令和3年7月26日（月） M1前期発表会

ダブルレイヤーLSTMを用いた翻訳システムの構築

ソフトウェアシステム研究グループ

陳　偉斉

発表の構成

1.はじめに

2.要素技術

3.データセット

4.実験の流れ

5.まとめと今後の課題

発表の構成

1.はじめに

2.要素技術

3.データセット

4.実験の流れ

5.まとめと今後の課題

はじめに

漫画に関する翻訳システム

識別

セリフ

セリフ

翻訳

台词

はじめに（研究目的）

手法

・Attentionメカニズムを使ったモデルは時間と性能の要求が高い,故にLSTMで実験する

課題

LSTMで機械翻訳を理解する

発表の構成

1.はじめに

2.要素技術

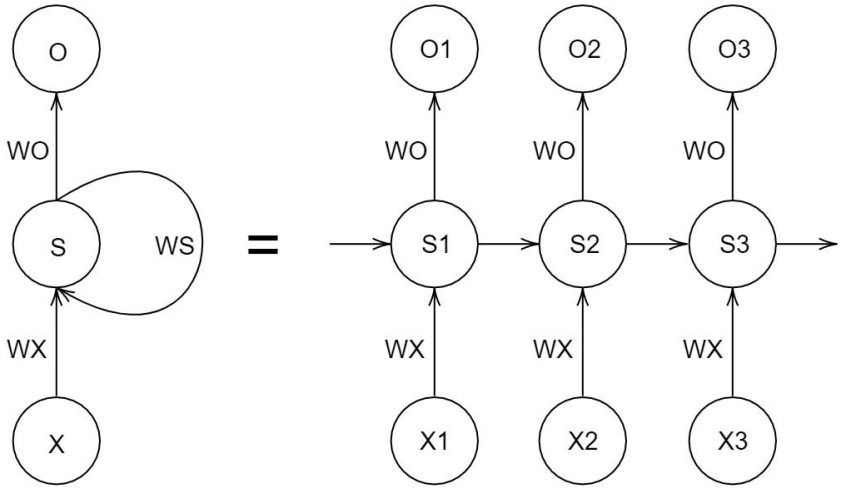
3.データセット

4.実験の流れ

5.まとめと今後の課題

要素技術

Recurrent Neural Netword (RNN)



$$O_t = g(WO \cdot S_t)$$

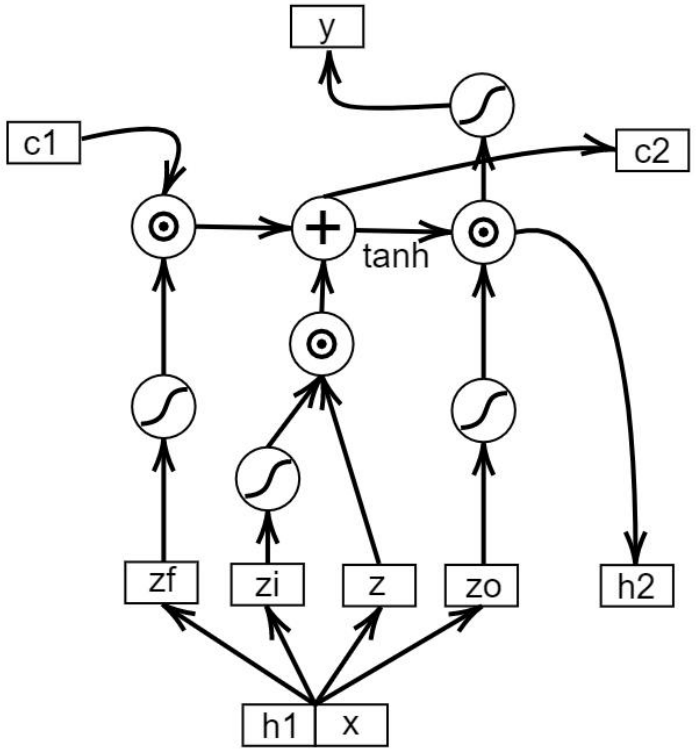
$$S_t = f(WX \cdot X_t + WS \cdot S_{t-1})$$

