K-016

ニューラルネットワークを用いた自由記述アンケートによる講義評価
Lecture Evaluation by Free Questionnaire using Neural Network

大谷　崇文†　　椎名広光‡
Takafumi Otani　Hiromitsu Shiina

1. はじめに

現在，大学でアンケートによる講義評価が行われている．アンケートには，さまざまな手法があるが，自由記述の評価の推定は難しい．そこで，これまで自由記述のコメントに対して，コメントと単語の評価ランクの頻度から比率を作り，それを再帰的に繰り返して評価する手法[1] を提案してきた．また，評価ランクの頻度を混合正規分布変更し，コメントや単語の評価を確率分布として考えるほうが，より正確な評価作成が可能ではないかと考えられる．

一方，ニューラルネットワークを利用した評判分析が行われてきており，RNN や Long Short Term Memory[2](以下 LSTM)が利用されている．同種類の方法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られ，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．しかし，LSTM を単純に利用した方法は，講義評価程度の少ない学習データが少ない懸念がある．そこで，テストデータの一部を順次 LSTM で評価し，それを学習に回す繰込み方法の方式も有用でないかと考えている．

本研究では，3 つの評価方式を利用して自由記述アンケートによる講義や教員の評価を行う． また，実験の設定は，2014 年度のコメントのうち 100 件を 12 人の評価者でそれぞれ評価しなおし，評価者ごとに学習器から他のコメントの評価や 2017 年度の評価を行っている．違う年度のアンケート結果に対して，経年変化による講義や教員の授業アンケートによる評価変化についても調査を行った．

2. 使用した講義アンケートデータ

本研究は，岡山理科大学総合情報学部情報科学科の次の 2 つ時期での講義アンケートを利用している．なお，講義アンケートのコメントは評価得点の取得をしていない．

(1) 2014 年度春学期(4 月～9 月)の中間段階(15 回中 8 回目の時期)，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 件．

(2)2017 年度の春 1 学期(4 月～6 月)の終了段階，調査対象とした教員は 8 人，講義の科目数は 9 科目，アンケート回答数は 754 件．

また， 2014 年度のアンケートから 100 件抽出し，評価者 12 人が人手で「とても悪い(ランク 1)」から「とても良い(ランク 6)」の 6 段階評価で評価し．シードデータとして使用している．評価のランクについては，良いから悪いをアンケートの評価でつけやすくかつ中央によりすぎないようにするため，偶数の 6 段階を設けている．なお，同じコメントであっても評価者によって差があり，提案手法で推定するコメントや単語のランク推定値も差があると考えられる．

† 岡山理科大学大学院総合情報研究科　Graduate School of Informatics, Okayama University of Science
‡ 岡山理科大学総合情報学部, Faculty of Informatics, Okayama University of Science

