Vol.2015-NL-224 No.18
2015/12/4

意味の構成性に基づく句の文脈を考慮したベクトル空間モデル

須田山 強真[1,a)]　阿辺川 武[2]　高野 明彦[2]

概要：言語学における分布仮説を根拠として単語の素性をその文脈に基づいて与えることは自然言語処理の分野で広く行われてきた。しかし感情分析において事前に用意した単語の素性の中から句を構成する単語に対応するものを与えるだけでは、句に対する極性の予測には不十分であることが知られていた。こうした問題に対して近年提案された教師あり学習による手法では、句構造文法に基づいた文の解析木を入力として与え、木の頂点として現れる句に対してその極性を正しく予測できるように訓練する。そこではベクトルと行列の対をパラメータとして単語から句の特徴ベクトルを構成的に計算していくが、その際テキストに現れる一部の句は、その意味を構成性の観点から説明するのが難しいという問題があった。この点について本研究では、他の句の極性を変化させる役割を果たす行列の割り当てに際して句の文脈を考慮するように拡張手法を提案する。そして感情分析のタスクによる評価の後に、構成性によって得られた句の特徴ベクトルについての分析を行う。

キーワード：ベクトル空間モデル, 意味の構成性

1. はじめに

言語学における分布仮説を根拠として、単語の素性を現れる文脈に基づいて与えることは自然言語処理の分野で広く行われてきた。単語の出現する文脈の情報に基づいて得られる素性は単語の分布表現と呼ばれ、特に低次元実ベクトルとして表されるものは分散表現という。2 つの単語に対応する分布表現のコサイン類似度は、人間による類似の程度に関する判定との相関があることが知られており、単語対の差の向きに関するコサイン類似度は統語的あるいは意味的な類似性と相関があることがわかっている。[5], [8]

またこうした表現を既存の教師あり学習に基づく NLP のシステムに対し、入力として与える素性に含めることで性能が改善するという報告がある。[11]

一方、入力に含まれる単語を個別に考慮するだけでは不十分なタスクの一つとして感情分析があげられる。このタスクでは従来 bag-of-words による素性が使われていたが、レビューの書き手の態度を予測する上で ”least interesting” / ”most interesting” の例のように修飾によって大きく意味が変化する表現を無視することはできない。そうした問題に対して近年提案されたアプローチの一つに、再帰的ニューラルネットワーク (RNN) を用いる教師あり学習手法がある。

RNN を用いたアプローチでは、扱う構造の各部分で正しい出力が得られるように訓練を行うが、自然言語の文とその統語的な構造に関してそのような手法を用いる論拠としては Montague による構成性原理が挙げられる。構成性に着目した単語の分散表現に関しては、形容詞と名詞からなる句に対してそれぞれに行列とベクトルを与えた上で、形容詞が名詞の意味をどのように変化させるかを扱う試みがある。[6]

構成性原理をより現実的な問題設定に応用するにあたっては、Socher らによって RNN を用いる手法が有効であることが報告されている。[9], [10] いくつかの RNN による手法の中でも MV-RNN [9] は、句の間の相互作用を捉えるために異なる単語ごとに割り当てられる行列パラメータを導入しているという点で他の手法とは異なる。

ところで MV-RNN では行列を割りあてる単位として単語を用いるために、訓練の際に最適化するパラメーターの次元は語彙の大きさ V に関して線形で増加するが、特にこれは行列パラメータを導入する場合には問題になる。同じ著者による RNTN[10] ではこの点に関して、単語ごとの行列パラメータを取り除くとともに、構成に用いる関数に手を加えることでさらに感情分析タスクでの性能を改善して

[1] 東京大学 大学院情報理工学系研究科
[2] 国立情報学研究所
[a)] sudayama@is.s.u-tokyo.ac.jp

いるが、ある句の極性のような性質がいくつかの構成要素からどのように得られるかという点については、より分析が難しくなっている。

この点に関して、本研究では上述の問題を避けつつ、構成性を持つ分散表現の性質についてより詳しく分析することを目指して、ある句が他の句の性質をどう変えるかを行列として明示的に扱うように MV-RNN の定式化を拡張する。続いていくつかの実験を通して句どうしの相互作用について調べる。

