半教師有り再帰的オートエンコーダの生体医学分野のテキストへの適用

橋本 和真 †　　三輪 誠 ‡　　鶴岡 慶雅 †　　近山 隆 †
† 東京大学 工学部電子情報工学科, ‡ マンチェスター大学
{hassy, tsuruoka, chikayama}@logos.t.u-tokyo.ac.jp
makoto.miwa@manchester.ac.uk

1　はじめに

近年, 再帰的ニューラルネットワーク (Recursive Neural Netowork, RNN) による単語, 句, 文のベクトル表現の学習が注目されている [1]. 単語のベクトル表現の研究は多く行われているが, RNN では再帰的にニューラルネットワークを適用することにより, 単語だけでなく, 単語によって構成される句や文も同じ次元のベクトルで表現することができる. それにより, パラフレーズ認識, 文の感情表現の推定などの様々な自然言語処理のタスクに RNN が適用されている.

一方で, 生体医学分野の文献数が急増しつつあり, それらの情報を人手で整理し, 利用可能にすることは難しくなってきている. これを解決するために, 自然言語処理による情報抽出などの試みが行われている.

そこで本研究では, 生体医学分野のテキストにおける RNN の適用を試みる. まず関連研究として, 再帰的オートエンコーダ (Recursive Autoencoder, RAE) の説明を 2 章で行う. 続いて 3 章で, 文タイプを分類するタスクと, タンパク質間相互作用を判定するタスクに RAE を適用する方法を提案し, それらの手法がタスク依存でなく幅広く適用可能であることも示す. そして 4 章で, 実験結果を示す. 最後に, 5 章で今後の研究の方針などについて述べる.

2　再帰的アートエンコーダ (RAE)

この章では, Socher らによって提案された RAE の学習法 [1] を説明する.

2.1　単語のベクトル表現

ここでは, 各単語 (トークン) は n 次元の実数値ベクトルで表現されているとし, それらを並べて構成する行列を $L \in \mathcal{R}^{n \times |V|}$ と表す. ただし, $|V|$ は語彙数とする.

2.2　オートエンコーダの学習

2 つの単語 (ベクトル表現を $x_1, x_2 \in \mathcal{R}^n$ とする) からなる句のベクトル表現 p をオートエンコーダによって求める. エンコード用のパラメータを $W_e \in \mathcal{R}^{n \times 2n}, b_e \in \mathcal{R}^n$ として,

$$p = f(W_e[x_1; x_2] + b_e) \tag{1}$$

と計算する. ただし, $[x_1; x_2] \in \mathcal{R}^{2n}$ は x_1 と x_2 の連結を表し, 非線形関数 f は $tanh$ などを用いる. 式 (1) により p は n 次元ベクトルで表現されるが, x_1, x_2 を再構成することでその表現の良さを評価する. p から x_1, x_2 を再構成するには, デコード用のパラメータ $W_d \in \mathcal{R}^{2n \times n}, b_d \in \mathcal{R}^{2n}$ を用いて

$$[x_1'; x_2'] = f(W_d p + b_d) \tag{2}$$

と計算する. そして, 再構成誤差

$$\frac{1}{2}\|[x_1; x_2] - [x_1'; x_2']\|^2 \tag{3}$$

を最小化するようにネットワークのパラメータ W_e, b_e, W_d, b_d を調整する.

2.3　教師無し RAE の学習

RAE では 2 分木の構造を用いるが, ここでは各文に対して構文解析器などによって 2 分木の構造が与えられているとする. 図 1 のように 2 分木の構造に従って, 再帰的に 2.2 節の通りに単語から句のベクトル表現を計算していき, 文全体のベクトル表現が得られるまで続ける. この時, ある文の木 t 全体の再構成誤差

$E_{rec}(t)$ は, t に属する全ての非終端ノード n の再構成誤差の和

$$E_{rec}(t) = \sum_{n \in t} \frac{1}{2} \| [x_1{}^{(n)}; x_2{}^{(n)}] - [x'_1{}^{(n)}; x'_2{}^{(n)}] \|^2 \quad (4)$$

