概要

本研究は動詞・形容詞の類義語の使い分けを教師あり機械学習を使用して行う．

類義語とは，語形は異なるが意義がほぼ同じである語のことである．類義語間においては，使い分けが必要な場合がある．例えば「探し回る」と「探し求める」という類義語対では,「探し回る」は「時間」「日」といった短い時間を表す場合に用いられることが多いが,「長年」「旅」といった長い時間を表す場合には「探し求める」が用いられる．また,「あちこち探し回る」とは表現するが,「あちこち探し求める」とは普通表現しない．ある類義語対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，その類義語対は特に使い分けの必要な類義語対とわかり，機械学習での性能が低かった類義語対は，それほど使い分けの必要がないと推定できる．

また，機械学習が使用した素性を分析して，動詞・形容詞の類義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既存の辞書から獲得できた動詞・形容詞の類義語対を対象に行う．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，動詞 22 対，形容詞 10 対の類義語対を用いた実験において，類義語のうち最も頻度の高い語を常に選択するベースライン手法の正解率が動詞では 0.77，形容詞では 0.70 であるのに対して，機械学習を用いる提案手法は動詞では 0.88，形容詞では 0.81 の正解率であった．提案手法には，ベースライン手法よりも高いという有用性がある．

2 つ目は，動詞・形容詞の類義語対ごとの機械学習の性能に基づき，動詞・形容詞の類義語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，特に使い分けが必要であるとされた動詞・形容詞の類義語対に「探し回る」と「探し求める」や「易しい」と「手軽い」などがあった．また，いくつかの類義語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

類義語とは，語形は異なるが意義がほぼ同じである語のことである．例としては「探し回る」と「探し求める」などがある．類義語に関する研究では，コーパスから類義語を獲得する研究 [1] や 西尾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 [2] などがある．また，小島らは異表記の使い分けを機械学習で行っている [3]．小島らが機械学習を用いて使い分けを行った対象である異表記とは， 同じ語の表記が異なるもののことであり,「しょう油」と「醤油」が異表記対の例となる．小島らの研究では，異表記の対を機械学習の対象としているが，類義語全般を対象とはしていない．また，強田，中瀬らは EDR 電子化辞書から得られる類義語を利用し，機械学習による分類性能の高い名詞・副詞の類義語の使い分けの研究を行っている [4, 5]．

本研究では，機械学習の性能や素性が類義語の使い分けに役立つと考え，機械学習を用いて動詞・形容詞の類義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類義語の選択，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．

本研究では EDR 電子化辞書から得られる動詞・形容詞の類義語を利用する．

類義語間においては，使い分けが必要な場合がある．類義語とは，語形は異なるが意義がほぼ同じである語のことである．例えば「探し回る」と「探し求める」という類義語対では,「探し回る」は「時間」「日」といった短い時間を表す場合に用いられることが多いが,「長年」「旅」といった長い時間を表す場合には「探し求める」が用いられる．また,「あちこち探し回る」とは表現するが,「あちこち探し求める」とは普通表現しない．

本研究では，機械学習によって動詞・形容詞の類義語の使い分けも目指すが， 動詞・形容詞の類義語の使い分けが特に必要なものとそれほど必要でないものの分類も試みる．機械学習によって類義語を推定しやすい場合は，類義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類義語の使い分けが明瞭でないということがわかる．これらの知見は，動詞・形容詞の類義語の使い分けに役立つと思われる．

本研究の主な主張点を以下に整理する．

- 本論文は類義語の使い分けのために機械学習を使用し，複数の動詞・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示した．

- 機械学習に基づき動詞・形容詞の類義語の使い分けを行った．動詞 22 対，形容詞 10 対の類義語対を用いた実験において，類義語のうち最も頻度の高い語を常に選択するベースライン手法の正解率が動詞では 0.77，形容詞では 0.70 であるのに対して，機械学習を用いる提案手法は動詞では 0.88，形容詞では 0.81 の正解率であった．提案手法には，ベースライン手法よりも高いという有用性がある．

- 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・形容詞の類義語の使い分けに重要な情報を把握することができる．いくつかの類義語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した．例として，「探し回る」と「探し求める」といった類義語対では「時間」「日」といった短い時間を表す場合は「探し回る」を用いるが,「長年」「旅」といった長い時間を表す場合は「探し求める」を使う．

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，本研究で使用する動詞・形容詞の類義語対の説明を行う．第 5 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 6 章では，第 5 章の結果から考察を行う．また，具体的な類義語対の考察も行い，どのような情報が類義語の使い分けに役立ったのかを明らかにする．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，西尾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，中瀬ら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．

2.1　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 [2]．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してもらう．それを年齢・性別・地域で分けてどのような選択の違いが見られたかを調べる．

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 の通りとする．

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

調査結果を簡潔に記すと，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，例えば「是非」と「ぜひ」など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 [3]．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 年から 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，更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する．

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．実験方法は各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する．

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ている [4]．

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，以下の条件を満たす名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[6] 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

中瀬らは，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ている [5].

中瀬らは副詞の類義語の使い分けのために機械学習を使用し，複数の副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.

中瀬らは EDR 電子化辞書と 1991 年から 1995 年の毎日新聞を使用し，以下の条件を満たす副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 年から 1995 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[6] 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 年から 1995 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 (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」,「ことごとく」と「すっかり」があった．

本研究ではこの先行研究をもとに実験を行う．先行研究では名詞・副詞の類義語を扱っているが本実験では動詞・形容詞の類義語を扱う．

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

使い分けをしたい類義語対 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，正しく類義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,「探し回る」と「探し求める」の例を以下に挙げる．

甲高い鳴き声で仲間に合図しながら、コケや植物の茎、実、葉などのエサを探し回る。 廃棄物の流れは、国と国の“ 間 ”で最も抵抗の少ない道を探し求める。

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．

甲高い鳴き声で仲間に合図しながら、コケや植物の茎、実、葉などのエサを X。 廃棄物の流れは、国と国の“ 間 ”で最も抵抗の少ない道を X。

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を利用する．

機械学習により類義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類義語の使い分けの再現率の高さごとに高・中・低を設定する．類義語対の語 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 8 割以上，中を再現率 8 割未満 5 割以上，低を再現率 5 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 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．類義語の使い分けでは，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: 動詞・形容詞の類義語対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 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　実験に用いる動詞・形容詞の類義語対

本章では実験に用いる動詞・形容詞の類義語対の説明を行う． 4.1 節で動詞・形容詞の類義語の認識方法を説明する．4.2 節で実験に用いる動詞・形容詞の類義語対の選定方法を説明する．

4.1　EDR 電子化辞書を用いた動詞･形容詞の類義語の認識

本研究では，2 つの語が類義語対であるかの判定に EDR 電子化辞書を利用する．

EDR 電子化辞書は 10 種類の辞書からなり，本研究ではその中の 1 つである,「日本語単語辞書」と「概念辞書」を使用する．日本語単語辞書には，約 26 万語収録されており，各語に対して「品詞」や「活用情報」など複数の情報が付与されている．その情報の 1 つに,「概念識別子」という情報がある．この概念識別子は 16 進整数で表されており，概念辞書に各識別子の意味が記述されている．このため，日本語単語辞書からは概念識別子を通して概念辞書を参照することにより語の意味を獲得できる．概念識別子が同じ語どうしを類義語対と判定する． 概念辞書には，約 41 万の概念が収録されている．日本語単語辞書によって語に与えられた概念識別子の例を表 4.1 に示す．また，概念辞書によって記述されている概念識別子と概念の関係の例を，表 4.1 の識別子を用いて表 4.2 に示す．

表 4.1: 語と概念識別子の対応例

語	識別子
探し回る	3cec34
探し求める	3cec34 0ec740 10c07a 1f239a 10e00a
むつまじい	3cfc19
近しい	3cfc19
ほったらかす	3ce673
怠る	3ce673 0ef497 0f2210 0fb918

表 4.2: 概念識別子と概念の対応例

識別子	概念
3cec34	探し求める
0ec740	企業などが採用すべき人を探し求める
10c07a	住む家を探し求める
1f239a	(未知のものを) 探し求めることができる
10e00a	奇怪なものや異常なことに興味を持って、それを探し求めること
3cfc19	親密だ
3ce673	するべきことをしないで、ほうっておくさま
0ef497	精進を怠る
0f2210	道理がわからず怠ること
0fb918	仕事を怠る