2. 関連研究

MV-RNN を拡張した手法として AdaMC[1] が提案されており、この手法では構成のための関数として事前に決めた数の候補を用意して、どれを用いるかを関数の入力として与えられる 2 つの句に基づいて決定する。しかし関数の候補は双方の句に依存して重み付けられるため、一方の句が他方の句に対してどのように作用するかという点について分析するのは容易ではないという点では RNTN と同様のことがいえる。

また直接比較を行うことはできないが、教師なしの設定で構成性を持つようなベクトル空間モデルについて扱った研究として、Fyshe らによる CNNSE[3] がある。CNNSE では非負値行列因子分解を拡張し、疎な正則化とともに構成性に関する制約を課している。

3. 手法

3.1 行列とベクトルの対による句の分散表現

構成性の成り立つような句の分散表現を扱った研究の中で、拡張を施すにあたっては MV-RNN[9] を参考にする。本論文での拡張について述べる上で、必要な事柄についてのみ簡単に説明する。以下では句または単語を指して、構成素と呼ぶことにする。

意味の構成性を考慮した分散表現に関する研究では、ある文とともに葉と内部ノードがそれぞれ単語とその文に含まれる句を表すような、句構造文法に基づく解析木を仮定することが多く、ここでもその慣習に従うことにする。

ある句に対応する内部ノードに着目したとき、その句に対応するベクトルは次のようにある句を構成するような、2 つの句に対応する行列とベクトルの対 $(x_l, M_l), (x_r, M_r)$ をととって元のベクトルと同じ次元にうつす関数 f を用いて計算する。

$$f((x_l, M_l), (x_r, M_r)) = g\left(W \begin{pmatrix} M_r \cdot x_l \\ M_l \cdot x_r \end{pmatrix}\right) \quad (1)$$
$$where \quad W \in R^{d \times 2d},\ b \in R^d$$
$$M_l, M_r \in R^{d \times d},\ x_l, x_r \in R^d$$

ここで g は構成から得られたベクトルに対して適用する変換を表す。再帰的ニューラルネットワークを用いたモデルの場合には、双曲線正接関数 (tanh) などが用いられる。

上述の式を用いると、解析木にそって構成素 x_l, x_r から句に対応するベクトル x_p を再帰的に得る。具体的には、2 つの構成素から句のベクトルを得るために $W \in R^{d \times 2d}$ を用いて、互いに作用を与えた後の 2 つの句のベクトルを元の d 次元ベクトルにうつす。ここで重みとバイアス W, b は文脈に依存しないパラメータである一方、残りの M_l, M_r, x_l, x_r については単語のパラメータから再帰的に構成される。

こうして得た句の分散表現を多クラス分類問題を解くのに用いる場合には、クロスエントロピー誤差とともにソフトマックス関数を用いることができる。この場合、句 x_i に対応するラベルを y_i と書くことにして、訓練に用いるデータ集合に対する誤差は次のように与えられる。

$$E^{(1)}(W^{(c)}) = -\frac{1}{N}\sum_{i=1}^{N}\sum_{j=1}^{C} 1[y_i = j] \log \frac{\exp(W_j^{(c)} \cdot x_i)}{\sum_{l=1}^{C} \exp(W_l^{(c)} \cdot x_i)} \quad (2)$$

ここで N はラベルのついた句の総数を表す。また (c) は分類に関する誤差関数に対して付加する。

3.2 句に対する行列の割り当て

事前に決めておいた句の行列の成分の数を K と表記する。

句に対応するベクトル表現 $x \in R^d$ が与えられたとし、句のベクトルに応じて重み付けられる、K 個の潜在行列を $\{M_k \in R^{d \times d}\}$ と表記する。このとき句に対して割り当てられる、各潜在行列の寄与 $\pi_k(x)$ を次のように定める。

$$\pi_k(x) = \frac{\tilde{\pi}_k(x)}{\sum_{j=1}^{K} \tilde{\pi}_j(x)} \quad (3)$$
$$\text{where} \quad \tilde{\pi}_k(x) = w_k^{(m)} \cdot x$$

