Word2Vec で表現された単語の意味の可視化に関する検討 Consideration of Visualizing Word Meanings in Word2Vec Space

森 祥恭 Yoshiyuki Mori 高間 康史 Yasufumi Takama

首都大学東京大学院システムデザイン研究科 Graduate School of System Design, Tokyo Metropolitan University

This paper investigates the meaning of words acquired by Word2Vec using visualization. Word embedding such as Word2Vec has been popular in different kinds of applications using text data. Word embedding represents words in a corpus as vectors, based on which calculation in terms of those meanings is possible. However, it is difficult to interpret dimensions of a vector space. In order to investigate word categories of which meaning is obtained by Word2Vec, and dimensions corresponding to the meaning, this paper examines the variance of word vectors belonging to a word category, and visualizes word vectors using PCA-based scatter plot and radar chart.

1. はじめに

近年,自然言語処理の分野で単語の分散表現を学習する手 法が注目を集めている.単語の分散表現は,大規模なテキスト コーパスを利用して学習することにより獲得され,コーパスに出 現する単語をベクトルで表現することができる.これを利用した 研究や様々な応用が試みられている.

分散表現の特徴の一つとして、単語間のベクトル演算を行う ことで国と首都や男性と女性といった、単語間の意味的な関係 を求めることができる.また、学習した単語ベクトルにコサイン類 似度を利用することで、単語同士の類似度を求めることもできる. しかし、単語の分散表現で生成された空間が高次元であること、 および各次元の意味が自明でないため解釈が困難といった問 題がある.どの様な意味が獲得されているか、ベクトル空間のど の次元が意味の表現に関係しているかなどが確認可能となれ ば、分散表現を利用した学習結果や分析結果の理解に貢献す ると考える.

本稿では、Word2Vec により得られた単語の分散表現が持つ, 単語の意味や単語間の関係を情報可視化手法を用いて分析 することを目的とする. WordNet の単語カテゴリを利用した分析 結果により,所属する単語に対応したベクトルの分散が小さいカ テゴリが存在することや,特定のカテゴリと関係の強い次元が存 在することを示す.

2. 関連研究

2.1 単語の分散表現

単語の分散表現とは、各単語を高次元の実数ベクトルで表現したものであり、大規模テキストコーパスを利用して求められる.

Word2Vec は 2 層からなるニューラルネットワークを用いて単 語の分散表現を学習する手法である. 学習する方法として CBoW(Countinues Bag-of-Words)モデル[2][3]と Skip-gram モ デル[1][2]の二つのモデルが提案されている. CBoW モデルは テキスト内の周辺にある単語を利用して対象となる単語を予測 し、単語の分散表現を学習する. Skip-gram モデルは CBoW モ デルとは逆に、ある単語が与えられた時に、その周辺の単語を 予測し、単語の分散表現を学習する. Word2Vec ではベクトルの 演算を行うことで、単語間の意味的な関係を求めることが可能 である. 例えば、フランスのベクトルからパリのベクトルを引いた 結果に東京のベクトルを加えると、日本のベクトルに近くなること などが知られている.

2.2 高次元データの可視化手法

高次元を対象とした可視化手法の代表的なものに、PCA(主 成分分析)などの次元圧縮による可視化手法や、平行座標など のように次元圧縮を行わない可視化手法が存在する.次元圧 縮による可視化手法は2次元空間などの低次元空間に高次元 空間を写像する.平行座標による可視化では、各次元をそれぞ れ直交する軸として平行に並べ、各データを一つの線グラフと して描画する.これにより、高次元データの情報を損なわずに、 次元間の関係の解釈が可能になる.

3. 提案手法

3.1 学習データ

本稿では、Word2Vec で求めた分散表現において、ある単語 カテゴリの意味やカテゴリ間の関係が獲得されているか分析す る. Word2vec の学習には、英語版 Wikipedia のオリジナルのテ キストデータに前処理を施した text8¹を利用する.このデータセ ットの単語総数は 253,854 単語である.

英語の概念辞書である WordNet²を利用する. WordNet では 単語が同義語のグループごとに分けられ,定義や他の同義語 のグループとの関係が記述されている.本稿では WordNet に おける反意語と同義語を利用して単語のカテゴリ分けを行う.

3.2 単語ベクトルの分散による分析

WordNet を利用して学習データに存在する単語をカテゴリに 分類し、カテゴリごとに所属する単語の重心ベクトルを求める. 重心ベクトルと各単語との距離から分散を求めることにより、カテ ゴリに属している単語が高次元空間において互いに近い位置 に存在しているのか、遠い位置に存在しているのかを判断する.

連絡先:高間康史,首都大学東京大学院システムデザイン研 究科,〒191-0065 東京都日野市旭が丘 6-6, ytakama@tmu.ac.jp

¹ <u>http://mattmahoney.net/dc/text8.zip -O text8.gz</u>

² <u>https://wordnet.princeton.edu/</u>

分散が小さい場合,当該カテゴリの意味が分散表現により獲得 されていると仮定する.

