パラグラフベクトルとNNによる 文間及びカテゴリ間の関係性を捉えたレーティング予測

Sentiment Classification Reflecting Relation among Sentences and Categories via Pragraph Vector and Neural Network

外山 洋太　三輪 誠　佐々木 裕
Yota Toyama　Makoto Miwa　Yutaka Sasaki
豊田工業大学
Toyota Technological Institute

{sd12056, makoto-miwa, yutaka.sasaki}@toyota-ti.ac.jp

1 序論

近年、評判分類において、ニューラルネットワークを用いた手法 [3][4][5] が提案されており、従来の手法を上回る正答率を達成している。ニューラルネットワークを分類問題に用いることの利点の一つは層の数を増やすことによって入力となる特徴量の深い意味を捉えられることである。多カテゴリの分類問題に適用すれば、カテゴリ間の関係性を捉えた分類が実現できる。さらに、文毎の特徴量を入力とすれば文間の関係性も捉えることができる。

文や文章の意味表現の学習手法として、単語と文章の分散表現を同時に学習するパラグラフベクトル [2] がある。これは評判分類問題に対して優れていることが示されている。

本研究は、複数カテゴリにおける評判分類について、パラグラフベクトル [2] とニューラルネットワークを用いて文間及びカテゴリ間の関係性を捉えた分類を実現することを目的とする。

提案手法では、パラグラフベクトル [2] によって生成された各レビューの文書ベクトルと文ベクトルをニューラルネットワークによる分類器において分類しレーティング予測を行う。楽天トラベルのデータセットを用いた実験において、提案手法は従来手法 [1] に対して 2pp 以上上回る正答率を示した。レーティングの平均二乗誤差 (RMSE) を元にした評価基準では、従来手法 [1] において弱点となっていたカテゴリについてそれを上回る結果を示した。

2 関連研究

2.1 隠れ状態を用いたホテルレビューのレーティング予測

藤谷ら [1] は複数のカテゴリにおける評判分類問題に対して、MIML-RE モデルを用いた手法を提案している。その手法では、レビュー内の各文毎に予測した隠れレーティングからレビュー全体のレーティングを予測する。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する際のカテゴリ間の繋がりを手動で変化させカテゴリ間の関係性を考慮している。各文の素性には Bag Of Words (BOW) または Bag Of n-gram を用いている。各文毎に隠れレーティングを予測することによって正答率が向上すること、また、カテゴリ間の繋がりによって正答率が変化することが示されている。

2.2 パラグラフベクトル

パラグラフベクトルは、文や文章といった大きな単位の言語表現の意味表現を学習する手法である。これは、Continuous Bag Of Words (CBOW) または Skip-gram[6] という単語の意味表現の学習手法を応用した手法である。ここでは CBOW を応用した Distributed Memory Model of Paragraph Vectors (PV-DM) について説明する。

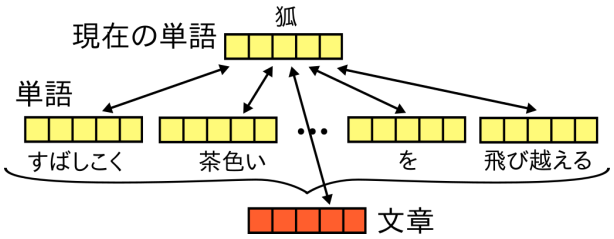

図 1: パラグラフベクトルの学習の概略

以下に具体的なアルゴリズムを示す。ここでは文章の意味表現を学習する場合について考える。学習の概略を図 1 に示す。まず、意味表現を学習する対象となる文章に含まれる単語を初めから一つずつ読んでいく。その際、現在の単語及びその周辺の単語、現在の文章について、式 1 に示す目的関数 L を最大化するように各パラメータの学習を行う。

$$L = \frac{1}{T}\sum_{t=k}^{T-k} \log p(w_t|w_{t-k},...,w_{t+k}), \tag{1}$$

$$p(w_t|w_{t-k},...,w_{t+k}) = \frac{e^{y_{w_t}}}{\sum_i e^{y_i}},$$

$$y = b + Uh(w_{t-k},...,w_{t+k},d;W,D)$$

ここで、d は文章、w_i は単語、W は全ての単語の分散表現を表す行列、D は全ての文章の分散表現を表す行列である。 k は片側のウィンドウサイズ、T は現在の文章に含まれる単語数である。ある単語の周辺を表す区間をウィンドウという。 p は softmax 関数により正規化された、文脈から現在の単語が導かれることの尤度である。 y は現在の単語とウィンドウ内の単語及び現在の文章から導出される正規化されていない対数確率である。 $h(w_{t-k},...,w_{t+k},d;W,D)$ は引数となるベクトルを平均したベクトルまたは結合したベクトルを返す関数である。