さらに各潜在行列に対する密度として、行列をベクトルとみなしたときに中心が単位行列であり、等方的な分散を持つ多変量ガウス分布を仮定する。

$$m_k | \mu_k, \sigma^2 I \sim \text{MvNormal}(\mu_k, \sigma^2 I) \quad (4)$$

ここで潜在行列の成分 M_k をベクトルとみなすときには m_k と記す。同様にその場合、句のベクトル x に対応する行列は $m(x)$ と表す。

このとき句のベクトル x に対応する行列の密度は次のように与えられる。

$$P(m(x) | \{\mu_k, \sigma^2 I, w_k^{(m)}\}) = \sum_{j=1}^{K} \pi_j(x) P(m_j | \mu_j, \sigma^2 I) \quad (5)$$

訓練集合に現れる句に対する負の対数尤度は次のようになる。

$$E^{(2)}(\{w_k^{(m)}\}) = -\frac{1}{N}\sum_{i=1}^{N} \left\{ \log \sum_{j=1}^{K} \pi_j(x_i) P(m_j | \mu_j, \sigma^2 I) \right\} \quad (6)$$

モデル	ベクトル	行列
MV-RNN	$\|V\|d$	$\|V\|cd$
拡張	$\|V\|d$	Kd^2

図 1: モデルごとのパラメータ次元

モデル	精度 (%)	ベクトルの次元
RNN	43.2	25 ∼ 35
MV-RNN	44.4	25 ∼ 35
AdaMC-RNN	45.8	25
拡張	42.6	5

図 2: 感情分析タスクでの性能

3.2.1 導入されるパラメータ

MV-RNN と本論文での拡張で導入されるパラメータに関しての比較を図 1 に示した。ここで d 次元の句または単語のベクトルに依存して重み付けられる K 個の行列は語彙の大きさ $|V|$ に比べて、小さくとれることに注意する。

訓練する必要のあるパラメータはいずれも線形に増加するが、MV-RNN では素朴に各単語に行列パラメータを割りあてると、単語のベクトルの次元による 2 乗の定数項の影響で、訓練に時間が増えるために対角成分とそうでない成分にわけて後者に低ランク近似を行っている。

3.3 訓練

3.3.1 誤差関数

拡張したモデルの訓練にあたっては式 (2), (6) の二つを合わせた次の誤差関数を上で導入したパラメータ $\{\mu_k, w_k^{(m)}\}, W^{(c)}$ に関して最適化する。

$$E(\Theta) = E^{(0)}(\Theta) + E^{(1)}(\Theta) + \frac{1}{2}\sum_{i\in I}\lambda_i\|\theta_i\|_2^2 \qquad (7)$$

最適化の際の学習率の調整には、AdaGrad[2] を用いた。

3.3.2 単語ベクトルの初期化

構成性を持つベクトル空間モデルにおいて、単語に対応するベクトルは別の教師なし学習手法によって得られた、単語のベクトル表現を初期化に用いる場合がある。[9] また形容詞を含む句の表現およびその間の類似度について扱った Baroni らの研究では、同じ文に現れる単語の共起行列を特異値分解したものを入力として用いている。[6]

ただ Socher らによる後の論文 [10] では、感情分析タスクを構造をともなう分類問題として、十分なデータとともに再帰的ニューラルネットワークを訓練する場合には、絶対値が十分に小さくなるような一様分布に従う擬似乱数を用いて、初期化しても検証時の性能に大きな差は見られなかったと報告している。

本論文でもこれに従い、単語ベクトルは無作為に初期化した。また訓練集合における頻度が非常に小さいものおよび訓練時に現れない単語全ては Out-of-vocabulary としてベクトルを一つ用意した。