4.2 実験で用いる類義語の選定

本研究では，新聞記事に出現する語について機械学習を用いた動詞・形容詞の類義語の使い分けを行う． 新聞記事には，動詞類義語対獲得には 1991 年から 1995 年の 5 年分の毎日新聞を使用し，形容詞類義語対獲得には加えて 2011 年から 2015 年の 10 年分の毎日新聞を使用する．

以下の条件をすべて満足する語の対を取り出し，実験に用いる動詞・形容詞の類義語対とする．

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[6] 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

条件 1 は，今回使用した EDR 電子化辞書において，同一の概念識別子は概念辞書により同一の概念として定義されており，同一の概念識別子をもつ語どうしは類義であるとみなせるため設定する．条件 2 は新聞内で多く使われている語について調査を行うためで，機械学習に用いる学習事例の数を大きくすることに繋がる．条件 3 の代表

表記が異なるものを扱うのは，異表記における使い分けを 本研究で扱わないようにするためである．異表記対は同じ代表表記を持つ． 異表記対の使い分けはすでに文献 [3] で扱われており， 本研究では扱わないため条件 3 を設けた．

名詞・副詞の類義語の使い分けには上記の条件に加えて，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 つであることといった条件があった．しかし，動詞・形容詞の類義語対の場合は多義語であることがほとんどであり，概念識別子が 1 つであるといった条件を加えると，実験に使用する類義語対の数が少なくなるため，この条件を省いた．例えば「探し回る」は概念識別子が 1 つのみであるが,「探し求める」は細かい意義の違いも含めると概念識別子が他に 4 つある．本実験では上記の理由より，これを考慮せず，実験と考察を行う．

これらの条件を満足する動詞の類義語対は 22 対あり，形容詞の類義語対は 10 対あった． これらを実験に用いる類義語対とする．

実際に獲得した動詞類義語対を表 4.3 に示し，形容詞類義語対を表 4.4 に示す．

表 4.3: 獲得した動詞類義語対

1	煎る	炒る
2	代わる	入れ替わる
3	そそぐ	言い逃れる
4	準じる	準ずる
5	似る	類する
6	奪い取る	分捕る
7	はみ出す	はみ出る
8	見限る	見捨てる
9	咲き誇る	咲き乱れる
10	投げ込む	投げ入れる
11	薄らぐ	薄れる
12	見定める	突き止める
13	さらけ出す	届け出る
14	ほったらかす	怠る
15	取りやめる	取り消す
16	群がる	群れる
17	取り去る	除く
18	いら立つ	焦る
19	探し回る	探し求める
20	せしめる	勝ち得る
21	買い求める	買い入れる
22	書き込む	書き入れる

表 4.4: 獲得した形容詞類義語対

1	みっともない	見苦しい
2	ずうずうしい	ずぶとい
3	けだるい	だるい
4	痛ましい	涙ぐましい
5	気高い	神々しい
6	注意深い	用心深い
7	眠い	眠たい
8	易しい	手軽い
9	近しい	むつまじい
10	気まずい	面はゆい

第5章　実験

本章では，本研究が行った実験方法を 5.1 節で説明し，実験結果を 5.2 節に示す．

5.1　実験方法

獲得した動詞類義語対 22 対・形容詞類義語対 10 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う．入力文は，動詞では 1991 年から 1995 年，形容詞では加えて 2011 年から 2015 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う．類義語対のうち出現頻度が多かった語を全ての問題の分類先とするものをベースライン手法とし，提案手法とベースライン手法の性能の比較を行う．また，素性を表 3.1 の素性 2 である対象後の前後 3 単語のみとした実験結果と提案手法の性能の比較も行う．

5.2　実験結果

5.2.1　動詞類義語対での実験結果

機械学習の再現率の高さごとに動詞類義語対を分類した割合を表 5.2 に示す．再現率の高さごとの分類は，高を再現率 8 割以上，中を再現率 8 割未満 5 割以上，低を再現率 5 割未満である．また，分類には類義語対のうち，再現率の低い方を基準とする．機械学習の再現率の高さごとに 22 対の動詞類義語対を分類した結果を表 5.1 に示す．提案手法とベースライン手法の動詞類義語対ごとの正解率の平均を表 5.3 に示す． 機械学習の再現率の高さごとの値も示している．提案手法と素性 2 のみを使う実験の正解率の平均を表 5.4 に示す．提案手法とベースライン手法の動詞類義語対ごとの正解率を 22 個の動詞類義語対で比較した結果を表 5.5 に示す．表 5.5 における「差なし」とは，提案手法とベースライン手法の再現率の差が± 0.01 以内であった動詞類義語対の数を示す.「提案手法〇」は「差なし」以外でありかつ提案手法の正解率の方が高かった動詞類義語対の数を,「ベースライン手法〇」は「差なし」以外でありかつベースラ

イン手法の正解率の方が高かった動詞類義語対の数を示す．同様に，提案手法と素性2のみで行う手法の動詞類義語対ごとの正解率を22個の動詞類義語対で比較した結果を表5.6に示す．

表 5.1: 動詞類義語対の再現率の高さごとの割合

再現率の高さ	割合
高	27.2% (6/22)
中	50.0% (11/22)
低	22.7% (5/22)

表 5.2: 再現率の高さごとに分類した動詞類義語対

再現率の高さ	再現率	動詞類義語対
再現率高	8割以上	「さらけ出す」と「届け出る」
		「探し回る」と「探し求める」
		「投げ込む」と「投げ入れる」
		「書き込む」と「書き入れる」
		「せしめる」と「勝ち得る」
		「取り去る」と「除く」
再現率中	7割以上8割未満	「奪い取る」と「分捕る」
		「そそぐ」と「言い逃れる」
		「薄らぐ」と「薄れる」
		「見定める」と「突き止める」
		「煎る」と「炒る」
	6割以上7割未満	「ほったらかす」と「怠る」
		「似る」と「類する」
		「群がる」と「群れる」
		「代わる」と「入れ替わる」
	5割以上6割未満	「買い求める」と「買い入れる」
		「準じる」と「準ずる」
再現率低	4割以上5割未満	「咲き誇る」と「咲き乱れる」
	3割以上4割未満	「いら立つ」と「焦る」
		「見限る」と「見捨てる」
	3割未満	「とりやめる」と「取り消す」
		「はみ出す」と「はみ出る」

表 5.3: 提案手法とベースライン手法の動詞類義語対ごとの正解率の平均

	再現率：高	再現率：中	再現率：低	すべての対
提案手法	0.97	0.88	0.77	0.88
ベースライン手法	0.76	0.76	0.78	0.77

表 5.4: 提案手法と素性 2 のみで行う手法の動詞類義語対ごとの正解率の平均

	再現率
提案手法	0.88
素性 2 のみ	0.85

表 5.5: 提案手法とベースライン手法の動詞類義語対ごとの正解率の比較結果

	再現率：高	再現率：中	再現率：低
提案手法○	5	10	2
ベースライン手法○	0	0	2
差なし	1	1	1

表 5.6: 提案手法と素性 2 のみで行う手法の動詞類義語対ごとの正解率の比較結果

提案手法○	9
素性 2 のみ○	4
差なし	9

5.2.2 形容詞類義語対での実験結果

機械学習の再現率の高さごとに形容詞類義語対を分類した割合を表 5.8 に示す．再現率の高さごとの分類は，高を再現率 8 割以上，中を再現率 8 割未満 5 割以上，低を再現率 5 割未満である．また，分類には類義語対のうち，再現率の低い方を基準とする．機械学習の再現率の高さごとに 10 対の形容詞類義語対を分類した結果を表 5.7 に示す．提案手法とベースライン手法の形容詞類義語対ごとの正解率の平均を表 5.9 に示す． 機械学習の再現率の高さごとの値も示している．提案手法と素性 2 のみを使う実験の正解率の平均を表 5.10 に示す．提案手法とベースライン手法の形容詞類義語対ご

との正解率を 10 個の形容詞類義語対で比較した結果を表 5.11 に示す．表 5.11 における「差なし」とは，提案手法とベースライン手法の再現率の差が± 0.01 以内であった形容詞類義語対の数を示す.「提案手法○」は「差なし」以外でありかつ提案手法の正解率の方が高かった形容詞類義語対の数を,「ベースライン手法○」は「差なし」以外でありかつベースライン手法の正解率の方が高かった形容詞類義語対の数を示す．同様に，提案手法と素性 2 のみで行う手法の形容詞類義語対ごとの正解率を 10 個の形容詞類義語対で比較した結果を表 5.12 に示す.

