Long Short-Term Memory に基づく Recurrent Auto-Encoder を用いた文の分散表現獲得手法に対する Attention 機構の導入

1　はじめに

近年，計算機の著しい発達に伴い，言葉や画像といった離散的な記号概念の分散表現を獲得する研究が盛んになされている．得られた分散表現は人工知能研究におけるさまざまなタスクに対して適用されるが，その精度は分散表現の性能に大きく依存する．それゆえに，分散表現の性能向上は人工知能研究の発展のために極めて重要な事項であるといえる．

自然言語処理の分野においては現状として，単語の分散表現獲得手法については Word2Vec[1] のような複数のタスクに対して高い性能が認められている優れた手法が開発されている．その応用として，文の分散表現の獲得手法に関するいくつかの先行研究が存在するが，いまだに決定的な手段は確立されているとは言い難い．

上記の事実を受けて本研究では，既存の文の分散表現獲得手法の改良を目的として，Long Short-Term Memory (LSTM) に基づく Recurrent Auto-Encoder (RAE) を用いたモデルに対して Attention 機構 [2] を導入する．また，獲得した分散表現を用いた文の連続性識別の実験を通して，それらの性能を Attention 機構の有無の観点から相対的に評価し，その導入の有効性を判断する．

2　Attention 機構

機械翻訳のタスクに対して考案された Encoder-Decoder モデルは可変長の文を固定長のベクトルにエンコードするため，長い入力文になるほど隠れ層のノード数が不足し，学習が難しくなる問題がある．そこで Bahdanau らにより提案されたのが Encoder 側で入力文の各単語の荷重を決定してエンコードすべき場所を制御する Attention 機構である．

Attention 機構では入力文の各単語 x_i に対する荷重 $\boldsymbol{\alpha}_t$ を計算することで，コンテキストベクトル $\boldsymbol{c}_t$ を得る．$\boldsymbol{\alpha}_t$ は Encoder で出力される全時刻の隠れ状態ベクトル $\overline{\boldsymbol{h}_i}$ と Decoder から出力される各時刻のベクトル $\boldsymbol{h}_t$ から算出される類似度を示す score を正規化することにより得られる．式 (1), (2), (3) にその具体的な算出方法を示す．

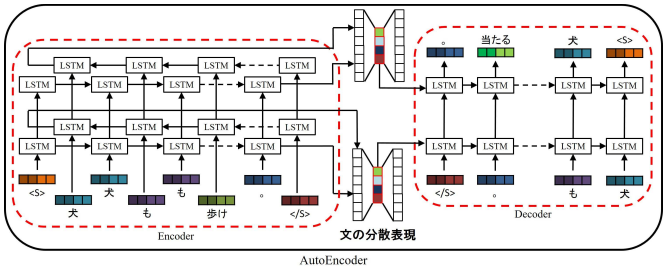

図 1: 先行研究のモデル概略図

$$\boldsymbol{\alpha}_t(i) = \frac{\exp(\mathrm{score}(\boldsymbol{h}_t, \overline{\boldsymbol{h}_i}))}{\sum_{j=1}^{n} \exp(\mathrm{score}(\boldsymbol{h}_t, \overline{\boldsymbol{h}_j}))} \tag{1}$$

$$\boldsymbol{c}_t = \sum_{i=1}^{n} \boldsymbol{\alpha}_t(i)\overline{\boldsymbol{h}_i} \tag{2}$$

$$\mathrm{score}(\boldsymbol{h_t}, \overline{\boldsymbol{h_i}}) = \boldsymbol{h}_t^{\top}\overline{\boldsymbol{h}_i} \tag{3}$$

3　先行研究

文の分散表現獲得手法の先行研究として，福田らにより提案された手法がある [3]．この手法では，自然言語処理の分野で有効性が示されている Encoder-Decoder モデルにおいて，Encoder の入力と Decoder の出力を同一にすることで RAE として用いている．その Encoder と Decoder の連結部における中間表現を対象文の分散表現とする．

この手法では，時系列データに対する有効性が示されている LSTM を用いて Encoder と Decoder のネットワークを構築している．LSTM を多層構造にして用いることで，各隠れ状態ベクトルに異なる複数の情報を保存することができると期待される．

