# 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，一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．また，新納ら [1] の研究により，多義性解消の誤りの原因の約 7 割が，学習データの不足によって起こっていることがわかった．そこで，学習データを増やすべきであると考えた．本研究では多義語の言い換えを利用することで，自動で学習データを作成し，データ数を増やす．また，その学習データに基づき機械学習を用いて多義性解消を行う．

実験の結果,「言い換えによって増えた学習データ」を SemEval2 の学習データの語義ごとのデータ数の比率に合うようにデータ数を変更した場合の方が,「言い換えによって増えた学習データ」をそのままの数で使用するよりも 4 単語すべての正解率では良い性能となった.「言い換えによって増えた学習データ」を SemEval2 の学習データの語義ごとのデータ数の比率に合うようにデータ数を変更した場合について述べる．最大エントロピー法では「SemEval2 の学習データと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，サポートベクトルマシン法では「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」が一番良い性能となった．正解率は最大エントロピー法で 0.76 となり，サポートベクトルマシン法で 0.78 となった．また，最大エントロピー法とサポートベクトルマシン法は，サポートベクトルマシン法の方が少し良い性能となったが，ほぼ同等の性能となった．最大エントロピー法は「SemEval2 の学習データのみを利用する手法」の正解率が 0.73 という性能に対して「SemEval2 の学習データと言い換えによって増えた学習データを利用する手法」が正解率 0.77 という性能で，言い換えによって増えた学習データを追加する前より性能が向上した．また,「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」でも正解率 0.76 という性能となり，この場合でもある程度とけることがわかった．サポートベクトルマシン法は「SemEval2 の学習データのみを利用する手法」の正解率が 0.74 という性能に対し「SemEval2 の学習データと言い換えによって増えた学習データを利用する手法」が正解率 0.77 という性能で，言い換えによって増えた学習データを追加する前より性能が

向上した．また,「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」でも正解率 0.78 という性能となり，この場合でもある程度とけることがわかった．

単語ごとについては「SemEval2 の学習データのみを用いる手法」の正解率が低かった「意味」と「子供」は,「SemEval2 の学習データ」に「言い換えによって増えた学習データ」を追加した場合，追加する前より両者 (ME と SVM) とも性能が向上した．「SemEval2 の学習データのみを用いる手法」の正解率が高かった「情報」については，追加した後のほうが両者とも少し性能が下がった．「SemEval2 の学習データのみを用いる手法」の正解率が低かった「他」は追加した後も両者とも性能は変わらなかった．

# 目 次

# 表 目 次

# 図 目 次

# 第1章　はじめに

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，一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．また，新納ら [1] の研究により，多義性解消の誤りの原因の約 7 割が，学習データの不足によって起こっていることがわかった．そこで，学習データを増やすべきであると考えた．

本研究では多義語の言い換えを利用することで，自動で学習データを作成し，データ数を増やす．また，その学習データに基づき機械学習を用いて多義性解消を行う．

本研究の主な主張点を以下に整理する．

- 多義性解消の誤りの原因である「学習データの量の不足」に着目し，本研究では，言い換えを利用して学習データの量を増やす．
- 今回の実験では「言い換えによって増えた学習データ」を SemEval2 の学習データの語義ごとのデータ数の比率に合うようにデータ数を変更した場合の方が，「言い換えによって増えた学習データ」をそのままのデータ数で使うよりも良い結果となった．データ数を変更する前の全単語の最高の正解率は，最大エントロピー法の「SemEval2 の学習データのみを利用する手法」で正解率 0.73，サポートベクトルマシン法の「SemEval2 の学習データのみを利用する手法」で正解率 0.74 であったのに対して，データ数を変更後は，最大エントロピー法の「SemEval2 の学習データと言い換えによって増えた学習データを利用する手法」で正解率 0.77，サポートベクトルマシン法の「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」で正解率 0.78 であった．
- 言い換えによって増えた学習データ数を変更した場合の全単語の正解率は，最大エントロピー法で「SemEval2 の学習データのみを利用する手法」の正解率が 0.73 という性能に対して「SemEval2 の学習データと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 0.77 という性能で，言い換えによって増えた学習データを追加する前より性能が向上した．

- 言い換えによって増えた学習データ数を変更した場合の全単語の正解率は，サポートベクトルマシン法で「SemEval2 の学習データのみを利用する手法」の正解率が 0.74 という性能に対し「SemEval2 の学習データと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 0.77 という性能で，言い換えによって増えた学習データを追加する前より性能が向上した．

- 単語ごとの特徴としては，言い換えによって増えた学習データ数を変更した場合に「SemEval2 の学習データのみを用いる手法」の正解率が低かった多義語は「SemEval2 の学習データ」に「言い換えによって増えた学習データ」を追加した場合，追加する前より両者 (ME と SVM) とも性能が向上した．「SemEval2 の学習データのみを用いる手法」の正解率が高かった多義語については，追加した後のほうが両者とも少し性能が下がるか，性能は変わらない結果となった．

- 素性分析を行った結果，有効である素性がわかった．例えば「子供」について「生徒」「教諭」といった単語が文中の近くに出現した場合，語義 1 の意味で使われることが多く，「父親」「夫」「妻」といった単語が文中の近くに出現した場合，語義 2 の意味で使われることが多かった．

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，今後の課題について述べる．第 6 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 [1] 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）を対応する事例を用いてベクトル化し，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[2] 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．WordNet 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．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 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 [3] 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．

## 2.4 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 [4] 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し，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 [2] 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，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 [3] 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 [2]．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．[1]

