2020年度（令和2年度）　卒業論文

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

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

B17T2082A 日笠　孝祐

概要

本研究は 3,4 組の単語に対して，教師あり機械学習を用いることにより，これらの単語の使い分けや知見獲得を行う．

ここでの 3,4 組の単語とは、「春・夏・秋・冬」「東・西・南・北」「上・中・下」といったような 3,4 組で纏められ，かつ類義語や対義語ではない単語のことである．このような単語組から各単語においてよく使用されている単語や文法上重要な単語を調査して得たものを知見とし，最終的にこれらの単語組が使い分けが必要かどうかを判断する．強田ら [1] は EDR 電子化辞書から得られた名詞の類義語を利用し，機械学習を用いた名詞の類義語の使い分けの研究を行い，中瀬 [2] は強田らと同様の手法で副詞の類義語の使い分けの研究を行った．また，赤江 [3] は使い方の分かる類語例解辞典 [8] 及び「擬音語・擬態語」使い分け帳 [9] から得られる類義語を利用して強田らや中瀬とは異なった類義語で使い分けの研究を行い，織金 [4] は強田らと同様の手法で動詞と形容詞の類義語の使い分けの研究を行った．ある 3,4 組の単語間での機械学習の性能が高く，より正確に使い分けを行えていた場合は，その単語組は特に使い分けの必要な単語であるとわかる．また，機械学習が使用した素性を分析して，各単語の使い分けに役立つ情報の考察を行う．このような実験と調査を自身で考えた単語を対象に行う．

本研究の成果は 2 つある．1 つは単語組の使い分けのために機械学習を使用し，3,4 組の単語 13 組 47 単語について実験を行った結果，データ数を同数に揃えた実験では提案手法が 0.68，同数におけるベースラインが 0.28 であったため，この提案手法自体が 3,4 組の使い分けに対して有用であると言えたことである．

もう 1 つは，いくつかの単語組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ことである．例えば「上・中・下」の場合,「上」では「事実」(事実上) や「インターネット」(インターネット上) といった単語が出現したのに対し,「中」では「活動」(活動中),「下」では「水面」(水面下),「体制」(体制下) のような単語が周辺に出現するということである．

この 2 つの成果は，日本語初学者に対する知見獲得や書き間違えの修正に役立つと考えられる．

目 次

表 目 次

図 目 次

第1章　はじめに

3,4 組の単語とは，それぞれが何らかに関連する単語ではあるが、類義語や対義語とは判断できない単語のことである．例としては「春・夏・秋・冬」(関連する単語：季節，四季) などが挙げられる．先行研究に類義語の使い分けの研究がある．類義語に関する研究では，西尾 [6] の人間の会話における類義語の使用傾向を調査し分析する研究などがある．また，小島ら [7] は「しょう油」と「醤油」のような同じ語の表織香ね記が異なる場合の使い分けを機械学習で行った．また，強田ら [1] は EDR 電子化辞書から得られる類義語から機械学習による名詞の類義語の使い分けを，中瀬 [2] は副詞の類義語の使い分けの研究を行った．そして，赤江 [3] は強田，中瀬らが扱っておらず，かつ言語学で議論となっている類義語の使い分けの研究を行い，織金 [4] は動詞と形容詞の類義語の使い分けを行った．本研究では，機械学習の性能や素性が類義語や対義語だけでなく，その他の表現の使い分けにも役立つと考え，機械学習を用いて 3,4 組の単語組の使い分けと知見獲得を行う．本研究の成果は，日本語初学者に対する知見獲得や書き間違えの修正に役立つと考えられる．

本研究では，私が考えたそれぞれが何らかに関連する単語ではあるが，類義語や対義語ではない 3,4 組に纏められる単語組を列挙したものを利用する．

機械学習によって単語を推定しやすい場合は，他の単語でも使い分けの必要な語とわかり， 逆に機械学習で推定しづらい場合は単語の使い分けが明瞭でないということがわかる．機械学習の素性を分析することで，使い分けに役立つ知見を得ることを目的とする．

本研究の主な主張点を以下に整理する．

- 単語組の使い分けのために機械学習を使用し，3,4 組の単語 13 組 47 単語について実験を行った結果，データ数を同数に揃えた実験では提案手法が 0.68，同数におけるベースラインが 0.28 であるため，この提案手法自体が 3,4 組の使い分けに対して有用であると言える．

- 実際に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することができ，使い分けに役立つ情報を明らかにした．例として,「上」の推定に役立つ素性には「インターネット」,「事実」などがあり,「下」の推定に役立つ素性には「水面」,「体制」などがあった．

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，本研究が行った使い分けの実験についての説明と，各単語における結果と考察について記述する．

第 5 章では，第 4 章の結果全体に対する考察を行う．

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，西尾 [6] が行った類義語に対するアンケート調査について記述する．2.2 節では，小島ら [7] が行った表記選択の研究について記述し，2.3 節では，強田ら [1] が行った類義語に対する機械学習を用いた名詞の使い分けについて記述する．2.4 節では，中瀬 [2] が行った類義語に対する機械学習を用いた副詞の使い分けについて記述する．2.5 節では，赤江 [3] が行った強田や中瀬が行っていない類義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．2.6 節では，織金 [4] が行った動詞や形容詞における類義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．

2.1　類義語間の選択についての調査

西尾は，同一の個人が状況や場面に応じて使い分ける類義語と，ある人はふつう一方の語を，他の人はふつうもう一方の語を使うというような類義語があるとし，今回は主に後者のような類義語についての選択を調査している [6]．調査方法は，調査対象者に意味の似た言葉の対を複数提示し，親しい人と話すときにどちらを使って話すかを回答させる．それを年齢・性別・地域で分類し，どのような選択の違いが見られたかを調べた．

調査した類義語対は，性質によって A から D に分類し，分類方法は表 2.1 の通りである．

表 2.1: 類義語対の分類

分類	性質	例
A	外来語を一方にもつ類義語対	デパートと百貨店
B	旧式語を一方にもつ類義語対	婚礼と結婚式
C	日常語と文章語の類義語対	双生児とふたご
D	その他	通信簿と通知表

調査結果を例として，選択の差が一番顕著に見られたのが年齢による区別で，選択の差があった類義語対としては「プレゼント」と「おくりもの」があった．この対は，若い世代へ移るほど「プレゼント」の割合が増加している傾向にあった．性別での差が見られた類義語対としては「後家」と「未亡人」という対があり，男性のほうが「後家」を用いる傾向にあり，女性は「未亡人」を使用する傾向にあった．また地域で差があった類義語対としては，それほど大きな差がみられた類義語対はなかったが，挙げるとすれば「車庫」と「ガレージ」という対で，大阪では「ガレージ」が用いられる傾向にあり，東京では「車庫」が用いられる傾向にあった．

この先行研究は，類義語の使い分けの調査という点では本研究と類似している部分がある．しかし先行研究は，人手によるアンケート調査であり，機械学習により類義語の使い分けを自動で推定する本研究とは違った角度からのアプローチである．

2.2 機械学習を用いた表記選択の難易度推定

小島らは，表記にゆれがある単語,「是非」と「ぜひ」などの単語について機械学習を用いて表記選択の難易度推定を行った [7]．機械学習によって高い正解率で表記選択を行えたものは人間による表記選択が容易で，機械学習によって充分な正解率を得られなかったものは人間による表記選択が困難であると考えている．この研究では，実験で用いるデータを 2005 年〜2007 年の毎日新聞の文章としている．JUMAN で形態素解析した結果得られる代表表記を用いて，表記のゆれが検出された単語 (15185 語) を対象とし，さらに条件を付与して得られた単語 (1877 語) の半分 (939 語) を実験対象としている．付与する条件は以下のものとした．

条件 1 対象の単語のすべての表記の合計出現頻度数が 100 以上であるもの

条件 2 対象の単語の曖昧性を避けるため，JUMAN の解析結果で@マークが一度もつかないもの

条件 3 対象の単語の各表記の出現頻度数上位 2 つが，どちらも 10 以上であるもの

なお条件 2 の JUMAN で@マークがつかないものとは，表記は違うが代表表記が同じものである．逆に@マークがつくものは，代表表記が別の語であることを示している．例えば,「けいじ」という語を JUMAN で解析すると代表表記が「啓示」のほかに,@マークがつき代表表記に「掲示」「刑事」「計時」が解析結果として出力される.「啓

示」「掲示」「刑事」「計時」はそれぞれ別の語である．JUMAN の解析では，読みは同じで代表表記が別の語がある場合は，先頭に@マークをつけて出力する．実験方法は単語ごとに機械学習を適用し，10 分割のクロスバリデーションを行う．なお，機械学習は表記のゆれがある単語の各表記の出現頻度数上位 2 つについて判定を行った．機械学習の再現率の高さごとに高・中・低を設定する．2 つの表記のうち，低いほうの再現率で分類を行い，再現率が 8 割以上のものを高，8 割未満 5 割以上を中，5 割未満を低とし，再現率高のものを適切な表記を選択できたものとした．