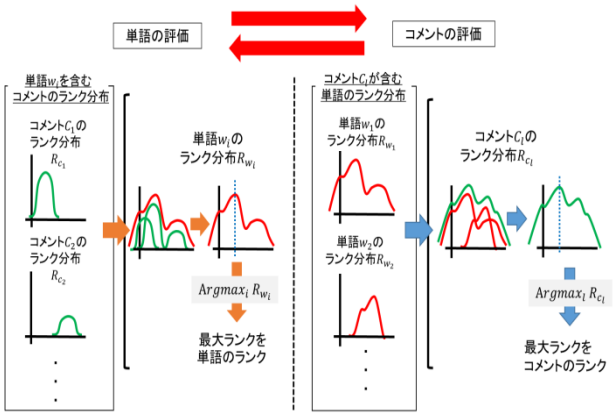

図 1　コメントと単語の再帰的評価概要

られる．

3. 確率分布の再帰的評価手法

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$\mathrm{w_k}$を含むコメントの評価ランクを$\mathrm{i}(=1,\dots,\mathrm{M},\mathrm{M}=6)$として評価ランクの出現ごとにランク$\mathrm{i}$を中心とした$\mu_\mathrm{i}(=\mathrm{i})$，分散を$\sigma^2$の正規分布$\phi(\mathrm{x};\mu_\mathrm{i},\sigma^2)$と，単語$\mathrm{w_k}$ごとの評価ランク頻度を$\mathrm{N_{wk}}(i)$を掛け合わせた$\phi(\mathrm{x};\mu_\mathrm{i},\sigma^2)\cdot\mathrm{N_{wk}}(i)$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$\mathrm{p_{wk}}(x)$を次式で定義する．初期値については，$\sum_{i=1}^{M}\alpha=1$となるように設定する．

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

(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．

$$\mathrm{R_{wk}}=\mathrm{argmax_{i=1,\dots,M}}\ p_{wk}(i)$$

(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．
(2-1)コメントc_lのコメントランク分布P_{cl}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$N_{cl}(i) = \prod_{w_k \in cl}(p_{w_k}(i) + 1)$，コメントのランクごとの分布$N_{cl}(1), N_{cl}(2), \ldots, N_{cl}(M)$を作る．
次に，コメントのランクごとの分布 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．

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

(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．

$$R_{cl} = \mathrm{argmax}_{i=1,\ldots,M}\ p_{cl}(i)$$

(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_{cl}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_{cl}をたし合わせて$\sum_{cl \in W(w_k)} P_{cl}(x)$を作成し，(2-1)と同様にランクごとの分布$N'_{wk}(1), N'_{wk}(2), \ldots, N'_{wk}(M)$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．

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

(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．

4. LSTM による評価手法

LSTM を用いた機械学習に学習させる．学習モデルは，入力されたコメントに対して 6 クラス分類

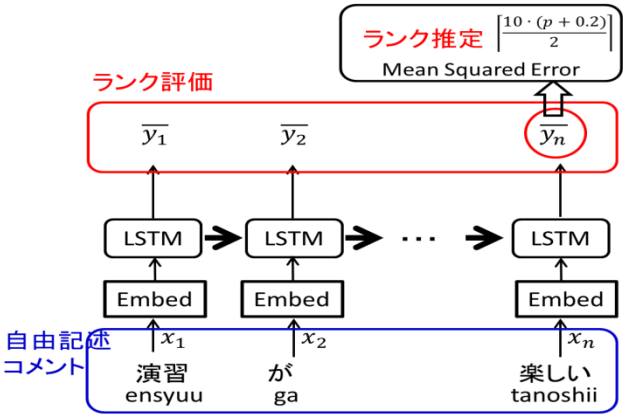

図 2　LSTM によるコメント評価

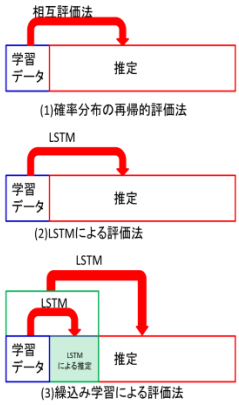

図 3　学習データと評価の関係

の結果の出力を行う分類モデルである．学習モデルは LSTM 層数を 1 として，入力層にコメントの単語を入力する．各 LSTM ブロックに接続された出力層で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．出力層は，1 つのノード(出力ノード)で構成しており，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 としている．
出力ノードからは，0 から 1 で極性値pを得る．その値pを

$$\text{変換関数}\lceil(10 \cdot (p + 0.2)/\ 2\rceil$$

で 6 クラスに分類し推定ランクとする．損失の計算は，出力された推定結果と評価ラベルとの平均二乗誤差より計算し，パラメータの更新を行っている．
単語の評価については，2.(1)のコメントを構成する単語のランク推定を利用して推定する．

5. 学習データの繰込みによる評価法

新たに評価するコメントを LSTM による手法で評価し，それを再び学習データに組み入れる．多く評価するデータがあるときは，k個ずつ取り出して評価し，その評価を学習データに繰込んでいく．学習データと評価データの関係について他の評価手法との比較を図 2 に示し，学習データの繰込み方式による機械学習の手順を述べる．

