評価対象とする。実験方法は3.3節で述べた手法を用いる。

### 5.1.1 Word2vec との比較

**手順1** Word2vec に4節で使用した実験データ(毎日新聞2007年から2018年)を学習させる。

**手順2** 手順1で作成したモデルに単語対(A, B)を入力し、単語間の類似度を求める。

**手順3** AとBとの類似度が高い単語500個を収集する。

**手順4** 4.2節で収集した素性と手順3で収集した単語の一致数を数える。

### 5.1.2 人手評価

#### 1. 連想語句による評価

**手順1** 人手で単語対から連想する単語を30個書き出す。

**手順2** 手順1で収集した連想語30個と4.2節で収集した素性500個との一致数を数える。

## 2. 有益性の評価

**手順1** どちらの手法で得られた素性であるか不明にするために4.2節で収集した単語対の手法ごとの素性100個をマージして意味ソート [8] する.

**手順2** 手順1でソートした単語対別の素性を表3.1の評価基準を基に評価する.

## 5.2 実験結果と考察

### 5.2.1 Word2vec を用いた評価実験

最大エントロピー法で得られた単語対の素性500個とWord2vecで得られた単語対の素性500個との一致数と符号検定で得られた単語対の素性500個とWord2vecで得られた素性500個との一致数を表5.1に示す. 表5.1より, 最大エントロピー法で得られた単語対の素性よりも符号検定で得られた単語対の素性の方がWord2vecで得られた各単語対の素性との一致数が多い. 先述したとおり, Word2vecは単語をベクトル化することで入力された単語と意味の近い単語を収集することができるので, 単語対と関連性のある単語を収集することができる. 表5.1の結果で各単語対のWord2vecとの一致数で最大エントロピー法と符号検定のt検定の両側検定により有意差検定を行った. p値が0.00647であり有意水準5%以下であるので有意差がみられた. 従って, 最大エントロピー法よりも符号検定の方が関連語句を取り出すのに適していることがわかる.

表 5.1: Word2vec で得られた素性 500 個との一致数

単語対	ME	符号検定
政治	35	94
経済	65	160
輸入	47	101
輸出	39	109
社会主義	5	4
資本主義	4	3
オリンピック	30	60
パラリンピック	20	31
平均	30.6	70.3

## 5.2.2 人手による評価実験

筆者が人手による評価実験を行った。あらかじめ単語対から連想する語句を人手で30個用意する。その連想語句と最大エントロピー法と符号検定で得た単語対の素性の一致数を表5.2に示す。次に最大エントロピー法と符号検定で収集した単語対の素性100個を先述した評価基準(表3.1)を基に評価した結果を表5.3に示す。評価実験の比率を表5.4に示す。また、その単語対別の評価例を表5.5, 表5.6, 表5.7, 表5.8, 表5.9, 表5.10, 表5.11, 表5.12に示す。表5.2より人が連想した単語対の連想語句は最大エントロピー法で得た単語対の素性よりも符号検定で得た単語対の素性の方が一致数が多いことがわかった。表5.3では、人が役に立つと考える単語対の素性は符号検定よりも最大エントロピー法の方が多く収集できていることがわかる。つまり、人が連想できる一般的な素性の収集においては符号検定の方が最大エントロピー法よりも優れており、新たな知見や役に立つ知見につながりやすい素性の収集は最大エントロピー法であるということがわかった。表5.3の結果における最大エントロピー法と符号検定の各単語対の◎の個数に有意差があるかを調べるために、t検定の両側検定による有意差検定を行った。p値が0.0473であり有意水準5%以下であるので有意差がみられた。

表 5.2: 連想語句一致数

単語対	ME	符号検定
政治	6	15
経済	11	18
輸入	4	6
輸出	2	9
社会主義	15	14
資本主義	12	12
オリンピック	3	2
パラリンピック	10	14
平均	7.9	11.3

表 5.3: 評価結果

単語対	ME			符号検定		
	◎	○	×	◎	○	×
政治	19	66	15	5	67	28
経済	14	63	23	11	75	14
輸入	14	67	19	10	75	15
輸出	24	60	16	2	87	11
社会主義	6	73	21	3	67	30
資本主義	10	59	31	15	48	37
オリンピック	6	58	36	5	73	22
パラリンピック	18	57	25	6	65	29
平均	13.9	62.9	23.3	7.1	69.6	23.3

