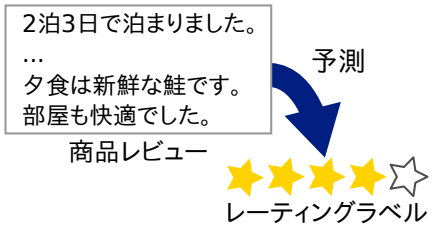
# レーティング予測によるフォントを基盤としたレビュー解析

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

- 対象タスク：表意・表語文字を含む言語におけるレーティング予測
- 応用例：企業における文書からの商品の評判分析
- 目的：文字の表層情報を利用したレーティング予測の実現




表語文字: 1文字で1語を表現

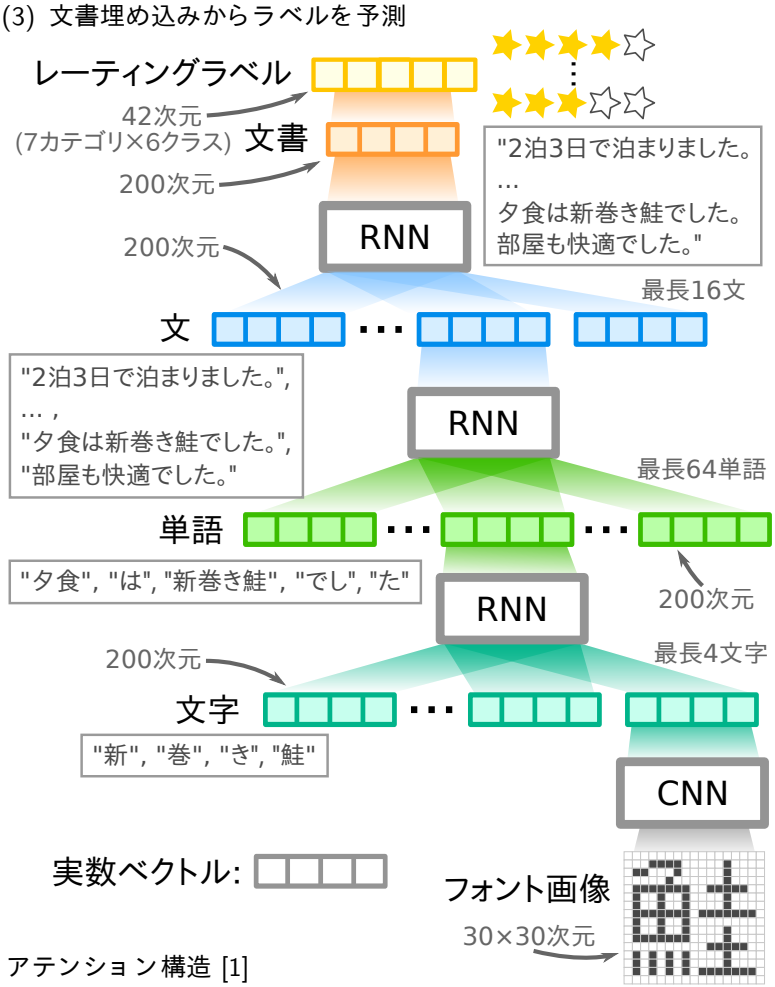
快 悦 悩　鯛 鱒 鯵

表意文字: 1文字で意味を表現

☺ ☹ ♥ ★ ♪

## 提案手法

- 入力：フォント画像で表現されたレビュー
- 出力：予測したレーティングラベル
- 特徴
  - フォント画像から文字の意味情報を抽出
  - HAN[1] の手法を基に文字からの文書の階層構造を利用
- 予測過程
  (1) 畳み込み NN (LeNet) によりフォント画像を文字埋め込みを生成
  (2) 文字から単語，単語から文，文から文書へ階層的に Gated Recurrent Unit (GRU) による RNN を用いて埋め込みを生成
  (3) 文書埋め込みからラベルを予測




- アテンション構造 [1]

$$u_i = \tanh(Wh_i + b)$$

$$\alpha_i = \frac{\exp(u_i^T u_{context})}{\sum_i \exp(u_i^T u_{context})}$$

$$\hat{h} = \sum_i \alpha_i h_i$$

- $u_{context}$：文脈ベクトル
- $\alpha_i$：アテンション
- $h_i$：下の階層の埋め込み
- $\hat{h}$：上の階層の埋め込み
- $W, b$：線形層のパラメータ

## 関連研究

- Hierarchical Attention Network (HAN) [1]
  - Attention 構造付きの Recurrent Neural Network (RNN) を用いた文書分類モデル
  - 文字または単語から文，文書までの階層構造を利用

→ 文字の表層情報の利用ができていない

- Radical-Enhanced Chinese Character Embedding [2]
  - 漢字-部首辞書を利用した漢字埋め込みの生成手法
  - 対象タスクと漢字の部首当てについて同時に学習

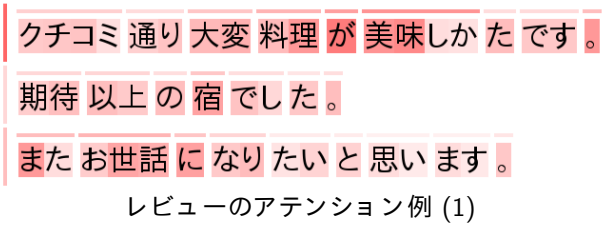
→ 漢字-部首辞書が余分に必要

## 実験

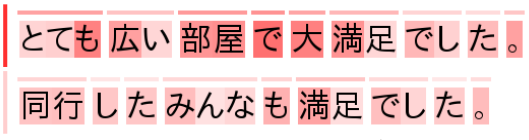
- 実験設定
  - 7 カテゴリにおける 0〜5 点のレーティング予測
  - データセット：楽天トラベルのレビュー 310,000 件 (訓練データ: 300,000 件，テストデータ： 10,000 件)
- 結果
  - 従来手法（パラグラフベクトルを応用した手法）より高い正答率

| 手法 | 正答率 |
|---|---|
| 従来手法 [3] | 0.503 |
| 提案手法 | **0.524** |

- 高いアテンションが付く表現
  - 「食」,「部屋」,「風呂」等のカテゴリを表すもの
  - 「広」,「満」,「良」,「悪」等の評価を表すもの
  - 「は」,「が」,「も」等の助詞




レビューのアテンション例 (1)




レビューのアテンション例 (2)

## まとめ

- フォント画像を用いたレーティング予測及びレビュー解析の手法を提案
- 提案手法による従来手法 [1] より高い正答率
- アテンションの可視化によるレビューの解析
- 今後の予定
  - フォント画像に対するアテンションの可視化の実装

参考文献

[1] Zichao Yang et al., Hierarchical Attention Networks for Document Classification. NAACL 2016, 2016.

[2] Yaming Sun et al., Radical-Enhanced Chinese Character Embedding. ICONIP 2014, 2014.

[3] 外山洋太ら，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 学士論文.