教師データを用いた語義の分散表現の構築

山木 翔馬　　新納 浩幸　古宮 嘉那子　佐々木 稔
茨城大学大学院　　茨城大学 工学部
理工学研究科　　情報工学科

{16nm724r, hiroyuki.shinnou.0828, kanako.komiya.nlp, minoru.sasaki.01}@vc.ibaraki.ac.jp

1　はじめに

本論文では単語の分散表現と教師データを用いて語義の分散表現を構築する手法を提案する．

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 (Word Embeddings) が注目されている．さらにその手法を応用して語義ごとの分散表現 (Multi-sense Embeddings) を得る研究がされており，多くの自然言語処理タスクにおいて有効な結果を残している．

現在研究されている手法は，単語の分散表現を構築する従来の教師なしの言語モデルを拡張したものがほとんどである・そのため，コーパス中に現れる語義の出現頻度の情報を得ることができないという問題点がある．

本論文では教師データを用いて，語義の出現頻度を考慮した語義の分散表現を構築する手法を提案する．具体的には，あらかじめ得た単語の分散表現を用いて教師データ中の各語義に対する用例の文脈ベクトルを求め，文脈ベクトルと語義の出現頻度に基づいて単語の分散表現を語義の分散表現に分解する．

実験では単語の分散表現として Wikipedia から学習したものと国立国語研究室が作成した分散表現 nwjc2vec[5] を用いた．教師データとしては SemEval-2 の日本語辞書タスクのデータを用いて語義の分散表現を構築し，語義曖昧性解消 (Word Sense Disambiguation; WSD) による評価実験と，語義の分散表現の類似単語による評価を行った．実験の結果，作成した語義の分散表現を WSD に用いた実験では精度の向上は確認できなかったが，作成した語義の分散表現と類似の単語を分析したところ，語義の分散表現が正しく作られていることが分かった．

2　関連研究

深層学習の手法を利用して単語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，自然言語処理の様々な分野で有効な結果を残している．単語の分散表現の構築には Feedforward Neural Network Language Model や Recurrent Neural Network Language Model などのニューラルネットワークに基づく言語モデルを用いる方法が多く研究されているが，なかでも Mikolov らが提案した skip-gram モデルと CBoW モデルは，言語モデルを単純化することでベクトル表現の学習の高速化に成功した [3]．これらのモデルをツール化した word2vec[1] は分散表現を獲得する手段として広く使われている．

単語の分散表現の学習モデルを拡張することで語義ごとの分散表現を構築する研究も多くされている．Neelakantan らの研究では 1 つの単語に語義ごとのベクトルを与えるモデルとして Skip-gram モデルを拡張した Multi Sense Skip-gram(MSSG) モデルを提案している [4]．また各単語にいくつの語義があるかを自動で決めるノンパラメトリックな MSSG(NP-MSSG) モデルも提案している．

MSSG モデルによる語義の分散表現の構築では，語義数をどのように決定するかが重要になる．Chen らは WordNet の辞書データを用いて WordNet の語義ごとの分散表現を学習する手法を提案している [1]．Li らは MSSG モデルに中華料理店過程 (CRP) を適応させた NP-MSSG モデルを提案している [2].

また Li らの研究では語義の分散表現を自然言語処理の様々なタスクに利用するためのパイプラインアーキテクチャを提案し， 語義の分散表現が part-of-speech tagging, semantic relation identification, semantic relatedness のタスクにおいて有効であることを確認している．

しかし一般的に MSSG モデルの学習は教師無しで

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

行われているため，語義の頻度の情報が得られないという問題点がある．本論文ではこの問題を回避するために教師データを用いた語義の分散表現の構築手法を提案し，得られた分散表現の効果を分析する．

3 教師データの利用

語義の分散表現を構築する上で語義の出現頻度は有力な情報であるが，前述の MSSG モデルでは語義の出現頻度が得られないため，対象単語の文脈ベクトルとあらかじめ設定した語義の文脈ベクトルとの類似度によって語義を推定している．

教師データを用いる場合も文脈ベクトルは有効な情報である．対象単語 w_t に対する教師データの i 番目の用例

$$w_1, w_2, \cdots, w_t, \cdots, w_m$$

の文脈ベクトル $\bar{\boldsymbol{u}}_i$ は

$$\bar{\boldsymbol{u}}_i = \frac{\boldsymbol{w}_1, \boldsymbol{w}_2, \cdots, \boldsymbol{w}_m}{m}$$

