ニューラルネットワークと相互評価法を用いた自由記述アンケートからの単語評価

大谷 崇文[1]　椎名 広光[2]　小林 伸行[3]
岡山理科大学大学院 総合情報研究科 情報科学専攻[1]
岡山理科大学 総合情報学部 情報科学[2]
山陽学園大学 地域マネジメント学部 地域マネジメント学科[3]

i18im02ot@ous.jp[1], shiina@mis.ous.ac.jp[2], kob_nob@sguc.ac.jp[3],

1　はじめに

大学の授業アンケートのような小規模であっても，自由記述式のアンケート評価は難しい．これまで，アンケートの自由記述を評価する手法について，一部のアンケートのシードデータから全体を評価する手法をコメントと単語の評価を相互に繰り返して評価する手法 [1] を提案している．また，ニューラルネットワークなどの様々な機械学習を利用した評判分析が新しく行われてきており，同様な方式でも講義のコメントを評価することができるのではないかと考えて，Long Short Term Memory[2](以下 LSTM) を利用した講義コメント評価に関する研究を進めている．

アンケート自体の評価の他に，キーワードなど単語の評価が人によって評価どのように変化するかなどの単語の評価に興味がもたれる。本研究では，自由記述アンケートの評価とそれに表れる単語の評価を抽出について提案する．なお，使用したデータは，岡山理科大学総合情報学部情報科学科の 2014 年度春学期 (4 月〜9 月) の中間段階 (15 回中 8 回目の時期) に講義アンケートで実施された自由回答項目を利用した．調査対象とした教員は 15 人，講義の科目数は 41 科目，アンケート回答数は 1678 個である．なお，講義アンケートのコメントはそれのみを取得しているため関連する評価値の取得をしていない．そこで，コメントのみの 100 件を 12 人の人手で 1 から 6 の 6 段階のランクを付して，シードデータとしている．

2　単語とコメントの相互評価法

コメントとコメントを構成する単語の評価は，相互に評価しあう方法を用いている．評価のシードから相互評価法について，以下に述べる．

(1) コメントを構成する単語のランク推定 : 評価コメントからコメントを構成する単語のランク推定を行う．

(1-1) シードデータで使用した評価コメントから名詞・動詞・形容詞の単語を抽出し，コメントの評価値をそれに含まれている単語の評価ランクとする．

(1-2) 単語ごとの評価ランクが複数の可能性があるので，単語ごとの評価ランク頻度から単語ランク分布を作成する．単語ランク分布は，対象となる単語 w_k を含むコメントの評価ランクを $i(=1,\cdots,M,\ M=6)$ として評価ランクの出現ごとにランク i を中心とした $\mu_i(=i)$，分散を σ^2 の正規分布 $\phi(x;\mu_i,\sigma^2)$ と，単語 w_k ごとの評価ランク頻度を $N_{w_k}(i)$ を掛け合わせた $\phi(x;\mu_i,\sigma^2)\cdot N_{w_k}(i)$ で求める．

(1-3) 全評価ランクの正規分布を合成して混合正規分布を作成し，単語ランク分布とする．混合正規分布の混合数 (ランク数と同じ) を M，パラメータ α_i をランク i に対する正規分布の重みとした混合正規分布 $p_{w_k}(x)$ を次式で定義する．初期値については，$\sum_{i=1}^{M}\alpha_i=1$ となるように設定する．

$$p_{w_k}(x)=\sum_{i=1}^{M}\alpha_i\cdot\phi(x;\mu_i,\sigma^2)\cdot N_{w_k}(i)$$

(1-4) 単語ランク分布から最大ランクを単語ランク推定値とする．また，単語「説明」を例に混合正規分布による単語ランク分布を図 2 に示す．

$$R_{w_k}=argmax_{i=1,\ldots,M}\ p_{w_k}(i)$$

(2) 未評価コメントのランク推定 : 単語ランク分布から未評価コメントのコメントランク分布を作成し，コメントランク分布から，最大ランクをコメントランク推定値とする．

