

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 偉齊

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
- 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

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



翻訳



セリフ



台詞

はじめに(研究目的)

手法

- ・LSTMの性能を確認するため,シングルレイヤーLSTMとダブルレイヤーLSTMで対照実験を実行

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

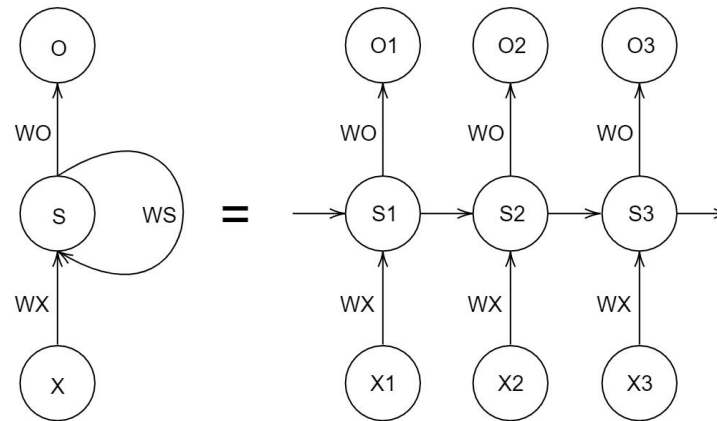
3.データセット

4.実験

5.まとめと今後の課題

要素技術

Recurrent Neural Network (RNN)

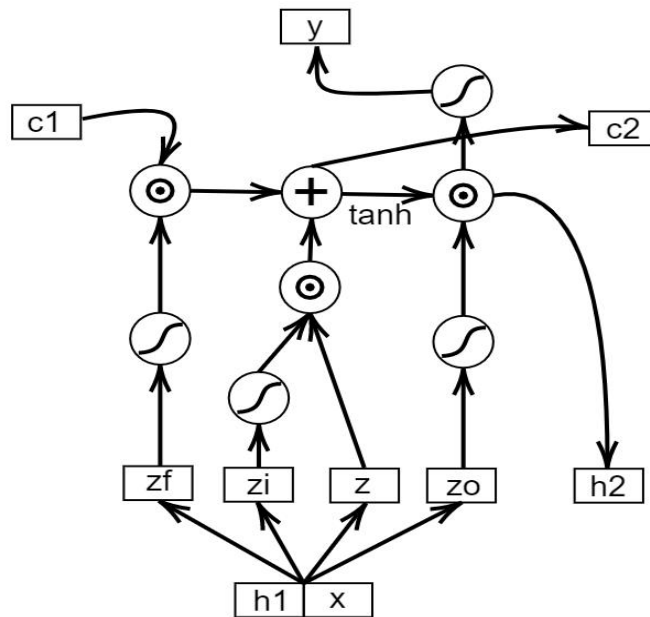


- ・回帰構造を持つニューラルネットワーク
- ・逆伝播による勾配消失と勾配爆発問題

・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.

要素技術

Long Short-term Memory (LSTM)



- ・ゲート構造で勾配を制御

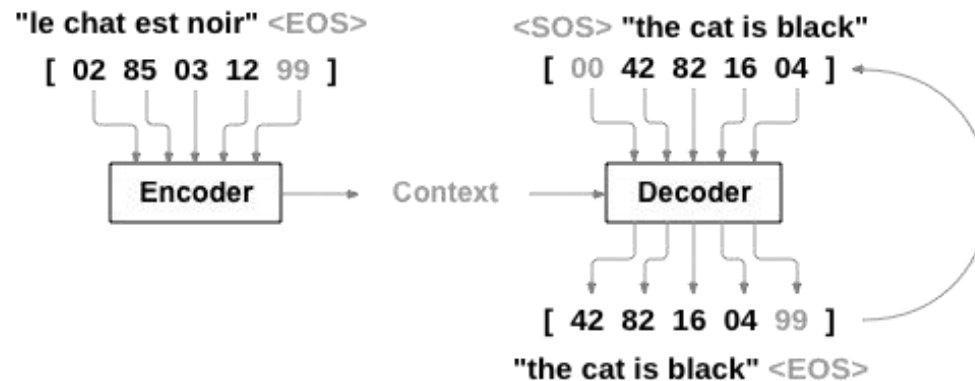
- ・メモリセル c で情報を記憶

- ・長期な記憶が可能

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

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
- ・本研究は,Pytorchチュートリアルのseq2seqモデル構造を使用

