文脈を考慮した単語ベクトル集合からの単語領域表現

山内 崇史　　梶原 智之　　荒瀬 由紀
大阪大学
{yamauchi.takashi, arase}@ist.osaka-u.ac.jp, kajiwara@ids.osaka-u.ac.jp

1 はじめに

単語分散表現は自然言語処理において基本的な単語の表現方法であり，多くのタスクで利用されている．しかし，fastText[1] などの一般的な単語分散表現は，各単語を 1 つのベクトルで表現するため，多義性や意味の広がりを表現できないという課題を持つ．これらの課題を解決するために，単語をベクトルではなくガウス分布に埋め込む領域表現 [1, 2] が注目されている．

Vilnis and McCallum [1] は各単語をガウス分布でモデル化し，コーパスから分布の平均と分散を学習する w2g 法を提案した．Athiwaratkun and Wilson [2] はその発展として，各単語を混合ガウス分布によってモデル化する w2gm 法により，各単語に対して複数の語義を割り当てることを可能にした．これらの手法は単語間の意味的類似度推定タスクなどのベンチマークで単語分散表現よりも高い性能を達成したが，どちらの手法も全ての単語に対して同じ語義数を割り当てる，訓練時に固定長の窓幅の文脈しか考慮しない，という課題がある．

一方で，ELMo [3] や BERT [4] に代表される文脈化された単語分散表現が多くの自然言語処理タスクで高い性能を示している．文脈化された単語分散表現では，各単語が出現する文全体を考慮した単語表現を獲得できる．反面，同じ単語であってもそれが出現する文全体の文脈に応じて異なる単語分散表現を獲得するため，意味の広がりを表現することができない．また，文脈が存在しないタスクでは利用できない．

そこで本研究では，ELMo や BERT によって得られる文脈化された単語ベクトル集合をクラスタリングすることにより領域表現を獲得する手法を提案する．これにより，各単語の出現文全体の文脈を考慮した単語表現が得られることを期待する．また，動的にクラスタ数を決定できるクラスタリング手法を用いることで，各単語に対して適切な語義数が割り当てられることを期待する．図 1 に提案手法によって得られた単語領域表現を可視化した例を示す．`plant`（工場，植物）がそれぞれ `factory` および `flower` と重なる領域を持ち，意味の広がりを表現できていることが分かる．

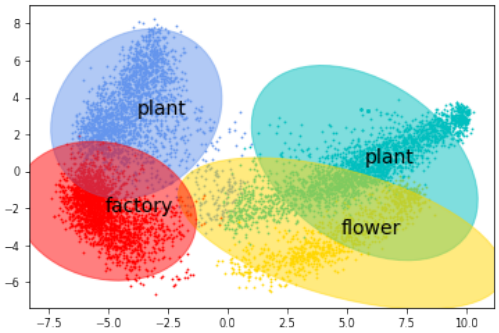

図 1: 単語領域表現の可視化の例

提案手法によって得られた単語表現を用いて，単語間の意味的類似度推定タスクおよび単語間の関係推定タスクによって評価を行った．その結果，文脈を考慮しないタスクでは既存手法と同等以上の性能を，文脈を考慮するタスクでは既存手法を上回る性能を示した．

2 文脈化された単語分散表現に基づく単語領域表現

本研究では文脈化された単語分散表現から領域表現を獲得する手法を提案する．単語を複数の領域で表現することにより，図 1 に示すように多義性や意味の広がりを捉えた単語表現を獲得することを目的とする．

まず，ELMo や BERT の訓練済みモデルを用いて，各単語についてコーパス内での出現回数個の文脈化された単語ベクトル集合を得る．そして，この単語ベクトル集合を Density Based Spatial Clustering of Applications with Noise (DBSCAN) [5] によりクラスタリングし，各クラスタを領域とみなした単語領域表現を獲得する．

[1] https://fasttext.cc/

DBSCAN は密度準拠型クラスタリングアルゴリズムである．その特徴として，事前にクラスタ数を指定する必要がなく，データに応じてクラスタ数を動的に決定できる点が挙げられる．これにより，各単語に適切な語義数の単語表現を割り当てられると期待する．

