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

山木 翔馬　新納 浩幸　古宮 嘉那子　佐々木 稔
茨城大学 工学部 情報工学科
10t4065y@hcs.ibaraki.ac.jp, hiroyuki.shinnou.0828@vc.ibaraki.ac.jp,
kanako.komiya.nlp@vc.ibaraki.ac.jp, minoru.sasaki.01@vc.ibaraki.ac.jp

1　はじめに

本論文では教師あり機械学習手法による語義曖昧性解消 (Word Sense Disambiguation; 以下 WSD と略す) に，分散表現から得た用例間の類似度を用いる手法を提案する．

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おり，様々な自然言語処理のタスクに分散表現が利用されている．

教師あり機械学習による WSD に分散表現を用いる研究として Sugawara の研究 [2] と，それを改善した我々の研究 [3] がある．我々が提案した手法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となる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．実験の結果，分散表現の利用法として Sugawara の手法よりも有効であることが分かった．

本論文では基本とする素性ベクトルに分散表現から求めた用例間の類似度ベクトルを結合したベクトル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を提案する．

実験では基本素性として SemEval-2 の baseline とされたシステム [1] で利用された素性を用い，基本素性と提案手法による素性の正解率の比較を行った．実験の結果，用例間の類似度を用いた提案手法の方が高い正解率となった．

2　WSD における分散表現の利用

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．この手法により対象単語の周辺の文脈情報をベクトルで表現することができるが，これらは 0 または 1 の 2 値で表される離散的な表現になるため，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．

この問題の解決策として，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．たとえば「本を出す」という文の「出す」という多義語について WSD を行う場合，離散的な表現を用いる手法であると「雑誌を出す」「小説を出す」といった訓練データは分類の手掛かりにならないが，シソーラスを利用すれば「本」「雑誌」「小説」が同じ上位概念を持つことから, それを手掛かりとして「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が，ここでは単語の分散表現をシソーラスとして用いることによって WSD の精度を高めることを目的としている．

3　用例間の類似度の利用

分散表現を用いた教師あり機械学習による WSD の先行研究として Sugawara の研究 [2] がある．Sugawara の手法では対象単語の前後 5 単語を用い，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を表現している．たとえば，対象単語の前 5 単語が $w_{-1}, w_{-2}, w_{-3}, w_{-4}, w_{-5}$，後ろ 5 単語が 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 $(\boldsymbol{v}_{w-1}, \boldsymbol{v}_{w-2}, \cdots, \boldsymbol{v}_{w6}, \boldsymbol{v}_{w5})$ を合わせたものが素性を表すベクトルとなる．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を出すことが確認されており，WSD のタスクに分散表現を用いることの有用性が示されている．

しかし Sugawara の手法には

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

という 2 つの問題点があると考えられ，我々はこれらの点を改善した手法として用例間の類似度を用いた素性を提案し，Sugawara の手法よりも精度が良くなることを確認した [3].

用例間の類似度とは，まず N 個の用例からなる訓練データの用例 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

$$\boldsymbol{V}_i = (\boldsymbol{v}_{wi-1}, \boldsymbol{v}_{wi-2}, ..., \boldsymbol{v}_{wi4}, \boldsymbol{v}_{wi5})$$

$$\boldsymbol{V}_j = (\boldsymbol{v}_{wj-1}, \boldsymbol{v}_{wj-2}, ..., \boldsymbol{v}_{wj4}, \boldsymbol{v}_{wj5})$$

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．

$$sim(i,j) = \frac{\sum_{\boldsymbol{v}_{iw}}^{\boldsymbol{V}_i} \sum_{\boldsymbol{v}_{jw}}^{\boldsymbol{V}_j} cos(\boldsymbol{v}_{iw}, \boldsymbol{v}_{jw})}{|\boldsymbol{V}_i| \cdot |\boldsymbol{V}_j|}$$

4 提案手法

本論文で提案する手法は，基本となる素性ベクトルに分散表現から得られた用例間の類似度ベクトルを加えたものを新たな素性ベクトルとして利用するという手法である．

実験で用いる基本素性は SemEval-2 の baseline とされたシステムの素性を用いる．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）

