隠れ状態を用いたホテルレビューのレーティング予測

Prediction of Ratings for Hotel Reviews Using Hidden States

藤谷 宣典　　三輪 誠　　佐々木 裕
Yoshinori Fujitani　　Makoto Miwa　　Yutaka Sasaki
豊田工業大学
Toyota Technological Institute
{sd11082, makoto-miwa, yutaka.sasaki}@toyota-ti.ac.jp

1 背景と目的

Web2.0 時代において SNS (Social Networking Service)やニュース，商品，サービスに対するコメントなど，個人が情報を発信する機会が増加している．一方，様々な企業が個人の発信した情報を分析し消費者の理解を試みている．しかし情報過多により有益な評判情報を利用者自身で発見するには多大な時間を要する．この問題を解決する手段の一つとして，推薦システム[1][2]があげられる．推薦システムとは，ユーザの嗜好と商品やサービスの特徴を比較して，これらを基にユーザに最適な商品やサービスを推薦するシステムのことである．推薦システムを用いることで，ユーザに最適な情報を容易に探しだすことが可能となり，欲しい情報を発見するまでの時間を短縮することができる．ユーザに推薦する主な基準は「レーティングが高いと予想される」とされている．ここでのレーティングとは発信された情報と共に付属されている評価値のことである．推薦システムの推薦精度を向上する為にはレーティングを正確に予測できなければならない．

本研究の目的は，ユーザの記述したレビューを用い，文を隠れ状態としたモデルを構築することでレーティングを高精度に予測することである．構築するシステムは，ユーザレビューを入力として，いくつかの観点におけるレーティングを予測する．

2 Multi-Instance Multi-Label Learning for Relation Extraction (MIML-RE)

MIML-RE モデル[3]は関係抽出のためのモデルであり，文が複数のラベルを所有することを仮定としている．MIML-RE モデルの概略図を図 1 に示す．

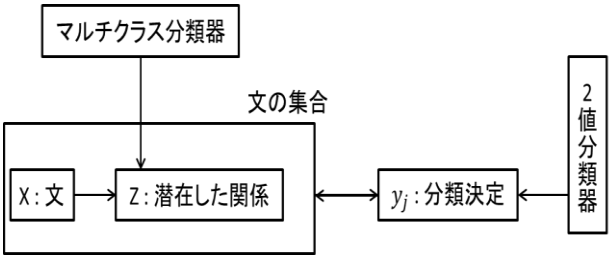

図 1　MIML-RE モデル概略図

全体の入力はエンティティのペアの集合で，出力は関係である．このモデルは，マルチクラス分類器と 2 値分類器と呼ばれる分類器を用いる．マルチクラス分類器は，関係ラベルのセットから個々のエンティティペアに潜在的な関係ラベルを割り当てる．また 2 値分類器は，入力としてエンティティペアレベルの分類を使用して，与えられたエンティティペアの集合ラベルかどうかを決定する．このモデルは文でのカテゴリラベルを隠れ状態 z として Expectation Maximization (EM) 法を用いて学習する．E step では，文レベルのマルチラベル分類器の重みベクトルW_z, 2 値分類器の重みベクトルW_yとして以下の式(1)により文レベルの分類z_iを推測する．

$$z_i^* = \arg\max_z p(z|y_i, x_i, W_y, W_z) \qquad (1)$$

しかし，上記の式(1)では計算量が多い為，以下の式で近似した式を扱う．

$$\begin{aligned} &\mathrm{p}\left(z_i^{(m)}|y_i, x_i, W_y, W_z\right) \\ &\propto \mathrm{p}(Y_i, z_i^{(m)}|x_i, W_y, W_z) \\ &\approx \mathrm{p}\left(z_i^{(m)}|x_i^{(m)}, W_z\right)\mathrm{p}\left(y_i|z_i', W_y\right) \\ &= \mathrm{p}\left(z_i^{(m)}|x_i^{(m)}, W_z\right) \prod_{r \in P_i \cup N_i} p\left(y_i^{(r)}|z_i', W_y^{(r)}\right) \end{aligned}$$

近似式より，文レベルの分類z_iは以下の式(2)で推測する．

$$z_i^{(m)*} = arg\max_z p\left(\mathrm{z}|x_i^{(m)}, W_z\right) \times \prod_{r\in P_i\cup N_i} p(y_i^{(r)}|z_i', W_y^{(r)}) \quad (2)$$