[1]最大エントロピー法は素性分析に便利であるので利用した．

## 3.2　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 3 つ持つものとする．まず，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している．定義文中の語を選定する場合が多いが，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場合もある．辞典の 1 番目の語義を特徴付ける語を x1，辞典の 2 番目の語義を特徴付ける語を x2，辞典の 3 番目の語義を特徴付ける語を x3 とする．そして x1，x2，x3 を含む文を新聞から抜き出す．そして，抜き出した文から x1，x2，x3 を X に言い換える．このとき x1 を X に言い換えた場合，言い換えた後の X は辞典の 1 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．

先行研究 [2] 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 [2] 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場合もあることが先行研究 [2] と異なる点である．

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壌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。 |

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 | 着陸の動機は明らかにされていない。 | 着陸の意味は明らかにされていない | 語義 2 |
| 価値 | 一票の価値が最も低い神奈川四区と… | 一票の意味が最も低い神奈川四区と… | 語義 3 |

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.

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．そして，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.

多義語を X とし，その多義語 X は語義が 2 つある場合を考える．まず，辞典の複数ある語義をそれぞれ特徴付けた語を人手で選定する．ここでは，辞典の 1 番目の語義を特徴付けた語を x1，辞典の 2 番目の語義を特徴付けた語を x2 とする．そして x1，x2 を含む文を新聞から抜き出す．そして，抜き出した文から x1，x2 を X に言い換える．このとき x1 を X に置き換えた場合，言い換えた後の X は辞典の 1 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とサポートベクトルマシンを利用する．[2]

## 3.3 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 [5, 6].

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

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

ただし，$A, B$ は分類と文脈の集合を意味し，$g_j(a,b)$ は 文脈 $b$ に素性 $f_j$ があってなおかつ分類が $a$ の場合 1 となりそれ以外で 0 となる関数を意味する．また，$\tilde{p}(a,b)$ は，既知データでの $(a,b)$ の出現の割合を意味する．

式 (3.1) は確率 $p$ と出力と素性の組の出現を意味する関数 $g$ をかけることで出力と素性の組の頻度の期待値を求めることになっており，右辺の既知データにおける期待値と，左辺の求める確率分布に基づいて計算される期待値が等しいことを制約として，エントロピー最大化 (確率分布の平滑化) を行って，出力と文脈の確率分布を求めるものとなっている [3].

## 3.4 サポートベクトルマシン法

サポートベクトルマシン法は，空間を超平面で分割することにより 2 つの分類からなるデータを分類する手法である．このとき，2 つの分類が正例と負例からなるもの

[2]最大エントロピー法は素性の分析に便利であるため，利用した.

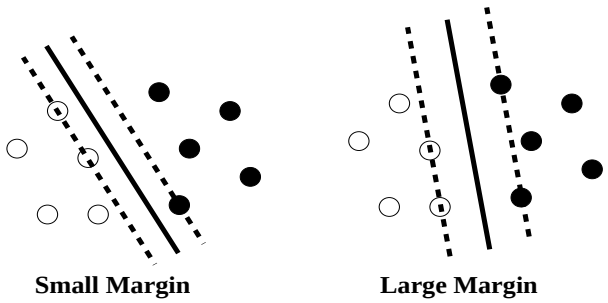


図 3.1: マージン最大化

とすると，学習データにおける正例と負例の間隔 (マージン) が大きいもの (図 3.1 参照[3]) ほどオープンデータで誤った分類をする可能性が低いと考えられ，このマージンを最大にする超平面を求めそれを用いて分類を行う．基本的には上記のとおりであるが，通常，学習データにおいてマージンの内部領域に少数の事例が含まれてもよいとする手法の拡張や，超平面の線形の部分を非線型にする拡張 (カーネル関数の導入) がなされたものが用いられる．この拡張された方法は，以下の識別関数を用いて分類することと等価であり，その識別関数の出力値が正か負かによって二つの分類を判別することができる [7, 8].

$$
\begin{aligned}
f(\mathbf{x}) &= sgn\left(\sum_{i=1}^{l} \alpha_i y_i K(\mathbf{x}_i, \mathbf{x}) + b\right) \\
b &= -\frac{max_{i,y_i=-1} b_i + min_{i,y_i=1} b_i}{2} \\
b_i &= \sum_{j=1}^{l} \alpha_j y_j K(\mathbf{x}_j, \mathbf{x}_i)
\end{aligned}
\tag{3.3}
$$

ただし，$\mathbf{x}$ は識別したい事例の文脈 (素性の集合) を，$\mathbf{x}_i$ と $y_i (i = 1, ..., l, y_i \in \{1, -1\})$ は学習データの文脈と分類先を意味し，関数 $sgn$ は，

$$
\begin{aligned}
sgn(x) = \quad & 1 \quad (x \geq 0) \\
& -1 \quad (otherwise)
\end{aligned}
\tag{3.4}
$$

であり，また，各 $\alpha_i$ は式 (3.6) と式 (3.7) の制約のもと式 (3.5) の $L(\alpha)$ を最大にする場合のものである．

$$
L(\alpha) = \sum_{i=1}^{l} \alpha_i - \frac{1}{2} \sum_{i,j=1}^{l} \alpha_i \alpha_j y_i y_j K(\mathbf{x_i}, \mathbf{x_j}) \tag{3.5}
$$

