Box Embedding による単語の分散表現獲得手法の検証

1　はじめに

深層学習など機械学習による自然言語処理は盛んな研究分野の一つである. 自然言語処理はもちろん機械学習においても単語の表現や処理方法は文章の意味を計算によって解析するために必要かつ根幹的な要素である. 近年, Word2Vec のような単語の分散表現ベクトルが広く使われる一方で単語の意味で表現しきれない要素があることが認識されている. その代表例として点でしかデータを表すことができないというものがある. 単語の埋め込みにおいてこのことは単語の意味の包含関係や階層関係といった集合的性質を自然に表現できないということに起因する. このような問題を解決するための表現として領域表現が提案されている. 本実験では始点と終点の組で表される「箱」への単語埋め込み表現である Box Embedding を扱う.

Box Embedding の獲得手法の一つである Word2Box による単語埋め込み表現を獲得し, その埋め込み表現が単語間の意味をどのように表しているかを実験により確認した.

2　要素技術

2.1　Box Embedding

Box Embedding は d 次元のベクトルの組からなる「箱」の領域で単語を表現するものであり, Vilnis Luke ら [1] によって提案された. 図 1 に Box Embedding の例を示す. 図 1 では mammal と carnivore の「箱」の共有部分にて cat の「箱」が存在するように, Box Embedding では「箱」同士の重なりによって単語の意味の包含関係や階層関係を表現する.

Box Embedding は「箱」の重なりを調整することで表現を獲得する. しかし, Vilnis Luke らによって提案された方法では「箱」が重ならないと勾配が計算できず最適化が難しくなる問題がある. この問題に対して「箱」の境界を滑らかだとして領域を計算することで「箱」同士が離れていても勾配を求められ最適化できる. 図 2 に Gumbel 分布を適用した 2 次元での「箱」を示す. Gumbel 分布を適用することで「箱」をうまく最適化することができ Box Embedding を獲得しやすくなる [2].

図 1: 2 次元での Box Embedding の例 (文献 [3] の Figure 1 から引用)

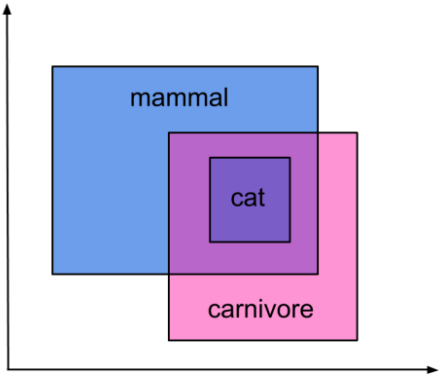
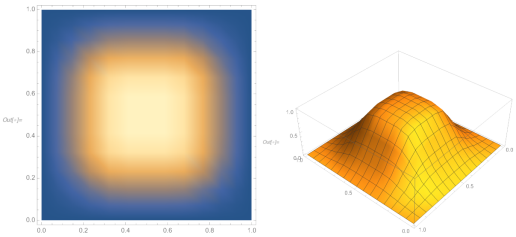


図 2: 2 次元での Gumbel Box の例 (文献 [4] の Figure 3 から引用)



2.2　Word2Box

Word2Box [5] は Sankar Dasgupta らによって提案された手法であり, 単語の領域表現の Box Embedding を教師なし学習で獲得するものである. 学習方法は Word2Vec [6] と同様で, 似た意味の単語は似た文脈に現れるという仮説をもとに対象単語とその周辺単語の関係を学習することで獲得する. また, 今回の実験では CBOW モデルのように対象の単語とその周辺単語との「箱」同士の重なりが大きくなるように学習する. 加えて,「箱」をただ大きくする最適化にならないように適当な語を負例として抽出し周辺単語との重なりを小さくなるように学習もする.

3　データセット

3.1　Penn Treebank データセット

Penn Treebank データセット [7] は Mitchell P. Marcus らによって作成された 450 万語を超える大規模英語コーパスであり, コーパスには品詞情報も付加されている. この実験では数字を全て “N” , 語彙に含まれない単語を “<unk>” , パディングを “<pad>” としてそれぞれ一つの単語と見なして語彙数が 1 万語になるように調整されたものを用いた.

3.2　単語類似度データセット

実験で単語の類似性を確かめるために単語類似度データセットとして Simlex-999 [8], WordSim-353 [9], YP-130 [10], MEN [11], MC-30 [12], RG-65 [13], VERB-143 [14], Stanford RW [15], Mturk-287 [16], Mturk-771 [17], SimVerb-3500 [18] を用いた. ただし, 各データセットの特徴に関しての詳細はそれぞれの文献を参照してもらいたい. また, WordSim-353 は関連する単語組で構成されるサブセットである WS-353 (Rel) と類似する単語組で構成されるサブセットである WS-353 (Sim) も用いた.