表 5.7: 形容詞類義語対の再現率の高さごとの割合

再現率の高さ	割合
高	20.0% (2/10)
中	30.0% (3/10)
低	50.0% (5/10)

表 5.8: 再現率の高さごとに分類した形容詞類義語対

再現率の高さ	再現率	動詞類義語対
再現率高	8 割以上	「易しい」と「手軽い」
		「近しい」と「むつまじい」
再現率中	6 割以上 7 割未満	「気高い」と「神々しい」
	5 割以上 6 割未満	「みっともない」と「見苦しい」
		「痛ましい」と「涙ぐましい」
再現率低	4 割以上 5 割未満	「注意深い」と「用心深い」
		「だるい」と「けだるい」
		「眠い」と「眠たい」
		「ずうずうしい」と「ずぶとい」
	3 割以上 4 割未満	「気まずい」と「面はゆい」

表 5.9: 提案手法とベースライン手法の形容詞類義語対ごとの正解率の平均

	再現率：高	再現率：中	再現率：低	すべての対
提案手法	0.93	0.76	0.80	0.81
ベースライン手法	0.59	0.64	0.78	0.70

表 5.10: 提案手法と素性 2 のみで行う手法の形容詞類義語対ごとの正解率の平均

	再現率
提案手法	0.81
素性 2 のみ	0.75

表 5.11: 提案手法とベースライン手法の形容詞類義語対ごとの正解率の比較結果

	再現率：高	再現率：中	再現率：低
提案手法○	2	3	3
ベースライン手法○	0	0	1
差なし	0	0	1

表 5.12: 提案手法と素性 2 のみで行う手法の形容詞類義語対ごとの正解率の比較結果

提案手法○	7
素性 2 のみ○	0
差なし	3

第6章　考察

本章ではまず，6.1 節で本研究の提案手法とベースライン手法の結果の比較について考察する．次に 6.2 節で今回実験を行った類義語対の中から，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．6.3 節では，これより得られた再現率の高さごとの傾向について考察する．6.4 節では人手による評価を行うとともにその結果について考察する．

6.1　提案手法とベースライン手法の比較

表 5.3 と表 5.9 のように， 提案手法とベースライン手法の正解率は，動詞では 0.88 と 0.77 で，形容詞では 0.81 と 0.70 であった． 提案手法はベースライン手法の正解率よりも高かった．また，表 5.5 と表 5.11 のように， 再現率高と中の類義語対は，動詞・形容詞ともに，ほとんどがベースライン手法よりも提案手法の正解率の方が高い結果であった.（一部差なし）これにより， 提案手法および機械学習で使用した素性は動詞・形容詞の類義語の判別に十分有用であることが言える．

しかし，再現率が低くなるほどベースライン手法と提案手法の差が小さくなり，ベースライン手法よりも提案手法のほうが低いものも見られた．原因として，類義語対の語の出現頻度に極端に差があることが考えられる．今回設定したベースライン手法は，類義語対のうち出現頻度の多い方の語を全て分類先とするものなので，出現頻度に極端に差があると再現率も極端に良くなる．このため，差が小さくなったと思われる．

6.2　類義語対ごとの考察

分類を行った再現率の高さごとに動詞類義語対，形容詞類義語対を 2 組ずつ例として挙げ，その類義語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類義語対ごとに 1 例ずつの計 4 例 (4 例無いものは 3 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． ここで示す素性のうち，「デフォルト素性」は常に利用されるデフォルトの素性であり， 他に情報がなければこの素性のみにより分類先が決定される．

6.2.1 再現率高の例「探し回る」と「探し求める」

（正解例 1）街中を探し回って、三日目、やっとイチゴジャムの缶詰を見つけることができた。

（正解例 2）長い間、探し求めていたものに巡り合えた』と直感しました」翌九一年、再来日して正式に入門した。

（誤り例 1）小説の処女作品集「信濃大名記」を探し求めているが、一度、古書目録で見つけて、注文したものの、手に入らなかった。

表 6.1: 機械学習の結果 (再現率高の例：「探し回る」と「探し求める」)

	再現率	適合率	総数
探し回る	1.00	0.97	87
探し求める	0.97	1.00	88

表 6.2: 機械学習が参考にした素性 (再現率高の例：「探し回る」と「探し求める」)

探し回る		探し求める	
素性	正規化 α 値	素性	正規化 α 値
素性 1:UNIGRAM	0.71	素性 1:を	0.60
素性 1:こと	0.55	素性 1:ため	0.55
素性 1:日	0.53	素性 1:長年	0.52

再現率高の例として,「探し回る」と「探し求める」という対がある． これらの語は両方，EDR 日本語単語辞書で概念識別子に 3cec34 が与えられており， EDR 概念辞書によるとこの識別子は「探し求める」を意味する．

表 6.2 の「日」のように短い時間を表す場合には「探し回る」,「長年」のように長い時間を表す場合には「探し求める」が使用される．また,「探し回る」は「探し回ること」のように使われやすく,「探し求める」は「探し求めるため」と使われることが多い傾向にある．これらより，この類義語対は使い分けを必要とする傾向にあると考えられる.「素性 1:UNIGRAM」とは，何も情報がない場合に使用されることが多いということである．

6.2.2 再現率高の例「せしめる」と「勝ち得る」

（正解例 1）便利さや豊かさだけを追求してきた生活が、現今の環境破壊を招来せしめたのである。

（正解例 2）その飾らない人柄が短い時間でチームメートの信頼を勝ち得た。

（誤り例 1）いうまでもなくXマスプレゼントや、お年玉でゲーム・ソフトをせしめた子どもたちが、そのゲーム攻略情報を仕入れるために買うのが、この手の本。

（誤り例 2）南アは昨年、デクラーク改革路線への評価により西欧各国から経済制裁解除を勝ち得たばかり。

表 6.3: 機械学習の結果 (再現率高の例：「せしめる」と「勝ち得る」)

	再現率	適合率	総数
せしめる	0.87	0.91	74
勝ち得る	0.89	0.84	55

表 6.4: 機械学習が参考にした素性 (再現率高の例：「せしめる」と「勝ち得る」)

せしめる		勝ち得る	
素性	正規化 α 値	素性	正規化 α 値
素性 1:UNIGRAM	0.57	素性 1:を	0.63
素性 1:修飾先が動詞	0.54	素性 1:信頼	0.53
素性 1:金	0.54	素性 1:修飾語の最後が名詞	0.53

再現率高の例として,「せしめる」と「勝ち得る」という対がある．これらの語は両方，EDR 日本語単語辞書で概念識別子に 3cfbcd が与えられており，EDR 概念辞書によるとこの識別子は「狙っていたものを手に入れる」を意味する．

表 6.4 の「修飾先が動詞」より，動詞を修飾する場合は「せしめる」が使われることが多いことがわかる．また,「金」,「信頼」とそれぞれにあることから，悪い印象の表現には「せしめる」，良い印象の表現には「勝ち得る」が使われることがわかる．これらより，この類義語対は使い分けを必要とする傾向にあると考えられる．

6.2.3 再現率中の例「見定める」と「突き止める」

（正解例 1）日本としてはソ連側の真意、出方を慎重に見定めながら、対処することになろう。

（正解例 2）船や飛行機などの位置を正確に突き止める方法として、米国防総省が開発した。

（誤り例 1）アラブ諸国やフランスなどの平和解決の努力の結果も見定めなければならない。

（誤り例 2）同研究所は、この技術で十数人の復顔に成功、五体の白骨死体の身元を突き止めている。

表 6.5: 機械学習の結果 (再現率中の例：「見定める」と「突き止める」)

	再現率	適合率	総数
見定める	0.71	0.83	207
突き止める	0.95	0.92	732

表 6.6: 機械学習が参考にした素性 (再現率中の例：「見定める」と「突き止める」)

見定める		突き止める	
素性	正規化 α 値	素性	正規化 α 値
素性 1:判断	0.62	素性 1:原因	0.69
素性 1:今後	0.62	素性 1:事件	0.63
素性 1:方向	0.61	素性 1:捜査	0.61

再現率中の例として,「見定める」と「突き止める」という対がある． これらの語は両方，EDR 日本語単語辞書で概念識別子に 3c4b79 が与えられており， EDR 概念辞書によるとこの識別子は「調べて明らかにする」を意味する．

表 6.6 の「判断」,「今後」,「方向」や正解例 1 から「見定める」は先の状況を想定し明らかにする場合に使われ,「原因」,「事件」,「捜査」や正解例 2 から「突き止める」は現在の状況を明らかにする場合に使われることがわかる．よって，この類義語対は使い分けを必要とする傾向にあると考えられる．

