平成27年度茨城大学工学部情報工学科
卒業研究論文

# 分散表現を用いた教師あり機械学習による語義曖昧性解消

平成28年2月9日提出

茨城大学 工学部情報工学科
10T4065Y 山木 翔馬
指導教員：新納 浩幸 教授

平成 27 年度茨城大学工学部情報工学科卒業研究

# 分散表現を用いた教師あり機械学習による語義曖昧性解消

氏名：10T4065Y 山木 翔馬
指導教員：新納 浩幸 教授

## 論文要旨

本論文では教師あり機械学習による語義曖昧性解消 (Word Sense Disambiguation, WSD) に分散表現を用いる手法する．

単語の意味をベクトル表現する場合，従来は Bag of Words(BoW) を用いて，高次元スパースなベクトルとして表現してきた．近年，深層学習の手法を利用して，単語の意味を低次元の密なベクトルで表現した分散表現が注目されている．分散表現により単語の意味をベクトル表現した場合，単語間の距離が BoW を用いるよりも，より正確に用いられるようになる．そのため，様々な自然言語処理のタスクに分散表現が利用され，有効な結果を残している．

WSD のタスクに関しては，単語の分散表現がその単語の語義の分散表現の和，つまり Bag of Senses になっていることに注目した研究がいくつかあるが，それらはどれも教師なし機械学習の枠組みであり，教師あり機械学習による WSD に分散表現を利用した研究は我々の知る限り Sugawara　のものだけである．Sugawara の手法は素性として周辺単語の他にその分散表現を加えた単純なものである．周辺単語のみを用いたモデルよりも正解率は有意に高かったが，以下の2 つの問題があると考えられる．

(1) 文脈上の単語の位置が規定される
(2) 自立語以外の単語も考慮している

本論文では上記 (1) と (2) を改善した分散表現の新しい利用法を提案する．具体的には訓練データとして N 個の用例があった場合，各用例との類似度を測り，その類似度を並べた N 次元のベクトルを基本の素性ベクトルに結合させ，それを新たな素性ベクトルとして学習と識別に利用するというものである．各用例との類似度を測る部分に分散表現を用いている．これは上記の (1) の問題を回避している．また類似度を測る際に自立語のみを用いることで上記の (2) の問題も回避できる．実験では SemEval-2 の日本語辞書タスクのデータを用いて，提案手法が Sugawara の手法よりも高い正解率を出すことを確認した．

また，基本とする素性ベクトルに用例間の類似度を並べたベクトルを結合させたものを新たな素性ベクトルとして学習と識別に利用する手法についての実験も行った．この実験でも同様に SemEval-2 の日本語辞書タスクのデータを用い，基本素性として SemEval-2 の baseline とされたシステムで利用された素性を用いた．実験の結果，用例間の類似度を用いた手法のほうが高い正解率となった．

# 目 次

**第 1 章　序論** **2**
1.1　概要 . . . . . 2
1.2　構成 . . . . . 2

**第 2 章　WSD における分散表現** **4**
2.1　語義曖昧性解消 . . . . . 4
2.2　離散的な表現を用いた従来の WSD . . . . . 4
2.3　WSD におけるシソーラスの利用 . . . . . 5
2.4　WSD における分散表現の利用 . . . . . 6

**第 3 章　word2vec** **7**

**第 4 章　用例間の類似度** **8**
4.1　Sugawara の手法 . . . . . 8
4.2　用例間の類似度を用いた提案手法 . . . . . 9

**第 5 章　実験** **11**
5.1　基本設定 . . . . . 11
5.2　実験結果 . . . . . 11

**第 6 章　考察** **15**

**第 7 章　おわりに** **20**

# 第1章　序論

## 1.1　概要

本論文では教師あり機械学習による語義曖昧性解消 (Word Sense Disambiguation, WSD) に分散表現を用いる手法する．単語の意味をベクトル表現する場合，従来は Bag of Words(BoW) を用いて，高次元スパースなベクトルとして表現してきた．近年，深層学習の手法を利用して，単語の意味を低次元の密なベクトルで表現した分散表現が注目されている．分散表現により単語の意味をベクトル表現した場合，単語間の距離が BoW を用いるよりも，より正確に用いられるようになる．そのため，様々な自然言語処理のタスクに分散表現が利用され，有効な結果を残している．WSD のタスクに関しては，単語の分散表現がその単語の語義の分散表現の和，つまり Bag of Senses になっていることに注目した研究がいくつかあるが，それらはどれも教師なし機械学習の枠組みであり，教師あり機械学習による WSD に分散表現を利用した研究は我々の知る限り Sugawara[1] のものだけである．Sugawara の手法は素性として周辺単語の他にその分散表現を加えた単純なものである．周辺単語のみを用いたモデルよりも正解率は有意に高かったが，以下の2 つの問題があると考えられる．(1) 文脈上の単語の位置が規定される (2) 自立語以外の単語も考慮している本論文では上記 (1) と (2) を改善した分散表現の新しい利用法を提案する．具体的には訓練データとして N 個の用例があった場合，各用例との類似度を測り，その類似度を並べた N 次元のベクトルを基本の素性ベクトルに結合させ，それを新たな素性ベクトルとして学習と識別に利用するというものである．各用例との類似度を測る部分に分散表現を用いている．これは上記の (1) の問題を回避している．また類似度を測る際に自立語のみを用いることで上記の (2) の問題も回避できる．実験では SemEval-2 の日本語辞書タスクのデータを用いて，提案手法が Sugawara の手法よりも高い正解率を出すことを確認した．

