

## 概要

自然言語処理における重要な問題の一つに、多義性解消がある。多義性解消とは、多義語(複数の語義を持つ語)が文中に出現したときに、その多義語の語義を、1つの語義に絞ることをいう。多義性解消は、翻訳や知識獲得に役立つ。先行研究[1]に言い換えを使った研究がある。言い換えを使うことで、多義性解消の誤りの原因である、学習データの不足を改善することができる。先行研究では、様々な手法を用いているが、単語ごとに適した方法を使用できていなかったという問題がある。

そこで、本研究は、様々な手法がある中、単語ごとに適した手法を自動的に選択するために10分割クロスバリデーションを用いる。10分割クロスバリデーションを用いることで、単語ごとに適した方法を使用できていなかったという問題が解決できる。10分割クロスバリデーションを用いて正解率の向上を試みる。

また、別の実験として SemEval2[2] の学習データ数を言い換えによって増えたデータ数の  $X$  倍(ここでは  $1/10, 1/5, 1/2, 1, 2, 5, 10$ ) のデータ数に変更することにより、性能の変化を調査する。SemEval2[2] は、多義性解消のコンテストで用意されたものであり、多義性解消の研究や実験を行いやすいように人手で作成されたものである。また、その対象の多義語は50個あり、多義語1語につき、学習データとテストデータがそれぞれ50個ずつ用意されている。

また、もう一つ別の実験として先行研究[1]で扱っていなかった動詞と形容詞の実験を行い、言い換えによって増えたデータを追加後、正解率の変化を調査する。

実験の結果、名詞の実験では、「手法1:SemEval2の学習データのみを利用する手法」の正解率が0.733に対して、「手法5:クロスバリデーションによる手法選択」の正解率が0.757となっており、正解率が向上した。また、「手法5:クロスバリデーションによる手法選択」の正解率0.757が最も良い正解率となった。

また、多義語6単語で合計12回の実験を行ったが、12回の実験中10回は正しく手法を選択することができた。

SemEval2の学習データ数を言い換えによって増えたデータ数の  $X$  倍( $1/10, 1/5, 1/2, 1, 2, 5, 10$ ) のデータ数に変更する実験では、「手法2:SemEval2と言い換えによって増え

たデータをを用いる手法」は、SemEval2の学習データ数が増えるにつれて性能が向上する傾向がみられた。

動詞の実験では、「手法1:SemEval2の学習データのみを利用する手法」の正解率0.913に対して、「手法2:SemEval2の学習データ+言い換えによって増えた学習データを利用する手法」の正解率0.893となり、正解率が低下した。

形容詞の実験では、「手法1:SemEval2の学習データのみを利用する手法」の正解率が0.90に対して、「手法2:SemEval2の学習データ+言い換えによって増えた学習データを利用する手法」の正解率が0.92となり、正解率はほぼ変わらなかった。

# 目次

第1章	はじめに	1
第2章	先行研究	3
2.1	多義性解消の誤りの原因	3
2.2	単義の同義語を利用した英語単語の学習データの増やし方	3
2.3	言い換えと機械学習を用いた日本語単語の多義性解消	4
2.4	日本語単語の多義性解消における種々の機械学習手法と素性の比較	4
2.5	日本語単語の多義性解消のための学習データの自動拡張	4
第3章	本研究の手法	5
3.1	本研究の多義性解消の方法	5
3.2	10分割クロスバリデーション	5
3.3	SemEval2の学習データ数を変更	6
3.4	言い換えを利用した学習データの増やし方	6
3.5	最大エントロピー法	8
第4章	実験	10
4.1	名詞の多義語での実験	10
4.2	単語の選定(名詞)	12
4.3	動詞の多義語での実験	13
4.4	単語の選定(動詞)	14
4.5	形容詞の多義語での実験	16
4.6	単語の選定(形容詞)	16
第5章	実験結果	18
5.1	名詞の多義語での実験結果	18
5.1.1	名詞の実験に使用したデータ	18

5.1.2	多義性解消の結果 . . . . .	21
5.1.3	有意差検定 . . . . .	27
5.1.4	考察 (クロスバリデーションによる手法選択) . . . . .	29
5.1.5	正解例と誤り例 . . . . .	30
5.1.6	SemEval2 の学習データ数を変更 . . . . .	34
5.1.7	考察 (SemEval2 の学習データ数を変更) . . . . .	35
5.2	動詞の多義語での実験結果 . . . . .	36
5.2.1	動詞の実験に使用したデータ . . . . .	36
5.2.2	考察 . . . . .	38
5.3	形容詞の多義語での実験結果 . . . . .	39
5.3.1	形容詞の実験に使用したデータ . . . . .	39
5.3.2	考察 . . . . .	40
第 6 章	今後の課題	41
第 7 章	おわりに	42

# 表 目 次

3.1 「内容」を含む文の例 . . . . .	7
3.2 「動機」を含む文の例 . . . . .	7
3.3 「価値」を含む文の例 . . . . .	8
3.4 言い換える前と言い換えた後の文 . . . . .	8
4.1 使用した素性 . . . . .	11
5.1 semeval2 の学習データ (名詞) . . . . .	19
5.2 semeval2 のテストデータ (名詞) . . . . .	19
5.3 semeval2 の学習データ (ランダム) . . . . .	19
5.4 semeval2 のテストデータ (ランダム) . . . . .	20
5.5 semeval2 の学習データ (ランダムにして交換) . . . . .	20
5.6 semeval2 のテストデータ (ランダムにして交換) . . . . .	20
5.7 言い換えによって増えたデータ数 (名詞) . . . . .	20
5.8 言い換えによって増えたデータ数 (名詞):データ数調整 . . . . .	21
5.9 semeval2 の学習データのみの評価 (ランダム) . . . . .	21
5.10 semeval2 のテストデータのみの評価 (ランダム) . . . . .	22
5.11 semeval2 の学習データのみの評価 (ランダムにして交換) . . . . .	23
5.12 semeval2 のテストデータのみの評価 (ランダムにして交換) . . . . .	24
5.13 semeval2 の学習データのみの評価 (名詞) . . . . .	25
5.14 semeval2 のテストデータのみの評価 (名詞) . . . . .	26
5.15 semeval2 のテストデータでの評価:6 単語全て . . . . .	27
5.16 有意差検定結果 . . . . .	28
5.17 手法 1 : 不正解で手法 5 : 正解の問題数 . . . . .	30
5.18 手法 1 : 正解で手法 5 : 不正解の問題数 . . . . .	31
5.19 手法 1 : 不正解で手法 5 : 正解の例 . . . . .	32
5.20 手法 1 : 正解で手法 5 : 不正解の例 . . . . .	33

5.21 semeval2 の学習データ数を変更 (ランダム) . . . . .	34
5.22 semeval2 の学習データ数を変更 (ランダム交換) . . . . .	35
5.23 semeval2 の学習データ (動詞) . . . . .	36
5.24 semeval2 のテストデータ (動詞) . . . . .	36
5.25 言い換えによって増えたデータ数 (動詞) . . . . .	37
5.26 利用する学習データとその正解率 (動詞) . . . . .	37
5.27 手法と正解率:5 単語全て . . . . .	38
5.28 SemEval2 のテストデータ (形容詞) . . . . .	39
5.29 SemEval2 の学習データ (形容詞) . . . . .	39
5.30 言い換えによって増えたデータ (形容詞) . . . . .	39
5.31 利用する学習データとその正解率 (形容詞) . . . . .	40

# 第1章 はじめに

現在，自然言語処理における重要な問題の一つに，多義性解消がある．多義性解消とは，多義語（複数の語義を持つ語）が文中に出現したときに，その多義語の語義を，1つの語義に絞ることをいう．多義性解消は，翻訳や知識獲得に役立つ．先行研究 [1] に言い換えを使った研究がある．言い換えを使うことで，多義性解消の誤りの原因である，学習データの不足を改善することができる．先行研究では，様々な手法を用いているが，単語ごとに適した方法を使用できていなかったという問題がある．

そこで，本研究は，様々な手法がある中，単語ごとに適した手法を自動的に選択するために10分割クロスバリデーションを用いる．10分割クロスバリデーションを用いることで，単語ごとに適した方法を使用できていなかったという問題が解決できる．そして，10分割クロスバリデーションを用いて正解率の向上を試みる．

また，別の実験として SemEval2[2] の学習データ数を言い換えによって増えたデータ数の X 倍（ここでは 1/10, 1/5, 1/2, 1, 2, 5, 10）のデータ数に変更することにより，性能の変化を調査する．SemEval2[2] は，多義性解消のコンテストで用意されたものであり，多義性解消の研究や実験を行いやすいように人手で作成されたものである．また，その対象の多義語は 50 個あり，多義語 1 語につき，学習データとテストデータがそれぞれ 50 個ずつ用意されている．

