

< 修 士 論 文 >

# 深層学習を使った雑談対話システム

滋 賀 大 学 大 学 院

デ ー タ サ イ エ ン ス 研 究 科

デ ー タ サ イ エ ン ス 専 攻

修了年度：2021 年度

学籍番号：6020105

氏 名：長田 帆貴

指導教員：市川 治

提出年月日：2022 年 1 月 12 日

## 内容

第1章 はじめに.....	1
1.1 研究の背景 .....	1
1.2 目的 .....	1
1.3 本論文の構成.....	2
第2章 提案手法 .....	3
2.1 提案手法の概要 .....	3
2.2 使用したデータ.....	3
2.2.1 対話破綻検出チャレンジ .....	3
2.2.2 テレビ字幕データ .....	6
第3章 対話破綻検出器によるデータ作成.....	8
3.1 LightGBM による対話破綻検出器 .....	8
3.1.1 特徴量及び教師信号の設計 .....	8
3.2 BERT による対話破綻検出器.....	11
3.2.1 回帰による方法 .....	11
3.2.2 ソフトラベルによる方法 .....	12
3.3 実験のまとめ・考察.....	13
第4章 対話モデルの作成 .....	14
4.1 字幕データからの対話抽出.....	14
4.2 Sequence To Sequence による対話モデルの作成 .....	14
4.3 対話モデルの対話例 .....	15
第5章 結論と今後の展望 .....	16
謝辞.....	17
参考文献 .....	18

# 第1章 はじめに

## 1.1 研究の背景

対話システムの研究においてタスク指向型と非タスク指向型の 2 種類の研究がされてきた。タスク指向型の対話システムは、ユーザ側に明確な達成目標があり、天気や道案内等の閉じたドメインで対話が行われることを仮定できる場合が多い。一方、非タスク指向型の雑談対話システムはユーザ側に明確な達成目標がなく対話自体が目的であり、対話システムは多様な話題を扱う必要がある。これに対して近年、非タスク指向型対話システムの研究において Twitter や Reddit のような SNS における人間同士のコミュニケーションからなる大量の会話データから対話システムにおける発話を自動獲得することが可能になっている [1]。特に、英語においては上記のような非タスク指向型対話システムの State-of-the-art に Facebook による Blender Bot がある。Blender Bot は英語掲示板 Reddit のデータを用いて大規模な Transformer encoder-decoder モデルを pre-train して、雑談対話において筆者が重要と考える 3 つのスキル(話者性・知識・共感)を運用できるように fine-tune したモデルであり、主観評価において人と同程度の性能を有していることが示されている [2]。日本語の雑談対話モデルにおいても、Blender Bot の日本語版相当モデルが対話システムライブコンペティション 3 の予選において 1 位を取り、大規模な対話データによって pretrain されたモデルの適用が進んでいる [3] [4]。

大規模コーパスによる対話システムの他のシステムに対する利点として、素朴な常識的推論を行うことができるとされていること [5]やデータ量に伴って性能が向上することが知られている。一方で、欠点としてはコーパスに存在しない新語に対応できないこと、コーパスが書き言葉である場合、雑談対話には不自然な発話が生成されること、口語調の対話ペアのテキストデータを適切な質と量を揃えることの困難さがあることが考えられる。非タスク指向型の対話モデルは雑談を通じてユーザの対話欲求を満たすことや親しみを持ってもらうことを目的とされることが多い。対話モデルが口語調の対話を行うことはこの目的を達成するために重要であると考えられる。

## 1.2 目的

既存の大規模コーパスを雑談対話モデルへの学習データとして用いる場合の問題点として、その言語資源が文語でということがある。ユーザの対話欲求を満たす雑談対話システムにはユーザが対話システムに親しみを持ち対話を続けたい言葉遣いが求められる。しかし、十分な質と量のある口語調の対話データを作成することは一般に高コストである。そこで本研究ではテレビ字幕データから対話ペアを抽出し、対話モデルの学習データとすることで口語調の発話が可能な対話モデルの作成を目指す。