で計算される. ただし, $x_1{}^{(n)}, x_2{}^{(n)}$ は非終端ノード n の子ノードのベクトル表現であり, $x'_1{}^{(n)}, x'_2{}^{(n)}$ はそれらを式 (2) に従って再構成したものとする. 以上から, 全訓練データの集合を T とすると最小化する目的関数 J_{rec} は,

$$J_{rec} = \frac{1}{N} \sum_{t \in T} E_{rec}(t) + \frac{\lambda}{2} \|\theta\|^2 \quad (5)$$

となる. ただし, N は T に属する非終端ノードの総数であり, $\theta = (W_e, b_e, W_d, b_d, L)$ はパラメータである. また, λ は正則化項の係数である. この勾配は,

$$\frac{\partial J_{rec}}{\partial \theta} = \frac{1}{N} \sum_{t \in T} \frac{\partial E_{rec}(t)}{\partial \theta} + \lambda \theta \quad (6)$$

となる. これを用いて, L-BFGS などの最適化アルゴリズムによってパラメータ θ を決定する.

2.4 半教師有り RAE の学習

2.3 節までは教師無し学習であるが, 教師情報を利用して半教師有り学習を行うことができる. 任意のノードにソフトマックス層を設定し, 教師情報との誤差の総和を式 (5) に加えることで実現される.

$$J = \alpha J_{rec} + (1 - \alpha) J_{sup} \quad (7)$$

ただし, J_{sup} はソフトマックス層の誤差の総和を最小化する目的関数であり, $\alpha \in [0, 1]$ は再構成誤差に対する重みである. 特に $\alpha = 0$ の時は, 単なる教師有り学習 (RNN) になる.

3 生体医学分野の自然言語処理のタスクへの RAE の適用

本章では, 2 つの生体医学分野の自然言語処理のタスクに RAE を適用する方法を説明する. ただし, 後に示すように本手法はタスク依存ではない.

3.1 文タイプの分類

科学的な論文の要約を構成する文のカテゴリを分類する研究が行われている [4]. そこで本研究では, MEDLINE の文献の要約の各文に, 4 種類の文タイプ (OBJECTIVE, METHOD, RESULTS, CONCLUSIONS) の情報がついたデータを用いて RAE の学習を行い, 文タイプの分類を試みる. 具体的には, 各 RAE の根ノードにソフトマックス層を乗せて教師情報を与え, 4 クラス分類問題として扱う. その際, 単に分類精度を評価するだけでなく, ネットワークの学習によって単語のベクトル表現が受ける影響が, 経験的な直感に従うことも確認する.

本研究では生体医学分野の文献のデータを用いるが, この手法は一般の文献にも適用でき, 当然, 文タイプの分類というタスクのために限定されるものではない.

3.2 タンパク質間相互作用 (PPI) の判定

文中のタンパク質間相互作用 (Protein-Protein Interaction, PPI) を判定するタスク [3] に RAE を適用する. このタスクは, 文中で指定された 2 つのタンパク質のペアが相互作用するかどうかを判定することを目的としている. この問題に RAE を適用するために, 図 1 のように, 2 つのタンパク質に対応する葉ノードを含む最小の部分木の根ノード (以下ではカバーノードと呼ぶ) に, 相互作用しているかどうかの教師情報 (0, 1) を与える. また, 図 1 に示す通り, 各タンパク質は区別せず, 共通のシンボルとして単語 “**PROTEIN**” を単語リストに追加する.

本研究では PPI を判定するタスクに取り組むが, この手法は他の関係抽出のタスクにも同様に適用可能である. ただし, 2 つの単語が同じ文中に存在するという前提が必要となる. 文中で特定の単語のペアが指定されると, 自動的にそれらのカバーノードを特定して教師情報を与える.

ただしこの手法には, 学習時に複数のペアがカバーノードを共有する可能性があるという点で問題が生じうる. 例えば, あるカバーノードが複数のタンパク質のペアに対応しており, 正例と負例の両方を含む場合に, 学習時にそのノードに教師情報として 0 または 1 を割り当てるのは望ましくない. そこで, あるカバーノードに対応する正例の数を P, 負例の数を N とするとき, 教師情報を式 (8) のように計算する.

$$\frac{P}{P+N} \times 1 + \frac{N}{P+N} \times 0 \quad (8)$$

これは, PPI の判定のような 2 値分類でなく, 一般の多クラス分類の際も同様に扱うことができる.