これにより、BOW と異なり、単語の並び順を考慮した文や文章の分散表現を生成することができる。

Quoc ら [2] によって、パラグラフベクトルは評判分類問題において BOW 等に比べて高い正答率を示すことが示されている。しかし、文書全体にパラグラフベクトルを用いる場合、文同士の位置関係が分類時に考慮できない。

2.3 ニューラルネットワークを用いた評判分析

ニューラルネットワークを用いた評判分析の手法が、Nal ら [3]、Rie ら [4]、Duyu ら [5] 等によって提案されている。

これらの方法に共通するのは、単語の意味表現から畳み込みニューラルネットワークと全結合ニューラルネットワークを用いて分類を行うことである。まず、単語の意味表現から畳み込みニューラルネットワークを用いて単語同士の関係を捉えた特徴量を抽出する。その後、そこから得られた文書全体の特徴量を全結合ニューラルネットワークの入力とし多値または二値分類を行う。また、Duyu ら [5] と Nal ら [3] の手法はニューラルネットワークのモデルの中にパラメータとして単語の意味表現を取り込んでいる。これにより、特定の分類問題に対してそれを直に微調整することが可能となる。

3 提案手法

提案手法では、PV-DM によってレビュー内の各文及び文章の分散表現を生成し、それらをニューラルネットワークの入力として分類を行う。これによって、文間及びカテゴリ間の深い関係性を捉える。

提案するレビュー分類手法の入力はレビューである文書と正解レーティングの組の集合、出力は各文書について予測されたカテゴリ毎のクラスである。

初めに、PV-DM を用いて、各レビューの文書ベクトルとそれに含まれる各文のベクトルを生成する。文書ベクトルと文ベクトルについては別々に学習し生成する。式 2 に示す目的関数 L_d を最大化するように学習を行う。

$$L_d = \sum_{t=n+1}^{T} \{\log \sigma(s(w_t, w_{t-n},...,w_{t-1},d)) + \sum_{w'_t \sim P_n}^{k} \log(1-\sigma(s(w'_t, w_{t-n},...,w_{t-1},d)))\}, \tag{2}$$

$$s(w_t, w_{t-n},...,w_{t-1},d) = W_{map}(w_t) \cdot \begin{bmatrix} W(w_{t-n}) \\ \vdots \\ W(w_{t-1}) \\ D(d) \end{bmatrix}$$

ここで、T は現在の文章内の単語数、t は現在の単語位置、d は現在の文章、w_i は位置 i にある単語、n はウィンドウサイズである。 $W(w_i)$ は単語 w_i に相当する単語ベクトルを単語行列 W から抜き出す関数を表す。 $D(d)$ は文章 d に相当する文章ベクトルを文章行列 D から抜き出す関数を表す。関数 $s(w_t, w_0,...,w_n,d)$ は

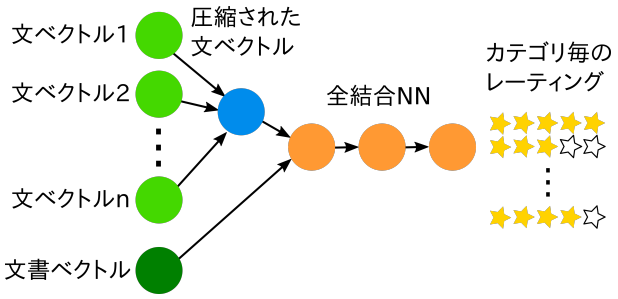

図 2: 全結合ニューラルネットワークによる分類器

ある単語とそれが出現する文脈との類似度を計算する。行列 W_{map} は内積によって文脈と単語との類似度を計算するための単語毎のベクトルを保持する。文章行列内の各文章ベクトルはレビュー全体を表す文書ベクトル、または、各レビュー内の文ベクトルを表す。σ はシグモイド関数である。また、式 2 の中括弧内の右項はネガティブサンプリングを表す。ネガティブサンプリングとは、文脈外の単語をデータセットにおける出現確率でサンプリングし、それらと文脈の意味が遠ざかるように学習する手法である。 $w'_t \sim P_n$ は文脈外の単語 w'_t を単語の出現頻度 P_n によってサンプリングすることを示す。ただし、出現頻度は極端に頻出する単語の影響を抑えるため各単語に対して 3/4 乗している。現在の単語と同じ単語や同一回の学習で一度サンプリングされた単語はサンプリングされない。

