

# 深層学習に基づく 4コマ漫画の感情推定と マルチモーダル化への検討

ソフトウェアシステム研究グループ

B4 高山 裕成

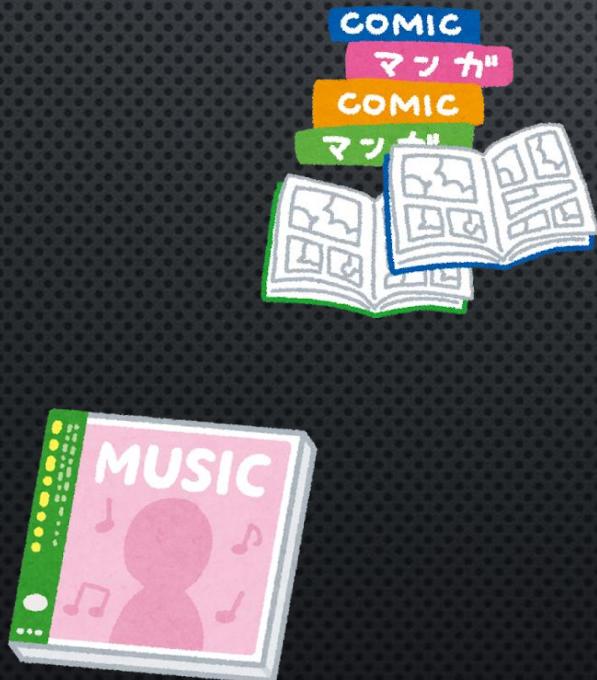
# 発表の流れ

- はじめに
- 提案手法
- データセット及び要素技術
- 実験手順・設定
- 結果・考察
- まとめと今後の課題

# 発表の流れ

- はじめに
- 提案手法
- データセット及び要素技術
- 実験手順・設定
- 結果・考察
- まとめと今後の課題

# はじめに



# はじめに

## 漫画



# はじめに

## 漫画



[自然言語]  
セリフ・オノマトペ

[画像]  
キャラクタ・背景

©作画：棟田ウメコ  
(シナリオ：(株) スポマ播村早紀/大阪工業大学 上野未貴)

# はじめに

## 漫画



[自然言語]  
セリフ・オノマトペ

[画像]  
キャラクタ・背景

©作画：棟田ウメコ  
(シナリオ：(株) スポマ播村早紀/大阪工業大学 上野未貴)

# はじめに

[研究目的]

人工知能を用いた漫画の内容理解

[タスク]

漫画における

キャラクタのセリフの感情推定



# はじめに

[アプローチ]

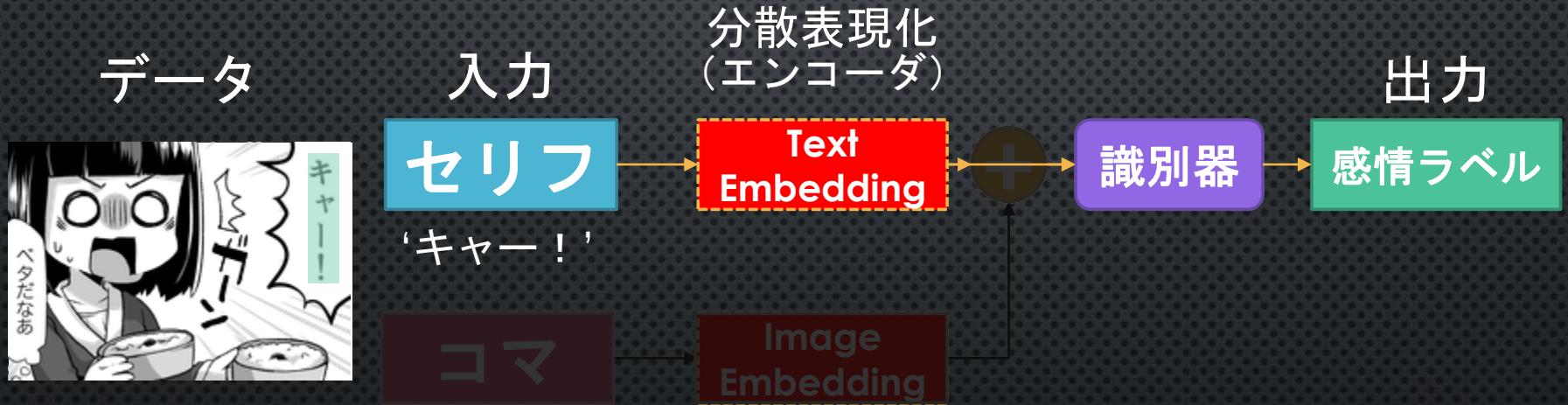
1. 自然言語処理のみを用いた推定
2. 画像情報を加えたマルチモーダルな推定



# 発表の流れ

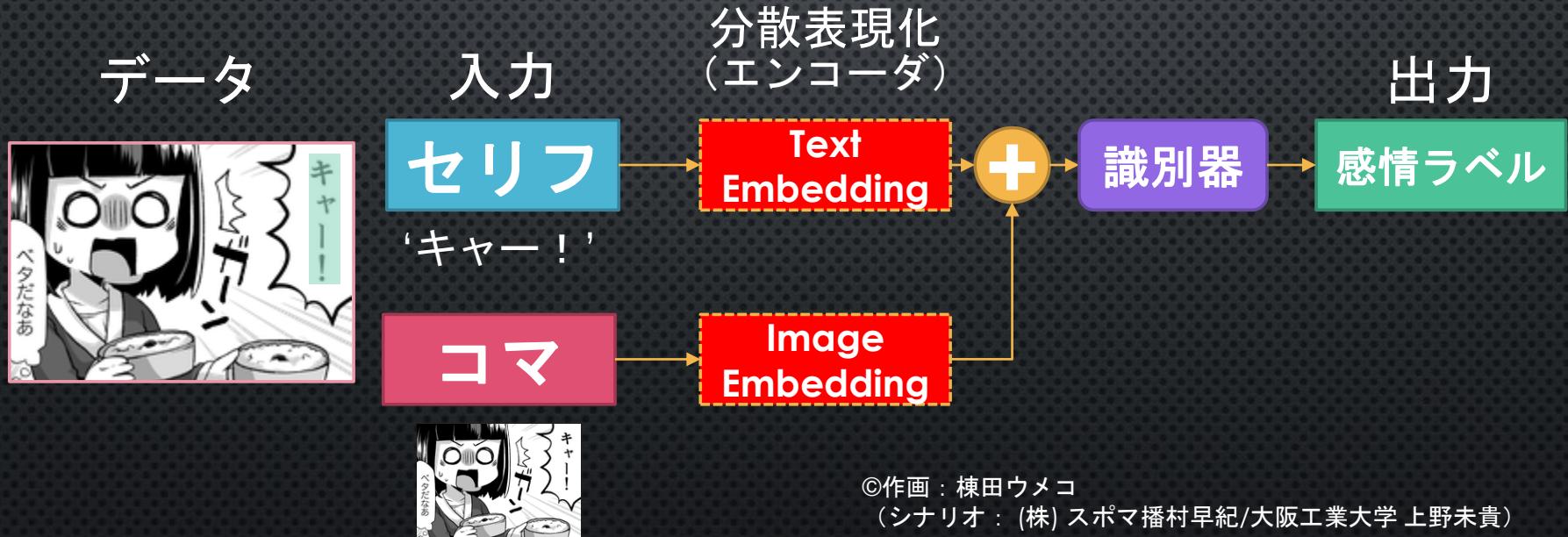
- はじめに
- 提案手法
- データセット及び要素技術
- 実験手順・設定
- 結果・考察
- まとめと今後の課題

# 漫画のセリフの マルチモーダルな感情推定手法



©作画：棟田ウメコ  
(シナリオ：(株) スポマ播村早紀/大阪工業大学 上野未貴)

# 漫画のセリフの マルチモーダルな感情推定手法



# 発表の流れ

- はじめに
- 提案手法
- データセット及び要素技術
- 実験手順・設定
- 結果・考察
- まとめと今後の課題

# 4コマ漫画ストーリーデータセット

## 少女漫画タッチ

感情:ニュートラル



感情:喜楽

©作画：高科りさ

(シナリオ：(株) スポマ播村早紀/大阪工業大学 上野未貴)

## 萌えタッチ

感情:喜楽



感情:喜楽

©作画：棟田ウメコ

(シナリオ：(株) スポマ播村早紀/大阪工業大学 上野未貴)

- 複数の漫画家が同一プロットの下、作成
- 複数のタッチが存在
- 作者によるセリフの感情アノテーション

4コマ漫画ストーリーデータセット（創作者と人工知能: 共作実現に向けた  
創作過程とメタデータ付与 4コマ漫画ストーリーデータセット） 上野未貴

- ・タッチ (全 5 種類)

ギャグタッチ 少女漫画タッチ 少年漫画タッチ  
青年漫画タッチ 萌えタッチ ※各 10 話ずつ

- ・感情ラベル (全 8 種類)

ニュートラル 喜楽 驚愕 恐怖 悲哀 憤怒 嫌惡 UNK

- ・タッチ (全 5 種類)

ギャグタッチ 少女漫画タッチ 少年漫画タッチ  
青年漫画タッチ 萌えタッチ ※各 10 話ずつ

- ・感情ラベル (全 7 種類)

ニュートラル 喜楽 驚愕 恐怖 悲哀 憤怒 嫌惡 UNK

# Data Augmentation

日本語 WordNet のシソーラスを用いてデータを拡張

※同時に変換するのは最大で1単語のみ

オリジナル:

Bさんはいつも優しいな

Bさんは何時も優しいな

Bさんはいつも思いやりがあるな

Bさんはいつでも優しいな

Bさんはいつも温かいな

:

:

:

:

:

:

# BERT

Text  
Embedding

- Bidirectional Encoder Representations from Transformer
- 複数の双向向 Transformer に基づく汎用言語モデル
- 事前学習済みモデルを他のタスクに転移させることが容易

Devlin, Jacob and Chang, Ming-Wei and Lee,  
Kenton and Toutanova, Kristina.  
BERT: Pre-training of Deep Bidirectional  
Transformers for Language Understanding. 2018.