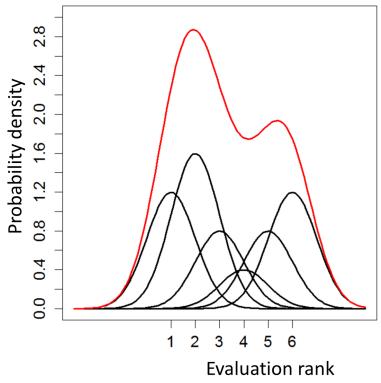

図 1: 単語「説明」の単語ランク分布

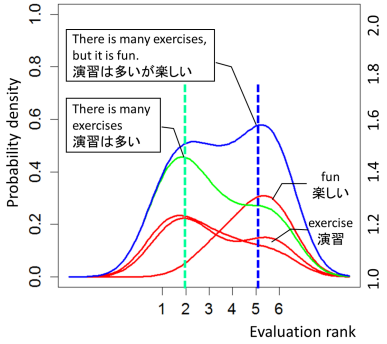

図 2: コメントランク分布

(2-1) コメント c_l のコメントランク分布 P_{c_l} を作るには，コメント内の係り受けを考慮する必要がある．まず構成している単語 w_k のランク i での単語ランク分布の確率 $p_{w_k}(i)$ に 1 を加算し，係り受け情報を反映させるため，コメント内に含まれる係り受け関係にある単語のランクごとに確率同士に重みを掛け合わせ $N_{c_l}(i) = \prod_{w_k \in c_l}(p_{w_k}(i)+1)$，コメントのランクごとの分布 $(N_{c_l}(1), N_{c_l}(2), \ldots, N_{c_l}(M))$ を作る．次に，コメントのランクごとの分布を EM アルゴリズムを用いてランク数を混合数とした混合正規分布で近似し，またその重みの合計 $\sum_{i=1}^{M} \beta_i = 1$ となるように正規化する．

$$P_{c_l}(x) = \sum_{i=1}^{M} \beta_i \cdot \phi(x; \mu_i, \sigma^2) \cdot N_{c_l}(i)$$

(2-2) コメントランク推定値 R_{cl} を (1-4) と同様にコメントランク分布から最大ランクを計算する．

$$R_{c_l} = argmax_{i=1,\ldots,M}\, P_{c_l}(i)$$

(3) 全コメントのランク推定：全コメントに対するコメントランク推定とそれを構成する単語に対する単語ランク推定を交互に繰り返して，全コメントランク推定値の改善がなくなるまで繰り返す．繰り返しの停止後，コメントランク分布と単語ランク分布から最大ランクを最終的なコメントと単語のランク推定値とする．
(3-1) 全コメントに対して (2-1) と同じように，コメント c_l のコメントランク分布 P_{c_l} を更新する．
(3-2) 全コメントに対するコメントランク分布を用いて，コメントを構成する単語の単語ランク分布を更新する．単語 w_k の属するコメント c_l のコメントランク分布 P_{c_l} をたし合わせて $\sum_{c_l \in W(w_k)} P_{c_l}(x)$ 作成し，(2-1) と同様にランクごとの分布 $(N'_{w_k}(1), N'_{w_k}(2), \ldots, N'_{w_k}(M))$ を求め，EM アルゴリズムを用いてランク数を混合数とした混合正規分布で近似する．なお，$W(w_k)$ は，単語 w_k を含むコメント集合を表し，混合正規分布は重みの合計が $\sum_{i=1}^{M} \gamma_i = 1$ となるように正規化する．コメント「演習は多いが，楽しい」のコメントランク分布を図 2 に示す．

$$p_{w_k}(x) = \sum_{i=1}^{M} \gamma_i \cdot \phi(x; \mu_i, \sigma^2) \cdot N'_{w_k}(i)$$

(4) ランク分布のパラメータ推定: 各コメントのランク推定を行うには，人手によるコメント評価との推定との差が小さいランク推定を行う必要がある．本研究では，(1) の重み α_i と正規分布の分散 σ^2 のパラメータを推定差が少なるように最急降下法による近似解で推定する．初期値はランダムで 5 回発生させ，最急降下法による近似解が最も良いものを使用している．

3 自由記述評価のための LSTM の構成

コメントデータを分かち書き処理を行い単語に区切り，コメントの 1 から 6 のランクを 0 から 1 に正規化する．学習モデルはバッチサイズを 10，初期の学習率を 0.001，LSTM 層数を 1 として，入力層にコメントデータを入力する．バッチサイズ分のコメントを単語単位で順に入力する．中間層は embed 層と LSTM 層で構成し，各ブロックで出力層にその時点での評価を出力する．また，次の時間方向に出力を受け渡す．LSTM 層には dropout を適用し，活性化関数にはシグモイド関数を用いる．出力層ではコメントの最後の