実験の結果，実験対象とした 939 語中 81 語が再現率高となった．また，再現率高となったものの例としては「手引」と「手引き」や,「うかる」と「受かる」など，中のものには「讃歌」と「賛歌」や,「冬物」と「冬もの」などがあり，低には「朝顔」と「あさがお」や,「倦怠」と「けん怠」などがあった．

この先行研究は，機械学習を適用した対象は違うが，手法などが本研究と類似している部分がある．

2.3 機械学習を用いた名詞の類義語の使い分け

強田らは，機械学習による分類性能の高い名詞の類義語の使い分けの研究を行った [1]．

類義語に関する研究では，類義語の使い分けに機械学習を用いた研究はない．強田らは名詞の類義語の使い分けのために機械学習を使用し，複数の名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．

強田らは EDR 電子化辞書と 1991 年の毎日新聞を使用し，名詞の類義語を獲得した．名詞の類義語を獲得する条件は以下の通りである．

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 その二つの語が両方とも，日本語単語辞書において，付与された概念識別子が 1 つであること

条件 3 その二つの語が両方とも，1991 年の毎日新聞で出現頻度が 50 回以上であること

条件 4 形態素解析システム JUMAN[10] 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

獲得した名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，1991 年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名詞の類義語対を，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) を分析することで類義語の使い分けに重要な情報を把握した．

強田らの研究の成果として，機械学習を用いた名詞の類義語の使い分けの手法自体が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名詞の類義語対とそれほど必要でない名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，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．使い分けが必要な名詞の類義語対として「貯金」と「貯蓄」,「メダル」と「賞碑」，使い分けが必要でない類義語対として「省エネ」と「省エネルギー」,「上期」と「上半期」があった．

2.4 機械学習を用いた副詞の類義語の使い分け

中瀬は，機械学習による分類性能の高い副詞の類義語の使い分けの研究を行った [2]．

中瀬は副詞の類義語の使い分けに機械学習を使用した．複数の副詞の類義語対を対象に，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．

中瀬は EDR 電子化辞書と 1991 年〜1995 年の毎日新聞を使用し，副詞の類義語を獲得した．副詞の類義語を獲得する条件は以下の通りである．

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 その二つの語が両方とも，日本語単語辞書において，付与された概念識別子が 1 つであること

条件 3 その二つの語が両方とも，1991 年〜95 年の毎日新聞で出現頻度が 50 回以上であること

条件 4 形態素解析システム JUMAN[10] 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

獲得した副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，1991 年〜95 年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副詞の類義語対を，高・中・低に分類し，機械学習における素性を分析することで類義語の使い分けに重要な情報を把握した．

中瀬の研究の成果として，機械学習を用いた副詞の類義語の使い分けの手法自体が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副詞の類義語対とそれほど必要でない副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，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．使い分けが必要な副詞の類義語対として「きわめて」と「だいぶ」,「そっくり」と「すっかり」，使い分けが必要でない類義語対として「さして」と「さほど」,「すっかり」と「ことごとく」があった．

2.5 強田や中瀬が行っていない単語での機械学習を用いた類義語の使い分け

赤江は，強田や中瀬が行っていない単語での機械学習を用いた類義語の使い分けの研究を行った．

赤江は使い方の分かる類語例解辞典 [8] および「擬音語・擬態語」使い分け帳 [9] から人手で選んだ 2 組から 5 組の類義語を利用した使い分けを行い，1991 年〜1995 年，2011 年〜2015 年の毎日新聞から類義語の組のいずれかの語を含む文を獲得した．

データ数は 1 語につき 100 文以上のものを実験を行い，11 組 29 単語を出現率に合わせた場合と同数に合わせた場合をベースライン手法と比較し，評価を 10 分割のクロスバリデーションで行った．

赤江の研究の成果として，機械学習を用いた類義語の使い分けは全ての単語組においてベースライン手法よりも提案手法の方がよりよい正解率が出ることを明らかにした．また，機械学習での性能に基づき使い分けが必要な類義語組とそれほど必要でない類義語組を明らかにし，各単語から有用な素性も得た．

使い分けが必要な類義語として「おかげ」と「せい」と「ため」,「はっきり」と「きっぱり」，使い分けが必要でない類義語対として「うろうろ」と「ぶらぶら」などがあった．

2.6 機械学習を用いた動詞と形容詞の類義語の使い分け

織金は，機械学習を用いて動詞と形容詞の類義語の使い分けの研究を行った．

織金は動詞と形容詞の類義語の使い分けに機械学習を使用した．複数の副詞の類義語対を対象に，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獲得にはEDR電子化辞書と1991年〜1995年の5年分の毎日新聞を，形容詞類義語対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，動詞と形容詞の類義語を獲得した．動詞，形容詞の類義語を獲得する条件は以下の通りである．

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 その二つの語が両方とも，1991年〜95年の毎日新聞で出現頻度が50回以上であること，また，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[10]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，動詞の類義語対獲得には1991年〜1995年の5年分の毎日新聞を，形容詞の類義語対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を分析することで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，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．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」,「近しい」と「むつまじい」，使い分けが必要でない動詞と形容詞の類義語対として「はみ出る」と「はみ出す」,「気まずい」と「面はゆい」があった．

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記述している．3.4 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

使い分けを行ったりや知見を得たりしたい 4 組の単語 A，B，C，D があるとする．各単語 A，単語 B，単語 C，単語 D 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，正しく単語を使い分けることができたと考えられる．具体的な例として，4 つの単語組「春」「夏」「秋」「冬」を例に以下に示す．

「春らしさ」が演出されたダイニングテーブルに心が浮き立つ。 １９８９年の夏 の甲子園で優勝。 １１月はカボチャなどをふんだんに使った「秋グラタン」（９４５円）など２種類。 夏は木の葉が陽光を遮り、冬 は落葉して陽光を通す。

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．

「Xらしさ」が演出されたダイニングテーブルに心が浮き立つ。 １９８９年のX の甲子園で優勝。 １１月はカボチャなどをふんだんに使った「Xグラタン」（９４５円）など２種類。 夏は木の葉が陽光を遮り、X は落葉して陽光を通す。

X とした箇所に対象語のうち，いずれの単語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，機械学習を用いて学習を行う．教師あり機械学習には最大エントロピー法を利用する．

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．再現率の高さごとの傾向の予測を図 3.1 に示す.

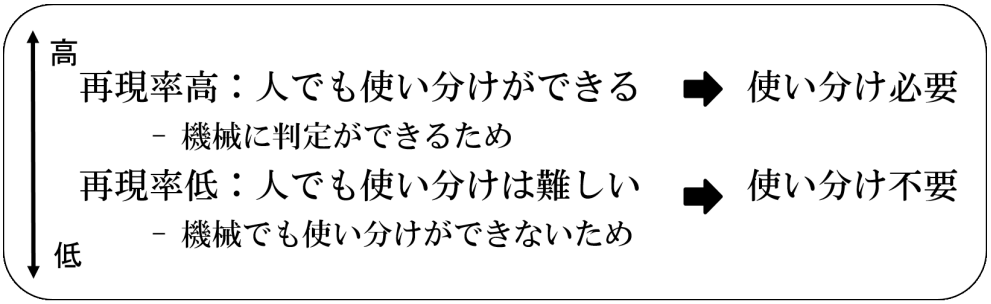

図 3.1: 再現率の高さごとの傾向

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 [11, 12, 13, 14].

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

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

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 素性

文献 [7][1] 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．3,4 組の単語の使い分けでは，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 4-45)．

表 3.1: 3,4 組の単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付属語
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付属語
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付属語
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に係る文節の自立語
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に係る文節の付属語
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に係る文節の最初の自立語
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最後の自立語
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最初の付属語
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後の付属語
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が係る文節の自立語
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が係る文節の付属語
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が係る文節の最初の自立語
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最後の自立語
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最初の付属語
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	対象語の 3,4 組の単語対が含まれる文節が係る文節の最後の付属語
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

第4章　使い分けの実験と考察

本章では，本研究で実験を行った 3,4 組の単語組を 4.1 節で説明し，本研究が行った実験方法を 4.2 節で説明し，実験結果を 4.3 節に示す.

4.1　実験データ

本研究では，自らが選出した単語組を利用する．表 4.1 に実験を行った単語組を示す．1991 年〜1995 年，2011 年〜2015 年の毎日新聞から単語組のいずれかの語を含む文を全てランダムに各単語 1000 文ずつ (「牛肉」「豚肉」「鶏肉」のみ 500 文ずつ) 獲得し, データ数を同数にした実験を行った．表 4.2 に 3,4 組の単語組においてデータ数を示す.