3.4 行列パラメータおよび共通の重みの正則化

文脈によらない、2 つの句ベクトルを受け取り、元の次元にうつすための重み W は、それぞれの句からの寄与が同等になるような点からの距離に関する L_2 正則化を行った。

$$\Omega(W) = \tfrac{1}{2}(\|W_l - 0.5I\|_2^2 + \|W_r - 0.5I\|_2^2) \qquad (8)$$
$$where \qquad W = \begin{pmatrix} W_l \\ W_r \end{pmatrix}$$

ここで I は単位行列を表す。一方、行列パラメータに関しては単位行列からの距離を考慮した。

4. 評価実験

4.1 感情分析タスク

拡張したモデルの評価は、比較のために Stanford Sentiment Treebank[10] を用いて行う。これは rottentomatoes.com 上の映画レビューを集めて Pang らが公開した既存のデータセット [7] を元に、Socher らが Stanford Parser[4] によって得られた解析木を二分木に変形したのち、Amazon Mechanical Turk を利用して、その木の頂点として現れる句に対して中立を含む否定的から肯定的の 5 段階のラベルを付加して公開した。

このデータセットは訓練 (8544)/開発 (1101)/テスト (2210) 集合に分割されて配布されていて、それぞれを目的通りに利用した。

4.2 比較する対象

評価にあたって、感情分析タスクに用いられるいくつかのモデルとともに、拡張手法を比較する。

RNN 行列を割り当てないが、構成的に句のベクトルを再帰的ニューラルネットワークによって計算する

MV-RNN RNN を単語の行列を扱うように拡張 [9]

AdaMC-RNN RNN における構成のための関数を適応的に拡張したもの [1]

4.3 極性の予測

はじめに Stanford Sentiment Treeebank をデータセットとして用いて訓練された、それぞれのモデルの性能を拡張手法とともに 2 に示した。

比較対象としたいずれのモデルも AdaGrad を用いて訓練された。[10], [1] それぞれのモデルでの単語ベクトルの次元は表の通りである。また AdaMC-RNN については適応させる関数の数は 15 に設定した際の精度である。

(1) "being funny"
(2) "Your response"
(3) "The notion"
(4) "special-effects-laden extravaganzas"
(5) "The attraction"
(6) "his character"
(7) "staggeringly well-produced"
(8) "for Godard"
(9) "lot more"
(10)"process or"

図 3: 作用の点で類似した句

4.4 類似した働きをする句の分析

MV-RNN[9] や AdaMC-RNN[1] と比べて、拡張手法では極性の予測を行う上である句が他の句にどのように働くかという点について、分析を行うのが容易になるという利点がある。実際にそのような例として "wildy fascinating" という 2-gram の句と行列の割り当てにおける割合が類似しているものを類似度の順に 3 に提示した。ここで行列の割り当てにおける割合の近さをはかるために、カルバックーライブラー情報量を対称となるように変更したものを用いた。