また，もう一つ別の実験として先行研究 [1] で扱っていなかった動詞と形容詞の実験を行い，言い換えによって増えたデータを追加後，正解率の変化を調査する．

本研究の主な主張点を以下に整理する．

- 10分割クロスバリデーションの実験では，複数手法がある中，クロスバリデーションによる手法選択が一番良い正解率となった．
- SemEval2 の学習データ数を言い換えによって増えたデータ数の X 倍 (1/10, 1/5, 1/2, 1, 2, 5, 10) のデータ数に変更する実験では，「手法 2: SemEval2 と言い換えによって増えたデータを用いる手法」は，SemEval2 の学習データ数が増えるにつれて性能が向上する傾向がみられた．

本論文の構成は以下の通りである．第2章では，先行研究について述べる．第3章では，本研究の手法について述べる．第4章では，本研究の実験について述べる．第5章では，本研究の実験結果について述べる．第6章では，今後の課題について述べる．第7章では，本研究の簡単なまとめを述べる．

## 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 多義性解消の誤りの原因

新納ら [3] は語義曖昧性解消の誤り原因のタイプ分けについて述べた．7名のメンバーが各自誤り分析を行い，誤り原因のタイプ分けを行った．各自の分析結果を人手で統合することは，困難であった．そこで統合処理を行うため，誤り原因（計75個）をベクトル化し，それらのクラスタリングを行った．クラスタリング結果を微調整することで誤り原因のタイプ分けを行った．誤り原因の主要な3つにより，語義曖昧性解消の誤りの9割が生じていることがわかった．その誤りの9割が生じている3つの原因は「訓練データの不足」「深い意味解析が必要」「シソーラスの問題」であった．分析対象が50事例ある中，「訓練データの不足」は36事例があてはまる．

### 2.2 単義の同義語を利用した英語単語の学習データの増やし方

Mihalcea[4]らは，単義の同義語を利用し，学習データを自動獲得する方法を提案した．英語の概念辞書（意味辞書）の同義語のうち，単義語や，定義文の一部を利用しWeb検索を行い，獲得したスニペット中の対象語に語義を付与し，テストデータに追加した．スニペットとは，検索エンジンによる検索結果の一部として表示される、Webページの要約文のことである。この方法であれば，テストデータにしか出現しない語義は，同義語を用いた訓練データの拡張を行うことで，推定できる可能性がある．

## 2.3 言い換えと機械学習を用いた日本語単語の多義性解消

戸田ら [1] は新納ら [3] の研究によって明らかとなった，多義性解消の誤りの原因の一つである「学習データの不足」に着目した．学習データを増やすため，言い換えを使った手法を提案した．言い換えを利用することで自動で学習データが増え，その学習データを利用することで，学習データを追加する前より性能が上がった．

## 2.4 日本語単語の多義性解消における種々の機械学習手法と素性の比較

村田ら [5] は，2001 年に行われた SENSEVAL2 コンテストの日本語辞書タスクでの取り組みについて述べた．村田らは，機械学習手法を用いるアプローチを採用した．数多くの機械学習手法と素性を比較検討し用いた．また，素性を変更した実験を行い，各素性の有効性，特徴を調査した．その結果，文字列素性のみを用いても比較的高い精度が得られるなどの興味深い知見が得られた．

## 2.5 日本語単語の多義性解消のための学習データの自動拡張

藤田ら [6] は，訓練データの自動拡張による多義性解消の精度向上方法について述べた．評価対象として，SemEval-2010 日本語語義曖昧性解消タスクを利用した．辞書の例文，配布データ以外のセンスバンク，ラベルなしコーパスなど，さまざまなコーパスを利用して，訓練データの自動拡張を試みた．実験の結果，異なる品詞体系，異なる辞書（語義）に基づいて構築されたセンスバンクであっても，自動的に学習データに追加することができることがわかった．また，精度向上に有効であることがわかった．

## 第3章 本研究の手法

### 3.1 本研究の多義性解消の方法

本研究では、先行研究 [4] で提案されている手法と類似した手法を用いる。言い換えと機械学習を利用して多義性解消を行う手法を用いる。多義性解消の入力は、多義語を含む文であり、出力は、複数ある語義のうち、どの意味で使われたかとする。先行研究 [5] と同様に学習データを用いた教師あり機械学習により、多義語の語義を1つに絞る。しかし、学習データが少ない場合、多義性解消を誤りやすい。そこで本研究では、言い換えを利用して学習データを自動で増やし、その増やした学習データを利用する [4]。言い換えを利用して学習データを増やすには、対象の多義語の類義語を含む文を抜き出し、その類義語を対象の多義語に言い換えることにより学習データを増やすことができる。学習に使用する素性は48種類で、文構造や文中にある単語などを素性とする。機械学習には最大エントロピー法を利用する。

### 3.2 10分割クロスバリデーション

本研究は様々な手法を用いているので、単語ごとに適した手法を自動的に選択する手法を用いる。具体的には、「手法1:SemEval2の学習データのみを用いる手法」「手法2:SemEval2と言い換えによって増えたデータを用いる手法」「手法3:言い換えによって増えたデータのみを用いる手法」「手法4:最頻出語義を常に出力する手法」の4つの手法を用いる。これらの手法ごとに学習データのみでの正解率を求める。ここでいう正解率とは、多義性解消の正解率である。学習データのみでの正解率を求める際10分割クロスバリデーションを用いる。10分割クロスバリデーションは、学習データを10個に分割し、そのうちの1つをテストデータとし、それ以外の9つを学習データにする。10分割されたデータがそれぞれテストデータとなるように10回繰り返す。そうすることで手法ごとの学習データ全体の傾向を知ることができる。そしてSemEval2の学習データでの10分割クロスバリデーションにより4手法の中で一番正解率が良かつ

た手法を自動選択する．本研究ではこの手法を「手法 5:クロスバリデーションによる手法選択」とする．

### 3.3 SemEval2 の学習データ数を変更

本研究は，10 分割クロスバリデーションとは別に，実験を行った．SemEval2 の学習データ数を言い換えによって増えたデータ数の  $X$  倍 (ここでは  $1/10, 1/5, 1/2, 1, 2, 5, 10$ ) のデータ数に変更することにより，性能の変化を調査する．

### 3.4 言い換えを利用した学習データの増やし方

言い換えを利用した学習データの増やし方を説明する．多義語を  $X$  とし，ここでは，その多義語  $X$  は語義を  $N$  個持つものとする．

- 手順 1 多義語  $X$  の語義ごとにその語義を特徴付ける語を人手で選定する．この選定では，語義の定義文中の語を参考に行っている．定義文中の語を選定するケースが多いが，定義文にはないが定義文から人が思いつく語を選定するケースもある．
- 手順 2 辞典の  $n$  番目の語義を特徴付ける語を  $X_n$  とする．( $n=1, 2, 3, \dots, N$ )
- 手順 3  $X_n$  を含む文を新聞から抜き出す．
- 手順 4 抜き出した文から  $X_n$  を  $X$  に言い換える．このとき  $X_n$  を  $X$  に言い換えた場合，言い換えた後の  $X$  は辞典の  $n$  番目の語義を持つ  $X$  となる．これを学習データとして新たに獲得することができる．これにより自動で学習データを増やすことができる．そして，その学習データを利用して  $X$  という単語の多義性解消を行う．学習に使用した素性は 48 種類あり，文構造や文中にある単語などを素性とする．機械学習には最大エントロピー法を利用する．

本研究では，言い換えを利用して増えた学習データ数を SemEval2 の学習データ数の比率になおし，実験を行う．

先行研究 [4] と本研究の違いは，対象としている言語が違っている．先行研究 [4] は英語，本研究は日本語を対象としている．また，本研究では，単語の選定をする際，人が思いつく語を選定するケースもあることが先行研究 [4] と異なる点である．

言い換えを利用した学習データの増やし方の具体例を以下に示す．

表 3.1: 「内容」を含む文の例

内容は別項の通りだが、男性二人と、女性一人がともに平壤市で暮らしていることを伝え、経済的に困窮していることを訴えていた。

表 3.2: 「動機」を含む文の例

着陸の動機は明らかにされていない。

例として多義語「意味」の学習データの増やし方を考える。多義語「意味」には、岩波国語辞典では以下の3つの語義がある。

- 語義 1: その言葉の表す 内容。意義。「辞書を引けば がわかる」
- 語義 2: 表現や行為の意図・動機。「どういう でそんなことをしたのか」
- 語義 3: 表現や行為のもつ 価値。意義。「そんな事をして も がない」