表 4.1: 実験を行った 3,4 組の単語組

1	「春」「夏」「秋」「冬」
2	「東」「西」「南」「北」
3	「香川」「徳島」「愛媛」「高知」
4	「小学」「中学」「高校」「大学」
5	「上」「下」「左」「右」
6	「上」「中」「下」
7	「牛肉」「豚肉」「鶏肉」
8	「大雨」「洪水」「大雪」
9	「風」「林」「火」「山」
10	「晴れ」「雨」「雪」
11	「朝」「昼」「夕」「夜」
12	「花」「鳥」「風」「月」
13	「赤」「青」「黄」

表 4.2: 3,4 組の単語において同数にした実験のデータ数

単語組	1 語のデータ数	全データ数
「春」「夏」「秋」「冬」	1000	4000
「東」「西」「南」「北」	1000	4000
「香川」「徳島」「愛媛」「高知」	1000	4000
「小学」「中学」「高校」「大学」	1000	4000
「上」「下」「左」「右」	1000	4000
「上」「中」「下」	1000	3000
「牛肉」「豚肉」「鶏肉」	500	1500
「大雨」「洪水」「大雪」	1000	3000
「風」「林」「火」「山」	1000	4000
「晴れ」「雨」「雪」	1000	3000
「朝」「昼」「夕」「夜」	1000	4000
「花」「鳥」「風」「月」	1000	4000
「赤」「青」「黄」	1000	3000

4.2 実験方法

獲得した 3,4 組の単語 13 組について， 単語組ごとに使い分けの実験を行う．入力文は，1991 年〜1995 年，2011 年〜2015 年の毎日新聞から獲得した，3,4 組の単語のうち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．また全ての単語を同数に揃えて実験を行っていることから，無作為に単語を挿入した場合に正解し得る確率，つまり 3 組の単語については 0.33(33%)，4 組の単語については 0.25(25%) をベースライン手法として提案手法との比較を行う．

4.3 実験結果

以下，実験結果について記述する．

機械学習の正解率の高さごとに単語組を分類した割合を表 4.3 に示す．データ数を同数に揃えた 3,4 組の単語，13 組 47 単語におけるベースライン手法と提案手法の結果を表 4.4 に示す．

表 4.4 のように正解率の平均は提案手法が 0.68，ベースライン手法が 0.28 であった．このことから提案手法の正解率はベースライン手法より正解率が高いと言えた．

表 4.3: データ数を同数に揃えた正解率の高さごとに分類した単語組

正解率	3,4 組の単語
8 割以上	「晴れ」「雨」「雪」
	「大雨」「洪水」「大雪」
7 割以上 8 未満	「風」「林」「火」「山」
	「花」「鳥」「風」「月」
	「赤」「青」「黄」
	「上」「中」「下」
	「上」「下」「左」「右」
6 割以上 7 未満	「小学」「中学」「高校」「大学」
	「牛肉」「豚肉」「鶏肉」
5 割以上 6 未満	「香川」「徳島」「愛媛」「高知」
	「東」「西」「南」「北」
	「朝」「昼」「夕」「夜」
	「春」「夏」「秋」「冬」

表 4.4: データ数を同数に揃えた時のベースライン手法と提案手法の正解率

3,4 組の単語	ベースライン手法	提案手法
「春」「夏」「秋」「冬」	0.25	0.51
「東」「西」「南」「北」	0.25	0.55
「香川」「徳島」「愛媛」「高知」	0.25	0.59
「小学」「中学」「高校」「大学」	0.25	0.70
「上」「下」「左」「右」	0.25	0.70
「上」「中」「下」	0.33	0.72
「牛肉」「豚肉」「鶏肉」	0.33	0.62
「大雨」「洪水」「大雪」	0.33	0.81
「風」「林」「火」「山」	0.25	0.76
「晴れ」「雨」「雪」	0.33	0.82
「朝」「昼」「夕」「夜」	0.25	0.54
「花」「鳥」「風」「月」	0.25	0.74
「赤」「青」「黄」	0.33	0.72
平均正解率	0.28	0.68

4.4　3,4 組の単語組ごとの考察

本節では，3,4 組の単語組ごとに使い分けに関する考察を行う．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単語組ごとにそれぞれ例を示す．下線が機械学習が判定した結果であり，括弧内が元の文の語である．

3,4 組の単語の使い分けにおいて，それぞれどのような素性が使い分けに役に立つのかを明らかにするために，素性の分析を行う．

以下の表において,「R1：」のような記号「Rn：」は単語組の直前 n 番目に素性として得られた単語があることを示し，同様に「F2：」のような記号「Fn：」は単語組の直後 n 番目に素性として得られた単語があることを示す．

また，再現率とは各単語において正しい文章であった確率を示し，適合率とは各単語を実際に機械学習によって挿入させた文のうち正解であった文章の確率を示す．

4.4.1　「春」「夏」「秋」「冬」

（正解例 1）新型車両Ｅ６系がデビューした秋田新幹線や、２０１５年春に金沢開業を予定する北陸新幹線に対抗する狙い。

（正解例 2）今年も節電の夏。

（正解例 3）秋に予定されている日本政府によるフランスからのプルトニウムの海上輸送問題が、国際的な波紋を巻き起こしている。

（正解例 4）町全域が警戒区域に指定され避難生活の終わりが見えない中、雪深い会津地方で冬を迎えるのを心配する人が多いことが一因だ。

（誤り例 1）東日本大震災直後に開催された１１年冬(春) も心に残る大会だ。

（誤り例 2）仙台を訪れれば、秋田竿燈（かんとう）まつり、盛岡さんさ踊り、福島わらじまつりなど東北の冬(夏) 祭りが楽しめる。

（誤り例 3）アプリには今年夏(秋) にビデオ通話機能を追加する。

（誤り例 4）小２から中２までドイツのデュッセルドルフにいて、夏(冬) の３カ月間、家に閉じこもりっきり。

表 4.5 に単語組「春」「夏」「秋」「冬」の機械学習の分類結果を示す．表 4.6 に単語組「春」「夏」「秋」「冬」の正規化 α 値に基づいた機械学習が参考にした素性を示す．

表 4.5: 機械学習の分類結果 (「春」「夏」「秋」「冬」)

	データ数	再現率	適合率
春	1000	0.53	0.52
夏	1000	0.44	0.47
秋	1000	0.48	0.48
冬	1000	0.59	0.56
総数	4000	0.51	0.51

表 4.6: 機械学習が参考にした素性 (正規化 α 値:「春」「夏」「秋」「冬」)

春		夏		秋		冬	
素性	正規化 α 値	素性	正規化 α 値	素性	正規化 α 値	素性	正規化 α 値
心	0.53	甲子園	0.48	11 月	0.53	R２：春	0.48
大学	0.50	最高	0.47	紅葉	0.48	練習	0.47
開業	0.49	福島	0.43	春	0.47	夏	0.45
桜	0.45	参院	0.42	問題	0.40	雪	0.40
北陸	0.40	大会	0.41	消費	0.38	ソ連	0.39

「春」からは想定されがちな「花」や「桜」のような植物に関する単語だけでなく,「開業」や「卒業」,「就職」のような「物事の区切り」や「心の入れ替わり」がなされるような単語も多く見受けられた．また，春は選抜高校野球選手権大会 (センバツ) が行われることから，対比関係になりやすい「秋」とほぼ同頻度で「夏」や「センバツ」という単語も見受けられた.

「夏」からは甲子園や大会，参院のような夏にある「大きなイベント」が頻出した一方で，東日本大震災後の新聞からは「福島」や「電力」,「節電」という単語も多く見受けられた．また,「最高の夏」や「最後の夏」という言い回しもよくされることがわかった.

「秋」からは「紅葉」や「収穫」など実りに関することが多く出現したのに対し,「問題」や「疑惑」という単語が多く出現する季節でもあった．この「問題」や「疑惑」が出現した文章を見ると,

- この問題に秋の臨時国会でピリオドを打つのが次の総裁の責務だ。
- 昨年秋に浮上した健康への懸念を払拭（ふっしょく）し、内外全ての問題に精通していることを改めてアピールした形だ。
- ４中全会は通常は秋に開かれるが、周氏の疑惑に対する関心が高まっていることから、８月末か９月上旬に開催を早め、疑惑について発表

することを検討しているという。

- 疑惑の原点、八九年秋。

のような文章から，疑惑や問題が浮上する季節だけでなく，これらの課題を解決させる季節であるということも知見として得ることができた.

また「消費」という単語については，秋に食べ物や物の消費がよくされるという旨の文章も見受けられたが，一番多かったのは「消費税増税」関連の記事であった.