単語の出力をコメントに対するランクの推定結果として出力する．損失関数には MSE(平均二乗誤差) を用いて出力結果と人手によるランクとの差からロス率を算出しパラメータの更新を行う．パラメータ更新における確率勾配法は Adam を用いる．学習を繰り返し行うエポック数は 10 とする．推定処理 (テスト段階) は，0 から 1 で極性値 p を得る．その値 p を変換関数 $\lceil \frac{10\cdot(p+0.2)}{2} \rceil$ で 6 クラスに分類し推定ランクとする．

人手で評価されているコメントを相互評価法と LSTM による評価法で評価するクローズドテストを行った．クローズドテストの結果のうち人手の評価と推定の相関と MSE を表 1 に示す．表 1 の相関係数の平均は相互的手法では 0.569，LSTM による評価法では 0.849 であった．また，MSE については，一貫して LSTM による評価法の方が小さい値となっており，相関と MSE の結果からも LSTM による評価法の方が良い精度であると考えらえる．一方，未評価コメントに対する評価結果 (表 2 の 2,3 列目) から LSTM の方が低く推定すると考えられる．

表 1: コメント評価のクローズドテスト (評価と推定値間の相関)

評価者	相関係数		MSE(平均二乗誤差)	
	相互評価法	LSTM	相互評価法	LSTM
A	0.773	0.864	0.062	0.024
B	0.482	0.813	0.120	0.018
C	0.359	0.894	0.072	0.010
D	0.573	0.813	0.102	0.023
E	0.521	0.846	0.071	0.011
F	0.475	0.850	0.114	0.014
G	0.284	0.841	0.067	0.008
H	0.734	0.907	0.075	0.019
I	0.779	0.756	0.050	0.028
J	0.535	0.877	0.120	0.017
K	0.657	0.889	0.047	0.012
L	0.661	0.841	0.058	0.016
平均	0.569	0.849	0.080	0.017

4　学習データの繰込み方式

未評価コメントへの LSTM による評価法は低い推定をする傾向があることから，新たに評価したコメントのうち k 個を取り出して，新しい学習データとして繰込んで学習し，学習と評価を繰り返していく，学習データの繰込み方式を提案する．以下に，学習データの繰込み方式による機械学習の手順を説明する．

(1) 学習データの繰込み方式による機械学習の手順 (1-1):初期学習としては，人手による評価の付いたコメントを学習器 LSTM を用いて学習し，学習器 LSTM で未評価コメントを評価する．

(1-2):学習データに新たに評価した k 個のコメントを併合して新しい学習データを作る．併合できる場合は (1-3) の学習と推定を行う．新しい学習 k 個が学習データに併合できなくなった場合は停止する．

(1-3):併合した新しい学習データを学習器 LSTM で学習を行い，新しい学習データに含まれない未評価コメントの推定を行う．推定を行った後，(1-2) に戻る．

本研究では，コメント総数 1678 個，人手による学習コメント 100 個，新たに学習データに繰込むコメント数 $k = 100$ 個として評価実験を行った．未評価コメントの評価推定値を表 2 の 4 列目に示す．　未評価コメントの全部の平均は，相互評価法と LSTM による手法を単独で使用した場合では平均 2.790 と 2.316 に対して，学習データの繰込み方式では 3.408 となっており，単独で使用するよりも高い評価となっている．

(2) コメントから単語の評価

コメントと単語の相互評価法でのコメントから単語への評価 (第 2 章 (1)) を用いて，LSTM で評価されたコメントから単語の評価を推定する．

相互評価法，学習データを繰込んだ LSTM による評価法の 2 種類の単語の評価の違いを表 3 に示す．

5　コメントと単語の推定評価

未評価コメントと単語の推定に対する評価を述べる．

(1) 未評価コメントのランク推定に関して

●未評価コメントへの推定の平均は学習データ繰込みによる LSTM が 3.48 であり，相互評価法の 2.790 と LSTM のみの 2.316 に対して全体的に高い評価になる傾向が見られた．