4　実験

Word2Box による Box Embedding が単語の意味関係を獲得できているかの確認を目的として, Word2Box で学習したモデルで類似単語組の類似度による順位を求めた. Word2Box の著者による実装[1]を用いて Penn Treebank コーパスから単語の埋め込みモデルを獲得をした. 表 1 に埋め込み表現を学習する際のモデルの学習パラメータを示す. 各単語類似度データセットをもとに類似度の高い単語の組について埋め込み表現同士での類似度を計算し, それぞれの単語においてもう一方の単語が類似度で学習した語彙の中で上位何番目に位置するかを求めた. 各類似度データセットは数値の基準がそれぞれ異なるため, 基準値を上回る単語組を実験に用いた. 表 2 に各データセットに対して定めた基準値を示す. ただし, 基準値は各データセットの類似度の最大値の 80 % としている. 得られた結果から Word2Box が 2 単語間の類似性を適切に学習し Box Embedding で表現できているかを確認した.

[1] https://github.com/iesl/word2box

表 1: Word2Box モデルの学習パラメータ

parameter	value
次元数	64
バッチサイズ	4096
エポック数	10
ウィンドウサイズ	5
ネガティブサンプル数	2
サブサンプル閾値	0.001
学習率	0.004204091643267762

5　結果と考察

表 3 に実験結果を示す. どのデータセットにおいても類似度が上位 1 から 99 番目に含まれる類似単語のペアはほとんど存在せず, 上位 1000 番目までにでも含まれる類似単語ペアの割合は少数となった.

続いて表 4 に類似度の順位が 1 から 99 番目に含まれていた単語組を示す. 含まれていた単語組と含まれなかった単語組との間に使われる文脈や意味において法則性を見出すことはできなかった. また, データセット間での特徴や法則性も見受けられなかった.

以上よりこの実験において Word2Box は単語同士の関係を適切に表現する Box Embedding を獲得していないように考えられる.

6　今後の課題

今回の実験結果は Word2Box によって獲得した Box Embedding について意味が近い単語間では似た分散表現になるという性質を示しておらず, その点において十分な表現能力があることを確認できなかった. 提案論文 [5] において学習パラメータはいくつかの数値からランダムに選択することで決定し, 実験とは異なる学習データを用いているため, 学習パラメータの調整や学習用の文章の変更によって性能の向上の可能性を模索する余地があると考えられる. 加えて今回は Word2Vec に代表されるようなベクトル埋め込み表現と Word2Box の性能差を確かめる実験ができなかったが他の領域埋め込み手法も加えて表現能力の比較検証の必要がある. 同じように今回検証できなかった Box Embedding による単語の意味の階層関係や包含関係の表現も確認したい. また, 日本語のデータに対して Word2Box と　Box

表 2: 類似度データセットから抽出した単語組

	Stanford RW	RG-65	YP-130	MEN	MC-30	Mturk-287	SimVerb-3500	SimLex-999	Mturk-771	WS-353(Sim)	WS-353(All)	WS-353(Rel)	VERB-143
基準値	8.0	3.2	3.2	40	3.2	4.0	3.2	8.0	4.0	8.0	8.0	8.0	3.2
該当組数	36	4	12	166	3	20	1053	68	85	16	31	14	0

表 3: 類似単語ペアが含まれる順位

順位	1-99	100-499	500-999
Stanford RW	0	7	0
RG-65	0	0	0
YP-130	0	2	2
MEN	6	24	19
MC-30	0	0	0
Mturk-287	0	5	0
SimVerb-3500	19	74	86
SimLex-999	5	8	9
Mturk-771	1	11	3
WS-353(Sim)	1	1	2
WS-353(All)	2	1	2
WS-353(Rel)	1	0	0
VERB-143	0	0	0

Embedding を適用した場合においても意味の包含関係や階層関係が表現されるのかを確認したい.

参考文献

[1] Luke Vilnis, Xiang Li, Shikhar Murty, and Andrew McCallum. Probabilistic embedding of knowledge graphs with box lattice measures. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 263–272, Melbourne, Australia, July 2018. Association for Computational Linguistics.

[2] Shib Sankar Dasgupta, Michael Boratko, Dongxu Zhang, Luke Vilnis, Xiang Lorraine Li, and Andrew McCallum. Improving Local Identifiability in Probabilistic Box Embeddings. *arXiv e-prints*, p. arXiv:2010.04831, October 2020.

[3] Xiang Li, Luke Vilnis, Dongxu Zhang, Michael Boratko, and Andrew McCallum. Smoothing the geometry of probabilistic box embeddings. In *International Conference on Learning Representations*, 2019.