BERT

Text  
Embedding

- ・ [京大 BERT]

京都大学 黒橋・河原研究室が公開している  
日本語 Wikipedia コーパスを用いたモデル

- ・ [hottoSNS-BERT]

hottolink 社が公開している  
大規模日本語 SNS コーパスを用いたモデル

柴田 知秀, 河原 大輔, 黒橋 祐夫,  
BERT による日本語構文解析の精度向上,  
言語処理学会 第25回年次大会, pp.205-208, 2019

Sakaki, Takeshi and Mizuki, Sakae and Gunji, Naoyuki ,  
BERT Pre-trained model Trained on Large-scale Japanese  
Social Media Corpus, 2019

# Illustration2Vec

Image  
Embedding

- Saito, Matsui らが提案した画像のベクトル化手法
- アニメや漫画のようなイラストに対する  
画像認識の難しさに取り組む
- 筆者らによって公開されている  
イラスト投稿サイトから約 100 万枚を用いて  
学習した事前学習済みモデルを使用

Masaki Saito, Yusuke Matsui,  
“Illustration2Vec: A Semantic Vector Representation  
of Illustrations”, 2015

# 発表の流れ

- はじめに
- 提案手法
- データセット及び要素技術
- 実験手順・設定
- 結果・考察
- まとめと今後の課題

# 実験

1. セリフ 1 文のみを入力とする感情推定
2. マルチモーダルな感情推定の検討

# 実験

正例クラス：喜楽　負例クラス：その他　とした  
2 クラスのセリフの感情推定を設定

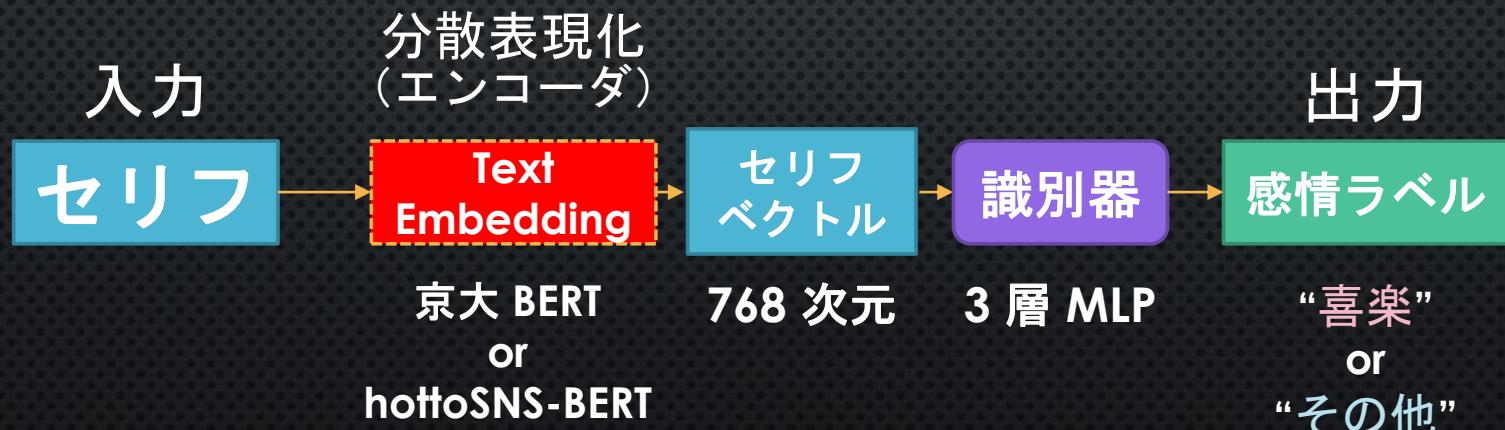
ニュートラル　喜楽　驚愕　恐怖　悲哀　憤怒　嫌悪

各セリフには 1 個の感情ラベルが付与

各タッチそれぞれについてモデルを作成・評価

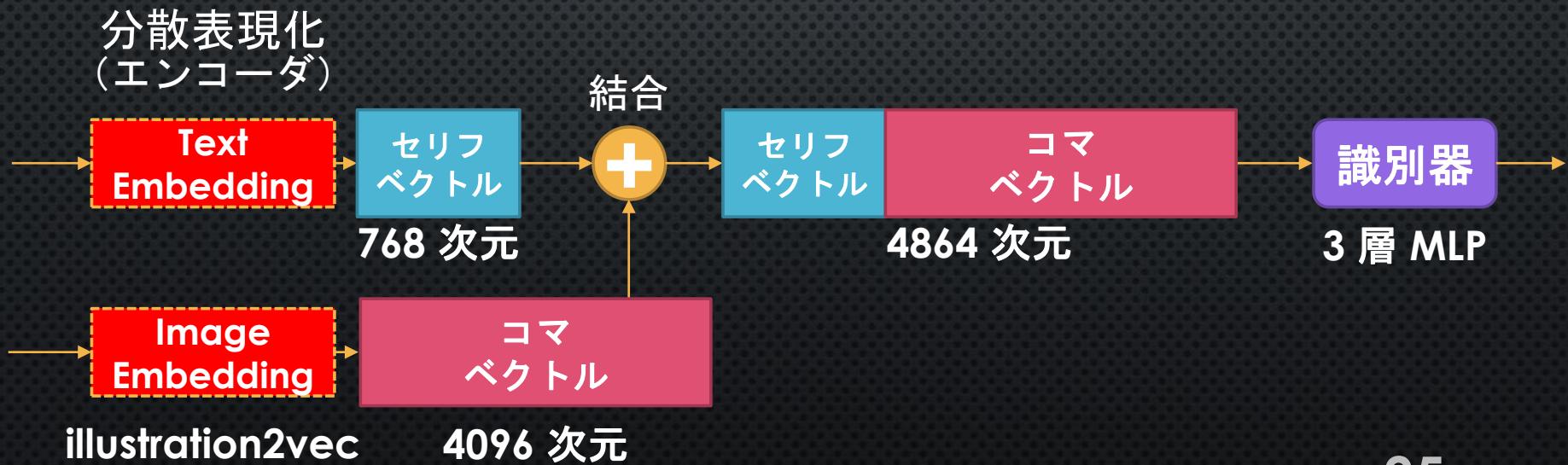
# 実験 1：セリフ 1 文のみを入力とする感情推定

- ・識別器は多層パーセプトロン (MLP)
- ・BERT モデルは最終層のみを fine-tuning
- ・京大 BERT と hottoSNS-BERT で精度比較



## 実験 2：マルチモーダルな感情推定の検討

- ・BERT モデルは最終層のみを fine-tuning
- ・京大 BERT と hottoSNS-BERT で精度比較
- ・コマベクトルは固定



# 実験設定 (実験 1・2)

- ・訓練用データ 1～5 話 (拡張後)  
・訓練：検証 = 8 : 2
- ・評価用モデル 検証用データにおける正例の F 値が最大となるパラメータを採用
- ・評価用データ 6～10 話  
(オリジナルのセリフのみ)

# データ数

	ラベル	ギャグ	少女漫画	少年漫画	青年漫画	萌え
訓練 + 検証	喜楽	15	39	15	18	25
	その他	40	26	45	44	35
	喜楽 (拡張後)	<b>1115</b>	<b>2575</b>	<b>940</b>	<b>998</b>	<b>1766</b>
	その他 (拡張後)	<b>2851</b>	<b>1391</b>	<b>3076</b>	<b>3145</b>	<b>2323</b>
評価	喜楽	10	38	12	14	22
	その他	56	29	52	51	42

訓練 + 検証用データ

Data Augmentation によって約 65 倍

# パラメータ

## [実験 1 | 2] 設定

エポック数	50
バッチサイズ	16
最適化手法	Adam
損失関数	cross entropy loss
クラス重み	ラベル比の逆比
学習率	optuna で調整

## [実験 1 | 2] 3層 MLP

入力層次元	768   4864
隠れ層次元	30   300
出力層次元	2
活性化関数	tanh <sup>*1</sup>
Dropout 率	0.5

\*1 出力層では softmax

Takuya Akiba, Shotaro Sano, Toshihiko Yanase,  
Takeru Ohta, and Masanori Koyama. Optuna:  
A next-generation hyperparameter optimization  
framework. In Proceedings of the 25rd ACM  
SIGKDD International Conference on Knowledge  
Discovery and Data Mining, 2019.

# 発表の流れ

- はじめに
- 提案手法
- データセット及び要素技術
- 実験手順・設定
- 結果・考察
- まとめと今後の課題

# 実験 1

	5 タッチ総合		
	Acc	P-Recall <sup>*1</sup>	P-F 値 <sup>*2</sup>
京大 BERT	0.733	0.521	0.535
hottoSNS-BERT	0.758	0.583	0.586
ベースライン <sup>*3</sup>	0.705	—	—

\*1  
P-Recall:

正例の再現率

\*2  
P-F 値:

正例の F 値

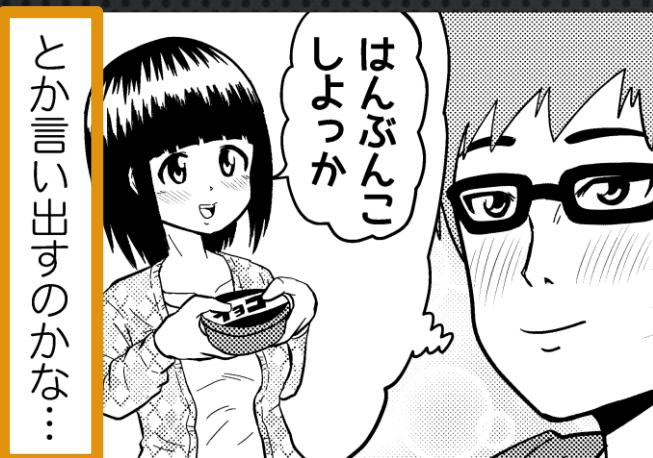
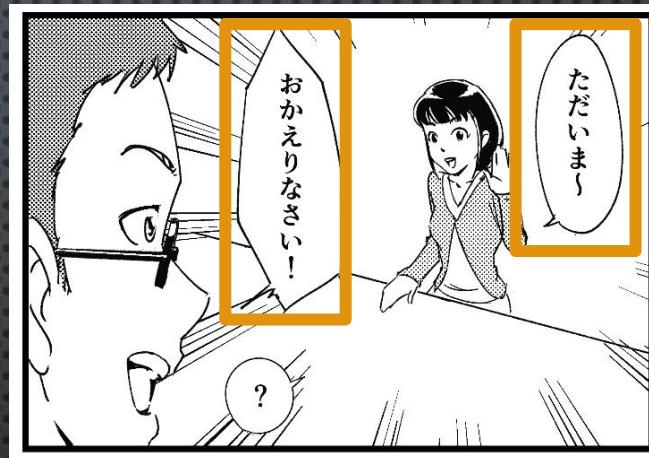
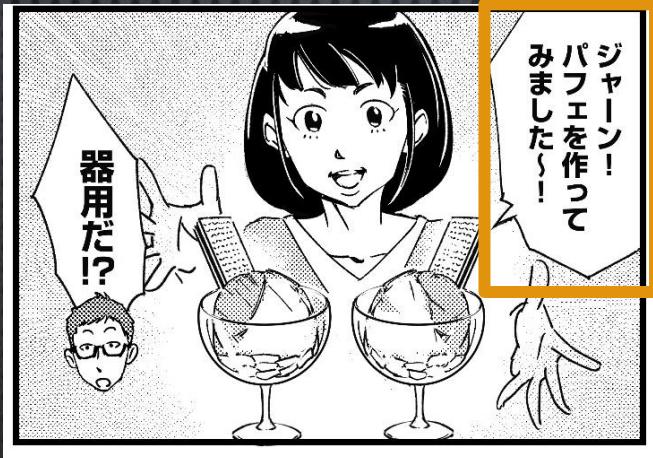
\*3  
ベースライン:

すべての予測値が  
負例の場合

# 実験 2

	5 タッチ総合		
	Acc	P-Recall	P-F 値
京大 BERT	0.715 ↓	0.583 ↑	0.546 ↑
hottoSNS-BERT	0.776 ↑	0.625 ↑	0.622 ↑
ベースライン	0.705	—	—

## 実験 2：マルチモーダル化による影響



# 発表の流れ

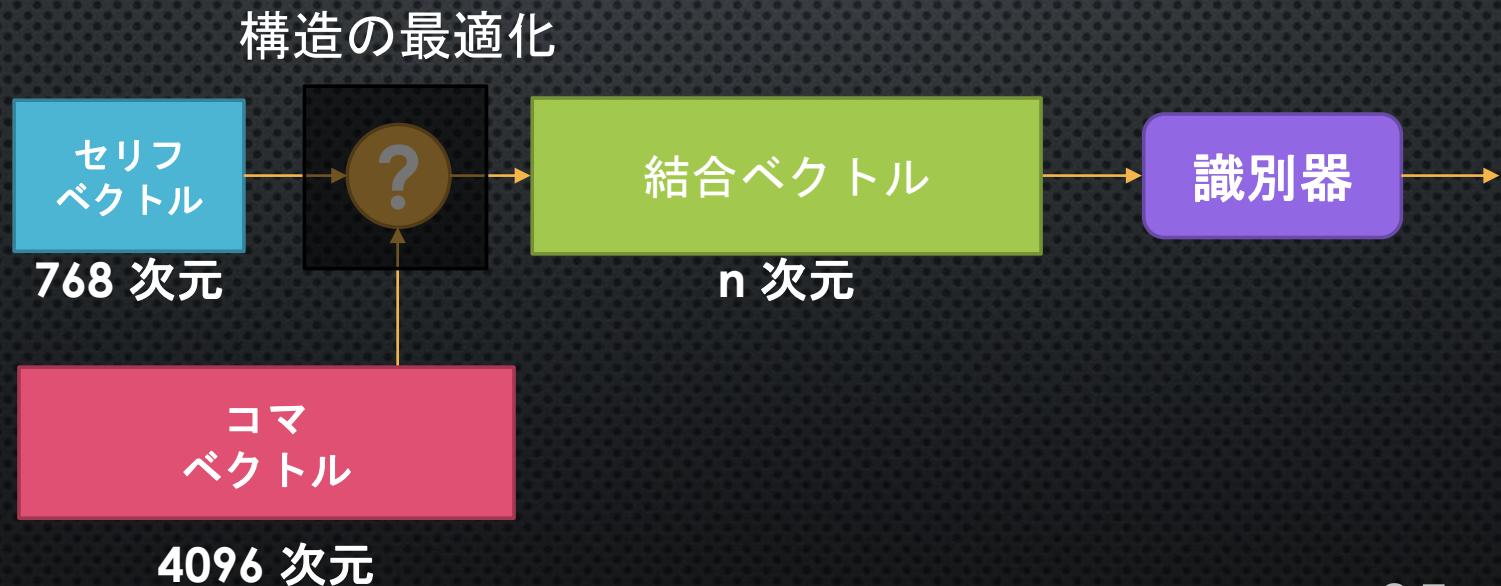
- はじめに
- 提案手法
- データセット及び要素技術
- 実験手順・設定
- 結果・考察
- まとめと今後の課題

# まとめ

- ・BERT で得た分散表現からセリフの感情推定を行った
  - ⇒ ・口語的なコーパスからなる hottoSNS-BERT の方が  
漫画のセリフのより合理的な分散表現が得られた
- ・画像情報を加味したマルチモーダルな感情推定を行った
  - ⇒ ・実験的にその有効性を確認
    - ・データの扱い方やネットワークの構築について  
更なる工夫が必要

# 今後の課題

- ・セリフベクトルとコマベクトルの結合方法の最適化  
⇒ Neutral Architecture Search (NAS) で探索



# 今後の課題

- ・データセットの問題

1. データ数の少なさ
2. アノテーション不備 (UNK ラベル)



## [解決策]

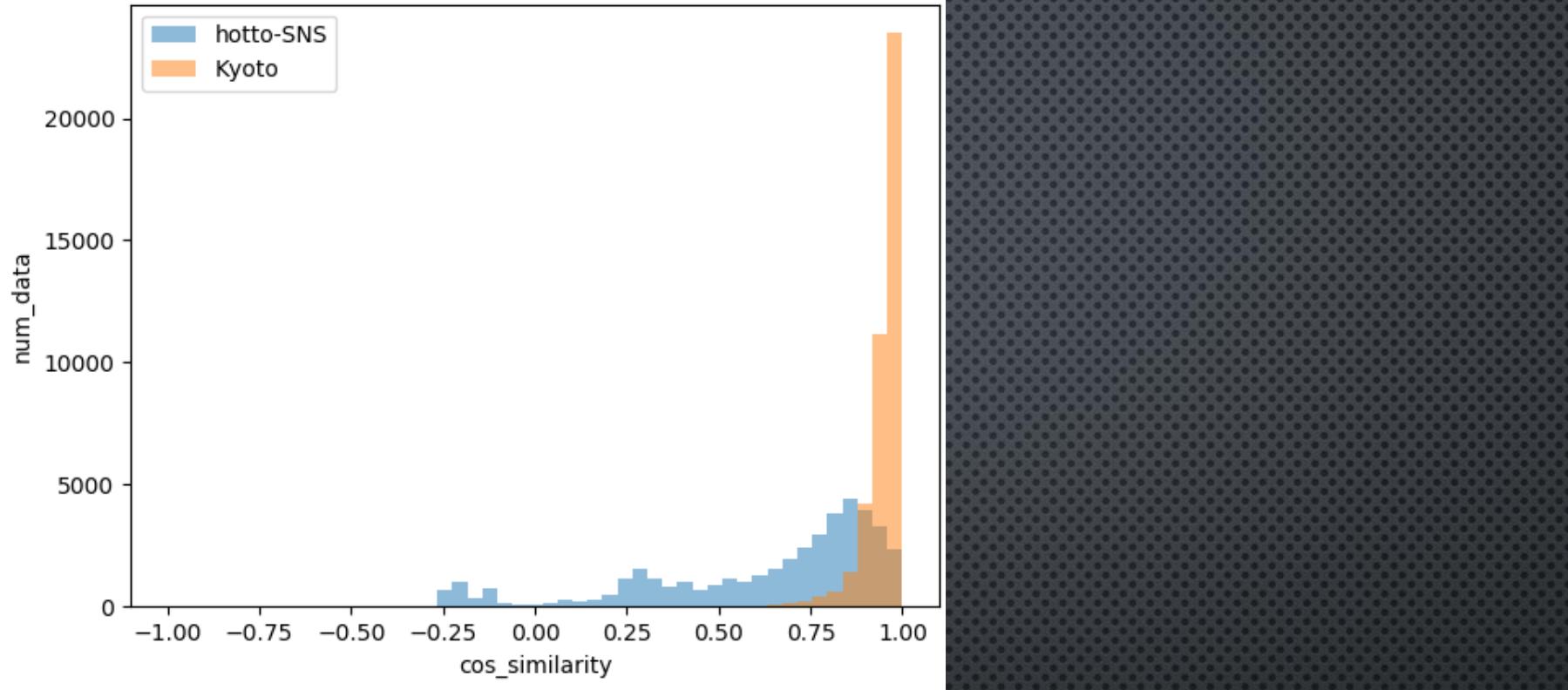
- ・Manga109 やその他データセットを併用した半教師あり学習
- ・人手によるデータセット拡張

ご清聴ありがとうございました



表 1: 拡張前データ

emotion	ギャグ	少女	少年	青年	萌え	合計
ニュートラル	43	8	55	33	30	169 (25.6%)
喜楽	25	77	27	33	47	209 (31.7%)
驚愕	19	16	17	29	20	101 (15.3%)
悲哀	25	12	13	16	13	79 (11.9%)
恐怖	6	11	8	8	9	42 (6.3%)
憤怒	4	5	2	7	2	20 (3.0%)
嫌悪	2	4	3	3	4	16 (2.4%)
UNK	7	3	5	2	6	23 (3.4%)
合計	131	136	130	131	131	659



オリジナル	拡張後のセリフ	京大 BERT	hottoSNS-BERT
去年は私が着たやつ	去年は私が着た若者	0.97	-0.19
ですか	正しくですか	0.89	-0.19
僕はいいんですけど、 気をつけてくださいね	僕はグーんですけど、 気をつけてくださいね	0.96	-0.16
	僕はいいんですけど、 真性をつけてくださいね	0.94	0.25