また，基本とする素性ベクトルに用例間の類似度を並べたベクトルを結合させたものを新たな素性ベクトルとして学習と識別に利用する手法についての実験も行った．この実験でも同様に SemEval-2 の日本語辞書タスクのデータを用い，基本素性として SemEval-2 の baseline とされたシステムで利用された素性を用いた．実験の結果，用例間の類似度を用いた手法のほうが高い正解率となった．

## 1.2 構成

本論文では，WSD における分散表現の利用について示し，分散表現から得た用例間の類似度ベクトルを用いる手法の有用性を調べる．2 章で WSD における分散表現の利用について説明をする．3 章では，分散表現を求める際に用いた Word2Vec の原理について述べる．4 章では分散表現から得た用例間の類似度を用いた提案手法について説明する．

# 第2章　WSDにおける分散表現

## 2.1　語義曖昧性解消

語義曖昧性解消 (Word Sense Disambiguation; WSD) は，複数の語義を持つ多義語について，文中にある多義語がどの語義をあらわしているのかを判断するタスクのことである．例えば「やる」という単語は以下のような複数の語義を持っている．

- 彼はその仕事をやった．(ある動作をする)
- その日はジャズをやった．(演奏/上演する)
- プレゼントとして時計をやった．(譲渡する)
- 机の上の本を向こうへやった．(どかす)
- 心配なので人をやった．(遣いを出す)
- 目を向こうへやった．(視線を投げる)

WSDのタスクは自然言語処理の様々なタスクにおいても重要な問題である．例えば機械翻訳のタスクにおいて，上の例文を英語に翻訳しようとした場合,「やる」という単語の英語訳はその語義によって「do, play, give」など異なってくる．そのため機械翻訳では単語の語義を正確に識別することが求められる．

## 2.2　離散的な表現を用いた従来のWSD

WSDのタスクへのアプローチとして，対象単語の周辺に出現した単語を素性とする手法がある．例えば以下の3つの文

(1) 彼は本を出す予定です.

(2) 彼は小説を出す予定です.

(3) 彼は車を出す予定です.

の「出す」という単語についてWSDを行うことを考える．上の例では(1)と(2)の「出す」は出版するという意味であり，(3)は用意するという意味である．この例であれば，WSDの手がかりとなるのは「本」「小説」「車」という3つの単語であることがわかり，また「本」と「小説」は似ている単語で、「車」だけは似ていない単語であるということがわかれば「出す」の語義を識別することができそうである．

ここで，この3つの文の対象単語の周辺文脈は以下のようにあらわすことができる．このように文中に出現する単語のリストを作り，各文の素性となる単語（(1)の文では

| | 彼 | は | 本 | 小説 | 車 | を | 予定 | です |
|---|---|---|---|---|---|---|---|---|
| (1) 彼は本を出す予定です | 1 | 1 | 1 | 0 | 0 | 1 | 1 | 1 |
| (2) 彼は小説を出す予定です | 1 | 1 | 0 | 1 | 0 | 1 | 1 | 1 |
| (3) 彼は車を出す予定です | 1 | 1 | 0 | 0 | 1 | 1 | 1 | 1 |

[彼，は，本，を，予定，です]）に対応するインデックスが1となったベクトルで表現する方法を Bag-of-Words(BoW) モデルという．

BoWは文脈を表現する方法として最もシンプルなものであり，自然言語処理の多くのタスクで用いられているが，上の例を見ると「本」「小説」「車」の単語の類似性はまったく表現されていない．つまり訓練データとして(1)と(3)の用例が与えられたときに，(2)の用例の「出す」の語義は(1)と(3)の用例の「出す」のどちらと同じか識別するようなWSDを考えた場合，上のようなBoWによる離散的な表現だけではうまく識別できない．

このように離散的な表現では，訓練データに出現しない単語に対応できないという欠点がある．

## 2.3 WSDにおけるシソーラスの利用

離散的な表現の問題点の解決策として，シソーラスを利用し単語の上位概念を素性として用いる手法が一般的に行われている．シソーラスとは単語の上位（下位）概念や同義関係，類似関係などによって単語を分類し，体系づけた辞書のことである．このシソーラスから得た情報を先ほどの離散的な素性ベクトルに加え，以下のような新しい素性を考える．

このとき「本」の上位概念と「小説」の上位概念が同じであれば，(1)と(2)の文脈が類似していることが表現されることになり，それを手がかりとして「出す」の語義を識別できる可能性がある．