また，Encoder 部分の LSTM を双方向にして用い，順方向と逆方向からデータを入力することで，直前までの情報だけでなく文全体の情報を考慮することを可能にした [4]．図 1 にその概略図を示す．

4　提案手法

本研究では，先行研究として示した 2 層の単方向 LSTM に基づく RAE を用いた文の分散表現の獲得

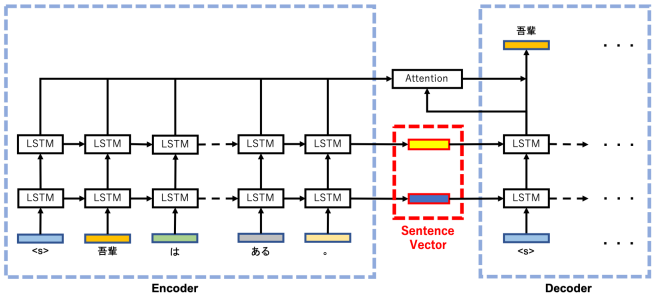

図 2: 提案手法のモデル概略図

手法に対して Attention 機構を導入したモデルを提案する．図 2 にそのモデル構造を示す．提案モデルでは最終的に得られたベクトルと，Decoder への入力に対応する単語の分散表現との誤差を最小化するように学習を進める．

先行研究および提案手法では，対象文を形態素解析により分割し，各単語の分散表現を入力および出力として用いる．形態素解析には日本語形態素解析エンジンである MeCab[1] を使用した．

また，単語の分散表現獲得手法としては，Keras により実装した Word2Vec を採用した．表 1 にその設定値を示す．学習用のデータとしては，日本語版「ウィキペディア (Wikipedia): フリー百科事典」[2] および，電子図書館である「青空文庫」[3]，小説投稿サイト「小説家になろう」[4]から収集したテキストデータを用いた．ただし，文章中に現れる頻度がしきい値以下の単語については未知語として処理した．

5　数値実験

提案手法の有効性を確認するため，獲得した文の分散表現を用いて文の連続性を識別する実験を通して，提案手法と従来手法とを比較した．従来手法として用いたのは，先に示した福田の手法と，Gensim により実装した Doc2Vec[5] である．

5.1　実験手順

以下に実験の手順を示す．

1. 夏目漱石の小説「坊っちゃん」「草枕」「三四郎」「それから」「門」「彼岸過迄」「行人」「こころ」「明暗」の 9 作品から，同段落内の連続する 3 文の組を抽出し，連続文セット $\boldsymbol{c}$ とする．また，それらの集合を連続文セットの集合 $\boldsymbol{C}_{\text{train}}$ とする．ただし各 $\boldsymbol{c}$ に含まれる文は，以下の各条件に従うものとする．
   1) n_{min} 単語以上 n_{max} 単語以下で構成されること．
   2) 感嘆符や疑問符のような記号を含まないこと．
   3) 各文はただ 1 つの $\boldsymbol{c}$ にのみ含まれること．
2. 谷崎潤一郎の小説「痴人の愛」「卍」，芥川龍之介の小説「河童」，宮沢賢治の小説「銀河鉄道の夜」「風の又三郎」，太宰治の小説「斜陽」「人間失格」の 7 作品から 1. の操作における 1), 2) の条件に合致する文を抽出し，その集合を $\boldsymbol{S}$ とする．
3. 1. の操作で生成した $\boldsymbol{C}_{\text{train}}$ に含まれる各 $\boldsymbol{c}$ の 2 文目を，2. の操作で生成した $\boldsymbol{S}$ の各文と置換することで非連続文セット $\boldsymbol{w}$ を生成し，その集合を $\boldsymbol{W}_{\text{train}}$ とする．
4. 夏目漱石の小説「吾輩は猫である」に対して，1. および 3. の各操作を施して $\boldsymbol{C}_{\text{test}}, \boldsymbol{W}_{\text{test}}$ を生成する．ただし，$\boldsymbol{W}_{\text{train}}$ の生成のために用いた $\boldsymbol{S}$ の各文は，$\boldsymbol{W}_{\text{test}}$ の生成には用いないものとする．
5. 4. で生成した各データの文についてそれぞれ分散表現を獲得し，それらを LSTM に入力し文の連続性に対する分類問題を解く．

