2022年度（令和4年度）　卒業論文

機械学習を用いた
上位語と下位語の使い分けと知見獲得

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

B19T2033Z 清原　詩央里

概要

本研究は上位語と下位語の使い分けを教師あり機械学習を使用して行う．

上位語とは，ある語についてより一般的で総称的な語のことであり，下位語とは，ある語についてより具体的なものを指す語のことである．本研究では，機械学習の性能や素性が上位語と下位語の使い分けに役立つと考え，自動的に上位語と下位語を使い分けることを目的とする．そのため，文中の上位語と下位語にあたる語を X とおき，機械学習を用いて X とした部分に上位語と下位語のどちらが入るのかを推定することで上位語と下位語の使い分けを行う．この際，再現率が高い場合は，機械でも判別できるため，その上位語と下位語は使い分けが必要であり，一方で，再現率が低い場合は，機械では判別できないため，その上位語と下位語は使い分けが不要であると考えられる．よって，機械学習の性能は使い分けの必要性と正の相関があると考えられる．本研究では，機械学習に最大エントロピー法と BERT を使用する．

また，機械学習が使用した素性を分析して，上位語と下位語の使い分けに役立つ情報の考察も行う．このような実験と調査を自らが選出した上位語と下位語の対を対象に行う．本研究の成果は 2 つある．1 つ目は，今回行った機械学習の性能がよく，31 対の上位語と下位語の対を用いた実験において，機械学習を用いた提案手法は，最大エントロピー法を用いた場合は 0.81，BERT を用いた場合は 0.85 の正解率であった．2 つ目は，機械学習での性能に基づき上位語と下位語の対を使い分けが必要なものとそれほど必要でないものに分類したことである．今回の実験で，使い分けが必要とされた上位語と下位語の対には「都道府県」と「鳥取」や「季節」と「秋」などの対があり，使い分けが不要とされた上位語と下位語の対には「穀物」と「小麦」や「乳製品」と「バター」などの対があった．また，いくつかの上位語と下位語の対について実際に使い分けに役立つ情報を明らかにした．

目 次

表 目 次

第1章　はじめに

上位語とは，ある語についてより一般的で総称的な語のことであり，下位語とは，ある語についてより具体的なものを指す語のことである．上位語と下位語の対の例としては「学校」と「大学」などがある．また，上位語と下位語の使い分けについて説明する．例えば,「この X には学科が 4 つあります」という文章があるとする．その際,まず，文章中の X に,「学校」と「大学」のどちらの単語を当てはめるべきかを考え,「文章中に『学科』という単語があるため,『大学』を当てはめる」のように推定する．このように，文の構造や文章中の単語などからより適切な単語を推定することを上位語と下位語の使い分けと言う．

単語対の使い分けに関する研究では，織金の動詞・形容詞の類義語の使い分け [1] や，日笠の 3，4 組の単語の使い分けの研究 [2] などがある．また，単語対の使い分けに似た，単語対の置き換え可否判定の研究として，佐々本の最大エントロピー法を用いた対義語の置き換え可否判定の研究 [3] や，小西の BERT を用いた対義語の置き換え可否判定の研究 [4] などがある．

本研究では，機械学習の性能や素性が上位語と下位語の使い分けに役立つと考え，機械学習を用いて上位語と下位語の使い分けを行う．本研究の成果は，文章を生成する際の上位語と下位語の選択，適切な表現の使い分けの提案などに利用できると考える．例えば,「食事をメニューにふさわしい X に盛り付ける」という文章を書く場合，X に「食器」と「皿」のどちらを当てはめるべきなのか分かれば，他者により伝わりやすい文章にすることができると考えている．

本研究では，新聞記事に出現する単語の中で，自ら考えた上位語と下位語の対を利用する．

本研究では，機械学習の性能や素性が上位語と下位語の使い分けに役立つと考え，自動的に上位語と下位語を使い分けることを目的とする．素性とは，教師あり機械学習が識別のために用いる情報のことであり，本研究では文中の単語を素性として使用した．本研究では，文中の上位語と下位語にあたる語を X とおき，次に機械学習を用いて，X とした部分に上位語と下位語のどちらが入るのかを推定することで上位語と

下位語の使い分けを行う．機械学習には，最大エントロピー法と BERT を使用する．

本研究の主な主張点を以下に整理する．

- 本論文は上位語と下位語の使い分けのために機械学習を使用し，複数の上位語と下位語の対について，どの程度使い分けが必要か，またどのような場合に使い分けが必要かなどを示した．使い分けが必要な場合として，機械学習を用いた上位語と下位語のそれぞれの使い分けにおいて参考にした素性は，下位語の素性には他の下位語が出現しやすいという知見が得られた．例としては,「バター」の素性である「脱脂粉乳」,「キャベツ」の素性である「レタス」などがある．

- 上位語と下位語の対 31 対について，機械学習を用いて，上位語と下位語の使い分けを行った．31 対の上位語と下位語の対を用いた実験において，機械学習を用いた提案手法は，最大エントロピー法を用いた場合は 0.81，BERT を用いた場合は 0.85 の正解率であった．また，本研究の機械学習の使い分けの正解率は一番低いものが 0.64，被験者実験の使い分けの正解率の一番低いものが 0.77 であった．これに対し，先行研究である織金 [1] の類義語対の使い分けにおける，機械学習の再現率低の一番低い分類は 3 割未満，被験者実験の使い分けの正解率の最低値は 0.45 と低かった．このことから，上位語と下位語の対は，先行研究で行われていた類義語対に比べて，全体的に使い分けが必要な単語対であると言える．

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，本研究が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，第 4 章の結果から素性分析による考察を行う．第 6 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，織金が行った類義語に対する機械学習を用いた動詞と形容詞の使い分けについて記述する．2.2 節では，日笠が行った 3，4 組の単語に対する機械学習を用いた使い分けについて記述する．2.3 節では，佐々本が行った対義語に対する機械学習を用いた置き換え可否判定について記述し，2.4 節では，小西が行った対義語に対する機械学習を用いた置き換え可否判定について記述する．

2.1　機械学習を用いた動詞・形容詞の類義語の使い分け

織金は，機械学習による動詞と形容詞の類義語の使い分けの研究を行った [1].

織金は動詞と形容詞の類義語の使い分けのために機械学習の最大エントロピー法を使用し，複数の動詞と形容詞の類義語対について，どの程度使い分けが必要か，またどのような場合に使い分けが必要かなどを新たに示した．

織金は動詞類義語対の獲得には EDR 電子化辞書と 1991 年から 1995 年の 5 年分の毎日新聞を使用し，形容詞類義語対の獲得には上記の年数に加えて 2011 年から 2015 年の 10 年分の毎日新聞を使用し，以下の条件を満たす動詞と形容詞の類義語を獲得した．

条件 1 その二つの語が，日本語単語辞書において，同一の概念識別子をもつこと

条件 2 その二つの語が動詞では 1991 年から 1995 年の 5 年分の新聞で出現頻度が 50 回以上であること，形容詞では 1991 年から 1995 年と 2011 年から 2015 年の 10 年分の新聞で出現頻度が 20 回以上であること

条件 3 形態素解析システム JUMAN[5] を用いて解析した結果，その二つの語の代表表記が異なること

獲得した動詞と形容詞の類義語対について，類義語対ごとに類義語の使い分けの実験を行った．入力文は，動詞では 1991 年から 1995 年の 5 年分の毎日新聞を，形容詞で

は上記に加えて 2011 年から 2015 年の 10 年分の毎日新聞から獲得した，類義語対のいずれかの語を含む文である．評価は 10 分割のクロスバリデーションで行った．機械学習の再現率の高さごとに動詞と形容詞の類義語対を，高・中・低に分類し，機械学習における素性 (学習に用いる情報のこと) を分析することで動詞と形容詞の類義語の使い分けに重要な情報を把握した．

織金の研究の成果として，機械学習を用いた動詞と形容詞の類義語の使い分けの手法自体が，動詞と形容詞の類義語の使い分けに有効であることを示した．更に，機械学習での性能に基づき使い分けが必要な動詞と形容詞の類義語対とそれほど必要でない動詞と形容詞の類義語対を明らかにした．また，実際に素性を分析した．使い分けに役立つ情報を明らかにし，さらにどのような場合に使い分けの必要があるかを明らかにした．特に使い分けが必要な動詞と形容詞の類義語対として「探し回る」と「探し求める」や「近しい」と「むつまじい」の対があり，使い分けが必要でない類義語対に「はみ出す」と「はみ出る」や「気まずい」と「面はゆい」の対があった．

2.2 機械学習を用いた 3，4 組の単語における使い分けと知見獲得

日笠は，機械学習による 3，4 組の単語の使い分けの研究を行った [2]．

日笠は 3，4 組の単語の使い分けのために機械学習の最大エントロピー法を使用し，獲得した 3，4 組の単語について，どの程度使い分けが必要か，またどのような場合に使い分けが必要かなどを新たに示した．

日笠は自らが選出した 3，4 組の単語を利用し，それらの 3，4 組の単語について，単語組ごとに使い分けの実験を行った．入力文は 1991 年から 1995 年と 2011 年から 2015 年の計 10 年分の毎日新聞から獲得した，3，4 組の単語のいずれかの語を含む文である．評価は 10 分割のクロスバリデーションで行った．機械学習の再現率の高さごとに 3，4 組の単語を，8 割以上，7 割以上 8 割未満，6 割以上 7 割未満，5 割以上 6 割未満に分類し，機械学習における素性 (学習に用いる情報のこと) を分析することで 3，4 組の単語の使い分けに重要な情報を把握した．

日笠の研究の成果として，機械学習を用いた 3，4 組の単語の使い分けの手法自体が，3，4 組の単語の使い分けに有効であることを示した．また，実際に素性を分析した．使い分けに役立つ情報を明らかにし，さらにどのような場合に使い分けの必要があるかを明らかにした．例えば,「上」,「中」,「下」では「事実上」や「活動中」,「水面下」

という使われ方をするの対し,「左」,「右」ではそのような使われ方がなされないということがわかった.