# 4 コマ漫画に関する研究

[1] Nomura, Arai

- ・遺伝的アルゴリズムに基づく感性解析

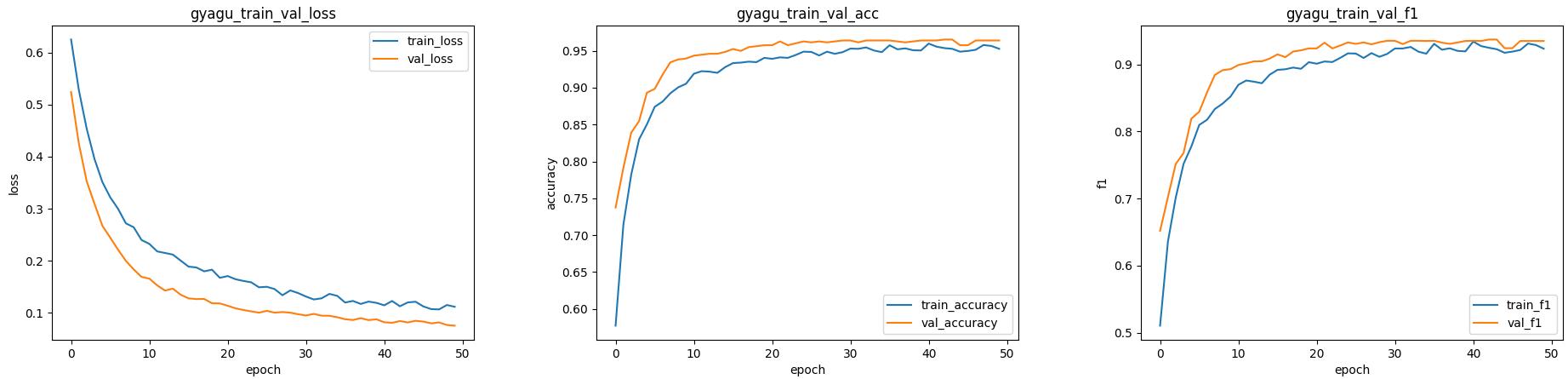
[2] Fujino, Mori, Matsumoto

- ・コマの順序識別に関する研究

[1] 野村俊太 & 荒井幸代. 進化計算を用いた人間の感性理解のための遺伝子解析法.  
In 第 31 回人工知能学会全国大会発表論文集, pages 3H2–OS–04b–2, 2017.

[2] 藤野紗耶, 森直樹, and 松本啓之亮. 3 分岐型畳み込みニューラルネットワークによる4 コマ漫画の順序識別. In 人工知能学会全国大会論文集 2018 年度人工知能学会全国大会(第32回)論文集, pages 1K1OS2a05–1K1OS2a05. 一般社団法人人工知能学会, 2018.

# 実験 1 学習推移



ギャグ漫画タッチ  
(hottoSNS-BERT)

# はじめに

[セリフの感情の種類]

- ・作者が設定した感情
- ・キャラクタ自身の感情
- ・読者が感じる感情

# Word2Vec

単語の分散表現を得る手法

- Continuous Bag-of-Words

周辺の単語から中心の単語を予測するように学習

- Skip-gram

ある単語から周辺の単語を予測するように学習

# Doc2Vec

Word2Vec を文章に拡張した手法

文章の分散表現を得る

- **Distributed Memory**

文書 ID + 周辺の単語から中心の単語を予測する  
ように学習

- Distributed Bag-of-Words

文書 IDのみから文書に含まれる単語を予測する  
ように学習

# Long short-term memory (LSTM)

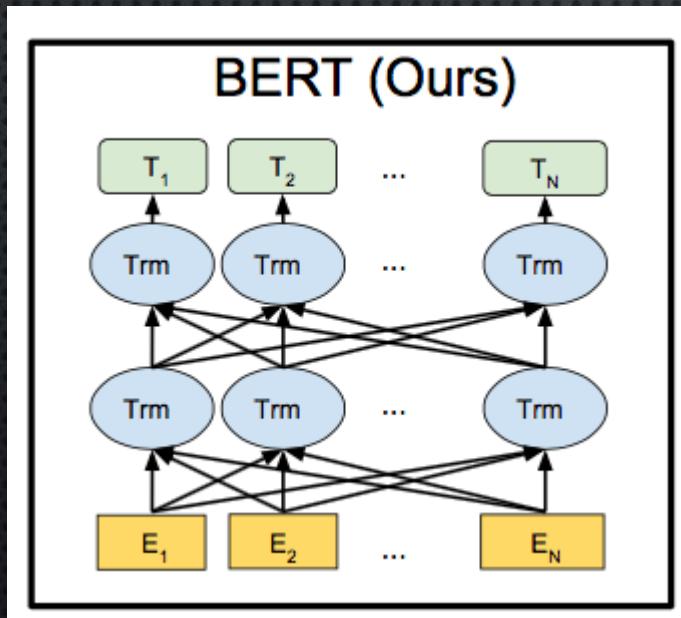


## [RNN 問題点]

- ・長期依存性
- ・勾配消失, 勾配爆発

# BERT

- Bidirectional Encoder Representations from Transformer
- 複数の双方向 Transformer に基づく汎用言語モデル
- 事前学習済みモデルを他のタスクに転移することが容易



Devlin, Jacob and Chang, Ming-Wei and Lee, Kenton and Toutanova, Kristina.  
BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. 2018.

# BERT

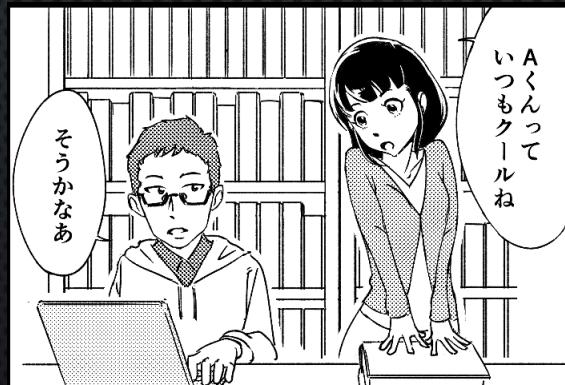
## 事前学習済みモデル

- 京都大学 黒橋・河原研究室 が公開
- 日本語 Wikipedia (約 1,800 万文) を用いて学習
- 設定値
  - Transformer の層数  $L = 12$
  - 隠れ状態ベクトルの次元数  $H = 768$
  - Multi-head Self-Attention のヘッド数  $A = 12$   
( 英語学習済みモデル BERT-Base と同一 )

# Illustration2Vec

Image  
Embedding

- 筆者らによって公開されている Danbooru・Safebooru から約 100 万枚のイラストを用いて学習した事前学習済みモデルを使用
- 本研究ではコマ画像のエンコーダとしてのみ用いる



タグ予測

('monochrome', 0.83),  
(‘1girl’, 0.72),  
(‘1boy’, 0.54),  
(‘glasses’, 0.44),  
(‘short hair’, 0.41) ...

## 未知語率 (JUMAN++ サブワード化なし)

		ギャグ	少女	少年	青年	萌え系
拡張前	総単語数	270	289	274	276	276
	未知語率	0.244	0.263	0.230	0.236	0.232
拡張後	総単語数	3030	3209	3089	3154	3154
	未知語率	0.681	0.688	0.681	0.682	0.682

## 未知語率 (JUMAN++ サブワード化あり)

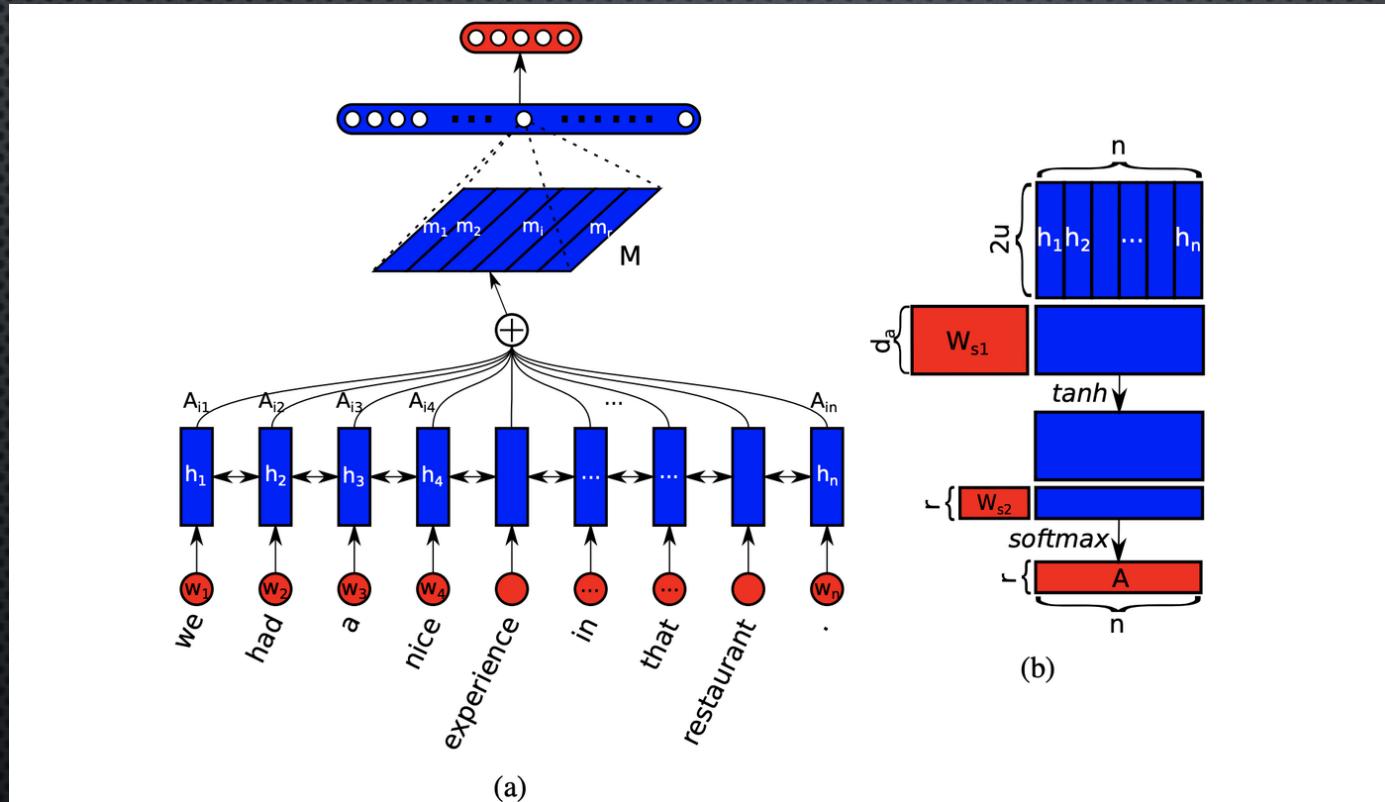
		ギャグ	少女	少年	青年	萌え系
拡張前	総単語数	311	331	315	316	316
	未知語率	0.026	0.021	0.019	0.022	0.025
拡張後	総単語数	2633	2728	2664	2705	2708
	未知語率	0.090	0.089	0.088	0.089	0.089

