

# 情報工学実験第 6 回レポート

平 智隆

November 14, 2022

## 1 課題 20221101-1

与えられた文章内の単語の分散表現を Word2Vec によって得た。

### 1.1 使用パラメータ

使用したパラメータを以下の Table 1 に示す。

使用パラメータ	値
vector_size	3
window	5
min_count	1
sg	0, 1

Table 1: 使用パラメータ

### 1.2 文章の前処理

文章内の半角記号を全て消去した。

### 1.3 最も類似している単語

- nothing (名詞) → nice (類似度: 0.9908)
- get (動詞) → wondered (類似度: 0.9982)
- tired (形容詞) → white (類似度: 0.9959)
- presently (副詞) → sticks (類似度: 0.9970)

### 1.4 単語の分散表現ベクトルの加減算

mouse + alice - knowledge の計算をした。類似度が高い順に、across (類似度: 0.9941), breve (類似度: 0.9907), about (類似度: 0.9896)... となった。

### 1.5 学習アルゴリズムによる違い

先程までは学習アルゴリズムとして CBoW を用いていたが、Skip-gram に変更して、これまでと同じ処理をした。

#### 1.5.1 最も類似している単語

- nothing (名詞) → showing (類似度: 0.9979)
- get (動詞) → pink (類似度: 0.9912)

- tired (形容詞) → make (類似度: 0.9850)
- presently (副詞) → name (類似度: 0.9962)

### 1.5.2 単語の分散表現ベクトルの加減算

mouse + alice - knowledge の計算をした。類似度が高い順に、across (類似度: 0.9919), how (類似度: 0.9847), about (類似度: 0.9837)... となった。

## 2 課題 20221101-2

日本語 Word2Vec(chiVe) を用いて、課題 1 と同様に、ある単語について最も類似度が高い単語を調べた。

### 2.1 最も類似度が高い単語

- 喉仏 (名詞) → 鎖骨 (類似度: 0.6448)
- とんずら (名詞) → 持ち逃げ (類似度: 0.5507)
- 逆らう (動詞) → 刃向かう (類似度: 0.6702)
- 騒々しい (形容詞) → 騒がしい (類似度: 0.8528)
- さっさと (副詞) → とっとと (類似度: 0.9674)

### 2.2 考察

課題 1 の結果と比べて、課題 2 では、単語の類似度をより正確に計算できていることがわかる。例えば課題 1 の "nothing" の最類似単語に、"nice" が選ばれているが、これは単に綴りが似ているだけであって、意味的には全く類似しているとは言えない。一方で、課題 2 の方では、例えば、「逆らう」という単語の最類似単語として、「刃向かう」という単語が選ばれている。これは意味的にほとんど同じ単語である。この精度の差の原因は、学習データ量の違いにあると考える。