辞典の3つの語義を特徴付けたものを人手で選定する。ここでは、「内容」「動機」「価値」とする。そして、「内容」「動機」「価値」を含む文を新聞から抜き出す。

表 3.1 から表 3.3 のように「内容」「動機」「価値」を含む文を新聞から抜き出す。そして、表 3.4 のように抜き出した文から「内容」「動機」「価値」をそれぞれ「意味」に言い換える。

このとき「内容」を「意味」に置き換えた場合、言い換えた後の「意味」は辞典に基づく語義 1 を持つ「意味」とする。これが学習データになるので、学習データを増やすことができる。その学習データを利用して「意味」という単語の多義性解消を行う。

表 3.3: 「価値」を含む文の例

一票の 価値 が最も低い神奈川四区と最も重い宮崎二区の格差は三・一八倍に広がった。

表 3.4: 言い換える前と言い換えた後の文

語	「内容」「動機」「価値」を含む文	「意味」に言い換えた文	語義
内容	内容は別項の通りだが...	意味は別項の通りだが...	語義 1
動機	着陸の <u>動機</u> は明らかにされていない。	着陸の <u>意味</u> は明らかにされていない	語義 2
価値	一票の <u>価値</u> が最も低い神奈川四区と...	一票の <u>意味</u> が最も低い神奈川四区と...	語義 3

### 3.5 最大エントロピー法

最大エントロピー法は、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] .

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

$$H(p) = - \sum_{a \in A, b \in B} p(a, b) \log(p(a, b)) \tag{3.1}$$

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  $f_j$  があつてなおかつ分類が  $a$  の場合 1 となりそれ以外で 0 となる指示関数を意味する。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。また、 $k$  は素性の総数を意味する。  
 $p(a, b)$  を求め、そこから  $p(a|b)$  を求める。(  $p(a|b) = p(a : \text{解答} | b : \text{問題})$  )

$$p(a|b) = \frac{p(a, b)}{\sum_i p(a_i, b)} \tag{3.2}$$

式 (3.1) は確率  $p$  と出力と素性の組の出現を意味する関数  $g$  をかけることで出力と素性の組の頻度の期待値を求めることになっており，右辺の既知データにおける期待値と，左辺の求める確率分布に基づいて計算される期待値が等しいことを制約として，エントロピー最大化 (確率分布の平滑化) を行って，出力と文脈の確率分布を求めるものとなっている [5] .

## 第4章 実験

### 4.1 名詞の多義語での実験

機械学習の入力は，多義語を含む文，出力は，複数ある語義のうち，どの意味で使われたかとする．本研究では，SemEval2[2]の対象単語50個のうち名詞「はじめ」「市場」「電話」「意味」「前」「子供」の計6個を実験に使用する多義語とする．SemEval2[2]は，多義性解消のコンテストで用意されたものであり，多義性解消の研究や実験を行いやすいように人手で作成されたものである．また，その対象の多義語は50個あり，多義語1語につき，学習データとテストデータがそれぞれ50個ずつ用意されている．

また，本研究では学習データとテストデータを合わせてから，ランダムにそれぞれ50個ずつ学習データとテストデータを作りなおし，実験を行う．また，ランダムに作りなおした後，学習データとテストデータの中身を全て交換し，実験を行う．こうすることで，6個の単語で12回実験を行うことができる．本研究では，毎日新聞1991年から1995年の5年分の新聞データを使用する．

また，本研究では言い換えによって増えたデータ数を，SemEval2の学習データの比率になおし実験を行う．

機械学習は最大エントロピー法を使用する．また，表4.1に実験に使用した素性(解析に用いる情報)を示す．表4.1は文献[9]を参考にしている．これらの素性を，対象語が含まれる文から取り出す．対象語とは，処理する多義語のことである．表4.1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．

表 4.1: 使用した素性

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に係る文節の自立語
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に係る文節の付属語
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に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に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に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に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

## 4.2 単語の選定(名詞)

本研究では、SemEval2の対象単語50個のうち名詞「はじめ」「市場」「電話」「意味」「前」「子供」の計6個を実験に使用する単語(多義語)とする。これらの名詞は、SemEval2の対象単語50個のうちランダムに取り出した10個に含まれていた名詞を実験対象の多義語とした。言い換えに利用する単語には、語義を特徴付けたものを人手で選定する。

「はじめ」についての選定例を示す。岩波国語辞典では「子供」という単語の語義は以下の2つがある。

- 語義1:物事ははじめた段階。それをし始めた、またはそれに関係を持った時。最初のころ。物事の起こり。最初の部分。
- 語義2:《「...を」「...を とする」の形で》多くあるものの中で前のもの。先のもの。また、おもだったものをあげる言葉。

「はじめ」の場合、語義2は、辞典の語義から選定できなかったため、この語義から人が思いつく語を選定した。語義1を「最初」、語義2を「主」とした。

「市場」についての選定例を示す。岩波国語辞典では「市場」という単語の語義は以下の3つがある。

- 語義1:いちば。
- 語義2:売行き先。
- 語義3:〔経済〕売手と買手とが規則的に出会う。取引を行う組織。

「市場」の場合、語義1と語義2は、辞典の語義から選定できなかったため、この語義から人が思いつく語を選定した。語義1を「売り場」、語義2を「販路」、語義3を「取引」とした。

「電話」についての選定例を示す。岩波国語辞典では「電話」という単語の語義は以下の2つがある。

- 語義1:電話機による通話。
- 語義2:「電話機」の略。

「電話」の場合，語義 1 を「通話」，語義 2 を「電話機」とした．

「意味」の選定は，2.2 節で示したように語義 1 を「内容」，語義 2 を「動機」，語義 3 を「価値」とした．

「前」についての選定例を示す．岩波国語辞典では「前」という単語の語義は以下の 2 つがある．

- 語義 1: 視線・顔が向いている方。空間的に、後ろ・横でない方向や場所。貴人のおん前の意で貴人をさしたことから、身分の高い女性の名に添えた語。
- 語義 2: 順序が先の方。本体より 先。時間的に早い方。現在に先立つ(ある)時。

「前」の場合，語義 1 は，辞典の語義から選定できなかったので，この語義から人が思いつく語を選定した．語義 1 を「方向」，語義 2 を「先」とした．

「子供」についての選定例を示す．岩波国語辞典では「子供」という単語の語義は以下の 2 つがある．

- 語義 1: 幼い子。児童。
- 語義 2: 自分のもうけた子。むすこ、むすめ。子。

「子供」の場合，語義 1 を「児童」，語義 2 を「息子」とした．

### 4.3 動詞の多義語での実験

戸田ら [1] の研究により，言い換えによって増えたデータを追加することで正解率が向上した．しかし，戸田らは名詞でしか実験を行っていない．そこで本研究では，動詞で実験を行った．言い換えによって増えたデータを追加することで正解率が向上するかを確認する．機械学習の入力は，多義語を含む文であり，出力は，複数ある語義のうち，どの意味で使われたかとする．本研究では，SemEval2 の対象単語 50 個のうち動詞「考える」「認める」「やる」の計 3 個を実験に使用する多義語とする．多義語 1 語につき，学習データとテストデータはそれぞれ 50 個ずつである．

また，本研究では言い換えによって増えたデータ数を，SemEval2 の学習データの比率になおし実験を行う．

本研究では，毎日新聞 1991 年から 1995 年の 5 年分の新聞データを使用する．

機械学習は最大エントロピー法を使用する．また，表 4.1 の素性(解析に用いる情報)を利用した．

## 4.4 単語の選定(動詞)

本研究では, SemEval2の対象単語50個のうち動詞「考える」「認める」「やる」「見る」「持つ」の計5個を実験に使用する単語(多義語)とする。これらの動詞は, SemEval2の対象単語50個のうちランダムに取り出した10個に含まれていた動詞を実験対象の多義語とした。言い換えに利用する単語には, 語義を特徴付けたものを人手で選定する。

「考える」についての選定例を示す。岩波国語辞典では「考える」という単語の語義は以下の2つがある。

- 語義1: あれやこれやと思いをめぐらす。その事について、心を知的に使って判断する。
- 語義2: 新たなものをくふうする。考案する。

「考える」の場合, 語義1を「判断する」, 語義2を「考案する」とした。

「認める」についての選定例を示す。岩波国語辞典では「認める」という単語の語義は以下の4つがある。

- 語義1: 目にとめる。有ることが確かだと見てとる。
- 語義2: 見て判断する。
- 語義3: 間違いないと判断する。差し支えないとして、許可する。
- 語義4: 見どころがあるものとして、目にとめる。

「認める」の場合, 語義4は辞典の語義から選定できなかったため, この語義から人が思いつく語を選定した。語義1を「見て取る」, 語義2を「判断する」, 語義3を「許可する」, 語義4は「評価する」とした。

