7P-06

リカーシブニューラルネットワークを用いたセンチメント分析とその日本語文への適用

赤井　龍一†　　渥美　雅保‡

創価大学工学部情報システム工学科†　　創価大学理工学部情報システム工学科‡

1.はじめに

センチメント分析の分野において，分散表現された単語の意味を構文木に沿って合成することを通じて句と文のセンチメントを構成的に分析する手法として，リカーシブニューラルネットワークRNTN(Recursive Neural Tensor Network)[1]が提案されている．RNTN は構文木の全てのノードに対してボトムアップにテンソルベースの関数を適用させることにより単語，句，文と辿ってセンチメント分析をしていく手法である．本論文では，RNTN を日本語文のセンチメント分析に適用する．そのために，センチメント分析のための英語文のコーパスStanford Sentiment Treebank[2]をもとに単語と文のみに教師ラベルがついた日本語文のコーパスを作成して，単語と文に対する教師ラベルのみから学習を行うようにした場合の日本語文に対するセンチメント分析の精度を評価する．

2.RNTN によるセンチメント分析

2.1.概要

RNTN[1]は様々な長さと構文タイプの句に対する合成ベクトル表現を単語の分散ベクトル表現からリカーシブに計算するニューラルネットワークモデルで，合成ベクトル表現は各句を分類する特徴として用いられる．センチメント分析では，これら合成分散ベクトルから単語・句・文のポジティブ・ネガティブクラス，並びに 5 センチメントクラス–強くポジティブ，ポジティブ，ニュートラル，ネガティブ，強くネガティブ–分類をソフトマックスクラシファイアにより行う．学習では，すべての単語・句・文に対するセンチメントラベルが教師ラベルとして用いられる．これにより，文中に現れる否定がセンチメント分析に反映されることもあり従来法に比べて高い精度が実現されているが，ラベル付けのコスト，特にすべての句に対するラベル付けのコストが高い．日本語文のセンチメント分析器の学習では，Stanford Sentiment Treebank[2]の英文を日本語文に変換し，それら文を形態素解析と構文解析により生成した構文木に対して，それらの単語と文に元の英文の対応するセンチメントラベルを与え，句に関しては簡易な規則により句のラベルを部分的に生成して用いる．

2.2.リカーシブニューラルネットワーク RNTN

RNTN は，単語埋め込み層，ニューラルテンソル層，センチメント分類層からなる．単語埋め込み層は，単語の one-hot ベクトルw に対して，単語埋め込み行列W_Eを用いて(1)式によりその分散ベクトル表現v_Wを計算する．

$$v_W = embedID(w) \quad (1)$$

ここで，v_Wはd-次元ベクトルとする．ニューラルテンソル層は，二分構文木の左と右の子ノードの分散表現をそれぞれ$\mathrm{v}_l, \mathrm{v}_r$とするとき，(2)式によりそれらの合成ベクトル$\mathrm{v}_p$を句のベクトル表現として計算する．

$$v_p = tanh\left(\begin{bmatrix}v_l\\v_r\end{bmatrix}^T W_B^{[1:d]}\begin{bmatrix}v_l\\v_r\end{bmatrix} + W_L\begin{bmatrix}v_l\\v_r\end{bmatrix}\right) \quad (2)$$

ここで，v_lとv_rはd–次元単語ベクトルまたは d–次元合成句ベクトル，$W_B^{[1:d]}$は$R^{2d\times 2d\times d}$の双線形テンソルである．これら合成ベクトルは単語から句，そして文へとリカーシブに計算される．センチメント分類層は，構文木の各ノードの出力，即ち単語ベクトル，合成句ベクトル，合成文ベクトルvに対して，(3)式によりセンチメントクラス次元のセンチメントラベルの確率分布を出力する．

$$p_v = softmax(W_o v) \quad (3)$$

学習におけるロスは，単語・句・文それぞれに対して与えられる教師ラベルtとソフトマックスクラシファイアの出力確率分布pとの間のクロスエントロピーの総和として(4)式により計算される．

$$L(\theta) = \sum_v L_v(\theta) = -\sum_v \sum_i t_{v,i}\, log\, p_{v,i} + \lambda\|\theta\|^2 \quad (4)$$

ここで，p_vは単語・句・文の分散ベクトルvに対する出力確率分布，t_vは one-hot ベクトルで与えられるその教師ラベル，$\theta = (W_E, W_B, W_L, W_O)$であり，θはロスの誤差逆伝播で最適化される．

2.3.単語教師ラベルからの句教師ラベルの設定

RNTN の高い精度はすべての句に対する教師ラベル付けによるところが大きいが，このラベル付けには膨大なコストがかかる．そこで，二分構文木の左と右の子ノードの教師ラベルから親ノードの教師ラ

Sentiment Analysis using Recursive Neural Network and its Application to Japanese Sentences
†Ryuichi Akai, Dept. of Information Systems Sci, Faculty of Eng, Soka University
‡Masaysu Atsumi, Dept. of Information Systems Sci, Faculty of Sci, and Eng, Soka University

