The 32nd Annual Conference of the Japanese Society for Artificial Intelligence, 2018

リカーシブニューラルテンソルネットワークの日本語文のセンチメント分析への適用

An Application of Recursive Neural Tensor Network to Sentiment Analysis of Japanese Sentences

赤井 龍一[*1] Ryuichi Akai
渥美 雅保[*1] Masayasu Atsumi

[*1] 創価大学理工学部情報システム工学科
Information Systems Science, Faculty of Science and Engineering, Soka University

Recursive Neural Tensor Network (RNTN) is a neural network model that recursively computes the synthetic distributed vector representation for phrases of various lengths and syntax types from the distributed vector representation of words along the syntax tree. Distributed vector representation is used as a feature to classify each phrase and it is used to classify sentiment of phrases in sentiment analysis. In this paper, we apply the RNTN to sentiment analysis of Japanese sentences. For this purpose, based on the Stanford Sentiment Treebank corpus for sentiment analysis, we first create a corpus of Japanese sentences with teacher labels only for words and sentences. Then we evaluate the accuracy of sentiment analysis on Japanese sentences when we learn from only teacher labels for words and sentences. We also consider the effect of attaching teacher labels of phrases by heuristic rules.

1. はじめに

リカーシブニューラルテンソルネットワーク RNTN(Recursive Neural Tensor Network)[Socher 13]は様々な長さと構文タイプの句に対する合成分散ベクトル表現を単語の分散ベクトル表現から構文木に沿って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，合成分散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，これら単語・句・文の分散ベクトル表現を用いてそれらをポジティブ・ネガティブの 2 クラス，もしくはより多くのクラスにソフトマックスクラシファイアにより分類する．本論ではこの 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[注 1]をもとに，まず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する．そして，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．また，句の教師ラベルをヒューリスティックな規則によりつけることの効果も検討する．

2. RNTN によるセンチメント分析

2.1 概要

リカーシブニューラルテンソルネットワーク RNTN を用いたセンチメント分析では，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．具体的には，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行っていく．本論では，センチメントを，「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」の 5 センチメントクラスに分類する場合を扱う．RNTN の学習では，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,文中に現れる否定語によるセンチメントの反転に関する効果が学習されて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 の英文を日本語文に翻訳し，それら文から形態素解析と構文解析により構文木を生成して，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る．また，句に関しては，教師ラベルとしてのセンチメントラベルは与えずに，簡易な規則により句のラベルを部分的に生成して与える．

2.2 リカーシブニューラルテンソルネットワーク 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．

$$v_W = embedID(w)$$

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v_l,v_rとするとき，(2)式によりそれらの合成ベクトルv_pを句のベクトル表現として計算する．

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

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．

$$p_v = softmax(W_o v)$$

連絡先：赤井 龍一，創価大学理工学部情報システム工学科，
〒192-8577 東京都八王子市丹木町 1-236,Tel: 042-691-2211
E-mail: m10777.t@outlook.jp

注 1: https://nlp.stanford.edu/sentiment/treebank.html

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．

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

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. 日本語コーパスの作成

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．まず，Stanford Sentiment Treebank の英文コーパスから，文センチメントラベルが付いた英文セットとセンチメントラベルが 2 以外の英単語の辞書を作る．次に，それらを機械翻訳により文センチメントラベル付き日本語文セットと単語センチメントラベル付き英和辞書に翻訳する．そして，それらに対して，意味の通らない文の修正・削除，英和辞書の修正・追加などを手作業で行う．最後に，それら日本語文を形態素解析・構文解析して構文木を生成し，その単語にセンチメントラベルを割り振る．日本語文の形態素解析と構文解析には Jigg[Noji 16]の Kuromoji と Jaccg を利用した．

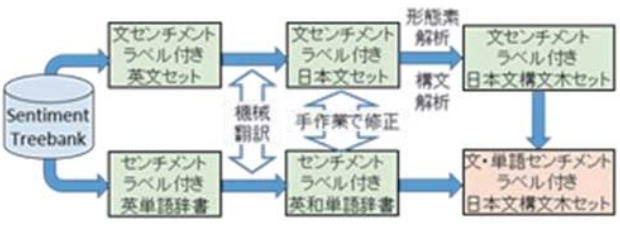

図 1　日本語文コーパス作成の流れ

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．図 2 と 3 に長文コーパスと短文コーパスの例文，表 1 にコーパスの文の数を示す．

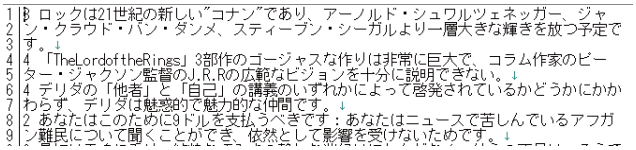

図 2　長文コーパスの例文

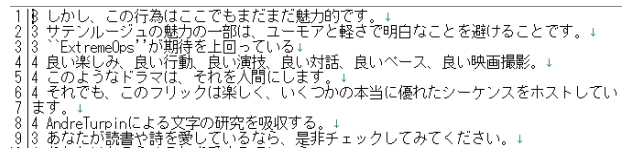

図 3　短文コーパスの例文

表 1　日本語文コーパス(文数)

	長文	短文	長文+短文
訓練用	1988	2211	4197
テスト用	102	100	100

また，「長文＋短文コーパス」から，「しかし」または「ない」を含む文だけを抽出して「否定文コーパス」を作成した．表 2 に否定文コーパスの文数を示す．

表 2　否定文コーパス(文数)