# VGG

畳み込み層と全結合層を連結し、  
シンプルに層を増やしたネットワーク構造  
構成する層の数に応じて, VGG-11 や VGG-16 など  
と呼ばれることが多い

Karen Simonyan and Andrew Zisserman",  
"Very Deep Convolutional Networks for Large-Scale  
Image Recognition", In "International Conference on  
Learning Representations", 2015

# Self-Attention



# Manga109



- 市販の 109 冊のコミックによって構成
- 画像周りのアノテーションは充実
- 作者によるセリフの感情アノテーションなし

- Y.Matsui, K.Ito, Y.Aramaki, A.Fujimoto, T.Ogawa, T.Yamasaki, K.Aizawa, Sketch-based Manga Retrieval using Manga109 Dataset, *Multimedia Tools and Applications*, Springer, 2017
- T.Ogawa, A.Otsubo, R.Narita, Y.Matsui, T.Yamasaki, K.Aizawa, Object Detection for Comics using Manga109 Annotations, arXiv:1803.08670



# 4コマ漫画におけるキャラクタのセリフの感情推定



# 評価指標

- ・F 値 ( $0 \leq F \leq 1$ )

再現率 (recall) , 適合率 (precision) の調和平均

$$\text{recall} = \frac{\text{TP}}{\text{TP} + \text{FN}}$$

$$\text{precision} = \frac{\text{TP}}{\text{TP} + \text{FP}}$$

$$F = \frac{2 \times \text{recall} \times \text{precision}}{\text{recall} + \text{precision}}$$

		予測	
		正例	負例
真値	正例	TP	FN
	負例	FP	TN

# コサイン類似度 とは

ベクトル同士の成す角度の近さを表現

- 1 に近ければ類似
- 0 に近ければ似ていない

$$\cos(\vec{q}, \vec{d}) = \frac{\vec{q} \cdot \vec{d}}{|\vec{q}| |\vec{d}|} = \frac{\vec{q}}{|\vec{q}|} \cdot \frac{\vec{d}}{|\vec{d}|} = \frac{\sum_{i=1}^{|V|} q_i d_i}{\sqrt{\sum_{i=1}^{|V|} q_i^2} \cdot \sqrt{\sum_{i=1}^{|V|} d_i^2}}$$

# Optuna とは

ハイパーパラメータの最適化を自動化するための  
ソフトウェアフレームワーク

ハイパーパラメータの値を自動的に調整  
優れた性能を発揮する値を決定

# 実験

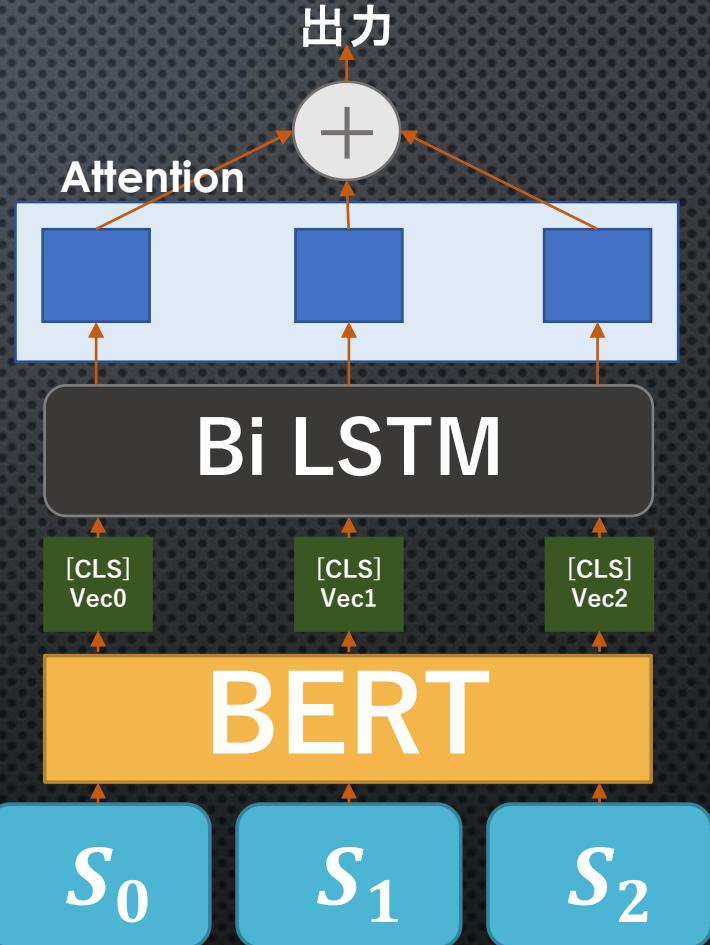
## セリフデータの前処理

1. JUMAN++ を用いて分かち書きにする
2. BERT 用にサブワード化し, 単語 ID 列に変換する  
(文頭に '[CLS]' , 文末に '[SEP]' を挿入)

## 実験 2：過去のセリフを考慮した感情推定

- => 連続したセリフ  $n$  文を順に入力 ( $2 \leq n \leq 5$ )
- => 末尾入力に対応する 1 つの感情ラベルを出力

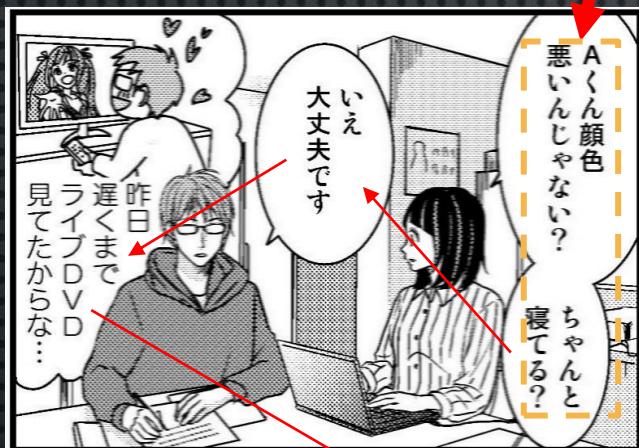
- ・識別器は Bi-LSTM ベースの Self-Attention
- ・BERT は最終層のパラメータのみをチューニング



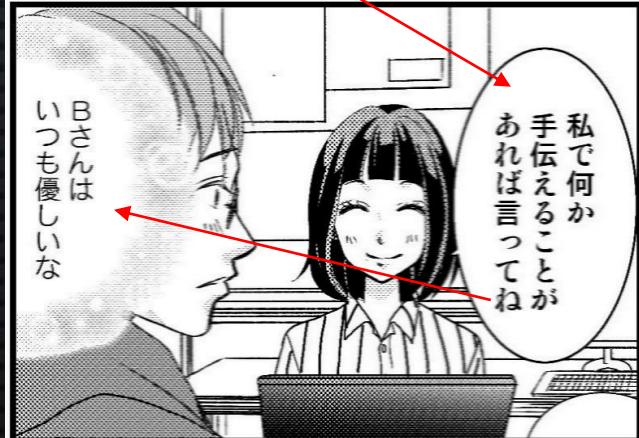
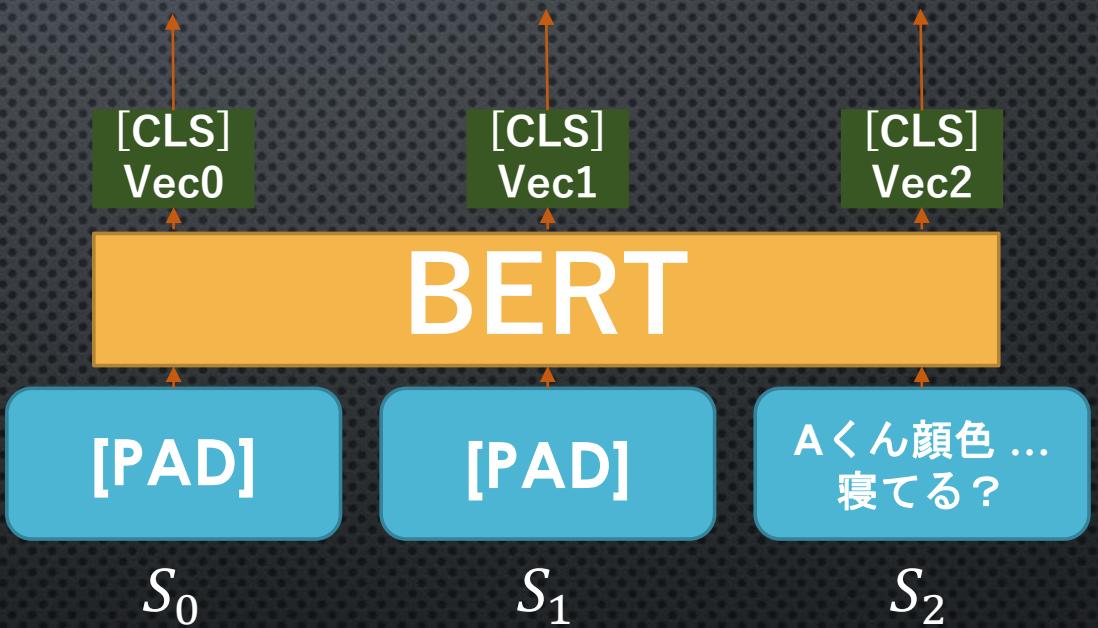
例：連続する 3 文のセリフの 単語 ID 列

# 実験 2

[PAD]  
[PAD]



