パラグラフベクトル

文ベクトルの重み付け平均の式

$$L = \frac{1}{T}\sum_{t=k}^{T} \log p(w_t | w_{t-k}, ..., w_{t-1}),$$

$$p(w_t | w_{t-k}, ..., w_{t-1}) = \frac{e^{y_{w_t}}}{\sum_i e^{y_i}},$$

$$y = b + Uh(w_{t-k}, ..., w_{t-1}, d; W, D)$$

- d：文章
- w_i：単語
- W：全ての単語の分散表現を表す行列
- D：全ての文章の分散表現を表す行列
- k：ウィンドウサイズ
- T：現在の文章に含まれる単語数
- ウィンドウ：ある単語の周辺を表す区間
- p：softmax 関数により正規化された，文脈から現在の単語が導かれることの尤度
- $h(w_{t-k}, ..., w_{t-1}, d; W, D)$：引数となるベクトルを結合したベクトルを返す関数

文間・カテゴリ間の関係

文間の関係

「とても良かった」の文が

- 食事に関する文の直後に存在
  ⇒ 食事◎
- 部屋に関する文の直後に存在
  ⇒ 部屋◎

食事に関する文

とても良かった。

……

部屋に関する文

とても良かった。

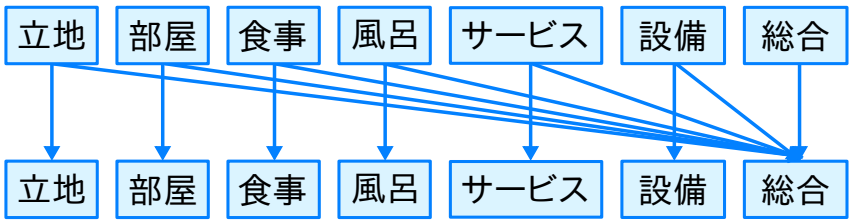
カテゴリ間の関係

- 他のカテゴリ◎
  ⇒ 「総合」カテゴリ◎

関連研究

隠れ状態を用いたホテルレビューのレーティング予測[1]（従来手法）

- 文毎のレーティングからレビュー全体のレーティングを予測
- カテゴリ間の繋がりを手調整で変化させて考慮



パラグラフベクトル[2]

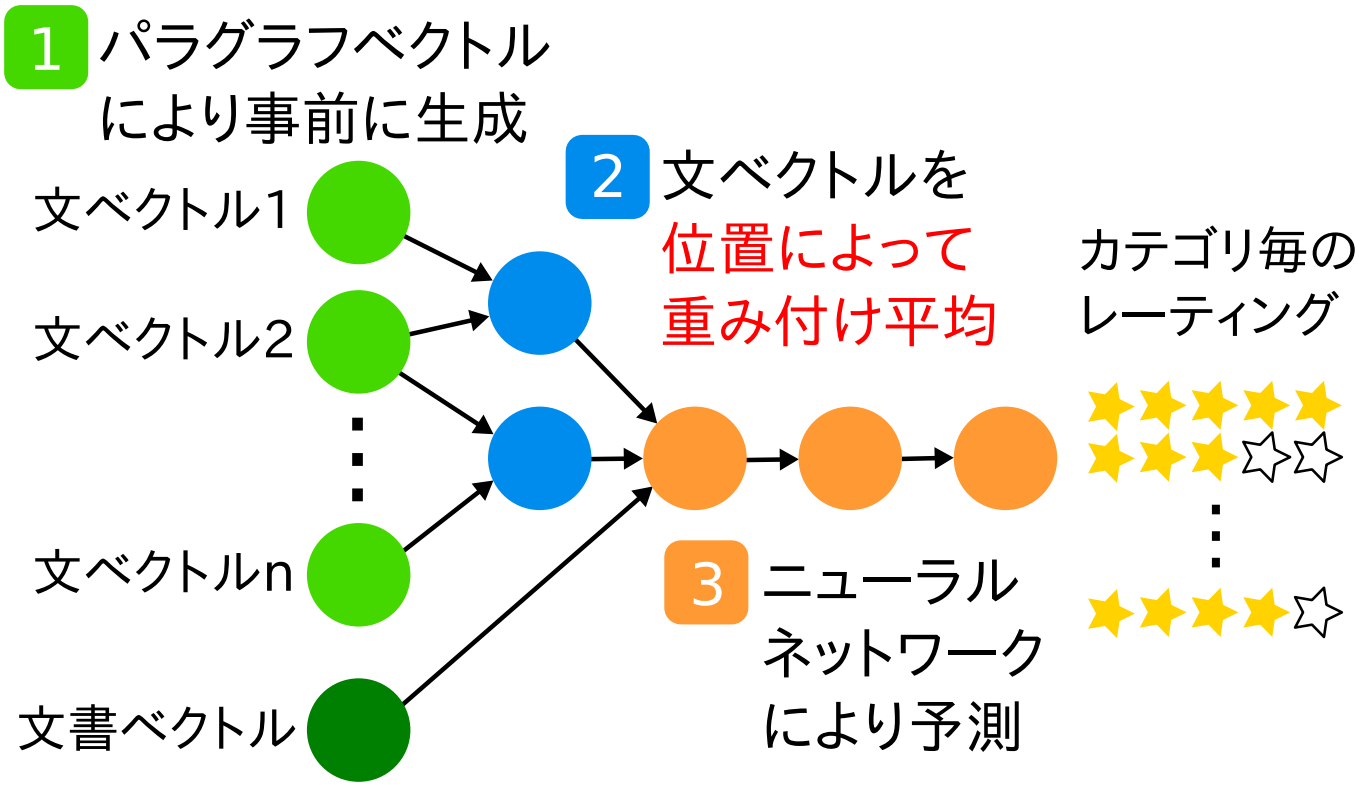
- 文や文書を実数ベクトルに変換する手法
- レーティング予測において優れた性能

[1] 藤谷宣典ら, 隠れ状態を用いたホテルレビューのレーティング予測. 言語処理学会第 21 回年次大会, 2015.

[2] Quoc Le et al., Distributed representations of sentences and documents. ICML 2014, 2014.

提案手法

- 文書・文間及びカテゴリ間の関係を自動で考慮したレーティング予測
- パラグラフベクトルと入出力間の複雑な関係を考慮できるニューラルネットワークを利用



実験

実験設定

- 7 カテゴリにおける 0〜5 点のレーティング予測の正答率を測定
- データセット ：楽天トラベルにおけるレビュー約 330,000 件

結果

- 提案手法が従来手法より高い正答率を示した

手法	正答率 [%]
従来手法	48.32
提案手法	**50.30**