このようにWSDにおいてシソーラスの利用は有効なアプローチであることが知られている．

| | 彼 | は | 本 | 小説 | 車 | を | 予定 | です | |
|---|---|---|---|---|---|---|---|---|---|
| (1) | 1 | 1 | 1 | 0 | 0 | 1 | 1 | 1 | 本の上位概念 |
| (2) | 1 | 1 | 0 | 1 | 0 | 1 | 1 | 1 | 小説の上位概念 |
| (3) | 1 | 1 | 0 | 0 | 1 | 1 | 1 | 1 | 車の上位概念 |

## 2.4 WSDにおける分散表現の利用

離散的な表現の問題を解決する方法として，従来単語を複数の連続値で表現する研究が行われてきたが，近年分散表現と言われる，単語間の特徴を密な実数値のベクトルで表現する手法が提案されている．分散表現により単語の意味をベクトル表現した場合，単語間の距離がBoWを用いるよりもより正確に求められるようになる．そのため，様々な自然言語処理のタスクに分散表現が利用され，有効な結果を残している．

本研究ではシソーラスの代わりに分散表現を利用することを目的としている．

# 第3章　word2vec

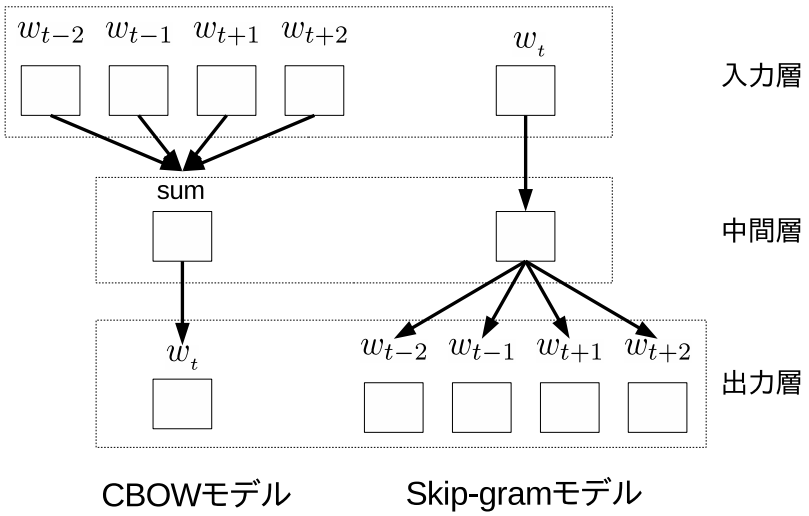


図 3.1: CBOW モデルと skip-gram モデルの仕組み

word2vec[1] は Tomas Mikolov らによって提案されたニューラルネットワークを用いた，単語に対する概念を低次元の密なベクトルとして獲得する手法である．word2vec は同じ文脈中に現れる単語同士は意味が似ている，という仮定に基づいており，テキスト中の各単語を周辺の単語から予測する擬似的な単語予測のタスクを設定し，それを大量のテキストからニューラルネットワークで学習することで，単語に対する分布表現を獲得する．

Mikolov らは word2vec を実現するネットワーク構造として，CBOW モデルと skip-gram モデルの 2 つのモデルを提案している (図 3.1)．CBOW モデルは単語周辺の文脈から単語を推定するモデルであり，skip-gram モデルは CBOW モデルの逆で，単語から文脈中の一単語を推定するモデルである．

これらのモデルをニューラルネットワークで学習させた際に得られる中間ノードの重みベクトルが，word2vec が最終的に生成する単語の分散表現となる．

[1] https://code.google.com/p/word2vec/

# 第4章　用例間の類似度

ここでは教師あり機械学習による語義曖昧性解消に分散表現を用いた先行研究として Sugawara の手法を説明し，その問題点，またその問題を回避した提案手法について述べる．

