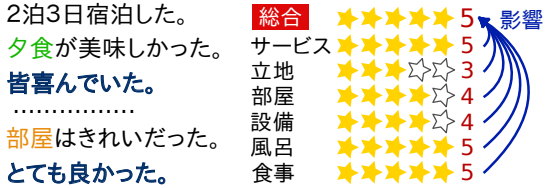
文書・文間及びカテゴリ間の関係を考慮したレーティング予測

豊田工業大学 知能数理研究室　外山洋太 三輪誠 佐々木裕

背景と目的

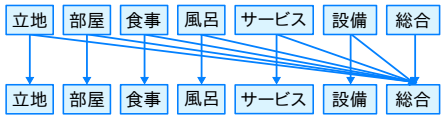
- 対象問題：多カテゴリにおける商品レビューのレーティング予測（多ラベル多クラス問題）
- 応用例：企業における文書からの商品の評判分析
- 目的：文書・文間の関係及びカテゴリ間の関係を考慮したレーティング予測の実現



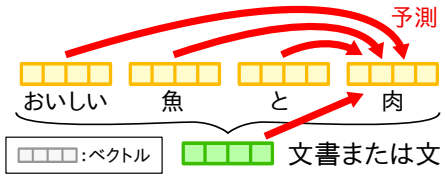
商品レビューの例

関連研究

- 隠れ状態を用いたレーティング予測 **[1]**
  - 文毎に隠れ状態を予測する Multiple-Instance Multiple-Label の手法を利用
  - 文毎の隠れ状態はカテゴリ別のレーティング
  - カテゴリ間の繋がりは手調整によって変化させ考慮



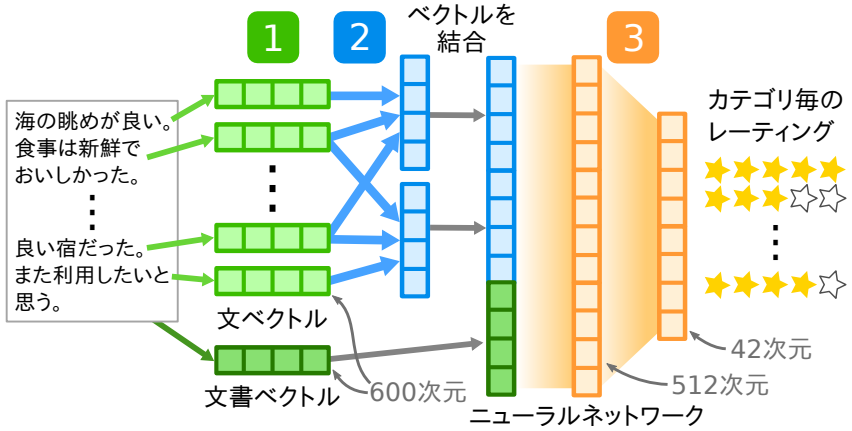
- パラグラフベクトル **[2]**
  - 文や文書を，その意味を表す実数ベクトルに変換
  - レーティング予測において優れた性能
  - 文書または文と周りの単語から現在の単語を予測するようにそれらのベクトルを学習



パラグラフベクトルの学習

提案手法

- 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

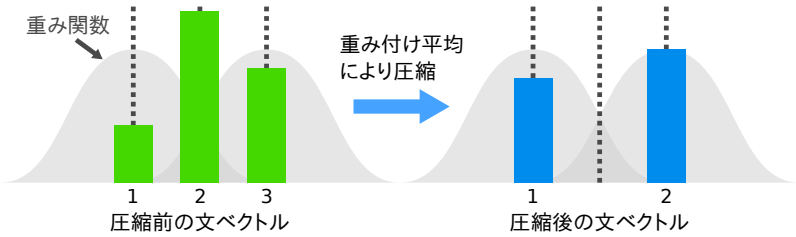


提案手法における予測モデル

1 パラグラフベクトルによる文書･文ベクトルの生成

2 重み付け平均による文ベクトルの圧縮

- 文の位置関係を考慮しつつ文ベクトルの数を **S 個 (2 個)** に固定



3 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複雑な関係を考慮
- 目的関数 E：カテゴリ毎に誤差を計算

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

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

u_{ck}：出力層のユニット
w：パラメータ
d_{nck}：文書 n がカテゴリ c でクラス k ならば 1，それ以外で 0
N：バッチサイズ（42 個）
C：カテゴリの総数（7 個）
K：クラスの総数（6 個）

実験

- 実験設定
  - データセット：楽天トラベルのホテルレビュー
    訓練：300,000 件，開発：10,000 件，評価：10,000 件
  - 7 カテゴリ 0〜5 点のレーティング予測の正答率を測定
  - 提案手法の分類器の入力を変更した 3 つの比較手法
    (1) Document Vector (DV)：文書ベクトル
    (2) Averaged Sentence Vector (ASV)：平均文ベクトル
    (3) Weighted ASV：重み付け平均した文ベクトル
  - 重み更新に Adam，L2 正則化，ドロップアウト（入力層：0.2，出力層：0.5）を利用

- 結果
  - 提案手法が従来手法より **0.020** 高い正答率を示した
  - 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用いることが有効

手法	正答率	RMSE	正答例 （風呂：3 点）
従来手法 [1]	0.483	0.81	… ただ洞窟風呂が人気でかなりの人が待っていましたが，… しかし，この 2 点以外は非常に素晴らしく，…
DV	0.498	0.74	
ASV	0.484	0.76	
Weighted ASV	0.487	0.76	
提案手法	**0.503**	**0.73**	

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 0.020 高い正答率を示した
- 今後の予定
  - 文間，単語間，文字間等のより多様な関係を考慮
  - レビューの文書について1 文字ずつ特徴を考慮したニューラルネットワークを利用
    → 文書・文ベクトルの生成と予測のモデルを統合

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.