- ・ Ilya Sutskever, Oriol Vinyals, and Quoc V. Le. Sequence to sequence learning with neural networks,2014.
- ・ https://pytorch.org/tutorials/intermediate/seq2seq_translation_tutorial.html

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

- ・ "Jieba" (Chinese for "to stutter") Chinese text segmentation: built to be the best Python Chinese word segmentation module.
<https://github.com/fxsjy/jieba>.

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
- 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-中国語文を使用
- 全データセットは英語-中国語本文24360ペア

▪ Bilingual Sentence Pairs Selected Sentences from the Tatoeba Corpus. <http://www.manythings.org/bilingual/>.

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
- 5.まとめと今後の課題

実験

データ処理

- ・データセットテキストを4:1の比率で分ける

DataSet	Training	Testing
Pairs	19488	4872

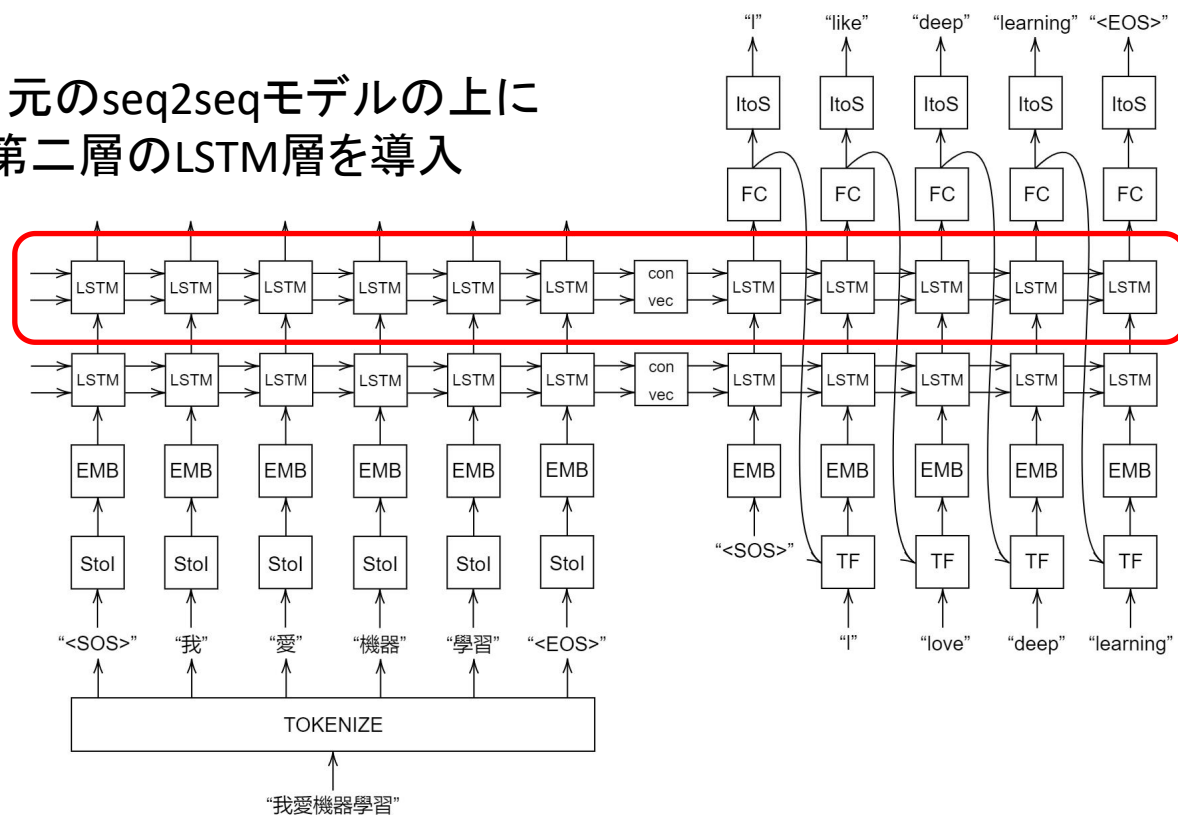
- ・Tokenizeで単語のディクショナリを構築する

DataSet	Training	Testing
Cmn_Vocab	12973	5814
Eng_Vocab	6750	3541

実験

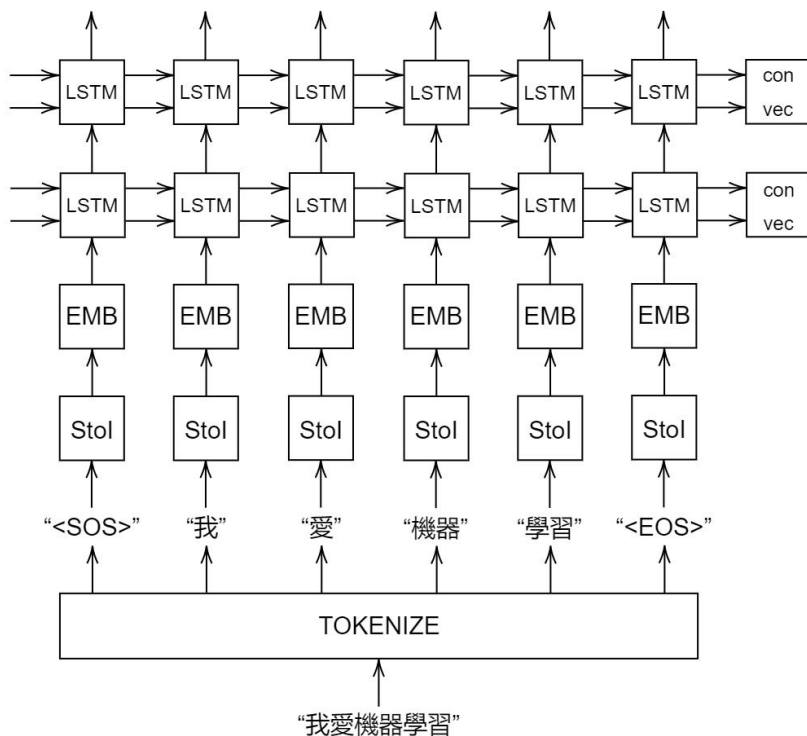
モデルの実装

- ・元のseq2seqモデルの上に第二層のLSTM層を導入



実験

Encoder



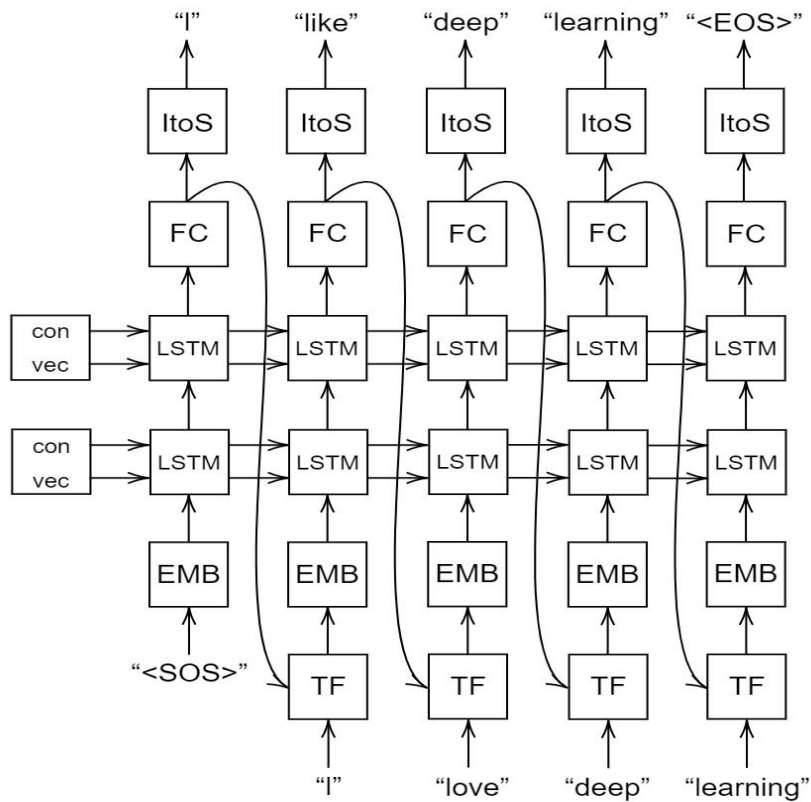
- ・始めと終わりのトークン<SOS>と<EOS>を加入

- ・二層LSTMで実装