「やる」についての選定例を示す。岩波国語辞典では「やる」という単語の語義は以下の4つがある。

- 語義1: 一方から他方へ移らせる。行かせる。進ませる。先方の事情が(はっきりとは)分からない時に差し向けるという気持を伴うことも多い。先方に構わずにするという気持の場合もある。思いを晴らす。
- 語義2: (同等以下のものに) 与える。

- 語義 3:《動詞連用形 + 助詞「て」を受けて》(恩恵的に)他人のために...する意を表す。(怒り・憎しみの気持で)進んで相手に不利益を与える意を表す。やけになってする意を表す。
- 語義 4:みずから(進んで)する。また、みずからの意志によらない(好ましくない)ことについても、わが身に引き取って、それがあたかも自分の意志の結果であるかのように表現することがある。本来は、うまくゆくかどうか分からない事をあえてする時に言う。

「やる」の場合、語義 3 は辞典の語義から選定できなかつたので、この語義から人が思いつく語を選定した。語義 1 を「行かせる」、語義 2 を「与える」、語義 3 を「あげる」、語義 4 は「する」とした。

「持つ」についての選定例を示す。岩波国語辞典では「持つ」という単語の語義は以下の 4 つがある。

- 語義 1:手に取る。
- 語義 2:自分のものにする。所持する。身につける。所有する。心にいだく。身にそなえる。担任する。受け持つ。負担する。
- 語義 3:成り立たせる。有する。保つ。
- 語義 4:五自状態・品質・働きが長い間変わらないで保たれる。

「持つ」の場合、語義 4 は辞典の語義から選定できなかつたので、この語義から人が思いつく語を選定した。語義 1 を「取る」、語義 2 を「所持する」、語義 3 を「保つ」、語義 4 は「続く」とした。

岩波国語辞典では「見る」という単語の語義は以下の 4 つがある。

- 語義 1:視覚を働かして、ものの存在・形・様子・内容をとらえる。目で認める。
- 語義 2:視覚に限らず広く、感覚を働かして、探りとらえる。感覚でとらえたものについて、判断・評価をする。
- 語義 3:《「...と」の形で》...と考える。理解する。推定する。見積もる。
- 語義 4:様子を見て世話をする。取り扱う。行う。

- 語義5:《動詞の連用形に助詞の「て」の付いた形を受けて》ためしに...する。《「...してみると」「...してみたら」「...してみれば」の形で》...すると。...したところが。

「見る」の場合、語義5は辞典の語義から選定できなかったため、この語義から人が思いつく語を選定した。語義1を「とらえる」、語義2を「判断する」、語義3を「考える」、語義4は「世話する」、語義5を「いる」とした。

## 4.5 形容詞の多義語での実験

戸田ら [1] の研究により、言い換えによって増えたデータを追加することで正解率が向上した。しかし、戸田らは名詞でしか実験を行っていない。そこで本研究では、形容詞で実験を行った。言い換えによって増えたデータを追加することで正解率が向上するかを確認する。機械学習の入力は、多義語を含む文であり、出力は、複数ある語義のうち、どの意味で使われたかとする。本研究では、SemEval2の対象単語50個のうち形容詞「強い」を実験に使用する多義語とする。多義語1語につき、学習データとテストデータはそれぞれ50個ずつである。

また、本研究では言い換えによって増えたデータ数を、SemEval2の学習データの比率になおし実験を行う。

本研究では、毎日新聞1991年から1995年の5年分の新聞データを使用する。

機械学習は最大エントロピー法を使用する。また、表4.1の素性(解析に用いる情報)を利用した。

## 4.6 単語の選定(形容詞)

本研究では、SemEval2の対象単語50個のうち形容詞「強い」を実験に使用する単語(多義語)とする。この形容詞は、SemEval2の対象単語50個のうちランダムに取り出した10個に含まれていた形容詞を実験対象の多義語とした。言い換えに利用する単語には、語義を特徴付けたものを人手で選定する。

「強い」についての選定例を示す。岩波国語辞典では「強い」という単語の語義は以下の2つがある。

- 語義1:積極的に働く力にあふれている。しとげる腕力・能力が十分にある。勢い・作用が激しい。

- 語義 2: 抵抗力に富み、簡単には壊れたりくずれたりしない。

「強い」の場合、語義 2 は辞典の語義から選定できなかったため、この語義から人が思いつく語を選定した。語義 1 を「激しい」、語義 2 を「頑丈だ」とした。

## 第5章 実験結果

### 5.1 名詞の多義語での実験結果

#### 5.1.1 名詞の実験に使用したデータ

「はじめ」「市場」「電話」「意味」「前」「子供」という多義語で実験を行った。機械学習手法として最大エントロピー法を用いた。表 5.1 から表 5.6 に SemEval2 の「はじめ」「市場」「電話」「意味」「前」「子供」についての事例数を示す。「意味」「市場」には、辞典に基づく語義が 3 つある。「はじめ」「電話」「子供」には、辞典に基づく語義が 2 つある。

本研究では学習データとテストデータを合わせてから、ランダムにそれぞれ 50 個ずつ学習データとテストデータを作りなおし、実験を行う。また、ランダムに作りなおした後、学習データとテストデータの中身を全て交換し、実験を行う。こうすることで、6 個の単語で 12 回実験を行うことができる。表中の「ランダム」とは、学習データとテストデータを合わせてから、ランダムにそれぞれ 50 個ずつ学習データとテストデータを作りなおしたもの。表中の「ランダム交換」はランダムに作りなおした後、学習データとテストデータの中身を全て交換したものである。

表 5.1 から表 5.6 に名詞の多義語 6 単語の「SemEval2 のテストデータ数」「SemEval2 の学習データ数」を示す。

表 5.1: semeval2 の学習データ (名詞)

語義	はじめ	市場	電話	意味	前	子供
語義 1	最初:15	売り場:5	通話:23	内容:25	方向:19	児童:26
語義 2	主:35	販路:10	電話機:27	動機:8	先:31	息子:24
語義 3	-	取引:35	-	価値:17	-	-
未知語義	-	-	-	1	-	-
合計	50	50	50	50	50	50

表 5.2: semeval2 のテストデータ (名詞)

語義	はじめ	市場	電話	意味	前	子供
語義 1	最初:20	売り場:3	通話:22	内容:27	方向:12	児童:18
語義 2	主:30	販路:12	電話機:28	動機:8	先:31	息子:32
語義 3	-	取引:35	-	価値:12	-	-
未知語義	-	-	-	1	7	-
合計	50	50	50	50	50	50

表 5.3: semeval2 の学習データ (ランダム)

語義	はじめ	市場	電話	意味	前	子供
語義 1	最初:18	売り場:4	通話:25	内容:28	方向:17	児童:22
語義 2	主:32	販路:11	電話機:25	動機:9	先:29	息子:28
語義 3	-	取引:35	-	価値:12	-	-
未知語義	-	-	-	1	4	-
合計	50	50	50	50	50	50

表 5.7 に言い換えによって増えたデータ数を示す。

表 5.4: semeval2 のテストデータ (ランダム)

語義	はじめ	市場	電話	意味	前	子供
語義 1	最初:17	売り場:4	通話:20	内容:24	方向:14	児童:22
語義 2	主:33	販路:11	電話機:30	動機:9	先:33	息子:28
語義 3	-	取引:35	-	価値:17	-	-
未知語義	-	-	-	-	3	-
合計	50	50	50	50	50	50

表 5.5: semeval2 の学習データ (ランダムにして交換)

語義	はじめ	市場	電話	意味	前	子供
語義 1	最初:17	売り場:4	通話:20	内容:24	方向:15	児童:22
語義 2	主:33	販路:11	電話機:30	動機:9	先:28	息子:28
語義 3	-	取引:35	-	価値:17	-	-
未知語義	-	-	-	-	7	-
合計	50	50	50	50	50	50

表 5.6: semeval2 のテストデータ (ランダムにして交換)

語義	はじめ	市場	電話	意味	前	子供
語義 1	最初:18	売り場:4	通話:25	内容:28	方向:20	児童:22
語義 2	主:32	販路:11	電話機:25	動機:9	先:30	息子:28
語義 3	-	取引:35	-	価値:12	-	-
未知語義	-	-	-	-	-	-
合計	50	50	50	50	50	50

表 5.7: 言い換えによって増えたデータ数 (名詞)

語義	増えた学習データ数					
	はじめ	市場	電話	意味	前	子供
語義 1	最初:370	売り場:78	通話:423	内容:6440	方向:10431	児童:6138
語義 2	主:865	販路:156	電話機:497	動機:2061	先:17019	息子:5666
語義 3	-	取引:546	-	価値:4379	-	-
未知語義	-	-	-	-	-	-
合計	1235	780	920	12880	27450	11804