(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．
(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(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(4)へ移る．
(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，(2)に戻る．
(4) コメントの評価推定ができているので，単語の評価については，2.(1)のコメントを構成する単語のランク推定を利用して推定する．

6. 精度評価

各評価者が評価した 100 個のデータに対するクローズドテストの結果として，評価者ごとの評価と推定の相関係数及び平均二乗誤差(MSE)を表 1 に示す．再帰的評価が最も悪く，LSTM による評価，繰込みによる評価の誤差が少なくなっている．また相関係数も同様な傾向が表れており，繰込み方式の相関係数は 1 となっている．

これに対して，シードデータ以外の 2014 度と 2017 年度のデータの評価を表 2 と表 3 に示す．

表 1　クローズドテストの評価

	相関係数			MSE		
評価者	相互評価	LSTM	繰込み	相互評価	LSTM	繰込み
E1	0.773	0.864	0.968	0.062	0.024	0.0092
E2	0.482	0.813	0.968	0.12	0.12	0.0092
E3	0.359	0.894	0.958	0.072	0.01	0.0092
E4	0.573	0.813	0.968	0.102	0.023	0.0064
E5	0.521	0.846	0.96	0.071	0.011	0.0064
E6	0.475	0.85	0.964	0.114	0.014	0.0068
E7	0.284	0.841	0.962	0.067	0.008	0.0044
E8	0.734	0.907	0.945	0.075	0.019	0.0156
E9	0.779	0.756	0.965	0.05	0.028	0.0076
E10	0.535	0.877	0.962	0.12	0.017	0.008
E11	0.657	0.899	0.984	0.047	0.012	0.0024
E12	0.661	0.841	0.98	0.058	0.016	0.0036
平均	0.569	0.85	0.965	0.0798	0.025	0.0075

表 2:2014 年アンケートコメントの評価

コメント	再帰的評価法	LSTM のみ	学習データ繰込み
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	4.330	3.083	4.333
黒板を消すのが速い	2.330	3.000	3.167
CG の作り方を学べる	3.330	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

表 3　　2017 年アンケートコメント評価

コメント	LSTM のみ	学習データ繰込み
解答を丁寧に書いてくれるのでうれしいです	2.833	4.167
基本的な解き方が分かりやすい	3.916	4.333
スライドと黒板を使い分けてわかりやすく授業を進めているところ	2.417	4.083
問題を解く時間をくれます．問題の解説をしてくれます．	4.417	3.667
説明が速い．	2.833	4.500
話が聞き取りにくい．話している生徒をあまり注意しないので授業が聞こえない	4.250	3.750
分からない所があったときに分からないところの内容を聞けば教えてくれる．すごく頭を使うのでとてもいい勉強になる	3.417	3.250
わからないことを丁寧に教えてくれる．エアコンがある．早く終わることもある．	4.833	3.500
解説がやや不十分　難しい	4.417	4.417
毎回授業開始時にまとめたプリントを配ってくれる．前の授業の最後に出された問題の解説をして授業に入ってくれる	2.000	3.167

