ダブルレイヤー LSTM を用いた翻訳システムの構築

1 はじめに

近年, 機械学習の発展に伴い, 自然言語処理も大きく発展している. 自然言語処理の 1 つタスクとして, 機械翻訳について新しい手法は次々に提案されている. 特に Attention メカニズムの出現は, 従来の逐次翻訳の手法を一変し, 時系列データの順番を問わず手法により, 翻訳の精度を大幅に向上している. しかし, 新しいモデルは高精度を持つ同時に, 時間と設備の要求も高くなる.

LSTM の可能性を探索するため, 本研究では, Attention メカニズムを使わず, Long Short-term Memory (LSTM) を基づいて実験をする.

2 要素技術

2.1 Long Short-term Memory

Recurrent Neural Network (RNN) [4] とは, 回帰構造を持つニューラルネットワークである.

通常のニューラルネットワークでは, レイヤの出力は, 次のレイヤの入力として利用されるが, RNN では, 同じニューラルに対して, 当時刻の時系列データを入力するだけでなく, 前時刻の出力と次時刻の時系列データを入力するニューラルネットワークである, 図 1 に RNN の構造を示す.

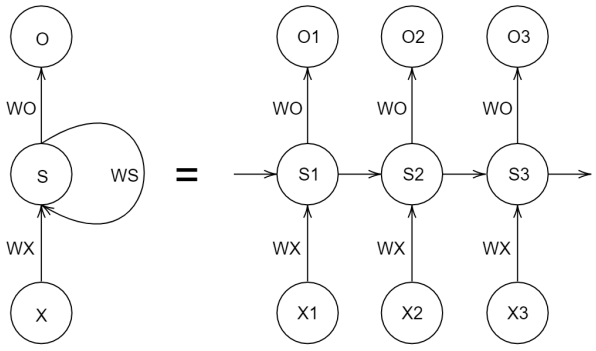

図 1: RNN の構造

誤差逆伝播法により RNN の訓練は, 逆伝播される勾配の消失 (勾配がゼロに収束) あるいは爆発 (勾配が無限に発散) 問題がある. この問題を解決するため, ゼップ・ホッフライターらは 1997 年に Long short-term memory (LSTM) [2] を提唱した. LSTM のアーキテクチャは Memory Cell と三つの Gate (Input Gate, Output Gate, Forget Gate) から構成される. LSTM は勾配をそのまま流れることが可能であるので, 勾配消失と勾配爆発の問題を解決できる. 図 2 に LSTM の構造を示す.

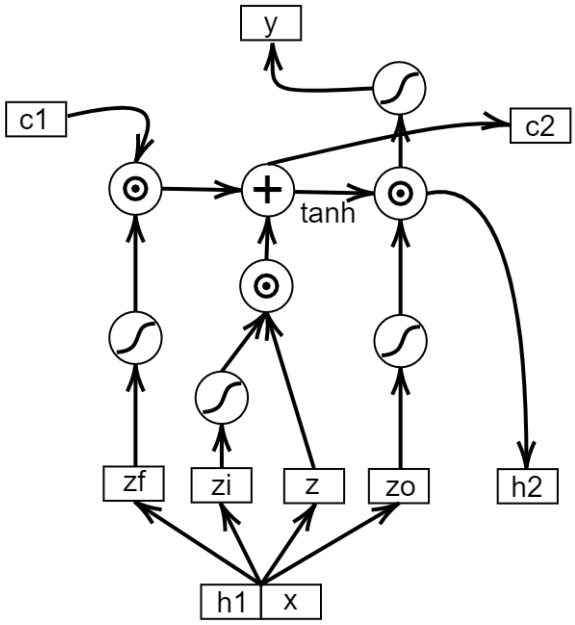

図 2: LSTM の構造

2.2 Sequence to Sequence

Sequence To Sequence (seq2seq) [5] とは, 2014 年に Google が発表した言語モデルである. 従来の Deep Neural Network (DNN) が扱いにくい時系列データ問題を解決するため, seq2seq は Encoder-Decoder という形式のモデル構造を導入した. Encoder は入力する時系列データをベクトルに圧縮し, そのベクトルを Decoder に渡し出力系列を生成する. 本実験の seq2seq モデルは RNN を利用したため, Decoder の出力は自動的に調整できる. 図 3 に本実験用の seq2seq モデルを示す.

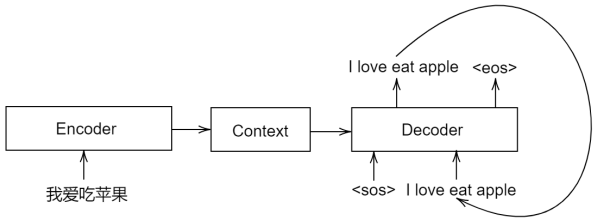

図 3: seq2seq モデル の構造

2.3 jieba

jieba [1] は 2013 年にリリースされ, 中国語 (簡体字と繁体字) 文章の分かち書きに専用するライブラリである. jieba の cut メソッドは精確モードと全モードがある. 精確モードは, 文章を jieba のディクショナリにより精確的に単語に分けるモード, そして, 全モードは文章の中に単語と見える部分を全部スキャンして, 分けるモードである. 英単語に対応するために, 本実験は精確モードで実行する.