3.3 節の研究を行うため SemEval2 の表 5.7 の言い換えによって増えたデータ数を調節した．そのデータ数を表 5.8 に示す．

表 5.8: 言い換えによって増えたデータ数 (名詞):データ数調整

語義	増えた学習データ数					
	はじめ	市場	電話	意味	前	子供
語義 1	最初:300	売り場:50	通話:230	内容:5750	方向:10260	児童:5200
語義 2	主:700	販路:100	電話機:270	動機:1840	先:16740	息子:4800
語義 3	-	取引:350	-	価値:3910	-	-
未知語義	-	-	-	-	-	-
合計	1000	500	500	11500	27000	10000

### 5.1.2 多義性解消の結果

表 5.3 と表 5.4 の実験データを用いた実験の結果を表 5.9 と表 5.10 に示す．

表 5.9: semeval2 の学習データのみの評価 (ランダム)

手法	はじめ	市場	電話	意味	前	子供
SemEval2 の学習データのみの利用する手法	0.90 (45/50)	0.74 (37/50)	0.82 (41/50)	0.56 (28/50)	0.76 (40/50)	0.56 (28/50)
SemEval2 の学習データ + 言い換えによって増えた学習データを利用する手法	0.96 (48/50)	0.70 (35/50)	0.76 (38/50)	0.70 (35/50)	0.74 (37/50)	0.66 (33/50)
言い換えによって増えた学習データのみの利用する手法	0.90 (45/50)	0.62 (31/50)	0.82 (41/50)	0.56 (28/50)	0.68 (34/50)	0.60 (30/50)
最頻出語義を常に出力する手法	0.64 (32/50)	0.70 (35/50)	0.28 (14/50)	0.56 (28/50)	0.58 (29/50)	0.56 (28/50)

「電話」は「手法 1:SemEval2 の学習データのみの利用する手法」と「手法 3:言い換えによって増えた学習データのみの利用する手法」が一番良い正解率となった。「市場」と「前」は「手法 1:SemEval2 の学習データのみの利用する手法」が一番良い正解率となった。「子供」と「意味」と「はじめ」は「手法 2:SemEval2 の学習データ + 言い換えによって増えた学習データを利用する手法」が一番良い正解率となった．

表 5.10: semeval2 のテストデータのみの評価 (ランダム)

手法	はじめ	市場	電話	意味	前	子供
SemEval2 の学習データのみを利用する手法	0.90 (45/50)	0.64 (32/50)	0.86 (43/50)	0.58 (29/50)	0.78 (39/50)	0.58 (29/50)
SemEval2 の学習データ + 言い換えによって増えた学習データを利用する手法	0.92 (46/50)	0.70 (35/50)	0.76 (38/50)	0.60 (30/50)	0.62 (31/50)	0.68 (34/50)
言い換えによって増えた学習データのみを利用する手法	0.86 (43/50)	0.70 (35/50)	0.70 (35/50)	0.54 (27/50)	0.62 (31/50)	0.68 (34/50)
最頻出語義を常に出力する手法	0.64 (32/50)	0.70 (35/50)	0.50 (25/50)	0.56 (28/50)	0.58 (29/50)	0.56 (28/50)

「電話」と「前」は「手法 1:SemEval2 の学習データのみを利用する手法」が一番良い正解率となった。「市場」と「子供」は「手法 2:SemEval2 の学習データ + 言い換えによって増えた学習データを利用する手法」と「手法 3:言い換えによって増えた学習データのみを利用する手法」が一番良い正解率となった。「意味」と「はじめ」は「手法 2:SemEval2 の学習データ + 言い換えによって増えた学習データを利用する手法」が一番良い正解率となった。

表 5.5 と表 5.6 の実験データを用いた実験の結果を表 5.11 と表 5.12 に示す。

表 5.11: semeval2 の学習データのみの評価 (ランダムにして交換)

手法	はじめ	市場	電話	意味	前	子供
SemEval2 の学習データのみを利用する手法	0.86 (43/50)	0.62 (31/50)	0.78 (39/50)	0.54 (27/50)	0.82 (41/50)	0.74 (37/50)
SemEval2 の学習データ + 言い換えによって増えた学習データを利用する手法	0.90 (45/50)	0.62 (31/50)	0.68 (34/50)	0.58 (29/50)	0.66 (33/50)	0.70 (35/50)
言い換えによって増えた学習データのみを利用する手法	0.88 (44/50)	0.72 (36/50)	0.70 (35/50)	0.50 (25/50)	0.62 (31/50)	0.70 (35/50)
最頻出語義を常に出力する手法	0.66 (33/50)	0.70 (35/50)	0.60 (30/50)	0.48 (24/50)	0.66 (33/50)	0.64 (32/50)

「電話」と「子供」と「前」は「手法 1: SemEval2 の学習データのみを利用する手法」が一番良い正解率となった。「市場」は「手法 3: 言い換えによって増えた学習データのみを利用する手法」が一番良い正解率となった。「意味」と「はじめ」は「手法 2: SemEval2 の学習データ + 言い換えによって増えた学習データを利用する手法」が一番良い正解率となった。

表 5.12: semeval2 のテストデータのみの評価 (ランダムにして交換)

手法	はじめ	市場	電話	意味	前	子供
SemEval2 の学習データのみを利用する手法	0.82 (41/50)	0.72 (36/50)	0.82 (41/50)	0.66 (33/50)	0.78 (39/50)	0.66 (33/50)
SemEval2 の学習データ + 言い換えによって増えた学習データを利用する手法	0.92 (46/50)	0.74 (37/50)	0.78 (39/50)	0.74 (37/50)	0.78 (39/50)	0.62 (31/50)
言い換えによって増えた学習データのみを利用する手法	0.92 (46/50)	0.68 (34/50)	0.58 (29/50)	0.66 (33/50)	0.68 (34/50)	0.56 (28/50)
最頻出語義を常に出力する手法	0.60 (30/50)	0.70 (35/50)	0.60 (30/50)	0.48 (24/50)	0.56 (28/50)	0.56 (28/50)

「電話」と「子供」は「手法 1:SemEval2 の学習データのみを利用する手法」が一番良い正解率となった。「市場」と「意味」は「手法 2:SemEval2 の学習データ + 言い換えによって増えた学習データを利用する手法」が一番良い正解率となった。「はじめ」は「手法 2:SemEval2 の学習データ + 言い換えによって増えた学習データを利用する手法」と「手法 3:言い換えによって増えた学習データのみを利用する手法」が一番良い正解率となった。「前」は「手法 1:SemEval2 の学習データのみを利用する手法」と「手法 2:SemEval2 の学習データ + 言い換えによって増えた学習データを利用する手法」が一番良い正解率となった。

表 5.9 と表 5.11 の合計 12 回の実験結果をまとめた正解率の平均を表 5.13 に示す

表 5.13: semeval2 の学習データのみでの評価 (名詞)

手法	正解率 (平均)
SemEval2 の学習データのみを利用する 手法	0.725
SemEval2 の学習データ + 言い換えによ って増えた学習データを利用する手法	0.722
言い換えによって増えた学習データの みを利用する手法	0.692
最頻出語義を常に出力する手法	0.590

6 単語全て、計 12 回の実験の正解率の平均では「手法 1:SemEval2 の学習データの  
みを利用する手法」が正解率 0.725 と最も良い正解率となった。

表 5.10 と表 5.12 の合計 12 回の実験結果をまとめた正解率の平均を表 5.14 に示す . 6

表 5.14: semeval2 のテストデータのみでの評価 (名詞)

手法	正解率 (平均)
SemEval2 の学習データのみを利用する 手法	0.733
SemEval2 の学習データ + 言い換えによ って増えた学習データを利用する手法	0.738
言い換えによって増えた学習データの みを利用する手法	0.682
最頻出語義を常に出力する手法	0.590

単語全て , 計 12 回の実験の正解率の平均では「手法 1:SemEval2 の学習データのみを  
利用する手法」が正解率 0.738 と最も良い正解率となった .

