

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

飯倉 陸^{†1,a)} 岡田 真^{†1} 森 直樹^{†1}

概要: 本研究では, 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 Attention 機構を導入し, その性能の向上を図った. 単語の順序識別および文章の連続性を判別する実験をした結果, それぞれで従来手法の精度を上回る高い精度を得られた. 単語の順序および文章の連続性を考慮するという観点から分散表現の性能向上を確認し, Attention 機構の導入に対する有効性を示す.

キーワード: 分散表現, Attention 機構, LSTM

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.

自然言語処理分野における現状として, 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. 一方で, 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が確立されているとは言い難い.

そこで本研究では, 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] を導入した手法を提案する. また, 獲得した分散表現を用いた単語の順序識別および文の連続性識別の実験により, 提案手法と従来手法の性能を Attention 機構の有無の観点から相対的に評価し, その有効性を判断する.

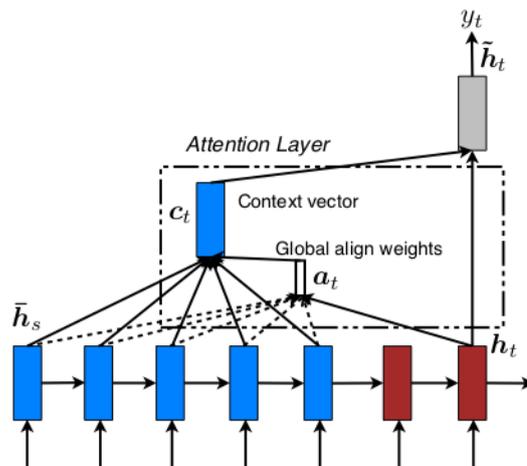


図 1: Attention 機構を備えたモデルの概略図

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手法が Attention 機構 [2] である. Attention 機構では, Encoder 側で入力文の各単語の荷重を決定してエンコードする際に注目すべき場所を制御する. 図 1 にその Attention 機構を備えたモデルの概略図を示す [3].

Attention 機構では入力文の各単語 x_i に対する荷重 α_t

^{†1} 大阪府立大学, 大阪府堺市中区学園町 1 番 1 号
Osaka Prefecture University, 1-1 Gakuen-cho, Naka-ku, Sakai,
Osaka, 599-8231 Japan