で表される．また教師データ中の語義 c_i に対する用例を C_i としたとき，語義 c_i の文脈ベクトル u^i は語義 C_i に対する用例の文脈ベクトルの平均で表し

$$\boldsymbol{u_i} = \frac{1}{|C_i|} \sum_{i \in C_i} \bar{\boldsymbol{u_i}}$$

となる．この文脈ベクトル $u^{(i)}$ には語義の出現頻度の情報が含まれていない．

本論文では語義の文脈ベクトル $u^{(i)}$ と語義の出現頻度 $|C_i|$ の 2 つの情報を考慮した語義の分散表現の構築法を提案する．

4 提案手法

単語の分散表現を $\boldsymbol{w}$, 語義（ここでは 3 つの語義があるとする）の分散表現を $\boldsymbol{e}_1, \boldsymbol{e}_2, \boldsymbol{e}_3$ としたとき，

$$\boldsymbol{w} = \boldsymbol{e_1} + \boldsymbol{e_2} + \boldsymbol{e_3}$$

が成り立つとすると，これらの k 次元目の値 w_k, e_k^1, e_k^2, e_k^3 においても

$$w_k = e_k^1 + e_k^2 + e_k^3$$

が成り立つ．ここで提案する手法は単語の分散表現と語義の文脈ベクトル $\boldsymbol{u_1}, \boldsymbol{u_2}, \boldsymbol{u_3}$ の k 次元目の値に注目し，u_k^1, u_k^2, u_k^3 の値に差が小さければ

$$e_k^i = \frac{|C_i|}{|C_1| + |C_2| + |C_3|} w_k$$

と w_k を語義の出現頻度で分解し，差が大きければ

$$\begin{aligned} e_k^i &= w_k \\ e_k^j &= 0 \ \ (j \neq i) \end{aligned}$$

とする．

値の差の大小を比較するための基準値を設定する．教師データの総数を N，教師データ中の最頻出語義 (Most Frequent Sense; MFS) の数を M としたとき，基準値 tr を

$$tr = \log \frac{M + 0.01}{N + 0.01}$$

とする．tr は決定リストのデフォルト規則適用の証拠の強さを表している．

u_k^1, u_k^2, u_k^3 の差の大きさは

$$m = |\max(u_k^1 w_k, u_k^2 w_k, u_k^3 w_k)|$$

とし，次のように正規化する．

$$\begin{aligned} n &= \sum_i |u_k^i w_k| \\ r &= \frac{N}{n} \\ d &= \log \frac{mr + 0.01}{nr + 0.01} \end{aligned}$$

tr と d を比較し，$tr > d$ であれば差が小さい，$tr \leq d$ であれば差が大きいと判定する．

この操作を分散表現のすべての次元に対して行い，得られた分散表現 $\boldsymbol{e}_1, \boldsymbol{e}_2, \boldsymbol{e}_3$ を語義の分散表現とする．提案手法の概要図を図 1 に示す．

5 語義の分散表現による WSD

WSD は文中のある多義語について，その単語がどの語義を表しているのかを判断するタスクである．通常，教師あり WSD においては教師データをもとに精度の高い語義の分類機器構築することが目的となるが，語義の分散表現 $\boldsymbol{e}_1, \boldsymbol{e}_2, \cdots, \boldsymbol{e_K}$ を利用すれば，テスト用例中の対象単語周りの文脈ベクトルと類似度が最大となる語義を選ぶことで WSD を行う．つまり対象単語 w_t の文脈ベクトルを $\boldsymbol{v}$ とすると，

$$\arg\max_i \cos(\boldsymbol{e}_i, \boldsymbol{v})$$

により分類する．

本論文では提案手法によって構築した語義の分散表現を評価するために WSD の実験を行う．また提案手法との比較のために教師データから SVM による分類器を作成した．素性ベクトルは語義の文脈ベクトルである．

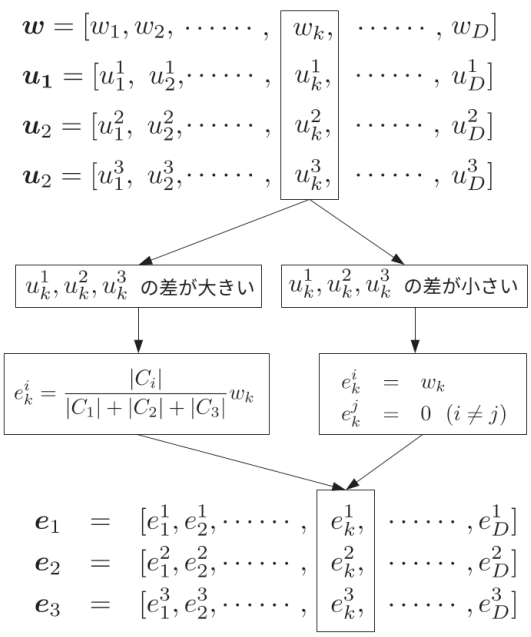

図 1: 提案手法の概要図

6 実験

6.1 実験設定

実験では教師データとして 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 用例，テストデータ 50 用例が用意されている．

単語の分散表現は分散表現の精度の違いによる比較を行うため，wikipedia の日本語記事を word2vec で学習した 200 次元ベクトルと，nwjc2vec の 200 次元ベクトルの 2 つを用いた．

また WSD による評価実験では比較のために文脈ベクトルを素性として SVM を学習させた．文脈単語は対象単語の前後 5 単語とし，SVM の学習には scikit-learn[2] の LinearSVC を用いた．