表 5.14 の結果をふまえ，SemEval2 の学習データでの 10 分割クロスバリデーションによる手法選択 (手法 5) と 4 手法の結果を表 5.15 に示す．

表 5.15: semeval2 のテストデータでの評価:6 単語全て

手法	正解率
手法 1:SemEval2 の学習データのみを利用する手法	0.733
手法 2:SemEval2 の学習データ + 言い換えによって増えた学習データを利用する手法	0.738
手法 3:言い換えによって増えた学習データのみを利用する手法	0.681
手法 4:最頻出語義を常に出力する手法	0.590
手法 5:クロスバリデーションによる手法選択	0.757

表 5.15 より、「手法 1:SemEval2 の学習データのみを利用する手法」の正解率が 0.733 に対して、「手法 5:クロスバリデーションによる手法選択」の正解率が 0.757 となっており，正解率が向上した．また「手法 5:クロスバリデーションによる手法選択」が最も良い正解率となった．

### 5.1.3 有意差検定

4 種類の学習データを利用する手法と 10 分割クロスバリデーションを利用する手法の計 5 種類の手法の性能の実験結果について，統計的に有意差があるかを調べた．6 単語すべての計 12 回の実験での 600 事例を利用して，符号検定を行った．表 5.16 に有意差検定の結果を示す．

表 5.16: 有意差検定結果  
 比べる手法

比べる手法		有意差の有無 (有意水準 0.05 の片側検定)
手法 1:SemEval2 の学習データのみを利用する手法	手法 2:SemEval2 の学習データ+言い換えによって増えた学習データを利用する手法	有意差なし (p 値:0.54)
手法 1:SemEval2 の学習データのみを利用する手法	手法 3:言い換えによって増えた学習データのみを利用する手法	有意差あり (p 値:0.00)
手法 1:SemEval2 の学習データのみを利用する手法	手法 4:SemEval2 の学習データの最頻出語義を常に出力する手法	有意差あり (p 値:0.00)
手法 2:SemEval2 の学習データ+言い換えによって増えた学習データを利用する手法	手法 3:言い換えによって増えた学習データのみを利用する手法	有意差あり (p 値:0.00)
手法 2:SemEval2 の学習データ+言い換えによって増えた学習データを利用する手法	手法 4:SemEval2 の学習データの最頻出語義を常に出力する手法	有意差あり (p 値:0.00)
手法 5:クロスバリデーションによる手法選択	手法 1:SemEval2 の学習データのみを利用する手法	有意差あり (p 値:0.03)
手法 5:クロスバリデーションによる手法選択	手法 2:SemEval2 の学習データ+言い換えによって増えた学習データを利用する手法	有意差なし (p 値:0.07)
手法 5:クロスバリデーションによる手法選択	手法 3:言い換えによって増えた学習データのみを利用する手法	有意差あり (p 値:0.00)
手法 5:クロスバリデーションによる手法選択	手法 4:SemEval2 の学習データの最頻出語義を常に出力する手法	有意差あり (p 値:0.00)

表 5.16 より、「手法 5:クロスバリデーションによる手法選択」と「手法 1:SemEval2 の学習データのみを利用する手法」または「手法 3:言い換えによって増えた学習データのみを利用する手法」または「手法 4:SemEval2 の学習データの最頻出語義を常に出力する手法」では、有意水準 0.05 の符号検定 (片側検定と両側検定の両方) で有意差があった。また、「手法 1:SemEval2 の学習データのみを利用する手法」と「手法 2:SemEval2 の学習データ+言い換えによって増えた学習データを利用する手法」、ま

たは「手法 2:SemEval2 の学習データ+言い換えによって増えた学習データを利用する手法」と「手法 5:クロスバリデーションによる手法選択」では、有意水準 0.05 の符号検定（片側検定）で有意差がなかった。しかし、「手法 2:SemEval2 の学習データ+言い換えによって増えた学習データを利用する手法」と「手法 5:クロスバリデーションによる手法選択」では、有意水準 0.05 の符号検定（両側検定）では、有意差があった。他の組み合わせは、有意水準 0.05 の符号検定（片側検定）で有意差があった。

#### 5.1.4 考察 (クロスバリデーションによる手法選択)

「手法 1:SemEval2 の学習データのみを利用する手法」の正解率 0.733 に対して、「手法 2:SemEval2 の学習データ+言い換えによって増えた学習データを利用する手法」の正解率 0.738 と、正解率がほとんど変わらなかった理由としては、言い換えによって増えたデータが必ずしも良いデータとは限らないということが考えられる。つまり、言い換えによって増えたデータが悪さをしていることがあるということである。このことから言い換えを利用したほうが良い多義語とそうでない多義語があることがわかる。

また、本研究は 6 個の単語で 12 回実験を行うことができる。SemEval2 の学習データでの 10 分割クロスバリデーションにより正しく手法選択できた実験は、12 回の実験のうち 10 回の実験であった。これは、学習データとテストデータで同じ手法の性能が良いという相関があることを示している。

### 5.1.5 正解例と誤り例

「手法 1:SemEval2 の学習データのみを利用する手法」が不正解で「手法 5:クロスバリデーションによる手法選択」が正解だった問題数を表 5.17 に示す。

表 5.17: 手法 1：不正解で手法 5：正解の問題数

単語	問題数
はじめ	8
市場	2
電話	0
意味	11
前	0
子供	11
合計	32

「手法 1:SemEval2 の学習データのみを利用する手法」が不正解で「手法 5:クロスバリデーションによる手法選択」が正解だった問題数の合計は 32 であった。

「手法 1:SemEval2 の学習データのみを利用する手法」が正解で「手法 5:クロスバリデーションによる手法選択」が不正解だった問題数を表 5.18 に示す。

表 5.18: 手法 1：正解で手法 5：不正解の問題数

単語	問題数
はじめ	2
市場	4
電話	0
意味	6
前	0
子供	6
合計	18

「手法 1:SemEval2 の学習データのみを利用する手法」が正解で「手法 5:クロスバリデーションによる手法選択」が不正解だった問題数の合計は 18 であった。

表 5.17 と表 5.18 より「手法 1:SemEval2 の学習データのみを利用する手法」が正解だった問題が不正解になることが 18 問あるが、「手法 1:SemEval2 の学習データのみを利用する手法」不正解だった問題が正解になることが 32 問あるので、総合的には、「手法 5:クロスバリデーションによる手法選択」は役立つことが多い。

「手法 1:SemEval2 の学習データのみを利用する手法」が不正解で「手法 5:クロスバリデーションによる手法選択」が正解だった問題の例を表 5.19 に示す。

表 5.19: 手法 1：不正解で手法 5：正解の例

単語	例文
はじめ	神風はじめ、幕内力士たちが住んでいたという借家や、久保田の寮の場所はどこなのか、それさえはっきりとわからなかった。
市場	こうした作業に筆者が集中していた千九百八十年は、日米間の半導体製品の輸出入額が逆転した記念すべき年であり、日本が半導体を中心に世界市場を席卷し始める最初の年でもあった。
電話	無し
意味	エミヤのように無理して平常を装う「やせがまん」も、これらの単語で意味が通じるよ。
前	無し
子供	街を歩けば前後左右から水のはねる音や子供の歓声が聞こえる。

「手法 1:SemEval2 の学習データのみを利用する手法」が正解で「手法 5:クロスバリ  
 デーションによる手法選択」が不正解だった問題の例を表 5.20 に示す。

表 5.20: 手法 1：正解で手法 5：不正解の例

単語	例文
はじめ	このほか、各地方公共団体においては、防災業務に精通した職員をはじめとした人員を夜間・休日においても二十四時間体制で配置するほか、災害時の職員の自はじめ参集基準の明確化や職場近郊の災害対応職員用宿舍の確保など災害初動体制の確立を図る必要がある。
市場	雇用は、人と仕事を結びつけること、ロンドンの排除された住宅団地に対する活動や、どのようにして彼らを主な市場、特に労働市場に上手に紹介するか、などに焦点を合わせています。
電話	無し
意味	表題に掲げた“リンク”には、つなぐとか、関連づけるなどの意味があります。
前	無し
子供	親の不安は募るばかりだが、かといって子供を家の中に閉じ込めておくわけにもいかない。

### 5.1.6 SemEval2の学習データ数を変更

SemEval2の学習データ数を言い換えによって増えたデータ数のX倍(1/10,1/5,1/2,1,2,5,10)のデータ数に変更した場合の結果を表5.21と表5.22に示す. また,表5.21と表5.22の「元データ」は, SemEval2の学習データ数を50のままで実験した結果である.