M step では，対数尤度の下界の最大化を行うことにより，分類器の重みW_z，W_yを探し出す．W_zとW_yを以下の式(3)と(4)でそれぞれ更新する．

$$W_z^* = arg\max_w \sum_{i=1}^{n}\sum_{m\in M_i} \log p\left(z_i^{(m)*}|x_i^{(m)}, W\right) \quad (3)$$

$$W_y^{(r)*} = arg\max_{\mathrm{w}} \sum_{1\le i\le n \text{ s.t. } r\in P_i\cup N_i} \log \mathrm{p}(y_i^{(r)}|\mathrm{z}_i^*, \mathrm{W}) \quad (4)$$

推論時には，初めにマルチクラス分類器にて下の式(5)で関係ラベルを獲得する．

$$z_i^{(m)*} = arg\max_z p(z|x_i^{(m)}, W_z^*) \quad (5)$$

その後，2 値分類器にて下の式(6)で最終の関係ラベルを決定する．

$$y_i^{(r)*} = arg\max_{y\in\{0,1\}} p(y|z_i^*, W_y^{(r)*}) \quad (6)$$

3　提案手法

本研究では，MIML-RE モデルを元に，レビュー単位にしか付与されていないレーティングが，レビューを構成する文がそれぞれ潜在的に持っているレーティングにより構成されると仮定したモデルを考える．図 2 に示すように，レビューを文単位に分割し，各文に隠れレーティングを与え EM 法により学習する．予測時には，文の隠れレーティングを予測し，それをレビュー単位で隠れレーティングを集め，レビューのレーティングを予測する．隠れレーティングの初期値は各レビューに付与されているレーティングをそのまま与える．図 2 に示すように隠れレーティングからレビューのレーティングを推測する時のカテゴリ間の繋がりは「総合」のみ全てのカテゴリと繋ぐ．素性には形態素，値には一文中にその形態素が出現した回数を用いる．上記の手法を基準手法とし，以下に示す方法で基準手法に変更点を加えた手法について提案する．

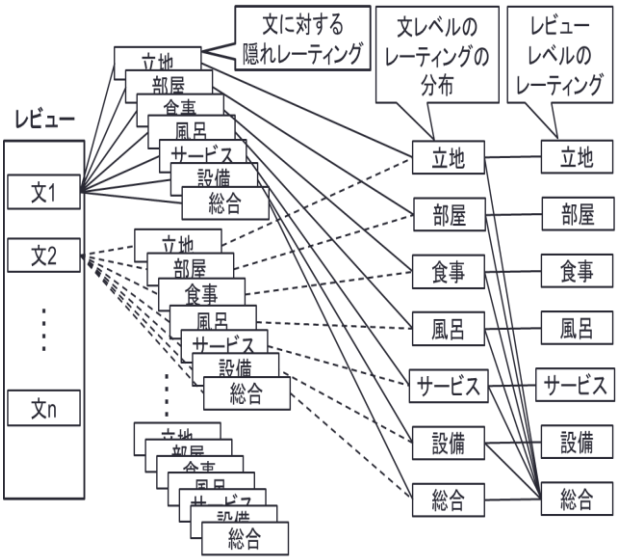

図 2　提案手法

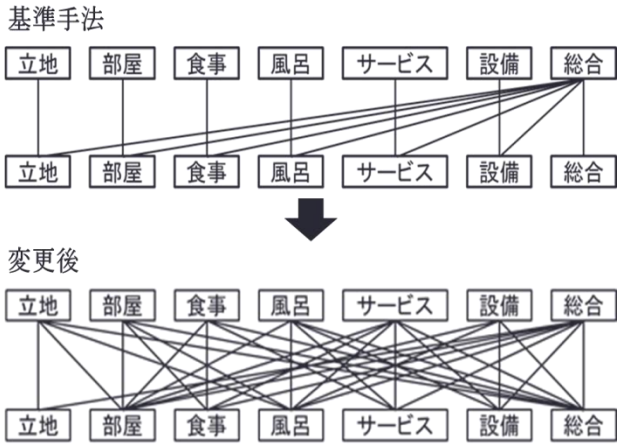

図 3　手調整の繋ぎ方の変更内容

繋ぎ方の変更：隠れレーティングからレビューのレーティングの推測時のカテゴリの繋がりを変更する．まず，ベースラインとして全てのカテゴリを全てのカテゴリと繋げる手法（**全対全**），一つのカテゴリをそのカテゴリのみと繋げる手法（**個対個**）を提案する．更に，関連性が高いと思われる繋がりを用いた手法（**手調整**）について提案する．手調整の接続内容を図 3 に示す．手調整の一例としては「食事」は，「部屋」や「食事」，「サービス」，「総合」との関係があると考えられる為，これらのカテゴリを図 3 の様に繋ぐ．