・回帰構造を持つニューラルネットワーク

・逆伝播による勾配消失と勾配爆発問題,故に長期的な記憶はできない

要素技術

Long Short-term Memory (LSTM)



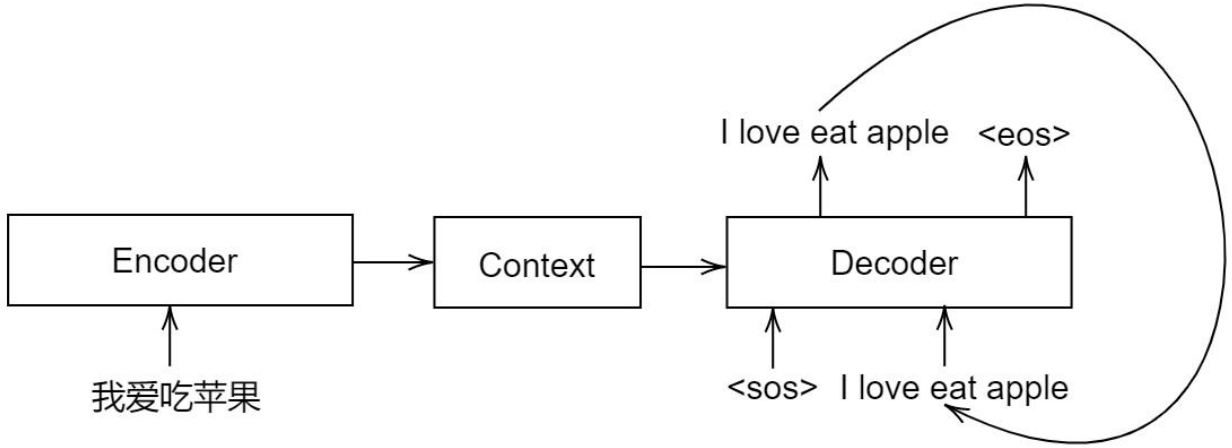
$$C_t = Zf \odot C_{t-1} + Zi \odot Z$$

$$h_t = Zo \odot tanh(C_t)$$

$$y_t = \sigma(Wh_t)$$

要素技術

Sequence to Sequence (seq2seq)



・時系列データを処理するネットワーク構造

要素技術

jieba(中国語テキスト分かち書き)

・全モード

我来到东京大学

我,来到,东京,东京大学,京大,大学

・精確モード

我来到东京大学

我,来到,东京大学

発表の構成

1.はじめに

2.要素技術

3.データセット

4.実験の流れ

5.まとめと今後の課題

データセット

ManyThingsデータセット

・ManyThings Bilingual Sentence Pairsの英語-中国語本文を使用

・全データセットは英語-中国語本文24360ペア，最長文本は33文字、最短文本は1

データセット

データセットの例

英語	中国語
Where are the strawberries	草莓在哪裡
What's the matter with you	你怎么了
You can count on her	你可以相信她
You don't need money	你不需要錢
We haven't lost hope	我们没有失望
Tom wanted to see me	汤姆想见我