次に、各レビュー内の全ての文ベクトルに対して重み付け平均を行い、圧縮された文ベクトルを生成する。この過程により、各レビューで疎らだった文の数が統一される。式 3 に重み付け平均によって圧縮された文ベクトル $t_{i_{part}}$ を示す。

$$\mathbf{t}_{i_{part}} = \sum_{i_{sent}} \frac{w(x_{i_{part}}(i_{sent}))}{|\sum_{i'_{sent}} w(x_{i_{part}}(i'_{sent}))|}\mathbf{s}_{i_{sent}}, \quad (3)$$

$$x_{i_{part}}(i_{sent}) = \frac{i_{sent} - i_{part}}{\#partitions},$$

$$w(x) = \begin{cases} \frac{1}{2}(\cos(\pi|x|)+1) & \text{if } |x| <= 1 \\ 0 & \text{otherwise} \end{cases}$$

ここで、i_{sent} はレビュー内の文のインデックス、$\#partitions$ は重み付け平均された後の文ベクトルの数、i_{part} は重み付け平均された後の文ベクトルのインデックス、$\mathbf{s}_{i_{sent}}$ は文ベクトルである。

次に、分類器によってレーティング予測を行う。分類器は全結合ニューラルネットワークによって構成される。図 2 に各層の結合の様子を示す。

入力層はレビュー毎の文書ベクトルと圧縮された文ベクトルの結合ベクトルである。ニューラルネットワークの活性化関数には、シグモイド関数を用いる。また、出力層はカテゴリの数とラベルの数の積だけのユニットを持ち、各ユニットの出力はあるカテゴリ内であるクラスが選ばれることの正規化されていない対数確率を表す。ニューラルネットワークは式 4 に示す目的関数 E を最小化するように学習を行う。

$$E = -\sum_{n=1}^{N}\sum_{c=1}^{C}\sum_{k=1}^{K} d_{nck} \log y_{ck}(x_n; w), \quad (4)$$

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

各ユニットの出力はカテゴリ毎にソフトマックス交差エントロピー関数によって損失に変換される。ここで、u_{ck} は出力層のユニットの出力値、y_{ck} はカテゴリ c においてクラス k が選ばれる対数確率、w はニューラルネットワークのパラメータである。 d_{nck} は n 番目の文書がカテゴリ c でクラス k ならば 1、それ以外で 0 となる値である。 N はミニバッチサイズ、C はカテゴリの総数、K はクラスの総数である。

4 実験

4.1 実験設定

提案手法と従来手法 [1] 及び基準手法を比較するため、実験を行った。実験は、各手法の正答率を測定するものと、その実験で最も正答率の高かった提案手法における予測レーティングと正解レーティングの RMSE を測定するものの 2 つを行った。

比較手法として、Quoc ら [2] によって提案されている基準手法を拡張した 2 つの手法、AveSV (Average Sentence Vector) と WAveSV (Weighted-Average Sentence Vector) を用いた。基準手法と 2 つの比較手法、提案手法に用いられる分類器の入力を表 1 に列挙する。

表 1: 各手法に用いられる特徴量

手法	特徴量
基準手法	レビュー全体の文書ベクトル
AveSV	レビュー内で平均した文ベクトル
WAveSV	レビュー内で重み付け平均によって圧縮された文ベクトル
提案手法	レビュー全体の文書ベクトル、レビュー内で重み付け平均によって圧縮された文ベクトル

正答率を測定する実験は基準手法と 2 つの比較手法、提案手法全てについて行った。データセットとしては、先行研究 [1] と同様に、ホテル予約サイト楽天トラベルにおけるレビュー 337,266 件からレビューの番号順に訓練データ 300,000 件、開発データ 10,000 件、評価データ 10,000 件を用いた。

レーティングの RMSE を測定する実験は、上記の正答率を測定する実験で得られた予測レーティングを用いて行った。正解レーティングが 0 点であるものは評価から省いている。