品詞素性の利用：データの素性に品詞を加える．1 文中に出現した回数を値とし，品詞を素性とする．

追加する素性は MeCab から品詞情報を得る．

N-gram 素性の利用：各データの素性を形態素と bi-gram と tri-gram で構成する．

ユーザ情報の利用：楽天トラベルレビューに記載されている情報の「ユーザ ID」，「同伴者」，「旅行の目的」を利用し，これを素性に加え各データを構成する．ユーザ情報から得た素性の値は 1 とする．また各データは，N-gram 素性の利用と同じく，形態素と bi-gram と tri-gram で構成する．

4　実験

4.1　実験設定

実験に使用した楽天トラベルレビューには，レビューとカテゴリ（「立地」，「部屋」，「食事」，「風呂」「サービス」，「設備」，「総合」）別のレーティングが記載されている．また，カテゴリのレーティングに対して無回答の場合は，レーティングに 0 点を付与した．各データは形態素解析システムの MeCab を用いて作成した．使用した楽天トラベルレビューの中には，【ご利用の宿泊プラン】という内容が多く記載されている．あまり意味を持たない為，各データのレビューからこれを省いた．実験の各データ数は学習データを 300,000，開発データを 10,000，テストデータを 10,000 とした．投稿番号順にソートし番号が若い方から学習データ，開発データ，テストデータにした．ロジスティック回帰には Liblinear の L2 正則化ロジスティック回帰を用いて，コストとバイアスは 1 とした．

表 1　各手法の正答率

	正答率（%）
隠れ状態なし	38.44
基準手法	47.76
繋ぎ方：全対全	47.52
繋ぎ方：個対個	47.72
繋ぎ方：手調整	47.43
品詞素性の利用	47.17
N-gram素性の利用	48.31
ユーザ情報の利用	48.32

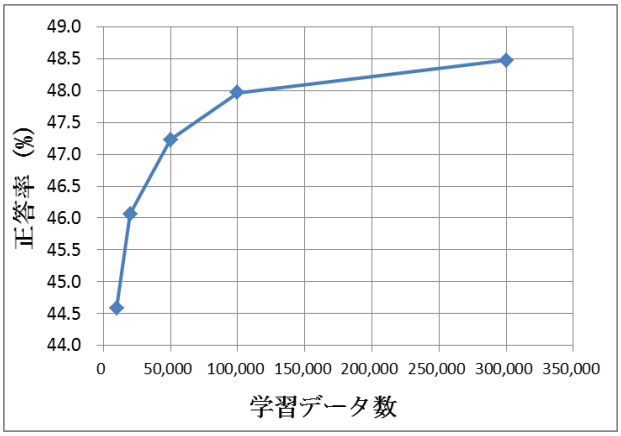

図 4　N-gram 素性利用の開発データの学習曲線

4.2　実験結果

EM 法の繰り返し回数は開発データで 5 回と決定した．各手法のテストデータでの正答率と隠れ状態なしの場合の正答率を表 1 に示す．また N-gram 素性の利用のカテゴリ別の正答率と平均二乗誤差を表 2 に示す．平均二乗誤差の算出は無回答の場合に 0 点を付与した為，0 点を省いて算出した．そして小堀らの研究結果[4]も表 2 に示す．学習曲線を図 4 に示す．

表 2　カテゴリ別の正答率と平均二乗誤差

	カテゴリ別正答率（%）	平均二乗誤差	小堀らの研究(平均二乗誤差)
立地	49.71	0.82	2.38
部屋	46.74	0.92	2.79
食事	49.08	3.23	2.36
風呂	41.44	1.45	2.85
サービス	49.21	0.85	2.89
設備	44.47	0.97	2.30
総合	57.53	0.67	1.84