DBSCAN では 2 つのハイパーパラメータである ε および $MinPts$ を指定することで密度に基づくクラスタリングを行う．V_w 内の各ベクトルはコア点，到達可能点および外れ値に分けられる．コア点とは，その点から距離 ε 以内に $MinPts$ 個以上の点が存在する点のことを指す．到達可能点とは，コア点から距離 ε 以内に存在し，自身がコア点でない点を指す．コア点と到達可能点に含まれるデータはいずれかのクラスタに属し，外れ値に含まれるデータはどのクラスタにも属さない．本研究では，クラスタの平均ベクトルを単語の意味ベクトル $\boldsymbol{\mu}_w^n$ として用いる．また，全てのベクトルが外れ値となった場合は V_w の平均ベクトルを単語の意味ベクトル $\boldsymbol{\mu}_w$ として用いる．

3 評価実験

提案手法の効果を検証するため，単語間の意味的類似度推定タスクおよび単語間の関係推定タスクによって評価を行った．

3.1 提案手法の設定

English Wikipedia[2]から無作為抽出した 10 万記事を用いて，文脈化された単語分散表現の集合を得た．また，文脈化された単語分散表現モデルとして，ELMo は Original モデル[3]の 2 層目を，BERT は Base (uncased) モデル[4]の最終層を用いた．

またクラスタリング手法の影響も検証するため，DBSCAN に代わり混合ガウスモデル (GMM) によるクラスタリングを行う手法についても評価を行う．混合ガウスモデルでは与えられたデータを複数のガウス分布の重ね合わせで表現する．単語 w のベクトル集合 V_w に対してクラスタ数 K を与え，EM アルゴリズムにより尤度関数が最大になるようクラスタの更新を繰り返す．これにより，V_w を K 個のガウス分布の重ね合わせで表現する．各データは割り当てられた分布のクラスタに属し，V_w^0 から V_w^{K-1} の K 個のクラスタを構成する．本研究では，クラスタ K_w^n の平均ベクトルを単語の意味ベクトル $\boldsymbol{\mu}_w^n$ として用いる．

単語分散表現とクラスタリング手法の組み合わせとして，ELMo+DBSCAN，ELMo+GMM，BERT+DBSCAN，BERT+GMM を比較する．DBSCAN および GMM における距離関数としては，コサイン距離を用いる.[5]

3.2 文脈を考慮しないタスク

3.2.1 単語間の意味的類似度推定

単語間の意味的類似度推定タスク[6]により，提案手法によって得られる単語表現の基本的な性能を評価する．このタスクは単語対が与えられ，それらの単語間の意味的類似度を推定するタスクである．モデルによって推定された類似度と人手で付与された類似度のスピアマンの順位相関係数によって評価を行う．本実験では，モデルによる推定値としてクラスタの平均ベクトル間のコサイン類似度を用いる．

各データセットには開発セットが存在しないため，WordSim-353 を提案手法のハイパーパラメータを設定するために使用する．DBSCAN については $\varepsilon = 0.5$，$MinPts = 10$，GMM については $K = 2$ となった．また各単語は複数のクラスタを持ち得るため，要素数が最大のクラスタを用いるものとする．これにより，モデルの推定するスコアは各単語の最も一般的な出現文脈に対する意味の類似度となると考えられる．

実験結果を表 1 に示す．w2g および w2gm の性能は文献 [2] で報告されている値であり，w2gm のクラスタ数は $K = 2$ である．提案手法の中では BERT+DBSCAN が一貫して最高性能を示し，8 つのうちの 6 つのデータセットでは既存手法よりも優れた性能を達成した．したがって，提案手法によって得られる単語表現によって，既存手法と同等以上の性能で単語の意味を表現することが可能であると言える．

3.2.2 単語間の意味的関係推定

単語間の意味的関係推定タスクにより，提案手法によって得られる単語表現が上位下位関係などの意味的関係を表現できるか評価する．このタスクは単語対が与えられ，それらの単語間の意味的関係を推定するタ

[2]https://dumps.wikimedia.org/enwiki/20190620/
[3]https://allennlp.org/elmo
[4]https://github.com/google-research/bert

[5]ユークリッド距離および word mover's distance も検証したが，コサイン距離が最も高い性能を示した．
[6]https://github.com/mfaruqui/eval-word-vectors/