## 1.3 本論文の構成

第2章では、はじめに本論文の提案手法を概観し、提案手法に必要なデータとその前処理について説明する。第3章では3種類の対話破綻検出器を提案し、各モデルの詳細の説明と比較を行う。第4章では、第3章で作成した対話破綻検出器を用いてテレビ字幕から対話ペアを学習データとして対話モデルの作成を行う。また、作成した対話モデルのユーザとの対話例を通じて学習に用いた対話データのコーパスとしての品質評価を行う。第5章では、今後の課題として本論文で未検証の事項や今後の展望を述べる。

## 第2章 提案手法

### 2.1 提案手法の概要

本論文ではテレビ字幕データから対話ペアを抽出し、対話モデルの学習データとすることで口語調の発話が可能な対話モデルの作成する方法を提案する。本論文は以下の3段階からなる。Dialogue Breakdown Detection Challenge (DBDC)については2.2.1節で説明する。

1. DBDCのデータを用いた対話破綻検出器の学習
2. テレビ字幕データから対話破綻検出器による対話ペアの抽出
3. 対話モデルの学習

字幕データはテレビの字幕からペアを作成したものである。口語調の対話を含むが、文脈に依存し対話ペアとしては破綻した応答をしているペアやナレーションや歌などでそもそも対話をしていないペアを含む。DBDCで提供されたデータを用いて対話破綻検出器を作成し、字幕データから対話として成立しているもののみを抽出する。また、抽出した対話ペアから対話モデルを作成し、対話ペアの有用性を検討する。以上の提案手法のアウトラインを図2-1に示す。

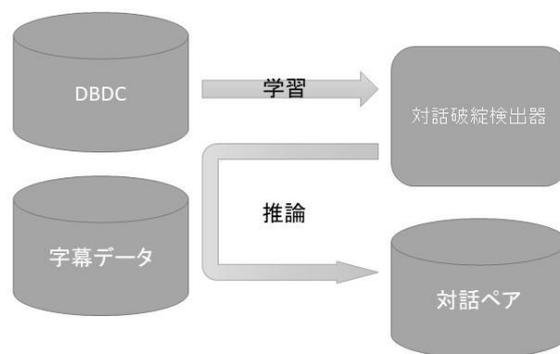


図 2-1

### 2.2 使用したデータ

#### 2.2.1 対話破綻検出チャレンジ

DBDCとは、非タスク指向型対話システムとユーザの対話ログに対して、対話の破綻を検出することを目的としたモデルを開発し、その精度を競う学術的なコンペティションである [6]。対話の

破綻を自動検出できれば、対話システム自身が発話前に対話破綻を認識し、リカバリすることで、より自然な応答を実現することができるとされている [7] [8]。

どのような場合に対話破綻をとするかは主観的な要素が強い。そこで、DBDC で提供されているデータは、一つの対話ペアに対して複数人のアノテータがいずれかのラベルを付与している。DBDC で提供されている対話ログは「あきらかにおかしいと思う発話, 破綻」「破綻とは言い切れないが, 違和感を感じる発話」「破綻でない」の 3 つクラスがある。過去の DBDC で提供された学習データは表 2-1 対話破綻検出器の学習に用いたデータの通りである。また、第 3 章で説明する対話破綻検出器の評価のために DBDC で提供されているデータを表 2-2 学習データとテストデータの分割に示す通り分割した。

表 2-1 対話破綻検出器の学習に用いたデータ

	対話システム の名前	対話システムの説明	対話ペアの数	アノテータ の人数
Chat dialogue corpus	DCM	ドコモが一般公開している雑談対話 API [9]を用いた雑談対話システム	1146	※
DBDC1	DCM	ドコモが一般公開している雑談対話 API [9]を用いた雑談対話システム	80	30
DBDC2	DCM	ドコモが一般公開している雑談対話 API [9]を用いた雑談対話システム	300	30
DBDC2	DIT	デンソーアイティラボラトリ提供の雑談対話システム [10]	300	30
DBDC2	IRS	オーガナイザが準備した, IR-STATUS [11]に準拠した用例ベースの雑談対話システム	300	30
DBDC3	DCM	ドコモが一般公開している雑談対話 API [9]を用いた雑談対話システム	50	30
DBDC3	DIT	デンソーアイティラボラトリ提供の雑談対話システム [10]	50	30
DBDC3	IRS	オーガナイザが準備した, IR-STATUS に準拠した用例ベースの雑談対話システム	50	30
Dialogue system live competition	Dialogue system live competition [12]において提出された 5 つの対話システム(ZNK,MRK,MMK,IRS,TRF)		146	30