[3]図の白丸，黒丸は，正例，負例を意味し，実線は空間を分割する超平面を意味し，破線はマージン領域の境界を表す面を意味する．

$$0 \leq \alpha_i \leq C \ (i = 1, ..., l) \tag{3.6}$$

$$\sum_{i=1}^{l} \alpha_i y_i = 0 \tag{3.7}$$

また，関数 $K$ はカーネル関数と呼ばれ，様々なものが用いられるが本稿では以下の多項式のものを用いる [3].

$$K(\mathbf{x}, \mathbf{y}) \ = (\mathbf{x} \cdot \mathbf{y} + 1)^d \tag{3.8}$$

# 第4章　実験

## 4.1　実験方法

本研究では，毎日新聞の 1 年分 (1991 年) のデータを用いる．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を出す．本研究では，SemEval2 の対象単語 50 個のうち，名詞 4 個を実験に使用する多義語とする．SemEval2 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，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．

実験に用いる学習データを変えることで，それぞれを用いた場合の結果を比較し，言い換えに基づく学習データの増加の有効性の確認を行う．用いる学習データは「SemEval2 の学習データのみ」「SemEval2 の学習データと言い換えによって増えた学習データ」「言い換えによって増えた学習データのみ」の 3 種類である．評価は，多義語の語義を 1 つに絞る際，その語義が正しいかを評価する．また,「言い換えによって増えた学習データ」をそのままのデータ数で実験を行う場合と，SemEval2 の学習データの語義ごとのデータ数の比率に合うようにデータ数を変更して実験を行う，2 種類の実験がある．

機械学習は最大エントロピー法とサポートベクトルマシンを使用する．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 |

## 4.2 単語の選定

本研究では，SemEval2 の対象単語 50 個のうち名詞「子供」「情報」の計 2 個を実験対象の単語 (多義語) とする．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．

「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．

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

- 語義 1 ある物事の事情についての 知らせ。
- 語義 2 それを通して何らかの 知識 が得られるようなもの。

「情報」の場合，語義 1 を「知らせ」，語義 2 を「知識」とした．

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

- 語義 1 ある規準・範囲に含まれない部分。
- 語義 2 それ以外ではないという気持ちで言う。

「他」の場合，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「よそ」，語義 2 を「しか」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．

## 4.3 実験結果

「意味」「子供」「他」「情報」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 (ME) またはサポートベクトルマシン (SVM) を用いた．表 4.2 から

表 4.5 に SemEval2 の「意味」「子供」「他」「情報」についての事例数を示す.「意味」には，辞典に基づく語義が 3 つある.「子供」「他」「情報」には，辞典に基づく語義が 2 つある.

表 4.2: 「意味」の事例数

| | 学習データ | テストデータ |
|---|---|---|
| 語義 1 | 25 | 27 |
| 語義 2 | 8 | 10 |
| 語義 3 | 17 | 12 |
| 未知語義 | 0 | 1 |
| 総数 | 50 | 50 |

表 4.3: 「子供」の事例数

| | 学習データ | テストデータ |
|---|---|---|
| 語義 1 | 26 | 18 |
| 語義 2 | 24 | 32 |
| 総数 | 50 | 50 |

表 4.4: 「他」の事例数

| | 学習データ | テストデータ |
|---|---|---|
| 語義 1 | 49 | 50 |
| 語義 2 | 1 | 0 |
| 総数 | 50 | 50 |

表 4.5: 「情報」の事例数

| | 学習データ | テストデータ |
|---|---|---|
| 語義 1 | 4 | 8 |
| 語義 2 | 46 | 42 |
| 総数 | 50 | 50 |

表 4.6 から表 4.9 に「意味」「子供」「他」「情報」についての毎日新聞 1 年分のデータを使用し，言い換えによって増えた学習データ数を示す.

表 4.6: 増えた学習データ：データ数そのまま (「意味」)

| | 増えた学習データ |
|---|---|
| 語義 1 | 4403 |
| 語義 2 | 370 |
| 語義 3 | 1177 |
| 総数 | 5950 |

表 4.7: 増えた学習データ：データ数そのまま (「子供」)

| | 増えた学習データ |
|---|---|
| 語義 1 | 997 |
| 語義 2 | 783 |
| 総数 | 1780 |

表 4.8: 増えた学習データ：データ数そのまま (「他」)

| | 増えた学習データ |
|---|---|
| 語義 1 | 490 |
| 語義 2 | 11708 |
| 総数 | 12198 |

表 4.9: 増えた学習データ：データ数そのまま (「情報」)

| | 増えた学習データ |
|---|---|
| 語義 1 | 157 |
| 語義 2 | 477 |
| 総数 | 634 |

表 4.10 から表 4.13 に「意味」「子供」「他」「情報」についての毎日新聞 1 年分のデータを使用し，言い換えによって増えた学習データ数を示す．なお，この学習データは，SemEval2 の学習データの語義ごとのデータ数の比率に合うようにデータ数を変更した．

表 4.10: 増えた学習データ：データ数変更 (「意味」)

| | 増えた学習データ |
|---|---|
| 語義 1 | 1156 |
| 語義 2 | 370 |
| 語義 3 | 786 |
| 総数 | 2312 |

表 4.11: 増えた学習データ：データ数変更 (「子供」)

| | 増えた学習データ |
|---|---|
| 語義 1 | 848 |
| 語義 2 | 783 |
| 総数 | 1631 |

表 4.12: 増えた学習データ：データ数変更 (「他」)

| | 増えた学習データ |
|---|---|
| 語義 1 | 490 |
| 語義 2 | 10 |
| 総数 | 500 |

表 4.13: 増えた学習データ：データ数変更 (「情報」)

| | 増えた学習データ |
|---|---|
| 語義 1 | 41 |
| 語義 2 | 471 |
| 総数 | 512 |

「言い換えによって増えた学習データ」をそのままの数で使用した場合において，3 種類の学習データを利用する手法と SemEval2 の学習データの最頻出語義を常に出力する手法の最大エントロピー法での多義性解消の結果を表 4.14 に示す.