ベルを導く簡易な規則を導入する．いま，5 つのセンチメントクラス「強くポジティブ」，「ポジティブ」，「ニュートラル」，「ネガティブ」，「強くネガティブ」をそれぞれ 4,3,2,1,0 で表し，左と右の子ノードのラベルをそれぞれl_{LC}とl_{RC}，親ノードのラベルをl_Pとする．このとき，子ノードの教師ラベルから親ノードの教師ラベルを導く規則として，

RULE1: $if\ l_{LC}$==2 $then\ l_P = l_{RC}$
RULE2: $if\ l_{RC}$==2 $then\ l_p = l_{LC}$
RULE3: $if\ l_{LC}$>2 $and\ l_{RC}$>2 $then\ l_P$= max(l_{LC}, l_{RC})

を導入する．これにより，単語の教師ラベルを与えることで部分的に句の教師ラベルをボトムアップに設定することができる．このとき，学習におけるロスは，単語，文と教師ラベルが設定された句の集合V^*に対するソフトマックスクロスエントロピーの和$\sum_v L_{v\in V^*}(\theta)$として計算される．

3.実験

3.1.概要

RNTN を用いた日本語文のセンチメント分析の実験的評価を行う．実験には，Stanford Sentiment Treebank の英文コーパスを利用して作成した日本語文の 3 種類のコーパスを用いる．句に対する教師ラベルを与えないで学習する場合と句に対する教師ラベルを「句教師ラベル設定規則」により設定して学習する場合の精度を比較する．

3.2.データセット

Stanford Sentiment Treebank の英文コーパスから日本語文のコーパスを作成する処理の流れを図 1 に示す．日本語文の形態素解析と構文解析には Jigg[3]の Kuromoji と Jaccg を利用した．

このコーパスから文の長さを基準に 30 単語で長文・短文にわけ，「長文コーパス」，「短文コーパス」，「長文＋短文コーパス」の 3 つのコーパスを作成した(表 1)

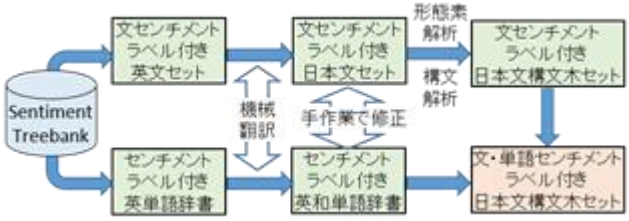

図 1　日本語文コーパス作成の流れ

表 1　日本語文コーパス(文数)

	長文	短文	長文+短文
train	899	1002	2004
dev	102	100	100

3.3.評価

日本語文コーパスに対して，句に対する教師ラベルなしと教師ラベルを「句教師ラベル設定規則」により設定して学習した場合の精度を表 2 に示す．

表 2 日本語文コーパスの結果

	5 分類:4,3,2,1,0			3 分類:(4,3),(2),(1,0)		
	全て	文	単語	全て	文	単語
長文	77.7	57.8	95.5	78.1	84.3	95.6
短文	77.9	53.0	96.3	78.7	79.0	96.3
長+短	79.1	54.0	96.8	79.8	85.0	96.8

(a)句ノード:教師なし

	5 分類:4,3,2,1,0			3 分類:(4,3),(2),(1,0)		
	全て	文	単語	全て	文	単語
長文	83.1	57.8	96.3	83.5	79.4	96.5
短文	89.2	49.0	97.2	90.2	78.0	97.2
長+短	96.7	45.0	97.0	97.4	75.0	97.0

(b)句ノード: 句教師ラベル設定規則でラベル付け

また，Stanford Sentiment Treebank の英文コーパスに対する結果を表 3 に示す．

表 3 Stanford treebank の英文コーパスの結果

5 分類:4,3,2,1,0			3 分類:(4,3),(2),(1,0)		
全て	文	単語	全て	文	単語
71.9	28.5	96.8	76.0	44.6	97.0

(a)句ノード:教師なし

5 分類:4,3,2,1,0			3 分類:(4,3),(2),(1,0)		
全て	文	単語	全て	文	単語
73.3	34.9	97.4	78.6	53.6	97.4

(b)句ノード: 句教師ラベル設定規則でラベル付け

日本語文の精度が英文より高いのは，手作業で意味の通らない文を削除したり整形をしたりしたためと思われる．句教師ラベル設定規則により全ノードでの精度は上がるが文に関してはそうでないため改良が必要である．英文コーパスで全句ノードに教師ラベルを与えた場合の精度は 79.1%であったが，教師ラベルを制限した上記の結果はそれより下がっており，句へのラベル付けの必要性が再確認された．

4. むすび

本論では日本語文に対するセンチメント分析の RNTN の適用について述べた。システム及びデータセットはまだ初期段階にあり改善が必要である．今後の課題として日本語訳の精度をより良くしていくことがあげられる．

参考文献

[1] Socher,R., et al.: Recursive Deep Models for Semantic Compositionality Over a Sentiment Treebank, EMNLP 2013.
[2] Sentiment Treebank, https://nlp.stanford. edu/sentiment/treebank.html
[3] Noji,H. and Miyao,Y.: Jigg: A Framework for an Easy Natural Language Processing Pipeline, Proc. of the 54th Annual Meeting of the ACL, pp.103-108, 2016.