発表の構成

1.はじめに

2.要素技術

3.データセット

4.実験の流れ

5.まとめと今後の課題

実験の流れ

データ処理

- データセット文本を4:1の比率で分ける

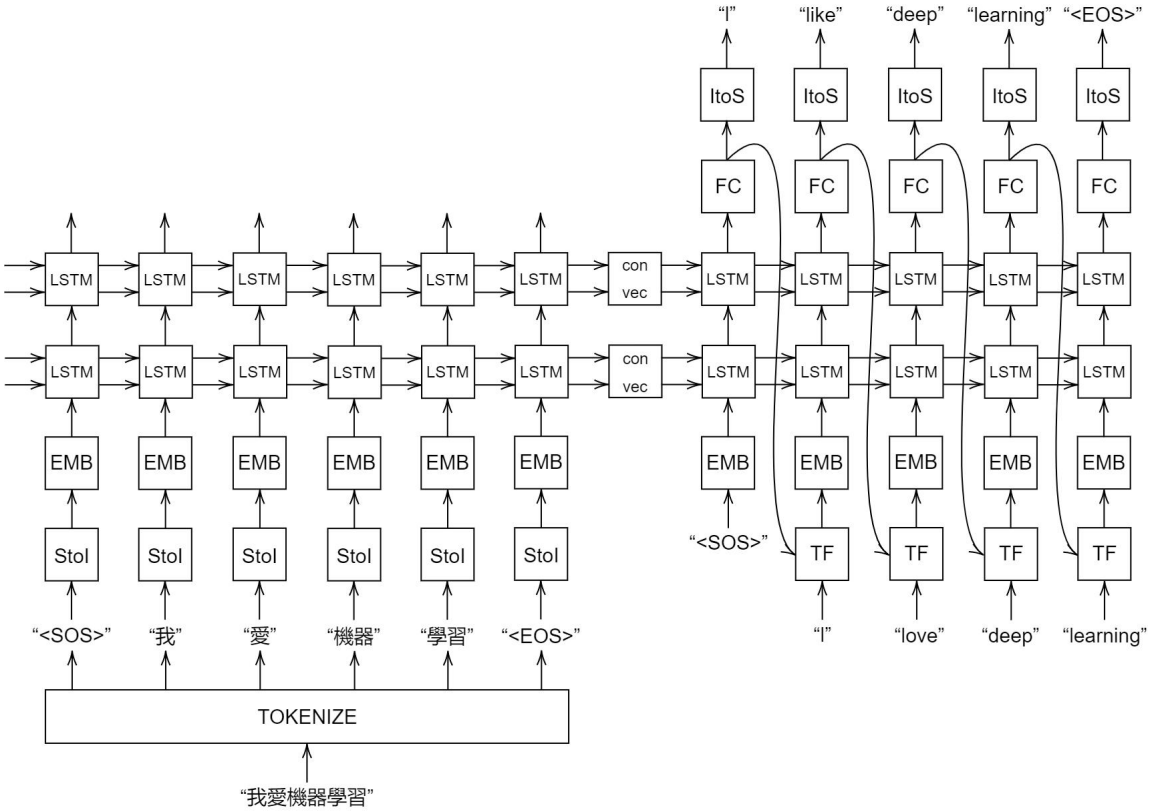
DataSet	Training	Testing
Pairs	19488	4872