## 4.1　Sugawara の手法

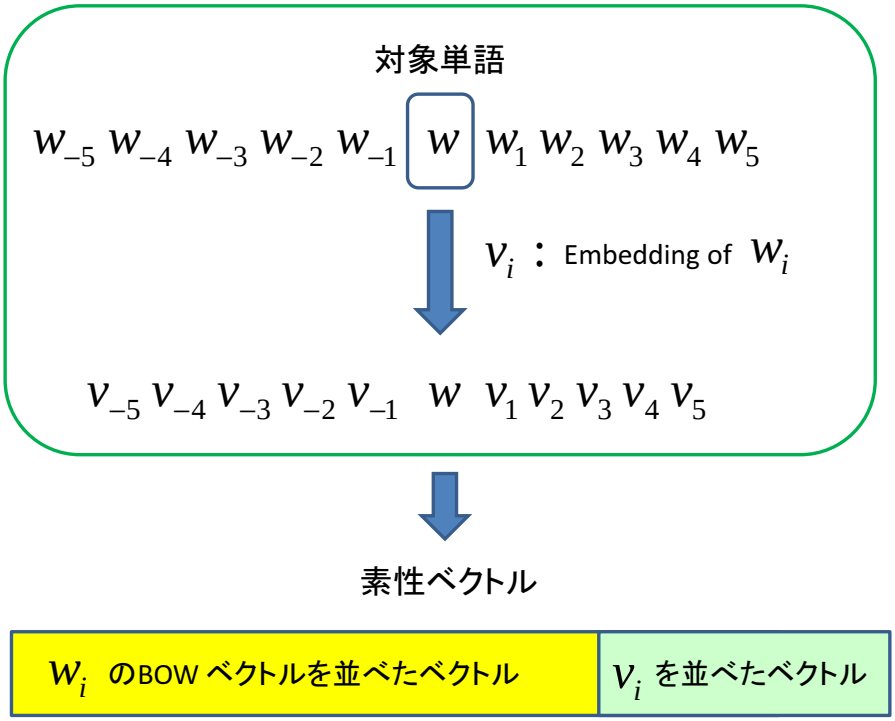


図 4.1: BoW+CWE の素性ベクトル

Sugawara の手法では対象単語の前後5単語を素性の単語として用い，BoW と分散表現 (Context-Word-Embeddings, CWE) によって得られたベクトルを組み合わせて素性として表現している．例えば，対象単語の前5単語が $w_{-1}, w_{-2}, w_{-3}, w_{-4}, w_{-5}$，後ろ5単語を

$w_1, w_2, w_3, w_4, w_5$ であった場合，BoWによって得られた2値ベクトル $(1, 0, 0, 1, 0, \cdots, 1)$ と CWEによって得られた embeddingを並べたベクトル $(v_{w-1}, v_{w-2}, \cdots, v_{w4}, v_{w5})$ を合わせたものが素性を表現するベクトルとなる (図 4.1)．Sugawara の実験ではこのBoW+CWEモデルがBoWモデルやCWEモデルよりも高い正解率となることが確認されている．

しかし BoW+CWEモデルには前述した以下の2つの問題点があると思われる．

(1) 文脈上の単語の位置が規定される

(2) 自立語以外の単語も考慮している

(1) 文脈上の単語の位置が規定されるという問題点については，例えば用例1の素性の $i$ 番目の単語 $w_{1i}$ と，用例2の素性の $j$ 番目の単語 $w_{2j}$ の embedding $v_{w1i}, v_{w2j}$ が類似していた場合であっても，$i \neq j$ であればその類似性が反映されない．

また (2) 自立語以外の単語も考慮しているという問題について，前述のように分散表現は単語の概念を表すシソーラスとして利用するため，自立語以外の単語は考慮に入れる必要はないと考えられる．

## 4.2 用例間の類似度を用いた提案手法

文脈上の単語の位置を規定しない素性として，用例間の類似度を用いる．まず訓練データの用例 $i$ と用例 $j$ について，Sugawara 手法での CWEと同様に各用例の素性となる単語の embeddingを求める．各用例の embeddingを並べたベクトルを

$$V_i = (v_{wi-1}, v_{wi_2}, \cdots, v_{wi4}, v_{wi5})$$

$$V_j = (v_{wj-1}, v_{wj-2}, \cdots, v_{wj4}, v_{wj5})$$

としたとき，用例間の類似度 $sim(i, j)$ は各用例の embeddingの cos類似度の平均とする．

$$sim(i, j) = \frac{\sum_{v_{iw}}^{V_i} \sum_{v_{jw}}^{V_j} sim(v_{iw}, v_{jw})}{|V_i| \cdot |V_j|}$$

自立語のみを素性として利用する場合は，自立語以外の単語の embeddingを $V_i, V_j$ から除外する．

ここで提案する手法では，用例 $i$ の素性を訓練データの用例 $j(1 \leq j \leq N;$ N は訓練データの用例数) との類似度 $sim(i, j)$ の値を並べたベクトル