6.2.4 再現率中の例「ほったらかす」と「怠る」

（正解例 1）かつて監督２５年目で初出場を果たす「家族を ほったらか しにしたワシが……」と絶句した監督もいました。

（正解例 2）政策内容をきちんと地元に提示する、コンセンサスを得る努力を 怠って 原子力行政を進める道はない。

（誤り例 1）国民のことは ほったらかす、一派閥の主導権争いばかり。

（誤り例 2）実際、ＩＷＣは２０年以上にわたり、利益ばかり優先させ、クジラの保護管理は 怠ってきました。

表 6.7: 機械学習の結果 (再現率中の例：「ほったらかす」と「怠る」)

	再現率	適合率	総数
ほったらかす	0.68	0.97	69
怠る	0.99	0.98	1255

表 6.8: 機械学習が参考にした素性 (再現率中の例：「ほったらかす」と「怠る」)

ほったらかす		怠る	
素性	正規化 α 値	素性	正規化 α 値
素性 1:まま	0.57	素性 1:を	0.64
素性 1:家族	0.56	素性 1:責任	0.55
素性 1:仕事	0.55	素性 1:努力	0.52

再現率中の例として,「ほったらかす」と「怠る」という対がある． これらの語は両方，EDR 日本語単語辞書で概念識別子に 3ce673 が与えられており， EDR 概念辞書によるとこの識別子は「するべきことをしないで，ほうっておくさま」を意味する．

表 6.8 の「責任」「努力」や正解例 2 から「怠る」には「するべきことをしないで，ほうっておくさま」の他にも「仕事を怠る」や「精進を怠る」といった意味を持っていることがわかった．また,「ほったらかす」にはそれ以上の意味はなく,「ほったらかす」は「怠る」に言い換えれることが多かった．よって「怠る」と「ほったらかす」は広義と狭義の関係にあると言える．また，テスト文の数に大きな差があり,「怠る」に偏った素性が得られたため，今後の課題としてテスト文を同数にして，実験を行う必要があると考える．

6.2.5 再現率低の例「見限る」と「見捨てる」

（正解例 1）このためボスニアの独立は空洞化して名目だけとなる、事実上ボスニア政府の命運を 見限る ことにもつながる。

（正解例 2）このままでは自民党ばかりか今の政治が、国民から 見捨て られる

（誤り例 1）政治のアマチュアである市民は、政治を 見限 ったはいない

（誤り例 2）十万人が五分で 見捨てる 番組も、一万人が五十分思い入れる番組も平均視聴率は全く同じだ。

表 6.9: 機械学習の結果 (再現率低の例：「見限る」と「見捨てる」)

	再現率	適合率	総数
見限る	0.36	0.42	110
見捨てる	0.79	0.75	274

表 6.10: 機械学習が参考にした素性 (再現率低の例：「見限る」と「見捨てる」)

見限る		見捨てる	
素性	正規化 α 値	素性	正規化 α 値
素性 1:政権	0.63	素性 1:UNIGRAM	0.66
素性 1:れる	0.60	素性 1:られる	0.66
素性 1:政治	0.56	素性 1: 社会	0.58

再現率低の例として,「見限る」と「見捨てる」という対がある． これらの語は両方，EDR 日本語単語辞書で概念識別子に 10a1ec が与えられており， EDR 概念辞書によるとこの識別子は「相手を見捨てる」を意味する．

表 6.10 より「見捨てる」の方が「見限る」に比べて一般的であることがわかった．しかし，どちらとも社会に関する同じような素性が多く，人間が判断して使い分けることが難しいと感じた．これより，この類義語対は特別使い分けが必要でないといえる．

6.2.6 再現率低の例「はみ出す」と「はみ出る」

（正解例 1）ヘルメットから はみ出 したパーマネントの髪が印象的。

（正解例 2）首都のそばに住む「枠から はみ出 た難民」は、伝染病流行の恐れなどから政府にとって、苦々しい存在だ。

（誤り例 1）少ないスポーツ施設の争奪戦で、三ツ沢球技場ではこの冬さっそくラグビーが はみ出 した。

（誤り例 2）違反はサークルの線を踏んだり、はみ出 て投げること。

表 6.11: 機械学習の結果 (再現率低の例：「はみ出す」と「はみ出る」)

	再現率	適合率	総数
はみ出す	0.94	0.87	341
はみ出る	0.24	0.45	62

表 6.12: 機械学習が参考にした素性 (再現率低の例：「はみ出す」と「はみ出る」)

はみ出す		はみ出る	
素性	正規化 α 値	素性	正規化 α 値
素性 1:UNIGRAM	0.78	素性 1:米	0.62
素性 1:上	0.60	素性 1:手	0.60
素性 1:社会	0.58	素性 1: 自分	0.59

再現率低の例として,「はみ出す」と「はみ出る」という対がある． これらの語は両方，EDR 日本語単語辞書で概念識別子に 1042ab が与えられており，EDR 概念辞書によるとこの識別子は「ある制限や範囲から食み出す」を意味する．

表 6.12 より「はみ出す」の方が「はみ出る」に比べて一般的であることがわかった．しかし使い分けに役立つ情報は得られなく，例文を見ても「はみ出す」と「はみ出る」は互いに置き換えられ，人間が見ても判断することが難しい．これより，この類義語対は特別使い分けが必要でないといえる．

6.2.7　再現率高の例「易しい」と「手軽い」

（正解例 1）技術的裏付けのない素人が参加出来るほど 易しい 作品ではない。

（正解例 2）検索機能もついているなど、手軽 さに加えた機能性が現代ふうなのだ。

（誤り例 1）ＴＢＳ「関口宏のサンデーモーニング」は、政治や社会、経済、世界の流れを毎回 易しい 言葉で説明してくれる、魅力ある番組だ。

表 6.13: 機械学習の結果 (再現率高の例：「易しい」と「手軽い」)

	再現率	適合率	総数
易しい	0.98	1.00	156
手軽い	1.00	0.97	97

表 6.14: 機械学習が参考にした素性 (再現率高の例：「易しい」と「手軽い」)

易しい		手軽い	
素性	正規化 α 値	素性	正規化 α 値
素性 1:修飾先が名詞	0.62	素性 1:さ	0.82
素性 1:修飾先が動詞	0.54	素性 1:UNIGRAM	0.59
素性 1:問題	0.52	素性 1:人気	0.57

再現率高の例として,「易しい」と「手軽い」という対がある． これらの語は両方, EDR 日本語単語辞書で概念識別子に 3cfb92 が与えられており， EDR 概念辞書によるとこの識別子は「容易く簡単にできるさま」を意味する.

表 6.14 より「修飾先が名詞」や「修飾先が動詞」の場合「手軽い」より「易しい」が使われることが多い．また,「手軽い」には正解例 2 のように「さ」が付く場合が多く見られた．これらより，この類義語対は使い分けを必要とする傾向にあると考えられる.

6.2.8　再現率高の例「近しい」と「むつまじい」

（正解例 1）愚かな人間たちの営む社会として「江戸」が現代にぐっと 近しい ものに思われてくる。

（正解例 2）お二人の中国訪問に同行する、最も印象に残ったのは仲 むつまじい 姿だった。

（誤り例 1）８月２９日の党代表選の開票後、元代表はグループの会合で敗北を一応認めたが、近しい 同僚議員の見方は異なる。

（誤り例 2）大海原では別々に行動していた前年のつがいが半年ぶりに再会するわけで、うらやましいような むつまじ さです。

表 6.15: 機械学習の結果 (再現率高の例：「近しい」と「むつまじい」)

	再現率	適合率	総数
近しい	0.88	0.84	75
むつまじい	0.88	0.90	101

表 6.16: 機械学習が参考にした素性 (再現率高の例：「近しい」と「むつまじい」)

近しい		むつまじい	
素性	正規化 α 値	素性	正規化 α 値
素性 1:修飾先が名詞	0.57	素性 1:仲	0.68
素性 1:被修飾先が副詞	0.55	素性 1:よう	0.61
素性 1:関係	0.53	素性 1:姿	0.58

再現率高の例として,「近しい」と「むつまじい」という対がある．これらの語は両方，EDR 日本語単語辞書で概念識別子に 3cfc19 が与えられており，EDR 概念辞書によるとこの識別子は「関係が深いさま」を意味する．

表 6.16 より「修飾先が名詞」や「被修飾先が副詞」の場合「むつまじい」より「近しい」を使うことが多い．例えば正解例 1 の「ぐっと」や他には「とても」などが多く見られた．また「むつまじい」は「仲むつまじい」と使われていることが多く，素性を見ても明らかであった．これらより，この類義語対は使い分けを必要とする傾向にあると考えられる．