2.3 機械学習を用いた対義語の置き換え可否判定

佐々本は，機械学習の性能や素性が対義語対の使い分けに役立つと考え，機械学習を用いて対義語対の使い分けを行い，その結果を用いて対義語対の置き換え可否を判定する研究を行った [3].

佐々本は対義語の使い分けのために機械学習の最大エントロピー法を使用し，複数の対義語対について，どの程度置き換えが可能なのか，また，どのような場合に置き換えが可能かなどを新たに示した.

佐々本は荻原らの研究 [6] でまとめられた，対義語データベースから被験者実験で 20 人中 17 人以上が対義語であると判断されたものを利用した．名詞，動詞，形容詞，副詞の対義語対の中から新聞記事の中に両方の対義語が 50 回以上出現したものからランダムに 50 対を抽出し，実験を行う．獲得した対義語対について，対義語対ごとに対義語の置き換え可否判定の実験を行った．入力文は，1991 年から 1995 年と 2011 年から 2015 年の計 10 年分の毎日新聞から獲得した，対義語の組のいずれかの語を含む文である．評価は 10 分割のクロスバリデーションで行った．機械学習の再現率の高さごとに対義語対を，高・中・低に分類し，機械学習における素性 (学習に用いる情報のこと) を分析することで対義語の置き換え可否に重要な情報を把握した.

佐々本の研究の成果として，機械学習の性能が高ければ置き換え可能であり，機械学習の性能が低ければ置き換え不可能であるという，機械学習の性能と置き換え可否に逆の相関があることが確認できた．更に，それぞれの対義語における有用な知見や対義語対の置き換え可否に関する知見などを獲得した.

2.4 BERT を用いた対義語の置き換え可否判定

小西は，機械学習の性能や素性が対義語対の使い分けに役立つと考え，機械学習を用いて対義語対の使い分けを行い，その結果を用いて対義語対の置き換え可否を判定する研究を行った [4].

小西は対義語の使い分けのために機械学習の BERT を使用し，複数の対義語対について，どの程度置き換えが可能なのか，また，どのような場合に置き換えが可能かな

どを新たに示した．

小西は荻原らの研究 [6] でまとめられた，対義語データベースから被験者実験で 20 人中 17 人以上が対義語であると判断されたものを利用した．名詞，動詞，形容詞，副詞の対義語対の中から新聞記事の中に両方の対義語が 50 回以上出現したものからランダムに 50 対を抽出し，実験を行う．獲得した対義語対について，対義語対ごとに対義語の置き換え可否判定の実験を行った．入力文は，1991 年から 1995 年と 2011 年から 2015 年の計 10 年分の毎日新聞から獲得した，対義語の組のいずれかの語を含む文である．評価は 10 分割のクロスバリデーションで行った．機械学習の再現率の高さごとに対義語対を，高・中・低に分類し，機械学習における素性 (学習に用いる情報のこと) を分析することで対義語の使い分けに重要な情報を把握した．

小西の研究の成果として，機械学習の性能が高ければ置き換え可能であり，機械学習の性能が低ければ置き換え不可能であるという，機械学習の性能と置き換え可否に逆の相関があることが，佐々本 [3] の最大エントロピー法と同様に BERT でも確認できた．更に，それぞれの対義語における有用な知見や対義語対の置き換え可否に関する知見などを獲得した．

第3章　問題設定と提案手法

本章では，本研究で扱う問題と提案手法の説明を記述する．3.1 節では，本研究で扱う問題設定について記述している．3.2 節では，提案手法の大まかな流れについて記述し，3.3 節では，本研究で使用する機械学習法である最大エントロピー法についての説明を，3.4 節では，本研究で使用する機械学習法である BERT についての説明を記述している．3.5 節では，機械学習で使用する素性について記述している．

3.1　問題設定

使い分けをしたい上位語と下位語 A，B があるとする．語 A と語 B のことを対象語と呼ぶ．対象語のいずれかを含む文を収集する．収集した文において対象語を削除し，対象語があった箇所に対象語のうちのどの語が存在したかを推定することが，本研究で扱う問題である．その文に元々あった方の語を選択できれば，正しく上位語と下位語を使い分けることができたと考える．具体的な例として,「都道府県」と「鳥取」の例を以下に挙げる．

この結果四十七都道府県のうち中央官僚と政府関係機関の出身者が二十七人になった。富山、石川、福井、京都、兵庫、鳥取の六府県から約百九十隻の沖合底引き網漁船が出漁した。

このように対象語を含んだ文を収集する．次にこれらの文から対象語を削除する．

この結果四十七 X のうち中央官僚と政府関係機関の出身者が二十七人になった。富山、石川、福井、京都、兵庫、 X の六府県から約百九十隻の沖合底引き網漁船が出漁した。

X とした箇所に対象語のうちどちらが存在したかを機械学習で推定する．

3.2 提案手法

本研究では，教師あり機械学習を利用して，対象語のうちのどの語が文中にあったのかを推定する．対象語のいずれかを含む文を学習データとして用いる．その文が含む対象語をその文の分類先として，学習を行う．教師あり機械学習には最大エントロピー法と BERT を利用する．

推定の結果の再現率が高い場合は，機械でも判別できるため，その上位語と下位語は使い分けが必要であり，一方で，再現率が低い場合は，機械では判別できないため，その上位語と下位語は使い分けが不要であると考えられる．よって，機械学習の性能は使い分けの必要性と正の相関があると考えられる．これに基づいて，機械学習により上位語と下位語の使い分けをより適切に行えたものとそうでないものにわけるために，機械学習の手法による上位語と下位語の使い分けの再現率の高さごとに高・中・低を設定する．上位語と下位語の対の語 A，語 B の再現率のうち，低い方の再現率で分類を行う．再現率の高さごとの分類は，高を再現率 9 割以上，中を再現率 8 割以上 9 割未満，低を再現率 8 割未満と設定する．分類に再現率を用いるのは， 再現率は機械学習が実験データのうちどれだけ正解を認識したかという指標であるためである．

3.3 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の説明を記述する．

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 [7, 8, 9, 10].

$$\sum_{a\in A,b\in B} p(a,b)g_j(a,b) = \sum_{a\in A,b\in B}(a,b)g_j(a,b) \quad for \quad \forall f_j(1 \leq j \leq k) \tag{3.1}$$

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

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 があってなおかつ分類が 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) を行なって，出力と文脈の確率分布を求めるものとなっている．

3.4 BERT

BERT とは，Bidirectional Encoder Representations from Transformers の略で,「Transformer による双方向のエンコード表現」と訳され，2018 年 10 月に Google の Jacob Devlin らの論文で発表された自然言語処理モデルである [11]．従来の自然言語処理では，大量のラベルのついたデータを用意させ，処理を行うことで課題に取り組む．しかし従来の手法に対し，BERT は事前学習でラベルのないデータをはじめに大量に処理を行う．その後，ファインチューニングで少量のラベルの付いたデータを使用することで課題に対応させる．

BERT 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．

BERT の素性分析の方法について述べる．BERT を用いてテストデータを 1 単語ずつに分ける．分けた 1 単語に対して BERT を用いることで，分けた 1 単語の分割，非分割に関するそれぞれ値が算出される．算出された値が分割の値が大きい場合は分割に関する素性，非分割の値が大きい場合は非分割に関する素性であると判断する．

3.5 素性

文献 [3][4] を参考にし，機械学習の素性には表 3.1 のものを用いる．これらの素性を，対象語が含まれる文から取り出す．表 3.1 中に記述されている分類語彙表の番号とは，分類語彙表によって与えられた語ごとの意味を表す 10 桁の番号である．上位語と下

位語の使い分けでは，文中に存在する語から使い分けに関する情報が得られると考え，素性 1 を設定する．その中でも対象語の前後の語に重要な情報があると考え，素性 2, 3 を設定する．また，対象語の存在する文構造にも情報があると考え，対象語の存在する文節の付属語，対象語の存在する文節が係る文節，対象語の存在する文節に係る文節の自立語と付属語をそれらの語彙情報とともに素性として設定する (素性 4-57).