●「数学が実際にどのように利用されているかがわかる」のような単語数の多いコメントはポジティブなコメントであっても，LSTM のみの推定では 1.750 と推定ランクが低くなる傾向があったが学習データ繰込みによる LSTM では 4.417 となっており評価が向上していると考えられる．

●「声が小さい．数字を入れた計算を教えてほしい」のように学習データ繰込みによる LSTM は 2.500，相互評価法は 1.167 であり，ネガティブなコメントに対する推定も高くなる傾向が見られる．

表 2: 未評価コメントのランク推定の平均

コメント	相互評価法	LSTM のみ	学習データ繰込みによる LSTM (繰込み数 $k = 100$)
わかりやすいと思います	4.500	3.500	4.667
板書がよい	2.167	3.167	2.750
授業が分かりやすい	4.333	3.083	4.333
黒板を消すのが速い	2.333	3.000	3.167
CG の作り方を学べる	3.333	3.000	3.583
数学が実際にどのように利用されているかがわかる	4.500	1.750	4.417
実技教科なので，演習や課題で技術が身につく	2.833	1.583	4.000
課題の答え合わせをしっかりやってほしい	1.167	1.500	3.000
声が小さい．数字を入れた計算を教えてほしい	1.167	1.583	2.500
声がとおっていない，ききづらい，生徒をみない	1.583	1.000	1.667
未評価コメントの平均	2.790	2.316	3.408

表 3: 単語ランク推定

	相互評価法									学習データ繰込みによる LSTM								
単語	A	B	C	···	J	K	L	平均	分散	A	B	C	···	J	K	L	平均	分散
課題	5	5	4	···	5	2	5	4.33	1.22	5	5	4	···	5	4	4	4.58	0.24
丁寧	5	5	4	···	5	5	4	4.75	0.19	5	5	4	···	5	4	4	4.67	0.22
声	1	2	1	···	1	2	2	1.75	0.35	1	4	3	···	2	2	3	2.58	0.91
交流	1	6	1	···	1	3	2	3.17	3.31	4	4	3	···	3	3	3	3.67	0.39
話	1	5	4	···	5	2	2	3.00	2.17	2	5	3	···	4	3	3	3.42	0.91
簡単	5	5	4	···	5	4	4	4.50	0.42	5	5	3	···	5	4	4	4.25	0.52
板書	2	2	3	···	2	2	2	2.58	0.91	5	4	4	···	4	3	3	4.08	0.58
教員	1	5	1	···	5	2	2	2.67	2.56	4	5	3	···	3	4	4	4.17	0.47

(2) 単語ランク推定に関して

●全体的に学習データ繰込みによる LSTM の方が推定ランクの平均は高くなり，分散は小さくなる傾向が見られる．

●「課題」の平均は相互評価法が 4.33，学習データ繰込みによる LSTM は 4.58 であり，「丁寧」の平均は相互評価法が 4.75，学習データ繰込みによる LSTM は 4.67 と，両手法による推定ランクの差はあまり見られない単語もある．

●「板書」の平均は相互評価法では 2.58，学習データ繰込みによる LSTM は 4.08 であり，学習データ繰込みによる LSTM の方がより高い評価となる単語も見られる．

●「教員」のように相互評価法ではランク平均は 2.67，分散は 2.56，学習データ繰込みによる LSTM の方ではランク平均は 4.17，分散は 0.47 であり学習データ繰込みによる LSTM の方がランク平均が高くなった単語の分散は学習データ繰込みによる LSTM の方が小さくなる傾向が見られる．

6 今後の課題

LSTM による評価と LSTM による繰込み学習では繰込み学習の方が評価が高くなる傾向が見られる．また，相互的手法と LSTM による繰込み学習の評価結果では，全体的に評価が高くなる傾向が見られるが 2 つの評価にあまり差は見られない．今後はコメントデータに含まれる単語に対して，単語の入力ごとに推定評価を抽出を行ったが，LSTM のみだけで単語の評価推定を行っていきたい．

参考文献

[1] 大谷, 椎名, “単語ランクに確率分布を用いた自由回答文解析,” 平成 29 年度（第 68 回）電気・情報関連学会中国支部連合大会，R17-27-05, 2017.

[2] Greff, K. et al, “LSTM:A Search Space Odyssey,” IEEE Transactions on Nenural Networks and Learnig Systems, Vol.28, Issue10, pp. 2222-2232, 2017.