教師あり・教師なし学習により構築した語義の分散表現を用いた語義曖昧性解消に関する一考察

山木 翔馬　　新納　浩幸
茨城大学大学院　理工学研究科　　茨城大学　工学部　情報工学科

{16nm724r, hiroyuki.shinnou.0828}@vc.ibaraki.ac.jp

1　はじめに

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) は, 今や自然言語処理の様々な分野で有効な結果を残している. 近年ではさらに, 分散表現の学習手法を拡張して語義ごとの分散表現 (Multi-sense Embeddings) を構築する研究がなされており，いくつかのタスクでは単語の分散表現よりも語義の分散表現を用いる方が有効であることを示す報告もある．

本論文では教師あり学習，教師なし学習により構築した語義の分散表現を用いて語義曖昧性解消 (Word Sense Disambiguation; WSD) を行い，その結果をもとに各手法における語義の分散表現について考察した．

具体的には，我々の提案した学習済みの単語の分散表現を教師データを用いて語義の分散表現に分解する教師あり学習の手法 [5] と，Huangらの提案した Multi Sense Skip-gram(MSSG) モデル [1]，Neelakantanらの提案した MSSG モデルを拡張し語義の数を自動的に決める Non-Parametric Multi Sense Skip-gram(NP-MSSG) モデル [4] の教師なし学習の 3 つの手法によってそれぞれ語義の分散表現を構築し，その分散表現を使って WSD を行った．

実験では, 教師あり学習での語義の分散表現の構築で必要な学習済みの単語の分散表現として，国立国語研究所が作成した分散表現 nwjc2vec[6] を用い，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た．教師なし学習での語義の分散表現の構築で使用するコーパスは毎日新聞の 1993 年から 1999 年の新聞記事データとした．これらの語義の分散表現を用いて SemEval-2 の日本語辞書タスクデータの名詞単語 20 単語を対象に WSD の実験を行った．なお WSD の精度の比較として対象単語の前後 5 単語の分散表現の平均を素性とした SVM の分類器を実装した．実験の結果，どの手法も SVM より平均正解率が低かったが，MSSG モデルによる手法が他の 2 つの手法に比べて著しく高い正解率となった．

対象単語ごとの正解率を見ると，語義の出現頻度の差が大きい単語に対しては提案手法が高い正解率となり，語義の出現頻度の差が小さい単語に対しては MSSG モデルが高い正解率となることが分かった．また NP-MSSG モデルは他の手法に比べ精度が悪かったが，教師データ中に出現しない語義の分散表現を構築できる可能性があることが分かった．

2　関連研究

単語の分散表現を学習する手法は Feedforward Neural Network Language Model や Reccurent Neural Netword 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研究が多くなされているが，なかでも Mikolv らが提案した skip-gram モデルと CBoW モデル [3] は word2vec としてツール化され，分散表現を得る手段として広く使われている．

近年この skip-gram モデルを拡張して語義ごとの分散表現を構築する研究が多くなされている．Huangらの研究では1 つの単語にあらかじめ指定した語義数のベクトルを与えるモデルとして MSSG モデルを提案している。Neelakantanらは MSSG モデルをさらに拡張し，語義の数を自動で決めるノンパラメトリックな NP-MSSG モデルを提案している．

Li らの研究では語義の分散表現が実際の自然言語処理で有効であることを示している．[2] この研究では MSSG モデル・NP-MSSG モデルによって構築された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， part-of-speech tagging, semantic relation

identification, semantic relatedness のタスクにおいて語義の分散表現が有効であることを示している．

このように MSSG モデルを主流とする教師なし学習での語義の分散表現を構築する手法は多く研究されている．これに対して我々は，学習済みの単語の分散表現を教師データを用いて語義ごとの分散表現に分解するという教師あり学習による手法を提案した．提案手法により構築した語義の分散表現を用いた WSD の実験では，精度の向上は確認できなかったが，いくつかの単語については語義の分散表現が正しく作られていることが分かった．

本論文では我々が提案した教師あり学習による手法と，MSSG モデル，NP-MSSG モデルの教師なし学習による手法を用いて語義の分散表現を構築し，語義の分散表現を用いた WSD の実験結果を分析する．

3 教師あり学習による語義の分散表現の構築

単語の分散表現を $\boldsymbol{w}$, 語義 C_1, C_2, C_3（ 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 時限目の値においても

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

が成り立つ．この仮定をもとにして，語義 C_i の分散表現 $\boldsymbol{e_i}$ を次元ごとに求める．

具体的には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，ある語義の文脈ベクトルの k 次元目の値と他の語義の文脈ベクトルの k 次元目の値に大きな差がある場合，その語義の分散表現の k 次元目の値 e_k^i を w_k とし，その他の語義の分散表現の k 次元目の値 $e_k^j (i \neq j)$ を 0 とする．反対に，文脈ベクトルの k 次元目の値に大きな差がない場合は教師データ中の語義の出現頻度によって各語義の分散表現の k 次元目の値を次のように求める．