表 3.1: 上位語下位語の判別に用いる素性

番号	素性の説明
素性 1	文中の名詞
素性 2	対象語の前後 3 語
素性 3	2 の分類語彙表の番号 7,5,4,3,2,1 桁
素性 4	対象語が含まれる文節の自立語
素性 5	4 の品詞
素性 6	4 の分類語彙表の番号 7,5,4,3,2,1 桁
素性 7	対象語が含まれる文節の最初の自立語
素性 8	7 の品詞
素性 9	7 の分類語彙表の番号 7,5,4,3,2,1 桁
素性 10	対象語が含まれる文節の最後の自立語
素性 11	10 の品詞
素性 12	10 の分類語彙表の番号 7,5,4,3,2,1 桁
素性 13	対象語が含まれる文節の付属語
素性 14	13 の品詞
素性 15	13 の分類語彙表の番号 7,5,4,3,2,1 桁
素性 16	対象語が含まれる文節の最初の付属語
素性 17	16 の品詞
素性 18	16 の分類語彙表の番号 7,5,4,3,2,1 桁
素性 19	対象語が含まれる文節の最後の付属語
素性 20	19 の品詞
素性 21	19 の分類語彙表の番号 7,5,4,3,2,1 桁
素性 22	対象語が含まれる文節に係る文節の自立語
素性 23	22 の品詞
素性 24	22 の分類語彙表の番号 7,5,4,3,2,1 桁
素性 25	対象語が含まれる文節に係る文節の付属語
素性 26	25 の品詞
素性 27	25 の分類語彙表の番号 7,5,4,3,2,1 桁
素性 28	対象語が含まれる文節に係る文節の最初の自立語
素性 29	28 の品詞
素性 30	28 の分類語彙表の番号 7,5,4,3,2,1 桁
素性 31	対象語が含まれる文節に係る文節の最後の自立語
素性 32	31 の品詞
素性 33	31 の分類語彙表の番号 7,5,4,3,2,1 桁
素性 34	対象語が含まれる文節に係る文節の最初の付属語
素性 35	34 の品詞
素性 36	34 の分類語彙表の番号 7,5,4,3,2,1 桁
素性 37	対象語が含まれる文節に係る文節の最後の付属語
素性 38	37 の品詞
素性 39	37 の分類語彙表の番号 7,5,4,3,2,1 桁
素性 40	対象語が含まれる文節が係る文節の自立語
素性 41	40 の品詞
素性 42	40 の分類語彙表の番号 7,5,4,3,2,1 桁
素性 43	対象語が含まれる文節が係る文節の付属語
素性 44	43 の品詞
素性 45	43 の分類語彙表の番号 7,5,4,3,2,1 桁
素性 46	対象語が含まれる文節が係る文節の最初の自立語
素性 47	46 の品詞
素性 48	46 の分類語彙表の番号 7,5,4,3,2,1 桁
素性 49	対象語が含まれる文節が係る文節の最後の自立語
素性 50	49 の品詞
素性 51	49 の分類語彙表の番号 7,5,4,3,2,1 桁
素性 52	対象語が含まれる文節が係る文節の最初の付属語
素性 53	52 の品詞
素性 54	52 の分類語彙表の番号 7,5,4,3,2,1 桁
素性 55	対象語の類義語対が含まれる文節が係る文節の最後の付属語
素性 56	55 の品詞
素性 57	55 の分類語彙表の番号 7,5,4,3,2,1 桁

第4章　実験

本章では，本研究で実験を行った対義語の組を 4.1 節で説明し，本研究が行った実験方法を 4.2 節で説明する．実験結果を 4.3 節に示し，被験者実験の結果を 4.4 節に示す．

4.1　実験データ

本研究では，自らが選出した上位語と下位語の対を利用する．表 4.1 に実験を行った上位語と下位語の対を示す．1991 年から 1995 年，2011 年から 2015 年の毎日新聞から上位語と下位語のいずれかの語を含む文を全てランダムに獲得し，データ数を同数にした実験を行った．表 4.2 に上位語と下位語の対においてデータ数を示す．被験者実験は，機械学習により求めた上位語と下位語のうち，正解率の低い方の性能をもとに上位語と下位語の対を 3 種類（高 9 割以上，中 8 割以上 9 割未満，低 8 割未満）に分類し，最大エントロピー法と BERT それぞれの高，中，低から 3 組ずつランダムに抽出し，合計 18 対（$3 \times 3 \times 2$）で実験を行う．表 4.3 に最大エントロピー法の結果をもとにして被験者実験を行った上位語と下位語の対を示し，表 4.4 に BERT の結果をもとにして被験者実験を行った上位語と下位語の対を示す．

表 4.1: 実験を行った上位語と下位語の対

1	「大学」 「学校」
2	「スポーツ」 「野球」
3	「動物」 「猫」
4	「人間」 「子供」
5	「果物」 「リンゴ」
6	「穀物」 「小麦」
7	「野菜」 「キャベツ」
8	「魚」 「サンマ」
9	「コンピューター」 「パソコン」
10	「文房具」 「鉛筆」
11	「都道府県」 「鳥取」
12	「職業」 「医者」
13	「酒」 「ワイン」
14	「病気」 「風邪」
15	「国」 「日本」
16	「乗り物」 「バス」
17	「海」 「太平洋」
18	「エネルギー」 「風力」
19	「道」 「歩道」
20	「植物」 「ひまわり」
21	「アクセサリー」 「ネックレス」
22	「食器」 「皿」
23	「季節」 「秋」
24	「家具」 「ソファ」
25	「鳥」 「鶏」
26	「楽器」 「ピアノ」
27	「洋服」 「コート」
28	「乳製品」 「バター」
29	「虫」 「セミ」
30	「行事」 「運動会」
31	「新聞」 「夕刊」

表 4.2: 実験に用いたデータ数

単語組	1語のデータ数	全データ数
「大学」 「学校」	1000	2000
「スポーツ」 「野球」	1000	2000
「動物」 「猫」	708	1416
「人間」 「子供」	1000	2000
「果物」 「リンゴ」	520	1040
「穀物」 「小麦」	324	648
「野菜」 「キャベツ」	258	516
「魚」 「サンマ」	132	264
「コンピューター」 「パソコン」	1000	2000
「文房具」 「鉛筆」	118	236
「都道府県」 「鳥取」	1000	2000
「職業」 「医者」	820	1640
「酒」 「ワイン」	808	1616
「病気」 「風邪」	571	1142
「国」 「日本」	1000	2000
「乗り物」 「バス」	130	260
「海」 「太平洋」	1000	2000
「エネルギー」 「風力」	127	254
「道」 「歩道」	550	1100
「植物」 「ひまわり」	120	240
「アクセサリー」 「ネックレス」	143	286
「食器」 「皿」	258	516
「季節」 「秋」	1000	2000
「家具」 「ソファ」	168	336
「鳥」 「鶏」	209	418
「楽器」 「ピアノ」	687	1374
「洋服」 「コート」	327	654
「乳製品」 「バター」	158	316
「虫」 「セミ」	153	306
「行事」 「運動会」	219	438
「新聞」 「夕刊」	327	654

表 4.3: 最大エントロピーの結果をもとに被験者実験を行った上位語と下位語の対

1	「都道府県」 「鳥取」
2	「季節」 「秋」
3	「行事」 「運動会」
4	「人間」 「子供」
5	「職業」 「医者」
6	「国」 「日本」
7	「魚」 「サンマ」
8	「穀物」 「小麦」
9	「酒」 「ワイン」

表 4.4: BERT の結果をもとに被験者実験を行った上位語と下位語の対

1	「都道府県」 「鳥取」
2	「道」 「歩道」
3	「海」 「太平洋」
4	「野菜」 「キャベツ」
5	「エネルギー」 「風力」
6	「植物」 「ひまわり」
7	「食器」 「皿」
8	「虫」 「セミ」
9	「鳥」 「鶏」

4.2 実験方法

獲得した上位語と下位語の対31対について，上位語と下位語の対ごとに上位語と下位語の使い分けの実験を行う．入力文は，1991年から1995年，2011年から2015年の毎日新聞から獲得した，上位語と下位語の対のいずれかの語を含む文である．評価は10分割のクロスバリデーションで行う．再現率ごとに，高中低の3種類に分類する．高が9割以上，中が8割以上9割未満，低が8割未満である．再現率が高のものは特に使い分けが必要なもの，低のものは比較的使い分けが必要でないもの，中のものはそれらの中間であるとする．

被験者実験についての説明をする．まず，上位語と下位語の対の使い分けの実験を行う．機械学習で使用した文から上位語と下位語の対ごとに10文ランダムに抽出し，被験者に上位語と下位語の使い分けを行わせる．その際，被験者には，まず上位語と下位語のどちらがよりふさわしいのかを選ばせ，そのうえで「どちらの語もあてはまりうる」と感じたものに対しては追加で「両方」の選択肢を選ばせる．