※100 対話(init100)は 24 人のアノテータによって、残りの対話(rest1046)は 2、3 人のアノテータによってアノテートされたものである。

表 2-2 学習データとテストデータの分割

学習データ	テストデータ
DBDC2 の学習データ・テストデータ DBDC3 の学習データ DBDC4 の学習データ・テストデータ DBDC5 の学習データ・テストデータ	DBDC3 のテストデータ

## 2.2.2 テレビ字幕データ

平成9年の放送法改正により放送事業者は平成 19 年度までに新たに放送する字幕付与可能な放送番組のすべてに字幕を付与することを目標とする放送努力義務が規定された。平成 19 年度実績では、字幕付与可能な放送時間に占める字幕放送時間の割合は、NHK(総合)で100%、民放キー5局平均で89.0%となった。

本研究で使用したテレビ字幕はガラポン株式会社の地上波デジタル放送のワンセグ帯を録画できる録画機である「ガラポンTV」を用いて取得した。本研究で用いた字幕データは2019年1月から2021年9月にワンセグ帯で放送された番組に付与された約1700万行のデータであり、字幕以外に以下の情報が含まれる。

- 放送形態
- 放送局
- 放送開始日時の unix time stamp
- 番組名
- 番組のジャンル

テレビ番組のジャンルによって対話が成立している事前確率が異なることが想定されるため、対話が成立している事前確率が高いと想定されるジャンルの番組に限定して、対話破綻検出器による対話の抽出をした。ジャンルはガラポンAPIによって付与されたものを用いた、抽出したジャンルは以下の通り。

- 討論・会談
- 国内ドラマ
- 海外ドラマ
- その他ドラマ
- トークバラエティ
- 旅バラエティ
- 洋画
- 邦画
- その他映画
- 特撮
- インタビュー・討論
- 現代劇・新劇

上記のジャンルに該当する番組のうち、対話の重複を防ぐために再放送や民放系列放送局ごとに一部放送局に限定して字幕データ対話抽出の対象とした。字幕は場面に対して付与されているため、発言の途中であっても区切られて複数の字幕とされている。発言の継続を示す記号が付与されている場合は字幕を結合し一つの発話として扱った。また、字幕中に歌を示す記号が含まれる場合や非 UTF-8 文字が含まれる字幕はその字幕を含む対話ペアを対話抽出の対象外とした。テキストの正規化は neologdn を用い、テキストのクリーニングとして登場人物の発言であることを示す記号やナレーションであることを示す記号をテキストから消去した。対話破綻検出器では対話ペアに対して対話が破綻しているかを検出するため、字幕データを時系列順でペアにして対話破綻検出器の入力とする。以上の処理を実施した字幕データのサンプルを下記に示す。

#### 字幕データのサンプル

歌は好きだったけど人見知りではあったんですって？
そうなんです。ものすごい人見知りで常に母親の後ろに隠れてました。あら。
もうろくに挨拶もできない…。はい。
よくその 13 歳でオーディション受かったからってそこで人の前で歌えるようになりましたね。
そうなんですよ。それも叔母にだまされて連れて行かれたんですね。
でまあ受かってしまったので。
その時ちょうど住んでたアパートがトイレもお風呂もなかったんです。
ってというような生活から…うん…。
少し私が歌手になる事で親孝行ができるのかなってその時思ったんですね。偉いね 13 歳ぐらいで
ずっとそういう生活でしたから。

本論文では対話破綻検出器の字幕データにおける性能評価のために評価用字幕データとして、筆者が字幕データより 1394 対話ペアのアノテーションを行った。評価用の字幕は対話破綻抽出に用いるジャンルの複数の番組を対象とした。評価用字幕データにおける各クラスのサンプルサイズは表 2-3 評価用字幕データの通りである。第 3 章では作成した評価用字幕データのうち「破綻とは言い切れないが、違和感を感じる発話」と「破綻でない」の 2 クラスを同一視して破綻と非破綻の 2 クラス分類問題としてモデル評価を行う。