- ・Context Vectorは出力 h と記憶 c を示す

実験

Decoder

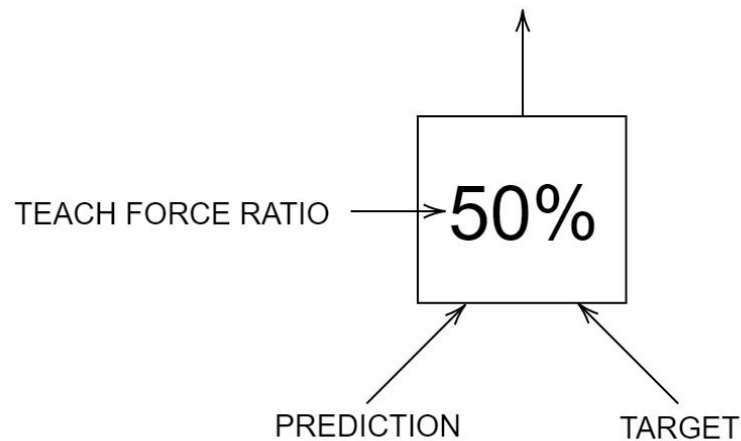


- Encoderの構造に Teach Force Ratioを導入

- 最初の入力は <SOS>だけ

実験

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

評価指標

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

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

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

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

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

実験

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

実験

実験結果

▪ Training Loss(Double LSTM)