「冬」からは「雪」や「スキー」のような雪に関することや「ソ連」や「北海道」など多く雪が降る地域が頻出したほか,「家族」や「旅行」のような「身内とだんらんを楽しむ」季節でもあることもわかった.

4.4.2 「東」「西」「南」「北」

（正解例 1）約十分後、現場から東約五キロで街路樹に二台の車がぶつかっているのを発見、関連を調べている。

（正解例 2）ＪＲ西は同日、１６年３月期連結決算の業績予想を上方修正し、売上高は従来予想より２００億円増の１兆４１１５億円、最終利益は９０億円増の９０５億円とした。

（正解例 3）南相馬市に転居して三十数年ぶりで福島市までバスに乗ったところ終点に着くと、運転手さんから「だいぶ前に乗っていましたよね」と言われ、びっくりしました。

（正解例 4）この後、大阪市北区の選挙事務所に戻って出陣式。

（誤り例 1）南 (東) の方を見上げれば、いつも浅間山の白い煙が澄み切った空に浮かんで見えました。

（誤り例 2）東(西) ヨーロッパ諸国も「長距離大気汚染防止条約」（七九年）を結び、汚染物質の排出を大幅に抑えている。

（誤り例 3）奈良・山の辺の道の西(南) の終点近くにある「つば市観音」。

（誤り例 4）最後の大正区のひったくり事件が起きた約二十分後、現場から東(北) 約三キロの駐車場で、犯行に使ったとみられる車が炎上した。

表 4.7 に単語組「東」「西」「南」「北」の機械学習の分類結果を示す．表 4.8 に単語組「東」「西」「南」「北」の正規化 α 値に基づいた機械学習が参考にした素性を示す．

表 4.7: 機械学習の分類結果 (「東」「西」「南」「北」)

	データ数	再現率	適合率
東	1000	0.51	0.52
西	1000	0.50	0.49
南	1000	0.57	0.57
北	1000	0.60	0.60
総数	4000	0.55	0.55

表 4.8: 機械学習が参考にした素性 (正規化 α 値：「東」「西」「南」「北」)

東		西		南		北	
素性	正規化 α 値	素性	正規化 α 値	素性	正規化 α 値	素性	正規化 α 値
西	0.67	東	0.71	相馬	0.63	F2:大阪	0.61
自_自立語単：福岡	0.51	地区	0.45	北	0.61	南	0.60
自_自立語単：東京	0.45	登板	0.44	スーダン	0.55	梅田	0.54
花巻	0.44	今治	0.41	アジア	0.43	北海道	0.44
写真	0.43	西武	0.41	平和	0.40	中東	0.42

「東」「西」「南」「北」全てに共通して「東」と「西」,「南」と「北」は相対的なものであることから双方に対する正規化 α 値が高くなっていることが分かる．

各々の頻出単語を見ると,「東」では福岡や東京，花巻などの地域を示すものに付くほか,「東 (あずま，ひがし)」のように名字としての使用が多く見受けられた．

また「東」同様,「西」でも地域や地域を示すものだけでなく名字で多く使用されていることが判明した．その他にも「西」にはプロ野球チームである「西武ライオンズ」の略称として扱われることもあり,「登板」や「ドーム」といった野球関連の単語も得ることができた．

「南」では対となる「北」よりも「相馬」という単語の正規化 α 値の方が高くなった．これはこの研究で用いた新聞において，東日本大震災で大きな被害の出た「南相馬市」に関するニュースが多く出現したからである．同様に「南」の頻出単語として独立を目指しスーダンとの衝突を繰り返す「南スーダン」や，グルジアからの独立を宣言している「南オセチア」の記事が多く見られたことに起因すると考えられる．また「平和」という単語は，南スーダンの「平和維持活動」(PKO) という使い方で頻出していた．

「北」では「大阪」や「梅田」のような大阪の地名が頻出していた．大阪や梅田が使用されている文章を確認したところ，梅田のある大阪北区や梅田の北側 (通称：ウメキタ) には多くの店だけでなく，会場やホール，ギャラリーがあり，イベントや公演の場となっていることがわかった．また，北は「北海道」や「北朝鮮」の略称として使用されることが多いということもわかった．

4.4.3 「香川」「徳島」「愛媛」「高知」

（正解例 1）四国地建は二十四日、瀬戸大橋中ほどの与島パーキングエリア（香川県坂出市）で、たそがれコンサートを開く。

（正解例 2）２００５年に４町の合併で発足した徳島県阿波市では思わぬ事態に。

（正解例 3）伊方原発（愛媛県伊方町）の建設・稼働を推進した。

（正解例 4）１３日まで４日連続４０度以上だった高知県四万十市は３８・６度だった。

（誤り例 1）鳥取の投票価値を１人１票とすると、高知(香川) は０・５９票となる。

（誤り例 2）高知(徳島) 県矢野では銅鐸を木の入れものに納めて埋めた穴の上に簡単な構造の小屋が建っていたことが分かりました。

（誤り例 3）徳島(愛媛) ２区、当選３回、４２歳。

（誤り例 4）研究グループはより多くのデータを集めるため香川(高知) 県歯科医師会にも協力を打診し、今後、元船員らに歯の提供を呼びかける。

表 4.9 に単語組「香川」「徳島」「愛媛」「高知」の機械学習の分類結果を示す．表 4.10 に単語組「香川」「徳島」「愛媛」「高知」の正規化 α 値に基づいた機械学習が参考にした素性を示す．

表 4.9: 機械学習の分類結果 (「香川」「徳島」「愛媛」「高知」)

	データ数	再現率	適合率
香川	1000	0.63	0.63
徳島	1000	0.56	0.56
愛媛	1000	0.61	0.62
高知	1000	0.56	0.56
総数	4000	0.59	0.59

表 4.10: 機械学習が参考にした素性 (正規化 α 値：「香川」「徳島」「愛媛」「高知」)

香川		徳島		愛媛		高知	
素性	正規化 α 値	素性	正規化 α 値	素性	正規化 α 値	素性	正規化 α 値
坂出	0.68	鳴門	0.81	今治	0.75	土佐	0.70
丸亀	0.62	三好	0.62	新居浜	0.69	明徳義塾	0.70
高松	0.61	阿波	0.59	伊方	0.59	安芸	0.49
ドルトムント	0.51	神戸	0.41	原発	0.50	地震	0.39
プレー	0.44	近畿	0.40	対策	0.47	キャンプ	0.39

四国四県を全てにおいてその県の有名な市や町の名前が有用な素性として得ることができた.

しかしながら,「香川」だけはプロサッカー選手である「香川真司」選手をはじめとした「香川」という名字の人も多く見受けられた．また，香川選手関連の情報として「プレー」や「ドルトムント」,「本田」(圭佑選手) などサッカーやチーム，他の選手に関連する素性も頻出していた.

「徳島」は徳島県に関する地名が一番多く出現していたが，隣県である兵庫県の「神戸」や「淡路島」だけでなく「近畿」といった単語も見受けられ，四国と本州を「鳴門」海峡で結ぶ県として重要な役割を担っていることがわかった.

「愛媛」における有用な素性は「原発」である.「新居浜」や「伊方」といった地名が多く見受けられ,「四国電力」や「対策」,「稼働」のような素性も出現していた.

「高知」ではプロ野球のキャンプ地になることが多いことから「キャンプ」やそのキャンプ地の場所，選手名が多く挙げられていた．また,「地震」という単語も素性としてあり，調べると高知県は四国の南端にあることから，東日本大震災を受け，南海トラフ地震への対策が行われていることがわかった.

4.4.4 「小学」「中学」「高校」「大学」

（正解例 1）夫の実家に行ったところ、小学２年の姪（めい）が遊びに来ていた。

（正解例 2）また、同市立桜宮中学は、生徒の訴え通り、生徒の頭髪をバリカンで刈ったことを認め、市教委は同中に対して「今後一切、バリカンで生徒の頭髪を切らないよう」指導した。

（正解例 3）私は高校を２年で中退し、１７歳で予備校通いをした。

（正解例 4）夢は大学の教授。

（誤り例 1）中学(小学) 時代、おてんば娘だったせいで、サッカーでは敵がどこから攻めてくるかわからないゴールキーパーばかりやらされました。

（誤り例 2）島根県吉賀（よしか）町教委は１４日、町立吉賀高校(中学) の給食に、調理で使用した包丁の破片が混入し、3年の男子生徒がのみ込んでいたと発表した。

（誤り例 3）古里・福岡や中学(高校) 時代を過ごした佐賀の関係者は「世界を舞台にもっともっと暴れてほしい」と声援を送った。

（誤り例 4）こうしたシステムは北海道の高校(大学) などで採用しているが、多機能で幅広く利用できるのは全国で初めてという。

表 4.11 に単語組「小学」「中学」「高校」「大学」の機械学習の分類結果を示す．表 4.12 に単語組「小学」「中学」「高校」「大学」の正規化 α 値に基づいた機械学習が参考にした素性を示す.

