Vol.2015-NL-224 No.17
2015/12/4

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

山木 翔馬[1,a)]　新納 浩幸[1,b)]　古宮 嘉那子[1,c)]　佐々木 稔[1,d)]

概要：近年，自然言語処理の多くのタスクに単語の分散表現が利用されている．教師あり機械学習による語義曖昧性解消に対しては Sugawara の研究が存在する．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，(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 つの問題があると思われる．ここでは上記 (1) と (2) を改善した分散表現の新しい利用法を提案する．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．

Yamaki Shoma[1,a)]　Shinnou Hiroyuki[1,b)]　Komiya Kanako[1,c)]　Sasaki Minoru[1,d)]

1. はじめに

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, 以下 WSD と略す）に分散表現を用いる手法を提案する．

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，単語の間の距離が 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．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着目した研究がいくつかあるが [2][3][1]，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 [4] 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

本論文では上記問題を回避した教師あり機械学習による WSD における分散表現の利用方法を提案する．

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) の問題を回避できている．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を用いた．Sugawara の手法が平均正解率 0.745 であったのに対して，提案手法は 0.754 となり，わずかではあるが改善された．

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

[1] 茨城大学工学部情報工学科
Ibaraki University, Nakanarusawa 4-12-1, Hiachi, Ibaraki 316–8511, Japan
[a)] 10t4065y@hcs.ibaraki.ac.jp
[b)] hiroyuki.shinnou.0828@vc.ibaraki.ac.jp
[c)] kanako.komiya.nlp@vc.ibaraki.ac.jp
[d)] minoru.sasaki.01@vc.ibaraki.ac.jp

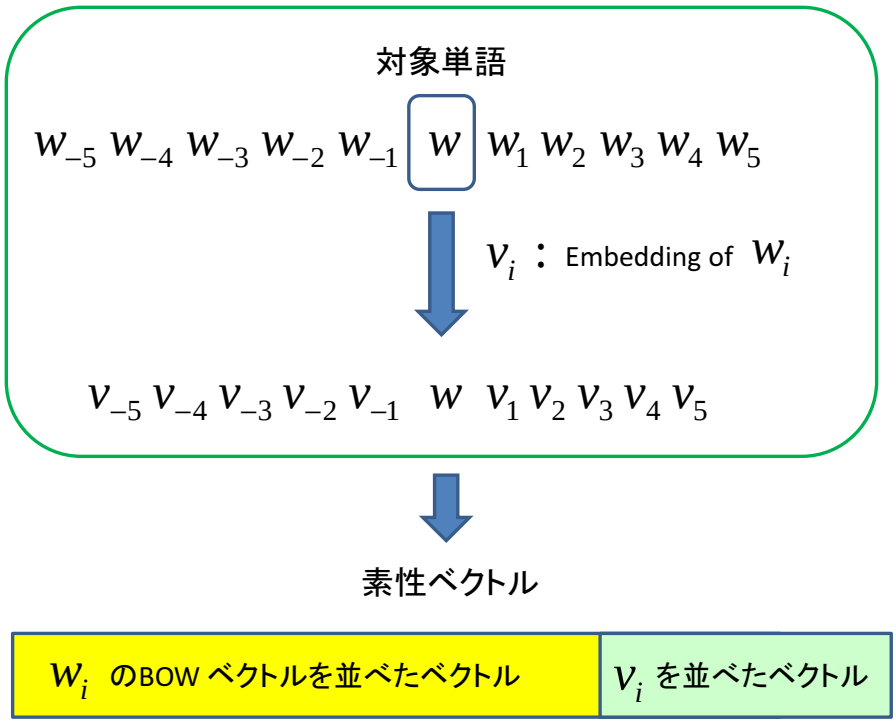

図 1 BoW+CWE の素性ベクトル

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. Sugawara の手法とその問題点

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 (図 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を出す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，たとえば用例 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 $\boldsymbol{v}_{w1i}, \boldsymbol{v}_{w2j}$ 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点について，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いれる必要はないと考えられる．

