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.提案手法

5.実験

6.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.提案手法

5.実験

6.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

・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

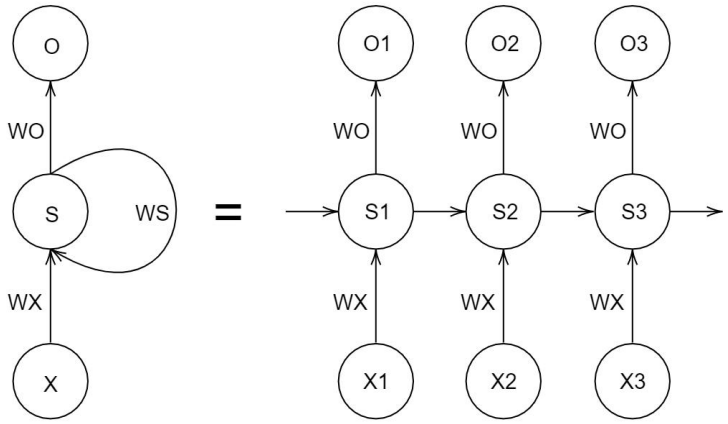
4.提案手法

5.実験

6.まとめと今後の課題

要素技術

Recurrent Neural Netword (RNN)



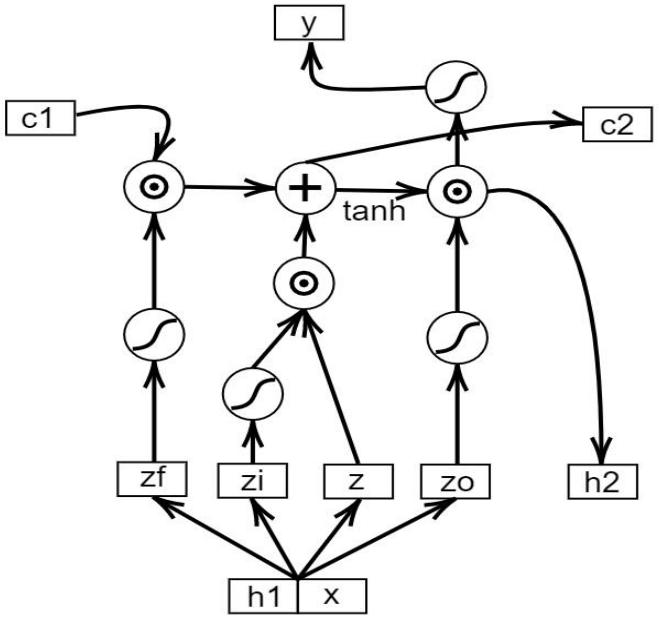
・回帰構造を持つニューラルネットワーク

・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

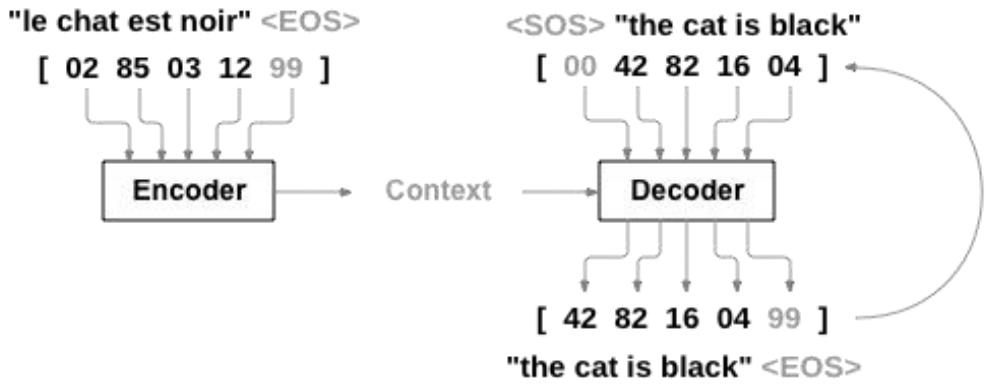
・メモリセルcで情報を記憶

・長期な記憶が可能

・Sepp Hochreiter and J¨urgen 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