	SL	MEN	MC	RG	YP	MT-287	MT-771	RW	micro 平均
w2g [2]	29.4	72.6	76.5	73.3	42.0	64.8	60.9	28.8	52.4
w2gm [2]	29.3	**73.6**	79.1	74.5	45.1	**66.6**	60.8	28.6	52.9
ELMo+GMM	45.5	60.7	61.6	64.1	38.7	57.1	55.5	50.5	55.1
ELMo+DBSCAN	45.5	61.6	63.8	64.4	39.2	58.1	57.4	53.4	56.4
BERT+GMM	42.5	65.6	65.4	64.0	50.8	56.6	54.7	55.7	58.3
BERT+DBSCAN	**47.1**	71.0	**85.5**	**78.6**	**59.3**	60.2	**62.8**	**61.2**	**64.1**

表 1: 単語類似度データセットにおけるスピアマンの順位相関係数

スクである．本実験では，単語間の意味的関係推定を Shwartz and Dagan [6] に従って分類問題として解き，F 値で評価する．分類モデルは文献 [6] で用いられている SVM を単語ベクトルの連結 $[\boldsymbol{\mu}^n_{w_1}, \boldsymbol{\mu}^m_{w_2}]$ で訓練したモデル (DS) と，1 層の隠れ層を持つニューラルネットワークモデルを単語ベクトルの連結 $[\boldsymbol{\mu}^m_{w_1}, \boldsymbol{\mu}^n_{w_2}]$ で訓練したモデル (DSh) を用いる．本実験で用いるデータセット[7]は BLESS，ROOT09，EVALution であり，データの分割は文献 [6] に従う．

教師あり学習によって関係推定を行う場合，Lexical Memorization という問題がある．この問題は Levy ら [7] によって報告されており，分類器が単語間の関係を学習せずに典型的な上位語を記憶することで，汎化性能が低下するという問題である．本実験で用いるデータセットのうち，ROOT09 ではこの問題を防ぐために，上位語にランダムな下位語を組み合わせたもの (例 : (apple, animal), (dog, fruit)) が負例として含まれている．

提案手法では各単語について複数の領域表現が存在し得るため，どの領域表現を用いるかを決定する必要がある．そこで上位下位関係にある単語は意味も近いと仮定し，各入力単語対の複数の領域表現について，重心間の距離が最も近いものをそれぞれ上位語・下位語の領域表現として用いる．DBSCAN のハイパーパラメータ（ε，$MinPts$）および GMM のクラスタ数（K）は開発データによって設定した．その結果，DBSCAN については $\varepsilon = 0.1$，$MinPts = 10$，GMM については $K = 2$ となった．

実験結果を表 2 に示す．DS および DSh の性能は文献 [6] で報告されている値である．提案手法の中では ELMo+DBSCAN が一貫して最高性能を示し，ROOT09 および EVALution のデータセットでは既存手法よりも優れた性能を達成した．特に，Lexical Memorization 問題の存在しない ROOT09 で提案手法が既存手法よりも高い性能を示したということは，分類器が単に典型的な上位語を記憶するだけではなく，提案手法で得られる単語表現によって単語間の関係を認識できているからだと言える．

model	BLESS	ROOT09	EVALution
DS [6]	0.811	0.646	0.525
DSh [6]	**0.889**	0.716	0.571
ELMo+GMM	0.852	0.734	0.575
ELMo+DBSCAN	0.854	**0.743**	**0.588**
BERT+GMM	0.834	0.655	0.587
BERT+DBSCAN	0.822	0.671	0.568

表 2: 単語間の意味的関係推定タスクにおける F 値

3.3 文脈を考慮するタスク

多義語に対する単語表現の性能を評価するために，Stanford's Contextual Word Similarities (SCWS) のデータセット[8]を用いて文脈を考慮した単語間の意味的類似度推定を行う．このタスクではターゲットの単語対が文脈とともに与えられ，多義性を考慮して単語間の意味的類似度を推定する必要がある．

本タスクでは入力単語に文脈が存在するため，ELMo および BERT を用いて分散表現を獲得し，入力単語が属するクラスタを決定することで，用いる領域表現を決定する．具体的には，GMM では，ELMo および BERT により出力された単語分散表現が与えられたとき，各クラスタについてそのベクトルが属する確率を得ることができるため，その確率が最大である領域表現を用いる．DBSCAN では，ELMo および BERT により出力された単語分散表現をベクトル集合の 1 つ

[7] https://github.com/vered1986/LexNET/tree/v1/datasets

[8] http://www-nlp.stanford.edu/~ehhuang/SCWS.zip

model	ρ
w2g [2]	0.662
w2gm [2]	0.655
ELMo	0.676
ELMo+GMM	0.660
ELMo+DBSCAN（$\varepsilon = 0.1$）	**0.680**
ELMo+DBSCAN（$\varepsilon = 0.5$）	0.669
BERT	0.617
BERT+GMM	0.645
BERT+DBSCAN（$\varepsilon = 0.1$）	0.650
BERT+DBSCAN（$\varepsilon = 0.5$）	0.644

