# 概要

本研究は同義語の使い分けを教師あり機械学習を使用して行う．同義語は使い分けの必要がないと思いがちだが，使い分けの必要な場合がある．例えば「衣類」と「衣料」という同義語対では,「衣料品」ということはあるが「衣類品」ということはない．ある同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，その同義語対は特に使い分けの必要な同義語対とわかり，機械学習での性能が低かった同義語対は，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，同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同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，この手法自体が同義語の使い分けに対して有用であることが挙げられる．2つ目は，同義語対ごとの機械学習の性能に基づき，同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同義語対に「貯金」と「貯蓄」などがあった．また，いくつかの同義語対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．

# 目 次

# 表 目 次

# 図 目 次

# 第1章　はじめに

同義語とは，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「即刻」と「即時」などがある．同義語に関する研究では，コーパスから同義語を獲得する研究 [1][2] や西尾の人間の会話における同義語の使用傾向を調査し分析する研究 [1] 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 [2]．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，同義語全般を対象とはしていない．そこで，本研究はそこに着目し，機械学習を用いて同義語全般の使い分けを行う．本研究の成果は，文章を生成する際の同義語の選択，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同義語を利用する．

同義語は意味がほぼ同じであり，一見同義語は使い分けが必要ないと思いがちだが，実は使い分けが必要な場合がある．例えば,「衣類」と「衣料」は EDR 電子化辞書では「体に着るもの」という意味で同義語とされているが，後ろに「品」をつけることができるのは「衣料」の方のみであり，後ろに「品」をつける場合は使い分けが必要となる．

本研究では，機械学習による性能の高い同義語の使い分けも目指すが，同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．機械学習によって同義語を推定しやすい場合は，同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同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，同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．

- 同義語に関する研究では，同義語の使い分けに機械学習を用いた研究はない．本論文は同義語の使い分けのために機械学習を使用し，複数の同義語対について，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．

- 機械学習に基づき同義語の使い分けを行った．45 個の同義語対を用いた実験にお

いて，同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 0.70 であるのに対して，機械学習を用いる提案手法は 0.86 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．

- 機械学習での性能に基づき同義語対を使い分けが特に必要なものとそれほど必要でないものに分類した．

- 機械学習における素性 (学習に用いる情報のこと) を分析することで同義語の使い分けに重要な情報を把握することができる．いくつかの同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．

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，本研究で使用する同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，具体的な同義語対の考察も行い，どのような情報が同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，小島らが行った表記選択の研究について記述し，2.2 節では，西尾が行った同義語に対するアンケート調査について記述する．2.3 節では，王らが行った同義語の自動獲得法について記述する．

## 2.1　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 [2]．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．

## 2.2 同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同義語と，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同義語があるとし，今回は主に後者のような同義語についての選択を調査している [1]．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．

調査した同義語対は，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: 同義語対の分類

| 分類 | 性質 | 例 |
|---|---|---|
| A | 外来語を一方にもつ同義語対 | デパートと百貨店 |
| B | 旧式語を一方にもつ同義語対 | 婚礼と結婚式 |
| C | 日常語と文章語の同義語対 | 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，選択の差があった同義語対としては「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同義語対としては「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

た同義語対としては，それほど大きな差がみられた同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，同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，機械学習により同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．

## 2.3 単語類似度ネットワークを通じた自動同義語獲得

王らは，コーパスから同義語を獲得する方法について新たな手法を提案している [3]．コーパスから同義語を獲得することは，テキスト処理の重要なリソースの 1 つであるシソーラスの品質を向上させることに繋がる．

王らが提案したコーパスから同義語を獲得する手法は単語類似度ネットワークを使用した方法である．同義語を獲得する方法には様々な方法が提案されており，それらは，同義語名詞なら似ている文脈情報をもつという分布仮説に基づいている．この仮説は基本的に 2 段階の処理で実行される．第 1 段階の処理では，コーパスから抽出された重要度の高い単語の文脈特徴における統計情報を抽出する．第 2 段階の処理では cosine 類似度などの類似性量度を選び，それをクエリ単語と同義語候補の単語対に適用して類似度を計算する．類似度の降順で各クエリ単語の同義語候補リストを作る．最後に同義語リストからトップ候補を選んで，クエリ単語の同義語と認定する．これまでが基本的な方法である．