6.2.9 再現率中の例「気高い」と「神々しい」

（正解例 1）「三橋さんの 気高い 魂に、芸術家の本当の在り方を教えられた」と語った。

（正解例 2）神々しい 名前の由来は、四つの生薬が「四神」と呼ばれたため、という説がある。

（誤り例 1）どんな困難の中でも他者をいたわる心は、世界に誇るべし日本人の気高い精神です。

（誤り例 2）富士山は神々しい 存在で、心が洗われたという。

表 6.17: 機械学習の結果 (再現率中の例：「気高い」と「神々しい」)

	再現率	適合率	総数
気高い	0.72	0.71	118
神々しい	0.69	0.70	113

表 6.18: 機械学習が参考にした素性 (再現率中の例：「気高い」と「神々しい」)

気高い		神々しい	
素性	正規化 α 値	素性	正規化 α 値
素性 1:いさ	0.73	素性 1:神	0.74
素性 1:悲しみ	0.69	素性 1:さ	0.64
素性 1:さん	0.63	素性 1:姿	0.61

再現率中の例として,「気高い」と「神々しい」という対がある． これらの語は両方, EDR 日本語単語辞書で概念識別子に 3cefa8 が与えられており， EDR 概念辞書によるとこの識別子は「優れて気高いさま」を意味する.

表 6.18 と正解例 2 より「神々しい」を選択する際に,「神」といった素性に高く依存していることが分かった．しかし誤り例 1 を見てみると,「精神」といった形で「神」が出現すると素性に依存して「神々しい」が選択されてしまうため有用とは言えなかった．また「悲しみ」といった単語が出現すると，それと同時に「気高い」が使われている文が多かった.「さん」が素性に出現していることから一般的に人物に対しては「気高い」を使うことが多いことが分かった.

よって，この類義語対は使い分けを必要とする傾向にあると考えられる.

6.2.10 再現率中の例「痛ましい」と「涙ぐましい」

（正解例 1）このたびの少女の痛ましい 事件は極めて遺憾だ。

（正解例 2）先生は学生の私語をやめさせようと、涙ぐましい努力をしている。

（誤り例 1）まぎれもない人道危機の進行を、国際社会はただ痛まし く見守るしかない。

（誤り例 2）涙ぐましい減食を続けても、やめると体はちゃんと適正体重を心得えていて、元の体重に戻ります。

表 6.19: 機械学習の結果 (再現率中の例：「痛ましい」と「涙ぐましい」)

	再現率	適合率	総数
痛ましい	0.98	0.92	448
涙ぐましい	0.50	0.86	75

表 6.20: 機械学習が参考にした素性 (再現率中の例：「痛ましい」と「涙ぐましい」)

痛ましい		涙ぐましい	
素性	正規化 α 値	素性	正規化 α 値
素性 1:UNIGRAM	0.72	素性 1:努力	0.72
素性 1:事件	0.58	素性 1:若者	0.59
素性 1:戦争	0.57	素性 1:本	0.58

再現率中の例として,「痛ましい」と「涙ぐましい」という対がある． これらの語は両方，EDR 日本語単語辞書で概念識別子に 3ceba7 が与えられており， EDR 概念辞書によるとこの識別子は「あわれで，気の毒なさま」を意味する．

表 6.20 を見ると,「痛ましい」の素性に「事件」,「戦争」があり,「涙ぐましい」の素性に「努力」がある．これは本来「あわれで気の毒なさま」といった同じ意味でも,「痛ましい」には正解例 1 でわかるようにかわいそうの意味が含まれており,「涙ぐましい」には感動の意味が含まれているからである．また「涙ぐましい」努力とは表現するが,「痛ましい」努力とは表現しない．

よって，この類義語対は使い分けを必要とする傾向にあると考えられる．

6.2.11 再現率低の例「注意深い」と「用心深い」

（正解例 1）今後、国際、国家、市民、草の根レベルでそれぞれが注意深 く見守る必要があります。

（正解例 2）上西さんは近所づきあいは少なく、訪問してくる知人も室内に入れない、など用心深い性格だったという。

（誤り例 1）聴衆は注意深く聴き入る、村上さんが時折交えるジョークに笑い声を上げた。

（誤り例 2）用心深い性格のシャミル首相は座視の構えだ、世論も慎重だ。

表 6.21: 機械学習の結果 (再現率低の例：「注意深い」と「用心深い」)

	再現率	適合率	総数
注意深い	0.96	0.92	577
用心深い	0.47	0.70	90

表 6.22: 機械学習が参考にした素性 (再現率低の例：「注意深い」と「用心深い」)

注意深い		用心深い	
素性	正規化 α 値	素性	正規化 α 値
素性 1:UNIGRAM	0.60	素性 1:前	0.65
素性 1:米	0.57	素性 1:性格	0.61
素性 1:見守る	0.56	素性 1: 慎重	0.60

再現率低の例として,「注意深い」と「用心深い」という対がある． これらの語は両方，EDR 日本語単語辞書で概念識別子に 3cf20c が与えられており， EDR 概念辞書によるとこの識別子は「細かいところまで，念を入れて注意すること」を意味する．

表 6.22 とデータ数の差より「注意深い」が「用心深い」に比べて一般的であることが分かったが，正解例 2 や誤り例 2 から人間の性格を表す場合は「用心深い」を使用することが分かった．これより,「注意深い」と「用心深い」は場合によって使い分けが必要であることが分かった．また使い分けが必要であるにも関わらず，再現率低に分類されたのは「用心深い」のデータ数が「注意深い」のデータ数に比べて低く精度が落ちたことが考えられる．

6.2.12 再現率低の例「気まずい」と「面はゆい」

（正解例 1）気まずい思いを抱えて別れたが、時間の経過が父子の葛藤を和らげていった。

（正解例 2）これほどほめられると 面はゆい が、環境が変わってもやれるという自信が感じられ、好感の持てる受け答えだった。

（誤り例 1）中には悪態をつく子や 気まず そうにたばこの火を消す子もいます。

（誤り例 2）面はゆ さを感じながらも、褒められると、うれしいものである。

表 6.23: 機械学習の結果 (再現率低の例：「気まずい」と「面はゆい」)

	再現率	適合率	総数
気まずい	0.92	0.80	142
面はゆい	0.31	0.60	47

表 6.24: 機械学習が参考にした素性 (再現率低の例：「気まずい」と「面はゆい」)

気まずい		面はゆい	
素性	正規化 α 値	素性	正規化 α 値
素性 1:人	0.62	素性 1:さん	0.66
素性 1:後	0.58	素性 1:とき	0.59
素性 1:なる	0.56	素性 1: 娘	0.58

再現率低の例として,「気まずい」と「面はゆい」という対がある． これらの語は両方，EDR 日本語単語辞書で概念識別子に 3d06a1 が与えられており，EDR 概念辞書によるとこの識別子は「照れくさいさま」を意味する．表 6.24 を見ると,「人」,「さん」,「娘」などから,「気まずい」,「面はゆい」はどちらも人間関係を表す際に多く使われ,使い分けが難しい.

これより，この類義語対は特別使い分けが必要でないといえる.

6.3 再現率の高さごとの傾向と考察

再現率高とした類義語対は，先に示した考察例のように，文中に出現する単語によって使い分けが必要なものが多かった．例としては「探し回る」と「探し求める」がある．また，修飾・被修飾の係先によってや使い分けができたり，変化形によって使い分ける必要があるものが多かった．例としては「近しい」と「むつまじい」がある．再現率中とした類義語対は再現率高とした類義語対と同じ傾向が見られた．また,「ほった

らかす」と「怠る」では「ほったらかす」の再現率がかなり低くなってしまったことから2つの知見を得られた．1つは「怠る」と「ほったらかす」はEDR辞書内で「するべきことをしないで，ほうっておくさま」といった同じ意味を持つが，意味合いとして「怠る」の方が「ほったらかす」に比べて広義であることである．もう1つはそれぞれの出現頻度に大きな差があったことである．これにより，今後の課題としてデータ数を揃えて，実験を行う必要があると考える．再現率低とした類義語対は，前節で考察したように，再現率高や中に分類されたものに比べて使い分けの必要のない類義語対が確認できた．例としては「見限る」と「見捨てる」がある．また再現率低では，類義語対のうちどちらを使っても良いが，片方が一般的である場合が見られた．例としては「はみ出す」と「はみ出る」がある．