6.2 実験結果

はじめに wikipedia の分散表現を用いた提案手法による WSD の分類結果を表 1 に示す．

次に nwjc2vec を用いた提案手法と SVM による WSD の分類結果を表 2 に示す．

実験の結果，いずれの手法も nwjc2vec を用いた方が wikipedia の分散表現を用いるよりも高い正解率となった．また手法別に見ると提案手法は MFS 値よりは高い正解率を出すものの，SVM よりはわずかに低い正解率となった．

表 1: wikipedia の分散表現を用いた提案手法による WSD の分類結果

手法	平均正解率
MFS	0.688
SVM	0.712
提案手法	0.6920

表 2: nwjc2vec を用いた提案手法と SVM による WSD の分類結果

手法	平均正解率
MFS	0.688
SVM	0.757
提案手法	0.736

また，構築した語義の分散表現がどのような意味を表しているのか調べるため，単語「意味」の各語義の分散表現と最も類似度の高い分散表現を持つ単語をそれぞれ 10 単語ずつ求めた．結果を表 3 に示す．

また岩波辞書における「意味」の語義は以下の通りである．

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

表 3: 単語「意味」の各語義の類似単語

2843-0-0-1	趣旨, 形式, 主旨, 自体, 面白味, 面白み, 意図, 2843-0-0-3, 要素
2843-0-0-2	ニュアンス, 比喩, 文脈, 暗喩, 真意, 婉曲, 恣意, 語弊, 含意, 隠喩
2843-0-0-3	本質, 意識, 美徳, 含意, アイデンティティ, 内実, 2843-0-0-1, 先入, 大局, 根底, 文脈

2843-0-0-1 その言葉の表す内容．意義.

2843-0-0-2 表現や行為の意図・動機.

2843-0-0-3 表現や行為のもつ価値．意義.

実験の結果，単語「意味」の語義の分散表現は正しく作られていると考えられる．

7 考察

WSD による実験では提案手法よりも SVM による分類結果の方が高い正解率となった．その原因として提案手法は訓練データから得た語義の文脈ベクトルを用いているため，SVM のように未知データに対する汎用性がないからだと考えられる．そこで我々は SVM を学習させた際の識別超平面を用いる手法を実装し，精度の改善が見られるか実験を行った．

具体的には提案手法における語義の文脈ベクトルの代わりに，実験で学習した SVM の識別超平面

$$g(x) = \boldsymbol{v}^T x + b$$

の重み $\boldsymbol{v}$ を用いて語義の分散表現を構築する．実験設定で述べたように SVM を学習させる際の素性は語義の文脈ベクトルとなっているため，その識別超平面は教師データ中の用例を語義ごとに分離する一つのベクトルである．このベクトルを用いることで汎用性のある語義の分散表現を構築できると考えた．この手法を用いた WSD の分類結果を表 4 に示す.

実験の結果，SVM の正解率より僅かに低い正解率となったが提案手法に比べると WSD の精度を改善することができた.

8 おわりに

本論文では教師データを用いて語義の分散表現を構築する手法を提案した．具体的には教師データから得た語義の文脈ベクトルと語義の出現頻度を用いて単語の分散表現を語義ごとの分散表現に分解するというものである．得られた語義の分散表現を用いて WSD の分類実験を行った結果，MFS 値よりは高い正解率を出したものの SVM よりは低い正解率となった．しかし単語「意味」の各語義の類似単語を分析したところ，提案手法による語義の構築が有効であることを確認した.

表 4: SVM の識別超平面を用いた提案手法による WSD の分類結果

手法	平均正解率
SVM	0.757
提案手法	0.736
提案手法+SVM	0.752

謝辞

本研究の一部は国立国語研究所の共同研究プロジェクト「all-words WSD システムの構築及び分類語彙表と岩波国語辞典の対応表作成への利用」の研究成果を報告したものである.

参考文献

[1] Xinxiong Chen, Zhiyuan Liu, and Maosong Sun. A Unified Model for Word Sense Representation and Disambiguation. In *EMNLP-2014*, pages 1025–1035, 2014.

[2] Jiwei Li and Dan Jurafsky. Do multi-sense embeddings improve natural language understanding? *CoRR*, abs/1506.01070, 2015.

[3] Tomas Mikolov, Ilya Sutskever, Kai Chen, Greg S. Corrado, and Jeff Dean. Distributed Representations of Words and Phrases and their Compositionality. *NIPS*, 26:3111–3119, 2013.

[4] Arvind Neelakantan, Jeevan Shankar, Alexandre Passos, and Andrew McCallum. Efficient Non-parametric Estimation of Multiple Embeddings per Word in Vector Space. *CoRR*, abs/1504.06654, 2015.

[5] 浅原正幸, 岡照晃. nwjc2vec:『国語研日本語ウェブコーパス』に基づく単語の分散表現データ．言語処理学会第 23 回年次大会, to apper, 2017.