表 2-3 評価用字幕データ

破綻ラベル	対話ペアの数
あきらかにおかしいと思う発話, 破綻	1394
破綻とは言い切れないが, 違和感を感じる発話	112
破綻でない	91

## 第3章 対話破綻検出器によるデータ作成

### 3.1 LightGBM による対話破綻検出器

#### 3.1.1 特徴量及び教師信号の設計

LightGBM による対話破綻検出器は、Universal Sentence Encode (USE) による文章の潜在表現と以下に示す特徴量をモデルの入力とした。

対話システムごとの破綻のパターンに注目し、特徴量を設計し DBDC2 において SoTA であった文献 [13] を参考に特徴量を作成した。作成した特徴量は表 3-1 LGBM を用いた対話破綻検出器の特徴量の通りである。

USE は 16 言語に対応した多言語モデルで、可変長の文を 512 次元のベクトルに埋め込む [14]。USE は TensorFlow Hub から 3 つの version の学習済みモデルが提供されており、本論文では Version 3 を用いた。対話ペアのテキストから USE によって得た 2 つの 512 次元のベクトルと下記の 9 つの特徴量を合わせた 1033 次元をモデルの入力とした。

表 3-1 LGBM を用いた対話破綻検出器の特徴量

特徴量	特徴量の説明
num_char	対象発話の文字数
num_word	対象発話の単語数
per_match_sentence	前の発話との単語一致率
char_ngram_sentence	前の発話との 4gram の一致率
wmd_sentence	前の発話との Word Mover's Distance を計算
wmd_noun	前の発話との Word Mover's Distance を名詞のみで計算
wmd_noun_verb	前の発話との Word Mover's Distance を名詞と動詞で計算
cos_sim	前の発話との対象発話のコサイン類似度
pos_cos_sim	前の発話と対象発話の品詞 id のコサイン類似度

学習データとして DBDC~DBDC4 までの学習データ・評価データと DBDC5 の学習データを用い、モデルの評価に DBDC5 の評価データを用いた。

DBDC のデータはそれぞれの対話に対して複数人のアノテータが「あきらかにおかしいと思う発話、破綻」・「破綻とは言い切れないが、違和感を感じる発話」・「破綻ではない」のいずれかにアノテーションしている。アノテーションに対して表 3-2 破綻ラベルとスコアの通りスコアをつけ、アノテータ

の人数で割り 0 から1の数値とし教師信号とした。

表 3-2 破綻ラベルとスコア

破綻ラベル	スコア
あきらかにおかしいと思う発話, 破綻	1
破綻とは言い切れないが, 違和感を感じる発話	0.5
破綻ではない	0

モデル評価データとして、DBDC3 の評価データと第 2 章で説明した字幕データを用いた。本論文では対話破綻検出器をテレビ字幕から対話ペアを抽出するために使用する。そのため、モデル評価時はモデルの出力を閾値以下なら非破綻、閾値より大きいなら破綻として扱い 2 クラス分類問題として評価した。

文献 [15]によると本モデルの学習データに含まれる Chat dialogue corpus の init 100 において 24 人のアノテータ間の一致の程度を測るため算出した Fleiss の  $\kappa$  は 0.276 であった。「破綻とは言い切れないが, 違和感を感じる発話」を「あきらかにおかしいと思う発話, 破綻」に含めると、Fleiss の  $\kappa$  は 0.396 とやや一致の具合が高まる。一方、「破綻とは言い切れないが, 違和感を感じる発話」を「破綻ではない」に含めると Fleiss の  $\kappa$  は 0.277 にしか改善されない。このことから「破綻とは言い切れないが, 違和感を感じる発話」は「破綻ではない」よりも「あきらかにおかしいと思う発話, 破綻」に近いクラスであると考えられることができる。よって、モデル評価時の正解データとして使用する DBDC3 の評価データは「破綻とは言い切れないが, 違和感を感じる発話」を「あきらかにおかしいと思う発話, 破綻」に含めて 2 クラスの正解データとした。

1 に近いほど破綻の度合いが高いため、閾値を下げすぎると非破綻と推論される対話が減り、閾値を上げすぎると実際に非破綻の対話を非破綻と推論する確率(precision)が下がることが考えられる。また対話破綻検出器を字幕データから対話ペアを抽出するために使用するため、accuracy や f1-score よりも、対話破綻検出器が非破綻とした対話ペアのうち実際に非破綻のである確率である precision を重視してモデル評価を行う。学習に用いたハイパーパラメータは表 3-3 に示した。

