概要

本研究は類義語に対して，教師あり機械学習を用いることにより，類義語の使い分けを行う．類義語の使い分けに関わる知見を得ることを目指す．

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「おおよそ」と「おおむね」という類義語の組では,「おおよその目安」とはいうが,「おおむねの目安」とはいわない．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った．また，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．ある類義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，その類義語の組は特に使い分けの必要な類義語とわかる．また，機械学習が使用した素性を分析して，類義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既存の辞書から獲得した類義語の組を対象に行う．

本研究の成果は 2 つある．1 つ目は，類義語 11 組について実験を行った結果，正解率がマクロ平均で約 8 割の性能が得られたため，この提案手法自体が類義語の使い分けに対して有用であることが挙げられる．もう 1 つは，いくつかの類義語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．特に，類義語の使い分けに関する文献に載っていないような新たな知見が多く得られた．例えば,「作成」は「表」「リスト」などを作る時に使われ,「作製」は「細胞」「遺伝子」などを作る時に使われるなどの素性を得られた．また，品詞間における類義語の使い分けに関する特徴も得られた．この 2 つの成果は，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に利用できる．

目 次

表 目 次

図 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「場合」と「際」などがある．類義語に関する研究では，西尾 [3] 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 [4] は異表記の使い分けを機械学習で行った．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，強田ら [1] は EDR 電子化辞書から得られる類義語を利用し，機械学習による名詞の類義語の使い分けの研究を行った．中瀬 [2] は強田らと同様に，EDR 電子化辞書から得られる類義語を利用し，機械学習による副詞の類義語の使い分けの研究を行った．しかし，強田，中瀬らの研究では名詞，副詞の類義語の一部でしか使い分けの研究を行っていないため，使い分けが必要な類義語は強田，中瀬らが扱った類義語の他にもまだ多数存在する．そこで，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う．

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，使い方の分かる類語例解辞典 [5] および「擬音語・擬態語」使い分け帳 [6] から得られる類義語を利用する．

類義語は意味がほぼ同じであり，一見類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,「おおよそ」と「おおむね」は文献 [5] によると「ほとんどすべてであるさま」という意味で類義語とされているが，後ろに「の目安」をつけることができるのは「おおよそ」の方だけであり，後ろに「良好」をつける場合は「おおむね」だけである．このように使い分けが必要な場合がある．

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．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．

- 類義語の使い分けのために機械学習を使用し，類義語 11 組について実験を行った結果，正解率のマクロ平均は「データ数を出現率に合わせた実験」では，提案手法が 0.84，ベースライン手法が 0.65，素性 2 のみの手法では 0.82 であり,「データ数を同数に揃えた実験」では，提案手法が 0.81，ベースライン手法が 0.42，素性 2 のみの手法では 0.78 であったため，この提案手法自体が類義語の使い分けに対して有用である．

- 実際に機械学習における素性 (学習に用いる情報のこと) を分析することで類義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「作成」の推定に役立つ素性には「表」「リスト」などがあり,「作製」の推定に役立つ素性には「細胞」「遺伝子」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，その結果と考察について記述する．第 5 章では，第 4 章の結果から素性分析による考察を行う．第 6 章では，機械学習を用いた文章の誤り訂正についての実験についての説明と，その結果について記述する．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 [3] 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 [4] 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．

2.1　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 [3]．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 [4]．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[7] 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[7] 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．

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の組 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，類義語の組「おおよそ」「おおむね」を例に以下に示す．

おおよそ の目安ですのでメーカーによって異なります。 おおむね 良好な近似を得られているので今回はその法則を書きます。

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 の目安ですのでメーカーによって異なります。 X 良好な近似を得られているので今回はその法則を書きます。

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，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.

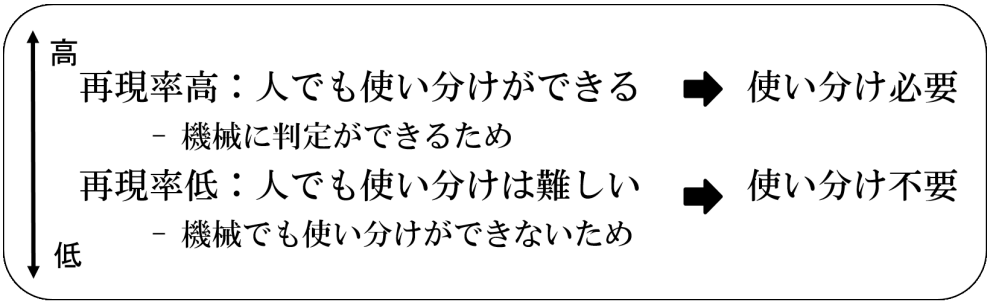

図 3.1: 再現率の高さごとの傾向

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 [8, 9, 10, 11].

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

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

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 [4][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．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: 類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　使い分けの実験と考察

本章では，本研究で実験を行った類義語の組を 4.1 節で説明し，本研究が行った実験方法を 4.2 節で説明し，実験結果を 4.3 節に示し，人手評価の結果を 4.4 節に示す．

4.1　実験データ

本研究では，使い方の分かる類語例解辞典 [5] および「擬音語・擬態語」使い分け帳 [6] から人手で選んだ類義語を利用する．表 4.1 に実験を行った類義語の組を示す．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．データ数は類義語の組内で同数にした実験と，類義語の出現率に合わせたたデータ数で実験を行った．表 4.2 に類義語の出現率に合わせた実験のデータ数を，表 4.3 に類義語の組においてデータ数を同数にした実験のデータ数を示す．データ数は 1 語につき 100 文以上で実験を行うことを条件とした．

表 4.1: 実験を行った類義語の組

1	「おかげ」「せい」「ため」
2	「場合」「時」「際」
3	「言う」「話す」
4	「発展」「発達」「進展」「進歩」
5	「作成」「作製」「制作」「製作」「製造」
6	「予想」「予測」
7	「おおよそ」「おおむね」
8	「がっかり」「がっくり」
9	「うろうろ」「ぶらぶら」
10	「はっきり」「きっぱり」
11	「うっすら」「ぼんやり」

表 4.2: 類義語の出現率に合わせたた実験のデータ数

類義語の組	分類先	データ数
「おかげ」「せい」「ため」	「おかげ」	105
	「せい」	114
	「ため」	4781
	全て	5000
「場合」「時」「際」	「場合」	3946
	「時」	3450
	「際」	2604
	全て	10000
「言う」「話す」	「言う」	6271
	「話す」	3729
	全て	10000
「発展」「発達」「進歩」「進展」	「発展」	5671
	「発達」	1030
	「進展」	2397
	「進歩」	902
	全て	10000
「作成」「作製」「制作」「製作」「製造」	「作成」	2538
	「作製」	368
	「制作」	2006
	「製作」	1309
	「製造」	3779
	全て	10000
「予想」「予測」	「予想」	7396
	「予測」	2604
	全て	10000
「おおよそ」「おおむね」	「おおよそ」	937
	「おおむね」	1920
	全て	2857
「がっかり」「がっくり」	「がっかり」	2243
	「がっくり」	396
	全て	2639
「うろうろ」「ぶらぶら」	「うろうろ」	124
	「ぶらぶら」	160
	全て	284
「はっきり」「きっぱり」	「はっきり」	9836
	「きっぱり」	702
	全て	10538
「ぼんやり」「うっすら」	「ぼんやり」	567
	「うっすら」	397
	全て	964

表 4.3: 類義語の組において同数にした実験のデータ数

類義語の組	1 語のデータ数	全データ数
「おかげ」「せい」「ため」	4186	12558
「場合」「時」「際」	4000	12000
「言う」「話す」	5000	10000
「発展」「発達」「進歩」「進展」	2500	10000
「作成」「作製」「制作」「製作」「製造」	1939	9695
「予想」「予測」	5000	10000
「おおよそ」「おおむね」	937	1974
「がっかり」「がっくり」	285	570
「うろうろ」「ぶらぶら」	124	248
「はっきり」「きっぱり」	702	1404
「ばんやり」「うっすら」	397	794

4.2 実験方法

獲得した類義語 11 組について， 類義語の組ごとに類義語の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，類義語の組のいずれかの語を含む文である．評価は 10 分割のクロスバリデーションで行う．類義語の組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 である対象語の前後 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．

4.3 実験結果

4.3.1 節ではデータ数を出現率に合わせた実験結果について記述する．4.3.2 節ではデータ数を同数に揃えた実験結果について記述する．

4.3.1 データ数を出現率に合わせた実験

機械学習の正解率の高さごとに類義語の組を分類した割合を表 4.4 に示す．データ数を出現率に合わせた類義語の組ごとの提案手法，ベースライン手法，素性 2 のみの手法での正解率を表 4.5 に示す．また，類義語の組のうち出現頻度が最も多かった語を下線で示す．データ数を出現率に合わせた提案手法，ベースライン手法の類義語の組ごとの正解率を類義語 11 組で比較した結果を表 4.6 に示す．データ数を出現率に合わせた提案手法，素性 2 のみの手法の類義語の組ごとの正解率を類義語 11 組で比較した結

果を表 4.7 に示す．表 4.6，表 4.7 における「差なし」とは，提案手法とベースライン手法，または素性 2 のみの手法の正解率の差が± 0.01 以内であった類義語の組の数を示す.「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類義語の組の数「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類義語の組の数「素性 2 のみの手法○」は「差なし」以外であり，かつ素性 2 のみの手法の正解率の方が高かった類義語の組の数を示す.