表 4.14: 利用する学習データとその正解率：データ数そのまま (ME)

| 手法 | 正解率 (意味) | 正解率 (子供) | 正解率 (他) | 正解率 (情報) | 正解率 (4 単語全て) |
|---|---|---|---|---|---|
| SemEval2 の学習データの最頻出語義を常に出力する手法 | 0.54 (27/50) | 0.36 (18/50) | 1.00 (50/50) | 0.84 (42/50) | 0.68 (137/200) |
| SemEval2 の学習データのみを利用する手法 | 0.50 (25/50) | 0.56 (28/50) | 1.00 (50/50) | 0.86 (43/50) | 0.73 (146/200) |
| 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 | 0.62 (31/50) | 0.64 (32/50) | 0.66 (33/50) | 0.82 (41/50) | 0.68 (137/200) |
| 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 | 0.60 (30/50) | 0.68 (34/50) | 0.52 (26/50) | 0.80 (40/50) | 0.65 (130/200) |

4 単語全ての正解率では，正解率 0.73 で「SemEval2 の学習データのみを利用する手法」が一番良い性能となった.

「言い換えによって増えた学習データ」を SemEval2 の学習データの語義ごとのデータ数の比率に合うようにデータ数を変更した場合において，3 種類の学習データを利用する手法と SemEval2 の学習データの最頻出語義を常に出力する手法の最大エントロピー法での多義性解消の結果を表 4.15 に示す.

表 4.15: 利用する学習データとその正解率：データ数変更 (ME)

| 手法 | 正解率 (意味) | 正解率 (子供) | 正解率 (他) | 正解率 (情報) | 正解率 (4 単語全て) |
|---|---|---|---|---|---|
| SemEval2 の学習データの最頻出語義を常に出力する手法 | 0.54 (27/50) | 0.36 (18/50) | 1.00 (50/50) | 0.84 (42/50) | 0.68 (137/200) |
| SemEval2 の学習データのみを利用する手法 | 0.50 (25/50) | 0.56 (28/50) | 1.00 (50/50) | 0.86 (43/50) | 0.73 (146/200) |
| 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 | 0.60 (30/50) | 0.64 (32/50) | 1.00 (50/50) | 0.84 (42/50) | 0.77 (154/200) |
| 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 | 0.54 (27/50) | 0.66 (33/50) | 1.00 (50/50) | 0.84 (42/50) | 0.76 (152/200) |

4 単語全ての正解率では，正解率 0.77 で「SemEval2 の学習データと言い換えによって増えた学習データを利用する手法」が一番良い性能となった.

「言い換えによって増えた学習データ」をそのままの数で使用した場合において，3 種類の学習データを利用する手法と SemEval2 の学習データの最頻出語義を常に出力する手法のサポートベクトルマシンでの多義性解消の結果を表 4.16 に示す.

表 4.16: 利用する学習データとその正解率：データ数そのまま (SVM)

| 手法 | 正解率 (意味) | 正解率 (子供) | 正解率 (他) | 正解率 (情報) | 正解率 (4 単語全て) |
|---|---|---|---|---|---|
| SemEval2 の学習データの最頻出語義を常に出力する手法 | 0.54 (27/50) | 0.36 (18/50) | 1.00 (50/50) | 0.84 (42/50) | 0.68 (137/200) |
| SemEval2 の学習データのみを利用する手法 | 0.56 (28/50) | 0.54 (27/50) | 1.00 (50/50) | 0.88 (44/50) | 0.74 (149/200) |
| 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 | 0.58 (29/50) | 0.68 (34/50) | 0.66 (33/50) | 0.82 (41/50) | 0.68 (137/200) |
| 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 | 0.68 (34/50) | 0.56 (28/50) | 0.66 (33/50) | 0.82 (41/50) | 0.67 (135/200) |

4 単語全ての正解率では，正解率 0.74 で「SemEval2 の学習データのみを利用する手法」が一番良い性能となった.

「言い換えによって増えた学習データ」を SemEval2 の学習データの語義ごとのデータ数の比率に合うようにデータ数を変更場合に置いて，3 種類の学習データを利用する手法と SemEval2 の学習データの最頻出語義を常に出力する手法のサポートベクトルマシンでの多義性解消の結果を表 4.17 に示す.

