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 [?] によって提案された. 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 [?] 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.

2.2 Word2Vec

Word2Vec [1] 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[?] 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. この実験では Dasgupta らによる実装である `https://github.com/iesl/word2box` を利用した.

3 データセット

3.1 Penn Treebank データセット

Penn Treebank データセットは

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

3.2 単語類似度データセット

3.2.1 SimLex-999

3.2.2 WordSim-353

3.2.3 YP-130

3.2.4 MEN

3.2.5 MC-30

3.2.6 RG-65

3.2.7 VERB-143

3.2.8 Stanford RW

3.2.9 Mturk-287

3.2.10 Mturk-771

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, 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.

5 結果と考察

6 今後の課題

今後の課題はうんぬんかんぬん〜

参考文献

[1] 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 2: Word2Box モデルの学習パラメータ

項目	SimLex-999	WordSim-353	YP-130	MEN	MC-30	RG-65	VERB-143	Stanford RW	Mturk-287	Mturk-771
基準値	8	8	3.2	40	3.2	3.2	3.2	8	4	4
該当組数										