表 4.4: データ数を出現率に合わせた正解率の高さごとに分類した類義語の組

正解率	類義語の組
9 割以上	「おかげ」「せい」「ため」
	「はっきり」「きっぱり」
	「おおよそ」「おおむね」
8 割以上 9 未満	「場合」「時」「際」
	「発展」「発達」「進展」「進歩」
	「予想」「予測」
	「がっかり」「がっくり」
	「うっすら」「ぼんやり」
7 割以上 8 未満	「言う」「話す」
	「作成」「作製」「制作」「製作」「製造」
6 割以上 7 未満	「うろうろ」「ぶらぶら」

表 4.5: データ数を出現率に合わせた，提案手法，ベースライン手法，素性 2 のみの手法の正解率

類義語の組	提案手法	ベースライン手法	素性 2 のみの手法
「おかげ」「せい」「ため」	0.97	0.96	0.97
「場合」「時」「際」	0.83	0.39	0.78
「言う」「話す」	0.74	0.63	0.76
「発展」「発達」「進展」「進歩」	0.85	0.57	0.82
「作成」「作製」「制作」「製作」「製造」	0.79	0.38	0.71
「予想」「予測」	0.83	0.74	0.82
「おおよそ」「おおむね」	0.92	0.67	0.88
「がっかり」「がっくり」	0.86	0.68	0.86
「うろうろ」「ぶらぶら」	0.68	0.56	0.69
「はっきり」「きっぱり」	0.98	0.93	0.98
「うっすら」「ぼんやり」	0.82	0.59	0.79
マクロ平均	0.84	0.65	0.82

表 4.6: データ数を出現率に合わせた提案手法とベースライン手法の類義語の組ごとの正解率の比較結果

提案手法○	11
ベースライン手法○	0
差なし	0

表 4.7: データ数を出現率に合わせた提案手法と素性 2 のみの手法の類義語の組ごとの正解率の比較結果

提案手法○	5
素性 2 のみの手法○	2
差なし	4

表 4.8: データ数を出現率に合わせた，提案手法，素性 2 のみの手法の比較

類義語の組	提案手法○素性 2 のみの手法×	提案手法×素性 2 のみの手法○
「おかげ」「せい」「ため」	50	36
「場合」「時」「際」	1088	552
「言う」「話す」	805	1029
「発展」「発達」「進展」「進歩」	759	467
「作成」「作製」「制作」「製作」「製造」	1366	544
「予想」「予測」	707	677
「おおよそ」「おおむね」	205	86
「がっかり」「がっくり」	38	38
「うろうろ」「ぶらぶら」	27	30
「はっきり」「きっぱり」	53	58
「うっすら」「ぼんやり」	88	59
合計	5186	3576

データ数を出現率に合わせた実験における機械学習の結果，提案手法で正解し，素性 2 のみの手法では不正解だった文の数と提案手法で不正解であり，素性 2 のみの手法で正解した文の数を表 4.8 に示す．この結果を基に，この 2 つの手法の有意差を検定した．

表 4.5 のように正解率のマクロ平均は提案手法が 0.84, ベースライン手法が 0.65, 素性 2 のみの手法が 0.82 であった．提案手法の正解率はベースライン手法や素性 2 のみの手法の正解率より高かった．また，表 4.6 より，類義語 11 組全てで，ベースライン手法よりも正解率が高い結果であった．表 4.7 より類義語 11 組で提案手法と素性 2 のみの手法を比較すると,「提案手法○」が 5 組であり,「素性 2 のみの手法○」が 2 組であり,「差なし」が 4 組であった．表 4.8 より「提案手法○素性 2 のみの手法×」の合計数と「提案手法×素性 2 のみの手法○」の合計数を用い二項分布に基づく片側検定の符号検定により有意水準 5%で有意差があった．これにより，提案手法および機械学習で使用した素性は類義語の判別に十分有用であるといえる．

今回設定したベースライン手法は類義語の組における，出現頻度が最も多い語を全て分類先とするものなので，データ数を出現率に合わせた実験では出現頻度に極端に差があるとベースライン手法での正解率が極端に良くなる．しかし，類義語の組内で出現頻度が極端に少ない語については機械学習による素性も少なくなってしまい，正解率も極端に低くなる．また，使い分けに役立つ素性も多く得られなかった．

4.3.2 データ数を同数に揃えた実験

機械学習の正解率の高さごとに類義語の組を分類した割合を表 4.9 に示す．データ数を同数に揃えた類義語の組ごとの提案手法，ベースライン手法，素性 2 のみの手法での正解率の結果を表 4.10 に示す．データ数を同数に揃えた提案手法，ベースライン手法の類義語の組ごとの正解率を類義語 11 組で比較した結果を表 4.11 に示す．データ数を同数に揃えた提案手法，素性 2 のみの手法の類義語の組ごとの正解率を類義語 11 組で比較した結果を表 4.12 に示す．表 4.11，表 4.12 における「差なし」とは，提案手法とベースライン手法，または素性 2 のみの手法の正解率の差が± 0.01 以内であった類義語の組の数を示す.「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類義語の組の数「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類義語の組の数「素性 2 のみの手法○」は「差なし」以外であり，かつ素性 2 のみの手法の正解率の方が高かった類義語の組

の数を示す.

表 4.9: データ数を同数に揃えた正解率の高さごとに分類した類義語の組

正解率	類義語の組
9 割以上	「おかげ」「せい」「ため」
	「はっきり」「きっぱり」
	「おおよそ」「おおむね」
8 割以上 9 未満	「発展」「発達」「進展」「進歩」
	「予想」「予測」
	「がっかり」「がっくり」
	「うっすら」「ぼんやり」
7 割以上 8 未満	「場合」「時」「際」
	「言う」「話す」
	「作成」「作製」「制作」「製作」「製造」
6 割以上 7 未満	「うろうろ」「ぶらぶら」

表 4.10: データ数を同数に揃えた，提案手法，ベースライン手法，素性 2 のみの手法の正解率

類義語の組	提案手法	ベースライン手法	素性 2 のみの手法
「おかげ」「せい」「ため」	0.90	0.33	0.86
「場合」「時」「際」	0.75	0.33	0.68
「言う」「話す」	0.77	0.50	0.78
「発展」「発達」「進展」「進歩」	0.81	0.25	0.78
「作成」「作製」「制作」「製作」「製造」	0.73	0.20	0.64
「予想」「予測」	0.81	0.50	0.82
「おおよそ」「おおむね」	0.90	0.50	0.83
「がっかり」「がっくり」	0.82	0.50	0.83
「うろうろ」「ぶらぶら」	0.63	0.50	0.65
「はっきり」「きっぱり」	0.93	0.50	0.91
「うっすら」「ぼんやり」	0.83	0.50	0.79
マクロ平均	0.81	0.42	0.78

表 4.11: データ数を同数に揃えた提案手法とベースライン手法の類義語の組ごとの正解率の比較結果

提案手法○	11
ベースライン手法○	0
差なし	0

表 4.12: データ数を同数に揃えた提案手法と素性 2 のみの手法の類義語の組ごとの正解率の比較結果

提案手法〇	7
素性 2 のみの手法〇	4
差なし	0

表 4.13: データ数を同数に揃えた，提案手法，素性 2 のみの手法の比較

類義語の組	提案手法〇素性 2 のみの手法×	提案手法×素性 2 のみの手法〇
「おかげ」「せい」「ため」	870	371
「場合」「時」「際」	1944	1047
「言う」「話す」	772	864
「発展」「発達」「進展」「進歩」	963	642
「作成」「作製」「制作」「製作」「製造」	1563	691
「予想」「予測」	1438	1497
「おおよそ」「おおむね」	211	79
「がっかり」「がっくり」	24	31
「うろうろ」「ぶらぶら」	21	26
「はっきり」「きっぱり」	53	31
「うっすら」「ぼんやり」	77	49
合計	7936	5328

データ数を同数に揃えた実験において機械学習の結果，提案手法で正解し，素性 2 のみの手法では不正解だった文の数と提案手法で不正解であり，素性 2 のみの手法で正解した文の数を表 4.13 に示す．この結果を基にこの 2 つの手法の有意差を検定した．

表 4.10 のように正解率のマクロ平均は提案手法が 0.81, ベースライン手法が 0.42, 素性 2 のみの手法が 0.78 であった．提案手法の正解率はベースライン手法，素性 2 のみの手法の正解率より高かった．また，表 4.11 より，類義語 11 組全てで，ベースライン手法よりも正解率が高い結果であった．表 4.12 より類義語 11 組で提案手法と素性 2 のみの手法を比較すると,「提案手法〇」が 7 組であり,「素性 2 のみの手法〇」が 4 組であった．表 4.13 より「提案手法〇素性 2 のみの手法×」の合計数と「提案手法×素性 2 のみの手法〇」の合計数を用い二項分布に基づく片側検定の符号検定により有意水準 5%で有意差があった．これにより，提案手法および機械学習で使用した素性は類義語の判別に十分有用であるといえる．

データ数を同数にしているのでベースライン手法の正解率は低くなった．また，データ数を類義語の組内で同数にすると，提案手法では類義語ごとの正解率に極端に差はなかった．また，得られる素性も類義語の組内で同程度得られた．

4.4 人手評価

データ数を同数に揃えた実験データから類義語 1 語につき類義語を含む文を 5 文ずつランダムに抜き出し, 類義語の組のうち，どの類義語が正しいかを被験者の 3 人が選び，類義語の組ごとの正解率を求めた．提案手法と被験者 3 人の正解率の結果を比較した結果を表 4.14 に示す.

表 4.14: 人手評価

類義語の組	提案手法	被験者 1	被験者 2	被験者 3
「おかげ」「せい」「ため」	0.73 (11/15)	0.60 (9/15)	0.80 (12/15)	0.93 (14/15)
「場合」「時」「際」	0.73 (11/15)	0.80 (12/15)	0.93 (14/15)	0.80 (12/15)
「言う」「話す」	0.60 (6/10)	0.60 (6/10)	0.50 (5/10)	0.40 (4/10)
「発展」「発達」「進展」「進歩」	0.85 (17/20)	0.45 (9/20)	0.70 (14/20)	0.55 (11/20)
「作成」「作製」「制作」「製作」「製造」	0.80 (20/25)	0.64 (16/25)	0.60 (15/25)	0.44 (11/25)
「予想」「予測」	1.00 (10/10)	0.70 (7/10)	0.80 (8/10)	0.80 (8/10)
「おおよそ」「おおむね」	0.80 (8/10)	0.50 (5/10)	0.70 (7/10)	0.80 (8/10)
「がっかり」「がっくり」	0.90 (9/10)	0.60 (6/10)	0.80 (8/10)	0.70 (7/10)
「うろうろ」「ぶらぶら」	0.80 (8/10)	0.70 (7/10)	0.90 (9/10)	0.70 (7/10)
「はっきり」「きっぱり」	0.90 (9/10)	0.90 (9/10)	1.00 (10/10)	0.90 (9/10)
「うっすら」「ぼんやり」	0.80 (8/10)	0.80 (8/10)	0.90 (9/10)	0.80 (8/10)
マイクロ平均	0.81 (117/145)	0.65 (94/145)	0.77 (111/145)	0.68 (99/145)
マクロ平均	0.81	0.66	0.78	0.71

表 4.14 からマクロ平均，マイクロ平均共に提案手法の正解率の方が高かったため, 提案手法および機械学習で使用した素性は類義語の判別に有用であることが言える.

被験者実験では「発展」「発達」「進展」「進歩」や「作成」「作製」「制作」「製作」「製造」のように類義語の組内で類義語の数が多い組ほど正解率の低い結果となった. 特に,「作製」という語は機械に比べ，被験者の正解率が大きく下回った．人手では「作製」という語は「作成」と誤って使われやすいことがわかった.