$$D(p,q) = \frac{1}{2}(D_K L(p||q) + D_K L(q||p)) \qquad (9)$$

where

$$D_{KL}(p||q) = \sum_j p_j \log \frac{p_j}{q_j}$$

ここで p, q は K 個の正の成分を持つ、和が 1 となるようなベクトルである。

例には単に肯定的な意味の句だけでなく、度合いを強めるような句が含まれていることがわかる。

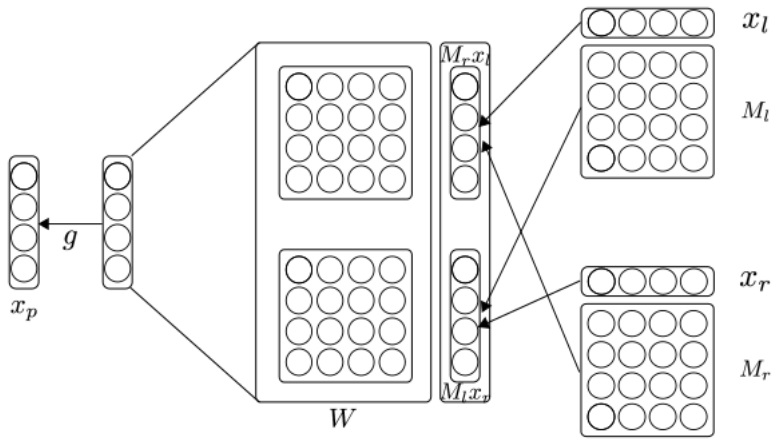
参考文献

[1] Dong, L., Wei, F., Zhou, M. and Xu, K.: Adaptive Multi-Compositionality for Recursive Neural Models with Applications to Sentiment Analysis, *AAAI 2014* (2014).
[2] Duchi, J., Hazan, E., and Singer, Y.: Adaptive Subgradient Methods for Online Learning and Stochastic Optimization, *Journal of Machine Learning Research* (2011).
[3] Fyshe, A., Wehbe, L., Talukdar, P., Murphy, B. and Mitchell, T.: A Compositional and Interpretable Semantic Space, *Proceedings of the 2015 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies* (2015).
[4] Klein, D. and D. Manning, C.: Accurate Unlexicalized Parsing, *Proceedings of the 41st Meeting of the Association for Computational Linguistics* (2003).
[5] Levy, O. and Goldberg, Y.: Linguistic Regularities in Sparse and Explicit Word Representations, *CoNLL 2014* (2014).
[6] Marco, B. and Zamparelli, R.: Nouns are vectors, adjectives are matrices: Representing adjective-noun constructions in semantic space, *Conference on Empirical Methods for Natural Language Processing* (2010).
[7] Pang, B. and Lee, L.: Opinion Mining and Sentiment Analysis, *Proceedings of the 43rd Annual Meeting on Association for Computational Linguistics* (2005).
[8] Schnabel, T., Labutov, I., Mimno, D. and Joachims, T.: Evaluation methods for unsupervised word embeddings, *Conference on Empirical Methods for Natural Language Processing* (2015).
[9] Socher, R., Huval, B., D. Manning, C. and Y. Ng, A.: Word representations: A simple and general method for semi-supervised learning, *Conference on Empirical Methods for Natural Language Processing* (2012).
[10] Socher, R., Perelygin, A., Y. Wu, J. and Chuang, J.: Recursive Deep Models for Semantic Compositionality Over a Sentiment Treebank, *Conference on Empirical Methods for Natural Language Processing* (2013).
[11] Turian, J., Ratinov, L. and Bengio, Y.: Word representations: A simple and general method for semi-supervised learning, *Proceedings of the 48th AAnnual Meeting of the Association for Computational Linguistics* (2010).

正誤表

下記に提出した予稿に対する訂正と補足を追加する。またこの差分を適用した予稿は *https://drive.google.com/open?id=0B08ah1go3XEXSEtXenVxV2xHRmc* **に公開する。**

訂正 2 **ページ** 3.1 **行列とベクトルの対による句の分散表現** **第** 3 **パラグラフ**	**ある句に対応する内部ノードに着目したとき、その句に対応するベクトルは次のようにある句を構成するような、**2 **つの句に対応する行列とベクトルの対** $(x_l, M_l), (x_r, M_r)$ **をととって元のベクトルと同じ次元にうつす関数** f **を用いて下図に示すように計算する。**