表 5.4: 評価実験の比率

ME			符号検定		
◎	○	×	◎	○	×
0.14	0.63	0.23	0.07	0.70	0.23

表 5.5: 政治

ME			符号検定		
◎	○	×	◎	○	×
パーティー ミサイル ネット 水俣病	家 官僚 省庁 法的	視野 おき 諸 本当	パーティー 虚偽 トランプ 発言	家 国民 政党 与野党	それ 私 何 変

表 5.6: 経済

ME			符号検定		
◎	○	×	◎	○	×
APEC インフレ EPA アベノミクス	大国 課 好転 統計	出さ 基礎 明かし 訪れ	日銀 原油 資源 アベノミクス	危機 悪化 エネルギー 緩和	野球 など 化 ・

表 5.7: 輸入

ME			符号検定		
◎	○	×	◎	○	×
LNG 銃 仮想	物価 規制 原材料	池 袋 学生	トランプ アルミニウム インフルエンザ	業者 関税 原材料	い 回収 患者

表 5.8: 輸出

ME			符号検定		
◎	○	×	◎	○	×
精液 受精卵 コレラ	OPEC 国税局 半導体	相違 ホワイト 工	債務 レアアース	危機 市場 半導体	感 実質 前日

表 5.9: 社会主義

ME			符号検定		
◎	○	×	◎	○	×
ナチス 東西 保障	国 市民 政党	彼ら ほとんど 基本	左派 中国共産党 キューバ	国 統一 政治	圏 ら 年

表 5.10: 資本主義

ME			符号検定		
◎	○	×	◎	○	×
バブル 競争 日本	文明 課税 金融	なき 東側 ず	バブル 格差 欲望	経済 利潤 金融	なき べき ます

表 5.11: オリンピック

ME			符号検定		
◎	○	×	◎	○	×
赤字 ロボット	IOC NOC	ダム 猫	憲章 アカデミー	IOC 処分	副 ファイル

表 5.12: パラリンピック

ME			符号検定		
◎	○	×	◎	○	×
WADA アルペンスキー	JPC 車椅子	弱 提出	ボッチャ アルペンスキー	義足 JPC	個 立

### 5.2.3 被験者による評価実験

より客観的な評価を得るため、筆者以外の学生3人による被験者実験を行った。被験者が、あらかじめ単語対から連想する語句を30個を書き出し、その連想語句と最大エントロピー法と符号検定で得た単語対の素性の一致数を表5.13に示す。また、被験者3人の連想語句一致数の平均を表5.14に示す。最大エントロピー法と符号検定で収集した単語対の素性100個を先述した評価基準(表3.1)を基に被験者が人手で評価した結果を表5.15に示す。また、表5.15より被験者による評価の平均を表5.16、比率を表5.17に示す。表5.14より、人が連想した単語対の連想語句は最大エントロピー法で得た単語対の素性よりも符号検定で得た単語対の素性の方が一致数が多いことがわかる。表5.17より、人が役に立つと考える単語対の素性の収集は最大エントロピー法の方が符号検定に比べて多く収集できている。また、表5.15の結果において最大エントロピー法と符号検定の各単語対の◎の個数の差が統計的に有意であるかを確認するために、被験者3名の単語対の合計24対の最大エントロピー法と符号検定の◎の個数でt検定の両側検定により有意差検定を行った。p値が0.0137であり有意水準0.05%以下であるので有意差がみられた。これらの結果は、5.2.2節で示した筆者による評価実験と同じ傾向があることがわかる。被験者が行った単語対別の評価の例を表5.18、表5.19、表5.20、表5.21、表5.22、表5.23、表5.24、表5.25に示す。

表 5.13: 被験者による連想語句一致数

単語対	被験者 1		被験者 2		被験者 3		平均	
	ME	符号検定	ME	符号検定	ME	符号検定	ME	符号検定
政治	6	12	5	16	4	8	5.0	12.0
経済	13	23	17	22	9	17	13.0	20.7
輸入	4	6	4	6	2	5	3.3	5.7
輸出	0	2	0	5	1	7	0.3	4.7
社会主義	15	12	11	15	9	14	11.7	13.7
資本主義	5	10	7	9	8	15	6.7	11.3
オリンピック	6	7	3	9	4	5	4.3	7.0
パラリンピック	11	12	6	11	11	12	9.3	11.7