また，被験者 3 人が誤り，機械が正解した例を以下に示す.

> 売上高は従来予想より５０億円減の８４００億円、営業利益は同３５億円減の１１５億円、最終（当期）利益は同２５億円減の４５億円となる見通しだ。

上記の文が元の文であり，被験者は「予想」を「予測」と誤った選択をした.「予測」は数値に関した文で使われやすく被験者は誤ったと考えられるが，機械学習では「利益」や被修飾語としての「売上高」は素性分析から「予想」の分類において，非常に有用な素性であったため，機械では正解したものと考えられる.「金」に関する文では

「予想」が使われることが多いことが素性分析からわかった．このようになぜ機械の方が正解率が高かったかをわかる例も発見できた．

第5章　素性分析による考察

本章では，素性分析による考察を記述する．節 5.1 では，類義語の組ごとの考察について記述する．節 5.2 では，品詞間における類義語の使い分けに関する特徴について記述する．

5.1　類義語の組ごとの考察

本節では，類義語の組ごとに使い分けに関する考察を行う．類義語ごとにデータ数を同数にした実験を基に考察を行った．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．

類義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．素性が全データでの出現率より偏って多くいずれかの分類先に出現しているかを，二項検定に基づく符号検定により求め，有意確率 p 値を求める．有意確率 p 値が 0.05 以下であり，学習データでの出現頻度が多い素性の例を表に示す．

また，正規化 α 値が高かった素性の例も表に示す．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．

5.1.1 「おかげ」「せい」「ため」

（正解例 1）「大勢の方が快く寄付してくださった おかげ」と感謝する。

（正解例 2）本人も「負けたのは自分の せい」と何度嘆いたことか。

（正解例 3）この ため、昨年６月に法律が改正されました。

（誤り例 1）福祉予算が削減されたり、スウェーデンの主張が通らなかった おかげ(ため) で、景気回復の遅れも原因らしい。

（誤り例 2）早起きの せい(おかげ) か、６０歳を超えてから一度も風邪をひいていない。

（誤り例 3）新進党にさまざまな批判が集まっているのも、切り替えができていない ため(せい) だ。

表 5.1 に類義語の組「おかげ」「せい」「ため」の機械学習の分類結果を示す．表 5.2 に類義語の組「おかげ」「せい」「ため」の有意差検定に基づいた機械学習が参考にした素性を示す．表 5.3 に類義語の組「おかげ」「せい」「ため」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 5.1: 機械学習の分類結果 (「おかげ」「せい」「ため」)

	データ数	再現率	適合率	F 値
おかげ	4186	0.90	0.88	0.89
ため	4186	0.95	0.95	0.95
せい	4186	0.87	0.88	0.87
総数	12558	0.90	0.90	0.90

表 5.2: 機械学習が参考にした素性 (有意差検定：「おかげ」「せい」「ため」)

おかげ		せい		ため	
素性	頻度 (p 値)	素性	頻度 (p 値)	素性	頻度 (p 値)
素性 2:対象語が文頭	523 (5.35×10^{-250})	素性 2:か (直後)	1351 (≈ 0)	素性 2:、(直後)	1323 (≈ 0)
素性 1:感謝	202 (2.26×10^{-89})	素性 2:その (直後)	367 (3.53×10^{-47})	素性 2:する (直前)	475 (1.45×10^{-197})
素性 2:対象語が文末	250 (1.81×10^{-64})	素性 2:も (直後)	271 (4.53×10^{-104})	素性 2:この (直前)	238 (5.15×10^{-84})
素性 34:できる	120 (3.44×10^{-43})	素性 1:自分	240 (1.99×10^{-24})	素性 1:死去	184 (3.17×10^{-87})
素性 1:皆さん	108 (1.94×10^{-45})	素性 1:不況	54 (1.35×10^{-13})	素性 2:防ぐ (直前)	56 (1.09×10^{-27})

表 5.3: 機械学習が参考にした素性 (正規化 α 値：「おかげ」「せい」「ため」)

おかげ		せい		ため	
素性	正規化 α 値	素性	正規化 α 値	素性	正規化 α 値
素性 2:対象語が文頭	0.80	素性 1:自分	0.60	素性 2:、(直後)	0.71
素性 1:感謝	0.70	素性 1:次男	0.53	素性 2:を (2 語前)	0.67
素性 2:対象語が文末	0.64	素性 4:？	0.53	素性 1:可能	0.59
素性 2:」(直後)	0.63	素性 1:不足	0.49	デフォルト素性	0.55
素性 1:チーム	0.61	素性 1:お前	0.49	素性 7:の	0.53

「おかげ」はまず文頭や文末にくることが多いが文頭で使われるのは「おかげ」だけである.「せい」は「〜せいか」「〜せいもある」「〜せい？」という表現が多い.「ため」は直後に読点がくることが多く,「ために」や「このため」といった表現が多かった.「おかげ」は「感謝」「できる」などプラスの表現の語があり,「せい」は「不況」「負ける」などマイナスの表現の語とともに使われる.「ため」はその中立であり,「〜のため死去。」のように理由を表す表現として幅広く使われる．また,「ため」には「防ぐため」「〜をするため」のように目的を表す用法も多く使われる.「ため」はデフォルト素性の正規化 α 値が高かったことから，特徴のない文の場合は「ため」が使用される傾向にあるとわかる.「おかげ」は「皆さん」「先生」,「せい」は「自分」「お前」などの人物を表す語とともに使われるが,「ため」は人物が続くことは少ない.

また,「使い方の分かる類語例解辞典」[5] によると以下のように書かれている.

> 「おかげ」は, 他から影響を受けた結果, 望ましい事態が成立したことを表し, 話し手の感謝の気持ちを伴う. 反対に, 望ましくない事態が成立したことを表すのは「せい」で, その責任を自分以外の他者 (人や物事) に押し付ける意味合いがある. 望ましくない事態の成立に「おかげ」が使われる

こともある. 本来なら「せい」を使うところにわざと反対の意味の「おかげ」を使い「せい」を使うより皮肉・非難の意味合いが強まり, 話し手のマイナス感情を強調することになる. 「ため」はマイナス感情もプラス感情も伴わないため, 望ましい事態か望ましくない事態下に関わらず広く中立的に用いられる表現である. 「おかげ」「せい」に比べて, 人物に直接には続きにくい. また目的を表す用法もある.

文献 [5] と素性分析の結果を比較し同じ結果となったことを以下に示す. 素性分析から「おかげ」は「感謝」「元気」などのプラス感情の語が多く,「せい」が「不況」「負ける」などのマイナス感情の語が多かったため,「おかげ」はプラス感情の文,「せい」はマイナス感情の文で使われ, その中立が「ため」であるということが分析からわかり, 文献と同じ結果となった. また,「ため」は「おかげ」「せい」に比べて, 人物に直接続きにくく, 目的を表す用法が多いということも同じであった.

文献に載っておらず, 素性分析から新たにわかったことを以下に示す. 文頭になるのは「おかげ」だけであったり,「〜せいもある」,「〜ため、」のようによく使われる用法など, 多くのことがわかった.

文献に載っているが素性分析からは得られなかったことを以下に示す.「おかげ」を皮肉・非難の意味で使う用法は素性分析からはわからなかった. また, 文献と違った点としては,「せい」は文献には「自分以外の他者に押し付ける」とあるが, 素性分析からは「自分」にも使われるという結果となった.

5.1.2 「場合」「時」「際」

（正解例 1）では、帰省しない 場合 はどうしたらよいのでしょうか。

（正解例 2）あの 時 は全く手探りの状態でした。

（正解例 3）まだ知事の時に知り合い、最近の訪日の 際 も話し合った。

（誤り例 1）市は東北大などと連携し震災時の津波被害の再現や将来発生した 場合(際) の予測を行ってきた。

（誤り例 2）小学校低学年の 時(場合) だと、本当は遠視なのに近視と間違えられていた、なんていうこともよくある。

（誤り例 3）このため捜査員が訪れた際(時) も、本当のことは話さず、一度は“ 対象外 ”になっていた。

表 5.4 に類義語の組「場合」「時」「際」の機械学習の分類結果を示す．表 5.5 に類義語の組「場合」「時」「際」の有意差検定に基づいた機械学習が参考にした素性を示す．表 5.6 に類義語の組「場合」「時」「際」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 5.4: 機械学習の分類結果 (「場合」「時」「際」)

	データ数	再現率	適合率	F 値
場合	4000	0.74	0.76	0.75
時	4000	0.74	0.75	0.75
際	4000	0.74	0.72	0.73
総数	12000	0.74	0.74	0.74

表 5.5: 機械学習が参考にした素性 (有意差検定：「場合」「時」「際」)

場合		時		際	
素性	頻度 (p 値)	素性	頻度 (p 値)	素性	頻度 (p 値)
素性 4:は	1166 (8.97×10^{-50})	素性 4:句点	151 (2.05×10^{-39})	素性 2:に (直後)	1262 (2.62×10^{-66})
素性 22:ない	349 (7.37×10^{-95})	素性 4:から	125 (6.77×10^{-55})	素性 22:する	1233 (2.21×10^{-71})
素性 31:ある	256 (1.72×10^{-44})	素性 2:歳 (2 語前)	109 (2.44×10^{-51})	素性 1:昨年	174 (9.27×10^{-20})
素性 31:よる	63 (6.53×10^{-27})	素性 2:あの (直前)	70 (1.97×10^{-34})	素性 1:事件	102 (1.06×10^{-14})
素性 1:以上	161 (7.68×10^{-23})	素性 2:そんな (直前)	48 (4.31×10^{-21})	素性 1:訪問	84 (3.52×10^{-25})

表 5.6: 機械学習が参考にした素性 (正規化 α 値:「場合」「時」「際」)

場合		時		際	
素性	正規化 α 値	素性	正規化 α 値	素性	正規化 α 値
素性 1:負担	0.72	デフォルト素性	0.77	素性 1:年	0.67
素性 1:想定	0.69	素性 31:来る	0.53	素性 1:先月	0.61
素性 1:予想	0.66	素性 2:いう (直前)	0.50	素性 1:一昨年	0.61
素性 1:以上	0.57	素性 1:出発	0.49	素性 2:、(直後)	0.58
素性 2:さん (2 語前)	0.56	素性 1:子供	0.48	素性 13:訪	0.52

「場合」は「場合による」「〜場合がある」のような表現が多く,「時」は「〜時から」や「〜時。」のように文末で使われる表現などが多く，また「際」は「〜する際、」「〜際に」のような表現が多かった.「場合」は「ない場合」「以上の場合」など物事の状態や条件を表す語と併せて使われやすい．特に,「以上」「%」「万」などの数量的な条件を表す語と共起しやすい．また,「〜さんの場合」など人物が前にくることも多い.一方,「際」は「昨年」「先月」などの具体的な時を表す語と共起しやすいことから出来事が起きた時点を表す文で使われることがわかった.「時」は「〜歳の時」のように具体的な人物の過去の話や「あの時」「そんな時」のように抽象的な時点を表す文によく使われる．また,「際」は「時」に比べ,「訪日」「有事」など堅苦しい文章に使われることが多い．デフォルト素性の正規化 α 値は分類先「時」でかなり高かったため,「時」は「場合」「際」よりも一般的に使われやすいことがわかる.