実際の例文を表4.5に示す．左の【】内がどちらがよりふさわしいのかを選ぶ際に用いる上位語と下位語の対で，右の【両方】が，どちらの語もあてはまりうると思った際に用いる選択肢である．

表 4.5: 被験者実験の例

「都道府県-鳥取」
1. 大会本部事務局には全国４７【都道府県・鳥取】【両方】高野連から１３８校（昨年より１校減）の推薦書類が届けられている。 2. 銀メダルを獲得した森下選手の【都道府県・鳥取】【両方】県八頭郡船岡町大江の実家では、父の大工、金治さん（５４）と母シゲ子さん（５６）が「一人息子のレースは静かに見たい」と夫婦二人きりでテレビ観戦。 3.【都道府県・鳥取】【両方】県・大山のふもと溝口町の白水川で三十一日、体長五十七センチ、重さ三キロもあるヤマメを同町のプロパン屋さんの高橋一行さん（６４）が捕まえた。 4. 限定付きながら、【都道府県・鳥取】【両方】レベルで公式大会への参加を認めたのは、滋賀県に次いで二例目。 5.【都道府県・鳥取】【両方】知事の指定を受けた保険医療機関の入院患者の給食費は一人一日千八百九十円と決められ、うち七百五十円前後が材料費。 6. 一方、【都道府県・鳥取】【両方】市の県立中央病院に入院していた琢磨ちゃんは七日夕退院する。 7. 通産省は同日夕、嶋野清社長を呼んで厳重注意、近く全国の関係【都道府県・鳥取】【両方】と関係業界に再発防止策の徹底を指示する。 8. １９９０年、麻薬及び向精神薬取締法が改正され、輸出入や製造は厚生大臣、小売業者は【都道府県・鳥取】【両方】知事の許可が必要。 9. 第二電電も、十二月四日に【都道府県・鳥取】【両方】市、同十一日に松江市のＰＯＩをそれぞれ開設し、日本テレコム同様に道北エリアを残すだけとなる。 10. 参院選挙区の議員一人当たり有権者数で見た「一票の格差」は、神奈川と【都道府県・鳥取】【両方】の間で最大六・五九倍となり、前回選挙時（一九八九年）の六・二五倍、昨年九月現在の有権者数による六・五三倍からさらに拡大した。

4.3 使い分けの実験結果

本章では機械学習を用いた上位語と下位語の使い分けの実験結果を示す．4.3.1 節では最大エントロピー法を用いた際の実験結果について，4.3.2 節では BERT を用いた際の実験結果について示す．

4.3.1 最大エントロピー法を用いた実験結果

再現率の高さごとに 31 対の上位語と下位語の対を分類した割合を表 4.6，再現率の高さごとの正解率の平均を表 4.7 に示す．また，再現率の高さごとに分類した上位語と下位語の対を表 4.8 に示す．それぞれの上位語と下位語ごとの結果を表 4.9 と表 4.10 に示す.

表 4.6: 最大エントロピー法を用いた場合の上位語と下位語の対の再現率の高さごとの割合

再現率の高さ	割合
高	0.10 (3/31)
中	0.42 (13/31)
低	0.48 (15/31)

表 4.7: 最大エントロピー法を用いた場合の再現率の高さごとの平均

再現率の高さ	再現率: 高	再現率: 中	再現率: 低	すべての対
平均	0.96	0.86	0.75	0.81

表 4.8: 最大エントロピーを用いた再現率の高さごとに分類した上位語と下位語の対

再現率の高さ	再現率	上位語と下位語の対
再現率高	9 割以上	「行事」「運動会」
		「都道府県」「鳥取」
		「季節」「秋」
再現率中	8 割以上 9 割未満	「海」「太平洋」
		「職業」「医者」
		「道」「歩道」
		「植物」「ひまわり」
		「国」「日本」
		「新聞」「夕刊」
		「エネルギー」「風力」
		「スポーツ」「野球」
		「動物」「猫」
		「楽器」「ピアノ」
		「人間」「子供」
		「果物」「リンゴ」
		「病気」「風邪」
再現率低	8 割未満	「酒」「ワイン」
		「乗り物」「バス」
		「学校」「大学」
		「家具」「ソファ」
		「食器」「皿」
		「洋服」「コート」
		「野菜」「キャベツ」
		「コンピューター」「パソコン」
		「穀物」「小麦」
		「虫」「セミ」
		「乳製品」「バター」
		「文房具」「鉛筆」
		「アクセサリー」「ネックレス」
		「鳥」「鶏」
		「魚」「サンマ」

表 4.9: 最大エントロピー法を用いた上位語と下位語の対の結果 1

上位語と下位語	再現率	データ数
学校	0.78	1000
大学	0.78	1000
スポーツ	0.86	1000
野球	0.81	1000
動物	0.81	708
猫	0.83	708
人間	0.84	1000
子供	0.80	1000
果物	0.80	520
リンゴ	0.84	520
穀物	0.71	324
小麦	0.71	324
野菜	0.76	258
キャベツ	0.74	258
魚	0.73	132
サンマ	0.64	132
コンピューター	0.78	1000
パソコン	0.74	1000
文房具	0.79	118
鉛筆	0.69	118
都道府県	0.97	1000
鳥取	0.95	1000
職業	0.87	820
医者	0.91	820
酒	0.79	808
ワイン	0.81	808
病気	0.81	571
風邪	0.80	571
国	0.86	1000
日本	0.91	1000
乗り物	0.78	130
バス	0.78	130

表 4.10: 最大エントロピー法を用いた上位語と下位語の対の結果 2

上位語と下位語	再現率	データ数
海	0.96	1000
太平洋	0.89	1000
エネルギー	0.90	127
風力	0.83	127
道	0.87	550
歩道	0.88	550
植物	0.93	120
ひまわり	0.87	120
アクセサリー	0.69	143
ネックレス	0.76	143
食器	0.78	258
皿	0.75	258
季節	0.93	1000
秋	0.91	1000
家具	0.83	168
ソファ	0.76	168
鳥	0.69	209
鶏	0.71	209
楽器	0.84	687
ピアノ	0.81	687
洋服	0.84	327
コート	0.75	327
乳製品	0.70	158
バター	0.78	158
虫	0.73	153
セミ	0.71	153
行事	1.00	219
運動会	0.98	219
新聞	0.90	327
夕刊	0.86	327

4.3.2 BERT を用いた実験結果

再現率の高さごとに 31 対の上位語と下位語の対を分類した割合を表 4.11，再現率の高さごとの正解率の平均を表 4.12 に示す．また，再現率の高さごとに分類した上位語と下位語の対を表 4.11 に示す．それぞれの上位語と下位語ごとの結果を表 4.14 と表 4.15 に示す．

表 4.11: BERT を用いた場合の上位語と下位語の対の再現率の高さごとの割合

再現率の高さ	割合
高	0.23 (7/31)
中	0.55 (17/31)
低	0.23 (7/31)

表 4.12: BERT を用いた場合の再現率の高さごとの平均

再現率の高さ	再現率: 高	再現率: 中	再現率: 低	すべての対
平均	0.94	0.86	0.76	0.85

表 4.13: BERT を用いた再現率の高さごとに分類した上位語と下位語の対

再現率の高さ	再現率	上位語と下位語の対
再現率高	9 割以上	「都道府県」「鳥取」
		「海」「太平洋」
		「職業」「医者」
		「国」「日本」
		「季節」「秋」
		「新聞」「夕刊」
		「道」「歩道」
再現率中	8 割以上 9 割未満	「植物」「ひまわり」
		「スポーツ」「野球」
		「人間」「子供」
		「動物」「猫」
		「エネルギー」「風力」
		「乗り物」「バス」
		「酒」「ワイン」
		「洋服」「コート」
		「病気」「風邪」
		「家具」「ソファ」
		「学校」「大学」
		「果物」「リンゴ」
		「コンピューター」「パソコン」
		「楽器」「ピアノ」
		「魚」「サンマ」
		「行事」「運動会」
		「野菜」「キャベツ」
再現率低	8 割未満	「食器」「皿」
		「文房具」「鉛筆」
		「鳥」「鶏」
		「アクセサリー」「ネックレス」
		「穀物」「小麦」
		「乳製品」「バター」
		「虫」「セミ」

表 4.14: BERT を用いた上位語と下位語の対の結果 1

上位語と下位語	再現率	データ数
学校	0.83	1000
大学	0.84	1000
スポーツ	0.88	1000
野球	0.88	1000
動物	0.87	708
猫	0.89	708
人間	0.88	1000
子供	0.90	1000
果物	0.82	520
リンゴ	0.85	520
穀物	0.73	324
小麦	0.77	324
野菜	0.81	258
キャベツ	0.80	258
魚	0.84	132
サンマ	0.81	132
コンピューター	0.82	1000
パソコン	0.82	1000
文房具	0.76	118
鉛筆	0.75	118
都道府県	0.98	1000
鳥取	0.97	1000
職業	0.92	820
医者	0.94	820
酒	0.85	808
ワイン	0.86	808
病気	0.87	571
風邪	0.84	571
国	0.92	1000
日本	0.95	1000
乗り物	0.86	130
バス	0.87	130

表 4.15: BERT を用いた上位語と下位語の対の結果 2

上位語と下位語	再現率	データ数
海	0.94	1000
太平洋	0.94	1000
エネルギー	0.87	127
風力	0.94	127
道	0.90	550
歩道	0.94	550
植物	0.91	120
ひまわり	0.89	120
アクセサリー	0.76	143
ネックレス	0.74	143
食器	0.78	258
皿	0.78	258
季節	0.94	1000
秋	0.92	1000
家具	0.85	168
ソファ	0.84	168
鳥	0.81	209
鶏	0.75	209
楽器	0.88	687
ピアノ	0.82	687
洋服	0.85	327
コート	0.85	327
乳製品	0.80	158
バター	0.73	158
虫	0.71	153
セミ	0.74	153
行事	0.81	219
運動会	0.90	219
新聞	0.92	327
夕刊	0.92	327

4.4 被験者実験

3 人の被験者による被験者実験の結果をそれぞれ表 4.16 から表 4.21 に示す．表中の「両方」は，どちらの語もあてはまりうると感じたものに対して追加で「両方」の選択肢が選ばれた割合である．また，3 名の再現率の高さごとの正解率の平均を表 4.22 と表 4.23 に示す.

