The 33rd Annual Conference of the Japanese Society for Artificial Intelligence, 2019

Word2vecを利用した文章校正支援に向けて

Toward proofreading support using Word2vec

丸山 正人[*1]
Masato Maruyama

竹川 高志[*1]
Takashi Takekawa

[*1] 工学院大学情報学部
Faculty of Informatics, Kogakuin University

Abstract: Word2vec is a method of for learning distributed vector representations that semantic word relationships. In this paper, we aim for detection words misused from a sentence by distributed vector representations built using Word2vec. As a result, using linear discriminant analysis improved the detection rate of words misused.

1. はじめに

自然言語で扱われる単語の意味を計算機に理解させることは自然言語処理の分野において重要な課題である. この解決策として, 分布仮説という同じ文脈で出現する単語は同じ意味を持つ傾向があるという事[1]に基づき, ある単語とその周辺に出現した単語を学習し単語間の関係をベクトルで表した分散表現という手法がある. これは現在では自動音声認識および機械翻訳への応用など広範囲の自然言語処理タスクへ応用されている[2]. 近年 Tomas Mikolov らによって研究された Word2vec は分散表現を得る手法であり, 従来の手法より高速な学習を可能にし[3], 単語同士の関係が vec(“Madrid”) - vec(“Spain”) + vec(“France”) の結果に最も近い単語ベクトルが vec(“Paris”)となるような線形構造を持つ[2]ことで近年話題になっている.

また, 自然言語処理のタスクとして文章から誤りを検出,訂正する作業として校正があるが, 検出, 訂正される誤りは事前に予測ができる習慣的なものや, 文から単語の分割を行う処理である形態素解析でわかるような表層的なものに限られており, 文脈や一般常識を用いた解析により判断できるような, 内容レベルの校正は困難とされている[4].

そこで Word2vec を用い, 実在している単語かつ文法として正しく用いられる固有名詞の誤字の検出を本研究の目的とする. これは実在している単語であり文法が正しいために事前の予測のみでの解決や, 形態素解析や文法からの検出が難しいと考えられるような内容レベルの校正になると考えられ, 助詞や代名詞などの,文脈に制限されずに出現が考えられる品詞ではなく特定の文脈でのみ出現するような固有名詞であるため, 分布仮説に基づいた分散表現を出力する Word2vec を用いる事が適切と考えられるためである.

以上のことから, 解析対象の文章に含まれていない固有名詞を, 文脈に出現しないような単語である誤字として, 文章内の固有名詞のベクトルの平均と各固有名詞のベクトルについてユークリッド距離を用いた誤字の検出を行い, それを通して固有名詞のベクトルの分布を調査した.

その結果, 文章内の固有名詞のベクトルの平均と各固有名詞のベクトルについてのユークリッド距離が文章内の固有名詞のベクトルと文章に含まれていない固有名詞のベクトルで分布が重なっていることが分かり, 文章内の固有名詞について主成分分析を行ったものについても分布が重なっていた. これは軸が表す意味を無視して分離を行なっているのが原因と考えられ, 文章内の固有名詞と文章に含まれない固有名詞の分布を分離するような特徴を持つ軸を切り出す線形判別分析を行うと 2 つの分布を分離することができた.

連絡先:竹川 高志, 工学院大学 情報学部, 〒163-8677 東京都新宿区西新宿 1-24-2 , 03-3340-0103, E-mail: takekawa@cc.kogakuin.ac.jp

2. 実験方法

本研究では, Word2vec を用いて学習を行った Web 上で公開されている分散表現のモデル[5]用い, モデルの学習に用いられた Wikipedia の日本語記事を解析対象の文章に用いる. また, 解析で用いる文章の単位は段落もしくは記事であり, 単語の分割, 品詞の特定についてはシステム辞書に mecab-ipadic-NEologd を用いた MeCab を用いた.

2.1 ユークリッド距離を用いた誤字の検出

Wikipedia の日本語記事から段落ごとに文章を取り出し, その中から固有名詞 1 つを段落に含まれていない固有名詞に交換された文章を解析対象とする. 文章内の固有名詞のベクトルの平均との差が最も大きかった固有名詞が交換された固有名詞であるか調べ, 段落内の固有名詞の数ごとにその正答率を計算した.