また，文献 [5] によると以下のように書かれている.

> いずれも, 動作状態を表す修飾語句を受けて用いられる. 「場合」は「時」と比べると, その物事, 状態が発生する具体的, 個別的な状態を指すことが多い.「際」は何かが行われる時間を予期した文脈で用いられることが多い.

文献 [5] と素性分析の結果を比較し同じ結果となったことを以下に示す.「場合」は「ない場合」「以上の場合」などの物事, 状態が発生する具体的, 個別的な状態を指すことが多く,「際」は「最近の訪日の際」などのように何かが行われる時間を表す文脈で用いられることが多いという点が同じであった.

文献に載っておらず，素性分析から新たにわかったことを以下に示す.「訪問」「訪日」など，訪れる時点を表す表現には「際」が使われ,「〜歳の時」「〜年生の時」のように人物に関する時期を表すものや「あの時」「その時」など抽象的な時点を表す表現には「時」が使われるというようなことが新たにわかった.

文献に載っているが素性分析から得られなかったことはなかった.

5.1.3 「言う」「話す」

（正解例 1） ２歳の私はね、まるで家畜を運ぶような引き揚げ列車の中で『おうちに帰ろ』って 言う んですって。

（正解例 2） 時代と芸術の価値判断への興味と、絵とゴッホへの興味ですね」と出演のきっかけを 話す。

（誤り例 1）スラスラと話すべき役なんですが、彼はリズム感をはずしトツトツと言う(話す)。

（誤り例 2）子どもが発言の機会を与えられているが、自分の意見を話す(言う) 機会がない場合。

表 5.7 に類義語の組「言う」「話す」の機械学習の分類結果を示す．表 5.8 に類義語の組「言う」「話す」の有意差検定に基づいた機械学習が参考にした素性を示す．表 5.9 に類義語の組「言う」「話す」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 5.7: 機械学習の分類結果 (「言う」「話す」)

	データ数	再現率	適合率	F 値
言う	5000	0.74	0.78	0.76
話す	5000	0.79	0.75	0.77
総数	10000	0.77	0.77	0.77

表 5.8: 機械学習が参考にした素性 (有意差検定:「言う」「話す」)

言う		話す	
素性	頻度 (p 値)	素性	頻度 (p 値)
素性 2:べき (直後)	131 (3.67×10^{-40})	素性 13:語	35 (7.55×10^{-6})
素性 2:って (直前)	41 (4.55×10^{-13})	素性 2:について (直前)	26 (7.62×10^{-6})
素性 2:通り (直後)	48 (9.84×10^{-12})	素性 2:機会 (直後)	20 (1.05×10^{-5})
素性 1:文句	26 (1.49×10^{-8})	素性 13:体験	17 (3.64×10^{-4})
素性 2:ものの (直後)	18 (3.81×10^{-5})	素性 13:きっかけ	10 (5.86×10^{-3})

表 5.9: 機械学習が参考にした素性 (正規化 α 値:「言う」「話す」)

言う		話す	
素性	正規化 α 値	素性	正規化 α 値
素性 2:』(2 語前)	0.85	素性 1:効果	0.83
素性 1:誇り	0.80	素性 2:」(2 語前)	0.82
素性 1:舞台	0.79	素性 2:について (直前)	0.81
素性 1:自ら	0.75	素性 1:将来	0.80
デフォルト素性	0.58	素性 1:当時	0.77

「言う」は「言うべき」「〜って言う」「言う通り」「言うものの」といった表現がよく使われる.「話す」は「〜語を話す」「〜について話す」「話す機会」のような表現が多い.「言う」は「文句を言う」など日常的な会話を表す文章に使われやすく,「話す」は「体験」「きっかけ」「将来」のことなどを口にする時に使われる．デフォルト素性は「言う」の方が高いため,「言う」の方が広義の意味で使われやすいとわかる.

また，文献 [5] によると以下のように書かれている.

> 思うことを口にして表現する.「言う」は，思ったことを言葉で表現する意だが，まとまった内容を表現する場合だけではなく，反射的に小さな叫び声を上げるような場合や文章表現などにも用いられ，広い用法をもつ語.「話す」は，相手と会話をするという意もある.

文献 [5] と素性分析の結果を比較し同じ結果となったことを以下に示す.「言う」は,広い用法をもつ語,「話す」は「〜語を話す」や「体験を話す」など，相手と会話をするという意もある，という点は同じだった.

文献に載っておらず，素性分析から新たにわかったことを以下に示す.「言うべき」「言う通り」「話す機会」のようによく使われる用法が素性分析から新たにわかった.

文献に載っているが素性分析から得られなかったことを以下に示す．反射的に小さな叫び声を上げるような「言う」の表現は素性分析からはわからなかった.

5.1.4 「発展」「発達」「進展」「進歩」

（正解例 1）ＡＳＥＡＮの平和的発展を願ってやまない。

（正解例 2）発達する低気圧の北側に現れる典型的な雲で、発達につれて曲がりがきつくなる。

（正解例 3）また開発途上国の民主化・平等化の 進展 を支援する観点からの援助の実施についても検討する。

（正解例 4）医学の 進歩 に伴って登場した新しいモノサシを、採用するべきかどうかである。

（誤り例 1）トナー氏は経済以外でも、地域的、世界的問題での協力のほか、人権問題での 発展(進展) を新指導部に促したい意向を示した。

（誤り例 2）日本の陶器の 発達(発展) 過程での中国の影響を知るうえで貴重」という。

（誤り例 3）核交渉合意を協力 進展(発展) につなげるべきだ」と述べた。

（誤り例 4）科学が 進歩(発達) すればするほど、呪術も発達する。

表 5.10 に類義語の組「発展」「発達」「進展」「進歩」の機械学習の分類結果を示す. 表 5.11 に類義語の組「発展」「発達」「進展」「進歩」の有意差検定に基づいた機械学習が参考にした素性を示す. 表 5.12 に類義語の組「発展」「発達」「進展」「進歩」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 5.10: 機械学習の分類結果 (「発展」「発達」「進展」「進歩」)

	データ数	再現率	適合率	F 値
発展	2500	0.79	0.80	0.80
発達	2500	0.81	0.83	0.82
進展	2500	0.85	0.86	0.85
進歩	2500	0.80	0.76	0.78
総数	10000	0.81	0.81	0.81

表 5.11: 機械学習が参考にした素性 (有意差検定:「発展」「発達」「進展」「進歩」)

発展		発達		進展		進歩	
素性	頻度 (p 値)	素性	頻度 (p 値)	素性	頻度 (p 値)	素性	頻度 (p 値)
素性 2:経済 (直前)	265 (5.68×10^{-158})	素性 1:気圧	353 (1.05×10^{-208})	素性 1:交渉	591 (9.58×10^{-291})	素性 1:技術	654 (1.94×10^{-159})
素性 13:関係	243 (9.72×10^{-15})	素性 1:子ども	126 (2.21×10^{-60})	素性 1:問題	501 (3.29×10^{-89})	素性 1:科学	206 (4.74×10^{-28})
素性 1:地域	189 (8.67×10^{-37})	素性 1:積乱雲	91 (1.63×10^{-55})	素性 1:協議	295 (2.00×10^{-144})	素性 1:医療	177 (2.67×10^{-37})
素性 1:産業	101 (9.85×10^{-14})	素性 2:未 (直前)	45 (8.08×10^{-28})	素性 31:ない	294 (9.95×10^{-48})	素性 1:医学	117 (4.35×10^{-33})
素性 2:的 (直前)	47 (4.57×10^{-11})	素性 1:精神	86 (2.64×10^{-26})	素性 13:化	224 (2.85×10^{-116})	素性 1:人類	70 (6.08×10^{-20})

表 5.12: 機械学習が参考にした素性 (正規化 α 値:「発展」「発達」「進展」「進歩」)

発展		発達		進展		進歩	
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
素性 1:基礎	0.58	素性 1:気圧	0.70	素性 1:交渉	0.76	素性 1:医療	0.50
素性 1:経済	0.51	素性 1:子ども	0.68	素性 2:化 (2 語前)	0.64	素性 1:発展	0.49
素性 1:維持	0.50	素性 1:障害	0.60	素性 1:協議	0.63	素性 1:治療	0.49
素性 1:地域	0.49	素性 1:体	0.48	素性 1:米	0.60	素性 1:技術	0.47
素性 1:世界	0.47	素性 1:段階	0.45	素性 1:北朝鮮	0.55	デフォルト素性	0.36

「発展」は「経済」「産業」「地域」「世界」などがより良くなることに使われる.「発達」は「気圧」「積乱雲」など気候に関することや「子ども」「体」「精神」など心身の成長などに関して使われる．また,「コンピューター」や「インターネット」にも「発達」が使われる.「進展」は「交渉」「協議」など話し合いが進むことに使われる.「進歩」は「医学」「技術」「科学」は学術的なことが進むことに使われる．また,「発展」は「経済的発展」のように「的」が使われる表現が多い.「進展」は「機械化の進展」のように「化」が使われる表現が多い.「発達」は「未発達」というが,「進展」は「進展がない」のようにいう．デフォルト素性は「進歩」が一番高かった.

また，文献 [5] によると，以下のように書かれている.

> 「発達」「発展」「進歩」の 3 語は，技術・学問・文化などについて，物事が進んで前より上の段階に入ることを意味する.「発達」は，発育してより完全な方向に近づくこと，より規模が大きくなることを表し，筋肉などの器官や台風などに関しても使われる．一方,「進歩」はより望ましい良い方向へ進むことに用い,「発展」は物事の勢いが伸び広がることの意に用いる.「進展」は，物事の状況が時間の経過とともに変化し，新たな局面を迎えることで，その進む方向については特に良否を問題にしない.