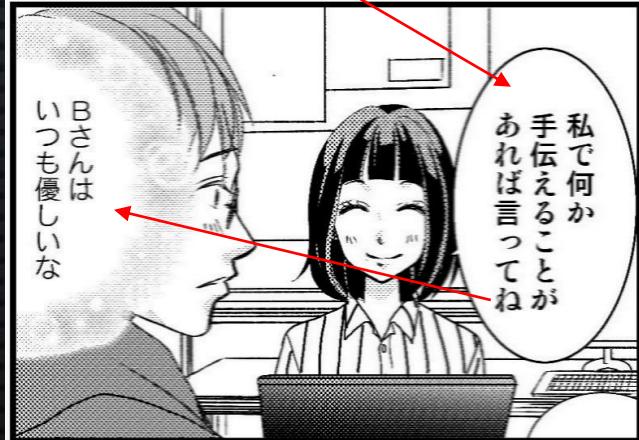
$n = 3$  の場合



※ 各 4 コマの初めに  $n - 1$  個の [PAD]

# 実験 2

[PAD]  
[PAD]



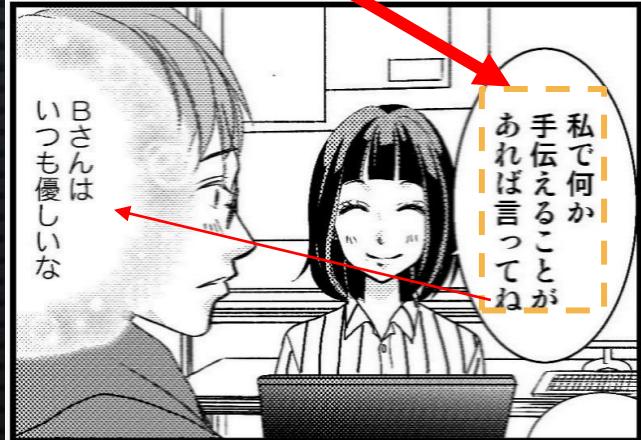
$n = 3$  の場合



# 実験 2

[PAD]

[PAD]



$n = 3$  の場合

[CLS]  
Vec0

[CLS]  
Vec1

[CLS]  
Vec2

BERT

いえ  
大丈夫です

昨日遅くまで  
…からな…

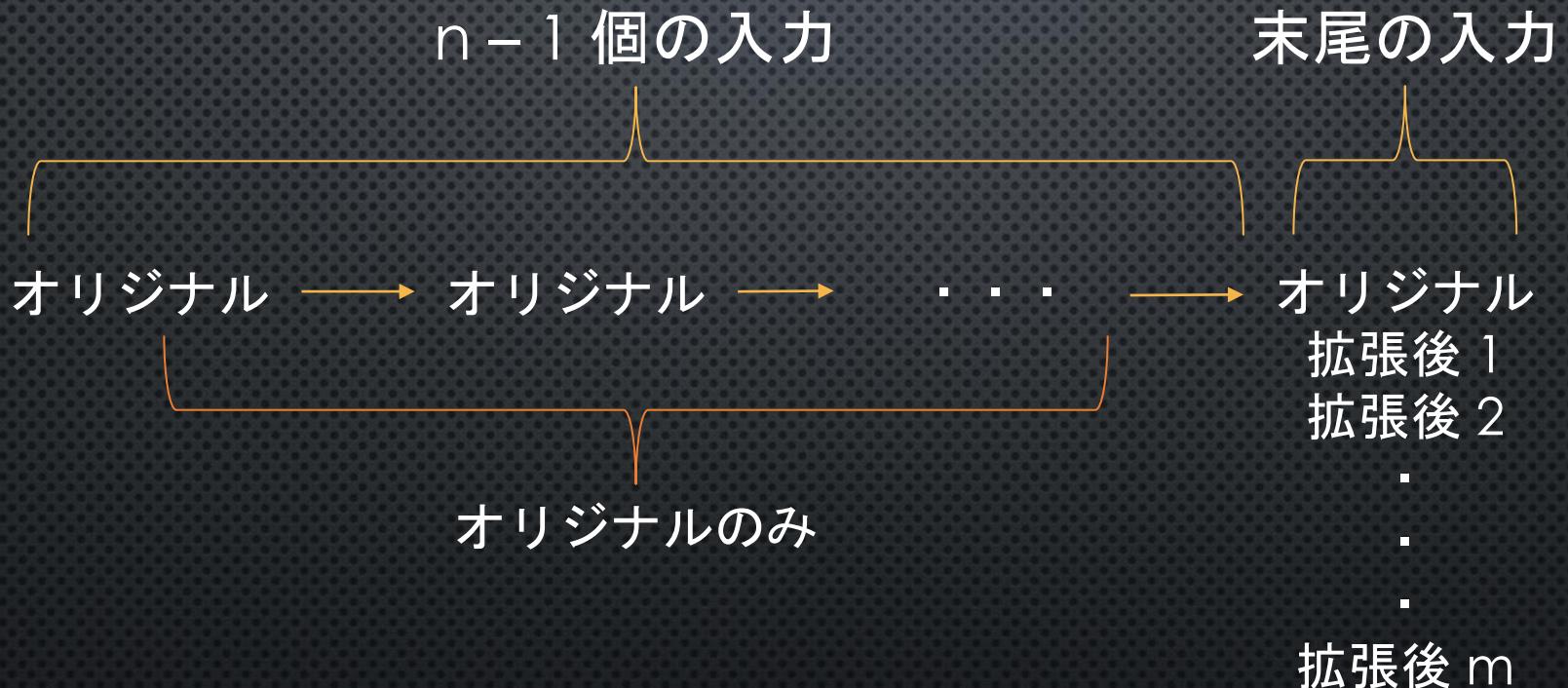
私で何か…  
言ってね

$S_0$

$S_1$

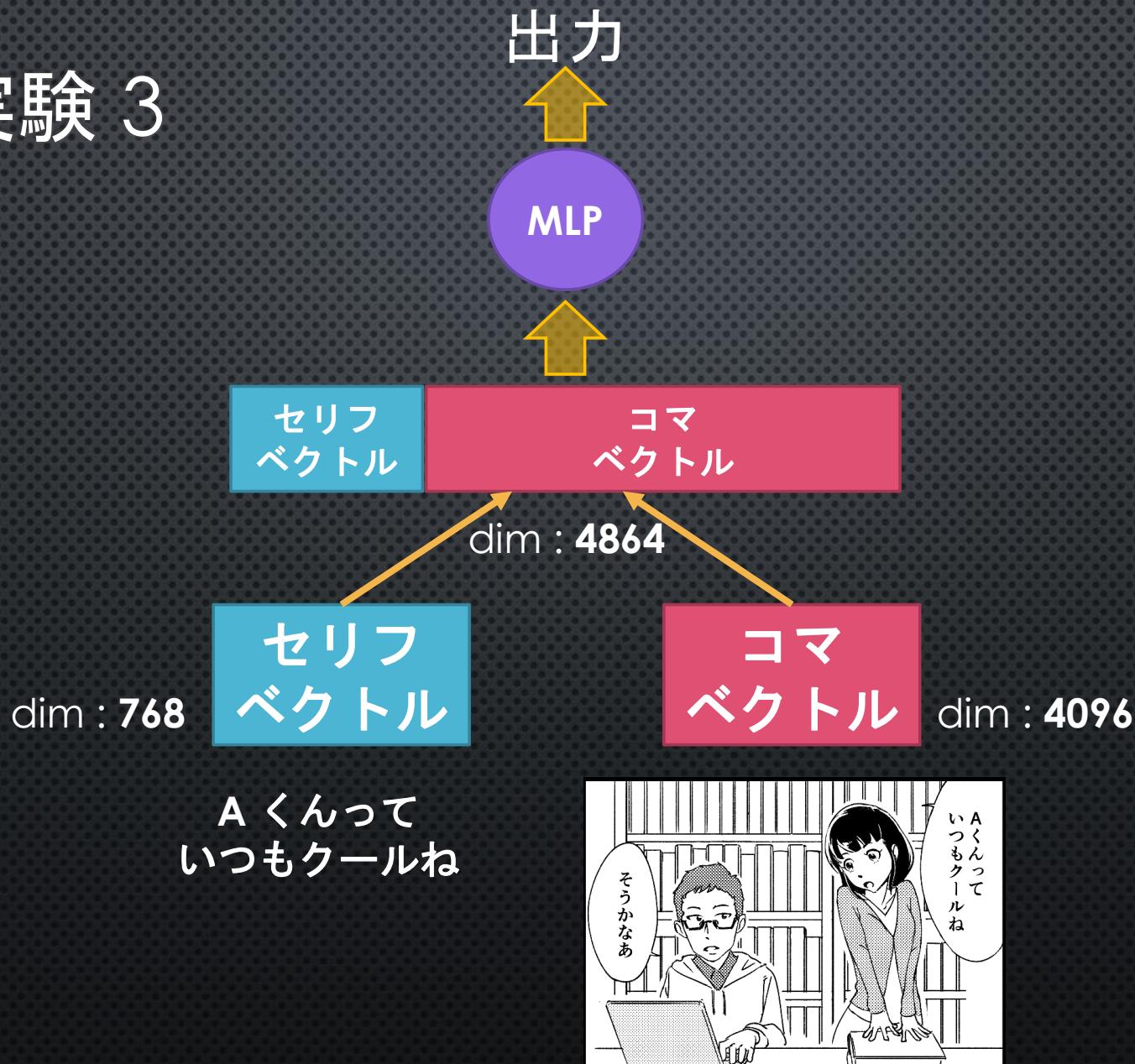
$S_2$

# 実験 2 の学習時における入力列の制限



実験 1 とのデータ数は一致

# 実験 3



# パラメータ（識別器）

## [実験 2] Self-Attention

Bi LSTM 入力次元	768
Bi LSTM 隠れ層次元	128
Self-Attention 入力次元	256
Self-Attention ヘッド数	3

# 実験 2

n	5 タッチ総合		
	Acc	P-Recall	P-F1 値
2	0.666	<b>0.427</b>	<b>0.429</b>
3	<b>0.635</b>	<b>0.552</b>	0.471
4	0.647	0.563	0.484
5	0.656	<b>0.656</b>	0.529
ベースライン	<b>0.705</b>	<b>0.000</b>	<b>0.000</b>