6.4 本章の問題点

表5.3と表5.9のように提案手法はベースライン手法に比べて高い精度が得られた．しかし，本研究で実験対象として使用した実験データには，類義語が基本形以外で出現するものも扱っている．基本形以外も扱うことは付属語によっての類義語の使い分けを判断することもできるが，一部の類義語対で実験が成立しない．その例を以下に示す．

時代と作品そのものの背景を、できる限り書き込んだ。 空欄には正解を、誤答は消しゴムで消して書き入れた。

これは動詞類義語対「書き込む」と「書き入れる」の例文である．この２文に出現している類義語部はどちらも基本形ではなく，過去形になっている.「書き込む」は過去形になる場合「書き込んだ」で語尾に「だ」がつくのに対して,「書き入れる」の過去形は「書き入れた」で語尾が「た」になる．本実験では本来前後の単語などの素性を用いて使い分けすることを目的としていた．しかし，この類義語対では「た」と「だ」を見るだけで使い分けができてしまうため問題であると言える．

第7章　追加実験

本章では，6.4 節の問題を考慮し，基本形のみで提案手法同様の実験を行った．実験に用いる類義語対の説明を 7.1 節に，実験結果を 7.2 節に示す．7.3 節と 7.4 節では，実験結果を元に考察を行う．また，7.5 節では，人手による評価を行う．

7.1　実験に用いる類義語対

本実験では実験データとして類義語が基本形で出現する文のみを扱うため，出現頻度の条件を同じにすると獲得できた類義語対の数が減少した．以下の表 7.1 は獲得できた動詞類義語対であり，表 7.2 は獲得できた形容詞類義語対である．数はそれぞれ 17 対と 9 対であった．

7.2　実験結果

7.2.1　動詞類義語対での実験結果

機械学習の再現率の高さごとに動詞類義語対を分類し，その割合を提案手法と比較したものを表 7.3 に示す．また，比較には同じ 17 対の類義語を使用し，再現率ごとの類義語対の分類には基本形のみを扱う手法を基準とする．再現率の高さごとの分類は，高を再現率 8 割以上，中を再現率 8 割未満 5 割以上，低を再現率 5 割未満である．また，分類には類義語対のうち，再現率の低い方を基準とする．機械学習の再現率の高さごとに 17 対の動詞類義語対を分類した結果を表 7.4 に示す．

提案手法と基本形のみで行う手法とベースライン手法の動詞類義語対ごとの正解率の平均を表 7.5 に示す． 機械学習の再現率の高さごとの値も示している．

提案手法と基本形のみで行う手法の動詞類義語対ごとの正解率を 17 個の動詞類義語対で比較した結果を表 7.6 に示す．表 7.6 における「差なし」とは，提案手法と基本形のみで行う手法の再現率の差が ± 0.01 以内であった動詞類義語対の数を示す.「提案手法○」は「差なし」以外でありかつ提案手法の正解率の方が高かった動詞類義語対の

表 7.1: 獲得した動詞類義語対

1	代わる	入れ替わる
2	準じる	準ずる
3	似る	類する
4	はみ出す	はみ出る
5	見限る	見捨てる
6	咲き誇る	咲き乱れる
7	投げ込む	投げ入れる
8	薄らぐ	薄れる
9	見定める	突き止める
10	さらけ出す	届け出る
11	取りやめる	取り消す
12	群がる	群れる
13	いら立つ	焦る
14	探し回る	探し求める
15	せしめる	勝ち得る
16	買い求める	買い入れる
17	書き込む	書き入れる

数を,「基本形のみ○」は「差なし」以外でありかつ基本形のみで行う手法の正解率の方が高かった動詞類義語対の数を示す．同様に，基本形のみで行う手法とベースライン手法の動詞類義語対ごとの正解率を 17 個の動詞類義語対で比較した結果を表 7.7 に示す.

表 7.3: 提案手法と基本形のみで行う手法の再現率の高さごとの割合比較結果

	再現率：高	再現率：中	再現率：低
提案手法	29.4% (5/17)	41.1% (7/17)	29.4% (5/17)
基本形のみ	41.1 %(7/17)	17.6% (3/17)	41.1% (7/17)

表 7.2: 獲得した形容詞類義語対

1	みっともない	見苦しい
2	ずうずうしい	ずぶとい
3	けだるい	だるい
4	痛ましい	涙ぐましい
5	気高い	神々しい
6	注意深い	用心深い
7	眠い	眠たい
8	近しい	むつまじい
9	気まずい	面はゆい

表 7.4: 再現率の高さごとに分類した動詞類義語対

再現率の高さ	再現率	動詞類義語対
再現率高	8 割以上	「探し回る」と「探し求める」
		「届け出る」と「さらけ出す」
		「書き込む」と「書き入れる」
		「投げ込む」と「投げ入れる」
		「買い求める」と「買い入れる」
		「勝ち得る」と「せしめる」
		「似る」と「類する」
再現率中	7 割以上 8 割未満	「代わる」と「入れ替わる」
	6 割以上 7 割未満	「準ずる」と「準じる」
	5 割以上 6 割未満	「突き止める」と「見定める」
再現率低	4 割以上 5 割未満	「咲き誇る」と「咲き乱れる」
	3 割以上 4 割未満	「焦る」と「いら立つ」
		「群がる」と「群れる」
	3 割未満	「見捨てる」と「見限る」
		「取り消す」と「とりやめる」
		「薄らぐ」と「薄れる」
		「はみ出す」と「はみ出る」

表 7.5: 提案手法と基本形のみで行う手法とベースライン手法の動詞類義語対ごとの正解率の平均

	再現率：高	再現率：中	再現率：低	すべての対
提案手法	0.94	0.88	0.80	0.87
基本形のみ	0.95	0.75	0.77	0.84
ベースライン手法	0.65	0.62	0.78	0.70

表 7.6: 提案手法と基本形のみで行う手法の動詞類義語対ごとの正解率の比較結果

	再現率：高	再現率：中	再現率：低
提案手法○	1	2	4
基本形のみ○	3	0	2
差なし	3	1	1

表 7.7: 基本形のみで行う手法とベースライン手法の動詞類義語対ごとの正解率の比較結果

	再現率：高	再現率：中	再現率：低
基本形のみ○	7	3	2
ベースライン○	0	0	3
差なし	0	0	2

7.2.2 形容詞類義語対での実験結果

機械学習の再現率の高さごとに形容詞類義語対を分類し，その割合を提案手法と比較したものを表 7.8 に示す．また，比較には同じ 9 対の類義語を使用し，再現率ごとの類義語対の分類には基本形のみを扱う手法を基準とする．再現率の高さごとの分類は，高を再現率 8 割以上，中を再現率 8 割未満 5 割以上，低を再現率 5 割未満である．また，分類には類義語対のうち，再現率の低い方を基準とする．

機械学習の再現率の高さごとに 9 対の形容詞類義語対を分類した結果を表 7.9 に示す．

提案手法と基本形のみで行う手法とベースライン手法の形容詞類義語対ごとの正解率の平均を表 7.10 に示す． 機械学習の再現率の高さごとの値も示している．

提案手法と基本形のみで行う手法の形容詞類義語対ごとの正解率を 9 個の形容詞類義語対で比較した結果を表 7.11 に示す．表 7.11 における「差なし」とは，提案手法と

基本形のみで行う手法の再現率の差が± 0.01 以内であった形容詞類義語対の数を示す.「提案手法○」は「差なし」以外でありかつ提案手法の正解率の方が高かった形容詞類義語対の数を,「基本形のみ○」は「差なし」以外でありかつ基本形のみで行う手法の正解率の方が高かった形容詞類義語対の数を示す．同様に，基本形のみで行う手法とベースライン手法の形容詞類義語対ごとの正解率を 17 個の形容詞類義語対で比較した結果を表 7.12 に示す.