文献 [5] と素性分析の結果を比較し同じ結果となったことを以下に示す.「発達」は,

「体」や「積乱雲」など体や気候に使われ,「進歩」は「医学」などがより望ましい良い方向へ進むことに使われ,「発展」は「地域」などの勢いが伸び広がることに使われ,「進展」は,「交渉」などの状況が時間の経過とともに変化し，新たな局面を迎えることに使われるという点が同じだった.

文献に載っておらず，素性分析から新たにわかったことを以下に示す.「進歩」は技術・学問に関して使われ,「発展」は文化について使われるとわかった．また,「経済的発展」のように「的」や「機械化の進展」のような「化」などの文献に載ってない使い分けに役立つ情報も素性分析から得られた.

文献に載っているが素性分析からは得られなかったことを以下に示す.「進展」の進む方向について良否を問題にしないということは，素性分析から考察することはできなかった.

5.1.5 「作成」「作製」「制作」「製作」「製造」

（正解例 1）職場では国会の資料 作成 に追われ、女性も当番日以外は午前３時まで勤務した。

（正解例 2）次にｉＰＳ細胞を 作製 する過程で、１５８種類から１種類ずつを加える実験を繰り返した。

（正解例 3）一社の意向だけで番組 制作 に影響を及ぼすなどは無理な話だ。

（正解例 4）すでに盧溝橋事件などをテーマにした映画の 製作 に入っている。

（正解例 5）報道によると、工場は車の車輪部品を 製造 しており、爆発は部品を研磨する作業場で発生した。

（誤り例 1）ＪＲ西日本のＩＣ乗車券「ＩＣＯＣＡ（イコカ）」をベースに 作成(作製)。

（誤り例 2）浮輪とランドマークタワー、海をデザインした新しいロゴマーク＝図＝も 作製(制作) した。

（誤り例 3）当初は耕作放棄田の再生をテーマにカメラを回し、ドキュメンタリー番組を 制作(製作) したが、東日本大震災の発生を受けて継続取材することにしたという。

（誤り例 4）昨年末に開催された予算 製作(作成) コンテスト「未来国会２０１０」を取材した。

（誤り例 5）カリフォルニア産木材で 製造(作製) したベンチには、習氏の名前と会談日が刻まれている。

表 5.13 に類義語の組「作成」「作製」「制作」「製作」「製造」の機械学習の分類結果を示す．表 5.14 に類義語の組「作成」「作製」「制作」「製作」「製造」の有意差検定に基づいた機械学習が参考にした素性を示す．表 5.15 に類義語の組「作成」「作製」「制作」「製作」「製造」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 5.13: 機械学習の分類結果 (「作成」「作製」「制作」「製作」「製造」)

	データ数	再現率	適合率	F 値
作成	1939	0.79	0.80	0.80
作製	1939	0.73	0.73	0.73
制作	1939	0.69	0.66	0.68
製作	1939	0.60	0.62	0.61
製造	1939	0.84	0.83	0.84
総数	9695	0.73	0.73	0.73

表 5.14: 機械学習が参考にした素性 (有意差検定:「作成」「作製」「制作」「製作」「製造」)

作成		作製		制作		製作		製造	
素性	頻度 (p 値)	素性	頻度 (p 値)	素性	頻度 (p 値)	素性	頻度 (p 値)	素性	頻度 (p 値)
素性 1:書	342 (1.32×10^{-165})	素性 1:細胞	1093 (≈ 0)	素性 1:番組	331 (2.90×10^{-172})	素性 1:映画	842 (≈ 0)	素性 2:業 (直後)	417 (≈ 0)
素性 1:案	157 (8.05×10^{-81})	素性 1:マップ	153 (1.82×10^{-95})	素性 1:ドラマ	130 (1.89×10^{-67})	素性 2:委員 (直後)	68 (1.26×10^{-42})	素性 1:工場	188 (2.44×10^{-101})
素性 1:文書	136 (8.93×10^{-80})	素性 1:地図	143 (1.07×10^{-75})	素性 2:共同 (直前)	86 (8.18×10^{-33})	素性 2:会 (2 語後)	69 (2.78×10^{-40})	素性 2:販売 (直後)	93 (7.30×10^{-61})
素性 1:報告	203 (4.59×10^{-72})	素性 1:遺伝子	72 (2.01×10^{-37})	素性 1:CM	61 (2.45×10^{-29})	素性 2:費 (直後)	89 (9.41×10^{-29})	素性 1:サリン	94 (4.70×10^{-59})
素性 1:計画	197 (2.56×10^{-60})	素性 1:図	95 (3.03×10^{-33})	素性 1:音楽	47 (2.11×10^{-15})	素性 2:新聞 (直前)	43 (3.29×10^{-24})	素性 1:部品	65 (2.13×10^{-32})

表 5.15: 機械学習が参考にした素性 (正規化 α 値:「作成」「作製」「制作」「製作」「製造」)

作成		作製		制作		製作		製造	
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	素性	正規化 α 値
素性 1:文書	0.62	素性 1:細胞	0.73	素性 1:番組	0.80	素性 1:映画	0.89	素性 1:工場	0.65
素性 1:書	0.59	素性 1:マップ	0.60	素性 1:ドラマ	0.66	素性 1:職人	0.35	素性 1:メーカー	0.46
素性 1:資料	0.48	素性 1:地図	0.54	素性 1:作品	0.43	素性 1:模型	0.34	素性 1:製品	0.40
素性 1:表	0.45	素性 1:ポスター	0.42	素性 1:美術	0.40	素性 1:新聞	0.33	素性 1:食品	0.38
素性 1:リスト	0.41	素性 1:チラシ	0.42	素性 1:アルバム	0.37	素性 1:はがき	0.30	素性 1:商品	0.37

「作成」は「文書」や「計画書」「報告書」などの書類を作る時に使われる．その他,「表」「リスト」「資料」にも使われる.「作製」は「細胞」「遺伝子」など生物学的なもの

を作る時に使われる．また,「マップ」「地図」などの図や「ポスター」,「チラシ」などにも使われる.「制作」は「番組」「ドラマ」「CM」などテレビ関係の映像を作る時に使われる．また，音楽や美術作品など芸術的なことにも使われることがわかった.「製作」は「映画」を作る時に使われる．他には「新聞」「模型」「はがき」などにも使われる.「製品」「商品」「食品」「部品」のように「品」のつくものや工場で作られているものに使われることがわかった.

また，文献 [5] によると，以下のように書かれている.

> 「製作」は，道具や機械などを用いて，物品を作ること.「制作」は，映画やテレビなどの番組を作ることや，そういう仕事，役割をいう.「作製」は，品物，機械，あるいは図面を作ることをいう.「作成」は書類や文書を作りあげる場合に使う.「製造」は原料を加工して大量に物を生産すること．主に，工場で生産するような場合に用いられる.

文献 [5] と素性分析の結果を比較し同じ結果となったことを以下に示す.「制作」は「ドラマ」などテレビ番組を作る時,「作製」は「マップ」などの図面を作る時,「作成」は「文書」,「計画書」などの書類を作る時,「製造」は「食品」「部品」など工場で大量に作るものに使われるということが同じであった.

文献に載っておらず，素性分析から新たにわかったことを以下に示す.「作製」は「細胞」,「遺伝子」を作る時に使われ,「作成」は「表」「リスト」を作る特に使われることなどが素性分析から新たにわかった.

文献に載っているが素性分析からは得られなかったことはなかった.

5.1.6 「予想」「予測」

（正解例 1）この判決に力を得て、不満の声がさらに噴き出すことも 予想 される。

（正解例 2）地震の発生確率を 予測 する政府の地震調査委員会。

（誤り例 1）その 予想(予測) 通り、翌日、近衛内閣を倒した東条陸相に組閣の大命が下った。

（誤り例 2）この区間を別ルートで通す新東名の開通で、今年の渋滞は２回と 予測(予想)、ほぼ解消されるとみている。

表 5.16 に類義語の組「予想」「予測」の機械学習の分類結果を示す．表 5.17 に類義語の組「予想」「予測」の有意差検定に基づいた機械学習が参考にした素性を示す．表 5.18 に類義語の組「予想」「予測」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 5.16: 機械学習の分類結果 (「予想」「予測」)

	データ数	再現率	適合率	F 値
予想	5000	0.81	0.82	0.81
予測	5000	0.82	0.81	0.81
総数	10000	0.81	0.81	0.81

表 5.17: 機械学習が参考にした素性 (有意差検定:「予想」「予測」)

予想		予測	
素性	頻度 (p 値)	素性	頻度 (p 値)
素性 2:さ (直後)	2034 (3.45×10^{-263})	素性 2:する (直後)	496 (1.10×10^{-49})
素性 2:れる (2 語後)	1430 (9.36×10^{-257})	素性 1:地震	373 (2.45×10^{-50})
素性 2:以上 (直後)	318 (3.00×10^{-94})	素性 1:システム	168 (1.90×10^{-33})
素性 2:外 (直後)	247 (2.43×10^{-65})	素性 1:データ	139 (2.01×10^{-23})
素性 1:業績	193 (6.09×10^{-35})	素性 2:拡散 (直前)	71 (4.24×10^{-22})

「予想」は「予想される」という表現が多いのに対し,「予測」は「予測する」という表現が多かった．また,「予想外」や「予想以上」のように直後に「外」「以上」がある場合,「予想」と書く使い分けが存在することがわかる.「予想」は「業績」「競馬」「値動き」など幅広く物事を推測する場合に使われ,「予測」は「地震」「豪雨」などの災害を推測することに使われやすい．また,「システム」「データ」「科学」などから「予測」は科学的根拠に基づく推測に使われるとわかる．

また，文献 [5] によると以下のように書かれている．

「予測」「予想」は将来のことを前もって見当をつけて，どうなるかを推測

表 5.18: 機械学習が参考にした素性 (正規化 α 値:「予想」「予測」)

予想		予測	
素性	正規化 α 値	素性	正規化 α 値
素性 1:相場	0.85	素性 1:在庫	0.85
素性 1:利益	0.84	素性 1:原発	0.84
素性 1:競馬	0.80	素性 1:科学	0.80
素性 1:値動き	0.74	素性 1:災害	0.76
デフォルト素性	0.57	素性 1:豪雨	0.75

することだが,「予測」のほうが，データなどに基づきより具体的であることが多い.

文献 [5] と素性分析の結果を比較し同じ結果となったことを以下に示す.「予想」は「システム」「データ」「値動き」のようにデータに基づいたことには「予測」を使うということが同じであった.