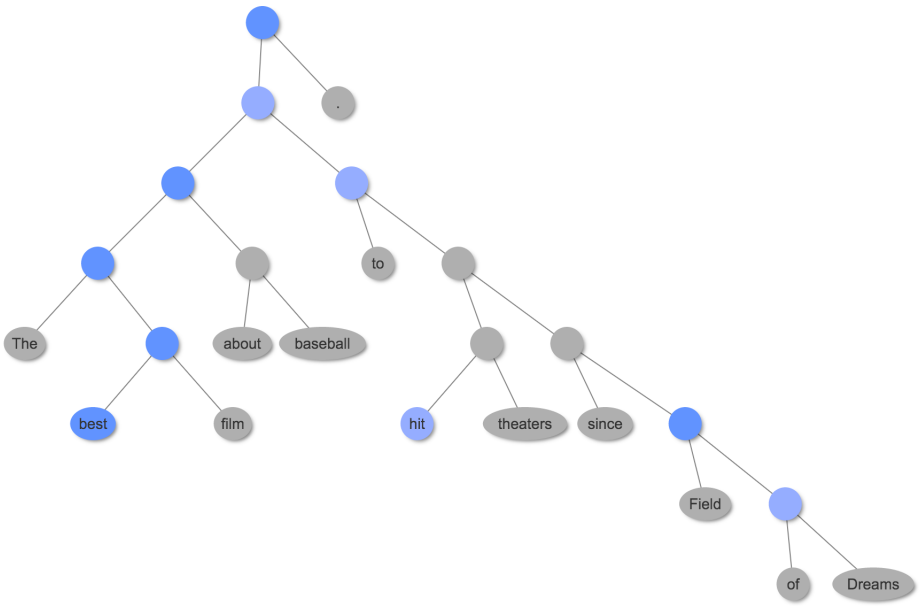


訂正 3 **ページ** 3.3.1 **誤差関数に現れる式**	$$E(\Theta) = E^{(0)}(\Theta) + E^{(1)}(\Theta) + \Omega(\Theta) \quad (1)$$ **ここで** $\Omega(\Theta)$ **は正則化項を表す。正則化項については** 3.3.1 **に続く節で説明している。**
挿入 3 **ページ** 3.3.1 **誤差関数の末尾**	**提案手法を含む一連のの再帰ニューラルネットワークに基づくモデルの訓練では、**Backpropagation Through Structure[2] **を用いて勾配を計算する。**

訂正
3 ページ
4.1 感情分析タスク

これは rottentomatoes.com 上の映画レビューを集めて Pang らが公開した既存のデータセット [5] を元に、Socher らが Stanford Parser[3] によって得られた解析木を二分木に変形したのち、Amazon Mechanical Turk を利用して、その木の頂点として現れる句に対してラベルを付加して公開した。ここでこのデータセットに含まれるラベルは、中立を含む否定的から肯定的の 5 クラスからなり、句の書き手の主観的な態度としてもっともらしいと思われるものを 3∼6 人のアノテーターが選んだ。

データセットに含まれる文の例として、句にラベルのついた、ある文の解析木を下図に示した。この図における頂点の色は青色/灰色がそれぞれ肯定と中立のラベルに対応し、青色についてはその 2 段階の濃さによって肯定の程度を区別する。また図には現れていないが、実際のデータセットには否定的なラベルのついた文も含まれる。
このデータセットは訓練 (8544)/開発 (1101)/テスト (2210) 集合に分割されて配布されていて、これを用いて既存手法と同様の設定で実験を行った。



挿入
3 **ページ**
4.2 **極性の予測の末尾**

この表から読み取れるように、構成的に句を計算するのに用いる関数を増やして重み付けた AdaMC-RNN **や拡張の際に参考にした** MV-RNN **と同等の性能を達成するには至らなかった。**
その要因として考えられるのは、AdaMC-RNN **で重み付けられているのが複数の関数である一方で、提案手法では行列を重み付けていることにより、訓練の過程で行列パラメータに対して摂動を加えた際に起きる、誤差への影響が大きくなってしまった可能性が考えられる。**

挿入
3 **ページ**
4.4 **類似した働きをする句の分析の直前**

再帰ニューラルネットワークを教師あり学習に用いるようないくつかのモデル [6], [1] **では、単語ベクトルは事前に用意した、教師なし学習によるものを用いて初期化し、その後にパラメータとして後の訓練で扱われる。**
いずれの研究でも語彙のサイズが性能に与える影響について詳しい言及はなく、この点について提案手法のもとで語彙に含まれる、頻度の低い単語の扱いによって性能にどのような影響を与えるかを実験によって評価した。
実験では、設定した閾値よりも低い頻度でしか現れない単語について、未知語として扱った。単語ベクトルの次元は 5 **に設定し、行列の成分の数についても** 5 **とした。この設定のもとで、根での予測をもとにした精度を下図に示した。**
また訓練集合の語彙に関して、その出現頻度が閾値よりも低い単語を数えて下図に併記した。語彙の影響を調べる際には単語ベクトルは MV-RNN **や** AdaMC **に比べて次元を小さく設定して行ったが、その場合には低頻度語の影響によって開発およびテスト集合における性能が低下していることがわかる。**
実験で用いたデータセットは映画レビューに関する文を含んでいるが、この中には映画の監督の名前やタイトルといった頻度の低い単語が含まれる。頻度の閾値を小さく設定するとそうした単語の影響によって、性能が劣化したと考えられる。このような現象は特にカテゴリをまたいだ極性予測のようなタスクを扱う場合には、性能への影響を与える可能性を排除できない。