^{a)} iikura@ss.cs.osakafu-u.ac.jp

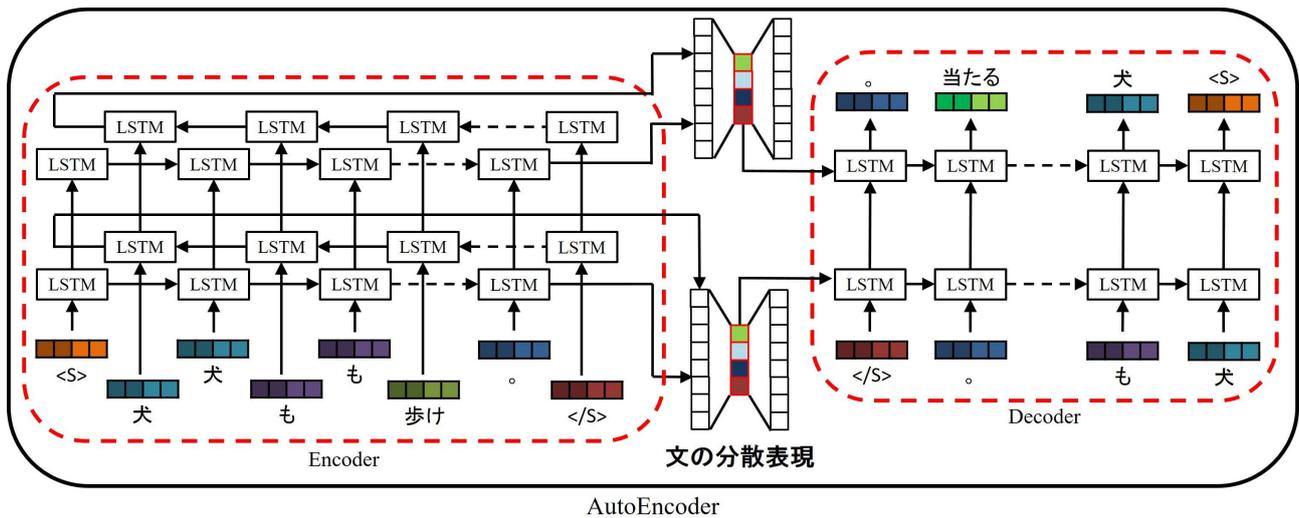


図 2: 福田らの手法のモデル概略図

を計算することで、コンテキストベクトル c_t を得る. α_t は Encoder で出力される全時刻の隠れ状態ベクトル \bar{h}_i と Decoder から出力される各時刻のベクトル 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.

$$\alpha_t(i) = \frac{\exp(\text{score}(h_t, \bar{h}_i))}{\sum_{j=1}^n \exp(\text{score}(h_t, \bar{h}_j))} \quad (1)$$

$$c_t = \sum_{i=1}^n \alpha_t(i) \bar{h}_i \quad (2)$$

$$\text{score}(h_t, \bar{h}_i) = h_t^\top \bar{h}_i \quad (3)$$

3. 先行研究

文の分散表現獲得手法の先行研究として、福田らの提案手法がある [4]. 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 モデルを用いており、そのモデルの Encoder の入力と Decoder の出力を同一にするように学習させることでこのモデルを RAE として用いている. 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.

福田らの手法では、時系列データに対して有効である 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を期待している.

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. 図 2 にその概略図 [5] を示す. このモデルでは、順方向、逆方向の LSTM それぞれにおける最終的な隠れ状態ベクトルを結合し、線形層に通過させることで次元圧縮する. そして得られたベクトルを対象文の分散表現としている.

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手法に対して Attention 機構を導入したモデルを提案する. 図 3 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. 文の分散表現とするのは、先行研究と同様に Encoder と Decoder の連結部における中間表現である.

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[6] 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」 [7] および、小説投稿サイト「小説家になろう」 [8] から収集した約 3.0 GB の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、以下の 2 つの実験により、提案手法と従来手法とを比較した.

- 単語の順序識別,
- 文章の連続性識別.

従来手法として用いたのは、先に示した福田らの手法と、Gensim[9] により実装された Doc2Vec[10] である. 表 2 に、Keras[11] により実装した提案手法および従来手法の設定値を示す. なお、学習用データとしては 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