表 4.17: 利用する学習データとその正解率：データ数変更 (SVM)

| 手法 | 正解率 (意味) | 正解率 (子供) | 正解率 (他) | 正解率 (情報) | 正解率 (4 単語全て) |
|---|---|---|---|---|---|
| SemEval2 の学習データの最頻出語義を常に出力する手法 | 0.54 (27/50) | 0.36 (18/50) | 1.00 (50/50) | 0.84 (42/50) | 0.68 (137/200) |
| SemEval2 の学習データのみを利用する手法 | 0.56 (28/50) | 0.54 (27/50) | 1.00 (50/50) | 0.88 (44/50) | 0.74 (149/200) |
| 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 | 0.60 (30/50) | 0.64 (32/50) | 1.00 (50/50) | 0.86 (43/50) | 0.77 (155/200) |
| 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 | 0.60 (30/50) | 0.72 (36/50) | 1.00 (50/50) | 0.82 (41/50) | 0.78 (157/200) |

4 単語全ての正解率では，正解率 0.78 で「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」が一番良い性能となった．

## 4.4 有意差検定

3 種類の学習データを利用する手法と SemEval2 の学習データの最頻出語義を常に出力する手法を用いる手法の計 4 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．4 単語すべての 200 事例を利用して，符号検定を行った．表 4.18 から表 4.21 に有意差検定の結果を示す．

最大エントロピー法で「言い換えによって増えた学習データ」をそのままの数で使用した場合について述べる (表 4.18).「SemEval2 の学習データの最頻出語義を常に出力する手法」と「SemEval2 の学習データのみ」または「SemEval2 の学習データと言い換えによって学習データ」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,「SemEval2 の学習データのみ」と「SemEval2 の学習データと言い換えによって学習データ」，または「SemEval2 の学習データの最頻出語義を常に出力する手法」と「SemEval2 の学習データと言い換えによって学習データ」では，有意水準 0.05 の符号検定（片側検定）で有意差があった．他の組み合わせでは，有意水準 0.05 の符号検定（片側検定）で有意差がなかった．

最大エントロピー法で「言い換えによって増えた学習データ」を SemEval2 の学習

表 4.18: 有意差検定結果：データ数そのまま (ME)

| 比べる手法 | | 有意差の有無 (有意水準 0.05 の片側検定) |
|---|---|---|
| 「SemEval2 の学習データのみ」 | 「SemEval2 の学習データ+言い換えによって増えた学習データ」 | 有意差なし |
| 「SemEval2 の学習データのみ」 | 「言い換えによって増えた学習データのみ」 | 有意差あり |
| 「SemEval2 の学習データ+言い換えによって増えた学習データ」 | 「言い換えによって増えた学習データのみ」 | 有意差なし |
| 「SemEval2 の学習データの最頻出語義を常に出力する手法」 | 「SemEval2 の学習データのみ」 | 有意差あり |
| 「SemEval2 の学習データの最頻出語義を常に出力する手法」 | 「SemEval2 の学習データ+言い換えによって増えた学習データ」 | 有意差あり |
| 「SemEval2 の学習データの最頻出語義を常に出力する手法」 | 「言い換えにって増えた学習データのみ」 | 有意差あり |

データの語義ごとのデータ数の比率に合うようにデータ数を変更した場合について述べる (表 4.19).「SemEval2 の学習データの最頻出語義を常に出力する手法」とそれ以外の 3 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．他の組み合わせでは，有意水準 0.05 の符号検定（片側検定）で有意差がなかった．

表 4.19: 有意差検定結果：データ数変更 (ME)

| 比べる手法 | | 有意差の有無 (有意水準 0.05 の片側検定) |
|---|---|---|
| 「SemEval2 の学習データのみ」 | 「SemEval2 の学習データ+言い換えによって増えた学習データ」 | 有意差なし |
| 「SemEval2 の学習データのみ」 | 「言い換えによって増えた学習データのみ」 | 有意差なし |
| 「SemEval2 の学習データ+言い換えによって増えた学習データ」 | 「言い換えによって増えた学習データのみ」 | 有意差なし |
| 「SemEval2 の学習データの最頻出語義を常に出力する手法」 | 「SemEval2 の学習データのみ」 | 有意差あり |
| 「SemEval2 の学習データの最頻出語義を常に出力する手法」 | 「SemEval2 の学習データ+言い換えによって増えた学習データ」 | 有意差あり |
| 「SemEval2 の学習データの最頻出語義を常に出力する手法」 | 「言い換えにって増えた学習データのみ」 | 有意差あり |

サポートベクトルマシンで「言い換えによって増えた学習データ」をそのままの数で使用した場合について述べる (表 4.20).「SemEval2 の学習データの最頻出語義を常に出力する手法」とそれ以外の 3 手法,「SemEval2 の学習データのみを利用する手法」と「言い換えによって学習データのみを利用する手法」では，有意水準 0.05 のの符号検定（前者は片側検定と両側検定の両方，後者は片側検定）で有意差があった．他の組み合わせでは，有意水準 0.05 の符号検定（片側検定）で有意差がなかった．

表 4.20: 有意差検定結果：データ数そのまま (SVM)

| 比べる手法 | | 有意差の有無 (有意水準 0.05 の片側検定) |
|---|---|---|
| 「SemEval2 の学習データのみ」 | 「SemEval2 の学習データ+言い換えによって増えた学習データ」 | 有意差なし |
| 「SemEval2 の学習データのみ」 | 「言い換えによって増えた学習データのみ」 | 有意差あり |
| 「SemEval2 の学習データ+言い換えによって増えた学習データ」 | 「言い換えによって増えた学習データのみ」 | 有意差なし |
| 「SemEval2 の学習データの最頻出語義を常に出力する手法」 | 「SemEval2 の学習データのみ」 | 有意差あり |
| 「SemEval2 の学習データの最頻出語義を常に出力する手法」 | 「SemEval2 の学習データ+言い換えによって増えた学習データ」 | 有意差あり |
| 「SemEval2 の学習データの最頻出語義を常に出力する手法」 | 「言い換えにって増えた学習データのみ」 | 有意差あり |