表 3　学習データの内訳

	0点	1点	2点	3点	4点	5点
立地	0	1,225	8,357	42,045	112,068	136,305
部屋	0	5,773	15,663	59,424	113,530	105,610
食事	108,079	4,288	10,801	39,080	66,836	70,916
風呂	13,332	6,782	20,423	89,400	91,983	78,080
サービス	0	6,222	10,036	68,898	111,043	103,801
設備	2,011	5,897	17,920	85,186	110,770	78,216
総合	0	4,997	9,811	36,221	140,624	108,347

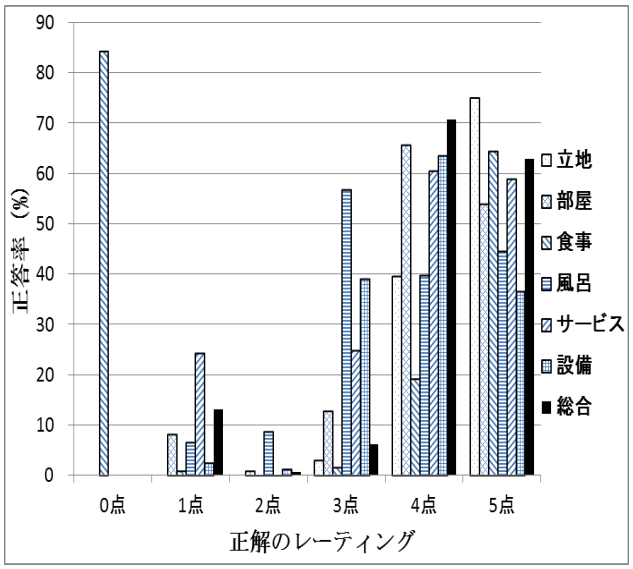

図 5　正解レーティング別正答率

5　考察

N-gram 素性の利用時の正解レーティング別正答率を図 5，表 4 に示す．これより，正解レーティングが 1 点，2 点の時の正答率が他と比べて低いことが確認できる．この要因の一つとして，学習データの偏りが挙げられる．表 3 に示す学習データの内訳の「立地」では，大半が 4〜5 点である．4〜5 点に予測が引きずられる為に 1 点，または 2 点と予測しにくいと考えられる．

6　まとめと今後の課題

ユーザの記述したレビューを文単位に分割し分割した文に隠れレーティングを与えて，レビュー全体のレーティングを予測させる手法を提案した．提案手法のレーティング予測は隠れ状態なしと比べて約 10%向上し，平均二乗誤差は小堀らの研究と比較して「食事」以外は大きく減少した．繋ぎ方の変更では，基準手法の「総合」のみ全てのカテゴリと繋ぐ方法が他の繋ぎ方と比べて最も正答率が高かった．また，品詞素性の利用では品詞を名詞や形容詞，動詞のみなどに変更しても効果はなく，基準手法と比べて正答率が減少した．ユーザ情報の「ユーザ ID」，「同伴者」，「旅行の目的」を素性に加えても N-gram 素性の利用時と比べて変化はなかった．これより，基準手法に上記の変更点を加えても大きな正答率の向上，平均二乗誤差の減少は確認できなかった．

今後の課題として，共参照解析を用いて，「これ」などの代名詞がレビュー内で何を示すかなどを特徴にうまく利用することが考えられる．

表 4　正解レーティング別正答率の詳細

	予測レーティングが正解した数/正解の合計					
	0点	1点	2点	3点	4点	5点
立地	0/0	0/43	0/253	40/1357	1496/3777	3433/4570
部屋	0/0	15/188	4/498	247/1949	2452/3735	1956/3630
食事	2903/3445	1/120	1/371	20/1314	455/2375	1528/2375
風呂	0/422	16/251	57/667	1671/2940	1225/3083	1175/2637
サービス	0/0	44/182	1/337	561/2258	2286/3781	2029/3442
設備	0/61	5/206	7/579	1067/2734	2408/3793	960/2627
総合	0/0	20/153	2/332	67/1103	3401/4813	2263/3599

参考文献

[1]Francesco Ricchi et al., Recommender Systems Handbook. Springer, 2011

[2]Bing Liu, Sentiment Analysis and Opinion Mining, Morgan & Claypool Publishers, 2012

[3]Mihai Surdeanu et al., Multi-instance Multi-label Learning for Relation Extraction. In Proceedings of EMLLP-CoNLL 2012. pp. 455-465, 2012

[4]小堀　脩豊，　評判分析に基づく推薦システムの研究，　豊田工大修論，　2013