表 3: SCWS データセットにおけるスピアマンの順位相関係数

の要素とみなしてクラスタリングを行い，そのベクトルが属する領域表現を用いる．クラスタリングの結果，外れ値と分類された場合は出力された単語分散表現をそのまま用いる．意味的類似度の推定にはコサイン類似度を用いた．これにより，提案手法の推定するスコアは文脈を考慮した類似度となると考えられる．本タスクでは開発データが存在しないため，DBSCAN のハイパーパラメータ（ε，$MinPts$）および GMM のクラスタ数（K）は 3.2 節および 3.3 節でそれぞれ設定した値を用いた．

実験結果を表 3 に示す．w2g および w2gm の性能は文献 [2] で報告されている値である．ELMo および BERT は，それぞれのモデルから出力される文脈化された単語分散表現をそのまま用いた結果である．表 3 の結果から，提案手法の ELMo+DBSCAN が既存手法よりも高い性能を示すことが確認できる．また，DBSCAN のハイパーパラメータについては $\varepsilon = 0.1$ が $\varepsilon = 0.5$ よりも本タスクでは高い性能を示しており，ハイパーパラメータはタスクに応じて適切な値を用いる必要があることがわかる．さらに ELMo や BERT をそのまま用いる場合よりも高い性能を示しており，領域表現を生成する有効性が示された．

3.4 考察

3.2 節および 3.3 節の評価において，提案手法は一貫して既存手法と同等以上の性能を示した．既存手法では固定長の窓幅の文脈のみを考慮するが，提案手法では文全体を考慮した単語表現が得られる．そのため，特に文脈に応じた単語の意味を推定する 3.3 節のタスクにおいて提案手法が既存手法よりも高い性能を示したと考えられる．

また，単語ベクトル集合に対して混合ガウスモデルでクラスタリングしたモデルよりも DBSCAN でクラスタリングしたモデルがより高い性能を示した．このことから，全ての単語に定数個のクラスタを割り当てるよりも，単語ごとに動的にクラスタ数を決定することが有効であると言える．

4 おわりに

本研究では単語の多義性と意味の広がりを捉えるために，文脈化された単語ベクトル集合から領域表現を獲得する手法を提案した．評価実験の結果，文脈を考慮しないタスクでは既存手法と同等以上の性能を示し，文脈を考慮するタスクでは既存手法よりも高い性能を達成した．

今後の課題としては，単語の意味の表現としてクラスタの平均ベクトルを用いるだけでなく，クラスタそのものの性質を利用した距離尺度を定義することが挙げられる．それによって，ベクトルでは表現できない意味の広がりというような，領域表現固有の特徴を単語間の関係推定に用いることができると考えられる．

参考文献

[1] Luke Vilnis and Andrew McCallum. Word Representations via Gaussian Embedding. In *Proc. of ICLR*, 2015.

[2] Ben Athiwaratkun and Andrew Wilson. Multimodal Word Distributions. In *Proc. of ACL*, pp. 1645–1656, 2017.

[3] Matthew E. Peters, Mark Neumann, Mohit Iyyer, Matt Gardner, Christopher Clark, Kenton Lee, and Luke Zettlemoyer. Deep contextualized word representations. In *Proc. of NAACL*, pp. 2227–2237, 2018.

[4] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. BERT : Pre-training of Deep Bidirectional Transformers for Language Understanding. In *Proc. of NAACL*, pp. 4171–4186, 2019.

[5] Martin Ester, Hans-Peter Kriegel, Jörg Sander, Xiaowei Xu, et al. A Density-Based Algorithm for Discovering Clusters in Large Spatial Databases with Noise. In *Proc. of KDD*, pp. 226–231, 1996.

[6] Vered Shwartz and Ido Dagan. Path-based vs. Distributional Information in Recognizing Lexical Semantic Relations. In *Proc. of CogALex*, pp. 24–29, 2016.

[7] Omer Levy, Steffen Remus, Chris Biemann, and Ido Dagan. Do Supervised Distributional Methods Really Learn Lexical Inference Relations? In *Proc. of NAACL*, pp. 970–976, 2015.