サポートベクトルマシンで「言い換えによって増えた学習データ」を SemEval2 の学習データの語義ごとのデータ数の比率に合うようにデータ数を変更した場合について述べる (表 4.21).「SemEval2 の学習データの最頻出語義を常に出力する手法」とそれ以外の 3 手法では，有意水準 0.05 のの符号検定（片側検定と両側検定の両方）で有意差があった．他の組み合わせでは，有意水準 0.05 の符号検定（片側検定）で有意差がなかった．

以上のことから，最大エントロピー法とサポートベクトルマシン法の両方で「言い換えによって増えた学習データ」を SemEval2 の学習データの語義ごとのデータ数の比率に合うようにデータ数を変更した場合，有意差検定で「言い換えによって増えた学習データ」の有効性はなかった．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）で有意差があった．実験に使用する多義語を増やして実験を行っていきたい．

表 4.21: 有意差検定結果：データ数変更 (SVM)

| 比べる手法 | | 有意差の有無 (有意水準 0.05 の片側検定) |
|---|---|---|
| 「SemEval2 の学習データのみ」 | 「SemEval2 の学習データ+言い換えによって増えた学習データ」 | 有意差なし |
| 「SemEval2 の学習データのみ」 | 「言い換えによって増えた学習データのみ」 | 有意差なし |
| 「SemEval2 の学習データ+言い換えによって増えた学習データ」 | 「言い換えによって増えた学習データのみ」 | 有意差なし |
| 「SemEval2 の学習データの最頻出語義を常に出力する手法」 | 「SemEval2 の学習データのみ」 | 有意差あり |
| 「SemEval2 の学習データの最頻出語義を常に出力する手法」 | 「SemEval2 の学習データ+言い換えによって増えた学習データ」 | 有意差あり |
| 「SemEval2 の学習データの最頻出語義を常に出力する手法」 | 「言い換えにって増えた学習データのみ」 | 有意差あり |

## 4.5 素性分析

最大エントロピー法では素性の重みを考察することで役立った素性がわかる．最大エントロピー法で「言い換えによって増えた学習データ」を SemEval2 の学習データの語義ごとのデータ数の比率に合うようにデータ数を変更した場合について，素性に基づく分析を行った結果，以下のことがわかった．

表 4.22 から表 4.23 に，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．表中の素性番

号は，4 章の表 4.1 を参考にしている．

表 4.22: 「意味」について機械学習が参考にした素性例

| 語義 1(「内容」) | | 語義 2(「動機」) | | 語義 3(「価値」) | |
|---|---|---|---|---|---|
| 素性 | 正規化 α 値 | 素性 | 正規化 α 値 | 素性 | 正規化 α 値 |
| 素性 1：具体 | 0.59 | 素性 1：犯行 | 0.61 | 素性 1：観 | 0.61 |
| 素性 1：検討 | 0.50 | 素性 1：容疑 | 0.54 | 素性 1：主義 | 0.53 |
| 素性 1：活動 | 0.47 | 素性 1：捜査 | 0.50 | 素性 1：付加 | 0.51 |

表 4.23: 「子供」について機械学習が参考にした素性例

| 語義 1(「児童」) | | 語義 2(「息子」) | |
|---|---|---|---|
| 素性 | 正規化 α 値 | 素性 | 正規化 α 値 |
| 素性 1：教諭 | 0.77 | 素性 1：父親 | 0.79 |
| 素性 1：小学校 | 0.74 | 素性 1：夫 | 0.67 |
| 素性 1：生徒 | 0.68 | 素性 1：妻 | 0.65 |

表 4.24: 「他」について機械学習が参考にした素性例

| 語義 1(「よそ」) | | 語義 2(「しか」) | |
|---|---|---|---|
| 素性 | 正規化 α 値 | 素性 | 正規化 α 値 |
| デフォルト素性 | 0.69 | 素性 31：ない | 0.56 |
| 素性 2：の | 0.54 | 素性 31：いる | 0.53 |

表 4.25: 「情報」について機械学習が参考にした素性例

| 語義 1(「知らせ」) | | 語義 2(「知識」) | |
|---|---|---|---|
| 素性 | 正規化 α 値 | 素性 | 正規化 α 値 |
| 素性 1：ビジネス | 0.58 | デフォルト素性 | 0.60 |
| 素性 1：午前 | 0.57 | 素性 1：的 | 0.58 |

「意味」は「具体」「検討」「活動」といった単語が文中の近くに出現した場合，語義 1(「内容」) の意味で使われることが多かった.「犯行」「容疑」「捜査」や助詞の「の」といった単語が文中の近くに出てきた場合，語義 2(「動機」) の意味で使われることが多かった.「主義」「付加」や見方，様子を表す「観」といった単語が文中の近くに出現した場合，語義 3(「価値」) の意味で使われることが多かった.

「子供」は「小学校」「生徒」「教諭」といった単語が文中の近くに出現した場合，語義 1(「児童」) の意味で使われることが多かった.「父親」「夫」「妻」といった単語が文中の近くに出現した場合，語義 2(「息子」) の意味で使われることが多かった.

「他」は助詞の「の」が後接したり，他に何も情報がなければ語義 1(「よそ」) の意味で使われることが多かった.「ない」「いる」といった単語が文中の近くに出現した場合，語義 2(「しか」) の意味で使われることが多かった.

「情報」は「ビジネス」「午前」といった単語が文中に出現した場合語義 1(「知らせ」) の意味で使われることが多かった.「者」「的」といった単語が文中に出現したり，他に何も情報がなければ語義 2(「知識」) の意味で使われることが多かった.

