情報工学実験 II 11/8 課題

1 課題 1

gensim の Word2Vec を用いて与えられたテキストファイルから学習を行い、学習したモデルに対していくつか単語与え学習結果を確認した.

1.1 学習条件

- 学習用テキストファイルの 1 行を 1 文として扱った.
- 前処理として文中の記号を除去した.
  記号として扱ったもの ., ”)(\!?:;-_
- vector_size を 10 に, min_count を 1 にとした.
- 上記以外のパラメータはデフォルト値を用いた.

1.2 結果

実験結果を次の表 1 から表 5 にまとめた.

表 1: ”alice” (名詞) と類似度が高い単語

アルゴリズム	CBoW	類似度が高い単語	sound	pegs	right	itself	about
		類似度	0.8146	0.6798	0.6726	0.6613	0.6493
	Skip-gram	類似度が高い単語	sound	pegs	itself	right	about
		類似度	0.8305	0.7223	0.7145	0.7125	0.7007

表 2: ”sister” (名詞) と類似度が高い単語

アルゴリズム	CBoW	類似度が高い単語	opportunity	after	latitude	first	herself
		類似度	0.7609	0.7209	0.7127	0.7094	0.6840
	Skip-gram	類似度が高い単語	opportunity	first	latitude	feel	near
		類似度	0.7613	0.6323	0.5862	0.5662	0.5579

表 3: ”remarkable” (形容詞) と類似度が高い単語

アルゴリズム	CBoW	類似度が高い単語	pleasure	watch	just	so	close
		類似度	0.8077	0.7770	0.7019	0.6988	0.6705
	Skip-gram	類似度が高い単語	pleasure	watch	just	so	close
		類似度	0.8217	0.8055	0.7200	0.7182	0.6964

表 4: ”say” (動詞) と類似度が高い単語

アルゴリズム	CBoW	類似度が高い単語	word	at	sleepy	my	without
		類似度	0.8088	0.7690	0.7350	0.7070	0.6750
	Skip-gram	類似度が高い単語	word	at	sleepy	could	fell
		類似度	0.8617	0.8380	0.7890	0.7685	0.7514

表 5: ”me+she-i” と類似度が高い単語

アルゴリズム	CBoW	類似度が高い単語	burning	look	her	eat	or
		類似度	0.7856	0.6811	0.6683	0.6465	0.6251
	Skip-gram	類似度が高い単語	her	should	eat	look	would
		類似度	0.8319	0.8297	0.8166	0.8166	0.8106

1.3 追加課題

Wikipedia2Vec のモデルを用いて同じ単語に対して類似度が高い単語を調査し, 結果を表 6 から表 10 にまとめた。

表 6: Wikipedia2Vec での ”alice” (名詞) と類似度が高い単語

類似度が高い単語	katherine	lucy	jane	emily	dorothy
類似度	0.8333	0.8299	0.8296	0.8290	0.8275

表 7: Wikipedia2Vec での ”sister” (名詞) と類似度が高い単語

類似度が高い単語	daughter	sibling	niece	stepsister	mother
類似度	0.7741	0.7597	0.7591	0.7552	0.7531

表 8: Wikipedia2Vec での ”remarkable” (形容詞) と類似度が高い単語

類似度が高い単語	unparalleled	impressive	surprising	astonishing	exceptional
類似度	0.8405	0.8361	0.8041	0.7998	0.7963

表 9: Wikipedia2Vec での ”say” (動詞) と類似度が高い単語

類似度が高い単語	know	outdance	believe	ENTITY/Wikipedia:FA	why
類似度	0.8962	0.8526	0.8490	0.8383	0.8325

表 10: Wikipedia2Vec での ”me+she-i” と類似度が高い単語

類似度が高い単語	herself	her	sally	rosie	rachel
類似度	0.7376	0.7104	0.6753	0.6692	0.6680

1.4 考察

与えられたテキストファイルで学習したモデルは CBoW と Skip-gram のどちらにおいても似たような結果になり, 両者とも似ている単語を適切に学習できているとは言えないように思える. これは学習データの量が不十分か

らだと考えられる．実際, Wikipedia2Vec に対する結果では関連する単語を示している．しかし, 動詞においては Wikipedia2Vec においても意味が類似する単語を適切に示せておらず用法が似ている単語を示している．また, 今回の結果では対義語や多義語による影響を見つけることはできなかった．単語ベクトルの演算に関しては Wikipedia2Vec では適切な結果が得られ, CBoW より Skip-gram の方が結果を残した．

2 課題 2

chiVe のモデルを用いていくつか単語を入力しその結果を確認した．

2.1 結果

結果を次の表 11 から表 15 にまとめた．

表 11: ”大阪” (名詞) と類似度が高い単語

類似度が高い単語	名古屋	神戸	阪	大阪市	関西
類似度	0.7537	0.7431	0.7317	0.7250	0.7056

表 12: ”妹” (名詞) と類似度が高い単語

類似度が高い単語	姉	弟	従妹	従姉	義妹
類似度	0.8511	0.8124	0.8106	0.8024	0.7994

表 13: ”早い” (形容詞) と類似度が高い単語

類似度が高い単語	遅い	早め	早まる	早める	早く
類似度	0.7753	0.6877	0.6807	0.6772	0.6327

表 14: ”言う” (動詞) と類似度が高い単語

類似度が高い単語	と	って	は	有る	思う
類似度	0.8314	0.8116	0.7894	0.7830	0.7821

表 15: ”神戸+宮城-関西” と類似度が高い単語

類似度が高い単語	仙台	宮城県	岩手	仙台市	気仙沼
類似度	0.6034	0.5843	0.5607	0.5605	0.5588

2.2 考察

名詞, 形容詞, 動詞, 単語ベクトルの演算それぞれの結果は課題 1 おおよそ同じ傾向を示した．動詞においては意味ではなく文脈上つながる単語が高い類似度を示し適切に働いていないと考えられる．単語ベクトルの演算の結果では東北地方の都市を示しており単語の意味的に適切な結果を得られていることが確認できる．ただし, 課題 2 では課題 1 と異なり ”早い” に対して ”遅い” という対義語が高い類似度を示すことが確認できる．これらのことから Word2Vec では意味が近い単語ではなく文脈上で似た役割や近い位置にある単語の類似度が高くなるようになっていることが確認できる．