文献に載っておらず，素性分析から新たにわかったことを以下に示す.「予想」は「される」,「予測」は「する」というような表現が多かった．また,「予想」は「相場」「利益」「競馬」「値動き」など「お金」が関わっていることによく使われ,「予測」は「地震」「豪雨」などの災害を推測することに使われることが新たにわかった.

文献に載っているが素性分析からは得られなかったことはなかった.

5.1.7 「おおよそ」「おおむね」

（正解例 1）ご相談に応じさせていただきますので おおよそ の目安とお考えください。

（正解例 2）おおむね 好評でしたのでこれからもこういうスタイルで開催させていただきます。

（誤り例 1）イメージ作成は おおよそ(おおむね) ５分程度なのでこの時間を惜しんで無理をする必要はない。

（誤り例 2）一番重要なところは おおむね(おおよそ) まとまったのでまあ後は適当でいいか。

y 表 5.19 に類義語の組「おおよそ」「おおむね」の機械学習の分類結果を示す．o 表 5.20 に類義語の組「おおよそ」「おおむね」の有意差検定に基づいた機械学習が参考に

した素性を示す．表 5.21 に類義語の組「おおよそ」「おおむね」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 5.19: 「おおよそ」「おおむね」の分類結果

	データ数	再現率	適合率	F 値
おおよそ	937	0.88	0.91	0.90
おおむね	937	0.92	0.88	0.90
総数	1874	0.90	0.90	0.90

表 5.20: 機械学習が参考にした素性 (有意差検定:「おおよそ」「おおむね」)

おおよそ		おおむね	
素性	頻度 (p 値)	素性	頻度 (p 値)
素性 32:名詞	705 (1.46×10^{-8})	素性 32:動詞	594 (1.60×10^{-12})
素性 5:助詞	466 (1.24×10^{-129})	素性 32:形容詞	109 (3.32×10^{-5})
素性 2:の (直後)	417 (6.50×10^{-122})	素性 31:妥当だ	18 (3.81×10^{-5})
素性 2:の (直前)	78 (8.24×10^{-15})	素性 2:好評 (直後)	16 (1.53×10^{-5})
素性 31:見当	20 (9.54×10^{-7})	素性 2:満足 (直後)	10 (9.77×10^{-4})

「おおよそ」は「の」などの助詞を付属語にし,「目安」「見当」のような名詞が修飾先となることが多く,「おおむね」は「好評」「満足」「妥当」などの単語が後ろに来やすいことがわかった．また,「おおよそ」は「のおおよそ」と言う表現も多かった．修飾先の品詞にも特徴が見られた.「おおよそ」は修飾先が助詞,「おおむね」は修飾先が形容詞，動詞が多くなることがわかった．助詞の例は「が」「は」「を」が特に多かった.「おおむね」が修飾する形容詞は「良好だ」「妥当だ」「適切だ」など，動詞は「認める」などであった．また,「おおよそ」は助詞，名詞が被修飾先としても多く使われていた

また，文献 [5] には以下のように書かれている．

「おおよそ」は、全体をざっと見渡して大体の見当をいう.「大凡」とも

表 5.21: 機械学習が参考にした素性 (正規化 α 値:「おおよそ」「おおむね」)

おおよそ		おおむね	
素性	正規化 α 値	素性	正規化 α 値
素性 1:の	0.82	素性 2:対象語が文頭	0.66
素性 2:ので (直前)	0.76	素性 1:案	0.65
素性 1:目安	0.68	素性 31:認める	0.64
素性 1:話	0.68	素性 1:範囲	0.64
素性 1:距離	0.62	素性 1:一致	0.59

書く.「おおむね」は、全体の中の主な点を大きくつかむことをいう．概要の意味でも使われる.「概ね」とも書く.

文献 [5] と素性分析の結果を比較し同じ結果となったことを以下に示す.「おおよそ」は「目安」「見当」などの大体を表すという点で同じであった.「おおむね」は「案」などの概要を表すことという点で同じであった.

文献に載っておらず，素性分析から新たにわかったことを以下に示す.「おおよそ」は修飾先が助詞，名詞,「おおむね」は形容詞，動詞が多くなることが新たにわかった.

文献に載っているが素性分析からは得られなかったことはなかった.

5.1.8 「がっかり」「がっくり」

(正解例 1) 私は３００枚近い年賀状を調べましたが、当選はお年玉切手シートがもらえる４等の３枚だけで、少し がっかり しました。

(正解例 2) 三位を表示する電光掲示板に がっくり と肩を落とした。

(誤り例 1) おかしいやないか」と“ 抗議 ”する人もいるが、新型導入の説明を聞き がっかり(がっくり) しているという。

(誤り例 2) 工場誘致に期待をかけてきた旧東独ドレスデンの関係者は がっくり(がっかり)。

表 5.22 に類義語の組「がっかり」「がっくり」の機械学習の分類結果を示す．表 5.23 に類義語の組「がっかり」「がっくり」の有意差検定に基づいた機械学習が参考にした素性を示す．表 5.24 に類義語の組「がっかり」「がっくり」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 5.22: 「がっかり」「がっくり」の分類結果

	データ数	再現率	適合率	F 値
がっかり	285	0.84	0.82	0.83
がっくり	285	0.81	0.83	0.82
総数	570	0.82	0.82	0.82

表 5.23: 機械学習が参考にした素性 (有意差検定:「がっかり」「がっくり」)

がっかり		がっくり	
素性	頻度 (p 値)	素性	頻度 (p 値)
素性 31:する	230 (1.67×10^{-26})	素性 2:と (直後)	56 (1.39×10^{-17})
素性 2:し (直後)	168 (6.96×10^{-22})	素性 1:肩	57 (3.82×10^{-11})
素性 34:ようだ	11 (4.88×10^{-4})	素性 31:落とす	54 (4.07×10^{-13})
素性 2:少し (直前)	6 (1.56×10^{-2})	素性 31:くる	24 (5.96×10^{-8})
素性 2:ので (直前)	6 (1.56×10^{-2})	素性 2:き (直後)	22 (2.38×10^{-7})

「がっくり」は,「がっくりと肩を落とす」や「がっくり来た」「がっくりとくる」というような表現が多かった.「がっかり」は,「少し」「大変」など程度を表す語が前に来やすいことが分かった.

また，文献 [6] によると以下のように書かれている.

> 「がっかり」も「がっくり」もともに落胆する様子です.「がっかり」は，精神的な状態だけを表すので，その失望感が外からは見えにくい．それに対して,「がっくり」は，精神的な状態が首を垂れるとか肩を落とすなどの動作として外にまで現れています．だから,「がっかり」はすぐ立ち直れそうな感じがするのに対して,「がっくり」はなかなか立ち直れない感じがする.

文献 [6] と素性分析の結果を比較し同じ結果となったことを以下に示す.「がっくり」は「肩を落とす」などの動作として外にまで現れるという点が同じであった.

文献に載っておらず，素性分析から新たにわかったことを以下に示す.「がっかり」は「がっかりだ」や「少し」「大変」などの程度を表す語と併せて使われやすく,「がっく

表 5.24: 機械学習が参考にした素性 (正規化 α 値:「がっかり」「がっくり」)

がっかり		がっくり	
素性	正規化 α 値	素性	正規化 α 値
素性 7:だ	0.60	素性 1:肩	0.72
素性 2:する (直後)	0.56	素性 1:力	0.69
素性 2:本当に (直前)	0.55	素性 34:句点	0.69
素性 2:大変 (直前)	0.54	素性 2:気持ち (3 単語後)	0.65
素性 2:あんなに (直前)	0.52	素性 2:で (直前)	0.62

り」は直後に「と」が来ることが多く,「がっくり来る」という表現が多いことも新たにわかった.

文献に載っているが素性分析からは得られなかったことを以下に示す.「がっくり」の「首を垂れる」という表現は素性分析からはわからなかった.

5.1.9 「うろうろ」「ぶらぶら」

(正解例 1) お年寄りが好きな所を うろうろ することがなぜ、問題行動として拘束させられるんでしょうか。

(正解例 2) 買い物に費やす半日を、ぶらぶら 歩いたり、もっと自由に使えば、どんなにいいかと思うんですけど。

(誤り例 1) いつも二人で夜のドライブに出かけ、昼間は うろうろ(ぶらぶら) と過ごす。

(誤り例 2) 動性も変化し、不活発となることが多いが、逆に落ち着かず ぶらぶら(うろうろ) することもある。

表 5.25 に類義語の組「うろうろ」「ぶらぶら」の機械学習の分類結果を示す. 表 5.26 に類義語の組「うろうろ」「ぶらぶら」の有意差検定に基づいた機械学習が参考にした素性を示す. 表 5.27 に類義語の組「うろうろ」「ぶらぶら」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.

「うろうろ」は「うろうろする」という表現があるのに対し,「ぶらぶら」は「ぶらぶらと歩く」や「ぶらぶらと時間を潰す」など時間に余裕がある場合は「ぶらぶら」が使われやすいことが分かった. また,「うろうろ」が文末になることが多いのに対し,

表 5.25: 「うろうろ」「ぶらぶら」の分類結果

	データ数	再現率	適合率	F 値
うろうろ	124	0.64	0.63	0.63
ぶらぶら	124	0.62	0.63	0.63
総数	248	0.63	0.63	0.63

表 5.26: 機械学習が参考にした素性 (有意差検定:「うろうろ」「ぶらぶら」)

うろうろ		ぶらぶら	
素性	頻度 (p 値)	素性	頻度 (p 値)
素性 2:する (直後)	29 (5.84×10^{-5})	素性 31:歩く	22 (2.67×10^{-4})
素性 20:名詞	20 (2.04×10^{-3})	素性 1:歳	13 (3.69×10^{-3})
素性 7:句点	12 (3.84×10^{-2})	素性 2:歩き (直後)	9 (1.95×10^{-3})
素性 2:対象語が文末	12 (3.84×10^{-2})	素性 2:家 (2 単語前)	5 (3.13×10^{-2})
素性 31:ます	10 (9.77×10^{-4})	素性 1:生活	5 (3.13×10^{-2})

「ぶらぶら」は文頭になりやすい.「うろうろ」「ぶらぶら」は他の類義語の組と比べ, データ数が少なく，わかったことは少なかった.

また，文献 [6] によると以下のように書かれている.