表 4.16: 最大エントロピー法の被験者実験の結果 (被験者 A)

	再現率：高	再現率：中	再現率：低
上位語下位語対	都道府県・鳥取 季節・秋 行事・運動会	人間・子供 職業・医者 国・日本	魚・サンマ 穀物・小麦 酒・ワイン
正解率	0.87	0.97	0.87
両方	0.17	0.00	0.10

表 4.17: BERT の人手評価の結果 (被験者 A)

	再現率：高	再現率：中	再現率：低
上位語下位語対	都道府県・鳥取 道・歩道 海・太平洋	野菜・キャベツ エネルギー・風力 植物・ひまわり	食器・皿 虫・セミ 鳥・鶏
正解率	0.93	0.87	0.87
両方	0.00	0.00	0.00

表 4.18: 最大エントロピー法の人手評価の結果 (被験者 B)

	再現率：高	再現率：中	再現率：低
上位語下位語対	都道府県・鳥取 季節・秋 行事・運動会	人間・子供 職業・医者 国・日本	魚・サンマ 穀物・小麦 酒・ワイン
正解率	1.00	0.93	0.80
両方	0.07	0.00	0.10

表 4.19: BERT の人手評価の結果 (被験者 B)

	再現率：高	再現率：中	再現率：低
上位語下位語対	都道府県・鳥取 道・歩道 海・太平洋	野菜・キャベツ エネルギー・風力 植物・ひまわり	食器・皿 虫・セミ 鳥・鶏
正解率	0.87	0.97	0.87
両方	0.17	0.10	0.13

表 4.20: 最大エントロピー法の人手評価の結果 (被験者 C)

	再現率：高	再現率：中	再現率：低
上位語下位語対	都道府県・鳥取 季節・秋 行事・運動会	人間・子供 職業・医者 国・日本	魚・サンマ 穀物・小麦 酒・ワイン
正解率	0.93	0.97	0.83
両方	0.13	0.07	0.20

表 4.21: BERT の人手評価の結果 (被験者 C)

	再現率：高	再現率：中	再現率：低
上位語下位語対	都道府県・鳥取 道・歩道 海・太平洋	野菜・キャベツ エネルギー・風力 植物・ひまわり	食器・皿 虫・セミ 鳥・鶏
正解率	0.77	0.97	0.80
両方	0.23	0.17	0.30

表 4.22: 最大エントロピー法の被験者実験の平均

	再現率：高	再現率：中	再現率：低
上位語下位語対	都道府県・鳥取 季節・秋 行事・運動会	人間・子供 職業・医者 国・日本	魚・サンマ 穀物・小麦 酒・ワイン
正解率	0.93	0.96	0.83
両方	0.12	0.02	0.13

表 4.23: BERT の被験者実験の平均

	再現率：高	再現率：中	再現率：低
上位語下位語対	都道府県・鳥取 道・歩道 海・太平洋	野菜・キャベツ エネルギー・風力 植物・ひまわり	食器・皿 虫・セミ 鳥・鶏
正解率	0.86	0.94	0.85
両方	0.13	0.09	0.14

表 4.22 と表 4.23 を見ると，正解率の平均は高・中・低がそれぞれ，最大エントロピー法では（0.93，0.96，0.83）となり，BERT では（0.86，0.94，0.85）となっていることが分かる．人手評価では機械学習の結果とは異なり，最大エントロピー法，BERT ともに「再現率中」の上位語と下位語の対の正解率が高くなった．

また，被験者実験に対して，測定の一致度を示す指標であるカッパ値を求めたところ，最大エントロピー法の被験者実験のカッパ値は 0.75(かなりの一致)，BERT の被験者実験のカッパ値は 0.69(かなりの一致) となった．このことから，最大エントロピー法の被験者実験と BERT の被験者実験のカッパ値はどちらも 0.60 を超えており，被験者間の一致度が十分高いことが分かる．

第5章　考察

本章では，考察を記述する．5.1 節では，今回実験を行った上位語と下位語の対の中から，再現率の高さごとにいくつかの対を挙げ，具体的にどのような使い分けに対する情報が得られたかを考察する．5.2 節では，被験者実験の結果について考察する．5.3 節では，実験結果全体から得られた傾向について考察する．

5.1　上位語と下位語の対ごとの考察

最大エントロピー法と BERT のそれぞれで，分類を行った再現率の高さごとに上位語と下位語の対を 2 組ずつ例として挙げ，その上位語と下位語の対の使い分けに関する考察を行う．それぞれの例には，機械学習が正しく判定した正解例と機械学習が誤って判定した誤り例を上位語と下位語の対ごとに 2 例ずつの計 4 例と，機械学習が判定を行う際に参考にした素性とその素性の正規化 α 値を示す．正規化 α 値とは，最大エントロピー法で求まる α 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．各素性の，分類先ごとに与えられた正規化 α 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．例えば，ある素性 S のある分類先 A に対する正規化 α 値が X とすると，その素性 S のみで分類を行った場合，分類先 A と推定する確率が X となることを意味する．

5.1.1　最大エントロピー法の再現率高の例「都道府県」と「鳥取」

（正解例 1）全国調査は、都道府県、市、東京二十三区の計七百二十二自治体を対象に昨年七月から実施。

（正解例 2）「政治改革を実現する若手議員の会」代表世話人の石破茂氏＝鳥取全県区＝も、八頭郡用瀬町で国会報告会を開き、「宮沢首相は政治改革をすると約束した。

（誤り例 1）都道府県では和歌山、愛媛両県がこの双方に入っている。

（誤り例 2）一方、同県内一のナシ産地・東伯郡東郷町の鳥取東郷農協によると、ひどい所ではナシの一〇%、平均で五%が落ちた。

表 5.1: 最大エントロピー法の結果 (再現率高の例：「都道府県」と「鳥取」)

	再現率	適合率
都道府県	0.97	0.95
鳥取	0.95	0.97

表 5.2: 最大エントロピー法で参考にした素性 (再現率高の例：「都道府県」と「鳥取」)

都道府県		鳥取	
素性	正規化 α 値	素性	正規化 α 値
全国	0.72	氏	0.74
市町村	0.67	岡山	0.71
調査	0.67	島根	0.66
都市	0.63	県警	0.65
自治体	0.63	さん	0.64

再現率高の例として,「都道府県」と「鳥取」という対がある．

「都道府県」には「都道府県」よりも大きな地域のくくりである「全国」という単語がよく見受けられた．これは，多くの大会や団体の名称に「全国」という単語が含まれており，それらの大会や団体の話をするうえで,「各都道府県」や「47 都道府県」などの形で「都道府県」という単語が用いられやすいからである．また,「市町村」や「自治体」と言った単語は，調査などの対象として都道府県と並列表記で用いられる場面も多く見受けられた．

「鳥取」には他の都道府県名である「岡山」や「島根」などの単語がよく見受けられた．これは，鳥取との比較などの形で都道府県名が並列表記をされることが多いからである．また，都道府県名の中でも，中国地方の県名が上位に見受けられるのは，鳥取と同じ地方にある都道府県であり，比較の対象になりやすいからである．他にも,「氏」や「さん」という単語は，主に人物名の後につく敬称として用いられており，人物の説明として出身地や所属団体名の中に含まれる「鳥取」とともによく見受けられた．

5.1.2　最大エントロピー法の再現率高の例「季節」と「秋」

（正解例 1）「古池や　蛙飛び込む水の音」などの約七百句をウクライナ語に訳し、季節ごとに分類して紹介。

（正解例 2）西日本は二十六日、晴天に恵まれ、行楽地は秋の風情を求める家族連れらでにぎわった。

（誤り例 1）不安定な天気を繰り返しながら、季節は春へ。

（誤り例 2）天皇賞から始まった秋のＧ１シリーズの大レースを毎週わくわくしながら見守っている。

表 5.3: 最大エントロピー法の結果 (再現率高の例：「季節」と「秋」)

	再現率	適合率
季節	0.93	0.91
秋	0.91	0.93

表 5.4: 最大エントロピー法で参考にした素性 (再現率高の例：「季節」と「秋」)

季節		秋	
素性	正規化 α 値	素性	正規化 α 値
今	0.67	日本	0.67
ごと	0.67	日	0.63
秋	0.67	回復	0.63
感	0.65	読書	0.62
野菜	0.64	改革	0.62

再現率高の例として,「季節」と「秋」という対がある.