発表の構成

1.はじめに

2.要素技術

3.データセット

4.提案手法

5.実験

6.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-中国語ペアを使用

・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

データセット

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.提案手法

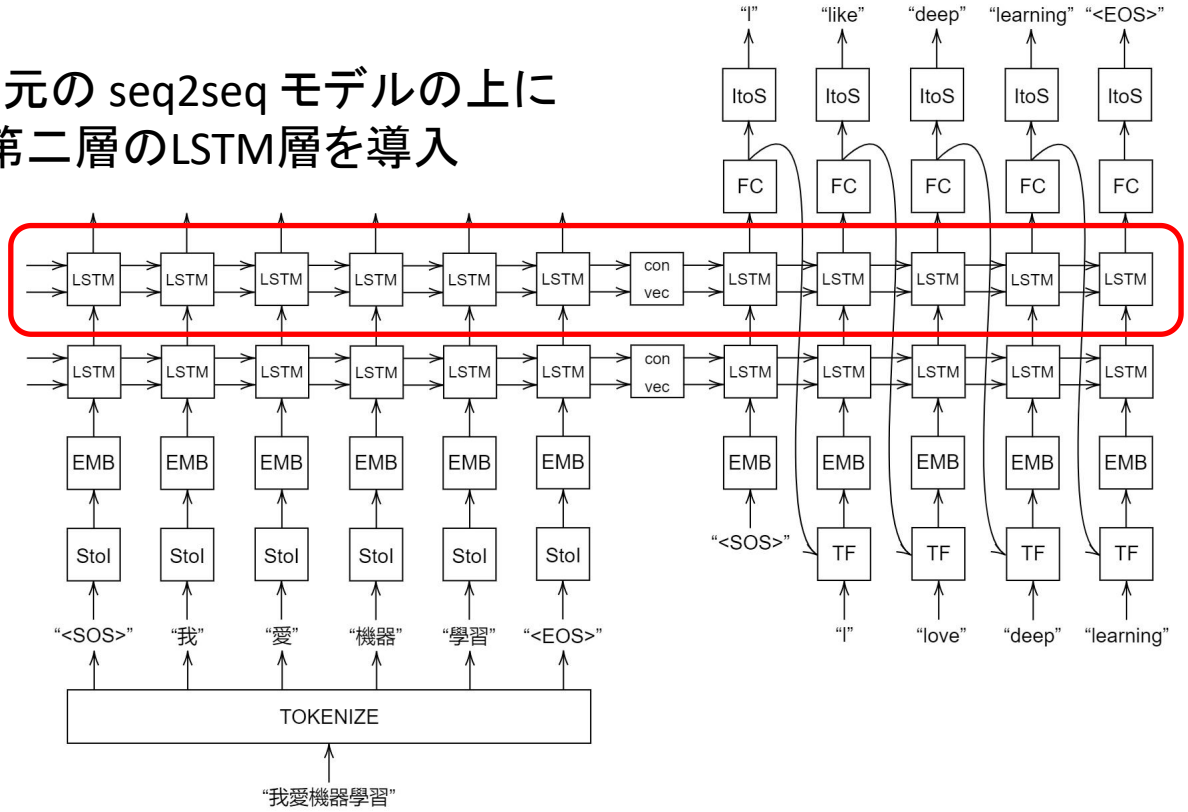
5.実験

6.まとめと今後の課題

提案手法

ダブルレイヤー LSTM

・元の seq2seq モデルの上に
第二層のLSTM層を導入



発表の構成

1.はじめに

2.要素技術

3.データセット

4.提案手法

5.実験