表 3-3 実験条件

ハイパーパラメータ	
boosting_type	gbdt
metric	mse
max_depth	8
min_child_weight	32
colsample_bytree	0.8

実験結果を表 3-4と表 3-5に示す。DBDC3による評価では閾値が0.1のとき precision が1.0となった。一方、字幕データによる評価では閾値0.1のとき、モデルの出力が0.1となった対話が無く precision が0となった。DBDC3評価データにおいては学習データと同一の対話モデルのログを含むことから、学習データに類する対話破綻の傾向を有していると考えられる。一方、字幕データにおける非破綻クラスの precision はDBDC3に比べ大幅に低下した。学習データに用いた対話システムとユーザの対話ログではなくテレビ字幕データを推論したことで異なる対話破綻の傾向があったと推測できる。

字幕データにおいてはDBDC3による評価に比べ高い accuracy が示されたが、第2章で示した通り評価用に作成した字幕データは8割以上が破綻としてアノテーションされている不均衡データである。そのため字幕データの accuracy の解釈には注意が必要である。

表 3-4 DBDC3 評価用データによる結果(提案法1 LGBM)

		<i>precision</i>	<i>recall</i>	<i>f1-score</i>	<i>accuracy</i>
threshold0.1	非破綻	1.000	0.010	0.020	0.611
	破綻	0.609	1.000	0.757	
threshold0.3	非破綻	0.882	0.076	0.140	0.633
	破綻	0.624	0.993	0.766	
threshold0.5	非破綻	0.550	0.512	0.530	0.643
	破綻	0.697	0.729	0.713	

表 3-5 字幕データによる結果(提案法 1 LGBM)

		<i>precision</i>	<i>recall</i>	<i>f1-score</i>	<i>accuracy</i>
threshold0.1	非破綻	0.000	0.000	0.000	0.873
	破綻	0.873	1.000	0.932	
threshold0.3	非破綻	0.308	0.020	0.037	0.870
	破綻	0.874	0.994	0.930	
threshold0.5	非破綻	0.130	0.296	0.181	0.659
	破綻	0.874	0.712	0.785	

## 3.2 BERT による対話破綻検出器

### 3.2.1 回帰による方法

回帰による方法では 0 から 1 の連続値を教師信号とした回帰モデルを作成し、モデルの推論時は出力値に閾値を設け、閾値以下を非破綻、閾値より大きな値を破綻の対話とした分類問題として評価した。

本モデルでは事前学習として BERT を使い、最終層に全結合層を追加したモデルを作成した。日本語の事前学習済みモデルとして Huggingface Transformers より提供されている東北大学の乾研究室が提供しているモデルを用いた [16]。回帰による方法では、LightGBM による対話破綻検出器と同様の方法で 0 から 1 の値を取る数値を作成し、教師信号とした。また、教師データと次元を揃えるため BERT の最終層に出力が 1 次元の全層結合層を設けた。

第 2 章で述べた通り DBDC で提供されているデータは各対話ペアに対して複数のアノテータが破綻ラベルを付与している。モデル評価の際に各対話のアノテーションを集計し破綻・非破綻の 2 クラスとして扱う。第 2 章で述べた通り、字幕データによるモデル評価は「破綻とは言い切れないが、違和感を感じる発話」と「破綻でない」の DBDC3 評価用データによる評価と同様に 2 クラスを同一視して破綻と非破綻の 2 クラス分類問題としてモデル評価を行う。

実験の結果を表 3-6 と表 3-6 表 3-7 に示す。

表 3-6 DBDC3 評価用データによる結果 (提案法 2 BERT 回帰)

		<i>precision</i>	<i>recall</i>	<i>f1-score</i>	<i>accuracy</i>
threshold0.1	非破綻	0.989	0.023	0.046	0.470
	破綻	0.463	1.000	0.633	
threshold0.3	非破綻	0.914	0.463	0.615	0.685
	破綻	0.598	0.948	0.733	
threshold0.5	非破綻	0.760	0.807	0.782	0.756
	破綻	0.752	0.697	0.723	

表 3-7 字幕データによる結果 (提案法 2 BERT 回帰)