表 1: Word2Vec の設定

パラメータ名	値
モデル	Skip-gram
高速化手法	Negative Sampling
文脈窓	10
ベクトルサイズ	200
サンプリングサイズ	15
Epoch 数	50
最適化手法	Adam
学習率	0.0025
頻度のしきい値	10
語彙数	300,000

[1] MeCab: https://taku910.github.io/mecab
[2] ウィキペディア: https://ja.wikipedia.org/
[3] 青空文庫: https://www.aozora.gr.jp
[4] 小説家になろう: https://syosetu.com

表 2: 提案手法および従来手法の設定

パラメータ名	値
$(n_{\min}, n_{\max})$	(12, 30)
$(m_{\min}, m_{\max})$	(10, 80)
学習データ数	22,415,243 文
Encoder 構造	2 層 LSTM
Encoder ユニット数 (第 1 層)	200
Encoder ユニット数 (第 2 層)	200
Decoder 構造	2 層 LSTM
Decoder ユニット数 (第 1 層)	200
Decoder ユニット数 (第 2 層)	200
損失関数	平均二乗誤差
Epoch 数	5
バッチサイズ	1,536
最適化手法	Adam
初期学習率	1.0×10^{-5}
Doc2Vec モデル	PV-DM
文脈窓	10
ベクトルサイズ	200
Epoch 数	5
頻度のしきい値	10

表 3: 各データのサイズ

データ	サイズ
$\boldsymbol{C}_{\text{train}} : \boldsymbol{W}_{\text{train}}$	3,092 sets : 3,092 sets
$\boldsymbol{C}_{\text{test}} : \boldsymbol{W}_{\text{test}}$	370 sets : 370 sets

5.2 実験条件

表 2 に提案手法および従来手法の設定値を示す．いずれの手法も，Keras により実装した．なお，学習データとしては Word2Vec の学習に用いたテキストデータにおいて，$m_{\min}$ 単語以上 $m_{\max}$ 単語以下の文を用いた．また表 3 に，分類問題として用いた各データのサイズを示す．

5.3 結果と考察

表 4 に，10 回の試行により得られた各手法に対する各精度の平均とその標準偏差を示す．提案手法の各層の分散表現に対して得られた精度は，いずれも各従来手法に対する精度を上回る結果となった．このことから，Attention 機構を導入することで文の連続性を判断する上でより重要な情報を含んだ分散表現の獲得に成功したと考えられる．

また，提案手法の各層の分散表現に対して得られた精度の差異は小さかった．それと同様に，各層における各文セットに対する識別結果についても大きな相違は見られなかった．表 5 に，提案手法の 2 層目の中間表現を文の分散表現とした場合について，すべての試行において同一の識別結果が得られた文セットの例と，その各セットにおける 2 文目とその前後の文の分散表現の cos 類似度を示す．ここで例 1, 2 はそれぞれ正しく識別されたある連続文セットと，それを元に生成された非連続文セットの例である．

例 1 について，2 文目とその前後の文の分散表現の cos 類似度が大きい方が連続文であると識別され，小さい方は非連続文であると識別されている．しかし，例 2 については例 1 と異なり，cos 類似度が小さい方が連続文，大きい方が非連続文であると識別した結果が得られた．このことから，各文の分散表現の cos 類似度は，文章の連続性を判別するための根拠となっていないことがわかる．

今回の実験では小説における連続した文章を正解データとして使用したが，実際には，連続する文章は一意に定まるものではない．筆者の主観ではあるが，例 3 に示したように人間でもその連続性を判別することが困難である文章が，データセット内に複数存在していた．この点を考慮に入れると，提案手法に対して得られた識別精度は十分に高いものであったと考察される．

6 まとめと今後の課題

本研究では，2 層の LSTM に基づく RAE による文の分散表現獲得手法に対して Attention 機構を導入した．その結果，文章の連続性を判断するという観点から分散表現の性能は向上し，その有効性を確認できた．

