# 機械学習を用いた同義語の使い分け

強田 吉紀 [*1]　村田 真樹 [*2]　三浦 智 [*2]　徳久 雅人 [*2]
[*1] 鳥取大学 知能情報工学科
[*2] 鳥取大学大学院 工学研究科 情報エレクトロニクス
{s092026, murata, s072052, tokuhisa}@ike.tottori-u.ac.jp

## 1　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 [3] 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 [4]．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分けも試みる．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．

- 機械学習に基づき同義語の使い分けを行った．50 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 0.72 であるのに対して，機械学習を用いる提案手法は 0.87 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分けした．
- 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．

## 2　問題設定と提案手法

### 2.1　問題設定

使い分けをしたい同義語対 X,Y があるとする．語 X と語 Y 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を補うべきかを推定することが，本研究で扱う問題である．その文に元に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．

### 2.2　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を補うべきかを推定する．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．

文献 [4] を参考にし，機械学習の素性には表 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 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 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 |

## 3 実験に用いる同義語対

本節では実験に用いる同義語対の説明を行う．3.1 節で同義語の認識方法を説明する．3.2 節で実験に用いる同義語対の選定方法を説明する．

### 3.1 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 (例えば，衣料の概念識別子は 0e504a であり，0e504a の意味は「体に着るもの」である)．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．

### 3.2 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 100 回以上であること

**条件 4** 形態素解析システム JUMAN 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 [4] 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 96 対あり，その中からランダムに取り出した 48 対を実験に用いる同義語対とする．

## 4 実験

### 4.1 実験方法

獲得した同義語対 48 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，同義語対のいずれかの語を含む，1991 年の毎日新聞の文である．毎日新聞には全く同じ文が存在しており，同じ文は一つだけ残し残りは削除して実験を行う．評価はクロスバリデーションで行う．

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 X, 語 Y 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．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との比較を行う．

### 4.2 実験結果

機械学習の再現率の高さごとに 48 個の同義語対を分類した結果を表 2 に示す．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 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．提案手法とベースライン手法の同義語対ごとの正解率を 48 個の同義語対で比較した結果を表 4 に示す．表 4 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 2 再現率の高さごとに分類した結果

| 再現率の高さ | 割合 |
|---|---|
| 高 | 0.10 % ( 5/48) |
| 中 | 0.56 % (27/48) |
| 低 | 0.33 % (16/48) |

表 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平均

| | 再現率：高 | 再現率：中 | 再現率：低 | 全ての対 |
|---|---|---|---|---|
| 提案手法 | 0.97 | 0.83 | 0.81 | 0.87 |
| ベースライン手法 | 0.68 | 0.67 | 0.80 | 0.72 |

表 4 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 | 27 | 7 |
| ベースライン手法○ | 0 | 0 | 2 |
| 差なし | 0 | 0 | 7 |

## 5 考察

### 5.1 提案手法とベースライン手法の比較

表 3 のように，提案手法とベースライン手法の正解率は，0.87 と 0.72 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 4 のように，再現率高と中の同義語対全てと低の同義語対 7 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 48 組のうちの約 8 割である．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 7 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．

### 5.2 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 1 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．

●再現率高の例：「婦女」と「女流」

(**正例** 1) 別件の窃盗、婦女 暴行については被告も争わず、判決は別件に限って有罪を認め、改めて懲役六年を宣告した。

(**正例** 2) 伝統文芸の短歌に現代の若者言葉を用いて幅広いファンを持つ 女流 歌人、俵万智さん（２８）らユニークな人材も加わっている。

(**負例** 1) 警察当局の最近の発表では、人口約三十万人の同国で昨年発生した 婦女 暴行事件は十一件と前年の二倍。

(**負例** 2) また、女流 招待で注目された佐藤紀子選手は一回戦で大会最年少の十八歳、斉藤杉太郎選手（山梨）に敗れて姿を消した。

表 5 機械学習の結果 (再現率高の例)

| | 再現率 | 適合率 | 総数 |
|---|---|---|---|
| 婦女 | 0.99 | 0.99 | 136 |
| 女流 | 0.97 | 0.97 | 63 |

表 6 機械学習が参考にした素性 (再現率高の例)

| 女流 | | 婦女 | |
|---|---|---|---|
| 素性 | 正規化 α 値 | 素性 | 正規化 α 値 |
| デフォルト素性 | 0.72 | 素性 2:暴行 | 0.72 |
| 素性 4:特殊 | 0.57 | 素性 1:容疑 | 0.58 |
| 素性 1:東京 | 0.56 | 素性 1:被告 | 0.57 |

再現率高の例として，「婦女」と「女流」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 0f1f60 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「女性であること」を意味する．表 8 より周辺に「暴行」「容疑」「被告」などの事件に関係する語が出現すると，「婦女」を用い，そうでない場合に「女流」を用いる使い分けが存在することがわかる．「婦女」を含む文は，大半が暴行事件に関する文章となっており，「婦女」という語を使用する場面自体が限られていると推測できる．「女流」を含む文では，芸術や競技に関する記事が多く見られた．また，素性 4 から，記号 (「」) などの特殊文字が使用されていることがわかり，文中で作品の引用などが行われていると推測できる．以上より，この同義語対は使い分けが必要と考えられる．

●再現率中の例：「衣料」と「衣類」

(**正例** 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。

(**負例** 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。

(**負例** 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。

表 7 機械学習の結果 (再現率中の例)

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

表 8 機械学習が参考にした素性 (再現率中の例)

| 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 |

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．正解例 1 や表 8 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．

●再現率低の例：「上期」と「上半期」

(**正例** 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。

(**負例** 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。

(**負例** 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 9 機械学習の結果 (再現率低の例)

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

表 10 機械学習が参考にした素性 (再現率低の例)

| 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．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．

### 5.3 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「諺い文句」と「キャッチフレーズ」があった．

## 6 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．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．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分けした．特に使い分けが必要とされた同義語対に「婦女」と「女流」などがあった．いくつかの同義語対について実際に素性を分析し，使い分けに役立つ情報を明らかにした．

## 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, 情報処理学会研究報告, SLP, 音声言語情報処理, p.7-14(2008).

[2] 伊藤 山彦, 相川 勇之, 鈴木 克志, コーパスからの同義語の獲得：スパース性への対処, 全国大会講演論文集, 第 56 回平成 10 年前期, p.241-242(1998).

[3] 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, 29 巻, p.161-182(1979).

[4] 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, p.300-303(2011).

[5] 黒橋禎夫, 河原大輔, 日本語形態素解析システム JUMAN version7.0 使用説明書, 京都大学大学院情報学研究科 (2012).

[6] 黒橋禎夫, 河原大輔, 日本語構文解析システム KNP version4.01 使用説明書, 京都大学大学院情報学研究科 (2012).