		<i>precision</i>	<i>recall</i>	<i>f1-score</i>	<i>accuracy</i>
threshold0.1	非破綻	0.235	0.020	0.036	0.867
	破綻	0.874	0.991	0.929	
threshold0.3	非破綻	0.129	0.118	0.123	0.786
	破綻	0.873	0.884	0.878	
threshold0.5	非破綻	0.141	0.350	0.201	0.646
	破綻	0.879	0.689	0.772	

### 3.2.2 ソフトラベルによる方法

ソフトラベルによる方法ではそれぞれのクラスにアノテーションされた数をアノテータの数で割り、3次元のベクトルを教師信号とした。またモデル最終層の全層結合の出力を3次元にした。事前学習済みモデルとして回帰による方法と同様に東北大学の乾研究室が提供しているモデルを用いた [16]。推論時は他のモデルと同様に分類問題として破綻ラベルを扱いモデル評価をした。

実験結果を表 3-8 と表 3-9 に示す。回帰モデルに比べ教師信号の次元が増えている分、教師信号から得られる情報は増加しているが、非破綻クラスの *precision* について、BERT による回帰モデルの対話破綻検出器に比較して、DBDC3 評価用データ・字幕データいずれも低い値が示された。原因としては、対話破綻ラベルを付与するアノテータ間の一致率が低いことが考えられる。文献 [15]によると本モデルの学習データに含まれる Chat dialogue corpus の init 100 において 24 人のアノテータ間の一致の程度を測るため算出した Fleiss の  $\kappa$  は 0.276 であった。この値はランダムではないが、よく一致しているともいえない」と解釈される。アノテータ間での評価が一致していないため、本モデルはアノテータ間の評価の不一致というノイズが回帰モデルよりも多く教師信号に含まれてしまった可能性がある。

表 3-8 DBDC3 評価用データによる結果 (提案法 3 BERT ソフトラベル)

		<i>precision</i>	<i>recall</i>	<i>f1-score</i>	<i>accuracy</i>
threshold0.1	非破綻	0.460	0.293	0.358	0.483
	破綻	0.494	0.667	0.567	
threshold0.3	非破綻	0.339	0.155	0.212	0.589
	破綻	0.638	0.831	0.722	
threshold0.5	非破綻	0.139	0.041	0.063	0.796
	破綻	0.830	0.949	0.885	

表 3-9 字幕データによる結果 (提案法 3 BERT ソフトラベル)

		<i>precision</i>	<i>recall</i>	<i>f1-score</i>	<i>accuracy</i>
threshold0.1	非破綻	0.151	0.123	0.136	0.800
	破綻	0.876	0.899	0.887	
threshold0.3	非破綻	0.134	0.227	0.168	0.715
	破綻	0.875	0.786	0.828	
threshold0.5	非破綻	0.134	0.404	0.201	0.593
	破綻	0.877	0.621	0.727	

### 3.3 実験のまとめ・考察

本章では DBDC で提供されている対話データを用い、3 種類の対話破綻検出器を作成・評価を行った。実験の結果、BERT による回帰モデルの対話破綻検出器が評価用字幕データにおいて最も *precision* が高かった。

DBDC の学習データ・評価データは対話モデルと人の対話ログである。一方、字幕データは、放送された字幕を時系列でつなぎ対話ペアにしたものである。字幕データにおける非破綻クラスの *precision* は DBDC3 評価用データにおける非破綻クラスの *precision* よりも大幅に低下している傾向が見られた。このことから字幕データと DBDC では、破綻のパターンが異なることが示唆される。

また、BERT による対話破綻検出器では、LightGBM モデルに比べその低下が少なく、BERT のほうが LightGBM よりも高い対話破綻検出の汎化性能を得ていることが確認できた。

## 第4章 対話モデルの作成

### 4.1 字幕データからの対話抽出

実験結果を踏まえ、BERT による回帰モデルの対話破綻検出器を用いて字幕データから対話抽出を行った。抽出した対話ペアを対話モデルの学習データとして用いて対話ペアの対話コーパスとしての有用性を検証する。対話破綻検出器の出力が 0.3 以下かつ文の単語最大長が 10 以下の対話を対話モデルの学習データとした。以上の条件を満たす字幕を約 1700 万行の字幕データから約 146 万対話ペアを抽出した。