> 「うろうろ」も「ぶらぶら」も，ともに辺りを徘徊する様子を表します. でも，それを見ている者が不審の念を抱くかどうかに違いがあります.
>
> 「うろうろ」は，どうして良いか分からず，ただむやみに動き回っている様子です．それを見ている者が,「何をしているんだろう」といぶかしく思う場合です.
>
> 一方,「ぶらぶら」は,「その年齢までまだ身もかため得ずにぶらぶらして居る」のように使います．行動する側に精神的・物理的な余裕があるので, それを見ている側も不審の念を抱かないのです.「ぶらぶら」は余裕のある徘徊です.

文献 [6] と素性分析の結果を比較し同じ結果となったことを以下に示す．素性分析か

表 5.27: 機械学習が参考にした素性 (正規化 α 値:「うろうろ」「ぶらぶら」)

うろうろ		ぶらぶら	
素性	正規化 α 値	素性	正規化 α 値
素性 1:駅	0.62	素性 1:買い物	0.63
素性 2:する (直後)	0.61	素性 1:時間	0.59
素性 1:境内	0.55	素性 2:対象語が文頭	0.58
素性 1:座席	0.54	素性 1:街	0.57
素性 31:動く	0.54	素性 1:わき道	0.56

ら「うろうろする」や「うろうろ動く」のようにどうして良いか分からず，ただむやみに動き回っている様子を表しているということがわかり文献と同じであった.「ぶらぶら」は「歩く」「買い物」「時間」などから余裕のある徘徊であることがわかり，また,「歳」「家」などの素性から「その年齢までまだ身もかため得ずにぶらぶらして居る」ということがわかり文献と同じであった.

文献に載っておらず，素性分析から新たにわかったことはなかった.

文献に載っているが素性分析からは得られなかったこともなかった.

5.1.10 「はっきり」「きっぱり」

（正解例 1）これまで明らかでなかった中国側の経緯が はっきり 分かる。

（正解例 2）センターは「購入の意思がなければ きっぱり 断ること」と助言している。

（誤り例 1）小さな声だが、はっきり(きっぱり) 決意を口にした。

（誤り例 2）これは旧暦の七月九日と きっぱり(はっきり) 日時が定まっている。

表 5.28 に類義語の組「はっきり」「きっぱり」の機械学習の分類結果を示す．表 5.29 に類義語の組「はっきり」「きっぱり」の有意差検定に基づいた機械学習が参考にした素性を示す．表 5.30 に類義語の組「はっきり」「きっぱり」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.

「はっきり」は「はっきりしない」という表現や「分かる」「示す」「見える」などの語が後に続くことが多い.「きっぱり」は「言い切る」「否定」「断る」のような語が後に続くことが多いとわかった．特に「断ち切る」「割り切る」のように「〜切る」という表現が多いことから「きっぱり」は「はっきり」に比べ強い意味合いで使われる

表 5.28: 「はっきり」「きっぱり」の分類結果

	データ数	再現率	適合率	F 値
はっきり	702	0.94	0.92	0.93
きっぱり	702	0.92	0.94	0.93
総数	1404	0.93	0.93	0.93

表 5.29: 機械学習が参考にした素性 (有意差検定:「はっきり」「きっぱり」)

はっきり		きっぱり	
素性	頻度 (p 値)	素性	頻度 (p 値)
素性 2:し (直後)	274 (1.18×10^{-57})	素性 2:対象語が文末	386 (1.20×10^{-112})
素性 2:ない (2 単語後)	96 (1.53×10^{-26})	素性 2:言い (直後)	32 (3.96×10^{-9})
素性 31:分かる	25 (2.98×10^{-8})	素性 2:否定 (直後)	31 (7.68×10^{-9})
素性 31:示す	20 (9.54×10^{-7})	素性 31:断る	18 (7.45×10^{-4})
素性 31:見える	19 (1.91×10^{-6})	素性 2:切っ(2 単語後)	17 (7.63×10^{-6})

語であることがわかった.「はっきり」幅広く使われ「きっぱり」の方が限定的な使われ方がされることが考察できた.

また，文献 [6] によると以下のように書かれている.

> 「きっぱり」と「はっきり」は，どちらも明確で確かな様子を表します. でも，態度に重点があるのか，発言内容に重点があるのかに違いがあります.「きっぱり」は「尋ねて来たら，きっぱりことわれば好い」（森鴎外『青年』）のように用います．断る態度が決然としているのです．取り付く島がありません.「きっぱり」は,「言う」「断る」「否定する」などの態度しか形容しません．いずれも，拒絶的な態度の形容です．だから,「きっぱり」の態度に出会うと，突き放されたようなショックを受けます．　一方,「はっきり」は,「洋食なら洋食，お蕎麦ならお蕎麦と，尋ねられればハッキリと喰べたい物を答えました.」（谷崎潤一郎『痴人の愛』）のように用います. 別に答える態度が決然としているわけではありません．他との区別が明ら

表 5.30: 機械学習が参考にした素性 (正規化 α 値:「はっきり」「きっぱり」)

はっきり		きっぱり	
素性	正規化 α 値	素性	正規化 α 値
素性 1:意見	0.64	素性 2:言い (直後)	0.68
素性 2:言う (直後)	0.64	素性 2:口調 (3 単語後)	0.66
素性 2:対象語が文頭	0.62	素性 31:あきらめる	0.55
素性 1:区別	0.54	素性 31:断れる	0.55
素性 31:説明	0.54	素性 31:断ち切る	0.54

かで確かな様子が発言内容から判断できるのです.「はっきり」は「言う」「断る」「否定する」などの動作の形容にも用いますが,「きっぱり」と違って，つねに発言の内容がほかと区別できるような明快さを持っているときです.「はっきり」はその他「分かる」「見える」「聞こえる」などの動作の形容にも使えます．いずれも内容が他と紛れることなく，確かに識別できるときに用います.

文献 [6] と素性分析の結果を比較し同じ結果となったことを以下に示す.「はっきり」は「分かる」「見える」など形容する語が幅広くあり,「きっぱり」は「断る」「否定する」などの拒絶的な態度を形容する時にしか使われないということが同じであった.

文献に載っておらず，素性分析から新たにわかったことを以下に示す.「はっきり」は「示す」,「きっぱり」は「あきらめる」という語が後ろにくることが多いことが新たにわかった．また，文頭は「はっきり」が，文末は「きっぱり」が多いことがわかった.

文献に載っているが素性分析からは得られなかったことを以下に示す.「はっきり」の素性分析から「聞こえる」という素性はなかった．また，文献と違った点としては，文献には「きっぱり」は「言う」を形容するとあるが，素性分析からは「言う」は「はっきり」の後に続くことが多く,「きっぱり」は「言う」よりも「言い切る」という表現が多かったという結果となった.

5.1.11 「うっすら」「ぼんやり」

（正解例 1）十九日朝、富士山山頂の火口付近が うっすら 雪化粧し、初雪が観測された。

（正解例 2）輪郭は情けないくらい ぼんやり している。

（誤り例 1）「近畿の水がめ」の湖面には柔らかにかすみがかかり、神が宿るという竹生島が、うっすら(ぼんやり) と見えた。

（誤り例 2）島北部の丘陵にある韓国軍哨所に立つと、北朝鮮の山並みがぼんやり(うっすら) と見渡せる。

表 5.31 に類義語の組「うっすら」「ぼんやり」の機械学習の分類結果を示す．表 5.32 に類義語の組「うっすら」「ぼんやり」の有意差検定に基づいた機械学習が参考にした素性を示す．表 5.33 に類義語の組「うっすら」「ぼんやり」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 5.31: 「うっすら」「ぼんやり」の分類結果

	データ数	再現率	適合率	F 値
うっすら	397	0.83	0.83	0.83
ぼんやり	397	0.83	0.83	0.83
総数	794	0.83	0.83	0.83

表 5.32: 機械学習が参考にした素性 (有意差検定:「うっすら」「ぼんやり」)

うっすら		ぼんやり	
素性	頻度 (p 値)	素性	頻度 (p 値)
素性 2:と (直後)	277 (4.52×10^{-9})	素性 2:し (直後)	100 (1.03×10^{-27})
素性 1:涙	68 (2.31×10^{-16})	素性 2:て (2 単語後)	99 (1.20×10^{-20})
素性 31:浮かべる	47 (7.11×10^{-15})	素性 2:いる (3 単語後)	39 (3.72×10^{-17})
素性 1:雪化粧	40 (9.09×10^{-13})	素性 31:眺める	25 (2.98×10^{-8})
素性 31:にじむ	11 (3.17×10^{-3})	素性 31:見る	24 (5.96×10^{-8})

「うっすら」は「うっすらと涙を浮かべる」という表現が多かった．修飾先は「浮かべる」の他に「見える」「光る」「にじむ」などであった．また「雪化粧」「色」「白」など色を表す時に「うっすら」が使われることが多かった.「ぼんやり」は「ぼんやりしている」という表現や「頭がぼんやりする」や「輪郭」がぼんやりしているのよう

表 5.33: 機械学習が参考にした素性 (正規化 α 値:「うっすら」「ぼんやり」)

うっすら		ぼんやり	
素性	正規化 α 値	素性	正規化 α 値
素性 31:見える	0.66	素性 1:頭	0.66
素性 1:色	0.63	素性 1:輪郭	0.62
素性 31:光る	0.62	素性 1:夢	0.59
素性 1:白	0.60	素性 31:抱く	0.59
素性 1:霧	0.60	素性 31:かすむ	0.58

な使われ方が多い.「ぼんやり」の修飾先としては「眺める」「抱く」「かすむ」などが多かった.「うっすら」は修飾先が「見える」が多かったことに対し,「ぼんやり」は修飾先が「見る」が多かったことから「うっすら」が「ぼんやり」よりも見える状態であることを表していることがわかった.

また，文献 [6] によると以下のように書かれている.

> 「うっすら」も「ぼんやり」も，ともに物が鮮明に見えない状態を表します．でも，その見え方に違いがあります.「うっすら」は，わずかですが，物の輪郭などが見える状態です．たとえば,「西空にうっすらと三日月が，はりついていた」(海野十三『海底都市』) のように使います．薄くかすかではありますが，三日月がでています.「薄い」の語と関係があることから,「覚えている」「見える」「感じる」など知覚のわずかである場合に使います．でも，かすかでも姿や像は見えているのです．同じく「薄い」から生まれた「うすら」「うすら笑い」「うすら寒い」などの形で使われますが，かすかという意味ですね．一方,「ぼんやり」は，輪郭自体もはっきりしない場合です.「雪が止むと，雲の間から薄明かりが漏れた．ぼんやりと地形を見ることができた」(新田次郎『八甲田山死の彷徨』) のように使います．暗くて物の形や輪郭がはっきりせずにかすんでいるのです．像を明確に結ばない事から,「思考力もなにも失ってしまって，ただもう，ボンヤリしていた」(江戸川乱歩『人間椅子』) のように，意識が十分に働かず，集中できてない様子を表します．さらに，そういう人は，気が利かない．だから「昼行灯と渾名されているほどのぼんやり者」(朝松健『元禄霊異伝』) のように，気が利かないという性質を表します.「ぼんやり」見えたのより,「うっすら」見えた方が頼りになります.