王らの手法では 2 段階の処理の後にさらにもう 1 つ処理を加える．2 段階の類似度によって形成されるネットワーク，すなわち単語をノード，類似度の順位が閾値以内の単語間にアークを持つとしてネットワークの構造を調べてみると，スケールフリーの性質を持っていることがわかった．これにより，クエリ単語の同義語である可能性が比較的高い単語だけを対象にする自動同義語候補選択のためのランク閾値を決める手法 RTS(Rank Threshold for synonym candidate Selection method) と単語類似度ネットワークの構造を活用する同義語候補の相互リランキング法 MRM(Mutual Re-ranking Method) を提案した．MRM でリランキングする際，スケールフリーネットワークにある類似度の降順でランクされた同義語候補はハブ単語と非ハブ単語をわけて扱う．同義語関係は対称だが，王らの MRM は対称ではない．以前の研究では単語類似度ネットワークの構造的な特性を使用していない．

王らの研究の成果として，提案したRTSで選択された単語ノードの単語類似度ネットワークがスケールフリーの特性をもっていることを示したということがある．更に，同義語候補リストを改良するためMRMを提案した．詳細な実験により，RTSで自動的に選択したランクの閾値が有効であり，MRMを加えてさらに有効性を示した．

この先行研究はコーパスから同義語を獲得する研究であり，同義語に関する研究であるということは本研究と同一であるが，それ以外は大きく異なっている．

# 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 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

使い分けをしたい同義語対 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，正しく同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,「衣料」と「衣類」の例を以下に挙げる．

| |
|---|
| 衣料 の防虫科学研究では国内の第一人者。<br>リュックサックに寝袋、衣類 を詰めての貧乏旅行だ。 |

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 の防虫科学研究では国内の第一人者。<br>リュックサックに寝袋、X を詰めての貧乏旅行だ。 |

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．

機械学習により同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同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．同義語対の語 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 $f_j(1 \leq j \leq k)$ 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 [4, 5, 6, 7].

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

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

ただし，$A, B$ は分類と文脈の集合を意味し，$g_j(a,b)$ は 文脈 $b$ に素性 $f_j$ 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，$\tilde{p}(a,b)$ 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 [2] 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．同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: 同義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 | 対象語の同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　実験に用いる同義語対

本章では実験に用いる同義語対の説明を行う．4.1 節で同義語の認識方法を説明する．4.2 節で実験に用いる同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同義語の認識

本研究では，2 つの語が同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同義語対と判定する．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 |
|---|---|
| 衣料 | 0e504a |
| 衣類 | 0e504a |
| 赤 | 0e29cb 1f8697 1f8698 1f8699 1f869b 1f869c 1f869d 1f86a0 1f86a3 |
| 青 | 0f91f3 1f6678 3bdae1 3c2c39 3c2c3a 3c2c49 3cfb7a |
| ランチ | 3bec74 3c0457 3c0458 3c58fd |
| 昼食 | 3bec74 |

表 4.2: 概念識別子と概念の対応例

| 識別子 | 概念 |
|---|---|
| 0e504a | 体に着るもの |
| 0e29cb | 赤の色をしているさま |
| 1f869c | 収支が赤字であること |
| 3bec74 | 昼の食事 |
| 3c0457 | 港内の人員輸送に用いる大型モーターボート |
| 3c58fd | 洋風の定食 |

## 4.2 実験で用いる同義語対の選定

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同義語の使い分けを行う．新聞記事には，1991 年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，実験に用いる同義語対とする．

**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，1 年分の新聞で出現頻度が 50 回以上であること

**条件 4** 形態素解析システム JUMAN[8] 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同義であるとみなせるため設定する．条件 2 は，多義語の場合は言語現象が複雑になると考え，扱わないようにするために設定する．例えば「ランチ」と「昼食」は同一の識別子 3bec74 をもつが，EDR 辞書によると「ランチ」は複数の識別子をもち,「昼食」とは違った意味 (識別子) をもつ場合がある．この違った意味で文章に記述されていた場合,「昼食」と同義であるとは言えないため，多義性のある語は省く必要がある．条件 3 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 4 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 [2] で扱われており，本研究では扱わないため条件 4 を設けた．