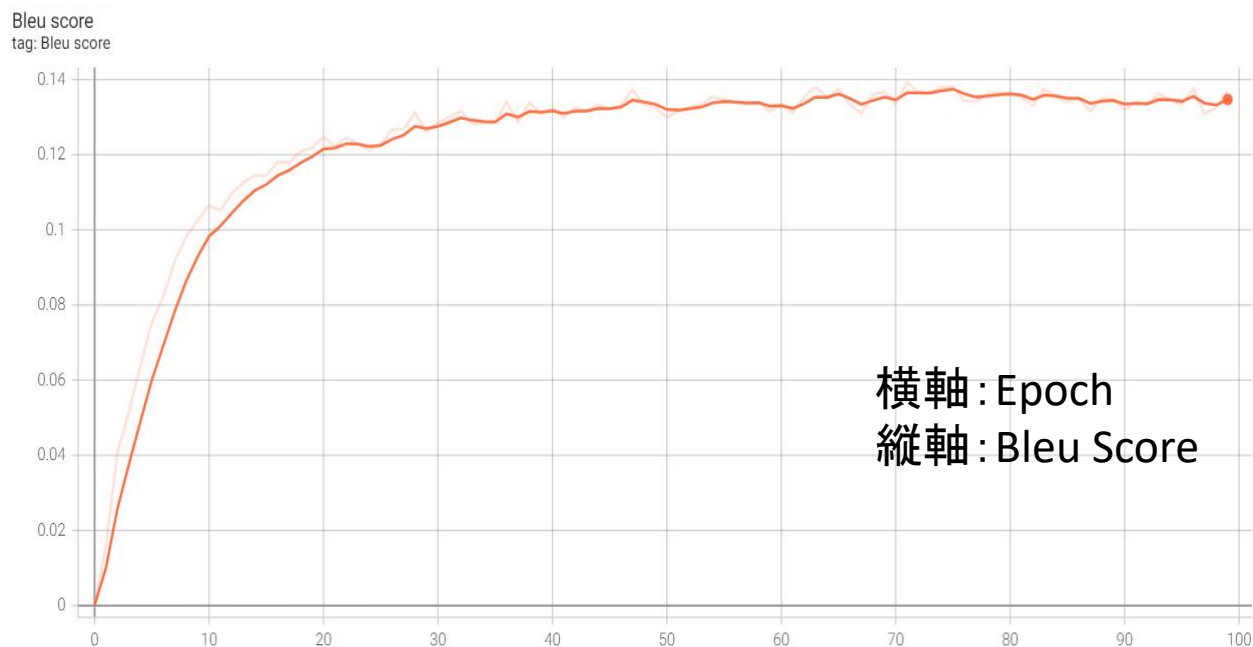


Training Loss 0.1995

実験

実験結果

▪ Testing Accuracy(Double LSTM)

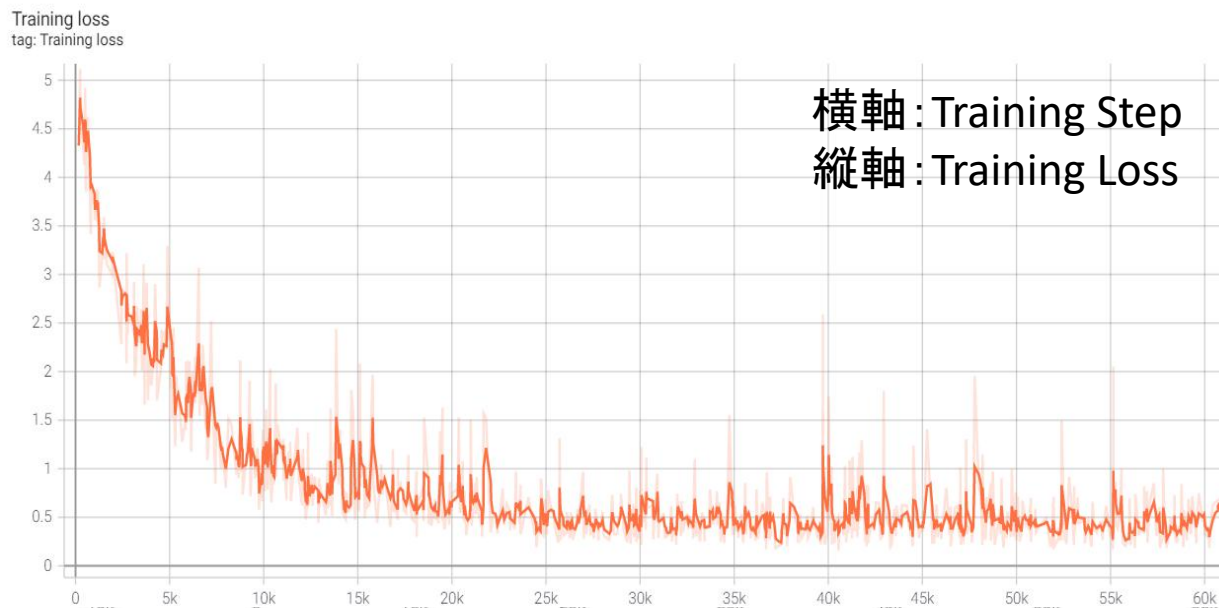


Bleu score: 0.1369

実験

実験結果

▪ Training Loss(Single LSTM)