表 4.11: 機械学習の分類結果 (「小学」「中学」「高校」「大学」)

	データ数	再現率	適合率
小学	1000	0.87	0.84
中学	1000	0.61	0.60
高校	1000	0.60	0.64
大学	1000	0.71	0.69
総数	4000	0.70	0.70

表 4.12: 機械学習が参考にした素性 (正規化 α 値：「小学」「中学」「高校」「大学」)

小学		中学		高校		大学	
素性	正規化 α 値	素性	正規化 α 値	素性	正規化 α 値	素性	正規化 α 値
年生	0.59	小学	0.78	大学	0.65	教授	0.65
R1:5	0.54	高校	0.73	県立	0.55	短大	0.51
ちゃん	0.51	夜間	0.55	全国	0.42	研究	0.51
児童	0.47	いじめ	0.42	中退	0.41	系	0.45
教室	0.39	興味	0.38	合格	0.40	学部	0.38

「小学」の素性として，小学生特有の「5 年生」や「6 年生」の存在や，学生ではなく「児童」,「○○君」や「○○ちゃん」というような，幼い子どもに対する表現が多くあった.

「中学」の素性として,「高校」とまとめて表記されることが多いことがわかった. また,「興味」という単語から中学時代は様々なものに関心を抱く年頃であるのにと対し,「夜間」や「いじめ」という単語も同時に見受けられ，詳しく確認すると夜間学校は義務教育未履修者や日本に来たばかりの海外の学生だけでなく,「いじめ」や「登校拒否」によって日中学校に通えない人のためのものであるということを知ることができた.

「高校」の素性として,「大学」,「受験」,「合格」といった受験に関連するものが挙げられるほか,「甲子園」をはじめとした「大会」や「レース」,「賞」といった全国規模の大きなイベントや大会が開かれる傾向にある年代であるということが分かった.

「大学」の素性として，今まで中学や高校になかった「教授」や「研究」の存在や「国立」や「短大」のような大学の種類,「(○○) 系」や「学部」のような「専攻」に関わる単語が多く見受けられたのが特徴であると言える.

4.4.5 「上」「下」「左」「右」

（正解例 1）（定数是正問題という）憲法上の制約は重大だが、致命的支障ではない。

（正解例 2）水面下のキーワードは、やはり「解散」だ。

（正解例 3）先頭の袴田が左ひじに死球を受けて負傷した。

（正解例 4）石本は一回に右アッパーでダウンを奪われ、その後も的確なパンチをもらった。

（誤り例 1）調べでは、中村さんは自室ベランダ側八畳間のベッドの下(上) で仰向きになって死亡。

（誤り例 2）顕微授精は顕微鏡の上(下) で卵細胞を操作する新しい不妊治療技術。

（誤り例 3）琴欧洲との一番は、脇腹への負担を減らすため徹底して右(左) から攻めた。

（誤り例 4）演技中の息子は楽しそうに見えたが、けがをした左(右) 脚が気になった。

表 4.13 に単語組「上」「下」「左」「右」の機械学習の分類結果を示す．表 4.14 に単語組「上」「下」「左」「右」の正規化 α 値に基づいた機械学習が参考にした素性を示す．

表 4.13: 機械学習の分類結果 (「上」「下」「左」「右」)

	データ数	再現率	適合率
上	1000	0.80	0.82
下	1000	0.77	0.77
左	1000	0.62	0.62
右	1000	0.62	0.60
総数	4000	0.70	0.70

表 4.14: 機械学習が参考にした素性 (正規化 α 値：「上」「下」「左」「右」)

上		下		左		右	
素性	正規化 α 値	素性	正規化 α 値	素性	正規化 α 値	素性	正規化 α 値
下	0.45	上	0.55	右	0.92	左	0.90
自_自立語最初単品：名詞	0.43	自_自立語最初単品：名詞	0.43	自_自立語最後単品：名詞	0.43	自_自立語最後単品：名詞	0.46
事実	0.39	政権	0.42	同右	0.42	同左	0.45
方	0.39	自_付属語最初単：、	0.42	三浦	0.39	守備	0.40
会社	35	水面	0.36	打率	0.38	ドリブル	0.40

東西南北と同様，上下左右でも「上」と「下」,「左」と「右」は相対的なものであることから双方に対する正規化 α 値が高くなっていることが分かる．

各々の頻出単語を見ると,「上」では「上方」など単純に物体の位置の情報だけではなく「事実上」や「憲法上」,「画面上」など「〜において」や「〜に関して」という言い回しや「希望を聞いた上で…」や「前置きした上で…」のように「何かをしたあとで」を意味する言い回しが頻出したことから,「上」は慣用的な表現を得ることができた．

また,「下」も「上」と同様,「水面下」や「憲法下」,「体制下」など「表に現れない様」という言い回しや「編集長の下，…」や「監督の下，…」のように「規則や支配

の及ぶところ」を意味する言い回しが頻出したことから文章の生成や使い方に役立つ素性を得ることができた.

「左」と「右」は守備や得点を入れたスポーツ選手の名前や，利き手足，攻撃した箇所，打球の行方，身体の負傷箇所などに多く用いられることがわかった．例として,左では「打率」,「安打」,「骨折」，右では「守備」,「ドリブル」,「投手」などが挙げられた.

「上」と「下」,「左」と「右」の大きな違いとして,「左」と「右」では「事実上」や「水面下」のように名詞の直後に付与させることで一つの慣用句として用いるという素性が見受けられなかった．以上のことから,「上下左右」のように方向としては一括りにできるものも,「上」,「下」と「左」,「右」では文中での使われ方に大きな差があることを知見として得ることができた.

4.4.6 「上」「中」「下」

（正解例 1）日本に対しては、「内需主導型成長を確保するため、必要に応じて財政上の措置の実施」を求めている。

（正解例 2）男性は窃盗容疑で勾留中、尿から覚醒剤の陽性反応が出たため、１０月１９日に逮捕された。

（正解例 3）戦時下で資材が制限され、群雄割拠状態だった美術団体も一元化された。

（誤り例 1）今世紀に起きた二つの世界大戦の下(上) に戦後五十年がある。

（誤り例 2）日米安保体制の下(中) で何もしないでじっとしている一国閉じこもり型だった。

（誤り例 3）規制委の中(下) に省庁や専門分野を超えた組織が必要と主張した。

表 4.15 に単語組「上」「中」「下」の機械学習の分類結果を示す．表 4.16 に単語組「上」「中」「下」の正規化 α 値に基づいた機械学習が参考にした素性を示す.

表 4.15: 機械学習の分類結果 (「上」「中」「下」)

	データ数	再現率	適合率
上	1000	0.71	0.73
中	1000	0.73	0.71
下	1000	0.72	0.71
総数	3000	0.72	0.72

表 4.16: 機械学習が参考にした素性 (正規化 α 値:「上」「中」「下」)

上		中		下	
素性	正規化 α 値	素性	正規化 α 値	素性	正規化 α 値
下	0.54	家	0.51	上	0.59
事実	0.52	活動	0.49	体制	0.58
インターネット	0.49	予定	0.46	手	0.51
安全	0.47	農業	0.46	不況	0.50
歳	0.45	高校	0.44	担当	0.49

前述の「上下左右」で用いた「上」「下」に「中」という単語を追加して実験を行った.

なおこの実験における「上」と「下」は再度無作為に選び直した,「上下左右」のものとは違う別の 1000 文である.

その結果, 上では「インターネット上」や「安全上」, 下では「不況下」や「担当下」のような先述の実験ではあまり出てこなかった単語が素性に浮上した.

また, 上や下だけでなく,「中」からも「活動中」や「予定中」,「農業中」のように「何かをし続けているさま」を表す言い回しが頻出したことから文法的に役立つ素性を得ることができた.

4.4.7 「牛肉」「豚肉」「鶏肉」

(正解例 1) 米国産牛肉の輸入制限緩和は前回の会談で大統領から進展を求められた課題。

(正解例 2) これについて、農水省は「家庭で食べるテーブルミートとしての豚肉の量は減っているが、ハム、ソーセージに加工した形での消費は伸びている。

(正解例 3) 同社はその後も苦戦続きで、昨年は取引先だった中国の工場で期限切れ鶏肉使用が発覚。

（誤り例 1）渡されたのは、鶏肉(牛肉) を甘辛く煮た唯一の得意料理。

（誤り例 2）牛肉(豚肉) は長時間煮込むことで、余分な脂が抜け落ち、軟らかくなります。

（誤り例 3）「モスバーガー」を展開するモスフードサービスはタイ産など輸入牛肉(鶏肉) を使用。

表 4.17 に単語組「牛肉」「豚肉」「鶏肉」の機械学習の分類結果を示す．表 4.18 に単語組「牛肉」「豚肉」「鶏肉」の正規化 α 値に基づいた機械学習が参考にした素性を示す.