$$sim(i, 1), sim(i, 2), \cdots, sim(i, i), \cdots, sim(i, N)$$

と BoWに基づく 2値ベクトルを合わせたベクトルで表現する (図 4.2)．

ここでは自立語以外の単語も素性として用いる手法を提案手法 (1)，自立語のみを素性として用いる手法を提案手法 (2) とする．

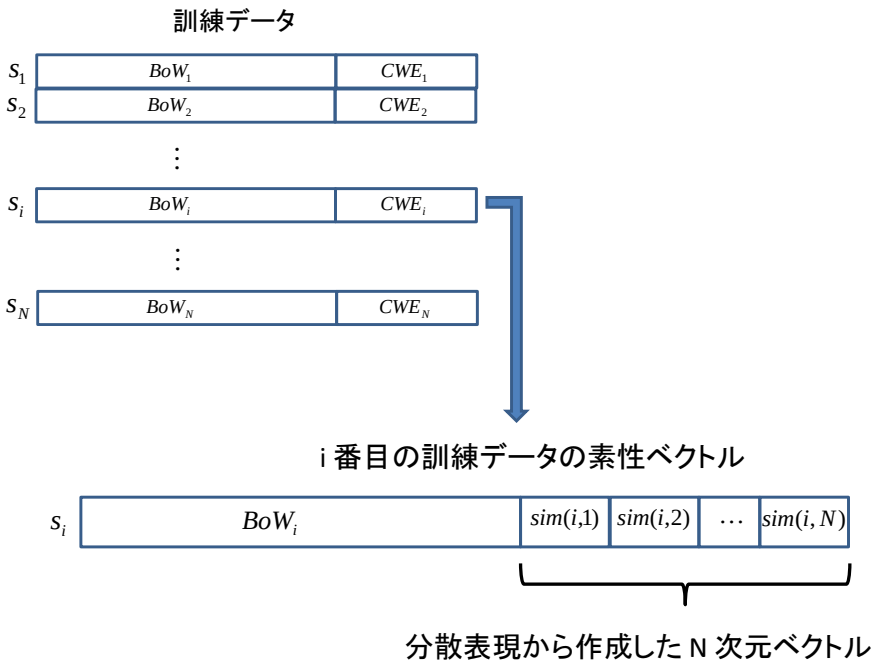


図 4.2: 提案手法による訓練データの素性ベクト ル

# 第5章　実験

## 5.1　基本設定

実験にはSemEval-2の日本語辞書タスクのデータを用いる．このデータは50個の異なる多義語で構成されており，各単語ごとに訓練データ50個，テストデータ50個が用意されている．訓練データ，テストデータは形態素解析されたXML形式となっている．

前述のCWEモデルで用いる単語の分散表現には，wikipediaの日本語記事（約5Gバイトのコーパス）をword2vecで学習した200次元のベクトルを使用した．

分類器の作成にはscikit-learn[1]のlinearSVCを使用し，正規化パラメータCは1.0に設定した．

また提案手法において自立語は，単語の品詞（第一分類）が名詞，動詞，形容詞，形状詞，副詞であるものとした．

## 5.2　実験結果

まずはじめに，Sugawaraの手法(BoW+CWE)が日本語タスクにおいても有効であることを確認する実験を行った．表5.2にBoW素性とBoW+CWE素性の正解率を示す．このことから，Sugawaraの手法が日本語タスクにおいても有効であることがわかった．

| 素性集合 | 正解率 |
|---|---|
| BoW | 0.716 |
| BoW + CWE | 0.745 |

表 5.1: BoWとBoW+CWEによる分類結果

次に，提案手法であるBoW+類似度ベクトルによる素性を用いた実験を行った．表5.2にBoW+CWEと提案手法の分類結果を示す．

[1]http://scikit-learn.org/stable/index.html

| 素性集合 | 正解率 |
|---|---|
| BoW + CWE | 0.745 |
| 提案手法 (1) | 0.753 |
| 提案手法 (2) | 0.754 |

表 5.2: BoW+CWE と提案手法による分類結果

実験の結果，提案手法が Sugawara 手法より高い正解率を出すことが確認できた．また，用例間の類似度を求める際に自立語である単語の分散表現のみを用いたほうが僅かながら精度が良くなっていることも分かった．

また各対象単語に対する正解率を表 5.3 にまとめた．太字のものはその対象単語に対する最高値のものである．また下線が引いてあるものは BoW+CWE と提案手法を比較して，strictly に大きい数値のものである．

表 5.3: 各対象単語に対する正解率 (1)

| 対象単語 | BoW | BoW+CWE | 提案手法 (1) | 提案手法 (2) | std-0 | std-1 |
|---|---|---|---|---|---|---|
| 相手 | **0.82** | **0.82** | **0.82** | **0.82** | 0.78 | 0.80 |
| 会う | 0.60 | 0.70 | 0.70 | 0.70 | 0.88 | **0.92** |
| 上げる | 0.36 | 0.36 | 0.44 | 0.42 | 0.44 | **0.52** |
| 与える | 0.64 | 0.64 | 0.66 | 0.68 | **0.76** | 0.70 |
| 生きる | **0.94** | **0.94** | **0.94** | **0.94** | **0.94** | **0.94** |
| 意味 | 0.38 | 0.52 | 0.64 | **0.68** | 0.48 | 0.44 |
| 入れる | 0.72 | **0.74** | **0.74** | **0.74** | **0.74** | **0.74** |
| 大きい | **0.94** | **0.94** | **0.94** | **0.94** | **0.94** | **0.94** |
| 教える | 0.22 | 0.34 | 0.38 | 0.38 | 0.36 | **0.52** |
| 可能 | 0.68 | **0.74** | 0.62 | 0.60 | 0.68 | 0.64 |
| 考える | **0.98** | **0.98** | **0.98** | **0.98** | **0.98** | **0.98** |
| 関係 | 0.82 | 0.88 | **0.96** | **0.96** | **0.96** | **0.96** |
| 技術 | 0.84 | 0.84 | **0.86** | **0.86** | 0.84 | 0.82 |
| 経済 | **0.98** | **0.98** | **0.98** | **0.98** | **0.98** | **0.98** |
| 現場 | 0.74 | 0.74 | 0.74 | 0.74 | 0.74 | **0.76** |
| 子供 | 0.60 | 0.54 | 0.44 | 0.42 | 0.60 | **0.62** |
| 時間 | 0.86 | 0.84 | **0.88** | **0.88** | 0.86 | 0.84 |
| 市場 | 0.58 | **0.64** | 0.60 | 0.60 | 0.52 | 0.56 |
| 社会 | **0.86** | **0.86** | **0.86** | **0.86** | **0.86** | **0.86** |
| 情報 | 0.70 | 0.76 | 0.82 | 0.82 | **0.86** | 0.84 |
| 進める | 0.44 | 0.58 | 0.86 | 0.86 | **0.92** | **0.92** |
| する | 0.54 | 0.66 | **0.72** | **0.72** | 0.64 | **0.72** |
| 高い | 0.86 | 0.86 | 0.86 | 0.86 | 0.86 | **0.88** |
| 出す | 0.40 | 0.46 | 0.40 | 0.40 | 0.40 | **0.50** |
| 立つ | 0.46 | 0.50 | 0.58 | **0.60** | 0.52 | 0.50 |
| 強い | **0.92** | **0.92** | **0.92** | **0.92** | **0.92** | 0.90 |
| 手 | **0.78** | **0.78** | **0.78** | **0.78** | **0.78** | **0.78** |
| 出る | 0.62 | **0.66** | 0.58 | 0.58 | 0.52 | 0.52 |
| 電話 | 0.78 | 0.78 | 0.78 | 0.78 | **0.84** | 0.78 |
| 取る | 0.24 | 0.26 | **0.32** | **0.32** | 0.26 | 0.28 |
| 乗る | 0.56 | 0.58 | 0.60 | 0.60 | **0.78** | **0.78** |

