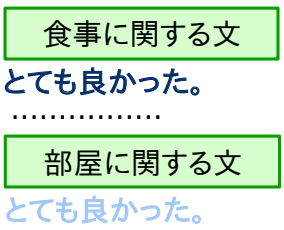
文書・文間及びカテゴリ間の関係を考慮したレーティング予測

知能数理研究室　12056 外山 洋太

1. 背景と目的

- 対象問題：多カテゴリにおける商品レビューのレーティング予測
- 目的：以下を考慮したレーティング予測の実現

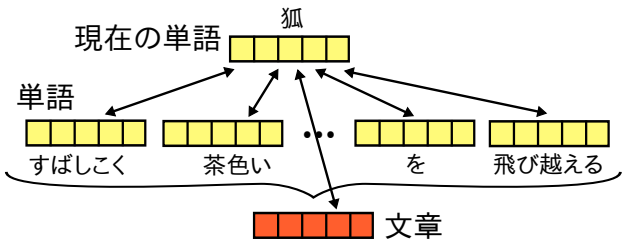


文章・文間の関係

カテゴリ間の関係

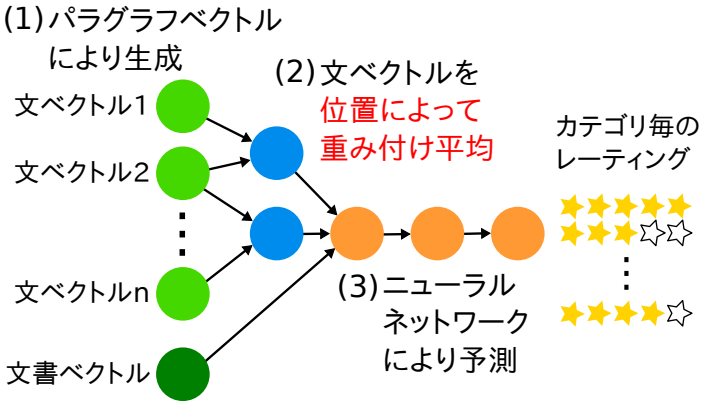
2. 関連研究

- 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]**
  - 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によって変化させその関係を考慮
- パラグラフベクトル **[2]**
  - 文や文書を，その意味を表す実数ベクトルに変換
  - レーティング予測において優れた性能



3. 提案手法

- 位置によって重み付け平均された文ベクトル
  → 文同士の位置関係を考慮
- ニューラルネットワークによる予測
  → 文書・文間及びカテゴリ間の関係を考慮



提案手法における予測モデル

- 重み付け平均された文ベクトル：$\mathbf{t}_{i_{part}}$

$$\mathbf{t}_{i_{part}} = \sum_{i_{sent}} \frac{w(x_{i_{part}}(i_{sent}))}{|\sum_{i'_{sent}} w(x_{i_{part}}(i'_{sent}))|} \mathbf{s}_{i_{sent}},$$

$$x_{i_{part}}(i_{sent}) = \frac{i_{sent} - i_{part}}{\#partitions},$$

$$w(x) = \begin{cases} \frac{1}{2}(\cos(\pi|x|)+1) & \text{if } |x| <= 1 \\ 0 & \text{otherwise} \end{cases}$$

i_{sent}：レビュー内の文のインデックス
$\#partitions$：重み付け平均後の文ベクトルの数
i_{part}：重み付け平均後の文ベクトルのインデックス
$\mathbf{s}_{i_{sent}}$：文ベクトル

- ニューラルネットワークの目的関数：E

$$E = -\sum_{n=1}^{N}\sum_{c=1}^{C}\sum_{k=1}^{K} d_{nck} \log y_{ck}(x_n; w),$$

$$y_{ck}(x_n; w) = \frac{e^{u_{ck}(x_n;w)}}{\sum_{j=1}^{K} e^{u_{cj}(x_n;w)}}$$

w：ニューラルネットワークのパラメータ
N：ミニバッチサイズ
C：カテゴリの総数
K：クラスの総数

4. 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 ：楽天トラベルのレビュー約 330,000 件
  - 分類器の入力が異なる 3 つの比較手法
    (1) Document Vector (DV)：
    レビュー全体の文書ベクトル
    (2) Averaged Sentence Vector (ASV)：
    平均した文ベクトル
    (3) Weighted ASV：重み付け平均した文ベクトル
- 結果
  - 提案手法が従来手法より高い正答率を示す
  - 文の並びが予測のために重要
  - 文書ベクトルと文ベクトルを同時に素性として用いることが有効

手法	正答率
従来手法 [1]	0.4832
DV	0.4980
ASV	0.4838
Weighted ASV	0.4867
提案手法	**0.5030**

5. まとめ

- 多カテゴリにおけるレーティング予測について，レビュー全体の文書ベクトルに加え重み付け平均された文ベクトルを用いた手法を提案
- 提案手法が従来手法 [1] より高い正答率を示した
- 今後の課題
  文間，単語間，文字間等のより多様で複雑な関係を考慮
  → レビューの特徴の抽出と分類のモデルを統合

参考文献

[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.