### 4.2 Sequence To Sequence による対話モデルの作成

字幕データの固有名詞は対話管理のためのスロットに格納し、[Food\_Other]や[Person]等のトークンに置換してから対話モデルに入力、対話モデルの出力のトークンを埋め戻し対話システムの発話とする (Slot Filling)。Slot Filling のための固有表現認識には spaCy から提供されている GiNZA を用いた。また、置換されたトークンを MeCab の辞書に登録することでトークンが単語分割されないように処理をした。

対話モデルは embedding 層と 2 層 LSTM、1 層の Attention 機構からなるモデルとして作成した。モデルの構造とパラメータは図 4-1 に示した通りである。

Layer (type)	Output Shape	Param #	Connected to
input_1 (InputLayer)	[(None, None)]	0	
input_2 (InputLayer)	[(None, None)]	0	
embedding (Embedding)	(None, None, 256)	17308416	input_1[0][0]
embedding_1 (Embedding)	(None, None, 256)	14615296	input_2[0][0]
lstm (LSTM)	[(None, None, 256),	525312	embedding[0][0]
lstm_1 (LSTM)	[(None, None, 256),	525312	embedding_1[0][0] lstm[0][1] lstm[0][2]
attention (Attention)	((None, None, 256),	0	lstm_1[0][0] lstm[0][0]
concatenate (Concatenate)	(None, None, 512)	0	lstm_1[0][0] attention[0][0]
dense (Dense)	(None, None, 57091)	29287683	concatenate[0][0]

=====  
Total params: 62,262,019  
Trainable params: 62,262,019  
Non-trainable params: 0  
=====

図 4-1 モデルの構造とパラメータ

### 4.3 対話モデルの対話例

作成したモデルの対話例を示す。[]で囲われている語句はモデル内でトークンとして扱われているものでそのまま出力してしまうと、意味が通らないので、実用に耐える対話モデルに改良するためには何らかの語彙でトークンを置換する必要がある。

非文や文脈と合わない発話を生成してしまうこともあったが、本研究の目的である口語調の発話が生成されていることが確認できた。

#### 対話例 1

user：こんにちは、よろしくお願ひいたします。

system：今回こそよろしくお願ひいたします。

#### 対話例 2

user：最近何か読みましたか？

system：「みんなやらしい」

#### 対話例 3

user：どちらから来ましたか

system：みんな今…[City]とかねえよやっぱりあの…

#### 対話例 4

user：昨日は雪が降っていましたよ

system：はいそうですね。

#### 対話例 5

user：朝は何を食べましたか。

system：突然ですね。

#### 対話例 6

user：何してるの。

system：えと楽しんでる？

#### 対話例 7

user：ワールドカップが始まりましたね

system：今回は[Province]かぁ。

## 第5章 結論と今後の展望

本論文では DBDC の対話データを用いて対話破綻検出器を作成し、テレビ字幕中に含まれる口語の対話ペアを抽出し、口語調の対話モデルの作成を目指した。作成した対話モデルが口語調の発話を生成できることを確認できた。しかし、対話モデルが一定割合で非文または文脈に合わない発話を生成することも確認した。

本研究では対話破綻検出器はテレビ字幕からの対話抽出のみに用いた。対話破綻検出器を対話システムの応答候補に対し対話破綻検出を適用し、応答性能の向上させる応用も試みられている [8]。

対話モデルの学習に使用したデータをテレビ字幕から抽出するために使用した対話破綻検出器の precision が低く学習データに対話ペアとして不適切なデータが混在してしまったためだと考えられる。つまり、対話モデルの学習に用いるデータのうち対話の破綻したペアの割合を下げる事ができれば、対話モデルが自然な応答を生成する割合が高まると考えられる。本論文では対話破綻検出器の作成において、DBDC で提供されていた学習データのみで学習をした。DBDC の学習データ・評価データは対話モデルと人の対話ログである。一方、字幕データは、放送された字幕を時系列でつなぎ対話ペアにしたものである。本論文の DBDC 評価データ・テレビ字幕データを用いた対話破綻検出器の実験では、テレビ字幕データの精度は DBDC 評価データよりも大幅に低下することが確認できた。このことから、字幕データと DBDC では破綻のパターンが異なることが想定される。よって、提案法 2 と提案法 3 においては、テレビ字幕データの一部を学習データとしてアノテーションし、DBDC のデータ学習した対話破綻検出器のファインチューニングを実施することが有効であると考えられる。