- Tokenizeで単語のディクショナリを構築する

DataSet	Training	Testing
Cmn_Vocab	12973	5814
Eng_Vocab	6750	3541

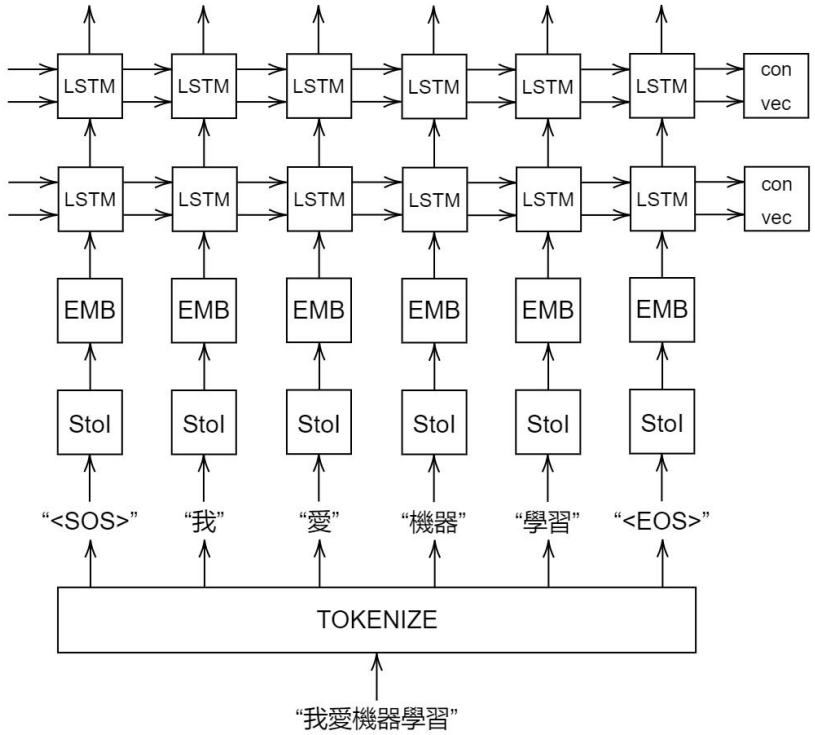
実験の流れ

モデルの実装



実験の流れ

Encoder



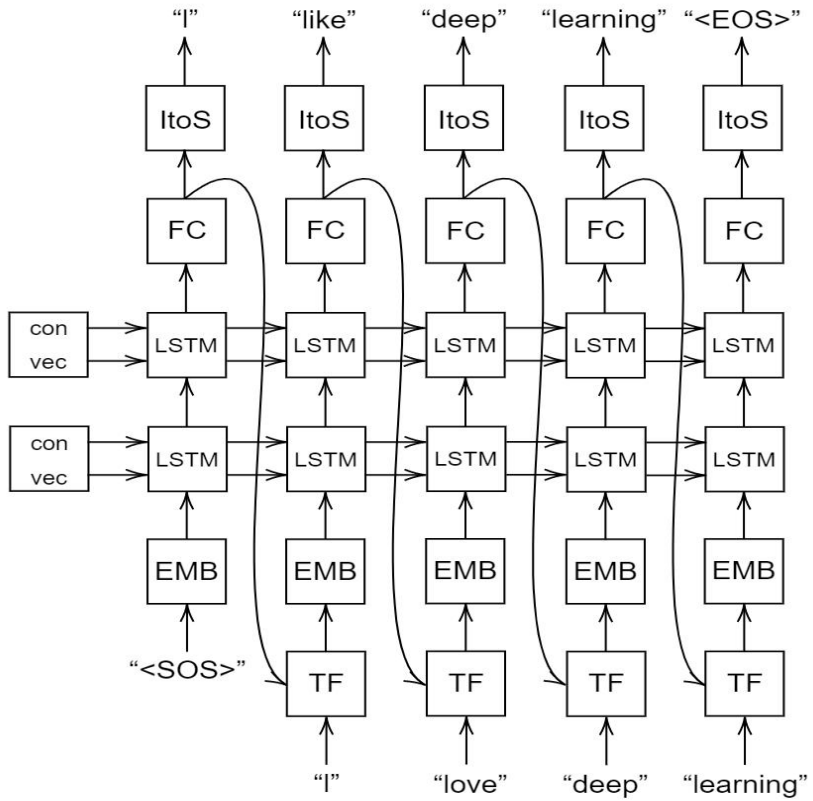
・始めと終わりのトークン<SOS>と<EOS>を加入

・二層LSTMで実装

・Context Vectorは出力hと記憶cを示す

実験の流れ

Decoder

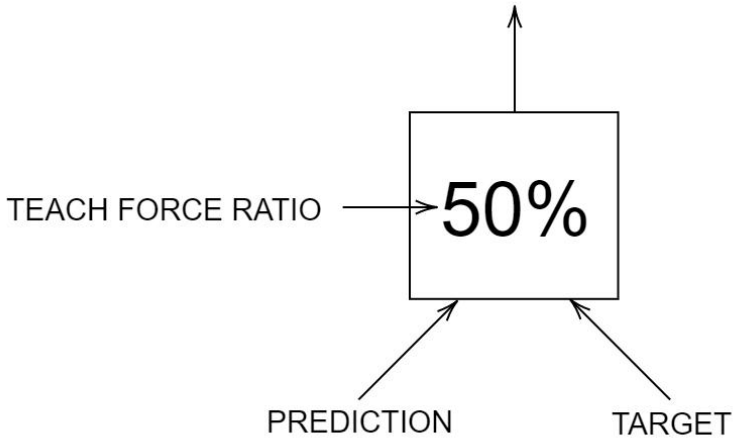


・Teach Force Ratioを導入

・最初の入力は<SOS>だけ

実験の流れ

Teach Force Ratio



・Encoderが予測した結果が間違えると,続きの実験によくない

・先生のように正確な答えをモデルに教える

実験の流れ

評価指標

- bilingual evaluation understudy(BLEU)
- 機械翻訳に広く使わる評価手法

$$\mathrm{BLEU}(\mathcal{H}, \mathcal{R}) = \mathrm{BP} \cdot \exp\left(\frac{1}{N}\sum_{n=1}^{N} \log P_n\right)$$