「季節」には「ごと」や「感」といった単語がよく見受けられた．これらの単語は,「季節ごと」や「季節感」といった形で用いられやすいため,「季節」とともによく見受けられている．また,「野菜」という単語も「季節ごと」「季節感」などの表現とともによく見受けられた.

「秋」には「日本」「回復」「改革」などの単語がよく見受けられた．これは,「今年秋」や「昨年秋」「来年秋」といった表現が政治や金融などのニュースでよく用いられているからである.

5.1.3　最大エントロピー法の再現率中の例「新聞」と「夕刊」

（正解例 1）二十三日未明までのテレビや新聞報道によると、フジモリ大統領の与党連合「新多数・カンビオ（変革）９０」（代表・ヨシヤマ氏）が過半数を辛うじて上回り、勝利した。

（正解例 2）土曜夕刊に掲載中の「トーク健康」面に加え、火、木、土曜の朝刊生活家庭面は健康のページ「生活　すこやか　家族」として新登場。

（誤り例 1）この流れのなかで、製紙工場の買い取り価格は昨年初頭から下がり始め、近畿地区では現在、新聞で一キロあたり十一円、雑誌で同七円という。

（誤り例 2）これが一新される毎日新聞夕刊の特徴です。

表 5.5: 最大エントロピー法の結果 (再現率中の例：「新聞」と「夕刊」)

	再現率	適合率
新聞	0.90	0.87
夕刊	0.86	0.89

表 5.6: 最大エントロピー法で参考にした素性 (再現率中の例：「新聞」と「夕刊」)

新聞		夕刊	
素性	正規化 α 値	素性	正規化 α 値
テレビ	0.68	朝刊	0.87
雑誌	0.66	毎日新聞	0.76
昨年	0.66	新聞	0.72
よう	0.66	日	0.71
毎日	0.65	原稿	0.69

再現率中の例として,「新聞」と「夕刊」という対がある.

「新聞」には「テレビ」や「雑誌」といったほかのメディア媒体の単語が見受けられた．これは,「テレビ」や「雑誌」が「新聞」との比較対象や,「新聞」との並列表記として用いられやすいからである.

「夕刊」には「朝刊」や「毎日新聞」といった単語が見受けられる．これらの単語は，引用の形で出てくることも多く,「朝刊」や「毎日新聞」からの引用がある場合に

は「夕刊」からの引用もある場合が多く，ともに出現しやすい．また,「朝刊」の場合,わざわざ「夕刊」という表現を用いる文章では「朝刊」と「夕刊」を区別する必要があることが多いため,「朝刊」という単語が出現しやすいのではないかと考えられた.

5.1.4 最大エントロピー法の再現率中の例「病気」と「風邪」

（正解例 1）「原因も治療方法も不明の病気に一矢を報いたい」と、作品の販売で一億円を集め、ＡＬＳ研究の団体や個人に助成金を出す「生命の彩（いのちのいろ）基金」を設立するためだ。

（正解例 2）自民党渡辺派会長の渡辺美智雄副総理・外相＝似顔絵＝が「風邪と過労」で入院後、同派議員にとっては渡辺氏の病状に気をもむ日々が続いている。

（誤り例 1）藤田さんは「塩素殺菌をやめると、病気が増えるだろう。

（誤り例 2）不破哲三委員長が質問する予定だったが、風邪のため急きょ変更した。

表 5.7: 最大エントロピー法の結果 (再現率中の例：「病気」と「風邪」)

	再現率	適合率
病気	0.81	0.80
風邪	0.80	0.81

表 5.8: 最大エントロピー法で参考にした素性 (再現率中の例：「病気」と「風邪」)

病気		風邪	
素性	正規化 α 値	素性	正規化 α 値
治療	0.76	病状	0.78
患者	0.69	体調	0.74
理由	0.68	ウイルス	0.73
障害	0.65	病院	0.70
私	0.63	インフルエンザ	0.67

再現率中の例として,「病気」と「風邪」という対がある.

「病気」からは想定されがちな「治療」や「患者」といった単語だけでなく,「理由」や「障害」と言った単語も見受けられた.「理由」は政治家や大統領など著名な人物に

関する出馬の辞退や，来訪のキャンセルなどの説明で用いられる場面が多かった.「障害」は「胃腸障害」や「摂食障害」といった病名に含まれているため,「病気」とともに見受けられやすかった.

「風邪」は「体調不良」などの形で用いられやすい「体調」と言った単語が見受けられやすかった．他には,「インフルエンザ」などの他の病名も見受けられやすく，これに伴って「ウイルス」も見受けられた.

5.1.5 最大エントロピー法の再現率低の例「学校」と「大学」

（正解例 1）一審の東京地裁判決も「生徒に改善の見込みがなく、学校外に排除することも教育上やむを得なかったとは、到底言えない」と述べていた。

（正解例 2）日本の大学には約二百人の盲学生がいる。

（誤り例 1）「配慮不足の指摘、申し訳ない」－－校長が会見事件直後に赴任した神戸高塚高校の衣川清馬校長は学校で記者会見。

（誤り例 2）主な著書に「大学の自治の歴史」「文学でつづる教育史」など。

表 5.9: 最大エントロピー法の結果 (再現率低の例：「学校」と「大学」)

	再現率	適合率
学校	0.78	0.78
大学	0.78	0.78

表 5.10: 最大エントロピー法で参考にした素性 (再現率低の例：「学校」と「大学」)

学校		大学	
素性	正規化 α 値	素性	正規化 α 値
生徒	0.85	学生	0.88
子供	0.83	教授	0.84
教師	0.77	企業	0.80
児童	0.74	入試	0.73
授業	0.74	採用	0.71

再現率低の例として,「学校」と「大学」という対がある.

「学校」には「生徒」や「教師」といった単語が見られた．それに対して「大学」には「学生」や「教授」といった単語が見られた．このことから,「学校」と「大学」では教わる立場や教える立場にある人の呼び方が異なることが分かる．他にも,「学校」は小学校や中学校などを指すことも多く,「児童」や「子供」と言ったより年齢の低い子供を指す単語も見受けられた.「大学」は卒業後就職をする人が多いため「企業」や「採用」などの就職活動にかかわるような単語も見られた.

5.1.6 最大エントロピー法の再現率低の例「コンピューター」と「パソコン」

（正解例 1）学校や病院、研究所などがコンピューターなどのハイテク機器の導入を拡大できるよう、公共投資の資金を振り向け、ハイテク不況に対応した公共投資の実施を求めるほか、情報ネットワークの基盤整備構想も検討している。

（正解例 2）送信側もパソコンと電話回線、ソフトウエアでトラック十台程度の管理なら三百万円程度という。

（誤り例 1）また会場とＴＡＭＡらいふ２１のイベント会場をＩＮＳネット６４で結び、コンピューターゲームの遠隔地対戦を実現する。

（誤り例 2）その後、松下は八四年に米ＩＢＭのパソコンを委託生産、八六年には米モトローラ社の元設計技師と組んでコンピューターメーカー、ソルボーン社を米国に設立するなど、コンピューター技術を蓄積してきた。

表 5.11: 最大エントロピー法の結果 (再現率低の例 :「コンピューター」と「パソコン」)

	再現率	適合率
コンピューター	0.78	0.78
パソコン	0.74	0.77

表 5.12: 最大エントロピー法で参考にした素性 (再現率低の例：「コンピューター」と「パソコン」)

コンピューター		パソコン	
素性	正規化 α 値	素性	正規化 α 値
パソコン	0.80	ソフト	0.83
研究所	0.73	コンピューター	0.77
上	0.73	ワープロ	0.77
編集	0.72	点字	0.75
技術	0.71	教室	0.70

再現率低の例として,「コンピューター」と「パソコン」という対がある.

「コンピューター」には「研究所」という単語がよく見られる．これは，様々な「研究所」で「コンピューター」を用いた調査などが行われているからだ．また,「上」は「画面上」や「仕事上」などの形でよく見受けられた.

「パソコン」には「ソフト」や「教室」という単語がよく見受けられた．これは，主に「パソコンソフト」や「パソコン教室」の形で見られた．また,「ワープロ」は「パソコン」と機能が似ているので並列表記され，見受けられることが多かった.