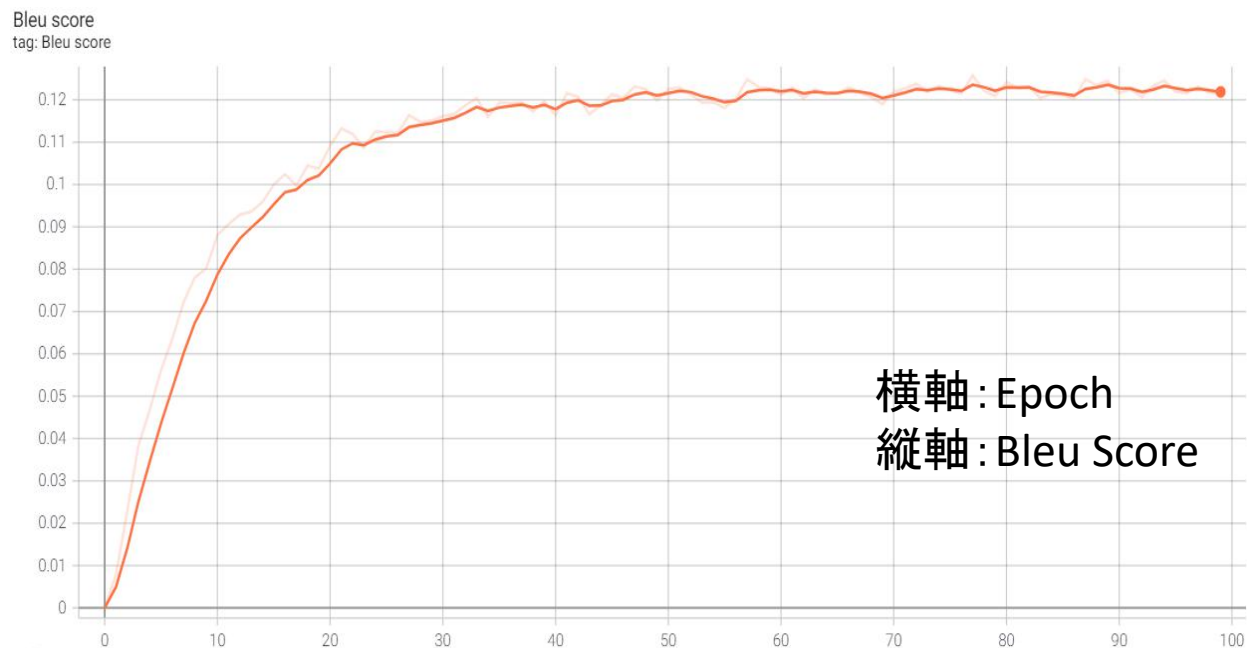


Training Loss 0.6991

実験

実験結果

▪ Testing Accuracy(Single LSTM)



Bleu score:0.1219

実験

翻訳結果

Double LSTM

Source Sentence	孩子們在公園裏玩。
Actual Translation	The children are playing in the park.
Prediction	the children were playing in the park .

Single LSTM

Source Sentence	孩子們在公園裏玩。
Actual Translation	The children are playing in the park.
Prediction	the children were having kites in the park .

実験

翻訳結果

Double LSTM

Source Sentence	我是超人。
Actual Translation	I am Superman.
Prediction	i'm left-handed .

Single LSTM

Source Sentence	孩子們在公園裏玩。
Actual Translation	I am Superman.
Prediction	i'm a new student .