これらの条件を満足する同義語対は 90 対あり，その中からランダムに取り出した 45 対を実験に用いる同義語対とする．

# 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

獲得した同義語対 45 対について，同義語対ごとに同義語の使い分けの実験を行う．入力文は，1991 年の毎日新聞から獲得した，同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．

同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．

## 5.2　実験結果

機械学習の再現率の高さごとに 45 個の同義語対を分類した結果を表 5.1 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．提案手法とベースライン手法の同義語対ごとの正解率の平均を表 5.2 に示す．機械学習の再現率の高さごとの値も示している．提案手法とベースライン手法の同義語対ごとの正解率を 45 個の同義語対で比較した結果を表 5.3 に示す．表 5.3 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 $\pm$ 0.01 以内であった同義語対の数を示す.「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同義語対の数を,「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同義語対の数を示す．

表 5.1: 再現率の高さごとに分類した結果

| 再現率の高さ | 割合 |
|---|---|
| 高 | 0.11 % ( 5/45) |
| 中 | 0.51 % (23/45) |
| 低 | 0.38 % (16/45) |

表 5.2: 提案手法とベースライン手法の同義語対ごとの正解率の平均

| | 再現率：高 | 再現率：中 | 再現率：低 | 全ての対 |
|---|---|---|---|---|
| 提案手法 | 0.95 | 0.84 | 0.80 | 0.86 |
| ベースライン手法 | 0.62 | 0.69 | 0.78 | 0.70 |

表 5.3: 提案手法とベースライン手法の同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 | 23 | 11 |
| ベースライン手法○ | 0 | 0 | 2 |
| 差なし | 0 | 0 | 4 |

# 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同義語対の中から，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．

## 6.1　提案手法とベースライン手法の比較

表 5.2 のように，提案手法とベースライン手法の正解率は，0.86 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 5.3 のように，再現率高と中の同義語対全てと低の同義語対 11 組では，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．これは実験に使用した同義語対 45 組のうちの 8 割以上である．これにより，提案手法および機械学習で使用した素性は同義語の判別に十分有用であることが言える．

しかし，再現率低においては，差なしが 4 組，ベースライン手法の再現率の方が高い同義語対が 2 組あった．原因として，同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，同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，ベースライン手法の方が良い場合があったものと思われる．

## 6.2　同義語対ごとの考察

分類を行った再現率の高さごとに同義語対を 2 組ずつ例として挙げ，その同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同義語対ごとに 1 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．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

**「貯金」と「貯蓄」**

**(正解例 1)** また、大口定期預金の預け入れ最低限度引き下げとともに郵便局の 貯金 預け入れ上限額は現行の七百万円から一千万円に引き上げられる。

**(正解例 2)** これは投資が弱いためだが、この原因は家計部門の 貯蓄 が低く、投資資金を国内で調達するのに限界があるからだ。

**(誤り例 1)** 今冬のボーナスの使い道はまだはっきりは決めていませんが、半分を 貯金、残りの半分を旅行費用に充てようと考えています。

**(誤り例 2)** 使い道のトップはやはり 貯蓄 で八二・七%。

表 6.1: 機械学習の結果 (再現率高の例：「貯金」と「貯蓄」)

| | 再現率 | 適合率 | 総数 |
|---|---|---|---|
| 貯金 | 0.89 | 0.89 | 403 |
| 貯蓄 | 0.86 | 0.86 | 332 |

表 6.2: 機械学習が参考にした素性 (再現率高の例：「貯金」と「貯蓄」)

| 貯金 | | 貯蓄 | |
|---|---|---|---|
| 素性 | 正規化 α 値 | 素性 | 正規化 α 値 |
| 素性 1:定額 | 0.74 | 素性 1:的 | 0.75 |
| 素性 1:郵便 | 0.73 | 素性 1:投資 | 0.70 |
| 素性 1:郵政省 | 0.73 | 素性 1:米国 | 0.64 |

再現率高の例として,「貯金」と「貯蓄」という対がある．これらの語は，EDR 日本語単語辞書で概念識別子に 0fbec6 のみが与えられており，EDR 概念辞書によるとこの識別子は「貯金」を意味する．

