進捗報告

1 今週やったこと

- 二層 LSTM で実験する
- テストデータを BLEU でチェックする

2 実験の流れ

2.1 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 1 にデータセットのペア数を示す.

表 1: Pairs

DataSet	Training	Testing
Pairs	19488	4872

データセットの処理は, jieba で行う. jieba は 2013 年にリリースされ, 中国語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単語に分けるモード, そして, 全モードは漢字に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.

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 2 は各データセットのボキャブラリー数を示す. 表 2 に各データセットの文章ペアを示す.

表 2: Pairs

DataSet	Training	Testing
Chinese	12973	5814
English	6750	3541

2.2 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 1 に Encoder の構造を示す.

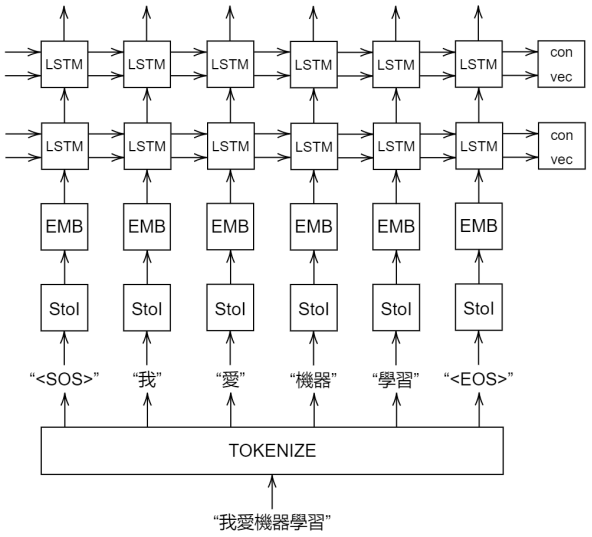

図 1: Encoder の構造

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 2 に Teach Force Ratio を示す.

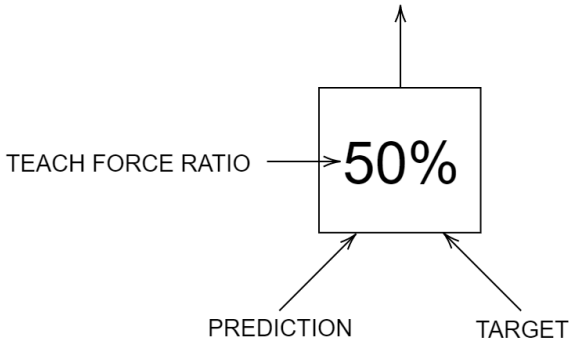

図 2: Teach Force Ratio

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 3 は *Decoder* の構造を示す.

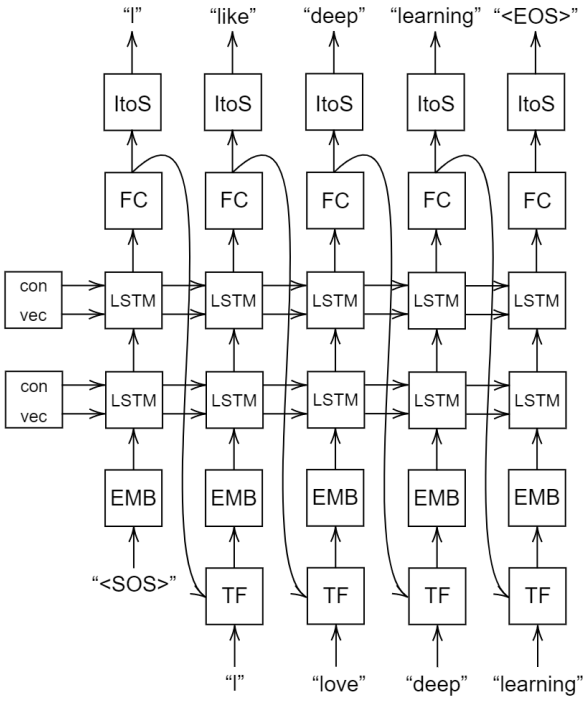

図 3: Decoder の構造

図 4 は Encoder と Decoder の構造を示す. 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.

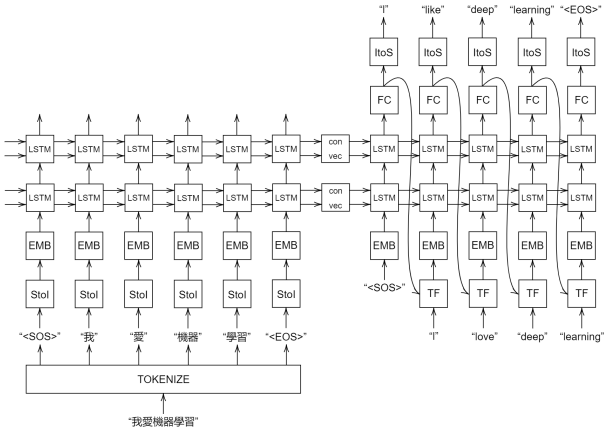

図 4: seq2seq の構造

2.3　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 ContextVector 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, *seq*2*seq* 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 3 に実験に用いたパラメータを示す.

表 3: 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

2.4　実験結果

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.

3　来週目標

- ASPEC-JC データセットで実験する