今後の課題としては，現在最先端の性能を示している言語モデルである Bidirectional Encoder Representations from Transformers (BERT)[6] との性能比較が挙げられる．また，LSTM の必要性や使用する Attention 層の数や種類などの観点からモデルの構造を検討することが必要であると考える．

表 4: 各手法の精度

	Accuracy		Precision		Recall		F-score	
モデル	mean.	std.	mean.	std.	mean.	std.	mean.	std.
提案手法 (第 1 層)	**0.836**	0.0050	**0.836**	0.012	**0.836**	0.014	**0.836**	0.0054
提案手法 (第 2 層)	**0.831**	0.0039	**0.826**	0.016	**0.839**	0.031	**0.832**	0.0080
Unidirectional LSTM Based RAE (第 1 層)	0.754	0.020	0.760	0.023	0.746	0.089	0.749	0.037
Unidirectional LSTM Based RAE (第 2 層)	0.695	0.0056	0.696	0.018	0.696	0.045	0.695	0.015
Bidirectional LSTM Based RAE (第 1 層)	0.739	0.0094	0.736	0.023	0.751	0.061	0.741	0.021
Bidirectional LSTM Based RAE (第 2 層)	0.720	0.014	0.729	0.029	0.711	0.086	0.715	0.036
Doc2Vec	0.613	0.0068	0.608	0.0083	0.635	0.011	0.621	0.0056

表 5: すべての試行において同一の識別をした例 (提案手法: 2 層目)

	連続性		cos 類似度		
例 No.	真値	予測値	1-2 文間	2-3 文間	文セット
1-a	✓	✓	0.99977	0.98617	はてな何でも容子がおかしいと、のそのそ這い出して見ると非常に痛い。吾輩は藁の上から急に笹原の中へ棄てられたのである。ようやくの思いで笹原を這い出すと向うに大きな池がある。
1-b	×	×	0.75133	0.63391	はてな何でも容子がおかしいと、のそのそ這い出して見ると非常に痛い。**と私は、私と並んでモッコをかついで歩いている若い娘さんにたずねた。**ようやくの思いで笹原を這い出すと向うに大きな池がある。
2-a	✓	✓	0.77122	0.40366	その中に月の光りが、大幅の帯を空に張るごとく横に差し込む。吾輩は前足に力を込めて、やっとばかり棚の上に飛び上がろうとした。前足だけは首尾よく棚の縁にかかったが後足は宙にもがいている。
2-b	×	×	0.98719	0.78272	その中に月の光りが、大幅の帯を空に張るごとく横に差し込む。**僕が小説を書けない振りをしたら、人々は僕を、書けないのだと噂した。**前足だけは首尾よく棚の縁にかかったが後足は宙にもがいている。
3	×	✓	0.99674	0.93814	しかも吾輩の方で少しでも手出しをしようものなら家内総がかりで追い廻して迫害を加える。**そうしてお顔色はいつも冴えず、お歩きになるのさえやっとのように見える日もある。**台所の板の間で他が顫えていても一向平気なものである。

参考文献

[1] Tomas Mikolov, Kai Chen, Greg S. Corrado, and Jeffrey Dean. Efficient estimation of word representations in vector space, 2013.

[2] Dzmitry Bahdanau, Kyunghyun Cho, and Yoshua Bengio. Neural machine translation by jointly learning to align and translate, 2014. cite arxiv:1409.0473Comment: Accepted at ICLR 2015 as oral presentation.

[3] 福田 清人, 森 直樹, and 松本 啓之亮. LSTM を用いた文の分散表現の獲得手法に関する一考察. **言語処理学会 第 24 回年次大会 発表論文集**, pages 1195–1198, 2018.

[4] 福田 清人. **計算機による物語の創発的生成に関する研究**. PhD thesis, 大阪府立大学, 2019.

[5] Quoc Le and Tomas Mikolov. Distributed representations of sentences and documents. In *Proceedings of the 31st International Conference on International Conference on Machine Learning - Volume 32*, ICML'14, pages II–1188–II–1196. JMLR.org, 2014.

[6] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. BERT: pre-training of deep bidirectional transformers for language understanding. *CoRR*, abs/1810.04805, 2018.