表は次ページに続く

前ページからの続き

| 対象単語 | BoW | BoW+CWE | 提案手法 (1) | 提案手法 (2) | std-0 | std-1 |
|---|---|---|---|---|---|---|
| 場合 | 0.86 | **0.88** | 0.84 | 0.84 | 0.84 | 0.84 |
| 入る | **0.66** | **0.66** | **0.66** | **0.66** | 0.54 | 0.56 |
| はじめ | 0.90 | **0.96** | **0.96** | **0.96** | 0.88 | 0.88 |
| 始める | 0.78 | 0.80 | 0.78 | 0.78 | **0.88** | 0.86 |
| 場所 | 0.94 | **0.96** | **0.96** | **0.96** | 0.90 | **0.96** |
| 早い | 0.58 | 0.66 | 0.62 | 0.62 | **0.70** | **0.70** |
| 一 | **0.92** | **0.92** | **0.92** | **0.92** | **0.92** | 0.90 |
| 開く | **0.90** | **0.90** | 0.88 | 0.88 | 0.78 | 0.84 |
| 文化 | **0.98** | **0.98** | **0.98** | **0.98** | **0.98** | **0.98** |
| 他 | **1.00** | **1.00** | **1.00** | **1.00** | **1.00** | **1.00** |
| 前 | 0.66 | 0.76 | **0.78** | **0.78** | 0.76 | 0.76 |
| 見える | 0.60 | 0.60 | 0.58 | 0.58 | 0.68 | **0.70** |
| 認める | 0.80 | 0.80 | 0.78 | 0.78 | 0.76 | **0.82** |
| 見る | **0.80** | **0.80** | **0.80** | **0.80** | 0.78 | 0.78 |
| 持つ | 0.64 | 0.74 | 0.76 | 0.76 | 0.78 | **0.80** |
| 求める | **0.76** | 0.74 | 0.74 | **0.76** | 0.64 | **0.76** |
| もの | **0.88** | **0.88** | **0.88** | **0.88** | **0.88** | **0.88** |
| やる | 0.94 | **0.96** | **0.96** | **0.96** | **0.96** | **0.96** |
| 良い | 0.36 | 0.40 | 0.38 | 0.38 | **0.56** | 0.54 |
| 平均 | 0.716 | 0.745 | 0.753 | 0.754 | 0.757 | **0.769** |

# 第6章　考察

WSD では利用するシソーラスの粒度の問題がある．[4] 一方，分散表現では単語間の距離が求まるので，シソーラスの粒度を 連続的なものとして利用できる．この点から，シソーラスの代わりに分散表現を 利用することで WSD の精度向上が期待できる．

実験では Sugawara の手法と 提案手法との比較を 行ったが，シソーラスの代わりに分散表現を 利用できるかどうかを 調べる．

標準的な手法として SemEval-2 の baseline とされたシステムを 実装した．学習アルゴリズムは線形の SVM であり，以下の 20 種類の素性を 利用した．

e1=二つ前の単語，e2=二つ前の品詞，e3=その細分類，

e4=一つ前の単語，e5=一つ前の品詞，e6=その細分類，

e7=問題の単語，e8=問題の単語の品詞，e9=その細分類，

e10=一つ後の単語，e11=一つ後の品詞，e12=その細分類，

e13=二つ後の単語，e14=二つ後の品詞，e15=その細分類，

e16=係り 受け

e17=ふたつ前の分類語彙表の値 (5 桁)

e18=ひとつ前の分類語彙表の値 (5 桁)

e19=ひとつ後の分類語彙表の値 (5 桁)

e20=ふたつ後の分類語彙表の値 (5 桁)

従来の baseline のシステムでは分類語彙表 ID の 4 桁と 5 桁を 同時に使う形になっていたが，ここでのシステムでは 5 桁のみとした．また，一般に一つの単語に対しては複数の分類語彙表 ID が存在するので，e17，e18，e19，e20 に対する素性は複数になる．