2.4 bilingual evaluation understudy

bilingual evaluation understudy (BLEU) [3] とは, 現在機械翻訳に対して, 最も広く使われている評価手法である. この手法はモデルの訳を翻訳者の訳と比べ, 近いければ近いほど精度が高いという評価手法である.

BLEU スコアは 0 から 1 の間の実数で表現され, その数値を 100 とかけると, 100 点が満点の形式評価できる.

3 データセット

ManyThings データセットは Tatoeba プロジェクトで収集され, 英語からフランス語や中国語などの 81 国の言語に対応するペアで集まるデータセットである. 収集されたデータは英語 - 他言語のペアを, 単語の少ない方から多いの方までソートされる.

本実験に使われるのは ManyThings データセットの英語 - 中国語データセットである. 英語 - 中国語データセットは 24,360 の英語 - 中国語文章ペアがあり, ペアの後ろに Tatoeba プロジェクトに関する情報があるので, すべての文章に対し, 特殊符号と無関係情報を除去した.

表 1 にデータセットの一部を示す.

表 1: ManyThings 英中データセット例

英語	中国語
To tell the truth, I don't like him.	老，我不喜他。
I still haven't finished my homework.	我没完成作。
Yesterday, the weather was very nice.	昨天天气非常好。

4 実験の流れ

4.1 データ処理

今回の実験データは英語と対応する中国語 24360 ペアがあるので, 先ずは実験データを四対一の比率でトレーニングデータとテスティングデータに分ける. 表 2 にデータセットのペア数を示す.

表 2: Pairs

DataSet	Training	Testing
Pairs	19488	4872

処理したデータは, 単語からインデックス (word2index) , インデックスから単語 (index2word) というディクショナリの形式で保存する. 表 3 は各データセットのボキャブラリー数を示す.

表 3: Vocab

DataSet	Training	Testing
Chinese	12973	5814
English	6750	3541

4.2 モデルの実装

実験に使う seq2seq モデルの Encoder は, 二層の LSTM で実装する. ソースシーケンスを分かち書きで単語のトークンに分け, トークンのインデックスをワードインベッドで相応しい行列に転換し (単語のボキャブラリー数かけるインベッドサイズ) , 転換した行列を二層の LSTM に入力する. LSTM は出力 h と記憶 c を出力する. 図 4 に Encoder の構造を示す.

Decoder の構造は Encoder の構造の上, 全結合層と Teach Force Ratio を導入した構造である. Teach Force Ratio とは, モデルが生成したの悪い結果とターゲットの正しい結果どちらを使うかを決めるパラメータである.

図 5 に Teach Force Ratio を示す.

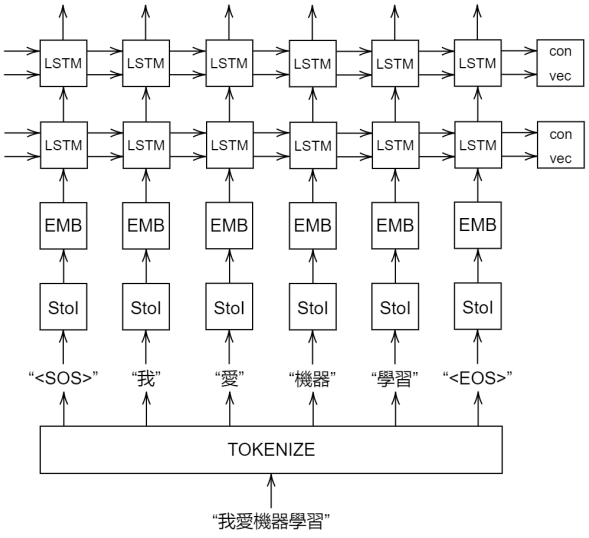

図 4: Encoder の構造

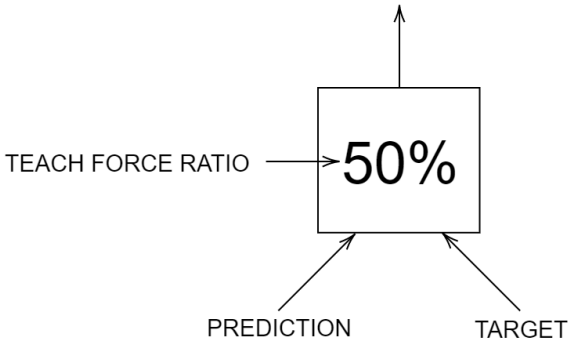

図 5: Teach Force Ratio

Decoder 最初の入力は始めを示すトークン 「SOS」と Encoder の出力である. 第一時系列の出力は, 次の時系列の入力として扱われる. 図 6 に Decoder の構造を示す.