表 2 に各手法におけるニューラルネットワークのパラメータ設定を示す。全ての手法において、中間層の数は 1、入力層及び中間層におけるドロップアウト率はそれぞれ 0.2 と 0.5 で共通である。 WAveSV と提案手法において圧縮された文ベクトルの数はそれぞれ 3 つと 2 つとした。全ての実験において文書及び文ベクトルについては、学習回数は 1,024 回、学習する単語の範囲は前 3 単語、単語の最少出現回数は 5 回、ベクトルの次元数は 600 次元に設定し学習したものを用いた。

4.2 結果と考察

表 3 に正答率を測定する実験の実験結果及び従来手法 [1] による結果を示す。表 4 に RMSE を測定する実験の実験結果を示す。

表 3 より、基準手法と 2 つの比較手法、提案手法全てが従来手法の正答率を上回っている。提案手法が従来手法 [1] の正答率を 0.0206 上回っていることから、提案手法が従来手法 [1] より正答率において優れていることが示されてた。また、WAveSV の正答率が AveSV の正答率を 0.018 上回っていることから、文の位置関係の考慮がレーティング予測に有効であることが示されてた。さらに、提案手法が基準手法や WAveSV に比べ高い正答率を示していることから、文書ベクトルと文ベクトルを同時に特徴量として用いることがレーティング予測に有効であることが示された。これは文書ベクトルと文ベクトルがいくらか異なる特徴を学習していることを示す。

表 2: 各手法のパラメータ設定

手法	学習回数	中間層でのユニット数
基準手法	20	512
AveSV	55	256
WAveSV	24	256
提案手法	30	512

表 4 より、提案手法は従来手法 [1] が苦手としていた食事と風呂のカテゴリにおいてそれぞれ 0.69 及び 0.45 だけ低い誤差を示した。藤谷ら [1] より、本実験で用いたデータセットには、食事のカテゴリにおいて 0 点が付与されたレビューが 108,079 件存在する。また、風呂のカテゴリでは 13,332 件、設備のカテゴリでは 2,011 件存在し、他のカテゴリでは 0 件である。一般に、0 点はユーザが何らかの理由でレーティング不可能と判断したことを示す。よって、提案手法は従来手法 [1] よりレビュー中の上記のような意味をよく捉えていると考えられる。しかし、その他のカテゴリにおいては提案手法と従来手法との間に有意な差は見られなかった。

5 結論

本研究では、多カテゴリにおける評判分類問題について、レビュー全体の文書ベクトルに加えレビュー内の各文ベクトルを用いた 3 つの手法を提案した。

実験では、従来手法 [1] 及びレビュー全体の文書ベクトルのみを用いた手法に比べ、文書ベクトルと重み付き平均によって圧縮した文ベクトルを同時に用いた提案手法が 2pp 以上高い正答率を示した。これはレビュー内の文の並びが評判分類に重要であること、及び、文書ベクトルと文ベクトルがレビューのいくらか異なる特徴を捉えていることを示す。

今後の課題は、文書や文の分散表現を生成する過程をニューラルネットワークによる分類器に統合することである。これによって、学習方法の柔軟性を高めると共にさらなる正答率の向上を目指す。

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.

表 3: 各手法における正答率

手法	正答率
従来手法 [1]	0.4832
基準手法	0.4969
AveSV	0.4848
WAveSV	0.4866
提案手法	**0.5038**

[2] Quoc Le, and Tomas Mikolov, Distributed Representations of Sentences and Documents. ICML 2014, 2014.

[3] Nal Kalchbrenner, Edward Grefenstette, and Phil Blunsom, A Convolutional Neural Network for Modelling Sentences. ACL 2014, 2014.

[4] Rie Johnson, and Tong Zhang, Effective Use of Word Order for Text Categorization with Convolutional Neural Networks. NAACL 2015, 2015.

[5] Duyu Tang, Bing Qin, and Ting Liu, Learning Semantic Representation of Users and Products for Document Level Sentiment Classification. ACL 2015, 2015.

[6] Yoshua Bengio et al, A Neural Probabilistic Language Model. The Journal of Machine Learning Research 3, 2003.

表 4: 提案手法と従来手法 [1] におけるレーティングの RMSE

手法	提案手法	従来手法 [1]
立地	0.87	0.82
部屋	0.88	0.92
食事	2.54	3.23
風呂	1.00	1.45
サービス	0.85	0.85
設備	0.90	0.97
総合	0.73	0.67