6.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

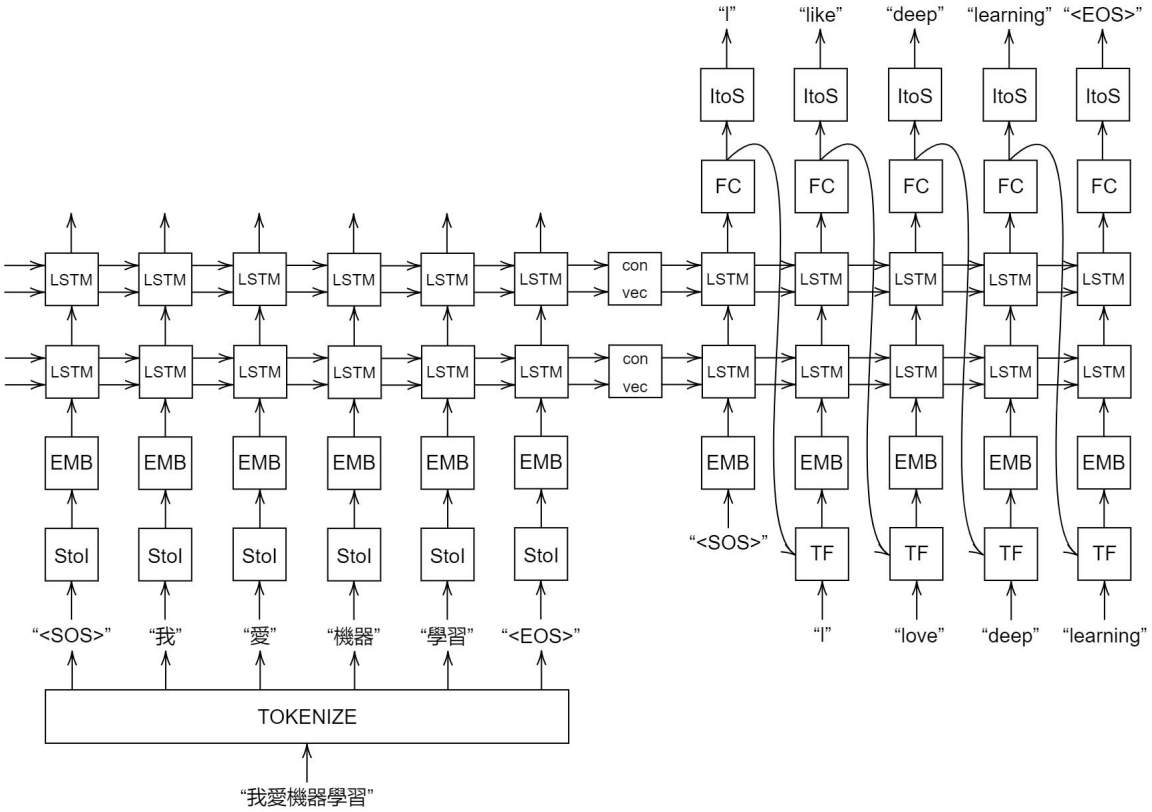
DataSet	Training	Testing
Pairs	19488	4872

・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

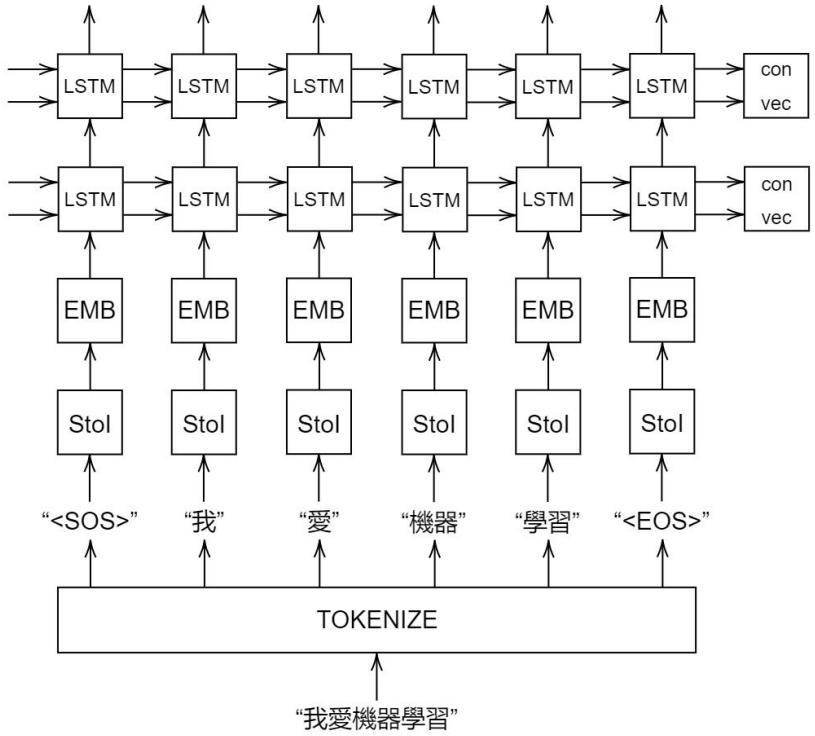
実験

モデルの実装


“I”
“like”
“deep”
“learning”
“<EOS>”
ItoS
FC
LSTM
con vec
EMB
StoI
“<SOS>”
“我”
“愛”
“機器”
“學習”
“<EOS>”
TF
“I”
“love”
“deep”
“learning”
TOKENIZE
“我愛機器學習”