2.2 ユークリッド距離と主成分分析を用いた単語の分布

Wikipedia の日本語記事から記事ごとに文章を取り出し, 記事内の固有名詞と記事で用いられていない固有名詞の分布をグラフに表し確認する. 分布をグラフに表す方法は, 記事内の固有名詞のベクトルの平均との距離を記事内の固有名詞と記事に用いられていない固有名詞の 2 群つについてヒストグラムに表す方法, 記事内の固有名詞のベクトルについて 2 次元へ圧縮する主成分分析を行い, 得られた変換行列を用いて記事内の固有名詞と記事に用いられていない固有名詞のベクトルに変換行列をかけ, 散布図にプロットする方法である.

2.3 線形判別を用いた分離

Wikipedia の日本語記事から記事ごとに文章を取り出し, 記事ごとに記事内の固有名詞と記事外の固有名詞を分離するような教師データとして線形判別分析を行い分離度を検証した.

3. 結果

3.1 ユークリッド距離を用いた誤字の検出

図 1 は段落内の固有名詞の数ごとに文章に含まれていなかった単語の正解率を出したグラフであり, エラーバーは誤差分散を表している. 文章内の交換されていない固有名詞の数が 2 個のときの正解率がおよそ 0.7 となり. 単語数が多くなるにつれ

て正解率は低下し，交換されていない固有名詞の数が 30 個の時には正解率はおよそ 0.5 となった．

3.2 ユークリッド距離と主成分分析を用いた単語の分布

図 2 は Wikipedia の“日本語”という記事に対して記事内の固有名詞と記事に含まれていない固有名詞について分離を行ったグラフである．上から順に記事内の固有名詞のベクトルの平均との距離を用いた分離方法，主成分分析を用いた分離方法のグラフである．ユークリッド距離を用いて分離を表したグラフは記事内の単語のヒストグラムと記事に含まれていない単語のヒストグラムの面積が等しくなる様にグラフの設定をおこなっているため，実際には記事内の単語のヒストグラムは小さくなる．二つのグラフとも文章内の固有名詞の分布と文章に含まれていない固有名詞の分布が重なっていることが分かる．

