B3 実験 11 月 8 日分課題

1 課題 1

1.1 実験内容

gensim の Word2Vec ライブラリを使い, 与えられたテキストを学習した. CBoW と Skip-gram の 2 つの学習アルゴリズムを用いて選択した単語の類似度, 分散表現ベクトルの結果を調べた.

1.2 前処理 と モデルのパラメータ

与えられたテキストにおいて .！？’ のみ残し, その他の半角記号を消去した. その後 .！？ の 3 つの記号に関しては前後に空白を挟むように調整し, 文章中の全単語を小文字にした. そして, Natural Language Toolkit (nltk)[1] を用いて分かち書きした. その後 .！？ の記号を文末と定義して 1 文ごとに分けて Word2Vec のモデルの学習を行えるようにした. また, 表 1 に今回作成した Word2Vec モデルのパラメータを示す

表 1: Word2Vec におけるパラメータ

パラメータ	分散表現の次元	学習時に利用される文脈の広さ	分散表現を獲得する単語の最小頻度
値	500	5	1

1.3 実験結果

Word2Vec においての学習アルゴリズムを CBoW , skip-gram に設定したときそれぞれの実験結果を示す.

1.3.1 単語の類似度

適当な名詞, 動詞, 形容詞を 1 つ選択しその単語と近い類似度の単語を 5 個調べた. 今回は名詞として ’alice’ , 動詞として ’think’ , 形容詞として ’good’ を選択した. 表 2 , 3 , 4 にそれぞれの結果を示す.

表 2: ’alice’ に最も類似度が高い上位 5 単語

CBoW	単語	’trouble’	’wish’	’what’	’little’	’then’
	類似度	0.130898	0.124550	0.104635	0.094204	0.092406
skip-gram	単語	’to’	’the’	’she’	’of’	’a’
	類似度	0.384755	0.380266	0.348892	0.348659	0.341373

表 3: ’think’ に最も類似度が高い上位 5 単語

CBoW	単語	’i’	’wish’	’quite’	’noticed’	’was’
	類似度	0.107442	0.105029	0.091723	0.090705	0.088661
skip-gram	単語	’i’	’to’	’very’	’in’	’a’
	類似度	0.334731	0.312466	0.312177	0.301978	0.286862

表 4: 'good' に最も類似度が高い上位 5 単語

CBoW	単語	'near'	'feel'	'herself'	'bat'	'dream'
	類似度	0.111693	0.108123	0.099820	0.096781	0.094528
skip-gram	単語	'to'	'of'	'that'	'with'	'very'
	類似度	0.275121	0.274885	0.244580	0.228482	0.213307

1.4 単語ベクトルの減加算

今回の実験では, 'she' + 'her' - 'sister' - 'dianh' - 'rabbit' の意味に近い上位 5 単語を調べた. 表 5 に結果を示す.

表 5: she' + 'her' - 'sister' - 'dianh' - 'rabbit' に最も類似度が高い上位 5 単語

CBoW	単語	'that'	'dreamy'	'sometimes'	'.'	'people'
	類似度	0.138724	0.129123	0.117495	0.113983	0.112256
skip-gram	単語	'that'	'it'	'.'	'dreamy'	'a'
	類似度	0.175264	0.156423	0.151874	0.142241	0.142241

2 課題 2

Word2Vec の学習済み日本語モデルとして chiVe を利用して, 課題 1 と同様に単語の情報を調べた.

2.1 実験結果

名詞として '進捗' , 動詞として, '生む' , 形容詞として ' 辛い' を選択した. 表 6 , 7 , 8 に結果を示す.

表 6: '進捗' に最も類似度が高い上位 5 単語

単語	'進捗状況'	'進行状況'	'進み具合'	'進捗度'	'進行具合'
類似度	0.876365	0.750939	0.656917	0.634026	0.587030

表 7: '生む' に最も類似度が高い上位 5 単語

単語	'生む'	'生み落とす'	'生み出す'	'生み育てる'	'生まれ出る'
類似度	0.771237	0.670780	0.644634	0.620251	0.548058

表 8: ' 辛い' に最も類似度が高い上位 5 単語

単語	'きつい'	'しんどい'	'苦しい'	'難い'	'けれど'
類似度	0.697457	0.683242	0.629626	0.609669	0.597894

2.2 考察

chiVe に関して名詞と動詞に関しては単語自体に似ている単語が類似度が高いと判断された. しかし,' 辛い' に関しては 5 番目に類似度が高い単語として ' けれど' が選ばれた. このことから Word2Vec は文脈情報も考慮して学習していることが分かる. 課題 1 に関しては選択した単語に類似している単語が選ばれているとは言い難い結果となった. これは学習に用いたテキストデータの量が少ないためと考えられる.

参考文献

[1] NLTK Project. Documentation, 2022. https://www.nltk.org/.