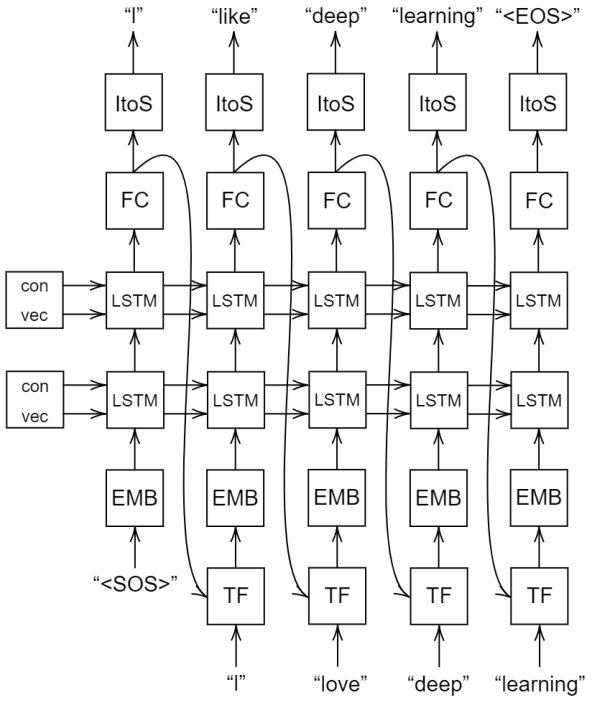

図 6: Decoder の構造

図 7 に Encoder と Decoder で構成した seq2seq モデルを示す. ミニバッチで実験するため, 実験は同時に複数の文章を処理する.

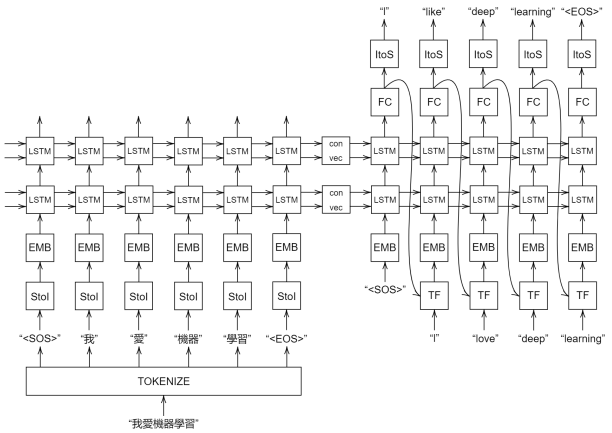

図 7: seq2seq モデルの詳細構造

4.3 トレーニング

トレーニングする前に, 分かち書きされたデータは順番に Encoder に入力し, ContextVector (出力 h と記憶 c) を出力する. Encoder が出力した Context Vector と文章の始めを示すトークン 「SOS 」を Decoder に入力し, 結果を再び Decoder に入力する. 文章の終わりを示すトークン 「EOS 」 が出力する場合停止する. 実験誤差は, seq2seq モデルの出力と, データセットの目標により, CrossEntrop で計算できる.

全エポックのランニングが終わると, テスティングデータセットでモデルを評価する. 評価するために, bilingual evaluation understudy(BLEU) スコアとは, 機械翻訳の評価方法である. 翻訳者の翻訳と近い程, 機械翻訳の精度が高い, 故に BLEU スコアも高いである. 実験は, テスティングデータで予測した結果とデータセットのターゲットにより, BLEU score メソッドで計算できる.

表 4 に実験に用いたパラメータを示す.

表 4: parameters

parameter	Value
input_size	12973
output_size	6750
hidden_size	1024
embedding_size	300
n_layer	2
batch_size	32
dropout	0.5
epoch	100
optimizer	Adam
loss	Cross-Entropy

4.4 実験結果

バッチサイズ 32 ,100 エポックでトレーニングロスが 0.47 に収束し, テストデータにより BLEU スコアは 12.21 になる.

5 まとめと今後の課題

本研究では, LSTM が機械翻訳分野に利用される可能性を探索するため, LSTM を使った seq2seq モデルを構築し, そして中国語英語の翻訳システムを作り, 評価した. 結果として二層の LSTM は確かに一層の LSTM よりよくのパフォーマンスがある.

今後の課題としては, Transformer を利用して機械翻訳システムの性能を比較し, そして漫画翻訳に利用できる可能性を探索する.

参考文献

[1] ”Jieba” (Chinese for ”to stutter”) Chinese text segmentation: built to be the best Python Chinese word segmentation module. `https://github.com/fxsjy/jieba`.

[2] Sepp Hochreiter and Jürgen Schmidhuber. Long short-term memory. *Neural Computation*, Vol. 9, No. 8, pp. 1735–1780, 1997.

[3] Kishore Papineni, Salim Roukos, Todd Ward, and Wei jing Zhu. Bleu: a method for automatic evaluation of machine translation. pp. 311–318, 2002.

[4] Alex Sherstinsky. Fundamentals of recurrent neural network (rnn) and long short-term memory (lstm) network. *Physica D: Nonlinear Phenomena*, Vol. 404, p. 132306, Mar 2020.

[5] Ilya Sutskever, Oriol Vinyals, and Quoc V. Le. Sequence to sequence learning with neural networks, 2014.