表 5.21: semeval2の学習データ数を変更(ランダム)

単語：手法	元データ	× 1/10	× 1/5	× 1/2	× 1	× 2	× 5	× 10
電話：手法1	0.86	0.86	0.86	0.86	0.86	0.86	0.86	0.86
電話：手法2	0.76	0.76	0.76	0.80	0.80	0.80	0.80	0.80
電話：手法3	0.70	0.70	0.70	0.70	0.70	0.70	0.70	0.70
電話：手法4	0.50	0.50	0.50	0.50	0.50	0.50	0.50	0.50
市場：手法1	0.64	0.64	0.64	0.64	0.64	0.64	0.64	0.64
市場：手法2	0.70	0.70	0.70	0.70	0.70	0.70	0.70	0.70
市場：手法3	0.70	0.70	0.70	0.70	0.70	0.70	0.70	0.70
市場：手法4	0.70	0.70	0.70	0.70	0.70	0.70	0.70	0.70
子供：手法1	0.58	0.60	0.60	0.60	0.58	0.58	0.58	0.58
子供：手法2	0.68	0.68	0.68	0.68	0.70	0.68	0.68	0.68
子供：手法3	0.68	0.68	0.68	0.68	0.68	0.68	0.68	0.68
子供：手法4	0.56	0.56	0.56	0.56	0.56	0.56	0.56	0.56
意味：手法1	0.58	0.58	0.60	0.60	0.60	0.58	0.58	0.58
意味：手法2	0.60	0.60	0.60	0.60	0.60	0.62	0.60	0.60
意味：手法3	0.54	0.54	0.54	0.54	0.54	0.54	0.54	0.54
意味：手法4	0.56	0.56	0.56	0.56	0.56	0.56	0.56	0.56
はじめ：手法1	0.90	0.90	0.90	0.90	0.90	0.90	0.90	0.90
はじめ：手法2	0.92	0.92	0.92	0.92	0.92	0.92	0.94	0.94
はじめ：手法3	0.86	0.86	0.86	0.86	0.86	0.86	0.86	0.86
はじめ：手法4	0.64	0.64	0.64	0.64	0.64	0.64	0.64	0.64
前：手法1	0.78	0.78	0.78	0.78	0.78	0.78	0.78	0.78
前：手法2	0.62	0.64	0.64	0.64	0.64	0.64	0.64	0.64
前：手法3	0.62	0.62	0.62	0.62	0.62	0.62	0.62	0.62
前：手法4	0.58	0.58	0.58	0.58	0.58	0.58	0.58	0.58

表 5.22: semeval2 の学習データ数を変更 (ランダム交換)

単語：手法	元データ	× 1/10	× 1/5	× 1/2	× 1	× 2	× 5	× 10
電話：手法 1	0.82	0.82	0.82	0.82	0.82	0.82	0.82	0.82
電話：手法 2	0.78	0.78	0.78	0.82	0.82	0.82	0.82	0.80
電話：手法 3	0.58	0.58	0.58	0.58	0.58	0.58	0.58	0.58
電話：手法 4	0.60	0.60	0.60	0.60	0.60	0.60	0.60	0.60
市場：手法 1	0.72	0.72	0.72	0.72	0.72	0.72	0.72	0.72
市場：手法 2	0.74	0.74	0.74	0.74	0.74	0.74	0.78	0.78
市場：手法 3	0.68	0.68	0.68	0.68	0.68	0.68	0.68	0.68
市場：手法 4	0.70	0.70	0.70	0.70	0.70	0.70	0.70	0.70
子供：手法 1	0.66	0.64	0.64	0.64	0.64	0.64	0.64	0.64
子供：手法 2	0.62	0.60	0.60	0.58	0.58	0.64	0.62	0.62
子供：手法 3	0.56	0.56	0.56	0.56	0.56	0.56	0.56	0.56
子供：手法 4	0.56	0.56	0.56	0.56	0.56	0.56	0.56	0.56
意味：手法 1	0.66	0.62	0.62	0.62	0.62	0.62	0.62	0.62
意味：手法 2	0.74	0.70	0.68	0.68	0.68	0.68	0.70	0.70
意味：手法 3	0.66	0.66	0.66	0.66	0.66	0.66	0.66	0.66
意味：手法 4	0.48	0.48	0.48	0.48	0.48	0.48	0.48	0.48
はじめ：手法 1	0.82	0.82	0.82	0.84	0.84	0.84	0.84	0.84
はじめ：手法 2	0.92	0.92	0.92	0.92	0.92	0.92	0.92	0.92
はじめ：手法 3	0.92	0.92	0.92	0.92	0.92	0.92	0.92	0.92
はじめ：手法 4	0.60	0.60	0.60	0.60	0.60	0.60	0.60	0.60
前：手法 1	0.78	0.78	0.78	0.78	0.78	0.78	0.78	0.78
前：手法 2	0.78	0.78	0.78	0.78	0.78	0.76	0.76	0.76
前：手法 3	0.68	0.68	0.68	0.68	0.68	0.68	0.68	0.68
前：手法 4	0.56	0.56	0.56	0.56	0.56	0.56	0.56	0.56

表 5.21 と表 5.22 より, SemEval2 の学習データ数を言い換えによって増えたデータ数の X 倍 (1/10,1/5,1/2,1,2,5,10) のデータ数に変更することにより「手法 2:SemEval2 と言い換えによって増えたデータを用いる手法」は, SemEval2 の学習データ数が増えるにつれ性能が向上する傾向がみられた。また × 1, × 2, × 5, × 10 は元データより, 良い正解率となった。

### 5.1.7 考察 (SemEval2 の学習データ数を変更)

SemEval2 の学習データ数を言い換えによって増えたデータ数の X 倍 (1/10,1/5,1/2,1,2,5,10) のデータ数に変更することにより「手法 2:SemEval2 と言い換えによって増えたデー

タを用いる手法」は、SemEval2 の学習データ数が増えるにつれ性能が向上する傾向がみられた。また  $\times 1$ 、 $\times 2$ 、 $\times 5$ 、 $\times 10$  は元データより、良い正解率となった。

## 5.2 動詞の多義語での実験結果

### 5.2.1 動詞の実験に使用したデータ

「考える」「認める」「やる」「持つ」「見る」という多義語で実験を行った。機械学習手法として最大エントロピー法を用いた。表 5.23 から表 5.25 に「考える」「認める」「やる」「持つ」「見る」についての「SemEval2 のテストデータ数」「SemEval2 の学習データ数」「言い換えによって増えた文数」を示す。「見る」には辞典に基づく語義が 5 つある。「認める」「やる」「持つ」には、辞典に基づく語義が 4 つある。「考える」には、辞典に基づく語義が 2 つある。

表 5.23: semeval2 の学習データ (動詞)

語義	考える	認める	やる	持つ	見る
語義 1	判断する:47	見て取る:2	行かせる:1	取る:9	とらえる:43
語義 2	考案する:3	判断する:6	与える:2	所持する:38	判断する:5
語義 3	-	許可する:40	あげる:3	保つ:3	考える:2
語義 4	-	評価する:2	する:44	続く:0	世話する:0
語義 5	-	-	-	-	いる:0
合計	50	50	50	50	50

表 5.24: semeval2 のテストデータ (動詞)

語義	考える	認める	やる	持つ	見る
語義 1	判断する:49	見て取る:1	行かせる:1	取る:14	とらえる:40
語義 2	考案する:1	判断する:10	与える:0	所持する:34	判断する:2
語義 3	-	許可する:35	あげる:2	保つ:1	考える:2
語義 4	-	評価する:4	する:47	続く:1	世話する:1
語義 5	-	-	-	-	いる:5
合計	50	50	50	50	50

表 5.23 から表 5.25 のデータを用いた実験の結果を表 5.26 に示す。

「考える」は「手法 1:SemEval2 の学習データのみを利用する手法」と「手法 2:SemEval2 の学習データ + 言い換えによって増えた学習データを利用する手法」と「手

表 5.25: 言い換えによって増えたデータ数 (動詞)

語義	考える	認める	やる	持つ	見る
語義 1	6439	29	259	135	4386
語義 2	411	87	518	570	510
語義 3	-	580	777	45	204
語義 4	-	29	11396	0	0
語義 5	-	-	-	-	0
合計	6850	725	12950	750	5100

表 5.26: 利用する学習データとその正解率 (動詞)

手法	考える	認める	やる	持つ	見る
手法 1: SemEval2 の学習データのみを利用する手法	0.98 (49/50)	0.80 (40/50)	0.96 (48/50)	0.84 (42/50)	0.80 (40/50)
手法 2: SemEval2 の学習データ + 言い換えによって増えた学習データを利用する手法	0.96 (48/50)	0.80 (40/50)	0.92 (46/50)	0.64 (32/50)	0.62 (31/50)
手法 3: 言い換えによって増えた学習データのみを利用する手法	0.98 (49/50)	0.72 (36/50)	0.88 (44/50)	0.58 (29/50)	0.60 (30/50)
手法 4: 最頻出語義を常に出力する手法	0.98 (49/50)	0.70 (35/50)	0.94 (47/50)	0.68 (34/50)	0.80 (40/50)

法 4:再頻出語義を常に出力する手法」が一番良い正解率となった。「認める」は「手法 1:SemEval2 の学習データのみを利用する手法」と「手法 2:SemEval2 の学習データ + 言い換えによって増えた学習データを利用する手法」が一番良い正解率となった。「やる」と「持つ」は「手法 1:SemEval2 の学習データのみを利用する手法」が一番良い正解率となった。「見る」は「手法 1:SemEval2 の学習データのみを利用する手法」と「手法 4: 最頻出語義を常に出力する手法」が一番良い正解率となった。