実験 1

BERT last layer

**0.733**

**0.521**

**0.535**

# 実験 2

n	5 タッチ総合		
	Acc	P-Recall	P-F1 値
2	0.666	<b>0.427</b>	<b>0.429</b>
3	<b>0.635</b>	0.552	0.471
4	0.647	0.563	0.484
5	0.656	<b>0.656</b>	0.529
ベースライン	<b>0.705</b>	<b>0.000</b>	<b>0.000</b>

↑ 増加傾向

実験 1

BERT last layer

**0.733**

0.521

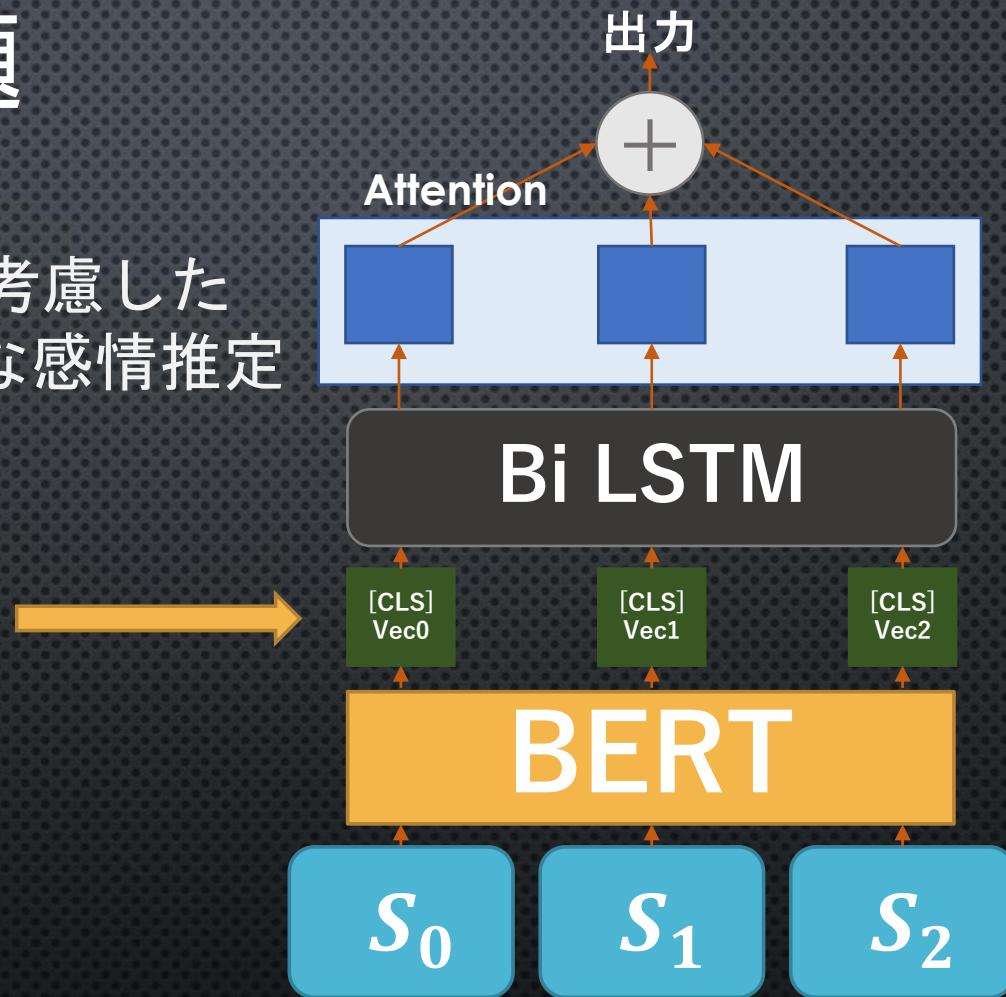
**0.535**

## 実験 学習時間 (GPU : USAGI)

1. 10 sec / (1 epoch • 1 タッチ)
2. 25 sec / (1 epoch • 1 タッチ) n=4
3. 15 sec / (1 epoch • 1 タッチ)

# 今後の課題

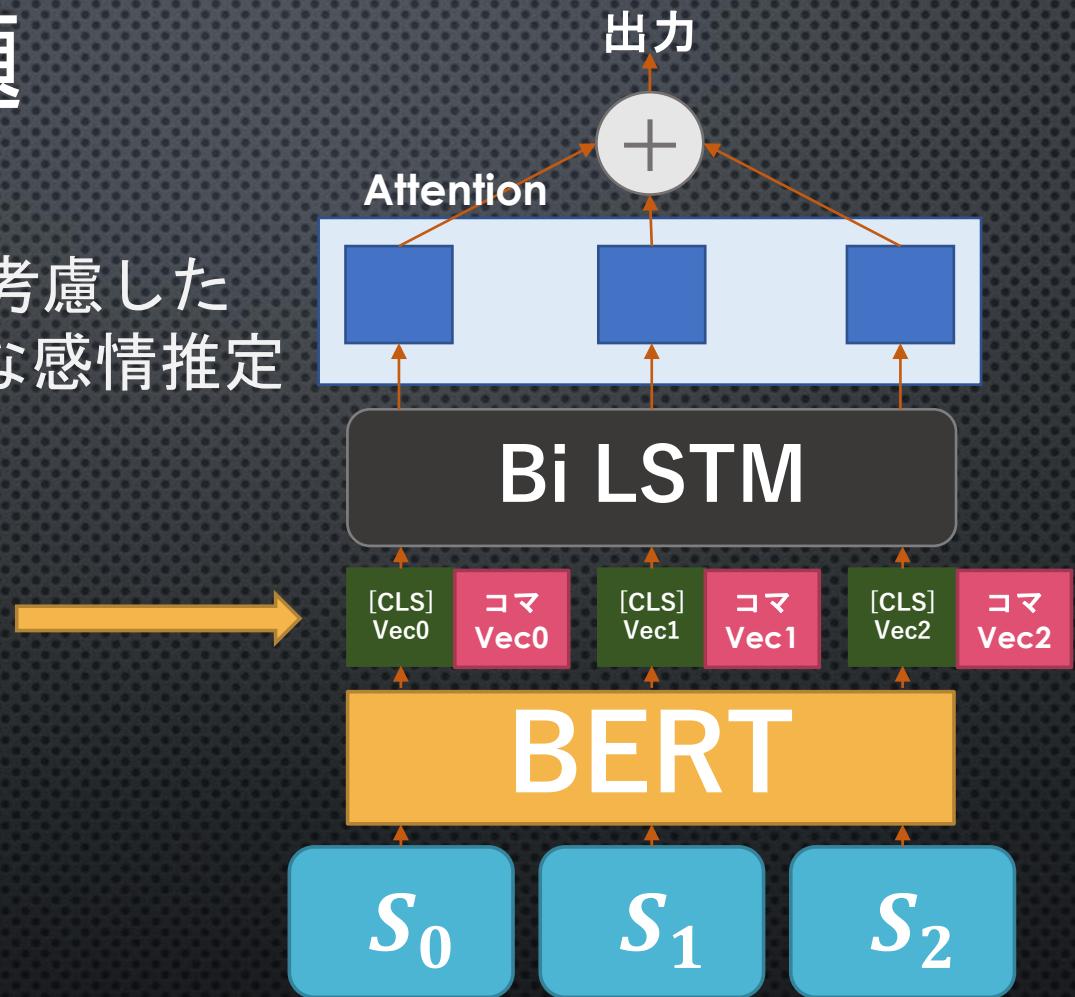
- ・過去のセリフを考慮したマルチモーダルな感情推定



例：連続する 3 文のセリフの 単語 ID 列

# 今後の課題

- ・過去のセリフを考慮した  
マルチモーダルな感情推定



例：連続する 3 文のセリフの 単語 ID 列

# 4 コマ漫画ストーリーデータセットの拡張

## 自作ツールの提案

エディター用メモ欄

エディター ID  
opu1234

A (男性)  
太郎

B (女性)  
花子

① Bさん\*

回答を入力

② 感情ラベル\*

選択

③ Aくん\*

回答を入力

④ 感情ラベル\*

選択

⑤ Bさん\*

回答を入力

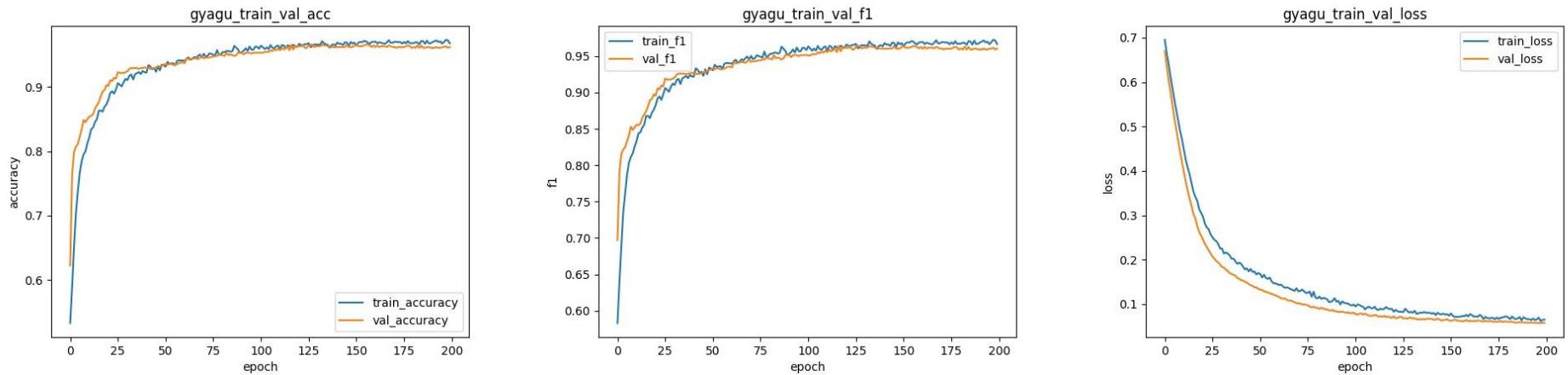
⑥ 感情ラベル\*

選択

1 / 20

- ・エディターに固有のID付与
- ・エディターには登場人物に名前を付けてもらう
- ・セリフと感情ラベルをセットで考えてもらう

## 実験 2 学習推移



ギャグ漫画タッチ

未知語率 (JUMAN++ サブワード化あり)

