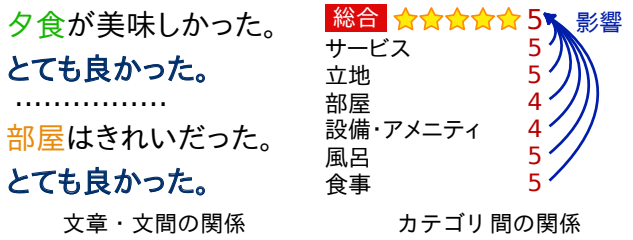
文書・文間及びカテゴリ間の関係を考慮したレーティング予測

知能数理研究室 12056 外山 洋太

背景と目的

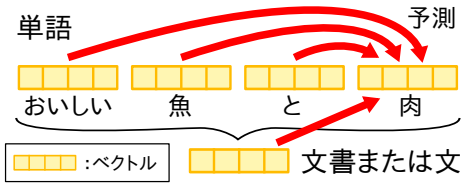
- 対象問題：多カテゴリにおける商品レビューのレーティング予測
- 応用例：企業における文書からの商品の評判分析
- 目的：文書・文間の関係及びカテゴリ間の関係を考慮したレーティング予測の実現



文章・文間の関係　　カテゴリ間の関係

関連研究

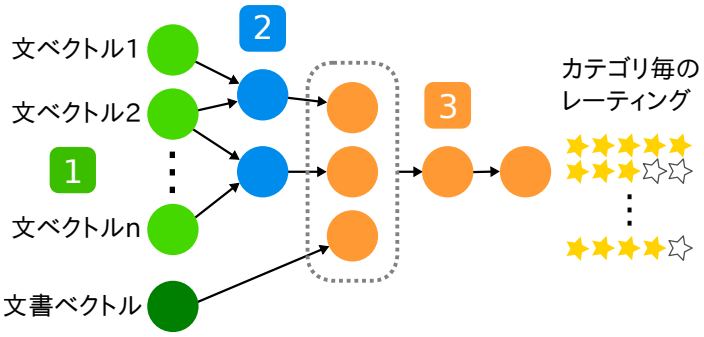
- 隠れ状態を用いたホテルレビューのレーティング予測 [1]
  - 文毎のレーティングからレビュー全体のレーティングを予測
  - カテゴリ間の繋がりを手調整によって変化させ考慮
- パラグラフベクトル [2]
  - 文や文書を，その意味を表す実数ベクトルに変換
  - レーティング予測において優れた性能
  - 文書または文と周りの単語から現在の単語を予測するようにそれらのベクトルを学習



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
  - 入力：訓練用レビューと正解レーティングの集合，及び，テスト用レビューの集合
  - 出力：各テスト用レビューのカテゴリ毎のレーティング
  - 訓練用レビューでニューラルネットワークを学習
    → テスト用レビューについてレーティングを予測



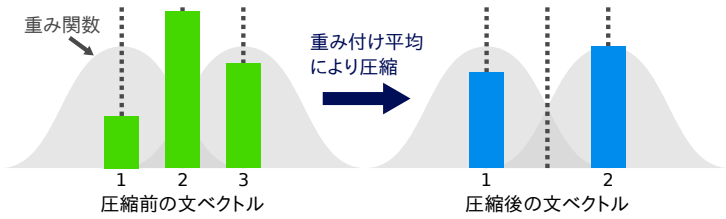
提案手法における予測モデル

1 パラグラフベクトルによる文書･文ベクトルの生成

- 文書・文の密なベクトル表現
- 訓練・テスト用レビュー全てについて予測の前に生成

2 文ベクトルの重み付け平均による圧縮

- 重み関数：*cos* 関数
- 文同士の位置関係を考慮しつつ文の数を統一



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：ミニバッチサイズ
C：カテゴリの総数
K：クラスの総数

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
  - 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)：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
従来手法 [1]	0.4832
DV	0.4980
ASV	0.4838
Weighted ASV	0.4867
提案手法	**0.5030**

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.