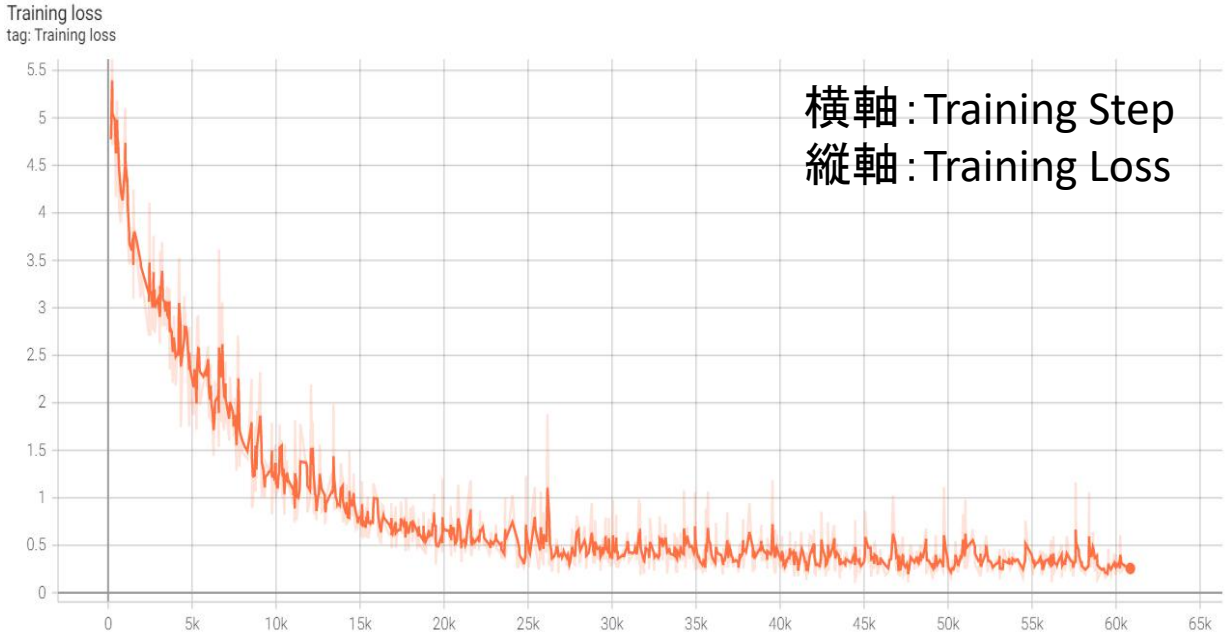
$$P_n = \frac{\sum_{i=1}^{S}\sum_{t_n \in h_i} \min(\mathrm{count}(h_i, t_n), \max\_\mathrm{count}(R_i, t_n))}{\sum_{i=1}^{S}\sum_{t_n \in h_i} \mathrm{count}(h_i, t_n)}$$

$$\mathrm{BP} = \min\left(1, \exp\left(1 - \frac{\mathrm{closest\_len}(\mathcal{R})}{\mathrm{len}(\mathcal{H})}\right)\right)$$

実験の流れ

実験結果

・Training Loss



横軸：Training Step
縦軸：Training Loss

Bleu score：16.94

実験の流れ

モデルパラメータ

パラメータ	値
Input_size	12973
Hidden_Size	1024
Output_size	6075
Embedding_size	300
Batch_size	32
Epoch	100
Loss Function	Cross Entropy
Optimizer	Adam
Learning Rate	0.001

発表の構成

1.はじめに

2.要素技術

3.データセット

4.実験の流れ

5.まとめと今後の課題

まとめと今後の課題

まとめ

・ダブルレイヤーLSTMで翻訳システムの実装

今後の課題

・日本語-中国語翻訳に関する取り組み

・Transformerの導入

・機械翻訳を漫画に利用する可能性の探索

ご清聴ありがとうございました