否定語を含む文	619
否定語を含まない文	3580

4. 実験

4.1 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「長文コーパス」，「短文コーパス」，「長文＋短文コーパス」，及び「否定文コーパス」を用いる．否定文コーパスは，否定語の有無がセンチメント分析の精度に対してどの程度影響を及ぼしているかを分析するために用いる．また，比較のために「英文コーパス」を用いた実験も行う．句に対する教師ラベルを与えないで学習した場合と，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した場合の精度を評価する．

4.2 実験結果

(1) 文の長さ

日本語の「長文コーパス」，「短文コーパス」，「長文＋短文コーパス」を用いたセンチメント分析実験により，文の長さの精度に及ぼす影響を評価した．表 3 に句ノードに対する教師ラベルなしの場合と「句教師ラベル設定規則」により教師ラベルを設定して学習した場合の精度を示す．長文コーパスのほうが短文コーパスよりも高い精度が得られる傾向がみられた．また，句教師ラベル設定規則の適用によりノード全体に対しては精度の向上

がみられたが，そのボトムアップな適用の影響が文まで適切に及んでいないことが確かめられた．

表 3 日本語文コーパス(文長別)の結果

	5 分類		
	全ノード	文	単語
長文コーパス	68.7	67.6	97.5
短文コーパス	68.8	56.0	94.6
長文＋短文コーパス	79.4	56.0	97.5

(a)句ノード：教師ラベルなし

	5 分類		
	全て	文	単語
長文	86.1	56.8	97.2
短文	82.7	54.0	95.2
長+短	83.1	50.0	96.4

(b)句ノード：教師ラベル設定規則でラベル付け

(2) 否定語の有無

日本語の「否定文コーパス」を用いたセンチメント分析実験により，「しかし」，「ない」といった否定語がセンチメント分析の精度に及ぼす影響を評価した．表 4 に句ノードに対する教師ラベルなしの場合と「句教師ラベル設定規則」により教師ラベルを設定して学習した場合の精度を示す． 否定語を含む文のほうが文に対する精度は下がるが，全ノードに対する精度はむしろ上がる傾向がみられた．また，この傾向は句教師ラベル設定規則の適用においてもみられた．句教師ラベル設定規則はセンチメントの反転を扱っていないが，句ノードへのラベル付けが何らかの効果をもたらしていることが示唆される．

表 4 日本語文コーパス(否定語の有無)の結果

	5 分類		
	全ノード	文	単語
否定語あり	83.0	46.0	95.0
否定語なし	81.7	54.0	96.3

(a)句ノード：教師ラベルなし

	5 分類		
	全ノード	文	単語
否定語あり	86.7	45.0	96.3
否定語なし	84.2	53.0	96.4

(b)句ノード:教師ラベル設定規則でラベル付け

(3) 英文の結果

Stanford Sentiment Treebank の英文コーパスに対するセンチメント分析の結果を比較のために表 5 に示す．

日本語文に対するセンチメント分析の精度が全般的に英文に対する精度より高くなっている．これには，日本語文のコーパスに対して手作業で意味の取りにくい文を整形したり削除したりした効果も考えられるが興味深い傾向である．また，英文コーパスに対しては句教師設定規則が機能して精度向上につながっている点も興味深い．

英文コーパスにおいてすべての句ノードに教師ラベルを与えた場合の全ノードに対する 5 分類精度は約 80%，文ノードに対する 5 分類精度は約 40%であった．これより，句へのラベル付けの必要性が再確認された．

表 5 Stanford treebank の英文コーパスの実験結果

5 分類		
全ノード	文	単語
71.9	28.5	96.8

(a)句ノード:教師なし

5 分類		
全ノード	文	単語
73.3	34.9	97.4

(b)句ノード: 句教師ラベル設定規則でラベル付け

(4) 考察

以上の実験結果をまとめると次のとおりである．

(a) 長文コーパスのほうが短文コーパスよりも高い精度が得られる傾向がみられた，
(b) 否定語を含む文のほうが文に対する精度は下がるが，全ノードに対する精度はむしろ上がる傾向がみられた，
(c) 日本語文に対するセンチメント分析の精度が全般的に英文に対する精度より高い傾向がみられた，
(d) ボトムアップな句教師ラベル設定規則の適用により全ノードに対する精度の向上はみられたが，文の精度の向上にはつながらなかった．一方，英文に対してはうまく機能した．

ボトムアップなセンチメント分析において，句ノードに対する適切なセンチメントの推定は，特に否定語が含まれる文における否定語によるセンチメントの反転を文のセンチメントに伝えるうえで重要な役割を果たす．今回の句教師ラベル設定規則にはこの反転を文に伝える機能がないため今後改良が必要である．

5. むすび

本論文では，リカーシブニューラルテンソルネットワーク RNTN の日本語文のセンチメント分析への適用について述べた．今後の課題として，句に対するラベル付けを自動的に行うより良い方法を開発していきたい．また，Stanford Sentiment Treebank の日本語訳の量と質を改善していくとともに，日本語文の別のセンチメント分析用のコーパスを作成して実験をしていきたい．

参考文献

[Socher 13] Socher,R., et al.: Recursive Deep Models for Semantic Compositionality Over a Sentiment Treebank, EMNLP 2013.

[Noji 16] 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.