表 5.26 の結果をまとめたものを表 5.27 に示す。

表 5.27: 手法と正解率:5 単語全て

手法	正解率
手法 1:SemEval2 の学習データのみを利用する手法	0.876
手法 2:SemEval2 の学習データ + 言い換えによって増えた学習データを利用する手法	0.788
手法 3:言い換えによって増えた学習データのみを利用する手法	0.752
手法 4:最頻出語義を常に出力する手法	0.820

表 5.27 より、「手法 1:SemEval2 の学習データのみを利用する手法」の正解率が 0.876 に対して、「手法 2:SemEval2 の学習データ + 言い換えによって増えた学習データを利用する手法」の正解率が 0.788 となっており、正解率が低下した。また、「手法 1:SemEval2 の学習データのみを利用する手法」が最も良い正解率となった。

## 5.2.2 考察

「手法 1:SemEval2 の学習データのみを利用する手法」の正解率 0.876 に対して、「手法 2:SemEval2 の学習データ + 言い換えによって増えた学習データを利用する手法」の正解率 0.788 と、正解率が低下した理由としては、言い換えによって増えたデータが悪さをしていることが考えられる。単語の選定がうまくできていなかった可能性も考えられる。

## 5.3 形容詞の多義語での実験結果

### 5.3.1 形容詞の実験に使用したデータ

「強い」という多義語で実験を行った．機械学習手法として最大エントロピー法を用いた．表 5.28 から表 5.30 に「強い」についての「SemEval2 のテストデータ数」「SemEval2 の学習データ数」「言い換えによって増えた文数」を示す．「強い」には，辞典に基づく語義が 2 つある．

表 5.28: SemEval2 のテストデータ (形容詞) 表 5.29: SemEval2 の学習データ (形容詞)

語義	データ数
語義 1	激しい:46
語義 2	頑丈だ:4
合計	50

語義	データ数
語義 1	激しい:45
語義 2	頑丈だ:5
合計	50

表 5.30: 言い換えによって増えたデータ (形容詞)

語義	文数
語義 1	激しい:702
語義 2	頑丈だ:78
合計	780

表 5.28 から表 5.30 のデータを用いた実験の結果を表 5.31 に示す。

表 5.31: 利用する学習データとその正解率 (形容詞)

手法	正解率
手法 1:SemEval2 の学習データのみを利用する手法	0.90 (45/50)
手法 2:SemEval2 の学習データ + 言い換えによって増えた学習データを利用する手法	0.92 (46/50)
手法 3:言い換えによって増えた学習データのみを利用する手法	0.94 (47/50)
手法 4:再頻出語義を常に出力する手法	0.92 (46/50)

表 5.31 より、「手法 1:SemEval2 の学習データのみを利用する手法」の正解率が 0.90 に対して、「手法 2:SemEval2 の学習データ + 言い換えによって増えた学習データを利用する手法」の正解率が 0.92 となり、正解率はほぼ変わらなかった。

### 5.3.2 考察

「手法 1:SemEval2 の学習データのみを利用する手法」の正解率が 0.90 に対して、「手法 2:SemEval2 の学習データ + 言い換えによって増えた学習データを利用する手法」の正解率が 0.92 となり、正解率はほぼ変わらなかった。正解率はほぼ変わらなかったが、「手法 3:言い換えによって増えたデータのみを利用する手法」の正解率が 0.94 となっており、正解率が一番良かった。言い換えによって増えたデータが少し役にたったといえる。1 単語でしか実験を行うことができていないので、単語を増やして実験を行いたい。

## 第6章 今後の課題

以下のことが今後の課題である。

- 形容詞の多義語が1単語でしかできていないので、数を増やして実験を行いたい。
- 名詞の多義語だけでも6単語でしかできていないので、実験規模の拡大を行いたい。
- 名詞の多義語は「手法5:クロスバリデーションによる手法選択」の実験を行ったが、動詞と形容詞では「手法5:クロスバリデーションによる手法選択」の実験を行えていない。動詞と形容詞でも実験を行いたい。
- 「言い換えによって増えた学習データ」の誤りが、正解率を下げていることが考えられるので、「言い換えによって増えた学習データ」の誤りの割合を出して、その割合が多い場合その学習データを修正していきたい。
- 言い換えによって増えた学習データを人手で評価し、分析を行いたい。
- 機械学習として最大エントロピー法を用いているが、別の機械学習としてサポートベクターマシン法を用いて実験を行いたい。
- 本研究は、最大エントロピー法を用いているので素性分析を行うことができる。素性分析を行えていないので、行いたい。

## 第7章 おわりに

本研究では、機械学習を用いて多義語の多義性解消を行った。名詞の実験では、「手法 1:SemEval2 の学習データのみを利用する手法」の正解率が 0.733 に対して、「手法 5:クロスバリデーションによる手法選択」の正解率が 0.757 となっており、正解率が向上した。また、「手法 5:クロスバリデーションによる手法選択」の正解率 0.757 が最も良い正解率となった。

また、多義語 6 単語で合計 12 回の実験を行ったが、12 回の実験中 10 回は正しく手法選択することができた。

SemEval2 の学習データ数を言い換えによって増えたデータ数の X 倍 (1/10, 1/5, 1/2, 1, 2, 5, 10) のデータ数に変更する実験では、「手法 2:SemEval2 と言い換えによって増えたデータを用いる手法」は、SemEval2 の学習データ数が増えるにつれて性能が向上する傾向がみられた。

動詞の実験では、「手法 1:SemEval2 の学習データのみを利用する手法」の正解率 0.913 に対して、「手法 2:SemEval2 の学習データ + 言い換えによって増えた学習データを利用する手法」の正解率 0.893 となり、正解率が低下した。

形容詞の実験では、「手法 1:SemEval2 の学習データのみを利用する手法」の正解率が 0.90 に対して、「手法 2:SemEval2 の学習データ + 言い換えによって増えた学習データを利用する手法」の正解率が 0.92 となり、正解率はほぼ変わらなかった。

# 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,研究の進め方や本論文の書き方など,御指導をいただきました,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,村上仁一准教授に心から御礼申し上げます。

## 参考文献

- [1] 戸田勇馬. 機械学習と言い換えを利用した日本語単語の多義性解消. 卒業論文, 鳥取大学知能情報工学科, 2016.
- [2] Okumura Manabu, Shirai Kiyooki, Kanako Komiya, and Yokono Hikaru. n semeval-2010 japanese wsd task. 自然言語処理, Vol. 18, No. 3, pp. 293–307, 2011.
- [3] 新納浩幸, 白井清昭, 村田真樹, 福本文代, 藤田早苗, 佐々木稔, 古宮嘉那子, 乾孝司. 語義曖昧性解消の誤り分析. 言語処理学会第 21 回年次大会ワークショップ「エラー分析ワークショップ」, 2015.
- [4] Rada Mihalcea and Dan I. Moldovan. An automatic method for generatingb sense tagged corpora. *In Proceedings of the American Association for Artificial Intelligence(AAAI-1999)*, pp. 461–466, 1999.
- [5] 村田真樹ら. SENSEVSAL2J 辞書タスクでの CRL の取り組み 日本語単語の多義性解消における種々の機械学習手法と素性の比較 . 自然言語処理, Vol. 10, No. 3, pp. 115–133, 2003.
- [6] 藤田早苗, Kevin Duh, 藤野昭典, 平博順, 進藤裕之. 日本語語義曖昧性解消のための訓練データの自動拡張. 自然言語処理, Vol. 18, No. 3, pp. 273–291, 2011.
- [7] Eric Sven Ristad. Maximum Entropy Modeling for Natural Language. ACL/EACL Tutorial Program, Madrid, 1997.
- [8] Eric Sven Ristad. Maximum Entropy Modeling Toolkit, Release 1.6 beta. <http://www.mnemonic.com/software/memt>, 1998.
- [9] 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 回年次大会, pp. 300–303, 2011.

# 外部発表の成果

これまでの研究成果を以下に示す。

- 戸田勇馬, 村田真樹, 馬青. 言い換えと機械学習を用いた日本語単語の多義性解消. 言語処理学会第 22 回年次大会, pp.172-175, 2016.3.
- Yuma Toda, Masaki Murata and Qing Ma, Japanese Word Sense Disambiguation Using Paraphrasing and Machine Learning, 2016 Joint 8th International Conference on Soft Computing and Intelligent Systems and 2016 17th International Symposium on Advanced Intelligent Systems (SCIS ISIS 2016), pp.966-971 Sapporo, Hokkaido, August 27, 2016.