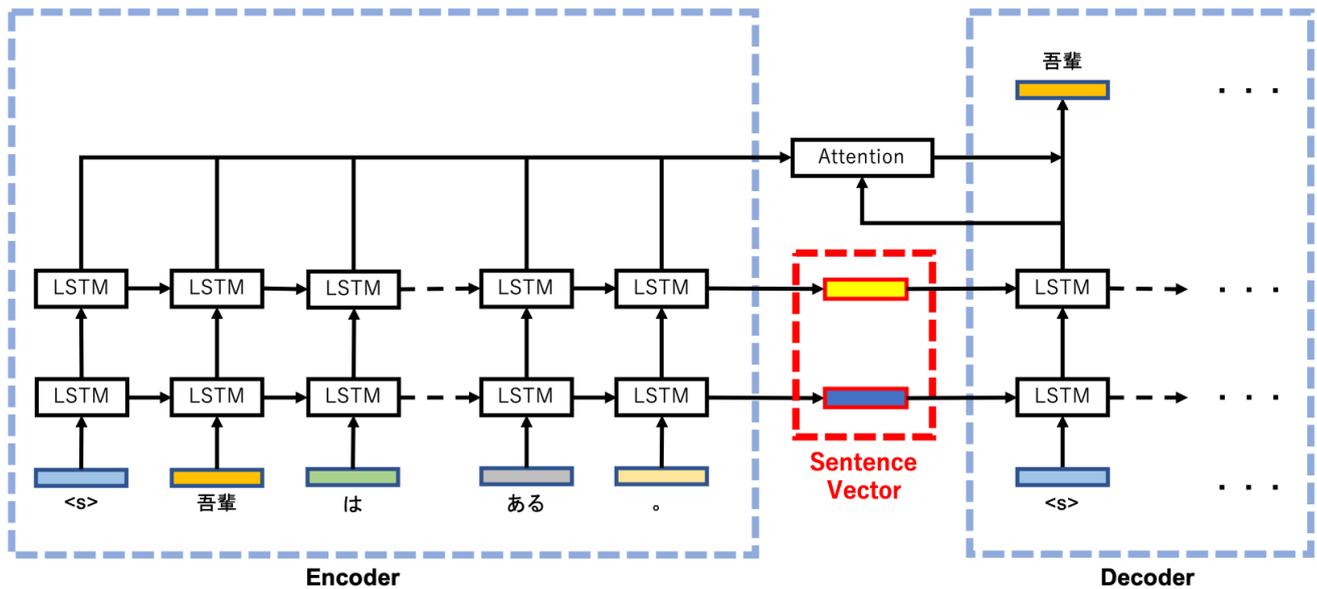


図 3: 提案手法のモデル概略図

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	5
語彙数	300,000

m_{\max} 単語以下の文を用いた。

5.1 実験 1: 単語の順序識別

5.1.1 実験 1: 実験手順

以下に実験の手順を示す。

- 電子図書館「青空文庫」[12] により管理される夏目漱石の小説「吾輩は猫である」「坊っちゃん」「草枕」「三四郎」「それから」「門」「彼岸過迄」「行人」「ころも」「明暗」の 10 作品から、会話を含まない n_{\min} 単語以上 n_{\max} 単語以下の文を抽出し、文集合 S を生成する。
- 文集合 S の各文 s を MeCab により単語に分割し、無作為に並び替えることで文法的に破綻した文、いわゆる非文 s' を生成しその集合を非文集合 S' とする。ただし、 s' は以下の条件に従うものとする。
 - 句点が文末にある。
 - 読点が文頭がない。

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	19,647,614 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	5

表 3: 実験 1: 各データのサイズ

データ	サイズ
$ S : S' $	24,000 : 24,000

- S, S' の各文について分散表現を獲得し、Multi Layer Perceptron による 2 クラス分類問題を解く。

(手順終わり.)

表 3 に S, S' の各集合のサイズを示す。

表 4: 実験 1: 各手法の精度

モデル	Accuracy	Precision	Recall	F-score
提案手法 (第 1 層)	0.836	0.889	0.769	0.824
提案手法 (第 2 層)	0.830	0.870	0.778	0.821
Bidirectional LSTM based RAE (第 1 層)	0.820	0.812	0.833	0.822
Bidirectional LSTM based RAE (第 2 層)	0.720	0.695	0.786	0.737
Unidirectional LSTM based RAE (第 1 層)	0.756	0.749	0.774	0.760
Unidirectional LSTM based RAE (第 2 層)	0.739	0.738	0.744	0.740
Doc2Vec	0.522	0.520	0.563	0.540

5.1.2 実験 1: 結果と考察

表 4 に 10 分割交差検証により得られた各手法に対する各精度を示す. 提案手法の各層における分散表現を用いた場合の Accuracy は, 各従来手法と比較して高い値が得られた. この結果から, Attention 機構の導入により単語の出現順序をさらに考慮した文の分散表現の獲得が可能になったといえる. また Precision については高く, Recall については低くなった. このことから, 提案手法は非文の検知に対して利点を有すると考えられる.

5.2 実験 2: 文章の連続性識別

5.2.1 実験 2: 実験手順

以下に実験の手順を示す.

- 「青空文庫」における夏目漱石の小説「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, 同一の段落内の連続する 3 文の組を抽出し, 連続文セット c とする. また, それらの集合を連続文セットの集合 C_{train} とする. ただし各 c に含まれる文は, 以下の各条件に従うものとする.
 - n_{\min} 単語以上 n_{\max} 単語以下で構成されること.
 - 感嘆符や疑問符のような記号を含まないこと.
 - 各文はただ 1 つの c にのみ含まれること.
- 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 R とする.
1. の操作で生成した C_{train} に含まれる各 c の 2 文目を, 2. の操作で生成した R の各文と置換することで非連続文セット w を生成し, その集合を W_{train} とする.
- 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 $C_{\text{test}}, W_{\text{test}}$ を生成する. ただし, W_{train} の生成のために用いた

表 5: 実験 2: 各データのサイズ

データ	サイズ
$ C_{\text{train}} : W_{\text{train}} $	3,092 : 3,092
$ C_{\text{test}} : W_{\text{test}} $	370 : 370

R の各文は, W_{test} の生成には用いないものとする.