このシステムによる 正解率は表 5.3 の std-0 と std-1 である．std-1 は素性として上記の 20 種類すべての素性を 用いた結果であり，std-0 は素性としてシソーラス情報 (e17，e18，e19，e20 を 除いた上記の 16 種類の素性を 用いた結果である．

提案手法の正解率は std-0 の正解率とほとんど差がなく，std-1 よりも劣る．しかし std-1 と std-0 の差 (0.0120) はシソーラスの利用の効果であり，BoW と提案手法の差 (0.0376) は分散表現の利用の差である．差の大きさから見ると分散表現の方が改善度が大きい．つまりシソーラスの代わりに分散表現を利用して，精度を改善できる可能性はあると考えられる．

このことを確認するための実験を行った．

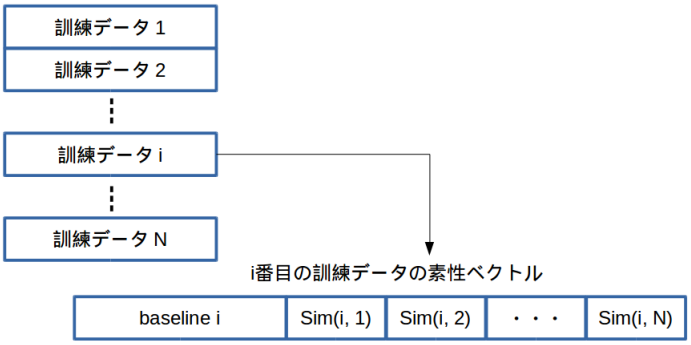


図 6.1: baseline の素性に用例間の類似度を加えた新たな素性

前述した SemEval-2 の baseline とされたシステムで用いられた素性 std-0 と std-1 を基本となる素性ベクトルとし，そこに用例間の類似度を並べたベクトルを結合させたものを新たな素性ベクトルとして利用する手法の実験を行った．

この baseline の素性を基本素性として用いた提案手法の素性を図 6.1 に示す．

実験設定はすべて前述の設定と同じである．基本素性 (std-0, std-1) と，基本素性に用例間の類似度を加えた素性 (std-0 + 用例間の類似度，std-1 + 用例間の類似度) による正解率を表 6.1 に示す．また，各対象単語に対する正解率を表 6.2 にまとめた．

| 素性集合 | 正解率 |
|---|---|
| std-0 | 0.757 |
| std-1 | 0.769 |
| std-0 + 用例間の類似度 | 0.761 |
| std-1 + 用例間の類似度 | 0.771 |

表 6.1: 基本素性と，基本素性に用例間の類似度を加えた素性による正解率

実験の結果，baseline の素性に用例間の類似度ベクトルを付け加えることで精度がよくなることが確認された．

しかし，シソーラス情報を含む素性 (std-1) とシソーラス情報を含めない素性 (std-0)+用例間の類似度による素性との正解率を見ると，std-1 での正解率 0.769 に対して，std-0+用例間の類似度での正解率は 0.761 と僅かに低い結果となっている．このことから実験では分散表現から求めた素性を用いるより，シソーラスを用いた方が改善度が高く，シソーラスの代わりに分散表現を用いることで精度が良くなるとは言えないことが分かった．しかし分散表現の利用方法は本論文で提案した用例間の類似度以外にも多数考えられること，また分散表現を求める際に用いるコーパスの量や質によって分散表現を用いた手法の精度を上げられることから，シソーラスの代わりに分散表現を用いて WSD の精度を向上させることは可能であると考えられる．

今後は分散表現のより良い利用方法を考案すること，また別のコーパスから学習した分散表現を用いることで分散表現を用いた手法の精度が改善されるかどうか調べる必要がある．

表 6.2: 各対象単語に対する正解率 (2)

| 対象単語 | std-0 | std-1 | std-0 + 類似度 | std-1 + 類似度 |
|---|---|---|---|---|
| 相手 | 0.78 | 0.80 | 0.78 | 0.80 |
| 会う | 0.88 | 0.92 | 0.90 | 0.92 |
| 上げる | 0.44 | 0.52 | 0.48 | 0.56 |
| 与える | 0.76 | 0.70 | 0.74 | 0.70 |
| 生きる | 0.94 | 0.94 | 0.94 | 0.94 |
| 意味 | 0.48 | 0.44 | 0.46 | 0.46 |
| 入れる | 0.74 | 0.74 | 0.74 | 0.74 |
| 大きい | 0.94 | 0.94 | 0.94 | 0.94 |
| 教える | 0.36 | 0.52 | 0.40 | 0.52 |
| 可能 | 0.68 | 0.64 | 0.68 | 0.64 |
| 考える | 0.98 | 0.98 | 0.98 | 0.98 |
| 関係 | 0.96 | 0.96 | 0.96 | 0.96 |
| 技術 | 0.84 | 0.82 | 0.84 | 0.82 |
| 経済 | 0.98 | 0.98 | 0.98 | 0.98 |
| 現場 | 0.74 | 0.76 | 0.74 | 0.76 |
| 子供 | 0.60 | 0.62 | 0.60 | 0.60 |
| 時間 | 0.86 | 0.84 | 0.86 | 0.86 |
| 市場 | 0.52 | 0.56 | 0.52 | 0.56 |
| 社会 | 0.86 | 0.86 | 0.86 | 0.86 |
| 情報 | 0.86 | 0.84 | 0.86 | 0.84 |
| 進める | 0.92 | 0.92 | 0.92 | 0.92 |
| する | 0.64 | 0.72 | 0.66 | 0.72 |
| 高い | 0.86 | 0.88 | 0.86 | 0.88 |
| 出す | 0.40 | 0.50 | 0.42 | 0.50 |
| 立つ | 0.52 | 0.50 | 0.52 | 0.52 |
| 強い | 0.92 | 0.90 | 0.92 | 0.90 |
| 手 | 0.78 | 0.78 | 0.78 | 0.78 |
| 出る | 0.52 | 0.52 | 0.52 | 0.52 |
| 電話 | 0.84 | 0.78 | 0.80 | 0.78 |
| 取る | 0.26 | 0.28 | 0.26 | 0.28 |
| 乗る | 0.78 | 0.78 | 0.78 | 0.78 |