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
- 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
- ・同じデータセットでダブルレイヤーLSTMとシングルレイヤーLSTMの対照実験を実行
- ・実験結果として,ダブルレイヤーLSTMは訓練誤差は0.19に収束し,テストアキュラシーは0.13に収束,故にダブルレイヤーLSTMは一定的に性能アップ

まとめと今後の課題

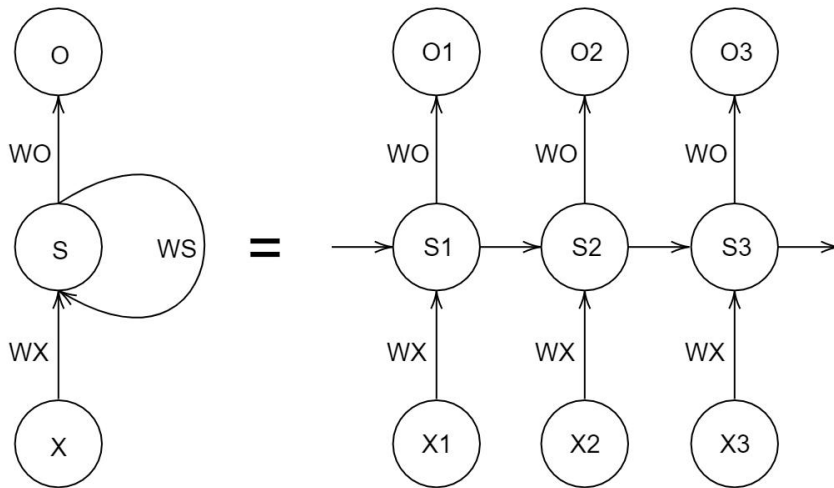
今後の課題

- ・ASPEC-JCデータセットで翻訳システムの構築
- ・最先端のモデル(Transformerや、BERTなど)の導入と性能の確認
- ・機械翻訳を漫画に利用する可能性を探索

ご清聴ありがとうございました

要素技術

Recurrent Neural Network (RNN)



皆分かってる
こういうモデルです

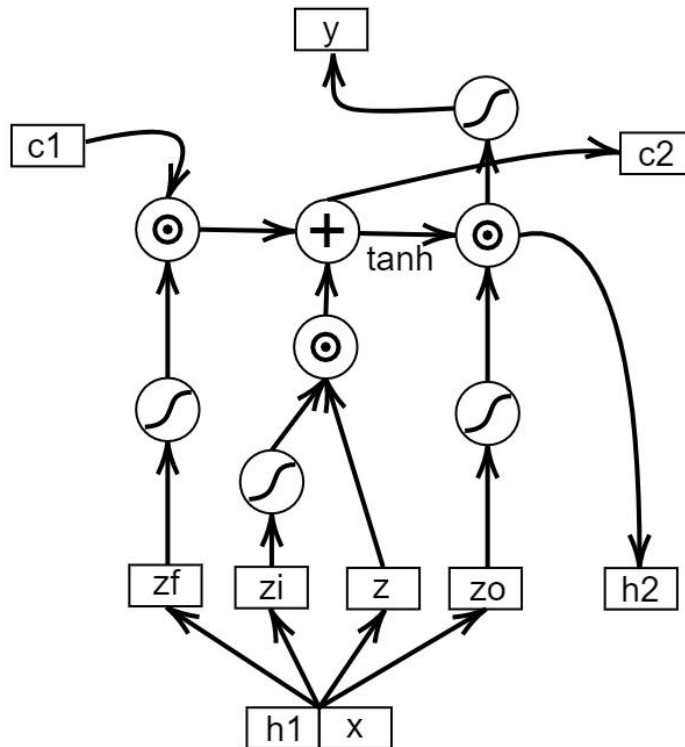
$$O_t = g(W_o \cdot S_t)$$

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

- ・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)



斜めしていけない
式のfrontチェック

$$C_t = Z_f \odot C_{t-1} + Z_i \odot Z$$

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

$$y_t = \sigma(Wh_t)$$