- $C_{\text{train}}, W_{\text{train}}$ を訓練データとし, $C_{\text{test}}, W_{\text{test}}$ をテストデータとする. 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.

(手順終わり.)

表 5 に使用した各データのサイズを示す.

5.2.2 実験 2: 結果と考察

表 6 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おける各文セットに対する識別結果について大きな相違は見られなかった. 表 7 に, 提案手法により獲得した文の分散表現を用いたとき, 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を示す.

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. 筆者の主観による評価ではあるが, 表 7 の例 3, 4 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 機構を導入したモデ

表 6: 実験 2: 各手法の精度

モデル	Accuracy		Precision		Recall		F-score	
	mean.	std.	mean.	std.	mean.	std.	mean.	std.
提案手法 (第 1 層)	0.861	0.013	0.847	0.012	0.882	0.043	0.863	0.017
提案手法 (第 2 層)	0.836	0.0079	0.819	0.020	0.865	0.028	0.841	0.078
Bidirectional LSTM based RAE (第 1 層)	0.816	0.0039	0.790	0.011	0.862	0.027	0.824	0.0069
Bidirectional LSTM based RAE (第 2 層)	0.771	0.011	0.761	0.015	0.794	0.050	0.776	0.019
Unidirectional LSTM based RAE (第 1 層)	0.797	0.011	0.768	0.013	0.851	0.035	0.808	0.014
Unidirectional LSTM based RAE (第 2 層)	0.755	0.0086	0.742	0.014	0.783	0.038	0.761	0.013
Doc2Vec	0.587	0.0072	0.580	0.0066	0.629	0.015	0.604	0.0086

表 7: すべての試行において同一の識別をした例 (提案手法: 1, 2 層目)

例 No.	連続性		文セット
	真値	予測値	
1	✓	✓	何でも薄暗いじめじめした所でニャーニャー泣いていた事だけは記憶している。吾輩はここで始めて人間というものを見た。しかもあとで聞くとそれは書生という人間中で一番獯悪な種族であったそうだ。
2	×	×	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柵の縁にかかったが後足は宙にもがいている。
3	✓	×	と今度は主人の方を見て顔色を窺う。悲しい事に力学と云う意味がわからぬので落ちつきかねている。しかしこれしきの事を尋ねては金田夫人の面目に関すると思っただけ、ただ相手の顔色で八卦を<unknown>見る。
4	×	✓	彼等はその強力を頼んで正当に吾人が食い得べきものを奪ってしまっている。私は黙っておナスに水をやってた。吾輩は教師の家に住んでいるだけ、こんな事に関する<unknown>よりもむしろ楽である。

ルを提案した。その結果、単語の出現順序および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、その有効性を確認することができた。

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)[13] を使用したモデル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 層の数や種類の観点からモデルの構造を検討することが必要である。

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] Thang Luong, Hieu Pham, and Christopher D. Manning. Effective approaches to attention-based neural machine translation. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*, pages 1412–1421, Lisbon, Portugal, September 2015. Association for Computational Linguistics.

[4] 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.

[5] 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.

[6] 工藤 拓, 山本 薫, and 松本 裕治. Conditional random fields を用いた日本語形態素解析. *情報処理学会研究報告 自然言語処理 (NL)*, 2004(47):89–96, may 2004.

[7] ウィキペディア. <https://ja.wikipedia.org/>.

[8] 小説家になろう - みんなのための小説投稿サイト. <https://syosetu.com>.

[9] Radim Řehůřek and Petr Sojka. Software Framework for Topic Modelling with Large Corpora. In *Proceedings of the LREC 2010 Workshop on New Challenges for NLP Frameworks*, pages 45–50, Valletta, Malta, May 2010. ELRA. <http://is.muni.cz/publication/884893/en>.

[10] 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.

[11] François Chollet et al. Keras. <https://keras.io>, 2015.

[12] 青空文庫. <https://www.aozora.gr.jp>.

[13] 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.