表 6.4 の「素性 1:郵便」や「素性 1:郵政省」からもわかるように，郵便貯金に関する文章では「貯金」が使用されている．また，正解例 1 でも「郵便局」という語が使用されており，この傾向が現れている.「貯蓄」を含む文章では，正解例 2 や表 6.2 の「素性 1:米国」などから，国の財政状況などを述べた文章の場合は「貯蓄」を使用する傾向にあると推測できる．また，今回機械学習では正しく識別できなかった誤り例として記載しているが，誤り例 1 のように，話し言葉の中では「貯金」が使用される傾向にあった．これらより，この同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．

**「メダル」と「賞牌」**

**(正解例 1)** 大会では団体戦のメンバーとして メダル を目指し、シングルスで世界ランク入りを狙う。

**(正解例 2)** 出版文化賞の著者に賞状と賞金５０万円、特別賞の著・編者には賞状、また両賞の出版社には賞状、賞牌 が贈られます。

**(誤り例 1)** 作品は小・中学生の部と一般の部に分け、最優秀賞各一点（賞状、メダル、盾、副賞）と優秀賞各四点（賞状、メダル、副賞）を選び、フェアのオープニングセレモニーで発表する。

**(誤り例 2)** 賞金三十万円と賞状、賞牌 が贈られる。

表 6.3: 機械学習の結果 (再現率高の例：「メダル」と「賞牌」)

| | 再現率 | 適合率 | 総数 |
|---|---|---|---|
| メダル | 0.97 | 0.97 | 98 |
| 賞牌 | 0.88 | 0.88 | 26 |

再現率高の例として,「メダル」と「賞牌」という対がある．これらの語は，EDR 日本語単語辞書で概念識別子に 103f04 のみが与えられており，EDR 概念辞書によるとこの識別子は「業績をほめたたえ記念として与えるメダル」を意味する．

表 6.4: 機械学習が参考にした素性 (再現率高の例：「メダル」と「賞牌」)

| メダル | | 賞牌 | |
|---|---|---|---|
| 素性 | 正規化 α 値 | 素性 | 正規化 α 値 |
| デフォルト素性 | 0.62 | 素性 1:円 | 0.70 |
| 素性 5:助詞 | 0.62 | 素性 1:賞金 | 0.63 |
| 素性 4:、 | 0.55 | 素性 1:副賞 | 0.57 |

表 6.4 を見てみると，機械学習が賞牌と推定するのに利用した素性には,「素性 1:円」や「素性 1:賞金」など賞金に関する素性が見られる．また正解例 2 や誤り例 2 からも確認できるように，何らかの大会などの結果，賞金などが授与されたことを示す文章では「賞牌」を使用するという使い分けがあることがわかる．また，機械学習がメダルと推定するのに参考にした素性に「デフォルト素性」が確認できる．これより，特徴のない文の場合は「賞牌」よりも「メダル」が使用される傾向にあるとわかる．

**「婦女」と「女流」**

### 6.2.2 再現率中の例

**「衣料」と「衣類」**

**(正解例 1)** 売上高の内訳では、紳士・婦人服を中心とする主力の 衣料 品は、各社とも、比較的高い伸び。

**(正解例 2)** 大阪府警泉大津署の調べでは、倉庫内には電化製品や 衣類、毛布などが大量にあり、ほとんどが焼けた。

**(誤り例 1)** 寄せられた募金は医薬品、テント、毛布、衣料、食料などの購入資金に充て、トルコ、イランの難民キャンプに送る。

**(誤り例 2)** それでも商店には食料も 衣類 もなく、長い行列の生活が続き、共働きの主婦にはあらゆる負担がのしかかり、「もう耐えられない」と言う。

再現率中の例として,「衣料」と「衣類」という対がある．この対の語は，EDR 日本語単語辞書で概念識別子に 0e504a のみが与えられており，EDR 概念辞書によるとこの識別子は「体に着るもの」を意味する．

表 6.5: 機械学習の結果 (再現率中の例：「衣料」と「衣類」)

| | 再現率 | 適合率 | 総数 |
|---|---|---|---|
| 衣類 | 0.63 | 0.63 | 106 |
| 衣料 | 0.75 | 0.75 | 160 |