[4] Michael Boratko, Javier Burroni, Shib Sankar Dasgupta, and Andrew McCallum. Min/max stability and box distributions. In Cassio de Campos and Marloes H. Maathuis, editors, *Proceedings of the Thirty-Seventh Conference on Uncertainty in Artificial Intelligence*, Vol. 161 of *Proceedings of Machine Learning Research*, pp. 2146–2155. PMLR, 27–30 Jul 2021.

[5] Shib Dasgupta, Michael Boratko, Siddhartha Mishra, Shriya Atmakuri, Dhruvesh Patel, Xiang Li, and Andrew McCallum. Word2Box: Capturing set-theoretic semantics of words using box embeddings. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 2263–2276, Dublin, Ireland, May 2022. Association for Computational Linguistics.

[6] Tomas Mikolov, Kai Chen, Greg Corrado, and Jeffrey Dean. Efficient Estimation of Word Representations in Vector Space. *arXiv e-prints*, p. arXiv:1301.3781, January 2013.

[7] Mitchell P. Marcus, Mary Ann Marcinkiewicz, and Beatrice Santorini. Building a large annotated corpus of english: The penn treebank. *Comput. Linguist.*, Vol. 19, No. 2, p. 313–330, jun 1993.

[8] Felix Hill, Roi Reichart, and Anna Korhonen. SimLex-999: Evaluating semantic models with (genuine) similarity estimation. *Computational Linguistics*, Vol. 41, No. 4, pp. 665–695, December 2015.

[9] Lev Finkelstein, Evgeniy Gabrilovich, Yossi Matias, Ehud Rivlin, Zach Solan, Gadi Wolfman, and Eytan Ruppin. Placing search in context: The concept revisited. In *Proceedings of the 10th International Conference on World Wide Web*, WWW '01, p. 406–414, New York, NY, USA, 2001. Association for Computing Machinery.

表 4: データセットごとのモデルによって類似度が高いとされた単語組

データセット	類似単語組
MEN	(flight, plane), (highway, traffic), (cold, frozen), (mountain valley), (bay, beach), (burger, meat)
SimVerb-3500	(let, release), (tell, testify), (release, go), (express, tell), (save, rescue), (release, relax), (tell, predict), (spend, use), (say, tell), (tell, respond), (fight, defend), (rescue, help), (leave, release), (release, drain), (release, loosen), (remove, release), (rescue, find), (gamble, bet), (gamble, risk)
SimLex-999	(essential, necessary), (area, zone), (pact, agreement), (physician, doctor), (expand, grow)
Mturk-771	(season, winter)
WS-353(All)	(king, queen), (opec, oil)
WS-353(Rel)	(opec, oil)
WS-353(Sim)	(king, queen)

[10] Dongqiang Yang and David Powers. Verb similarity on the taxonomy of wordnet. *Proceedings of the 3rd International WordNet Conference (GWC)*, pp. 121–128, 01 2006.

[11] Elia Bruni, Nam Khanh Tran, and Marco Baroni. Multimodal distributional semantics. *J. Artif. Int. Res.*, Vol. 49, No. 1, p. 1–47, jan 2014.

[12] George A. Miller and Walter G. Charles. Contextual correlates of semantic similarity. *Language and Cognitive Processes*, Vol. 6, No. 1, pp. 1–28, 1991.

[13] Herbert Rubenstein and John B. Goodenough. Contextual correlates of synonymy. *Commun. ACM*, Vol. 8, No. 10, p. 627–633, oct 1965.

[14] Simon Baker, Roi Reichart, and Anna Korhonen. An unsupervised model for instance level subcategorization acquisition. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pp. 278–289, Doha, Qatar, October 2014. Association for Computational Linguistics.

[15] Thang Luong, Richard Socher, and Christopher Manning. Better word representations with recursive neural networks for morphology. In *Proceedings of the Seventeenth Conference on Computational Natural Language Learning*, pp. 104–113, Sofia, Bulgaria, August 2013. Association for Computational Linguistics.

[16] Thang Luong, Richard Socher, and Christopher Manning. Better word representations with recursive neural networks for morphology. In *Proceedings of the Seventeenth Conference on Computational Natural Language Learning*, pp. 104–113, Sofia, Bulgaria, August 2013. Association for Computational Linguistics.

[17] Thang Luong, Richard Socher, and Christopher Manning. Better word representations with recursive neural networks for morphology. In *Proceedings of the Seventeenth Conference on Computational Natural Language Learning*, pp. 104–113, Sofia, Bulgaria, August 2013. Association for Computational Linguistics.

[18] Daniela Gerz, Ivan Vulić, Felix Hill, Roi Reichart, and Anna Korhonen. SimVerb-3500: A large-scale evaluation set of verb similarity. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pp. 2173–2182, Austin, Texas, November 2016. Association for Computational Linguistics.