表 5.14: 被験者による連想語句一致数の平均

ME	符号検定
6.7	10.4

表 5.15: 被験者による評価結果

	被験者 1						被験者 2						被験者 3					
	ME			符号検定			ME			符号検定			ME			符号検定		
単語対	◎	○	×	◎	○	×	◎	○	×	◎	○	×	◎	○	×	◎	○	×
政治	7	59	34	3	67	30	8	45	47	5	59	36	7	59	34	3	67	30
経済	12	60	28	10	74	16	14	63	23	11	75	14	5	49	46	3	75	22
輸入	3	38	59	2	50	48	8	36	56	2	50	43	10	21	69	6	40	54
輸出	3	39	58	2	41	57	21	28	51	19	43	38	15	17	68	3	36	61
社会主義	11	55	34	9	58	33	12	48	39	7	48	45	13	38	49	11	40	49
資本主義	5	43	52	7	44	59	2	35	63	2	38	59	5	28	67	4	32	64
オリンピック	1	57	42	3	71	26	4	39	56	2	54	44	6	45	49	6	61	33
パラリンピック	5	49	46	4	50	46	12	30	59	7	36	55	15	26	59	10	40	50
平均	5.9	50.0	44.1	5.0	56.8	38.1	10.0	40.5	49.3	6.9	50.4	41.8	9.5	35.4	55.1	5.8	48.9	45.4

表 5.16: 被験者による評価実験の平均

単語対	ME			符号検定		
	◎	○	×	◎	○	×
政治	7.3	54.3	38.3	3.7	64.3	32.0
経済	10.3	57.3	32.3	8.0	74.5	17.3
輸入	7.0	31.7	61.3	3.3	46.7	48.3
輸出	13.0	28.0	59.0	8.0	40.0	52.0
社会主義	12.0	47.0	40.7	9.0	48.7	42.3
資本主義	4.0	35.3	60.7	4.3	38.0	60.7
オリンピック	3.7	47.0	49.0	3.7	62.0	34.3
パラリンピック	10.7	35.0	54.7	7.0	42.0	50.3