文献 [6] と素性分析の結果を比較し同じ結果となったことを以下に示す.「うっすら」は「うっすらと見える」など知覚がわずかである場合に使い,「ぼんやり」は「頭がぼんやりする」のように意識が十分に働かず，集中できてない様子を表しますことが同じであった.

文献に載っておらず，素性分析から新たにわかったことを以下に示す.「うっすら」は直後に「と」が多く,「雪化粧」や「霧」を表す時に使い，修飾先には「浮かべる」「見える」「光る」などが多いことがわかった．また,「ぼんやり」は修飾先としては「眺める」「抱く」「かすむ」などが多いことが新たにわかった.

文献に載っているが素性分析からは得られなかったことを以下に示す.「覚えている」「感じる」は「うっすら」の素性分析からはわからなかった.

5.2 品詞間における類義語の使い分けに関する特徴

本研究では，品詞間における類義語の使い分けに関する特徴も得られた.

まず，名詞の類義語では文中の名詞が使い分けに関し，重要な情報となることが多いことがわかった．例として,「予想」「予想」という名詞の類義語の組では，表 5.18 より「予想」の有用な素性としては「相場」「利益」「競馬」,「予測」の有用な素性は「在庫」「科学」「災害」などのように文中の名詞が多くあった．このように文中の名詞から類義語の使い分けに関する傾向がわかる.

また，副詞の類義語では修飾先の語，または修飾先の語の品詞が使い分けに関し，重要な情報になることが多いことがわかった．例として,「おおよそ」「おおむね」という副詞の類義語の組では，表 5.20 より,「おおよそ」は修飾先が名詞になることが多く,「おおむね」は修飾先が形容詞，名詞となることが多いことがわかった．このように副詞の類義語は修飾先の語から使い分けに重要な情報が多いことがわかる.

先行研究の強田ら [1] の研究では「貯金」「貯蓄」という名詞の類義語対の使い分けにおいて有用だった素性について「貯金」の素性では「定額」「郵便」「郵政省」,「貯蓄」の素性では「投資」「米国」などであり，文中の名詞から使い分けの考察を行なっている．また，中瀬 [2] の研究では「きわめて」「だいぶ」という副詞の類義語対の使い分けにおいて有用だった素性について「きわめて」の素性では修飾先の最後が形容詞,「だいぶ」の素性では修飾先の最後が動詞などであり，修飾先の品詞から使い分けの考察を行なっている．他の類義語でも同じような傾向であった．このように強田，中瀬らの研究からも，本研究から得た品詞間における類義語の使い分けに関する特徴の傾

向が確認できた.

第6章　機械学習を用いた文章の誤り訂正実験と考察

本章ではWebの文章に誤った文章があることを仮定し，新聞の文章を学習データ，Webの文章をテストデータとして機械学習を行うことでWebの文章の誤り訂正に利用できると考えその実験と考察を示す．

6.1　「おかげ」「せい」「ため」での実験

クロスバリデーションでの正解率が9割以上だった類義語「おかげ」「せい」「ため」の新聞データを学習データとし，Webデータをテストデータにして実験を行った．表6.1に新聞データでのデータ数とクロスバリデーションでの正解率の結果を示す．

表 6.1: 「おかげ」「せい」「ため」の正解率 (新聞データ)

	データ数	正解率
おかげ	4186	0.90
ため	4186	0.95
せい	4186	0.87
総数	12558	0.90

表6.2にWebデータでのデータ数と新聞データを学習データとした時の元の文章の正解率を示す．正解率が0.88という推定結果となった．この実験は誤り訂正と見なすことができ，機械の推定結果が元データの分類と異なっている場合，機械の推定のように修正することで誤り訂正を実現できる．この考え方に基づき表6.2の結果を誤り訂正として考察した．

表 6.2: 「おかげ」「せい」「ため」の正解率 (Web データ)

	データ数	正解率
おかげ	500	0.89
ため	500	0.85
せい	500	0.92
総数	1500	0.88

表 6.3 に．機械の出力と元の文の語が異なっている文の出力例を示す．出力の形式は「機械が出した答え，○か×，元の文の語，　…．@元の文」となっている．

表 6.3: 機械の出力と元の文の語が異なっている文の出力例 (「おかげ」「せい」「ため」)

せい, ×, おかげ,0.9648,0.0349, おかげ@184@184@syosi1@syosi2@syosi3@そのおかげかどうか知らないけど、明日の休出は声かからなかったのでよかったですな。
せい, ×, おかげ,0.9701,0.0299, おかげ@610@610@syosi1@syosi2@syosi3@薬のおかげで、ぶり返した咳も治まったので大丈夫だろうとは思うけど。
おかげ, ×, せい,0.9439,0.0561, せい@382@382@syosi1@syosi2@syosi3@前作がとても面白かったのでそのせいで、ＩＩがつまらなく思えました。
おかげ, ×, せい,0.9893,0.0106, せい@439@439@syosi1@syosi2@syosi3@年末に休みをもらったせいで、去年のうちに終わらせるべき仕事が少し残っている状態でしたので。
ため, ×, おかげ,0.9817,0.0110, おかげ@307@307@syosi1@syosi2@syosi3@昨日の疲れと風邪気味な体調のおかげでかなり辛い、でも仕事はあるのでなんとか出社。
おかげ, ×, ため,0.9521,0.0358, ため@923@923@syosi1@syosi2@syosi3@でも、私は不利になる情報を与える目的では行なっていませんので念のため。
ため, ×, せい,0.7553,0.2446, せい@1471@1471@syosi1@syosi2@syosi3@情けないが全て自分の金銭管理が甘いせいなのでなにもいえない。
せい, ×, ため,0.9580,0.0420, ため@146@146@syosi1@syosi2@syosi3@完全予約制を取っているためか呼び込みとかないので入りづらい面はあります。

機械の出力のように訂正する場合より，元の Web の文章のほうが正しいことがほとんどであり，この実験はうまく行かなかった．

また，学習データ，テストデータを同じにして Web データ 12,000 文で実験を行い，その結果 2 文が機械の出力と元の文が異なっている文であったのでその例を表 6.4 に示す．

元の文の語，機械の出力した語のどちらでも良いような文章なので誤り訂正ができたとは言えない．

表 6.4: 学習データとテストデータを同じにした実験の機械の出力と元の文の語が異なっている文の出力例 (「おかげ」「せい」「ため」)

せい, ×, おかげ,0.5192,0.4773, おかげ@1413@1413@syosi1@syosi2@syosi3@酔っ払ったときはいつも近くの教会で「なーむー」とお参りをするのでそのおかげかもしれない。
せい, ×, ため,0.4989,0.4965, ため@11591@11591@syosi1@syosi2@syosi3@今回はビデオ収録があったみたいなのでそのためかとも思うけど。

6.2 「場合」「時」「際」での実験

クロスバリデーションでの正解率が約 0.83 であった類義語「場合」「時」「際」の新聞データを学習データとし，Web データをテストデータにして実験を行った．表 6.5 に新聞データでのデータ数とクロスバリデーションでの正解率の結果を示す.

表 6.5: 「場合」「時」「際」の正解率 (新聞データ)

	データ数	再現率
場合	3946	0.88
時	3450	0.81
際	2604	0.80
総数	10000	0.83

表 6.6 に Web データでのデータ数と新聞データを学習データとした時の元の文章の正解率を示す．正解率が 0.71 という推定結果となった．同様に表 6.6 の結果を誤り訂正として考察した.

表 6.6: 「場合」「時」「際」の正解率 (Web データ)

	データ数	正解率
場合	1000	0.89
時	1000	0.66
際	1000	0.59
総数	3000	0.71

表 6.3 に．機械の出力と元の文の語が異なっている文の出力例を示す．出力の形式は「機械が出した答え，○か×，元の文の語，　…．@元の文」となっている．

表 6.7: 機械の出力と元の文の語が異なっている文の出力例 (「場合」「時」「際」)

場合, ×, 時,0.8926,0.0991, 時@1@1@syosi1@syosi2@syosi3@が、流石にＴＲＰＧをやると喉に負担掛けてしまうのでこの前の日曜の宇宙堂コンの時は顔だけ出しました。
時, ×, 際,0.9126,0.0579, 際@8@8@syosi1@syosi2@syosi3@屋根のある広場があるので雨の際にも都合がいい。
時, ×, 際,0.8143,0.1557, 際@17@17@syosi1@syosi2@syosi3@交通費とか郵送費はコインの入手の際には必要経費ですのでコイン収集を楽しむための税金とお考えください。
場合, ×, 際,0.9310,0.0586, 際@18@18@syosi1@syosi2@syosi3@オンライン・サインアップの際は説明書によると、サインアップ中の接続料と通話料は無料とのことですので時間をかけて落ち着いて登録しましょう。

「おかげ」「せい」「ため」の実験と同様に機械の出力のように訂正する場合より，元の Web の文章のほうが正しいことがほとんどであり，この実験はうまく行かなかった．

第7章　おわりに

本研究では機械学習を用いて類義語の使い分けの研究を行った．本研究の成果は2つある．

第1の成果として，類義語11組について実験を行った結果，正解率のマクロ平均は「データ数を出現率に合わせた実験」では，提案手法が0.84，ベースライン手法が0.65，素性2のみの手法では0.82であり，「データ数を同数に揃えた実験」では，提案手法が0.81，ベースライン手法が0.42，素性2のみの手法では0.78であったため，機械学習を用いる提案手法の正解率が最も頻度の高い語を常に選択するベースライン手法と機械学習の素性を前後の3単語のみとした手法よりも，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，類義語の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，「作成」は「表」「リスト」などを作る時に使われ，「作製」は「細胞」「遺伝子」などを作る時に使われることが多い．また，「おおよそ」は助詞の付属語を伴い，修飾先が名詞の時に使われることが多いのに対し，「おおむね」は形容詞が修飾先の時に使われることが多いなどの文献[5][6]に載っていないような使い分けに関する情報が多く得られた．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, 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，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] 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.

[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, pp. 300–303, 2011.

[5] 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.

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

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

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

[9] 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.

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

[11] 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.