The 33rd Annual Conference of the Japanese Society for Artificial Intelligence, 2019

LSTM-RNNを用いた文体変換手法

A sentence style conversion method using LSTM-RNN

下地　健太[*1]　　森田　和宏[1]　　泓田　正雄[1]
Kenta Shimoji　　Kazuhiro Morita　　Masao Fuketa

[1] 徳島大学大学院先端技術科学教育部
Department of information Science and Intelligent Systems, Tokushima University

Abstract:　This paper describes a sentence style conversion method using recurrent neural network with long short-term memory cells (LSTM-RNN). In the proposed method, LSTM-RNN is used to learn direct style sentences vectorized by one-hot expressions. Then, the sentence end expression of the distal style sentence is removed and the vectorized one is input into the learned model. The next word is predicted until sentence ends, and the obtained word vector sequence is added to the end of the input vector sequence. A direct style sentence is converted by decoding the generated vector sequence into the form of the natural language. We experimented to evaluate the accuracy of the proposed method. As a result, it turned out that a sentence style can be converted by the method.

1. はじめに

近年，文書の電子化が進んでおり，コンピュータを使った文書作成の機会が増加している．電子媒体の文書にもアナログ媒体のものと同様にヒューマンエラーがしばしば見られる．その一つとして文体の統一に関する誤りが挙げられる．さらに，文体の統一の中でも多くの誤りが発生するのが，同一文書における常体文(「だ」・「である」調)と敬体文(「です」・「ます」調)の混在である．文書作成においてこのような誤りを訂正する作業は必要不可欠である．しかし，文書の作成者が自ら訂正作業をおこなうことは，その内容を理解しているため，客観性に欠ける．また，人手による訂正作業では文書作成と同様に一つの見落としもないという保証はなく，文書が長くなればなるほどヒューマンエラーは増えていく．そこで，コンピュータによる文体変換システムが必要とされる．本稿では LSTM-RNN を用いた文体変換手法について述べる．また，提案手法の変換精度評価のために実験をおこなう．

林らの研究[1]では，人手で作成した変換ルールを用いた手法で書き言葉から話し言葉への文体変換をおこなっている．しかし，この手法ではルールでの対応が困難な場合があり柔軟な変換がおこなえないことが問題であった．本研究ではルールでは難しい場合に対応するために，LSTM-RNN を用いることにより，それぞれの文体に特有の表現や単語の並びを学習させ，未知の入力に対し生成した学習モデルの単語予測を利用して文体の変換をおこなう．

2. LSTM-RNN

RNN(Recurrent Neural Nerwork)は時系列データの学習に特化した深層学習手法の一つである．時系列データとは，ある要素が時間的順序を追って並んでいるデータのことであり，文章，声，映像，株価情報等がこれに当たる．そのため，主に自然言語処理や音声言語処理等の分野で利用されている．内部にループ構造を持つことが特徴であり，これにより前のステップの情報を後続のステップに渡すことが可能となっている．また，本研究では RNN に加え，LSTM(Long Short-Term Memory)[2]を使用する．LSTM は学習に必要な情報の取捨選択をおこなうことにより，時系列データの長期依存性を学習することを可能にする RNN のユニットの一つである．LSTM を中間層に用いることにより，通常の RNN が持つ逆誤差伝播による学習における勾配消失に問題を解決し，より高い精度で時系列予測をおこなうことができる．本研究では，入力した文字列の次に来るであろう単語を予測するために RNN を用い，自然言語という比較的データ長の長い時系列データを扱うため，LSTM を採用する．本稿では LSTM を中間層に用いた RNN を LSTM-RNN と呼ぶこととする．

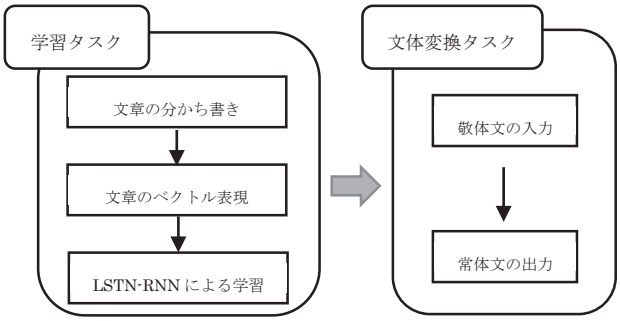

図 1:提案手法

3. 提案手法

学習データに常体文のみを使用することにより，常体文の単語の並びを予測することができるので，これを利用して敬体文を常体文に変換する．これに関する提案手法の流れを図 1 に示す．また，本章では提案手法のそれぞれのステップについて解説する．

連絡先：下地 健太，knt.smj@gmail.com

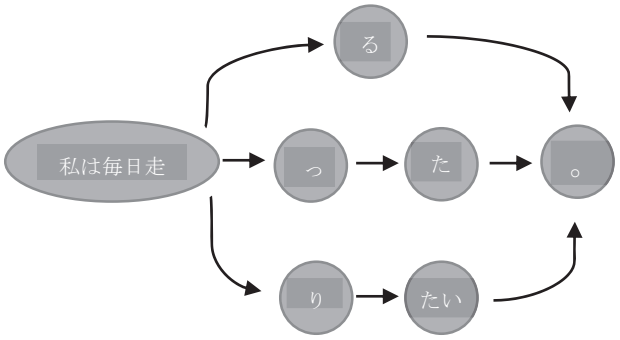

図 2:文末予測の例

3.1 学習タスク

本節では文章を学習しモデルを生成するタスクについて解説する.