3.3 次元圧縮による単語ベクトルの可視化

Word2Vec で得られたベクトル空間に PCA を適用し、2 次元 空間に可視化する. 3.2 節で求めた分散が小さいカテゴリに属 する単語や,関係のある 2 カテゴリにそれぞれ属する単語間の 関係を 2 次元空間で可視化することで,カテゴリの意味やカテ ゴリ間の関係を視覚的に確認する.

3.4 単語カテゴリ特有の次元の分析

分散表現によりその意味が獲得されたと判断した単語カテゴ リやカテゴリ間の関係について、その意味と関連する単語ベクト ルの次元について考察する.同じカテゴリに属する単語ベクトル、 あるいは単語間の演算により得られたベクトルの要素が特定の 次元に集中しているかを確認するために、要素値の平均値、分 散に基づき分析を行うほか、平行座標やレーダーチャートなど により視覚的に確認する.

4. 分析結果

単語の分散表現として、Word2Vec を以下の条件で学習したものを使用した。

- 単語ベクトルの次元数:200
- テキストコーパス内での出現回数が5回未満の単語を無視
- 予測する単語の前後5単語を学習に使用
- 学習モデル:CBoW

WordNet で分類した各カテゴリ内に存在する単語ベクトルの 分散を表 1 に示す.各行に、意味的な関係を持つカテゴリの分 散,両カテゴリでの対応関係にある単語ベクトルの差について 求めた分散を示す.全単語から求めた分散は 9.713 であり、国、 首都,原作者、文章,善はこれよりも分散が小さいことがわかる. 分散が小さいカテゴリは、単語間の意味的なまとまりが強いと考 えられる.また、差についてはどの対も分散が小さくなっている ことがわかる.

カテゴリ名	分散	カテゴリ名	分散	差の分散
玉	7.829	首都	6.458	6.782
男性	17.83	女性	14.58	3.297
始まり	14.46	終わり	11.36	4.385
原作者	9.157	文章	5.546	6.303
善	8.658	悪	13.15	3.093

表 1:分類したカテゴリとカテゴリ内の分散



男性及び女性カテゴリに属する単語について,それぞれ対応関係にある単語の差を PCA により可視化した結果を図1に示す. 図において,すべての単語のペアについて男性の単語が同じ位置に来るように平行移動している. 単語間の対応関係は,WordNetの反意語に基づき判断している. "Female"と"Girl"を除いて男性と女性の位置関係が似た傾きで存在していることがわかる.

男性と女性カテゴリの単語集合の重心ベクトルそれぞれについて、その絶対値が大きい10次元を求めた結果を表2に示す. 表に示す数値は次元のIDであり、同じ値は同じ次元に対応する.男性-女性は、関係のある単語同士で求めた差ベクトルの重心から求めたものである.男性、女性カテゴリにおいて、8次元が一致しており、性別に関係すると考察できる.また、女性の重心ベクトルのみに存在している2次元(ID:189, ID:23)が男性-女性の上位10次元内に存在しており、女性としての意味が強い次元と考察できる.男性と女性に共通して存在する次元のうち、ID:94が男性-女性の次元にも存在しており、男性と女性の違いを反映していると考察できる.

順位	1	2	3	4	5
男性	62	97	83	81	44
女性	94	62	44	83	97
男性-女性	189	86	197	94	123
順位	6	7	8	9	10
順位 男性	6 110	7 94	8 147	9 33	10 98
順位 男性 女性	6 110 189	7 94 147	8 147 98	9 33 33	10 98 23

表 2:重心ベクトルの絶対値が大きい 10 次元

5. おわりに

本稿では、単語の分散表現を学習する手法である Word2Vec により得られた、単語の意味や単語間の関係につい て分析を行った.同一カテゴリに属する単語ベクトルの分散を 求めることで、カテゴリに属する単語の意味的なまとまりを分析し た.また、各カテゴリの意味を表現する特有の次元が存在するこ とも示した.

今後は、より多くのカテゴリに関して同様に分析し、分散表現 で獲得された意味や関係の理解を深めていくことで、単語の意 味を利用する各種応用に貢献すると考える.また、学習に用い るデータや学習モデルの違いが、獲得される意味などに与える 影響を分析することも有用と考える.

参考文献

- [Mikolov 2013] T. Mikolov, I. Sutskever, K. Chen, G. Corrado, J. Dean: Distributed Representations of Words and Phrases and their Compositionality, NIPS2013, pp.3111-3119, 2013.
- [Mikolov 2013] T. Mikolov, K. Chen, G. Corrado, J. Dean: Efficient Estimation of Word Representations in Vector Space, ICLR2013, 2013.
- [Le 2014] Q. Le, T. Mikolov: Distributed Representations of Sentences and Documents, ICML2014, pp.1188-1196, 2014.