表 4.17: 機械学習の分類結果 (「牛肉」「豚肉」「鶏肉」)

	データ数	再現率	適合率
牛肉	500	0.65	0.65
豚肉	500	0.59	0.62
鶏肉	500	0.62	0.59
総数	1500	0.62	0.62

表 4.18: 機械学習が参考にした素性 (正規化 α 値:「牛肉」「豚肉」「鶏肉」)

牛肉		豚肉		鶏肉	
素性	正規化 α 値	素性	正規化 α 値	素性	正規化 α 値
乳製品	0.55	味付け	0.64	鶏	0.79
割	0.53	関税	0.63	中国	0.61
輸出	0.52	豚	0.61	期限切れ	0.57
日本	0.52	イスラム教	0.51	タマネギ	0.54
オレンジ	0.50	ベーコン	0.49	安全	0.46

「牛肉」からは「米国」,「オーストラリア」,「日本」のように肉牛が多く飼育されている国が多く見受けられたほか,「輸出」,「自由」,「値上げ」など関税に関係する単語や,「自動車」や「オレンジ」など貿易の対象となるものも多く含まれていた.

「豚肉」の素性として豚肉の加工食品である「ソーセージ」や「ベーコン」,「ハム」などのほか，豚がイスラム教において神聖なものであることから「イスラム教」,「禁止」,「戒律」などの単語も見受けられた．また，豚肉の素性には「味付け」,「コショウ」,「みそ」など味付けに関する単語があり，新聞記事にも料理の作り方が掲載され

ている頻度が多く見受けられたので，家庭料理でも簡単に調理することができる手軽な肉であることがわかった.

「鶏肉」においては「期限切れ」という言葉がかなりの頻度で出現していた．これは2014 年 7 月に起きたマクドナルドの取引先だった中国の食品会社が「消費期限切れ」の鶏肉を使用していた事件によるもので，同様に「マクドナルド」や「加工」,「安全」といった単語も多く出現していた．また,「豚肉」同様，新聞記事に多くの調理法が掲載されており,「タマネギ」や「ゴボウ」,「豆」など一緒に調理されやすい食材も多く見受けられた.

4.4.8 「大雨」「洪水」「大雪」

（正解例 1）気象庁は暴風や高波、大雨に厳重な警戒を呼びかけている。

（正解例 2）タイの大洪水で首都バンコクの王宮にも水が迫った。

（正解例 3）同気象台は、青森、岩手、福島各県の一部を除く東北地方に大雪注意報を発令した。

（誤り例 1）近畿地方でも２５日から２６日にかけて大雪(大雨) や強風に見舞われる危険があり、気象庁は警戒を呼びかけている。

（誤り例 2）低温や大雨(洪水) など世界的な異常気象の原因はエルニーニョ現象のためという説がある。

（誤り例 3）洪水(大雪) の影響は十三日も残り、近畿各地の高速道路は阪神高速道が十二日から二日間にわたって通行止めになり、開通した路線でも渋滞の列。

表 4.19 に単語組「大雨」「洪水」「大雪」の機械学習の分類結果を示す．表 4.20 に単語組「大雨」「洪水」「大雪」の正規化 α 値に基づいた機械学習が参考にした素性を示す.

表 4.19: 機械学習の分類結果 (「大雨」「洪水」「大雪」)

	データ数	再現率	適合率
大雨	1000	0.79	0.80
洪水	1000	0.83	0.83
大雪	1000	0.82	0.80
総数	3000	0.81	0.81

表 4.20: 機械学習が参考にした素性 (正規化 α 値:「大雨」「洪水」「大雪」)

大雨		洪水		大雪	
素性	正規化 α 値	素性	正規化 α 値	素性	正規化 α 値
台風	0.83	タイ	0.85	雪	0.85
洪水	0.70	死亡	0.61	日本海	0.65
九州	0.67	工場	0.51	甲信	0.57
梅雨	0.65	汚染	0.49	寒波	0.56
浸水	0.55	森林	0.48	停電	0.55

「大雨」,「洪水」,「大雪」全てにおいて正解率が約 8 割となり，かなり使い分けが必要な単語であることがわかった.

「大雨」の素性として「雨」や「梅雨」,「台風」など雨を伴う気象のほか,「土砂」,「浸水」,「避難」など大雨による災害に関する単語も頻出していた．また，大雨が多く降った地方も素性に出ていたが，一番正規化 α 値が高かったのは「九州」であった.

「洪水」の素性として,「タイ」や「バンコク」が挙げられる．タイでは 2011 年に起きた大洪水によって多くの死者が多く出たことが素性を読み取ることで得ることができた．また，洪水による被害は企業や自然にも及び，素性として「工場」,「汚染」,「森林」,「米」などの単語も多く出現していることがわかった.

「大雪」の素性として，やはり「日本海」,「甲信」,「北海道」など，雪が降る地域が多く出現していた．また,「大雨」と同様,「停電」,「事故」,「孤立」など大雪時に起こりうる災害も多数出現していた.

4.4.9 「風」「林」「火」「山」

（正解例 1）あれほど問題になりながら「どこ吹く風」の不可解。

（正解例 2）小中学校で、ＢＧＭとして流すほか、ブナ林保護の音楽祭の開催も検討している。

（正解例 3）第一巻の内容をみると、「火の用心」と米などの入札関係が全体の三分の二を占める。

（正解例 4）ここに、もうひとつの日本一低い山があるらしい。

（誤り例 1）雨、林(風) の中をあえてオープンで走るところに粋を感じるようなところがある。

（誤り例 2）クロマツ山(林) や芝生の緑に囲まれ、面積は２２・７ヘクタール、砂浜の延長は約１キロある。

（誤り例 3）林(火) おこしや、空き缶でご飯を炊く方法、段ボールを使ったベッドの作り方などを学んだ。

（誤り例 4）米など食品への放射能汚染が懸念され、それは火(山) の木々、海にも広がる。

表 4.21 に単語組「風」「林」「火」「山」の機械学習の分類結果を示す．表 4.22 に単語組「風」「林」「火」「山」の正規化 α 値に基づいた機械学習が参考にした素性を示す．

表 4.21: 機械学習の分類結果 (「風」「林」「火」「山」)

	データ数	再現率	適合率
風	1000	0.75	0.74
林	1000	0.82	0.84
火	1000	0.82	0.84
山	1000	0.66	0.67
総数	4000	0.76	0.76

表 4.22: 機械学習が参考にした素性 (正規化 α 値：「風」「林」「火」「山」)

風		林		火		山	
素性	正規化 α 値	素性	正規化 α 値	素性	正規化 α 値	素性	正規化 α 値
F1：嘉	0.56	自_自立語単品：名詞	0.58	F1：(	0.50	風	0.53
雨	0.49	ブナ	0.45	金	0.44	自_自立語最初単品：名詞	0.44
波	0.42	世界	0.45	油	0.40	海	0.41
方向	0.37	人工	0.36	たばこ	0.38	調査	0.38
音	0.35	自然	0.35	聖火	0.36	双葉	0.36

「風」「林」「火」「山」全てにおいて，名詞単体で自立語や付属語になりやすいことが多いことが言える．

各単語の素性として,「風」は「方向」や「音」など風を感じる表現が多く含まれていたほか,「吹く」,「乗る」という動詞が後に続く傾向があるという文法における素性があった．また,「立ち」という動詞も多く出現していることがわかり，記事を詳しく調べると宮崎駿監督の映画「風立ちぬ」に関する情報が多く出現していることが判明した.

「林」の素性として,「林」さんのように名字として多く出現するほか,「ブナ」や「スギ」,「カラマツ」など木の種類に関する単語や「人工」や「原生」,「保護」を前後につける使われ方をすることがわかった.

「火」の素性として,「火曜日」を示すものが多く出現し，曜日が「(火)」のように記されているほか,「出る」,「かける」という動詞が助詞の後に続く傾向があり，また「祭り」や「用心」という名詞も続く傾向があるということがわかった.

「山」の素性として,「スキー」のように自然を活かした活用がある反面,「火事」や「事故」,「不明」のような災害が多く起こりうる場所であるということがわかった.

また,「風」,「山」両方の素性として,「時津風」や「双葉山」のような相撲取りのしこ名に使用され，主に新聞記事の取組結果欄に多く出現することがわかった.