閾値	**訓練** (%)	**開発** (%)	**テスト** (%)	**語彙の大きさ**
2	49.52	38.51	38.14	9543
3	48.68	41.59	38.50	12411
4	51.31	42.41	40.67	13807
5	49.07	43.86	42.03	14726
全て	-	-	-	18278

挿入
3 **ページ**
4.4 **類似した働きをする句の分析の末尾**

この”wildly fascinating”**はもともと訓練集合に**”A winning and wildly fascinating work.”**として現れており、名詞を肯定的に修飾している句である。一方で提案手法によって類似していると予測された、**”a beautifully”**や**”a nicely”**も文法的に同様に用いられる** 2-gram **である。**
先行研究である、AdaMC-RNN **や** RNTN **では構成的に句のベクトルを得るための手段として、複数の重み付けられた関数あるいはパラメータを増やした関数を一つ用いることによって、少数の文を用いて計算した勾配による訓練を可能にするとともに性能を向上させたが、文に対する予測の上で句がどのような役割を果たしているかを分析するのは容易ではなかった。**

2-gram	divergence	**コサイン類似度**
a beautifully	1.1821e-10	0.954
layered richness	2.0241e-10	0.905
reasonably intelligent	4.9271e-10	0.737
ultimately worthwhile	5.6434e-10	0.968
Eye See	1.8284e-09	0.874
chilling style	5.1045e-09	0.636
distinct rarity	5.9595e-09	0.695
a nicely	8.7071e-09	0.750
thoroughly entertaining	8.8592e-09	0.993
Sometimes entertaining	1.1533e-08	0.873

挿入
4 **ページ**
5. **おわりにの末尾**

今後の課題としては訓練集合には同一の句は現れないが、それに類似した句は現れる場合には、ラベルがついていない文に対する解析木に現れる句の情報を利用することによって、性能の改善につながる可能性がある。また前後の文脈に基づいて単語の素性を得る Mikolov **らの手法における負例の考慮** [4] **と同様に、訓練中に現れる解析木の一部を改変して負例として補助的に用いることで、構成的な句の計算における相互作用をよりよく捉えるという点でも工夫の余地はあると考えられる。**

参考文献

[1] Dong, L., Wei, F., Zhou, M. and Xu, K.: Adaptive Multi-Compositionality for Recursive Neural Models with Applications to Sentiment Analysis, *AAAI 2014* (2014).

[2] Goller, C. and Kchler, A.: Learning Task-Dependent Distributed Representations by Backpropagation Through Structure, *Proc. of ICNN96* (1996).

[3] Klein, D. and D. Manning, C.: Accurate Unlexicalized Parsing, *Proceedings of the 41st Meeting of the Association for Computational Linguistics* (2003).

[4] Mikolov, T., Sutskever, I., Chen, K., Corrado, G. and Dean, J.: Distributed Representations of Words and Phrases and their Compositionality, *Advances in Neural Information Processing Systems 26 (NIPS 2013)* (2013).

[5] Pang, B. and Lee, L.: Opinion Mining and Sentiment Analysis, *Proceedings of the 43rd Annual Meeting on Association for Computational Linguistics* (2005).

[6] Socher, R., Huval, B., D. Manning, C. and Y. Ng, A.: Word representations: A simple and general method for semi-supervised learning, *Conference on Empirical Methods for Natural Language Processing* (2012).