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 勇馬 [*1]　村田 真樹 [*2]　馬 青 [*3]
[*1] 鳥取大学 工学部 知能情報工学科
[*2] 鳥取大学大学院 工学研究科 情報エレクトロニクス専攻
[*3] 龍谷大学 理工学部 数理情報学科
[*1,*2]{s122033,murata}@ike.tottori-u.ac.jp
[*3] qma@math.ryukoku.ac.jp

1　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，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．

2　本研究の手法

2.1　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]

2.2　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 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 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。

[*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．

表 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

表 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。

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 1 から表 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 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.

3 実験

3.1 実験方法

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 2 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.

機械学習は最大エントロピー法を使用する．また，表 5 に実験に使用した素性 (解析に用いる情報) を示す．表 5 は文献 [4] 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 5 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.

3.2 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。

表 5: 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

「情報」の場合，語義 1 を「知らせ」，語義 2 を「知識」とした．

「他」「意味」については多義性解消の実験は行っていないが，単語の選定は行った．

「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．

「他」「意味」の単語についての多義性解消の実験は今後の課題である．

3.3 実験結果

「子供」「情報」という多義語で実験を行った．表 6 から表 7 に SemEval2 の「子供」「情報」についての事例数を示す．「子供」「情報」には，辞典に基づく語義が 2 つある．

表 6: 「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 7: 「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 8 から表 9 に「子供」「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 8: 増えた学習データ (「子供」)

	増えた学習データ
語義 1	997
語義 2	783
総数	1780

表 10: 利用する学習データとその正解率

手法	正解率 (子供)	正解率 (情報)	正解率 (2 単語全て)
ベースライン	0.36(18/50)	0.84(42/50)	0.60(60/100)
SemEval2 の学習データのみを利用する手法	0.54 (27/50)	0.86 (43/50)	0.73(70/100)
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	0.66 (33/50)	0.82 (41/50)	0.74(74/100)
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	0.70 (35/50)	0.80 (40/50)	0.75(75/100)

表 9: 増えた学習データ (「情報」)

	増えた学習データ
語義 1	157
語義 2	477
総数	634

3 種類の学習データを利用する手法とベースラインを用いる手法の多義性解消の結果を表 10 に示す．なお，表中のベースラインとは，学習データ中で最も頻度が高いものを常に出力するものとする．

3.4 考察

2 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．この場合の正解率は 0.75 であり，この場合でもある程度は解けることがわかった．

単語ごとの正解率について考察した．「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．

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．素性に基づく分析を行った結果，以下のことがわかった．「子供」は「小学校」「学校」「生徒」「教諭」といった単語が文中に出てきた場合，語義 1(「児童」) の意味で使われることが多かった．また，「父親」「父」「妻」といった単語が文中に出てきた場合，語義 2(「息子」) の意味で使われることが多かった．

4 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，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，2 単語すべての正解率では，「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．この場合の正解率は 0.75 であり、この場合でもある程度は解けることがわかった．「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，追加する前より性能が向上した．「情報」については，追加した後のほうが少し性能が下がった．今回の実験では使用する単語が少なく，まだはっきりしたことはわからない．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] 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.