5.1.7 BERT の再現率高の例「国」と「日本」

（正解例 1）一自治体が国を訴えたのだからニュースである。

（正解例 2）暫定取り決めは、向こう三カ月間の日米両国の武力進出を凍結し日本は南仏印から撤収するその代わりに石油を含む交易を限定的に再開する――というもので、その間に本交渉を煮詰めていく手はずになっていた。

（誤り例 1）自由貿易圏を通じて投資と貿易の拡大、雇用増を期待するメキシコ政府は、来年中の協定締結を目指しているが、国内には新たな国の進路への不安も出ている。

（誤り例 2）今や「大国」の日本は、世界中からどれだけ役立つことをしてくれるか、勝手な期待の目で見られている。

表 5.13: BERT の結果 (再現率高の例：「国」と「日本」)

	再現率
国	0.92
日本	0.95

表 5.14: BERT で参考にした素性 (再現率高の例：「国」と「日本」)

国		日本	
素性	正規化 α 値	素性	正規化 α 値
地方自治体	1.00	両国	1.00
自治体	1.00	他国	1.00
民有	1.00	総領事館	1.00
自治労	1.00	世界銀行	1.00
敵する	1.00	経団連会館	1.00

再現率高の例として,「国」と「日本」という対がある.

「国」では「国」と比較する形で良く用いられる「地方自治体」や「自治体」などの素性が得られた. また,「国」と「地方自治体」や「自治体」が対比として用いられる場面では,「国」がそのまま「日本国」を指している場合多かった.「日本」は「両国」や「他国」「総領事館」など日本以外の国を話題に挙げる際に用いられやすい単語の近くによく現れた.

5.1.8 BERT の再現率高の例「道」と「歩道」

(正解例 1) このままでは自民党が多数を前提に予算委員会での審議を打ち切り、強行採決しか道はないか、と緊張が高まってきたヤマ場の日の朝、どの政党の国会対策委員会の部屋でも、話題はその日の朝刊各紙の社説である。

(正解例 2) しかし、横断歩道がないため通産省に通じる地下道か、虎ノ門交差点まで足を延ばさなければならず、官僚泣かせの難所だった。

(誤り例 1) 静かな歓迎車列が進む道の街路灯すべてでマレーシア国旗と日の丸が揺れる。

(誤り例 2) この栗の木歩道は、昭和五十九年度にスタート。

表 5.15: BERT の結果 (再現率高の例：「道」と「歩道」)

	再現率
道	0.90
歩道	0.94

表 5.16: BERT で参考にした素性 (再現率高の例：「道」と「歩道」)

道		歩道	
素性	正規化 α 値	素性	正規化 α 値
訪韓	1.00	地下道	1.00
改憲	1.00	道端	1.00
動力炉・核燃料開発事業	1.00	横道	1.00
復学	1.00	道幅	1.00
富国強兵	1.00	通りかかっ	1.00

再現率高の例として,「道」と「歩道」という対がある.

「道」は慣用的な表現や比喩として用いられることも多く,「訪韓」や「改憲」「動力炉・核燃料開発事業」など，政治にかかわる単語の近くに現れやすい.「歩道」は「地下道」や「道幅」「通りかかっ」などの近くに現れており，慣用的表現にはほとんど用いられていない.

5.1.9 BERT の再現率中の例「動物」と「猫」

（正解例 1）実際に撮影された動物園内の映像を立体的に処理し、見たい動物を選択できて子供たちの興味を引き、精神的な安定や学習意欲の高まりがみられたという。

（正解例 2）次いで「迷い猫だったから」（２５%)、「野良猫だったから」（２０%）と両方を合わせると半数近くが結局拾われてきたというもの。

（誤り例 1）犬だけでなく小鳥も、熱帯魚もいて動物は身近な存在だったが、猫っ可愛がりはしないタイプだった。

（誤り例 2）純粋種のなかで最も人気が高いのはシャム猫。

表 5.17: BERT の結果 (再現率中の例：「動物」と「猫」)

	再現率
動物	0.87
猫	0.89

表 5.18: BERT で参考にした素性 (再現率中の例：「動物」と「猫」)

動物		猫	
素性	正規化 α 値	素性	正規化 α 値
園内	1.00	拾わ	1.00
園児	1.00	チンチラ	1.00
園	1.00	可愛がり	1.00
愛護	1.00	泊まっ	1.00
園子	1.00	言い出し	1.00

再現率高の例として,「動物」と「猫」という対がある.

「動物」は「園内」「園」などと合わせて「動物園」の形で使われることが多い.「猫」は具体的な種類である「チンチラ」や，野良猫に対する記事内で「拾う」などの単語とともに現れやすい.

5.1.10 BERT の再現率中の例「乗り物」と「バス」

（正解例 1） ＪＲ総研によると、これまで乗り心地についての研究はあったが、乗り物酔いの本格研究は初めて。

（正解例 2） 近畿運輸局の調べでは、午後一時現在までにスト継続中の会社を含め、電車、バス計千四十本が運休、約四万人に影響した。

（誤り例 1） 皇太子さまと小和田雅子さんの「結婚の儀」の後のパレードで、お二人の乗り物に宮内庁所有のオープンカーを使うことが十八日、決まった。

（誤り例 2） 一方、ある意味で運転手確保よりも深刻なのが都心でのバス集合場所の確保問題。

表 5.19: BERT の結果 (再現率中の例：「乗り物」と「バス」)

	再現率
乗り物	0.86
バス	0.87

表 5.20: BERT で参考にした素性 (再現率中の例：「乗り物」と「バス」)

乗り物		バス	
素性	正規化 α 値	素性	正規化 α 値
酔い	1.00	近畿運輸局	0.99
気力	0.99	エアターミナル	0.99
盲導犬	0.99	行き	0.99
悔い	0.99	野田線	0.99
品物	0.99	京浜急行電鉄	0.99

再現率高の例として,「乗り物」と「バス」という対がある.

「乗り物」は「乗り物酔い」などの形で「酔い」とともに現れやすく，他にも交通機関と盲導犬の関係などの記事で「盲導犬」などの近くにも現れた.「バス」は料金改定や調査を実施する記事などで「近畿運輸局」の近くに現れやすいほか，細かな行先や路線名都ともに記載される関係から「行き」や「野田線」のそばにも現れた.

5.1.11 BERT の再現率低の例「穀物」と「小麦」

（正解例 1）同副委員長は会見で前年と比ベソ連の国民総生産（ＧＮＰ）が一五%、石油生産が一〇%、穀物生産量が二五%低下していると言明。

（正解例 2）米価は二年連続の据え置きで六十キロ当たり一万八千百二十三円、麦価は二年ぶりの引き下げで国内産小麦が同二千六百六十二円。

（誤り例 1）さらにエリツィン大統領は近く外国から穀物二千万トンを輸入、共和国民にパンを安定供給することを約束した。

（誤り例 2）日本の小麦は、品種改良が世界で最も進み、これ以上の早生化は無理とされていた。

表 5.21: BERT の結果 (再現率低の例：「穀物」と「小麦」)

	再現率
穀物	0.73
小麦	0.77

表 5.22: BERT で参考にした素性 (再現率低の例：「穀物」と「小麦」)

穀物		小麦	
素性	正規化 α 値	素性	正規化 α 値
国民総生産	0.00	フェニトロチオン	0.97
減産	0.99	米価	0.97
肥育	0.99	米食	0.96
世界銀行	0.99	混ぜる	0.96
財政難	0.99	蒸し	0.95

再現率高の例として,「穀物」と「小麦」という対がある.

「穀物」は国民総生産に穀物生産の項目があるため「国民総生産」の近くに現れることが多く，また「穀物肥育」の形で「肥育」の近くにも現れた.「小麦」は麦価と米価が比較されることが多い関係で「米価」の近くに現れやすく，小麦食と米食が比較されやすい関係から「米食」の近くにも現れやすかった.

5.1.12 BERT の再現率低の例「乳製品」と「バター」

（正解例 1）まずは、乳製品、デンプンなど輸入自由化していない農産物や、高率関税をかけている農産物加工品を含めた農業分野全体で二国間交渉を進め、コメへの影響を最小限にとどめたいとの思惑がある。

（正解例 2）農水省は二十日、バター、脱脂粉乳などの原料となる加工原料乳の政府保証価格（乳価、農家手取り価格）を二年連続で引き下げる方針を固めた。

（誤り例 1）また、コメや乳製品などの欄を空欄にしても関税化に賛成の国からの反発が出てくるものとみられる。

（誤り例 2）肉やバターは慢性的な品不足。

表 5.23: BERT の結果 (再現率低の例：「乳製品」と「バター」)

	再現率
乳製品	0.80
バター	0.73

表 5.24: BERT で参考にした素性 (再現率低の例：「乳製品」と「バター」)

乳製品		バター	
素性	正規化 α 値	素性	正規化 α 値
として	0.94	脱脂粉乳	0.99
関税	0.93	粉ミルク	0.99
品目	0.92	蒸す	0.99
欧州共同体	0.92	パン生地	0.99
特に	0.92	混ぜる	0.99

再現率高の例として,「乳製品」と「バター」という対がある.

「乳製品」は輸出入の関係から「関税」「品目」等の近くに現れやすい.「バター」はレシピ記事などで「粉ミルク」「混ぜる」と言った材料や調理工程で頻出の動詞などの近くに現れやすい.