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

この操作を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．

4 教師なし学習による語義の分散表現の構築

skip-gram モデルが単語のベクトルと単語周りのコンテキストベクトルの類似度が高くなるように学習するのに対して，MSSG モデルは単語周りのコンテキストベクトルから平均コンテキストを求め，あらかじめ決められた単語の意味候補の中から一番コンテキストに類似した語義を選択する．選択された語義ベクトルとコンテキストベクトルの類似度が高くなるように学習することで，語義ベクトルが得られる．

つまり skip-gram における目的関数は，単語 w_t のベクトル $v(w_t)$ とコンテキスト c のベクトル $v(c)$ の類似度を

$$P(D = 1|v(w_t), v(c)) = \frac{1}{1 + e^{v(w_t)^T v(c)}}$$

としたとき，

$$\begin{aligned} J(\theta) &= \sum_{(w_t, c_t) \in D^+} \sum_{c \in c_t} \log P(D = 1|v(w_t), v(c)) \\ &+ \sum_{(w_t, c_t') \in D^-} \sum_{c' \in c_t'} \log P(D = 0|v(w_t), v(c')) \end{aligned}$$

を最大化するモデルである．MSSG モデルの目的関数は，平均コンテキストに最も類似する語義 s_t のベクトル $v_s(w_t, s_t)$ を用いることで

$$\begin{aligned} J(\theta) &= \sum_{(w_t, c_t) \in D^+} \sum_{c \in c_t} \log P(D = 1|v_s(w_t, s_t), v_g(c)) \\ &+ \sum_{(w_t, c_t') \in D^-} \sum_{c' \in c_t'} \log P(D = 0|v_s(w_t, s_t), v_g(c')) \end{aligned}$$

となり，このモデルを学習することで語義のベクトル $v_s(w_t, s_t)$ が得られる．

MSSG モデルでは語義の数はあらかじめ設定する必要があるが，NP-MSSG モデルは自動で語義の数を決める．具体的には，現在単語 w_t に割り当てられている語義の数を $k(w_t)$，w_t の平均コンテキストベクトルを $v_{context}$,k 番目の語義のコンテキストの中心を $\mu(w_t, k)$ としたとき，現在割り当てられている語義のベクトルと新たなコンテキストベクトルの最大類似度

$$\max_{k=1,2,\ldots,k(w_t)} sim(\mu(w_t, k), v_{context}(c_t))$$

が閾値 λ 以下であれば新しい語義番号を割り当て，そうでなければ $s_t = k_{max}$ とする．λ はハイパーパラメータである．

s_t が決まれば MSSG モデルと同様にして語義ベクトルを求めるとができる．

5 実験

構築された語義の分散表現の評価として，SemEval-2 の日本語辞書タスクのデータを用いて語義曖昧性解消を行った。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。このうち名詞である 20 単語について実験を行った．

また教師あり学習の手法で必要な学習済みの単語の分散表現は，国立国語研究所が作成した分散表現 nwjc2vec を用いた．

教師あり学習· 教師なし学習で用いるコーパスは毎日新聞の 1993 年から 1999 年の新聞記事データを用いた．

なお，構築する分散表現の次元数は，nwjc2vec に合わせて 200 次元とした．

WSD の識別は，テストデータの用例の平均文脈ベクトルを $\boldsymbol{u_j}$ としたとき，$\boldsymbol{u_j}$ との類似度が最大となる語義を選ぶことで行う．

$$argmax_i cos(\boldsymbol{e_i}, \boldsymbol{u_j})$$

ここで平均文脈ベクトル $\boldsymbol{u_j}$ とは，テストデータの j 番目の用例の対象単語の前後 5 単語の分散表現を $\boldsymbol{w_1}, \boldsymbol{w_2}, \ldots, \boldsymbol{w_{10}}$ とした時，

$$\boldsymbol{u_j} = \frac{\boldsymbol{w_1} + \boldsymbol{w_2} + \cdots + \boldsymbol{w_{10}}}{10}$$

とする．

また，比較のために訓練データの平均文脈ベクトルを素性として SVM で学習させた分類器も実装した．

MSSG モデルと NP-MSSG モデルによって構築された語義の分散表現を用いて WSD を行う場合，構築された語義の分散表現がテストデータのどの語義と対応するか判断できない．そのため本実験において語義の分散表現とテストデータの語義の対応は，各単語ごとに平均正解率が最も高くなるようにした．