表 6.6: 機械学習が参考にした素性 (再現率中の例：「衣料」と「衣類」)

| 衣料 | | 衣類 | |
|---|---|---|---|
| 素性 | 正規化 α 値 | 素性 | 正規化 α 値 |
| 素性 1:品 | 0.78 | 素性 2:対象語が文頭である | 0.70 |
| 素性 3:aa | 0.66 | 素性 1:電気 | 0.62 |
| 素性 2:品 | 0.62 | 素性 1:物品 | 0.61 |

正解例 1 や表 6.6 から，直後に「品」がある場合「衣料」と書く使い分けが存在することがわかる．すなわち,「衣類品」という表現は一般に使わず「衣料品」という表現が使われる使い分けがある．しかし，その他に目立った特徴はなく，正解例 2 や誤り例 2 よりわかるように，名詞を並列して記述する場合でも，どちらを使用するべきだとは断定しがたい．これらより，この同義語対は，使用方法によっては使い分けを必要とするが，通常使い分けの必要がないと推測できる．

**「評論」と「論評」**

**(正解例 1)** 「保守道政の奪還」を掲げる自民は、昨年暮れ、元東大助教授で政治 評論 家の舛添要一氏（４２）に出馬を要請。

**(正解例 2)** 米政府当局は、この報道についての 論評 を控えているが、米政府関係者は、爆撃の結果フセイン大統領自身に被害を与えることもありうるとしている。

**(誤り例 1)** 戦後、長いこと続いた「冷戦」という言葉で呼ばれる東西対立の終えん、米ソ二極体制の崩壊は、人によっては、これを「民主主義の勝利」「市場経済原理の勝利」として世界はいよいよ希望に満ちた時代になるのかなということが、昨年来、いろいろ 評論 をされたわけでございます。

**(誤り例 2)** その多岐にわたる文化現象を 論評 したのが本書である。

表 6.7: 機械学習の結果 (再現率中の例：「評論」と「論評」)

| | 再現率 | 適合率 | 総数 |
|---|---|---|---|
| 評論 | 0.98 | 0.96 | 797 |
| 論評 | 0.78 | 0.91 | 141 |

表 6.8: 機械学習が参考にした素性 (再現率中の例：「評論」と「論評」)

| 評論 | | 論評 | |
|---|---|---|---|
| 素性 | 正規化 α 値 | 素性 | 正規化 α 値 |
| 素性 2:家 | 0.61 | 素性 1:大統領 | 0.62 |
| 素性 1:小説 | 0.61 | 素性 1:韓国 | 0.66 |
| 素性 1:文学 | 0.60 | 素性 1:政府 | 0.58 |

再現率中の例として,「評論」と「論評」という対がある．この対の語は，EDR 日本語単語辞書で概念識別子に 10ed0c のみが与えられており，EDR 概念辞書によるとこの識別子は「論評した文章」を意味する．

まず，表 6.8 から，機械学習が「評論」と推定する際に使用した素性として,「素性 2:家」がある．家の直前に使用するのは「論評」ではなく「評論」であり,「論評家」とは言わず「評論家」と言う使い分けがある．また,「評論」と推定する際に参考とした素性では，文学や小説など，文学に関連する語が確認できた.「論評」と推定した際に使用した素性としては,「大統領」や「政府」など，各国の情勢を表している文章であると推定できる語が並んでいる．これより「論評」は，国の情勢を記述した文章で使用される傾向にあると考えられる．しかしこれらの文章内でも,「評論家」という語が使用されている可能性があるため，その際は注意しなければならない．