表 5.17: 被験者による評価実験の比率

ME			符号検定		
◎	○	×	◎	○	×
0.085	0.42	0.49	0.059	0.52	0.42

表 5.18: 政治

ME			符号検定		
◎	○	×	◎	○	×
天皇 軍人	介入 結成	スタイル つなげる	トランプ 虚偽	家 国民	元 い

表 5.19: 経済

ME			符号検定		
◎	○	×	◎	○	×
リーマン 原子力	証券 バブル	基地 一段	原子力 資源	通貨 株価	会議 発表

表 5.20: 輸入

ME			符号検定		
◎	○	×	◎	○	×
化石 トランプ	ワクチン 規制	ハマス サウス	トランプ 鉄鋼	冷凍 関税	患者 回収

表 5.21: 輸出

ME			符号検定		
◎	○	×	◎	○	×
サバ IKEA	OPEC 半導体	精米 名誉	悪化 武器	危機 半導体	基調 堅調

表 5.22: 社会主義

ME			符号検定		
◎	○	×	◎	○	×
クーデター 共和	国 政党	ごろ 伝え	ソ連 ベトナム	国家 政治	圏 年

表 5.23: 資本主義

ME			符号検定		
◎	○	×	◎	○	×
発達 バブル	文明 マネー	理事 街	成長 バブル	格差 マネー	なき ます

表 5.24: オリンピック

ME			符号検定		
◎	○	×	◎	○	×
勇気 平和	IOC NOC	主義 知的	憲章 総会	IOC 処分	数学 ス

表 5.25: パラリンピック

ME			符号検定		
◎	○	×	◎	○	×
ブラインド バリアフリー	義足 アルペンスキー	ぜひ 未明	車椅子 バリアフリー	エンブレム メダル	年 機能

## 5.2.4 評価実験に対する考察

本節では、5.2.1 節、5.2.2 節、5.2.3 節に示した評価実験の結果に対する考察を記述する。

5.2.1 節で記述した Word2vec を用いた評価実験では、最大エントロピー法で収集した単語対の素性と Word2vec で収集した単語対の素性との一致数の平均が 30.6 個、符号検定で収集した単語対の素性と Word2vec で収集した単語対の素性との一致数の平均値が 70.3 個であり、符号検定で収集した単語対の素性の方が最大エントロピー法で収集した素性よりも Word2vec で収集した単語対の素性との一致数が多かった。また、5.2.2 節と 5.2.3 節で記述した人手評価における連想語句との一致数においては、筆者の連想語句一致数の平均が最大エントロピー法が 7.9 個、符号検定が 11.3 個、被験者 3 名による連想語句との一致数の平均は最大エントロピー法が 6.7 個、符号検定が 10.4 個であった。よって、符号検定は人が考えつくような一般的な語句の収集には優れていることがわかった。単語対に関わる一般的な語句は、単語対と共起する回数が多くなるため、単に共起回数を基に統計的に収集した符号検定の方が最大エントロピー法よりも一般的で人が考えつくような語句の収集には優れていると考えた。

5.2.2 節と 5.2.3 節で記述した有益性における人手評価においては、筆者の最大エントロピー法の評価の比率は、◎が 0.14、○が 0.63、×が 0.23 で、符号検定の評価の比率は、◎が 0.07、○が 0.70、× 0.23 であった。また、被験者の最大エントロピー法の評価の比率は、◎が 0.085、○が 0.42、×が 0.49 で符号検定の評価の比率は、◎が 0.059、○が 0.52、×が 0.42 であった。最大エントロピー法の方が符号検定よりも◎の比率が多いため、最大エントロピー法が符号検定に比べて役立つ単語対の素性の収集には優れていることがわかった。最大エントロピー法は、学習を繰り返し、単語対の分類の推定を行うことで、その手がかりとなる素性を取り出しているため、単に単語対と共起する単語を収集している符号検定に比べれば、より多くの役に立つ素性の収集ができるのではないかと考えた。評価方法の妥当性を調べるために、被験者 3 名の評価結果の一致度を表す Kappa 値 [9] を計算した。Kappa 値は 0.44 であった。役に立つと考えられる素性は人によって異なると考えられるが、被験者間の評価は適度に一致していることがわかり、評価方法にはある程度の妥当性があったことがわかる。

筆者と被験者の有益性の評価で一致数が多かった例として、最大エントロピー法で得た政治の素性では「パーティー」や「ネット」や「天皇」、経済では「リーマン」や「アベノミクス」が筆者、被験者ともに◎の個数が多く、役に立つ素性として考えられ

た。また、符号検定を用いた有益性の評価では、政治では「パーティー」や「トランプ」や「虚偽」、経済では「日銀」や「原油」や「資源」が筆者、被験者ともに◎の個数が多く、役に立つ素性として考えられた。

表 5.13 より輸出の連想語句と最大エントロピー法で収集した素性の一致数が 0.3 個で最も低いが、表 5.16 より、最大エントロピー法で収集した輸出の素性の◎の個数が 13.0 個で最も多かった。このことから、輸出が最も最大エントロピー法の知見獲得の有用性を示すことのできたキーワードであることがわかる。最大エントロピー法で得た輸出の素性の中で被験者が◎として評価した例は、「サバ」や「精液」や「受精卵」などがあった。主な輸出品以外のあまり知られていない輸出品などが多く、新たな知見に繋がるような語句が多く収集できたのではないかと考えた。

有益性の評価において、符号検定と比較することで最大エントロピー法の有用性を示すことができたが、最大エントロピー法の◎の比率が 0.10 付近であり、絶対的に有用な方法とは言い難い。最大エントロピー法を用いてより多くの役に立つ素性を得るために、正解率を向上させた学習データを用いることが今後の課題である。本研究では、Word2vec を用いた評価実験を行うために 1 単語の素性で機械学習を行ったが、1 から 3 単語連続の素性を利用することで正解率の向上を期待できるのではないかと考えた。また、1 単語のみではなく 1 から 3 単語連続の単語であれば人が考えつく一般的な知見よりも多くの役に立つ知見が得られるのではないかと考える。