表は次ページに続く

前ページからの続き

| 対象単語 | std-0 | std-1 | std-0 + 類似度 | std-1 + 類似度 |
|---|---|---|---|---|
| 場合 | 0.84 | 0.84 | 0.84 | 0.84 |
| 入る | 0.54 | 0.56 | 0.54 | 0.56 |
| はじめ | 0.88 | 0.88 | 0.88 | 0.88 |
| 始める | 0.88 | 0.86 | 0.88 | 0.86 |
| 場所 | 0.90 | 0.96 | 0.92 | 0.96 |
| 早い | 0.70 | 0.70 | 0.72 | 0.72 |
| 一 | 0.92 | 0.90 | 0.92 | 0.90 |
| 開く | 0.78 | 0.84 | 0.80 | 0.84 |
| 文化 | 0.98 | 0.98 | 0.98 | 0.98 |
| 他 | 1.00 | 1.00 | 1.00 | 1.00 |
| 前 | 0.76 | 0.76 | 0.76 | 0.76 |
| 見える | 0.68 | 0.70 | 0.68 | 0.70 |
| 認める | 0.76 | 0.82 | 0.78 | 0.82 |
| 見る | 0.78 | 0.78 | 0.78 | 0.78 |
| 持つ | 0.78 | 0.80 | 0.78 | 0.80 |
| 求める | 0.64 | 0.76 | 0.68 | 0.76 |
| もの | 0.88 | 0.88 | 0.88 | 0.88 |
| やる | 0.96 | 0.96 | 0.96 | 0.96 |
| 良い | 0.56 | 0.54 | 0.56 | 0.54 |
| 平均 | 0.757 | 0.769 | 0.761 | 0.771 |

# 第7章　おわりに

本論文では教師あり機械学習による語義曖昧性解消に分散表現を用いる手法を提案した．自然言語処理のタスクに分散表現を利用した研究は多いが，教師あり機械学習による語義曖昧性解消に分散表現を利用した研究は，我々の知る限り Sugawara のものだけである．Sugawara の手法は (1) 文脈上の単語の位置が規定される，(2) 自立語以外の語も考慮している，という 2 つの問題があると考えられる．ここではそれらの問題を回避した分散表現の利用法を提案した．実験では SemEval-2 の日本語辞書タスクを用い，Sugawara の手法よりも高い正解率を出すことができ，分散表現の利用方法としては改善できた．

また，基本となる素性ベクトルに用例間の類似度を並べたベクトルを結合させたものを新たな素性ベクトルとして学習と識別に利用する手法についての実験では，用例間の類似度を用いた手法の方が高い正解率となった．しかし，シソーラスと用例間の類似度との改善度を比較するとシソーラスを用いた手法の方が高い改善度を示し，シソーラスの代わりに分散表現を用いることで精度を改善できるとは言えない結果となった．

今後は分散表現のより良い利用方法を考察すること，また別のコーパスから学習した分散表現を用いることで分散表現を用いた手法の精度が改善されるかどうか調べる必要がある．

# 謝辞

卒業研究にあたり，熱心にご指導いただいた情報工学科の新納教授に深い感謝の意を表します．また，多くのご意見ご指摘を頂きました自然言語処理研究室の皆様に感謝します．

# 関連図書


[1] Hiromu Sugawara, Hiroya Takamura, Ryohei Sasano, and Manabu Okumura. Context representation with word embeddings for wsd. In *PACLING-2015*, pages 149–155, 2015.

[2] 山木翔馬，新納浩幸，古宮嘉那子，佐々木稔. 分散表現を用いた教師あり機械学習による語義曖昧性解消. In **情報処理学会自然言語処理研究会**, pages NL–224–17, 2015.

[3] 山木翔馬，新納浩幸，古宮嘉那子，佐々木稔. 分散表現から得た用例間類似度を素性に加えた語義曖昧性解消. In **言語処理学会第** *22* **回年次大会**, 2016.

[4] 新納浩幸，佐々木稔，古宮嘉那子. 語義曖昧性解消におけるシソーラス利用の問題分析. In **言語処理学会第** *21* **回年次大会**, pages P1–15, 2015.