### 6.2.3 再現率低の例

**「上期」と「上半期」**

**(正解例 1)** 八八年までは毎年十件に満たなかったが、八九年は二十四件に急増、九〇年は十九件、今年は 上半期 だけで十一件にのぼる。

**(正解例 2)** 湾岸ショックの逆風が弱まり、今年度 上期 をボトムに下期から上向くとみているからだ。

**(誤り例 1)** 大手証券四社の九一年度 上半期（四 九月）の債券売買益で、大和証券が業界トップの野村証券を抜き、半期ベースながら初めて首位に立ったことが十一日、明らかになった。

**(誤り例 2)** 都銀の年度 上期 中の預金残高の減少は、全銀協が調査を開始した一九五四年以来初めて。

表 6.9: 機械学習の結果 (再現率低の例：「上期」と「上半期」)

| | 再現率 | 適合率 | 総数 |
|---|---|---|---|
| 上期 | 0.45 | 0.56 | 60 |
| 上半期 | 0.83 | 0.75 | 124 |

表 6.10: 機械学習が参考にした素性 (再現率低の例：「上期」と「上半期」)

| 上期 | | 上半期 | |
|---|---|---|---|
| 素性 | 正規化 α 値 | 素性 | 正規化 α 値 |
| 素性 1:下期 | 0.77 | 素性 1:市場 | 0.66 |
| 素性 1:決算 | 0.70 | 素性 1:生産 | 0.62 |
| 素性 1:調査 | 0.64 | 素性 1:今年 | 0.62 |

再現率低の例として,「上期」と「上半期」という対がある．この対の語は，EDR 日本語単語辞書で概念識別子に 0ea538 のみが与えられており，EDR 概念辞書によるとこの識別子は「会計年度などの 1 年を 2 期にわけた前半の 6ヶ月」を意味する．

表 6.10 の機械学習が参考にした素性を見ると，お互いの素性に経済や金融と関連のある「決算」や「市場」が素性として現れており，どちらも経済や金融のことについて記述された文章で使用されることがわかる．例文を見ても，経済や金融に関する記事がかなり多く，人間による判定も難しいと考えられる．唯一判定の基準になるものとして,「上期」を含む文中には「下期」が対となって記述されている文章があるということがわかり,「下期」が文中にない場合，どちらを使用してもよいと推測できる．これにより，この同義語対は特定の場合を除き，使い分けが必要でないと言える．

**「うたい文句」と「キャッチフレーズ**

**(正解例 1)** 日米関係再構築が うたい文句 だが、実際は経済問題で米国の要求を突き付けられる場面が予想される。

**(正解例 2)** それならばと「環境にやさしい」を キャッチフレーズ に、研がずに炊ける「無洗米」が登場し始めた。

**(誤り例 1)** 公平な税制という うたい文句 は、どこに行ったのだろう。

**(誤り例 2)** 今春、東京都大田区にできた中高齢者専用マンション「シニアハウス久が原」は「(お年寄りの) 自立ある生活環境」が キャッチフレーズ だ。

表 6.11: 機械学習の結果 (再現率低の例 :「うたい文句」と「キャッチフレーズ」)

| | 再現率 | 適合率 | 総数 |
|---|---|---|---|
| うたい文句 | 0.29 | 0.50 | 56 |
| キャッチフレーズ | 0.90 | 0.78 | 160 |

表 6.12: 機械学習が参考にした素性 (再現率低の例:「うたい文句」と「キャッチフレーズ」)

| うたい文句 | | キャッチフレーズ | |
|---|---|---|---|
| 素性 | 正規化 α 値 | 素性 | 正規化 α 値 |
| 素性 1:日 | 0.64 | デフォルト素性 | 0.69 |
| 素性 1:の | 0.61 | 素性 2:」 | 0.62 |
| 素性 1:建設 | 0.59 | 素性 1:者 | 0.61 |

再現率低の例として,「うたい文句」と「キャッチフレーズ」という対がある．この対の語は，EDR 日本語単語辞書で概念識別子に 0ec55c のみが与えられており，EDR 概念辞書によるとこの識別子は「短かい宣伝文句」を意味する．

表 6.12 から，機械学習が「キャッチフレーズ」と推定するのに参考にした素性として「デフォルト素性や『素性 2:」』などがある．まずデフォルト素性が「キャッチフレーズ」と推定するのに有効であると判断された理由としては，一般に「うたい文句」よりも「キャッチフレーズ」がよく使用されているからであると推定できる．また『素性 2:」』は，正解例 2 や誤り例 2 のように，キャッチフレーズの具体的な内容を先に述べてから「キャッチ