4. 提案手法

文脈中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

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

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

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．

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

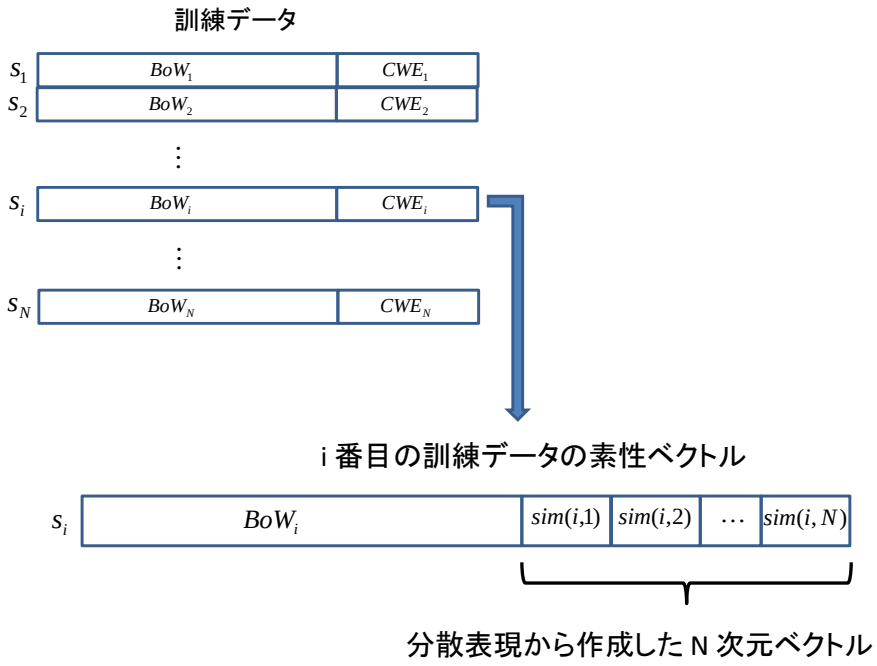

図 2 提案手法による訓練データの素性ベクトル

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 $\boldsymbol{V}_i, \boldsymbol{V}_j$ 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

$$sim(i,1), sim(i,2), ..., sim(i,i), ..., sim(i,N)$$

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 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．

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 (2) における自立語は，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．

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

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 1 に BoW 素性と BoW+CWE 素性の正解率を示す．

素性集合	正解率
BoW	0.716
BoW + CWE	0.745

表 1 BoW と BoW+CWE による分類結果

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分かった．

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 2 に Sugawara 手法と提案手法の分類結果を示す．

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 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方が僅かながら精度がよ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 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．

6. 考察

WSD では利用するシソーラスの粒度の問題がある [5]．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，ここではシソーラスを用いた標準的な手法と提案手法とを比較することで，シソーラスの代わりに分散表現が利用できるかどうかを調べる．

標準的な手法として SemEval-2 のコンペの際に 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）

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．

正解率を表 3 に示す．Std-1 は素性として上記の 20 種類全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の情報（`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効果である．差の大きさから見ると分散表現の方が改善の度合いが大きい．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．

本論文では Sugawara の手法との差を調べるために，BoW をベースとした素性を利用したが，ベースとなる素性は Std-0 のものであっても問題ない．Std-0 の素性に，ここで提案した分散表現による素性を加えた実験を行い，シソーラスの代わりに分散表現が利用して精度向上するかどうかを確認することが今後の課題である．

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，の 2 つの問題があると考えられる．ここではそれら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．今後は分散表現をシソーラスの代わりに利用して，精度向上が可能かどうかを調べる必要がある．

参考文献

[1] Bhingardive, S., Singh, D., Murthy, V. R., Redkar, H. and Bhattacharyya, P.: Unsupervised Most Frequent Sense Detection using Word Embeddings, *HLT-NAACL-2015*, pp. 1238–1243 (2015).
[2] Chen, X., Liu, Z. and Sun, M.: A Unified Model for Word Sense Representation and Disambiguation, *EMNLP-2014*, pp. 1025–1035 (2014).
[3] Neelakantan, A., Shankar, J., Passos, A. and McCallum, A.: Efficient Non-parametric Estimation of Multiple Embeddings per Word in Vector Space, *EMNLP-2014*, pp. 1059–1069 (2014).
[4] Sugawara, H., Takamura, H., Sasano, R. and Okumura, M.: Context Representation with Word Embeddings for WSD, *PACLING-2015*, pp. 149–155 (2015).
[5] 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，言語処理学会第 21 回年次大会，pp. P1–15 (2015).

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**
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**

表 3 各対象単語に対する正解率