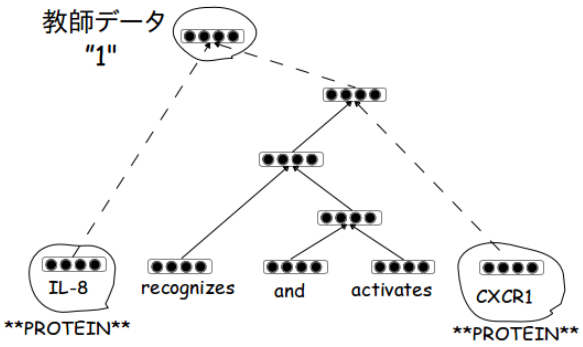

図 1: 2 つのタンパク質のカバーノード

4 評価

本研究の RAE の学習においては, 2 分木の構造として構文解析器 enju[1]の解析結果を用いた. ただし, enju

4.1 文タイプの分類

4.1.1 単語のベクトル表現の学習

単語のベクトル表現の学習結果の例を表 1 に示す. 単語間の距離はコサイン距離で測り, 文タイプの予測には学習されたソフトマックス層を用いた. ただし, 各単語のベクトル表現の初期値には乱数を用いて次元は 100 とした. また, 教師情報の影響を確認するために $\alpha = 0.01$ とした. ソフトマックス層の出力値は, 各文タイプに所属する度合いを表している. つまり, 値が大きいほどその文タイプに出現しやすい.

表 1 を見ると, 各文タイプにおいて出現しやすい単語のベクトル表現の距離が近くなっていることがわかる. そのなかで, diseases と disorders (次点が cancer), reduced と decreased, DNA と RNA のように, 単語の意味的な近さや関連性も実現されている. また, 学習に用いたデータが論文の要約であることを考えると, 現在形の are と is が METHOD に出現しやすく, 過去形の was と were が RESULTS に出現しやすいという結果は自然である.

4.1.2 文タイプの分類精度

RAE を用いた文タイプの分類精度を評価した. ただし, 各単語のベクトル表現の次元は 50 とし, 初期化には Collobert と Weston らのモデル [2] で学習された 50 次元のベクトル表現を用いた. RAE による文

[1]http://www.nactem.ac.uk/enju/index.ja.html

表 1: 単語の最近傍と, 文タイプ予測 (OBJECTIVE, METHOD, RESULTS, CONCLUSIONS の順) の例

単語	最近傍	ソフトマックス層出力			
Previously	Recently	83.6	5.6	5.2	5.6
suggests	indicates	21.2	45.3	17.7	15.9
revealed	showed	23.3	11.5	44.0	21.1
Overall	Lastly	5.6	2.9	29.6	61.9
diseases	disorders	35.3	31.1	18.3	15.3
reduced	decreased	21.5	15.0	42.0	21.6
DNA	RNA	8.3	76.1	10.0	5.7
are	is	18.5	37.9	19.3	24.3
was	were	24.9	20.5	42.5	12.1

タイプの予測には, 各 RAE の根ノードのベクトル表現 (文全体を表すベクトル) を入力としたソフトマックス層の出力値を用いた. つまり, 出力された 4 次元ベクトルの要素のうち最大値に対応するタイプを推定タイプとした. 実験の際には, 用意したデータから訓練データを 150,000 文, 開発データを 50,000 文, そしてテストデータを 50,000 文それぞれサンプリングした. ただし, 各データ集合における各タイプの文の数は同じになるようにし, 開発データを用いて α, λ などのパラメータをチューニングした. ただし, 今回はテストデータによる評価は行わなかった.

表 2 に結果を示す. 提案手法による結果を評価するために, bag-of-words (BoW) を特徴ベクトルとして, SVM で学習した分類器による分類精度を示す. また, 文タイプごとにデータ数に偏りが無いのでランダムに答えた結果は 25% になる. RAE による結果は, ランダムに答えるシステムに比べると学習の効果が見て取れるが, 非常に単純な BoW と SVM による分類と同程度の分類精度しか出ていない. 原因としては, 長い文のベクトル表現がうまくできていない可能性が考えられ, その場合, 文全体のベクトル表現のみを用いて分類を行う本手法にはある程度の限界があることになる. 残り 30% のデータで正解できない理由の原因の考察を行い, より長い文にも対応しうる学習方法を考える必要がある.