5.2 被験者実験の考察

機械学習の結果と人手評価の正解率の平均をまとめたものを表 5.25 に示す.

表 5.25: 機械学習と人手評価の正解率の平均

	再現率：高	再現率：中	再現率：低	すべての対
最大エントロピー法	0.96	0.86	0.75	0.81
被験者実験 (最大エントロピー法)	0.93	0.96	0.83	0.91
BERT	0.94	0.86	0.76	0.85
被験者実験 (BERT)	0.86	0.94	0.85	0.88

表 5.25 を見ると，正解率の平均は，高では機械学習の方が良く，中や低では被験者実験の方が良い結果となっていることが分かる．このことから，今回の実験の結果には正の相関がないことが分かる．これは，人手と機械では使い分けを得意とする上位語と下位語の対や参考にする素性が異なるためである．以下で被験者実験よりも機械学習の方が性能が良かった再現率高の特徴について示す．そのために，再現率高の上位語と下位語の対である,「都道府県」「鳥取」と「季節」「秋」の最大エントロピー法での素性分析の結果と,「新聞」「夕刊」と「海」「太平洋」の BERT での素性分析の結果を表 5.26 から表 5.29 に示し，それぞれ考察する．

表 5.26: 「都道府県」と「鳥取」が最大エントロピー法で参考にした素性

都道府県		鳥取	
素性	正規化 α 値	素性	正規化 α 値
全国	0.72	氏	0.74
市町村	0.67	岡山	0.71
調査	0.67	島根	0.66
都市	0.63	県警	0.65
自治体	0.63	さん	0.64

「都道府県」の素性には,「全国」という単語がよく見受けられた．これは，多くの大会や団体の名称に「全国」という単語が含まれており，それらの大会や団体の話をするうえで,「各都道府県」や「47 都道府県」などの形で「都道府県」という単語が用いられやすいからである.「鳥取」の素性には,「岡山」や「島根」などの単語がよく見

受けられた．これは，鳥取との比較などの形で都道府県名が並列表記をされることが多いからである．これらのことから「都道府県」と「鳥取」においては，全国規模の大会名などの付近に「都道府県」が出現しやすいことや,「岡山」や「島根」などの他の都道府県名の付近に「鳥取」が出現しやすいことが人手では分かりにくいと考えられる．

表 5.27: 「季節」と「秋」が最大エントロピー法で参考にした素性

季節		秋	
素性	正規化 α 値	素性	正規化 α 値
今	0.67	日本	0.67
ごと	0.67	日	0.63
秋	0.67	回復	0.63
感	0.65	読書	0.62
野菜	0.64	改革	0.62

「季節」の素性には，は「ごと」や「感」といった単語がよく見受けられた．これらの単語は,「季節ごと」や「季節感」といった形で用いられやすいため,「季節」とともによく見受けられている.「秋」の素性には,「日本」「回復」「改革」などの単語がよく見受けられた．これは,「今年秋」や「昨年秋」「来年秋」といった表現が政治や金融などのニュースでよく用いられているからである．これらのことから「季節」と「秋」においては,「ごと」や「感」と共に「季節」が用いられやすいことや,「秋」が「日本」や「改革」などの政治や金融に関する単語の付近に出現しやすいことが，人手よりも機械の方が分かりやすいと考えられる．

表 5.28: 「新聞」と「夕刊」が BERT で参考にした素性

新聞		夕刊	
素性	正規化 α 値	素性	正規化 α 値
学習研究社	1.00	日経新聞	1.00
科学技術庁	1.00	産経新聞	1.00
創価学会	1.00	日本経済新聞	1.00
中小企業退職金共済事業団	1.00	読売新聞	1.00
UNICEF	1.00	北海道新聞	1.00

「新聞」の素性には,「科学技術庁」や「UNICEF」などの団体や会社の名前がよく見受けられた．これは，それらの団体が新聞で取り上げられることが多いからである．

「夕刊」の素性には,「日経新聞」や「読売新聞」などの新聞の名前がよく見受けられた．これは,「夕刊」について話す際には，どこの新聞社が出した夕刊なのかなどの詳細も記載される場合が多いからである．これらのことから「新聞」と「夕刊」においては，団体や会社名の付近で「新聞」が出現しやすいことや,「夕刊」が具体的な新聞名の付近に出現しやすいことが人手では分かりにくいと考えられる．

表 5.29: 「海」と「太平洋」が BERT で参考にした素性

海		太平洋	
素性	正規化 α 値	素性	正規化 α 値
押し流さ	1.00	三井海上火災	1.00
車止め	1.00	南シナ海	1.00
流れ着き	1.00	中央アジア	1.00
押し倒し	1.00	ベーリング海	1.00
締め出さ	1.00	日本海	1.00

「海」の素性には,「押し流さ」「流れ着き」などの単語がよく見受けられた．これは,「海」が津波に関する話題で使われやすく,「押し流さ」「流れ着き」などの単語は津波の話題で用いられやすい単語だからである.「太平洋」には,「南シナ海」や「日本海」などの他の海の名前がよく見受けられた．これは，海流などの説明で海の名前が用いられたり，並列表記をされることが多いからである．このことから「海」と「太平洋」においては，津波に関する記事内で用いられやすい「押し流さ」や「流れ着き」の付近に「海」が出現しやすいことや,「南シナ海」や「日本海」などの他の海の名前の付近に「日本海」が出現しやすいことが，人手よりも機械の方が分かりやすいと考えられる．

5.3 実験結果全体の傾向と考察

上位語と下位語の対は，機械学習の使い分けの正解率は一番低いものが 0.64，被験者実験の使い分けの正解率の一番低いものが 0.77 であった．これに対し，先行研究である織金の類義語対の使い分けにおける，機械学習の再現率低の一番低い分類は 3 割未満，被験者実験の使い分けの正解率の最低値は 0.45 と低かった．このことから，上位語と下位語の対は，先行研究で行われていた類義語対に比べて，全体的に使い分けが必要な単語対であると言える．また，機械学習を用いた上位語と下位語のそれぞれ

の使い分けにおいて参考にした素性は，下位語の素性には他の下位語が出現しやすいという知見が得られた．例としては,「バター」の素性である「脱脂粉乳」,「キャベツ」の素性である「レタス」などがある．他にも，定型表現を用いた場合は正解率が上がりやすいこともわかった．例えば,「季節ごと」「季節感」などの定型表現がこれに当てはまる．

第6章　おわりに

本研究では機械学習を用いて上位語と下位語の対の使い分けを行った．

第 1 の成果として，31 対の上位語と下位語の対を用いた実験において，機械学習を用いた提案手法は，最大エントロピー法を用いた場合は 0.81，BERT を用いた場合は 0.85 の正解率であった．これにより，今回提案した手法が上位語と下位語の使い分けに対して有用だと考えられる．

第 2 の成果として，機械学習での性能に基づき上位語と下位語の対を使い分けが必要なものとそれほど必要でないものに分類した．今回の実験で再現率高に分類したものは特に使い分けが必要であると考えられる．使い分けが必要とされた上位語と下位語の対には「都道府県」と「鳥取」や「季節」と「秋」などの対があり，使い分けが必要でない上位語と下位語の対に「食器」と「皿」や「虫」と「セミ」などの対があった．また，いくつかの上位語と下位語の対について実際に素性を分析し，使い分けに役立つ情報を明らかにすることができた．

謝辞

本研究を進めるに当たり，終始に渡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に感謝の意を表します．

参考文献

[1] 織金和希. 機械学習を用いた動詞・形容詞の類義語の使い分け. 鳥取大学工学部卒業論文, 2017.

[2] 日笠考祐. 機械学習を用いた 3,4 組の単語における使い分けと知見獲得. 鳥取大学工学部卒業論文, 2020.

[3] 佐々本暖久. 機械学習を用いた対義語の置き換え可否判定. 鳥取大学工学部電気情報系学科, 2019.

[4] 小西択磨. BERT を用いた対義語の置き換え可否判定. 鳥取大学工学部卒業論文, 2021.

[5] Juman version7.0: http://nlp.ist.i.kyoto-u.ac.jp/index.php?cmd=readpage=juman.

[6] 萩原亜彩美, 森山菜々美, 浅原正幸, 加藤祥, , 山崎誠. 『分類語彙表』に対する反対語情報. 言語処理学会第 25 回年次大会, 2018.

[7] Eric Sven Ristad. Maximum entropy modeling for natural language. In *ACL/EACL Tutorial Program, Madrid*, 1997.

[8] 村田真樹, 内山将夫, 内元清貴, 馬青, 井佐原均. 種々の機械学習手法を用いた多義解消実験. 電子情報通信学会言語理解とコミュニケーション研究会, pp. 7–14, 2001.

[9] Masao Utiyama. Maximum entropy modeling packagen: http://www.nict.go.jp/x/x161/members/mutiyama/software.htmlmaxent. 2006.

[10] 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. Vol. 2, pp. 272–279, 2010.

[11] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*, 2018.