

Box Embedding による単語の分散表現獲得手法の検証

1 はじめに

機械学習や深層学習による自然言語処理が盛んである。自然言語処理において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を意味する。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」による埋め込み表現獲得手法 Box Embedding を単語に適用した Word2Box の単語表現と近い意味を持つ単語

このような領域表現のうちの単語へと適用したものが Word2Box である。本実験では Word2Box での単語埋め込みによる表現と単語の概念の包含・階層関係の確認し, 加えて Word2Vec との比較実験をした。

2 要素技術

2.1 Box Embedding

2.2 Word2Vec

Word2Vec 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 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。Box Embedding では d 次元空間において d 個の始点と終点の組で「箱」を表現し, 「箱」にデータを埋め込む。Box Embedding の学習は「箱」同士の重なりを調整することで実現されるが, Word2Box では Gumbel 分布を用いて「箱」の端をなめらかにしており, 「箱」が重ならない場合の学習の最適化をしている。Word2Box において単語の学習は Word2Vec で用いられている CBOW と同じ考え方によって実現され, 文から単語の学習データである「箱」を抽出している。

3 データセット

提案手法はうんたらかんたら ~

4 実験

実験をかくかくしかじか ~

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] 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.