4.4.10 「晴れ」「雨」「雪」

（正解例 1）晴れ舞台で更なる高みを見据えた。

（正解例 2）西日本と東日本では５日、東北地方では６日にかけて断続的に雷を伴った非常に激しい雨が降る所がある見込み。

（正解例 3）翌二日はカチンカチンに凍った雪を持ち帰り溶かしてみた。

（誤り例 1）この原稿が皆さんに届く頃には、冬雪(晴れ) にと復活を祈るばかりだが、なんだか嫌な予感がする。

（誤り例 2）雪(雨) の前兆か、少し湿った風が、息つきしながら街道を抜けて行く。

（誤り例 3）地元大工の男性（６３）は「一晩に５０～６０センチの雨(雪) が降り積もる。

表 4.23 に単語組「晴れ」「雨」「雪」の機械学習の分類結果を示す．表 4.24 に単語組「晴れ」「雨」「雪」の正規化 α 値に基づいた機械学習が参考にした素性を示す.

表 4.23: 機械学習の分類結果 (「晴れ」「雨」「雪」)

	データ数	再現率	適合率
晴れ	1000	0.93	0.97
雨	1000	0.76	0.75
雪	1000	0.77	0.75
総数	3000	0.82	0.82

表 4.24: 機械学習が参考にした素性 (正規化 α 値:「晴れ」「雨」「雪」)

晴れ		雨		雪	
素性	正規化 α 値	素性	正規化 α 値	素性	正規化 α 値
舞台	0.76	雪	0.67	氷	0.57
予報	0.57	試合	0.61	積雪	0.57
天気	0.54	酸性	0.52	センチ	0.56
西日本	0.52	ミリ	0.49	白	0.54
後	0.48	避難	0.47	スキー	0.53

「晴れ」,「雨」,「雪」の組が全ての組の中で一番正規化 α 値が高かった.

「晴れ」の素性として一番多かったのは「晴れ舞台」という表現であった. また, 晴れ舞台のように天候の様子に限らず,「疑惑が晴れる」や「表情が晴れない」のような慣用句として有用な素性も得られた.

「雨」の素性として,「雪」や「台風」のように天候の様子についての単語の頻度が高く, また降水量の単位である「ミリ」や「傘」,「中止」や「避難」のような単語も得られていた.

「雪」の素性として,「氷」や「積雪」などはもちろんだが,「雨」はミリに対して「雪」は「センチ」で表される点に面白さを感じた. また雪の素性に「父」という単語があり, 詳しく調べると

- 雪が降る中、かんじきを履いた今は亡き父の背中におんぶされ、自宅から４キロほど離れた町の歯科医院に向かいました。
- 雪国の人にとっては、その年の冬の雪が大雪か小雪かは生活の、というよりは死活の問題であったから、いろいろな予知情報が父祖の時代から伝えられている。

のような記事が出現したことから, 冬は父親の助けや教訓が活きる季節なのではないかと考えられる.

4.4.11 「朝」「昼」「夕」「夜」

（正解例 1）九日朝、本社のほかにトップの自宅にまで家宅捜索が及び、関係者に新たな緊張が走った。

（正解例 2）兵庫県警東灘署によると、男性は当日昼ごろ、倉庫で冷凍の魚やドライアイスを保冷車に積み込んで出発。

（正解例 3）同日夕、県庁で臨時記者会見を開き、正式表明する。

（正解例 4）ＡＰ通信は、イスラエル空軍が２９日夜から３０日未明にかけ、シリア領内で、レバノンのイスラム教シーア派武装組織ヒズボラ向けに地対空ミサイルを輸送していたトラックの隊列を空爆したと報じた。

（誤り例 1）昼(朝) の光は、輝くばかりの霜の白さを、軟障の内にも見せました。

（誤り例 2）朝(昼) から街頭で「反ＴＰＰ（環太平洋パートナーシップ協定)、反消費増税を訴える。

（誤り例 3）センターは１５日昼(夕) 時点で、影響を受けたのは４５６５人と発表したが、再調査で人数が増えた。

（誤り例 4）辞任報道に見舞われたロシアのチェルノムイルジン首相は十九日朝(夜)、南部の保養地ソチでの休暇を切り上げモスクワに戻った。

表 4.25 に単語組「朝」「昼」「夕」「夜」の機械学習の分類結果を示す．表 4.26 に単語組「朝」「昼」「夕」「夜」の正規化 α 値に基づいた機械学習が参考にした素性を示す．

表 4.25: 機械学習の分類結果 (「朝」「昼」「夕」「夜」)

	データ数	再現率	適合率
朝	1000	0.53	0.57
昼	1000	0.61	0.62
夕	1000	0.56	0.51
夜	1000	0.46	0.47
総数	4000	0.54	0.54

表 4.26: 機械学習が参考にした素性 (正規化 α 値：「朝」「昼」「夕」「夜」)

朝		昼		夕		夜	
素性	正規化 α 値	素性	正規化 α 値	素性	正規化 α 値	素性	正規化 α 値
捜索	0.52	夜	0.77	交渉	0.45	朝	0.53
学校	0.51	夕	0.51	帰国	0.43	世界	0.48
港	0.49	自民党	0.47	分間	0.42	晩さん	0.44
君	0.44	八月	0.43	修飾_自立語最後単：会見	0.41	未明	0.43
出勤	0.39	出発	0.40	正式	0.41	判明	0.40

「朝」の素性として,「出勤」や「通学」,「学生」などの単語が多いほか,「朝早く」という表現や「捜索」,「見つかる」のような事件の捜索や発見に関する単語も多く見受けられた.

「昼」の素性として,「昼から夕方」や「昼から夜」にかけてという表現が多いほか, 政党や議員，主席などが会談やどこかしらに「出発」する時間であることもわかった.

「夕」の素性として，著名人や自衛隊等が「帰国」したり,「会見」を行う時間であることがわかった．また記事を確認すると，夕方に行われる会見では「正式な発表」がなされることが多いということが知見として得られた.

「夜」の素性として,「昼」と同様に「夜から未明」,「夜から朝」にかけてという表現が多いほか,「朝」と同様に「夜遅く」という表現や「判明」という単語が頻出しているなど，朝と夜という相対的な単語と思われているものが実はよく似た素性を持っているという知見を得た.

4.4.12 「花」「鳥」「風」「月」

（正解例 1）雪が残る２月ごろから花をつけることから名付けられた雪割り桜。

（正解例 2）愛知県豊橋市の養鶏農場の鶏の大量死は、高病原性鳥インフルエンザの感染と２７日断定された。

（正解例 3）どこかから風に乗って、競馬中継のかん高いアナウンスが聞こえてきた。

（正解例 4）太陽が月に隠されてリング状に輝いて見える金環日食が２１日、日本の広い範囲で観測可能になる。

（誤り例 1）だが、実際、風(花) が盛りになっても、どこかにうつろな気持ちが残るのはなぜだろう。

（誤り例 2）それだけに、宗教的背景からも権威の誇示からも無縁に、「月(鳥) の背に乗って大空へ」という作者の個人的な、素朴な夢の表現に見えてうれしかった。

（誤り例 3）旧帝国ホテルの設計者として知られるライト鳥(風) の建築である。

（誤り例 4）米国では、民間企業が花(月) や火星の探査に乗り出そうとするなど、宇宙探査の裾野が広がり始めた。

表 4.27 に単語組「花」「鳥」「風」「月」の機械学習の分類結果を示す．表 4.28 に単語組「花」「鳥」「風」「月」の正規化 α 値に基づいた機械学習が参考にした素性を示す．

表 4.27: 機械学習の分類結果 (「花」「鳥」「風」「月」)

	データ数	再現率	適合率
花	1000	0.72	0.70
鳥	1000	0.70	0.71
風	1000	0.75	0.74
月	1000	0.82	0.84
総数	4000	0.74	0.74

表 4.28: 機械学習が参考にした素性 (正規化 α 値：「花」「鳥」「風」「月」)

花		鳥		風		月	
素性	正規化 α 値	素性	正規化 α 値	素性	正規化 α 値	素性	正規化 α 値
貴	0.41	羽	0.65	雨	0.48	地球	0.48
一つ	0.40	魚	0.46	メートル	0.44	夜	0.47
葉	0.40	感染	0.40	出版	0.42	太陽	0.41
梅	0.38	バード	0.40	波	0.40	修飾_付属語単：(	0.39
愛	0.31	保護	0.40	韓国	0.37	回	0.38

「花」の素性として,「梅」や「桜」,「チューリップ」などの植物の名前が挙げられたほか，花には「咲く」や「好き」という動詞が後に続く傾向があるという文法上の素性があった.「風林火山」の時と同様に，力士の「しこ名」に多く用いられることがあることもわかった.