フレーズ」という単語が使用される傾向にあることを示している．この同義語対にはこのような傾向があるということがわかったが，表 6.12 にある他の素性を見ても，特定の分野の記事ではどちらの語を使用するというような特徴のある素性はなく，全体的な使い分けの必要は少ないと考えられる．

## 6.3 再現率の高さごとの傾向

再現率高とした同義語対は，先に示した例のように，使い分けが必要なものが多かった．再現率中の同義語対は，先の例のようにある語が直前または直後にくる場合にどちらか一方のみを使用するものや，同義語対が広義と狭義の関係にあるものなどが多く含まれていた．後者の例としては,「宴」と「披露宴」がある．EDR 辞書内ではどちらも「宴会」と定義されていたが，国語辞書を引くと「宴」は広く「宴会」の意味をもち,「披露宴」には「めでたい事柄を発表するための宴」とあり,「ひろめの宴」という狭義の宴であることが示されていた．再現率低の同義語対では，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同義語対が確認できた．また，再現率低となった同義語対の中には，ある語とその略語が対となっているものや，日本語と外来語の対がいくつか含まれていた．略語と対になっていたものの例としては「省エネ」と「省エネルギー」，日本語と外来語の対では「謳い文句」と「キャッチフレーズ」があった．

# 第7章　おわりに

本研究では機械学習を用いて同義語対の使い分けを行った．本研究の成果は2つある．

第1の成果として，実験により，機械学習を用いる提案手法の正解率 (0.87) が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 (0.72) 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同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同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同義語対に「貯金」と「貯蓄」の対や「賞牌」と「メダル」などがあった．また，いくつかの同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．

# 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，鳥取大学工学部知能情報工学科計算機工学講座 C 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，徳久雅人講師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] 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.

[2] 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.

[3] 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. 情報処理学会研究報告, pp. 7–14, 2008.

[4] Eric Sven Ristad. Maximum Entropy Modeling for Natural Language. ACL/EACL Tutorial Program, Madrid, 1997.

[5] 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, 2001.

[6] Masao Utiyama. Maximum Entropy Modeling Package. http:// www.nict.go.jp/ x/ x161/ members/ mutiyama/ software.html#maxent, 2006.

[7] Masaki Murata, Kiyotaka Uchimoto, Masao Utiyama, Qing Ma, Ryo Nishimura, Yasuhiko Watanabe, Kouichi Doi, and Kentaro Torisawa. Using the maximum entropy method for natural language processing: Category estimation, feature extraction, and error correction. Cognitive Computation, Vol. 2, No. 4, pp. 272–279, 2010.

[8] 日本語形態素解析システム juman version7.0. http://nlp.ist.i.kyoto-u.ac.jp/index.php?cmd=read&page=JUMAN.

# 付 録A　付録

本研究で使用したEDR辞書から獲得した同義語対の一部を，機械学習の再現率の高さごとに示す.

表 A.1: 実験で使用した同義語対

| 再現率の高さ | 再現率 | 同義語対 |
|---|---|---|
| 再現率高 | 9割以上 | 「婦女」と「女流」 |
| | | 「民社」と「民社党」 |
| | 8割以上9割未満 | 「賞牌」と「メダル」 |
| | | 「貯金」と「貯蓄」 |
| 再現率中 | 7割以上8割未満 | 「論評」と「評論」 |
| | | 「私邸」と「自宅」 |
| | | 「所管」と「管轄」 |
| | 6割以上7割未満 | 「衣料」と「衣類」 |
| | | 「宴会」と「宴」 |
| | | 「西岸」と「西海岸」 |
| | 5割以上6割未満 | 「知識人」と「識者」 |
| | | 「コンテスト」と「コンクール」 |
| | | 「青少年」と「青年」 |
| 再現率低 | 4割以上5割未満 | 「上期」と「上半期」 |
| | | 「新春」と「新年」 |
| | 3割以上4割未満 | 「電算」と「コンピューター」 |
| | | 「慶應大」と「慶大」 |
| | 2割以上3割未満 | 「うたい文句」と「キャッチフレーズ」 |
| | | 「出費」と「支出」 |
| | 1割以上2割未満 | 「初旬」と「上旬」 |
| | 1割未満 | 「現況」と「現状」 |