6 実験結果

提案手法，MSSG モデル，NP-MSSG モデルによる平均正解率を表 1 に示す．

実験の結果，いずれの手法も SVM よりも低い正解率となった．特に語義数を自動的に決める NP-MSSG モデルは，他の手法に比べて著しく低い正解率となった．

また，提案手法, MSSG モデル, NP-MSSG モデルによる WSD の各単語に対する正解率を表 2 にまとめた．

手法	平均正解率
SVM	0.774
提案手法	0.699
MSSG モデル	0.737
NP-MSSG モデル	0.572

表 1: 各手法における平均正解率

対象単語	SVM	提案手法	MSSG	NP-MSSG
相手	0.68	0.58	0.72	0.48
意味	0.42	0.34	0.38	0.30
可能	0.58	0.50	0.54	0.50
関係	0.96	0.96	0.96	0.68
技術	0.78	0.68	0.76	0.52
経済	0.98	0.98	0.90	0.76
現場	0.76	0.62	0.74	0.42
子供	0.58	0.52	0.60	0.50
時間	0.76	0.72	0.72	0.58
市場	0.58	0.40	0.48	0.34
社会	0.88	0.84	0.82	0.80
情報	0.78	0.64	0.72	0.58
手	0.64	0.46	0.48	0.40
電話	0.80	0.68	0.70	0.56
場合	0.74	0.66	0.78	0.62
場所	0.90	0.90	0.90	0.78
一	0.92	0.92	0.92	0.42
文化	1.00	1.00	1.00	0.88
他	1.00	1.00	1.00	0.80
前	0.66	0.58	0.62	0.52
平均	0.774	0.699	0.737	0.572

表 2: 各単語に対する正解率

7 考察

WSD による実験では SVM に比べて提案手法，MSSG モデル，NP-MSSG モデルのいずれの手法も低い正解率となった．MSSG モデルと NP-MSSG モデルの正解率が低い原因としては，語義の分散表現の学習において別のコーパスを用いており，SemEval-2 の教師データを用いていないためであると考えられる．一方教師データを用いている提案手法の正解率が低い原因としては，単語の分散表現を語義の分散表現に分解する際に，特徴的な一つの語義に対して重みを与えているため，各語義の間で特徴の少ない単語について正しい学習ができていないのではないかと考えられる．

例えば「経済」「関係」「一」「文化」「他」のような特定の語義の出現頻度が際立って多い単語については SVM と同程度の高い正解率となっていることが確認できる．

また NP-MSSG モデルの特徴として，教師なし学習であるため，教師データ中に出現しない語義に対しても語義の分散表現を構築するというものがある．例

えば「 前」という単語の各語義の頻度は表 3 のようになっている．教師データ中に出現しない語義の用例がテストデータ中には 7 個あるが，NP-MSSG モデルでは単語「 前」に対して 3 つのベクトルを作っており，48488-X-X-X の語義の用例 7 個のうち 3 つの用例で正しい識別を行っていた．

このことから，NP-MSSG モデルは WSD の精度は低いものの，教師データ中に出現しない未知の語義に対する分散表現を構築できる可能性があると考えられる．

	教師データ	テストデータ
48488-0-0-1	19	12
48488-0-0-2	31	31
48488-X-X-X	0	7

表 3: 単語「 前」の語義の頻度

8 おわりに

本論文では教師あり学習，教師なし学習により構築した語義の分散表現を用いて WSD の実験を行った．実験の結果，どの手法も単純な SVM の分類器よりも平均正解率が低かったが，MSSG モデルによる手法は提案手法，NP-MSSG モデルの手法よりも高い平均正解率となることが分かった． また対象単語ごとに正解率を見ると「 一」や「 文化」「 他」など語義の出現頻度の差が大きい単語では提案手法が高い正解率となることが確認できた．

NP-MSSG モデルは他の手法と比べて正解率が低かったものの，教師データ中に出現しない未知の語義の分散表現を構築できる可能性があることが分かった．

これらの結果から，語義の分散表現を用いて WSD の精度を向上させるには語義の数と出現頻度の情報が重要であることが考えられる．そのため MSSG モデルはあらかじめ語義数を指定することで他の手法に比べ高い正解率となったが，学習過程で語義の出現頻度の情報を上手く使うことでより精度の良い語義の分散表現を構築することができるのではないかと考えられる．

参考文献

[1] Eric H Huang, Richard Socher, Christopher D Manning, and Andrew Y Ng. Improving word representations via global context and multiple word prototypes. In *Proceedings of the 50th Annual Meeting of the Association for Computational Linguistics: Long Papers-Volume 1*, pp. 873–882. Association for Computational Linguistics, 2012.

[2] Jiwei Li and Dan Jurafsky. Do multi-sense embeddings improve natural language understanding? *arXiv preprint arXiv: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. In *Advances in neural information processing systems*, pp. 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. *arXiv preprint arXiv:1504.06654*, 2015.

[5] 山木翔馬, 新納浩幸, 古宮嘉那子, 佐々木稔. 教師データを用いた語義の分散表現の構築. 言語処理学会第 23 回年次大会発表論文集, 2017.

[6] 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 flapper, 2017.