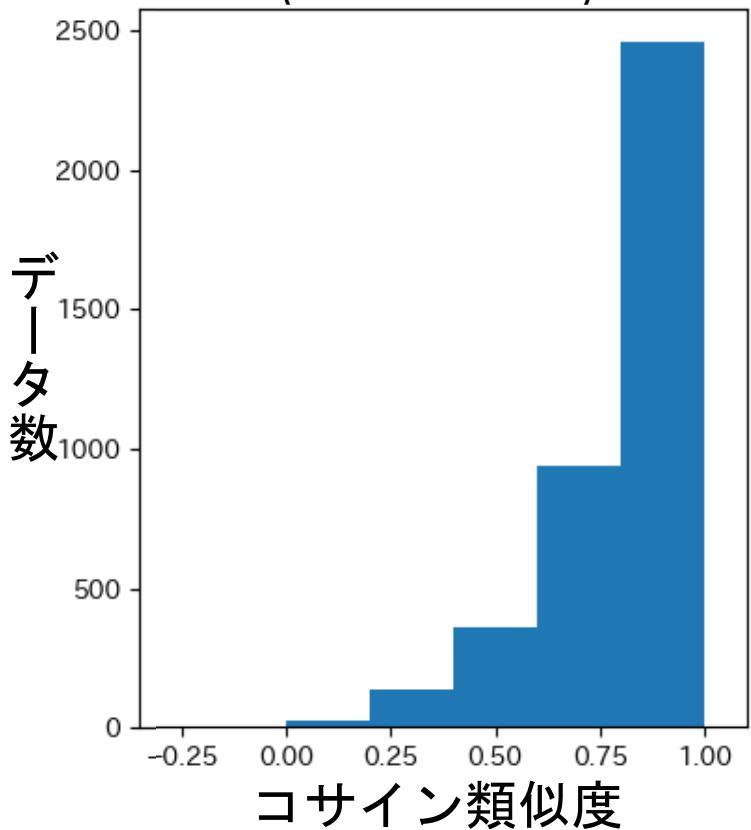
'A', 'DVD',

'B', 'GPU',

! ! !  
, , ■

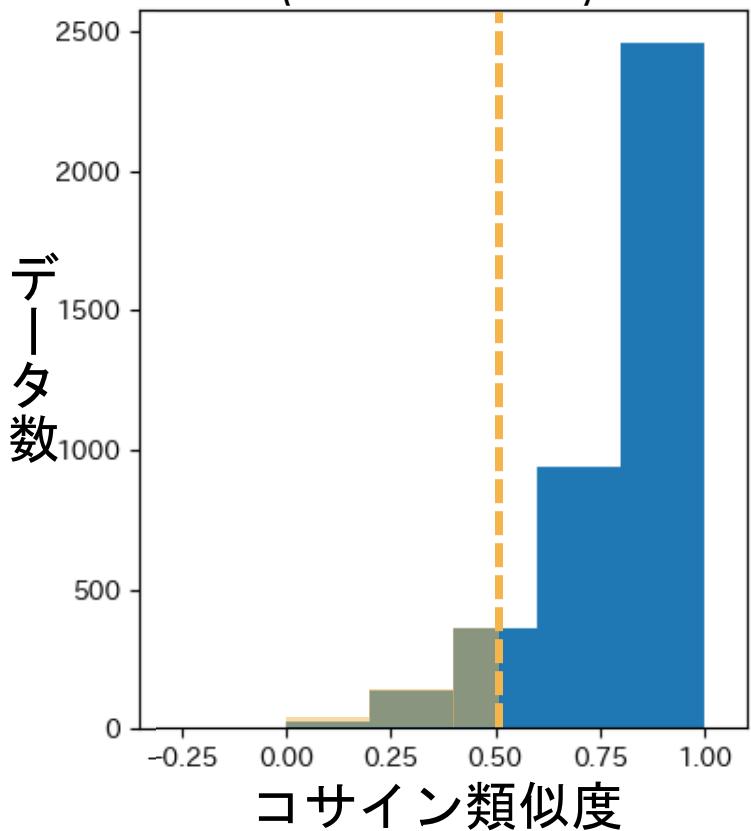
# Data Augmentation の妥当性

オリジナル↔拡張 コサイン類似度  
(ギャグタッチ)



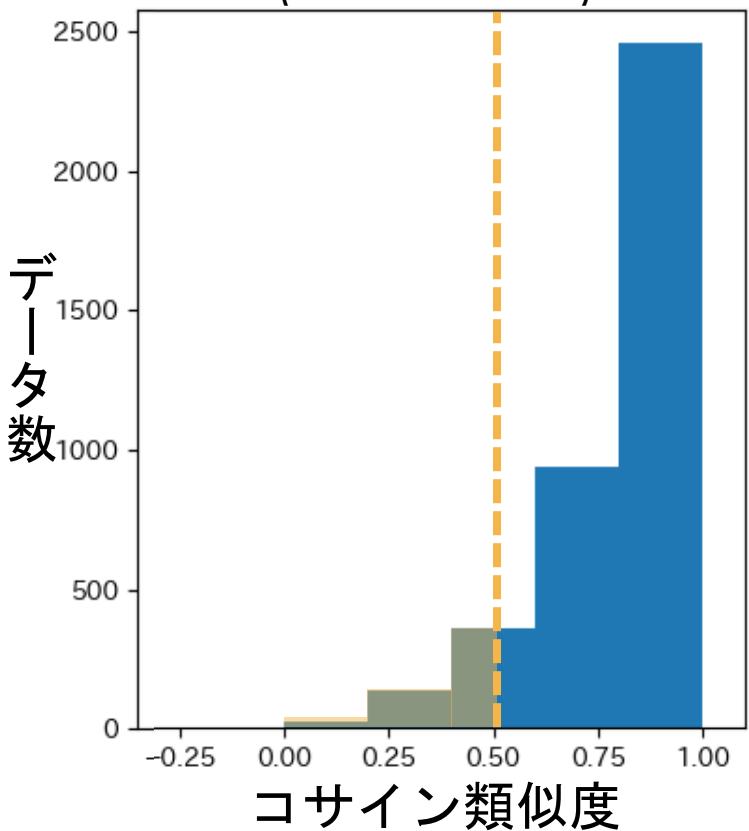
# Data Augmentation の妥当性

オリジナル↔拡張 コサイン類似度  
(ギャグタッチ)



# Data Augmentation の妥当性

オリジナル↔拡張 コサイン類似度  
(ギャグタッチ)



オリジナル：  
内緒です 気をつけてください

↓

拡張後：  
奥義です 魂胆をつけてください

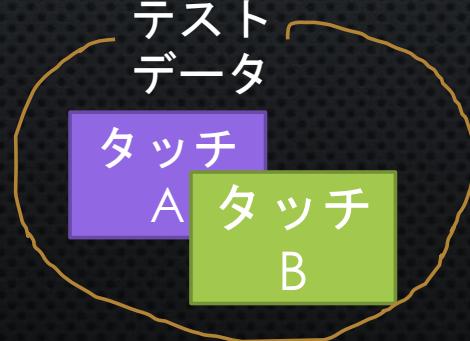
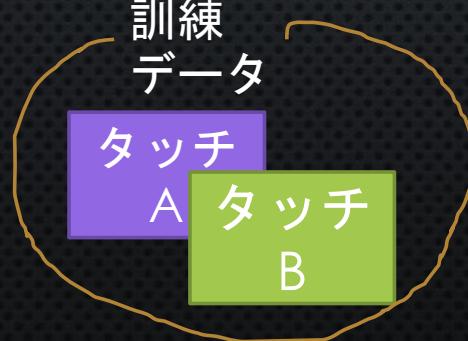
# 今後の課題

- ・マルチモーダル化による恩恵

非マルチモーダル (セリフのみ)



マルチモーダル (セリフ + コマ)



# 今後の課題

- ・データセット拡張手法の提案



セリフの情報を抜いたコマデータに対して

- ・登場人物の名前の一貫性
- ・ある程度の物語の一貫性

上記を踏まえた上でセリフを作成し、  
対応する感情ラベルを付けてもらう



「作者によるセリフの感情アノテーション」  
という元データセットの特徴を担保

# 今後の課題

- hottoSNS-BERT

hottolink 社が公開している大規模日本語 SNS  
コーパスによって事前学習させた BERT モデル

	コーパス	言語	分かち書き	単語数
hottoSNS-BERT	ツイート	ja	SentencePiece	32,000
京大 BERT	Wikipedia	ja	Juman++ / BPE	32,000

漫画に含まれるセリフは口語的であることから  
hottoSNS-BERT の方が本研究に適していると推測

# 今後の課題

- hottoSNS-BERT の特徴 : <mention> トークン

[通常ツイートデータ]

ゆめさんが、ファボしてくるあたり、世代だなって思いました( ˘˘ )笑

[リプライデータ

<mention> 90秒に250円かけるかどうかは、まあ個人の自由だしね()



セリフデータを加工することで、より会話の流れを加味した学習が可能になると推測できる

# 今後の課題

- ・ illustration2vec モデルの fine tuning
  - コマ画像データに対するメタ情報タグを付与
- ・ 過去のセリフを考慮したマルチモーダルな感情推定
  - 各データの結合方法やネットワーク構造の最適化

## ・データセット拡張手法の提案



萌えタッチ

萌えタッチ'

# ・コマ画像データに対するメタ情報

恐怖



喜楽

ニュートラル

感情ラベル

ニュートラル	喜楽	驚愕	恐怖	悲哀	憤怒	嫌悪	UNK
1	1	0	1	0	0	0	0

登場人物

Aくん	Bさん	その他
1	1	0

# BERTへの入力形式 改善案

## 現状

Sequential な  
識別器



## 改善案

MLP や SVM などの  
識別器



# 卒論発表までの課題

## 1. hottoSNS-BERT vs 京大 BERT

1-1 セリフ 1 文の感情推定

1-2 セリフ 1 文のマルチモーダルな感情推定

1-3 過去のセリフを考慮した感情推定

1-4 過去のセリフを考慮したマルチモーダルな感情推定

## 2. 半教師あり学習を用いた場合の変化

## 3. コマベクトルのみを fine-tuning した場合の変化