まず, 学習データには常体文のみを用いるため, 常体文のみを収集する必要がある. 敬体文には助動詞「です」,「ます」のどちらかが使われることが主であるため, これらの助動詞が含まれない文という条件で常体文のみを収集した. そして, 学習に用いる文章に対し形態素解析をおこない, 分かち書きをおこなう. ここで, 動詞については,「走る」,「走り(ます)」のように同じ意味を表していても敬体文と常体文で活用が異なる場合があるため, 形態素をさらに語幹で分割する. 前述の例の場合,「走/る」,「走/り」のように分割する. 形態素解析には MeCab[3]を使用し, 語幹はその動詞の原形と前方一致する部分とする. その後, One-hot 表現を用いて単語のベクトル化をおこなう. One-hot 表現とは, 1 つのビットにだけ 1(High)を, 他のビットには 0(Low)を割り当てるベクトル表現である. これにより, 文章を単語ベクトルの集合として表現する. 最後に, ベクトル化した文章について 1 文毎に LSTM-RNN を用いて単語の順番を学習し, モデルを生成する.

3.2 文体変換タスク

本節では入力文を 3.1 節で生成した学習モデルにより文体変換をおこなうタスクについて解説する.

敬体文から常体文への変換が目的であるため, 文体変換時の入力データには敬体文を用いる. 3.1 節と同様にして入力となる敬体文に対し分かち書きをおこなった後, 文末表現を除去する. 文末表現は文の最後に現れる名詞もしくは動詞の語幹の次の文字から句点までとする. 例えば,「私は毎日走ります。」という文の文末表現は「ります。」なので, 除去後の文字列は「私は毎日走」となる. そして, これに対しベクトル化をおこない 3.1 節で生成した学習モデルに入力する. 次に来る確率の高い文末語までの単語の組み合わせを複数生成し, その中で最も文全体としての生成確率が最も高いものを出力とする. 例の場合,「私は毎日走」に続く単語の組み合わせとして,「る。」,「った。」,「りたい。」が候補となり, その中から「る。」が出力として選ばれる. (図 2 参照)これを元の入力文の末尾に追加し, できたベクトル列を自然言語の形に変換することで常体文を得る. 前述の例では「私は毎日走る。」という常体文が得られる.

表 1:実験結果

	○	△	×
分類数	23	59	18

表 2:変換例

入力文	出力文	評価
実状の説明に適しているかも知れません。	実状の説明に適しているかも知れない。	○
まじめな顔をして言いました。	まじめな顔をして言った。	○
笑わせるのには事を欠きませんでした。	笑わせるのには事を欠いた。	△
まるで見当つかないのです。	まるで見当つかないのか。	△
完全に落第でした。	完全に落第る。	×
これがまた家中の大笑いでした。	これがまた家中の大笑いがある。	×

4. 評価実験

提案手法の精度を確認するために評価実験をおこなった. 学習データには, 読売新聞記事内の常体文を 26,366 文使用した. テストデータには青空文庫[4]の作品内の敬体文を 100 文使用し, 提案手法を用いて文体変換をおこなった. 本実験では埋め込み層, 2 つの LSTM 層と全結合層からなるニューラルネットワークを使用した. LSTM 層の次元数は 200 次元とし, 損失関数には活性化関数には softmax 関数を用いた. 評価は○(正しい), △(常体に変換できているが元の敬体文の意味情報が保持されていない), ×(文の意味が通っていない)の三段階評価を人手によりおこなった.

評価実験の結果を表 1 に示す. 8 割以上が意味に破綻のない文として出力された. 本実験における変換例を表 2 に示す. 評価が△のものについて, 1 つ目の「笑わせるのには事を欠きませんでした。」は「笑わせるのには事を欠かなかった。」と変換されるべきであるが, 提案手法では学習した単語の順番を元に予測をおこなうため, 変換の過程で否定文であるという情報が失われている. また, 2 つ目についても同様に肯定文であるという情報が失われている. この問題を解決するには, 入力文の種類が何であるかを判定し, それと同様の種類の出力文を生成する必要があると考える. ×のものについては, 学習データの不足が原因であると考えられるため, 十分な量のコーパスを用意することが必要である.

5. おわりに

本稿では, LSTM-RNN を用いた文体変換手法の提案をおこなった. 100 の敬体文に対し本手法を適用することにより, 精度評価をおこなった. その結果 8 割以上が意味に破綻がない文として出力されたが, 変換の過程で入力文の意味情報が失われる場合があるという課題が明らかになった.

今後は今回の結果を踏まえ, 課題に関して改良をおこなうことで変換精度を向上させていく予定である.

参考文献

[1] 林由紀子, 松原茂樹. 自然な読み上げ音声出力のための書き言葉から話し言葉へのテキスト変換. 情報処理学会研究報告, pp. 49-54,2007

[2] Sepp Hochreiter and J¨urgen Schmidhuber. Long short-term memory. Neural computation, Vol. 9, No. 8, pp. 1735–1780, 1997.

[3] T. Kudo. Mecab : Yet another partof-speech and morphological analyzer. http://taku910.github.io/mecab/

[4] 青空文庫. https://www.aozora.gr.jp/