## 第6章 今後の課題

本研究では、教師あり機械学習と統計的検定を用いて、様々な分野に関する知見の獲得とその評価実験を行ったが、いくつか問題がある。本章では、その問題を今後の課題として以下にまとめる。

- 本研究では、実験データに毎日新聞を用いて知見獲得を行ったが、新聞データ以外で知見獲得ができるのかを調査したい。
- 教師あり機械学習による知見の獲得は、有益な素性が得られた一方で、一部単語対に関係のない素性が得られた。より有益性のある素性が得られるように、正解率を向上させた学習データを用いたい。
- 知見の獲得に教師あり機械学習の最大エントロピー法を用いたが、その他の教師あり機械学習を用いた知見の獲得とその評価が行えていない。教師あり機械学習である BERT や SVM を用いて知見獲得を行い、知見獲得の評価実験を行いたい。

## 第7章 おわりに

本研究では、教師あり機械学習と統計的検定を用いて様々な分野の知見の獲得とその評価実験を行った。

知見獲得における研究では、教師あり機械学習と統計的検定を用いて素性分析を行い、様々な分野に関する知見を獲得することができた。

知見獲得の評価における研究では、教師あり機械学習と統計的検定による知見獲得の評価を行った。Word2vec で得た単語対の素性 500 個とそれぞれの手法で得た単語対の素性 500 個との一致数の平均は、最大エントロピー法が 30.6 個、符号検定が 70.3 個であった。被験者が行った、人手で連想した各単語対の語句 30 個とそれぞれの手法で得た単語対の素性 500 個との一致数の平均は、最大エントロピー法が 7.9 個、符号検定が 11.3 個であった。また、同様の被験者実験での結果では、最大エントロピー法が 6.7 個、符号検定が 10.4 個であった。それぞれの手法で得た各単語対の素性 100 個を筆者が人手で評価した結果の有益性の比率は、最大エントロピー法の◎が 0.14、符号検定の◎が 0.07 であった。また、同様の被験者実験での結果では、最大エントロピー法の◎が 0.85、符号検定の◎が 0.059 であった。単語対と類似度が高い素性や人が連想できる語句など一般的な語句の収集には統計的検定が優れていることがわかった。一方、人が役に立つと感じる、知見につながるような語句の収集には教師あり機械学習が優れていることがわかった。

# 謝辞

本研究を進めるに当たり, 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, 鳥取大学工学部電気情報系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へ感謝の意を表します.

## 参考文献

- [1] 村田真樹, 中原裕人, 馬青. パターンと教師あり機械学習と素性分析を利用したウェブと新聞からの株式相場に関わる知見獲得. 情報処理学会第 82 回全国大会, 2020.
- [2] 赤江涼太. 機械学習を用いた類義語の使い分けに関する知識獲得. 卒業論文, 鳥取大学工学部知能情報工学科, 2018.
- [3] 元田浩, 鷺尾隆. 機械学習とデータマイニング (大規模データベースからの知識獲得). 人工知能学会誌, Vol. 12, No. 4, pp. 505–512, jul 1997.
- [4] Eric Sven Ristad. Maximum entropy modeling for natural language. In *ACL/EACL Tutorial Program, Madrid*, 1997.
- [5] Masao Utiyama. Maximum entropy modeling packagen: <http://www.nict.go.jp/x/x161/members/mutiyama/software.htmlmaxent>, 2006.
- [6] Masaki Murata, Kiyotaka Uchimoto, Masao Utiyama, Qing Ma, Ryo Nishimura, Yasuhiko Watanabe, Kouichi Doi, and Kentaro Torisawa. Using the maximum entropy method for natural language processing: Category estimation, feature extraction, and error correction. *Cognitive Computation*, Vol. 2, No. 4, pp. 272–279, 2010.
- [7] 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*, p. 3111–3119, 2013.
- [8] 村田真樹, 神崎享子, 内元清貴, 馬青, 井佐原均. 意味ソート msort : 意味的並べかえ手法による辞書の構築例とタグつきコーパスの作成例と情報提示システム例. 自然言語処理, Vol. 7, No. 1, pp. 51–66, 2000.
- [9] Ebm のための臨床疫学入門講座. <http://kamiyacho.org/ebm/ce201.html>.