実験

Encoder



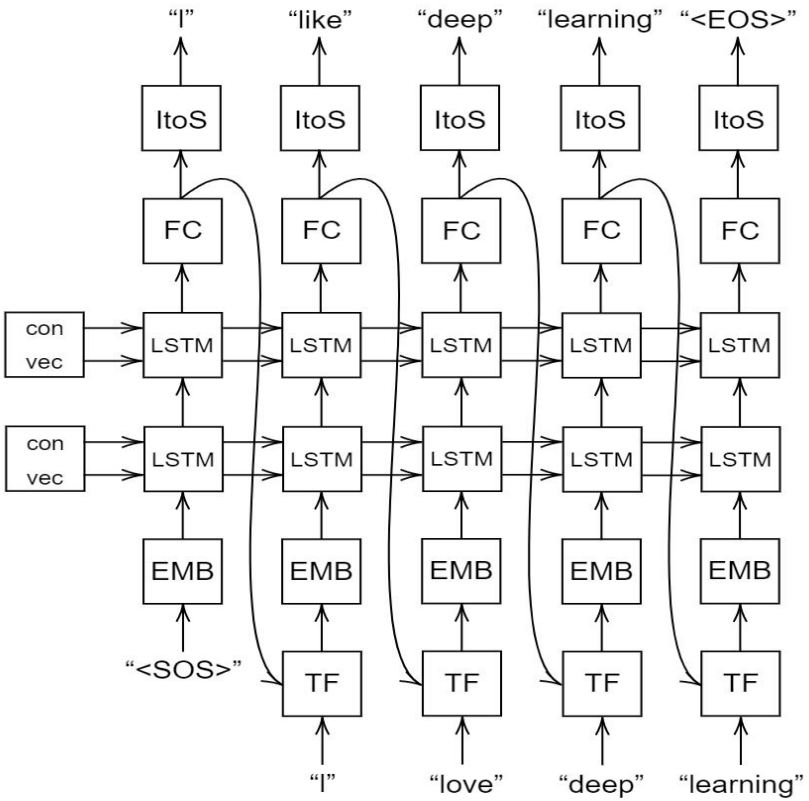
・始めと終わりのトークン <SOS> と <EOS> を加入

・二層 LSTM で実装

・Context Vector は出力 **h** と記憶 **c** を示す

実験

Decoder

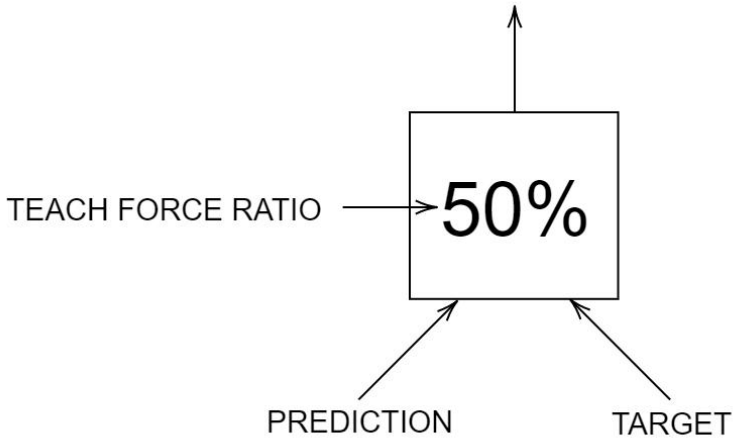


・Encoder の構造に Teach Force Ratio を導入

・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

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

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

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

・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

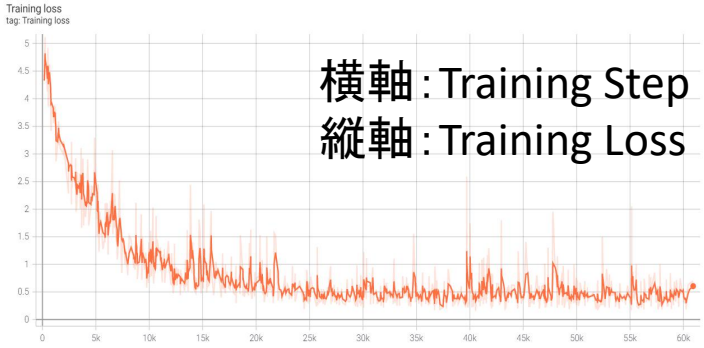
実験結果

・Training Loss

Double LSTM

Training Loss 0.1995

Single LSTM



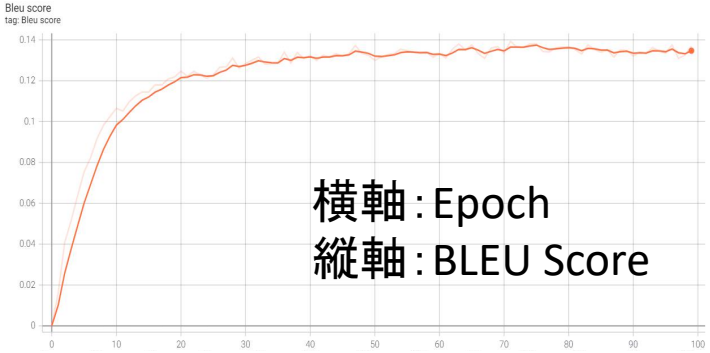
Training Loss 0.6991

実験

実験結果

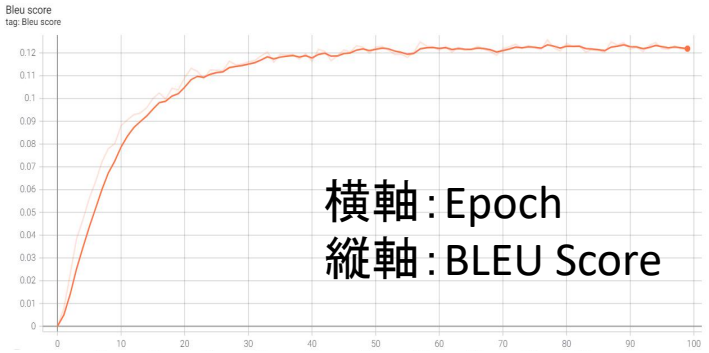