表 7.8: 提案手法と基本形のみで行う手法の再現率の高さごとの割合比較結果

	再現率：高	再現率：中	再現率：低
提案手法	11.1% (1/9)	44.4% (4/9)	44.4% (4/9)
基本形のみ	11.1% (1/9)	33.3% (3/9)	55.5% (5/9)

表 7.9: 再現率の高さごとに分類した形容詞類義語対

再現率の高さ	再現率	動詞類義語対
再現率高	8 割以上	「近しい」と「むつまじい」
再現率中	7 割以上 8 割未満	「だるい」と「けだるい」
		「用心深い」と「注意深い」
	5 割以上 6 割未満	「気高い」と「神々しい」
		「痛ましい」と「涙ぐましい」
再現率低	4 割以上 5 割未満	「気まずい」と「面はゆい」
		「見苦しい」と「みっともない」
	3 割未満	「ずぶとい」と「ずうずうしい」
		「眠い」と「眠たい」

表 7.10: 提案手法と基本形のみで行う手法とベースライン手法の形容詞類義語対ごとの正解率の平均

	再現率：高	再現率：中	再現率：低	すべての対
提案手法	0.99	0.85	0.72	0.79
基本形のみ	0.89	0.82	0.66	0.76
ベースライン手法	0.54	0.69	0.68	0.67

表 7.11: 提案手法と基本形のみで行う手法の形容詞類義語対ごとの正解率の比較結果

	再現率：高	再現率：中	再現率：低
提案手法○	0	2	4
基本形のみ○	1	0	0
差なし	0	2	0

表 7.12: 基本形のみで行う手法とベースライン手法の形容詞類義語対ごとの正解率の比較結果

	再現率：高	再現率：中	再現率：低
基本形のみ○	1	4	1
ベースライン○	0	0	2
差なし	0	0	1

7.3 類義語対ごとの考察

分類を行った再現率の高さごとに動詞類義語対，形容詞類義語対を 1 組ずつ例として挙げ，その類義語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類義語対ごとに 1 例ずつの計 4 例 (4 例無いものは 3 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． ここで示す素性のうち，「デフォルト素性」は常に利用されるデフォルトの素性であり， 他に情報がなければこの素性のみにより分類先が決定される．

7.3.1 再現率高の例「書き込む」と「書き入れる」

（正解例 1）膨大なデータもパソコンなら、一枚のＣＤ・ＲＯＭ（読み出し専用メモリー）に書き込むことができるという。

（正解例 2）筆で一つ一つ模様や献灯者名を 書き入れる など細やかな手作業が続いている。

（誤り例 1）書き入れる 予定がぐっと少なくなり、白紙が目立つ手帳をながめ、「五十代になった時に何をやっているかだ」と言う。

表 7.13: 機械学習の結果 (再現率高の例：「書き込む」と「書き入れる」)

	再現率	適合率	総数
書き込む	1.00	0.99	507
書き入れる	0.93	1.00	29

表 7.14: 機械学習が参考にした素性 (再現率高の例：「書き込む」と「書き入れる」)

書き込む		書き入れる	
素性	正規化 α 値	素性	正規化 α 値
素性 1:UNIGRAM	0.83	素性 1:修飾先が名詞	0.57
素性 1:を	0.60	素性 1:筆	0.55
素性 1:欄	0.52	素性 1:予定	0.52

再現率高の例として,「書き込む」と「書き入れる」という対がある．これらの語は両方，EDR 日本語単語辞書で概念識別子に 3d06fc が与えられており，EDR 概念辞書によるとこの識別子は「本や台帳や紙切れに書き込みを入れる」を意味する．

表 7.14 の UNIGRAM と表 7.13 の総数の差より,「書き込む」は「書き入れる」に比べてかなり一般的であると言え，たいていの場合は「書き込む」を使う．また，正解例 1 のように CD などで記録を残す場合は「書き込む」としか表現しない．一方,「書き入れる」は「筆」の場合や「予定」といった書き加えるものが対象の場合に使われる．これらより，この類義語対は使い分けが必要だと言える．

7.3.2 再現率中の例「見定める」と「突き止める」

（正解例 1）今回の大会が、新聞の将来への誤りのない針路を 見定める 機会になることを希望する」と述べた。

（正解例 2）一刻も早く精密検査を受け、原因を 突き止める べきだ。

（誤り例 1）見えにくい中で、自分の目で争点を 見定める。

（誤り例 2）抗体ができる仕組みを解明するには、分子の正体を 突き止める 必要がある。

表 7.15: 機械学習の結果 (再現率中の例：「見定める」と「突き止める」)

	再現率	適合率	総数
見定める	0.56	0.63	103
突き止める	0.74	0.68	131

表 7.16: 機械学習が参考にした素性 (再現率中の例：「見定める」と「突き止める」)

見定める		突き止める	
素性	正規化 α 値	素性	正規化 α 値
素性 1:大切	0.59	素性 1:調査	0.58
素性 1:今後	0.54	素性 1:研究	0.58
素性 1:将来	0.53	素性 1:原因	0.58

再現率中の例として,「見定める」と「突き止める」という対がある．これらの語は両方，EDR 日本語単語辞書で概念識別子に 3c4b79 が与えられており， EDR 概念辞書によるとこの識別子は「調べて明らかにする」を意味する．

表 7.16 から分かったことは基本形以外も扱っている提案手法と変わりなかった．これより，提案手法と同様で，この類義語対は使い分けが必要だと言える．

7.3.3 再現率低の例「薄れる」と「薄らぐ」

（正解例 1）物資が出回ると、賞品の魅力も 薄れる 一方。

（正解例 2）大江健三郎さんのノーベル文学賞受賞を除けば、この年の記憶は 薄らぐ 一方ではないか。

（誤り例 1）議会の総与党化傾向が原因で、行政との緊張感は 薄れる 一方だ。

（誤り例 2）議席数で労働党と大差がついた結果、二大政党の連立の可能性も 薄らぐ ことになった。

表 7.17: 機械学習の結果 (再現率低の例：「薄れる」と「薄らぐ」)

	再現率	適合率	総数
薄れる	0.97	0.88	398
薄らぐ	0.06	0.21	51

表 7.18: 機械学習が参考にした素性 (再現率低の例：「薄れる」と「薄らぐ」)

薄れる		薄らぐ	
素性	正規化 α 値	素性	正規化 α 値
素性 1:UNIGRAM	0.67	素性 1:思い	0.64
素性 1:時間	0.62	素性 1:復興	0.62
素性 1:関係	0.56	素性 1:実感	0.61

再現率低の例として,「薄れる」と「薄らぐ」という対がある．これらの語は両方, EDR 日本語単語辞書で概念識別子に 3c300e が与えられており， EDR 概念辞書によるとこの識別子は「物事の程度が弱くなる」を意味する.

表 7.18 と表 7.17 より,「薄らぐ」に比べて「薄れる」の方が一般的であると言える. また例文や素性からも特に使い分けの特徴を得ることができなかった．これより，この類義語対は「薄れる」が一般的だが，特別に使い分けが必要ということではないといえる.

7.3.4 再現率高の例「近しい」と「むつまじい」

（正解例 1）人間関係が閉ざされ、深い孤独のなかにいた彼自身の世界と近しいものを感じたのであろう。

（正解例 2）お二人の中国訪問に同行し、最も印象に残ったのは仲むつまじい姿だった。

（誤り例 1）日本と韓国とは本来、近しい国。

（誤り例 2）一見、むつまじい男女の仲を表していそうな、この「我一君一念」の文字皿。

表 7.19: 機械学習の結果 (再現率高の例：「近しい」と「むつまじい」)

	再現率	適合率	総数
近しい	0.90	0.90	53
むつまじい	0.88	0.88	44

表 7.20: 機械学習が参考にした素性 (再現率高の例：「近しい」と「むつまじい」)

近しい		むつまじい	
素性	正規化 α 値	素性	正規化 α 値
素性 1:修飾先が名詞	0.83	素性 1:仲	0.65
素性 1:人	0.54	素性 1:よう	0.56
素性 1:関係	0.52	素性 1:姿	0.55

再現率高の例として,「近しい」と「むつまじい」という対がある．これらの語は両方，EDR 日本語単語辞書で概念識別子に 3cfc19 が与えられており， EDR 概念辞書によるとこの識別子は「関係に深いさま」を意味する．

表 7.20 から分かったことは基本形以外も扱っている提案手法と変わりなかった．これより，提案手法と同様で，この類義語対は使い分けが必要だと言える．

7.3.5 再現率中の例「だるい」と「けだるい」

（正解例 1）身体は だるい のに頭だけ冴え、寝つくことができなかった。

（正解例 2）けだるい ベンチの空気を一変させたのは薮の力投だった。

（誤り例 1）１時間に１本は吸わないと何をするにも だるい 気分になってしまう。

（誤り例 2）だが、全体を覆う暗い、けだるい ようなムードがなんとも言えない。

表 7.21: 機械学習の結果 (再現率中の例：「だるい」と「けだるい」)

	再現率	適合率	総数
だるい	0.96	0.91	77
けだるい	0.75	0.87	28

表 7.22: 機械学習が参考にした素性 (再現率中の例：「だるい」と「けだるい」)