「鳥」の素性として,「羽」や「鳴き声」,「虫」など鳥の生態に関係する単語が多く出現していた．また「花」と同様に,「ペンギン」や「トキ」のように鳥類の名前も出現していたが,「花」ほど多くは得られなかった．また鳥は人間と関わりが深い動物でもあるため,「繁殖」や「人工」,「野生」,「保護」のような鳥を保護したり，繁殖を高

めることを連想させる単語も多い反面,「インフルエンザ」や「感染」のような単語も多いことがわかった.

「風」の素性を得た文章は「風林火山」で用いた「風」という単語とは違う 1000 文をランダムに選び直した．しかし,「風林火山」の時に現れた素性とは違い,「韓国風」や「山小屋風」のように「何かに似たもの」という意味や,「このような風に」のような「様子」を意味する「風」の言い回しが多く素性として得られた．また,「ゴルフ」や「ヨット」などの風の影響が試合を左右するスポーツも素性として存在した.

「月」の素性として，天体の「月」だけでなく，1ヵ月という意味の「月」，月曜日という意味の「月」が混在していた．また，1ヵ月を表す「月」の周辺単語には「回」や「万」,「円」のように単位や金額が周辺単語に現れる頻度が高いこと，曜日の「月」は (月) のように表記されやすいことがわかった.

4.4.13 「赤」「青」「黄」

（正解例 1）だが辞書を引き始めて８カ月がたつと辞書の中は赤ペンで真っ赤になった。

（正解例 2）青学では、私のような元気のいい女の子がたくさんいました。

（正解例 3）ライオンの幼獣にみられる斑紋とトラの黒いしま模様のある黄褐色の幼獣は「ライガー」と呼ばれ、「世紀の珍獣だ」と騒がれた。

（誤り例 1）本来は白と黄(赤) の２種類だが、当面は白１色に生産を絞る。

（誤り例 2）事故当日の午前十時十四分、信楽駅上り出発信号が赤(青) に変わらなくなるトラブルが発生。

（誤り例 3）種類は多く、色は白、ピンク、紫、ブルー、青(黄) など。

表 4.29 に単語組「赤」「青」「黄」の機械学習の分類結果を示す．表 4.30 に単語組「赤」「青」「黄」の正規化 α 値に基づいた機械学習が参考にした素性を示す.

表 4.29: 機械学習の分類結果 (「赤」「青」「黄」)

	データ数	再現率	適合率
赤	1000	0.73	0.71
青	1000	0.67	0.66
黄	1000	0.77	0.81
総数	3000	0.72	0.72

表 4.30: 機械学習が参考にした素性 (正規化 α 値:「赤」「青」「黄」)

赤		青		黄	
素性	正規化 α 値	素性	正規化 α 値	素性	正規化 α 値
青	0.66	音	0.56	韓国	0.63
壁	0.51	赤	0.54	褐色	0.52
バラ	0.51	自_自立語単：シソ	0.50	山	0.51
新月	0.49	黄色	0.49	ボーダーライン	0.51
ペン	0.48	運転	0.47	Ｒ１：信号	0.50

「赤」「青」「黄」すべての色において，国旗や信号，物のカラーバリエーションを表す文章に多く出現していたため,「赤と青」,「青と黄」,「赤と黄」，もしくは「赤，青，黄」すべてが文中に含まれている記事が多くあった.

各々の素性を見ていくと,「赤」の素性として色の要素では「青」や「黒」が多く見受けられ，赤色のものとして「ペン」や「バラ」,「富士」などの素性があった.

また，素性の一つに「新月」があったため詳しく記事を調べると，

- 赤十字がキリスト教を連想させるとの理由から、イスラム教圏では赤新月を用いる。
- 赤十字国際委員会は、シリア赤新月社などの救助隊が地区へ２日に、食料や医薬品を運び込むと発表した。

という文を確認し，私は赤十字社がイスラム教圏に対する配慮を行っているということを初めて知ることができたため，いい知見が獲得できたのではないかと考える.

「青」の素性として，色の要素では「赤」や「黄色」が多く見受けられ，青色のものとして「青ジソ (シソ)」や「青切符」,「青イソメ」などの素性があった.

記事を調べると「青切符」とはさほど重大でない交通反則切符のことを指し,「青イソメ (アオイソメ)」は釣りでよく使われる餌のことであることがわかり，こちらも初耳であったため，いい知見が獲得できたのではないかと考える.

「黄」の素性として，色の要素では「褐色」や「水色」が多く見受けられた．黄色のものは「黄斑」,「ユズ」,「黄色ブドウ球菌」などの素性があった．また「黄」において出現率が高かった「韓国」や「中国」では「黄」という名字が多く存在することがわかった．

第5章　全結果における考察とまとめ

今回の実験において過去の研究結果[1]から,「晴れ」「雨」「雪」のような天気に関するものは正解率が約 0.8 を上回ったため，使い分けが必要な単語組であると思われる反面,「春」「夏」「秋」「冬」のような単語では正解率は約 0.5 と，無作為に単語を挿入した場合よりは高いものの実験したものの中では低く，使い分けがあまり必要ではない単語であるとわかった.

また「東」と「西」や「上」と「下」のようにお互いが対となるような単語が単語組内に存在する場合,「事実上」や「水面下」,「晴れ舞台」のように該当単語が慣用句として使用される場合,「火」[2]や「月」[3]などで述べたように曜日や短縮形を括弧などで示す場合，正規化 α 値が約 0.8 と高くなる傾向があることがわかった.

今回の実験で得られた素性として,「上」「下」「左」「右」などにおいて,「事実上」や「水面下」というような使われ方が存在するが,「事実右」や「水面左」のような使われ方はしないことから，物体における「上」や「下」という一般的に予想される意味だけではなく，文中での使われ方に大きな差がある単語も存在するということを知見として得ることができた.

他にも「秋」の素性として「疑惑」や「問題」,「中学」の素性として「いじめ」や「興味」,「夕」の素性として「帰国」や「正式」など，元の単語からは連想されにくい単語や，南の素性として「スーダン」,「鶏肉」の素性として「マクドナルド」，赤の素性として「赤新月」など，実験に用いた新聞記事によって初めて知り得た興味深い単語や時事も存在したので，本研究は非常に有益なものであったと考えられる.

[1]第 2 章 先行研究を参照
[2]第 4 章 4.9 節を参照
[3]第 4 章 4.12 節を参照

第6章　おわりに

本研究では機械学習を用いて 3,4 組の単語の使い分けと知見獲得の研究を行った．本研究の成果は 2 つある．

第 1 の成果として，3,4 組の単語組 13 組 47 単語について実験を行った結果，提案手法が 0.68，ベースライン手法が 0.28 であったため，機械学習を用いる提案手法の正解率が無作為に単語を挿入した際に正解となるベースライン手法よりも高いことを確認した．これにより，今回提案した手法自体が 3,4 組の単語の使い分けに対して有用であると考えられる．

第 2 の成果として，3,4 組の単語組について素性を分析し，使い分けに役立つ素性が多く得られた．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．また,「晴れ，雨，雪」では「晴れ」のみが「晴れ舞台」や「疑惑が晴れる」というような慣用句として有用な素性を多く得ることができた．

そのほか，該当単語を抜き出した新聞記事から当時の情勢や事件，言葉の使い方も知見として得られた．

今後の課題として，今回の実験では実験データを同数のみでしか研究を行っておらず，また各単語 1000 文または 500 文のみでしか行っていないため，データ数を同数ではなく単語の出現率に依存した場合での研究や 1000 文以上を対象にした研究も行う必要があると考えている．また，今回の研究では被験者実験も行っていない為，こちらも併せて今後の課題としたい．

謝辞

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

参考文献

[1] 強田吉紀, 村田真樹, 三浦智, 徳久雅人. 機械学習を用いた同義語の使い分け. 言語処理学会第 19 回年次大会, 2013.

[2] 中瀬充暁. 教師あり機械学習を用いた副詞の類義語の使い分け. 卒業論文, 鳥取大学工学部知能情報工学科, 2015.

[3] 赤江涼太. 教師あり機械学習を用いた同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, 2017.

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

[5] 佐々本暖久. 教師あり機械学習を用いた対義語の置き換え判定. 卒業論文, 鳥取大学工学部, 2019.

[6] 西尾寅弥. 同義語間の選択についての調査. 群馬大学教育学部紀要, 人文社会科学編, Vol. 29, pp. 161–182, 1979.

[7] 小島正裕, 村田真樹, 南口卓哉, 渡辺靖彦. 機械学習を用いた表記選択の難易度推定. 言語処理学会第 17 年次大会発表論文集, pp. 300–303, 2011.

[8] 小学館辞典編集部. 使い方の分かる類語例解辞典. 小学館, 1994.

[9] 山口仲美, 佐藤有紀. 「擬音語・擬態語」使い分け帳. 山海堂, 2006.

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

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

[12] 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.

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

[14] 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.