## 4.6 考察

### 4.6.1 最大エントロピー法で 4 単語すべての正解率 (言い換えによって増えた学習データ数：そのまま)

最大エントロピー法で「言い換えによって増えた学習データ」をそのままの数で使用した場合について考察した．4 単語全ての正解率では,「SemEval2 の学習データのみを利用する手法」が一番良い性能となった．これは「言い換えによって増えた学習データ」が有効ではではないことを示している．例えば,「他」の「言い換えによって増えた学習データ」は語義 1 が 490，語義 2 が 11708 という偏った学習データ数となり，語義 2 という答えを出す可能性が高くなったと考えられる.

### 4.6.2 最大エントロピー法で単語ごとの正解率の考察 (言い換えによって増えた学習データ数：そのまま)

単語ごとの正解率について考察した.「意味」と「子供」の場合では,「SemEval2 の学習データのみを利用する手法」の正解率は低かった．このように,「SemEval2 の学習データのみを利用する手法」の正解率が低い場合,「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」の結果の方が良い性能となった.「他」と「情報」の場合は,「SemEval2 の学習データのみを利用する手法」の正解率は高かった．このように,「SemEval2 の学習データのみを利用する手法」の正解率が高い場合,「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」の方が劣る性能となった.

### 4.6.3 最大エントロピー法で 4 単語すべての正解率 (言い換えによって増えた学習データ数：変更)

最大エントロピー法で「言い換えによって増えた学習データ」を SemEval2 の学習データの語義ごとのデータ数の比率に合うようにデータ数を変更した場合について考察した．4 単語全ての正解率では,「SemEval2 の学習データと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」が一番良い性能となった.「SemEval2 の学習データのみを利用する手法」の正解率が 0.73 という性能に対し「SemEval2 の学習データと言い換えによって増えた学習データを利用する手法」が正解率 0.77 という性能で，言い換えによって増えた学習データを追加する前より性能が向上した．また,「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」でも正解率 0.76 という性能となり，この場合でもある程度とけることがわかった.

### 4.6.4 最大エントロピー法で単語ごとの正解率の考察 (言い換えによって増えた学習データ数：変更)

単語ごとの正解率について考察した.「意味」と「子供」の場合では,「SemEval2 の学習データのみを利用する手法」の正解率は低かった．このように,「SemEval2 の学習データのみを利用する手法」の正解率が低い場合,「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」の結果の方が良い性能となった.「情報」の場合は,「SemEval2 の学習データのみを利用する手法」の正解率は高かった．このように,「SemEval2 の学習データのみを利用する手法」の正解率が高い場合,「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」の方が劣る性能となった.「他」の場合は,「SemEval2 の学習データのみを利用する手法」と「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」の正解率は同じとなった.

### 4.6.5 サポートベクトルマシン法で 4 単語すべての正解率 (言い換えによって増えた学習データ数：そのまま)

サポートベクトルマシンで「言い換えによって増えた学習データ」をそのままの数で使用した場合について考察した．4 単語全ての正解率では,「SemEval2 の学習データのみを利用する手法」が一番良い性能となった．これは「言い換えによって増えた学習データ」があまり良いデータではないことを表している．例えば,「他」の「言い換えによって増えた学習データ」の数は語義 1 が 490，語義 2 が 11708 という偏った学習データ数となり，語義 2 という答えを出す可能性が高くなったと考えられる.

### 4.6.6 サポートベクトルマシン法で単語ごとの正解率の考察 (言い換えによって増えた学習データ数：そのまま)

単語ごとの正解率について考察した.「意味」と「子供」の場合では,「SemEval2 の学習データのみを利用する手法」の正解率は低かった．このように,「SemEval2 の学習データのみを利用する手法」の正解率が低い場合,「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」の結果の方が良い性能となった.「他」と「情報」の場合は,「SemEval2 の学習データのみを利用する手法」の正解率は高かった．このように,「SemEval2 の学習データのみを利用する手法」の正解率が高い場合,「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」の方が劣る性能となった.

### 4.6.7 サポートベクトルマシン法で 4 単語すべての正解率 (言い換えによって増えた学習データ数：変更)

サポートベクトルマシンで「言い換えによって増えた学習データ」を SemEval2 の学習データの語義ごとのデータ数の比率に合うようにデータ数を変更した場合について考察した．4 単語全ての正解率では,「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」が一番良い性能となった.「SemEval2 の学習データのみを利用する手法」の正解率が 0.74 という性能に対し「SemEval2 の学習データと言い換えによって増えた

学習データを利用する手法」が正解率 0.77 という性能で，言い換えによって増えた学習データを追加する前より性能が向上した．また,「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」でも正解率 0.78 という性能となり，この場合でもある程度とけることがわかった.

### 4.6.8 サポートベクトルマシン法で単語ごとの正解率の考察 (言い換えによって増えた学習データ数：変更)

単語ごとの正解率について考察した.「意味」と「子供」の場合では,「SemEval2 の学習データのみを利用する手法」の正解率は低かった．このように,「SemEval2 の学習データのみを利用する手法」の正解率が低い場合,「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」の結果の方が良い性能となった.「情報」の場合は,「SemEval2 の学習データのみを利用する手法」の正解率は高かった．このように,「SemEval2 の学習データのみを利用する手法」の正解率が高い場合,「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」の方が劣る性能となった.「他」の場合は,「SemEval2 の学習データのみを利用する手法」と「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」の正解率は同じとなった.