だるい		けだるい	
素性	正規化 α 値	素性	正規化 α 値
素性 1:修飾先が動詞	0.55	素性 1:修飾先が名詞	0.60
素性 1:体	0.55	素性 1:よう	0.53
素性 1:症状	0.53	素性 1:ムード	0.52

再現率中の例として,「だるい」と「けだるい」という対がある．これらの語は両方, EDR 日本語単語辞書で概念識別子に 3cea9b が与えられており， EDR 概念辞書によるとこの識別子は「気分がなんとなくおっくうで疲れたさま」を意味する．

表 7.22 の「体」や正解例 1 から身体の疲れを表す場合には「だるい」が使われる．対して,「ムード」や正解例 2 から気分の疲れを表す場合には「けだるい」が使われる．また正解例 2 のように「けだるい」は「修飾先が名詞」になる傾向がある．これより, この類義語対は使い分けが必要だと言える．

7.3.6 再現率低の例「見苦しい」と「みっともない」

（正解例 1）社会党がこの数カ月にみせた七転八倒は、見苦しい ものではなかった。

（正解例 2）国際社会の結束が問われている時、元首相が行く行かないで日本が混乱したのは みっともない。

（誤り例 1）なんと 見苦しい 姿をさらしてしまったことか。

（誤り例 2）米国と欧州が最後まで世界中に迷惑をかけた交渉の内容は、せんじ詰めれば現金の取り合いでコメよりずっとエゴの塊で みっともない。

表 7.23: 機械学習の結果 (再現率低の例：「見苦しい」と「みっともない」)

	再現率	適合率	総数
見苦しい	0.45	0.50	115
みっともない	0.71	0.66	176

表 7.24: 機械学習が参考にした素性 (再現率低の例：「見苦しい」と「みっともない」)

見苦しい		みっともない	
素性	正規化 α 値	素性	正規化 α 値
素性 1:あまりに	0.59	素性 1:こと	0.61
素性 1:言葉	0.59	素性 1:交渉	0.56
素性 1:姿	0.58	素性 1:考え	0.55

再現率低の例として,「見苦しい」と「みっともない」という対がある．これらの語は両方，EDR 日本語単語辞書で概念識別子に 10a2c5 が与えられており， EDR 概念辞書によるとこの識別子は「見ていて不快な感じになるさま」を意味する．

表 7.18 を見ると， α 値に関わらず，どの素性も「見苦しい」,「みっともない」の両方で現れそうなことがわかる．また例文を見ても,「見苦しい」と「みっともない」は置き換えられ，人間が判断しても使い分けが難しいことがわかる．これより，この類義語対は特別使い分けの必要はないと言える．

7.4 提案手法との比較と考察

提案手法と類義語の出現形に基本形のみを扱う手法の再現率自体は動詞では 0.87(提案手法) と 0.84(基本形のみ) であり，形容詞では 0.79(提案手法) と 0.76(基本形のみ) と提案手法の方が少し再現率が高いものの，大きな差はなかった．また再現率の高さごとの考察でも使い分けに用いた素性に特徴の違いはほとんど見られなかった．しかし，例として，提案手法のみで考察した「易しい」と「手軽い」では,「手軽い」は「手軽さ」と表現されることが多く，使い分けの参考にした素性として変化形が大きく影響していた．そのため，基本形のみを扱う手法は使い分けを行うにおいて，提案手法とは違った有用な情報を取り出せると考える．そのほかの類義語対では，使い分けの参考にした素性に大きな違いは見られなかった．

7.5 人手評価

再現率高・中・低から動詞・形容詞類義語対それぞれ 2 例ずつ (形容詞の「高」のみ 1 例」) 類義語を選定し，人手評価を行った．人手評価には類義語を含む文をそれぞれ 10 文ずつランダムに抜き出し，どちらの類義語が正しいかを選び，正解率を求めた．

回答者は 3 名である．3 名のそれぞれの結果を表 7.25 から表 7.27 に示す．また，3 名の再現率の高さごとの正解率の平均を表 7.28 に示す．

表 7.25: 人手評価の結果 (被験者 A)

	再現率：高	再現率：中	再現率：低
動詞	探し回る・探し求める 書き込む・書き入れる	準じる・準ずる 見定める・突き止める	見限る・見捨てる はみ出す・はみ出る
正解率	0.85	0.80	0.50
形容詞	近しい・むつまじい	だるい・けだるい 痛ましい・涙ぐましい	気まずい・面はゆい みっともない・見苦しい
正解率	0.90	0.80	0.70

表 7.26: 人手評価の結果 (被験者 B)

	再現率：高	再現率：中	再現率：低
動詞	探し回る・探し求める 書き込む・書き入れる	準じる・準ずる 見定める・突き止める	見限る・見捨てる はみ出す・はみ出る
正解率	0.75	0.80	0.60
形容詞	近しい・むつまじい	だるい・けだるい 痛ましい・涙ぐましい	気まずい・面はゆい みっともない・見苦しい
正解率	0.90	1.00	0.70

表 7.27: 人手評価の結果 (被験者 C)

	再現率：高	再現率：中	再現率：低
動詞	探し回る・探し求める 書き込む・書き入れる	準じる・準ずる 見定める・突き止める	見限る・見捨てる はみ出す・はみ出る
正解率	0.65	0.90	0.45
形容詞	近しい・むつまじい	だるい・けだるい 痛ましい・涙ぐましい	気まずい・面はゆい みっともない・見苦しい
正解率	1.00	0.95	0.50

表 7.28: 再現率の高さごとの正解率の平均

	再現率：高	再現率：中	再現率：低
動詞	探し回る・探し求める 書き込む・書き入れる	準じる・準ずる 見定める・突き止める	見限る・見捨てる はみ出す・はみ出る
正解率	0.75	0.83	0.51
形容詞	近しい・むつまじい	だるい・けだるい 痛ましい・涙ぐましい	気まずい・面はゆい みっともない・見苦しい
正解率	0.93	0.91	0.63

表 7.28 を見ると，正解率の平均は高・中・低がそれぞれ，動詞では (0.75・0.83・0.45) となり，形容詞では (0.93・0.91・0.63) となっていることが分かる．人手評価では提案手法とは違い，動詞・形容詞とも「再現率中」の類義語対の正解率がかなり高くなっていた．表 7.28 より，特に動詞ではその傾向にあることがわかる．これは「見定める」と「突き止める」の正解率が高くなっていることから，おそらく概念識別子が 1 つのみの類義語対に絞らなっかたことが原因として挙げられる．例えば「見定める」では今後の状況を明らかにする場合に,「突き止める」は現在の状況を明らかにする場合に用いる．このことが，人手による判断を容易にさせたと考える．また，形容詞の「だるい」と「けだるい」なども同様である．それ以外にはそれほど提案手法の結果と比べて差異がなかったことより，今後は概念識別子にも着目する必要があると考える．

第8章　おわりに

本研究では機械学習を用いて動詞・形容詞の類義語対の使い分けを行った．

第1の成果として，動詞22対，形容詞10対の類義語対を用いた実験において，類義語のうち最も頻度の高い語を常に選択するベースライン手法の正解率が動詞では0.77，形容詞では0.70であるのに対して，機械学習を用いる提案手法は動詞では0.88，形容詞では0.81の正解率であった．これにより，今回提案した手法自体が動詞・形容詞の類義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，機械学習での性能に基づき動詞・形容詞の類義語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．特に使い分けが必要とされた類義語対に「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．また，いくつかの動詞・形容詞の類義語対について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することができた．例えば「探し回る」と「探し求める」という類義語対では,「探し回る」は「時間」「日」といった短い時間を表す場合に用いられることが多いが,「長年」「旅」といった長い時間を表す場合には「探し求める」が用いられる．また,「あちこち探し回る」とは表現するが,「あちこち探し求める」とは普通表現しない．

使い分けが必要とされた類義語対は文中に出現する名詞によって使い分けの判断ができるものが多く，使い分けが必要でない類義語対は互いに置き換えられるものや，片方が一般的であることが多かった．

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 である強田吉紀さんに協力をいただきました．また，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赤江涼太君を初めとする計算機工学講座 C 研究室の皆様に感謝の意を表します．

参考文献

[1] 王玉馨, 清水伸幸, 吉田稔. 単語類似度ネットワークを通じた自動同義語獲得. pp. 7–14, 2008.

[2] 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.

[3] 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.

[4] 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, pp. 584–587, 2013.

[5] 中瀬光暁. 教師あり機械学習を用いた副詞の類義語の使い分け. 鳥取大学工学部卒業論文, 2015.

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

[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.