カテゴリ間の関連性を利用した多層ニューラルネットワークによる文書分類

外山洋太　三輪誠　佐々木裕　　豊田工業大学 工学部 先端工学基礎学科

背景と目的

- タスク
  文書を複数のカテゴリについて多値分類
  - カテゴリ：ラベルが付けられる各項目
  - 従来の手法では...
    - カテゴリ同士の関連性を手動で変化させ考慮
    - 文書の数値表現である BoW は文書内の語順を無視
- 目的
  - 多層ニューラルネットワークによりカテゴリ間の関連性を自動的に考慮
  - パラグラフベクトルの使用により語順や単語の位置関係を考慮

例. カテゴリ毎のラベルが付いた文書（商品レビュー）

ユーザ：ytoyama
ホテルの雰囲気はとてもよく食事もおいしかったです。部屋についても、窓からの見晴らしがよく海がとても綺麗でした。チェックイン当日、入口のフロアの汚れが気になりましたが、翌日にはきちんと清掃されていました。機会があれば、また利用したいと思います。

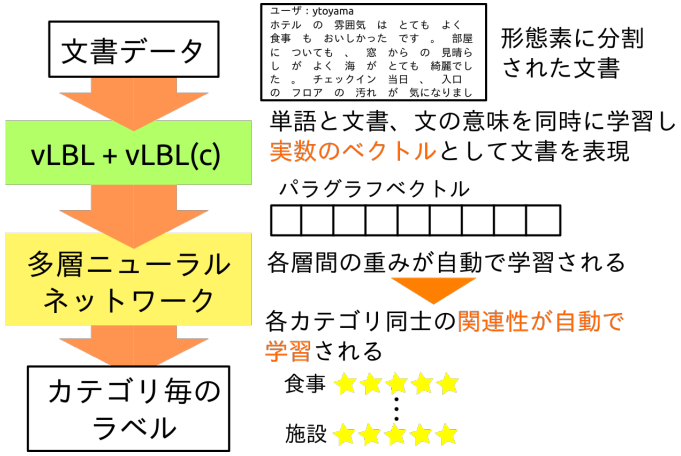
総合	★★★★☆ 4
サービス	3
立地	5
部屋	4
設備・アメニティ	4
風呂	3
食事	-

関連研究

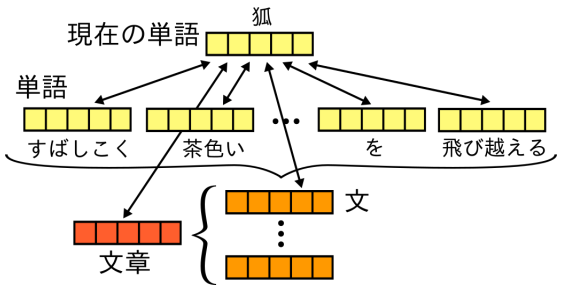
- 隠れ状態を用いたホテルレビューのレーティング予測 **[1]**
  文書内の各文に対して推定した隠れレーティングとレビュー全体のレーティングとの繋がりを手動で変化させる
  → カテゴリ間の関連性を考慮
- パラグラフベクトル **[2]**
  語順を考慮した文書の数値表現。文書分類に有用であることが実験により示されている。
- **vLBL+vLBL(c)[3]**
  単語同士の位置関係を考慮した単語ベクトルの学習手法。パラグラフベクトルまたは文ベクトルも同時に学習可能。

提案手法

- パラグラフベクトルに加え文ベクトルを導入した **vLBL+vLBL(c)** を提案語順と単語同士の位置関係を考慮した文書の数値表現を生成
  → 文書の意味をより正確に表現
- 分類器としての多層ニューラルネットワーク
  → カテゴリ間の関連性を自動的に考慮



文書データからのカテゴリ毎のラベル推定



vLBL+vLBL(c) による単語及び文、文章ベクトルの学習

$$g = \sum_t \left\{ \log \sigma(s(t)) + \sum_{t' \sim P_n}^{K} \log(1 - \sigma(s(t'))) \right\}$$

$$s(t) = \mathbf{c}_t \cdot \mathbf{w}_t + \mathbf{c}_t^{loc} \cdot \mathbf{w}_t^{loc} + b_t$$

$$\left(\begin{array}{c} t：\text{現在の単語の位置} \\ c_t, w_t：\text{文脈、単語を表すベクトル} \\ s：\text{位置関係を考慮した単語と文脈の類似度} \\ \sigma：\text{シグモイド関数} \end{array} \right)$$

- 現在の単語 w_t → 文脈との意味を近く
- 文脈外の単語 $w_{t'}$ → 文脈との意味を遠く
- $\mathbf{c}_t^{loc} \cdot \mathbf{w}_t^{loc}$ の項により単語同士の位置関係を考慮

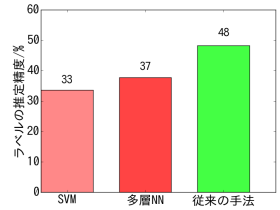
予備実験

vLBL+vLBL(c) と SVM または多層 NN を用い、従来の手法 [1] と同じ多値分類問題の精度を測定

- 目的
  - パラグラフベクトルの有効性の調査
  - 従来手法との比較による目標設定
- 実験設定
  - 宿泊予約サイト楽天トラベルのホテルレビューデータを利用
  - 入力データは、各レビューのコメント部分と 7 カテゴリのレーティングの組（各カテゴリのレーティングは評価なしを含む 6 段階評価）
  - 訓練データ：300,000 件、評価データ：10,000 件
  - 多層 NN の入力は位置を考慮した及び考慮していない 2 つのパラグラフベクトル
- 結果及び考察
  より表現力の高い文書の数値表現の評価や、多層 NN のパラメータ最適化が必要

プログラムのパラメータ設定

項目	値
学習する単語の範囲	前後 5 単語
単語の最少出現回数	5 回
ベクトルの次元数	400
中間層の数	1
入力層でのニューロン数	800 個
中間層でのニューロン数	200 個



各手法における点数推定精度

まとめと今後の課題

- パラグラフベクトルと多層ニューラルネットワークとを組み合わせただけでは精度が不十分
- 課題
  - vLBL+vLBL(c) における文ベクトルの評価
  - 多層ニューラルネットワークのパラメータ最適化
  - 提案手法の有用性の評価

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.

[3] 森洸樹ら, 英文穴埋め問題における文章ベクトルと学習データの質の影響. 第 222 回自然言語処理研究会, 2015.