### 4.6.9 最大エントロピー法とサポートベクトルマシン法の比較

今回の実験では，4 単語すべての正解率はサポートベクトルマシン法の方が最大エントロピー法よりも少しではあるが良い性能となった．正解率は，前者が「SemEval2 の学習データと言い換えによって増えた学習データを利用する手法」が 0.77,「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 0.78 となり，後者が「SemEval2 の学習データと言い換えによって増えた学習データを利用する手法」が 0.77,「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 0.76 となった.

### 4.6.10 言い換えによって増えた学習データ数

「言い換えによって増えた学習データ」を SemEval2 の学習データの語義ごとのデータ数の比率に合うようにデータ数を変更した場合の方が，データ数をそのままで実験を行うよりも良い正解率となった．データ数をそのままの数で実験を行った場合「言い換えによって増えた学習データ」が有効ではないことを示している．例えば,「他」

の「言い換えによって増えた学習データ」は語義 1 が 490，語義 2 が 11708 という偏った学習データ数となり，語義 2 という答えを出す可能性が高くなるので「言い換えによって増えた学習データ」を SemEval2 の学習データの語義ごとのデータ数の比率に合った方が良い性能となると考えられる．

# 第5章　今後の課題

以下のことが今後の課題である．

- 「言い換えによって増えた学習データ」を SemEval2 の学習データに追加した場合，性能は向上したが，有意差検定では，有意差がなかった．実験に使用する多義語が 4 つと少ないので，今後は増やして実験を行いたい．
- 実験対象の多義語を増やす際，単語の選定を行いたい．
- 名詞でしか実験を行えていないので，他の品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．
- 正しい「言い換えによって学習データ」で正解率がどれくらい上がるのかを実験を行い,「言い換えによって学習データ」の有効性を確認したい．これは,「言い換えによって学習データ」でどれくらい性能が上がるのかを確認するためである．
- 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今後検討し，それを用いて性能を上げていきたい．
- 今回の実験では，新聞の 1 年分のデータでしか行えていないので，言い換えによって増えた学習データを増やすために 5 年分のデータで実験を行っていきたい．

# 第6章　おわりに

本研究では，機械学習を用いて多義性解消を行った．また，本研究では多義語の言い換えを利用することで自動で学習データを作成し，自動で学習データ数を増やし，その学習データに基づき機械学習を用いて多義性解消を行った．

実験の結果,「言い換えによって増えた学習データ」を SemEval2 の学習データの語義ごとのデータ数の比率に合うようにデータ数を変更した場合の方が,「言い換えによって増えた学習データ」をそのままの数で使用するよりも 4 単語すべての正解率では良い性能となった.「言い換えによって増えた学習データ」を SemEval2 の学習データの語義ごとのデータ数の比率に合うようにデータ数を変更した場合について述べる．最大エントロピー法では「SemEval2 の学習データと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，サポートベクトルマシン法では「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」が一番良い性能となった．正解率は最大エントロピー法で 0.76 となり，サポートベクトルマシン法で 0.78 となった．また，最大エントロピー法とサポートベクトルマシン法は，サポートベクトルマシン法の方が少し良い性能となったが，ほぼ同等の性能となった．最大エントロピー法は「SemEval2 の学習データのみを利用する手法」の正解率が 0.73 という性能に対し「SemEval2 の学習データと言い換えによって増えた学習データを利用する手法」が正解率 0.77 という性能で，言い換えによって増えた学習データを追加する前より性能が向上した．また,「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」でも正解率 0.76 という性能となり，この場合でもある程度とけることがわかった．サポートベクトルマシン法は「SemEval2 の学習データのみを利用する手法」の正解率が 0.74 という性能に対し「SemEval2 の学習データと言い換えによって増えた学習データを利用する手法」が正解率 0.77 という性能で，言い換えによって増えた学習データを追加する前より性能が向上した．また,「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」でも正解率 0.78 という性能となり，この場合でもある程度とけることがわかった．

単語ごとについては「SemEval2 の学習データのみを用いる手法」の正解率が低かった「意味」と「子供」は,「SemEval2 の学習データ」に「言い換えによって増えた学

習データ」を追加した場合，追加する前より両者 (ME と SVM) とも性能が向上した．「SemEval2 の学習データのみを用いる手法」の正解率が高かった「情報」については，追加した後のほうが両者とも少し性能が下がった．「SemEval2 の学習データのみを用いる手法」の正解率が低かった「他」は追加した後も両者とも性能は変わらなかった．今後は，実験で使用する単語の数を増やして実験を行っていきたい．

# 参考文献


[1] 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. クラスタリングを利用した語義曖昧性解消の誤り原因のタイプ分け. 自然言語処理, Vol. 22, No. 5, pp. 319–362, 2015.

[2] Rada Mihalcea and Dan I. Moldovan. An automatic method for generating sense tagged corpora. *In Proceedings of the American Association for Artificial Intelligence(AAAI-1999)*, pp. 461–466, 1999.

[3] 村田真樹ら. SENSEVSAL2J 辞語タスクでの 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.

[4] 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.

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

[6] Eric Sven Ristad. Maximum Entropy Modeling Toolkit, Release 1.6 beta. http://www.mnemonic .com/software/memt, 1998.

[7] Nello Cristianini and John Shawe-Taylor. *An Introduction to Support Vector Machines and Other Kernel-based Learning Methods*. Cambridge University Press, 2000.

[8] Taku Kudoh. TinySVM: Support Vector Machines. http://cl.aist-nara.ac.jp/ taku-ku// software/TinySVM/ index.html, 2000.

[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.