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不可欠な根幹的な要素である. 近年, Word2Vec のような単語のベクトル表現が広く使われる一方で単語の意味で表現することができない要素があることが認識されている. その代表的なもの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である. 単語のベクトル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な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おり, ガウス分布を用いたものや双曲空間を用いたものがある. 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 の獲得手法の一つとして Word2Box がある. この実験では 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 [2] によって提案された. 図 1 が示す 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. しかし, Box Embedding は「箱」の重なりを調整することで表現を獲得するが, 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. Sankar Dasgupta ら [3] によって提案された手法はそのうちの一つであり, Gumbel 分布を用いるものである. Gumbel 分布を適用した 2 次元での「箱」は図 2 のようになる.

図 1: 2 次元での Box Embedding の例 (文献 [?] Figure 1 引用)

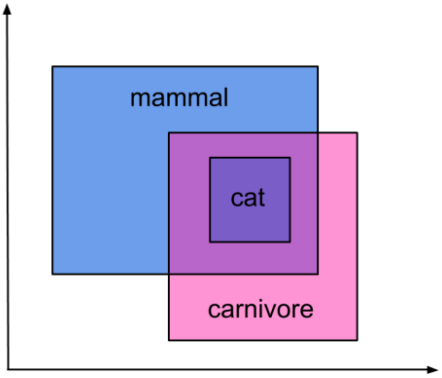
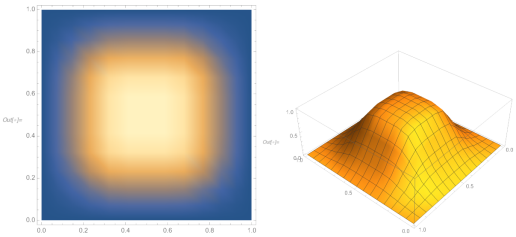


図 2: 2 次元での Gumbel Box の例 (文献 [1] Figure 3 引用)



2.2　Word2Vec

Word2Vec [4] は 2013 年に Google の Tomas Mikolov らによって発表された単語のベクトル埋め込み表現の生成モデルである. Word2Vec は 2 層ニューラルネットワークであり, 対象とする単語の周辺に現れる単語の頻度をもとに類似する周辺単語を持つ単語とのベクトルの類似度が大きくなるように学習する. 周辺の単語から対象の単語を予測し学習する CBOW モデルと対象の単語から周辺の単語を予測し学習する Skip-gram モデルがある.

2.3 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 の CBOW モデルと同様に対象の単語とその周辺単語の「箱」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するように学習する.

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 [6] 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

3.2.1 SimLex-999

SimLex-999 [7] は Hill Felix らによって発表された英語の 999 組の単語間の関連性ではなく類似性に基づいて, 単語間の類似度を 0 から 10 の範囲で定量化したデータセットである.

3.2.2 WordSim-353

WordSim-353 [8] は Finkelstein Lev らによって作成された英語の 350 個の単語組をその関係度で 0 から 10 の範囲で評価したデータセットである. このデータセットでは関連性と類似性による評価が混在しており, 関連する 252 単語組のサブセットである WS-353(Rel) と 類似する 203 単語組のサブセットである WS-353(Sim) がある.

3.2.3 YP-130

YP-130 は () らによって作成された英語の動詞の 130 個の単語組についての類似性を 0 から 4 の範囲で評価したデータセットである.

3.2.4 MEN

MEN は () らによって作成された画像情報なども活用した英単語 3000 組についての意味的類似性

4 実験

Word2Box の著者による実装 `https://github.com/iesl/word2box` を用いて表 1 のパラメータで Penn Treebank コーパスから単語の埋め込みモデルを獲得をした. 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, 表 2 に示す基準値を上回る単語組を実験に用いた. 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. 求めた結果から 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.

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

実験結果は表 3 のようになった. どのデータセットにおいても類似度が 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しなかった. それどころか上位 1000 番目までに含まれる類似単語ペアの割合は少数であった. また, 表 4 のようにデータセットの種類にかかわらずほとんどの単語ペアにおいて類似度の順位を求める対象単語を入れ替えると順位が大きく変動することが確認された. ただし, ここで大きく変動したとする基準は順位が 1000 異なる場合とした. この大きく変動する組においても単語組が類似性を示す結果とはならなかった.

表 2: 類似度データセットから抽出した単語組

項目	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

次に類似度の順位が 1 から 99 番目に含まれていた単語組を表 5 に示す. 含まれていた単語組と含まれなかった単語組との間に法則性を見出すことはできなかった.

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

表 4: 対象単語入れ替えによる順位変動

データセット	変動が大きい組	組の総数
Stanford RW	28	36
RG-65	4	4
YP-130	9	12
MEN	123	166
MC-30	3	3
Mturk-287	14	20
SimVerb-3500	847	1053
SimLex-999	54	68
Mturk-771	53	85
WS-353(Sim)	14	16
WS-353(All)	24	31
WS-353(Rel)	10	14
VERB-143	0	0

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. 提案論文において [5] 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. また, Box Embedding の特徴の一つでもある集合による単語の意味の包含関係の表現や多義語に対する表現の検証を今回の実験ではできなかった.

参考文献

[1] 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.

[2] 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.

[3] 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.

[4] 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.

表 5: 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

[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] 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.

[7] 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.

[8] 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.