・Testing Accuracy

Double LSTM



BLEU Score 0.1396

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 , テストアキュラシィは
0.13 に収束 , 故にダブルレイヤー LSTM はある
程度に性能が上回る

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

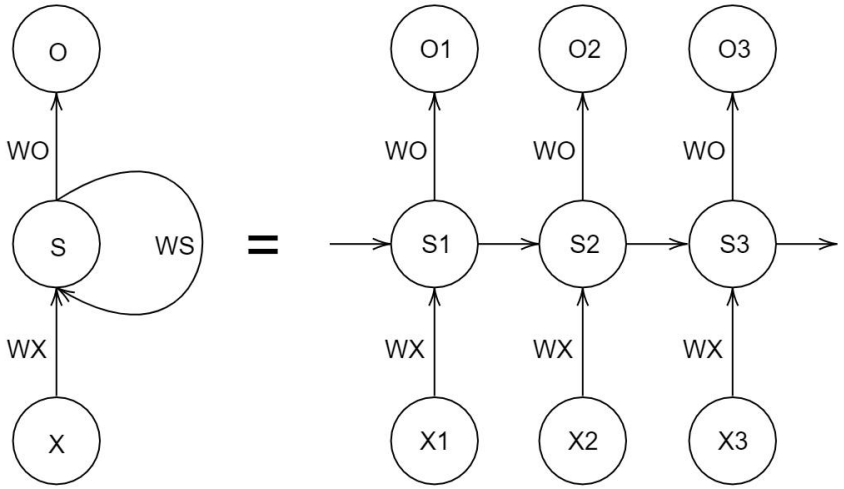
・最先端のモデル (Transformer や , BERT など) の導入と性能の確認

・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 Netword (RNN)



皆分かてる
こういうモデルです

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

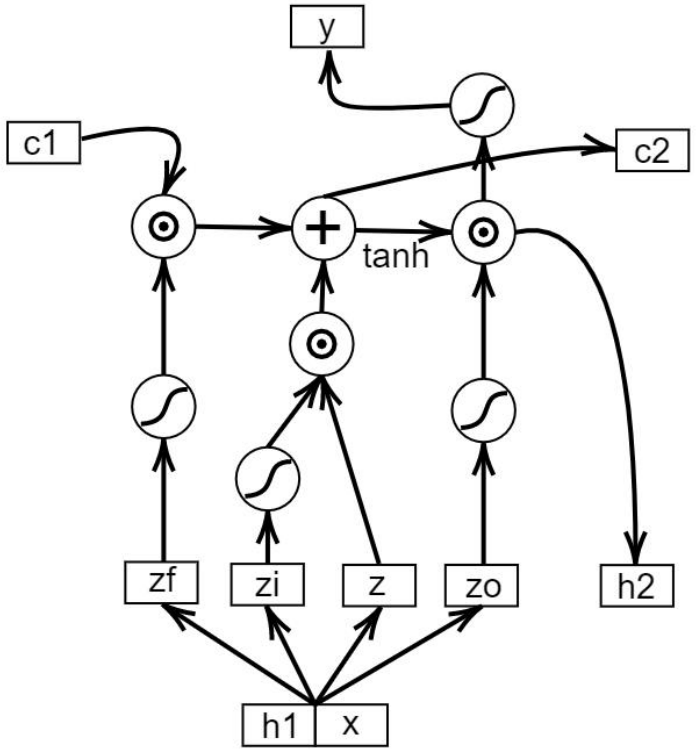
$$S_t = f(W_x \cdot X_t + W_s \cdot S_{t-1})$$

・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

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

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

$$y_t = \sigma(W h_t)$$