4.2 タンパク質間相互作用の判定

先行研究 [3] と同様に, PPI のデータセットとして, LLL, IEPA, HPRD50, BioInfer, AIMed の 5 つのコーパスを用いた. それぞれ, 正例または負例のタンパク

表 2: 文タイプの分類の正解率 (%)

手法	開発データ	テストデータ
ランダム	25.0	25.0
BoW+SVM	69.1	-
RAE	68.4	-

表 3: PPI の各データセットのスコア (%) (括弧内は標準偏差)

	RAE		[3]	
データ	F	AUC	F	AUC
LLL	80.8 (13.5)	85.5 (10.6)	80.5	86.0
IEPA	68.7 (7.6)	77.6 (8.2)	74.4	85.6
HPRD50	64.1 (9.8)	75.1 (9.0)	69.7	82.8
BioInfer	52.8 (6.2)	73.0 (5.2)	67.6	86.1
AIMed	40.2 (10.1)	71.7 (6.5)	64.2	89.1

質のペアが与えられており, これらのデータセットにおいて標準的に行われている 10 分割交差検定により分類精度を評価した. その際, F 値と AUC 値を評価の指標として用いた.

表 3 に結果を示す. 三輪らの研究結果 [3] を比較対象として挙げたが, LLL に関しては同程度の結果が出ている. その他の 4 つのデータセットに関しては結果が劣っており, 特に AIMed は著しく劣っている. まず考えられる原因としては, 3.2 で述べたように, 1 つのカバーノードが複数のタンパク質ペアに対応しうることである. そこで, 各データセットについて, 正例と負例が競合しているカバーノードに属するタンパク質ペアの数を表 4 に示す. ただし, enju による解析が失敗した文に含まれる事例は除いてあるので, 実際に与えられたデータ数より少なくなっている. 表 4 を見ると, 多くのタンパク質ペアに対して 0 または 1 というはっきりとした教師情報を与えられていないことがわかる. よって, このようなタスクにおいては, 本手法が必ずしも有効ではないと言えるので, さらに工夫した学習方法を考える必要がある.

5 おわりに

本研究では, 生体医学分野の自然言語処理の 2 つタスクへの RAE の適用を試みた. 文タイプの分類においては, 直感的に良いと思われる単語表現の学習に成功した一方で, 分類精度という点からは優れた結果が得られなかった. また, PPI においても, 分類精度が優れていなかった.

表 4: 正例と負例の競合があるタンパク質ペア数

データ	競合ペア数	全ペア数	割合 (%)
LLL	77	330	23.3
IEPA	64	433	14.8
HPRD50	56	816	6.9
BioInfer	1687	9654	17.5
AIMed	910	5668	16.1

これらの結果は RAE 単体のものなので, 今後の課題として他のシステムの特徴と組み合わせることによる性能向上の可能性が考えられる. さらに, 教師情報のついていない大量の文を混ぜて半教師有り学習を行うことや, PPI の 5 つのコーパスを組み合わせて学習することなども今後の研究課題とする.

参考文献

[1] R. Socher, Jeffrey Pennington, Eric Huang, A. Y. Ng, and Christopher D. Manning. Semi-Supervised Recursive Autoencoders for Predicting Sentiment Distributions. EMNLP, 2011.

[2] Ronan Collobert and Jason Weston. A unified architecture for natural language processing: Deep neural networks with multitask learning. ICML, 2008

[3] Makoto Miwa, Rune Stre, Yusuke Miyao, and Jun'ichi Tsujii. A Rich Feature Vector for Protein-Protein Interaction Extraction from Multiple Corpora. In Proc. of the 2009 Conference on Empirical Methods in Natural Language Processing. Singapore, pp. 121–130, ACL, 2009.

[4] K. Hirohata, N. Okazaki, S. Ananiadou, and M. Ishizuka. Identifying sections in scientic abstracts using conditional random elds. In Proc. of the IJCNLP 2008.