表 4　評価者 E1 のコメント評価推定

教員	2014 年アンケート						2017 アンケート			
	再帰的評価法		LSTM のみ		学習データ繰込み		LSTM のみ		学習データ繰込み	
	評価ランク平均	分散	評価ランク平均	分散	評価ランク平均	分散	評価ランク平均	分散	評価ランク平均	分散
A	3.804	408.806	4.346	105.806	3.981	268.472	―	―	―	―
B	4.056	239.139	3.408	11.472	3.983	82.333	2.708	1242.806	4.674	2243.556
C	3.282	754.472	3.660	17.000	3.484	250.917	―	―	―	―
D	3.010	194.667	3.863	11.806	3.344	144.000	―	―	―	―
E	3.724	312.889	3.904	46.472	3.741	292.889	3.796	91.806	4.375	92.000
F	3.357	322.472	3.752	32.139	3.809	235.139	―	―	―	―
G	4.042	126.333	4.469	25.806	4.041	64.806	3.618	245.472	4.295	273.222
H	3.176	91.583	3.294	11.917	3.294	44.917	―	―	―	―
I	3.530	99.333	4.629	74.556	3.614	60.556	2.860	83.222	4.755	185.472
J	3.958	1809.556	3.730	61.583	3.932	742.472	―	―	―	―
K	3.295	491.889	4.464	210.556	3.902	294.222	―	―	―	―
L	3.835	726.806	3.683	40.139	3.942	427.139	3.846	228.222	4.236	176.333
M	3.414	1376.806	3.906	151.583	3.684	770.139	3.512	553.472	4.258	522.333
N	3.470	90.667	3.625	3.556	3.672	57.222	―	―	―	―
O	3.524	217.889	4.113	80.889	3.669	237.556	―	―	―	―
P	―	―	―	―	―	―	3.407	175.472	4.422	349.000
Q	―	―	―	―	―	―	3.323	103.139	5.000	20.000
平均	3.565	484.220	3.923	59.019	3.739	264.852	3.384	340.451	4.502	482.740

7. 教員評価と単語評価

12 人の評価者のうち評価者 E1 による 2014 年度と 2017 年度のアンケートに対する評価法の違いによる評価ランクの平均と分散を表 4 に示す．講義担当や採用の関係で，評価されていない教員いるため表では，－で表している．また，評価者 E1 による 2014 年度と 2017 年度アンケートの評価から推定される単語の評価ランクを表 5 に示す．

評価者 E1 はよるコメントの 2014 年度から 2017 年度への変化は，

(1) 評価ランクが高く推移している
(2) 2014 年で低い評価を受けた教員は 2017 年では評価にない傾向がある．
(3) 2017 年の評価推定は分散が大きくなっている。
(4) 2017 年で新たに評価された教員は高い評価を受けている．

また，単語については一部評価の変化があるが，おおむね同じような評価になっている．

8. 今後の課題

本研究では，コメントの評価の精度を改善するようにニューラルネットワークの手法を用いた．ニューラルネットワークの方法では直接単語の評価ができないが，再帰的な手法で用いて間接的に評価している．しかし，単語の係り受け関係で評価が変わりやすいものあるので，共起による単語の評価についても評価が今後の課題である．

表 5 単語評価推定

単語	2014 年 再起的評価法		2014 繰込み学習		2017 繰込み学習	
	平均	分散	平均	分散	平均	分散
課題	4.33	1.22	4.42	0.91	4.17	0.31
丁寧	4.75	0.19	4.75	0.19	4.08	0.41
声	1.75	0.35	1.58	0.41	4.5	0.25
交流	3.17	3.31	3.17	3.31	―	―
話	3	2.17	3	2.17	4.08	0.24
簡単	4.5	0.42	4.58	0.41	4.17	0.31
板書	2.58	0.91	2.17	0.31	3.83	0.14
教員	2.67	2.56	2.17	0.64	3.83	0.47
速い	2.25	0.52	2.25	0.52	4.42	0.24
講義	4.5	0.75	4.75	0.19	3.08	2.24
面白い	4.92	0.58	4.92	0.58	4.58	0.41
やる気	4.17	1.64	4.67	0.39	4.17	1.64
プリント	4.67	0.22	4.67	0.22	4.5	0.25

参考文献

[1] Asami Shiwaku, Nobuyuki Kobayashi, Hiromitsu Shiina, “Word and Comments Evaluation using Recursive Evaluation in Lecture Questionnair”, INFORMATION AND SYSTEMS IN EDUCATION Vol.16, No.1, pp.1-6, 2017.
[2] K. Greff et al., “LSTM:A Search Space Odyssey," IEEE Trans. Nenural Netw. Learn. Syst. Vol. 28, Issue. 10, pp. 2222-2232, 2017.