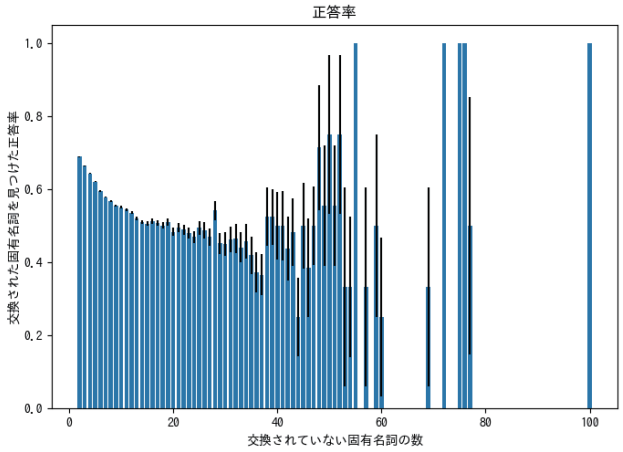

図 1. ユークリッド距離を用いて文章に含まれていなかった固有名詞を検出できた正答率

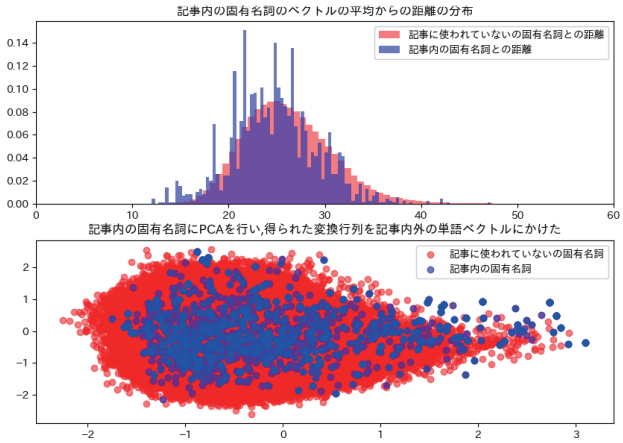

図 2. “日本語”記事に対しユークリッド距離を用いた分離，主成分分析を用いた分離を行なった分布

3.3 線形判別分析を用いた分離

図 3 は Wikipedia の“日本語”という記事に対して，図 4 は“石川県”という記事に対して記事に含まれる固有名詞と含まれていない固有名詞を教師データとして線形判別分析を行った結果である．この二つのグラフについても 2 つの分布の面積が等しくなる様にグラフの設定を行っているため，実際には記事に含まれる固有名詞のヒストグラムは小さくなる．ユークリッド距離を用いた分離や主成分分析を用いた分離より 2 つの分布の重なりが少ない事がわかる．

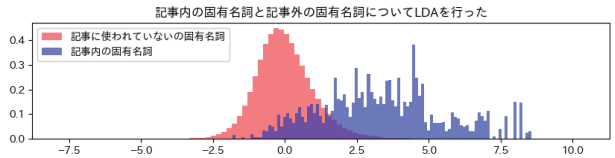

図 3. “日本語”の記事に対して線形判別を行った．

4. まとめ

ユークリッド距離を用いた誤字の検出では期待値よりは大きな正解率となったが，文章校正を行う上では不十分であった．そこで実際にベクトル同士の距離についての分布と主成分分析を行った分布について調査したところ，分離を行いたい 2 つの分布が重なっている事がわかった．これは Word2vec のモデルが 100 次元と高次元になっているのにも関わらず，距離を用いた分離では軸ごとの重みを考慮していないこから，主成分分析についても軸ごとの分布の意味を考慮していない事から起因するものだと考えられる．そこで 2 つの分布が分離する様な軸で次元を圧縮するために線形判別分析を行うと上の二つの手法よりよく分離している事がわかった．

また，図 5 は“日本語”の記事に対して線形判別分析を行い，得られた軸上に存在する単語を出力したグラフである．グラフに記載されている単語は横軸は線形判別分析により得られた軸であるが，縦軸に意味はない．このグラフにプロットされている単語は横軸で正の方向に進むほど文章の内容をよく表している単語と考えられられ，単語と文章の内容の関連度を数値化する事が期待できると考えている．

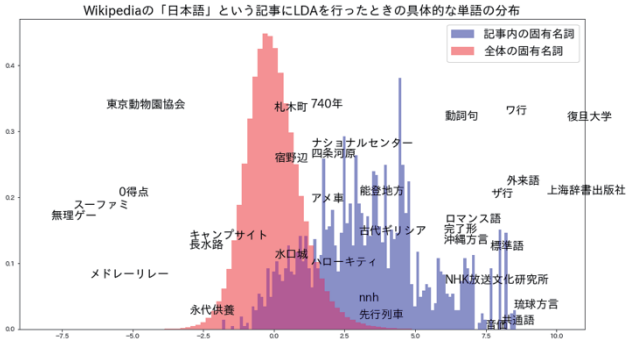

図 5. “日本語”の記事に対して線形判別分析を行った軸に対して単語をプロットした．

参考文献

[1] Zellig S. Harris: Distributional structure, WORD, 10:2-3 .pp.146-162, 1954.

[2] Tomas Mikolov, Ilya Sutskever, Kai Chen, Greg Corrado and Jeffrey Dean. Distributed Repersentations of Words and Phrases and their Compositionality. Advances in Neural Information Processing Systems 26. pp.3111-3119. 2013.

[3] Tomas Mikolov, Kai Chen, Greg Corrado and Jeffrey Dean. Efficient Estimation of Word Representations in Vector Space. In Proceedings of Workshop at International Conference on Learning Representations, 2013.

[4] 池原 悟, 小原 永, 高木 伸一郎. 文書校正支援システムにおける自然言語処理. 情報処理. 34 巻. 10 号. pp.1249-1258. 1993.

[5] 鈴木 正敏, 松田 耕史, 関根 聡, 岡崎 直観, 乾 健太郎. Wikipedia 記事に対する拡張固有表現ラベルの多重付与，言語処理学会第 22 回年次大会. pp.A5-2. 2016.