本来の 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 baseline の素性を基本素性として用いた提案手法の素性を図 1 に示す．

また実験では，シソーラス情報の有無による正解率の違いを確認するために，e1 から e16 までのシソーラス情報を含まない素性を std-0， e1 から e20 までのシソーラス情報を含めた素性を std-1 とし，それぞれを基本素性として比較を行った．

5 実験

5.1 実験設定

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結果の 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[1] 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[2] 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，baseline の素性において e1 から e16 までのシソーラス情報を含まない素性 (std-0) と，e1 から e20 までのシソーラス情報を含めた素性 (std-1) での実験を行った．結果を表 1 に示す．

表 1: baseline におけるシソーラスありとシソーラスなしの素性による分類結果

素性集合	正解率
std-0	0.757
std-1	0.769

次に std-0 と std-1 のそれぞれに用例間の類似度ベクトルを加えた素性ベクトルでの実験を行った．結果を表 2 に示す．

なお表 1 と表 2 の「正解率」は 50 個の各対象単語に対する正解率の平均，つまりマクロ平均による正解率である．

また，各対象単語に対する正解率を表 3 にまとめた．

[1] https://code.google.com/p/word2vec/
[2] http://scikit-learn.org/stable/index.html

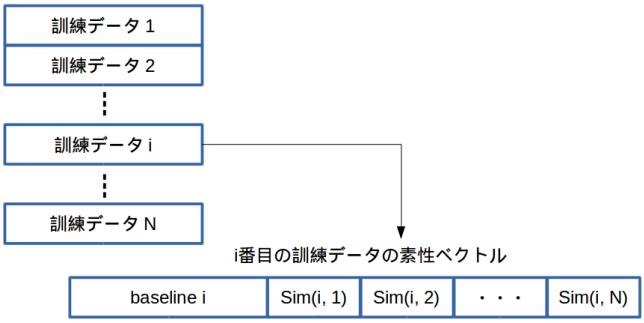

図 1: baseline の素性に用例間の類似度を加えた新たな素性

表 2: 用例間の類似度を付け加えた素性による分類結果

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

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．

6 考察

シソーラスなしの基本素性 (std-0) と std-0 + 類似度の正解率を対象単語別で比較すると，std-0 + 類似度の方が std-0 よりも高い正解率となっている単語が 10 個，低い正解率となっている単語が 3 個となっている．同様にシソーラスありの基本素性 (std-1) と std-1 + 類似度の正解率を比較すると，std-1 + 類似度の方が std-1 よりも高い正解率となっている単語が 5 個，低い正解率となっている単語が 1 個である．このことから提案手法が WSD の精度向上に有効であると考えられる．

また，WSD に分散表現を利用する目的の一つに，シソーラスの代わりに分散表現を利用することで WSD の精度向上を図るというものがある．実験結果のうち，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)+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+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．このことから本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手法が有効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手法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提案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．

7 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単語の分散表現から求めた用例間の類似度を用いる手法を提案した．具体的には基本となる素性ベクトルに分散表現から求めた用例間の類似度を付け加えたものを新たな素性として学習，識別に用いるというものである．実験では基本素性として SemEval-2 の baseline とされたシステムで利用された素性を用い，基本素性と提案手法による素性の正解率の比較を行った．実験の結果，用例間の類似度を用いた提案手法の方が高い正解率となり，提案手法が WSD の精度向上に有効であることを確認した．今後は別のコーパスから学習した分散表現を用いることで提案手法の正解率が改善されるか調べる必要がある．

参考文献

[1] Manabu Okumura, Kiyoaki Shirai, Kanako Komiya, and Hikaru Yokono. On SemEval-2010 Japanese WSD Task. 自然言語処理, Vol. 18, No. 3, pp. 293–307, 2011.

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

[3] 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. 情報処理学会自然言語処理研究会, pp. NL–224–17, 2015.

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
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

表 3: 各対象単語に対する正解率