## 謝辞

本研究を進めるにあたり多くの方々からのご協力・ご指導を頂きました。指導教員である滋賀大学の市川治教授に深く感謝申し上げます。市川治教授には本研究に必要な字幕データの提供を頂いただけでなく、分野に明るくない私に懇切丁寧にご指導を頂きました。

市川ゼミの徳永一輝さん、水口綾乃さんに深く感謝いたします。モデルの実装に関して有益な情報を提供してくださいました。今井貴史助教に深く感謝いたします。対話破綻検出器の学習に関して有益なアドバイスを頂きました。

修士課程の2年間で、ともに切磋琢磨し有益なものにしてくださいましたデータサイエンス研究科の同期の皆様に深く感謝いたします。

最後に、本研究ならびに学業全般にわたって経済的・心身的に支援してくださった両親に深く感謝し、お礼を申し上げます。

## 参考文献

- [1] 稲葉通将, 神園彩香, 高橋健一, “Twitter を用いた非タスク指向型対話システムのための発話候補文獲得,” 人工知能学会論文誌 29.1 (2014): 21-31., 2014.
- [2] Weston, Stephen RollerEmily DinanNaman GoyalDa JuMary WilliamsonYinhan Liu\* Jing XuMyle OttKurt ShusterEric M. SmithY-Lan BoureauJason, “Recipes for building an open-domain chatbot,” 2020.
- [3] 杉山弘晃,成松宏美,水上雅博,有本庸浩,千葉祐弥,目黒豊美,中嶋秀治, “Transformer encoder-decoder モデルによる趣味雑談システムの構築,” 第 90 回言語・音声理解と対話処理研究会(第 11 回対話システムシンポジウム), 2020.
- [4] 東中竜一郎, 船越孝太郎, 高橋哲朗, 稲葉通将, 角森唯子, 赤間怜奈, 宇佐美まゆみ, 川端良子, 水上雅博, 小室允人, ドルサ・テヨルス, “対話システムライブコンペティション 3,” 人工知能学会研究会資料 言語・音声理解と対話処理研究会 90 回 (p. 23), 2020.
- [5] Oriol Vinyals, Quoc V. Le, “A Neural Conversational Model,” 2015.
- [6] Ryuichiro Higashinaka, Kotaro Funakoshi, Michimasa Inaba, Yuiko Tsunomori, Tetsuro Takahashi, Nobuhiro Kaji, “Dialogue Breakdown Detection Challenge 4,” 2016.
- [7] 東中竜一郎, 船越孝太郎, 小林優佳, & 稲葉通将, “対話破綻検出チャレンジ,” 第 75 回言語・音声理解と対話処理研究会(第 6 回対話システムシンポジウム),人工知能学会研究会資料 SIG-SLUD-75-B502, pp. 27-32, 2015.
- [8] 稲葉通将, 高橋健一, “対話破綻検出の対話システムへの適用,” 人工知能学会論文誌, 34.3: B-I64\_1-8., 2019.
- [9] Onishi, K., Yoshimura, T., “asual conversation technology achieving natural dialog with computer,” NTT DOCOMO Technical Jouranl15(4), 16-21, 2014.
- [10] Tsukahara, H., Uchiumi, K., “System utterance generation by label propagation over association graph of words and utterance patterns for open-domain dialogue systems,” Proc.PACLIC, pp. 323-331, 2015.
- [11] Alan Ritter, Colin Cherry, and William B. Dolan., “Data-driven response generation in social media,” Proc. EMNLP, pp. 583-593, 2011.
- [12] Higashinaka, R., Funakoshi, K., Inaba, M., Tsunomori, Y., Takahashi, T., Akama, R, “Dialoguesystem live competition: identifying problems with dialogue systems through live event.,” Proc. IWSDS, 2019.
- [13] Sugiyama, Hiroaki, “ Dialogue breakdown detection based on estimating

appropriateness of topic transition,” Dialog system technology challenges (DSTC6), 2017.

- [14] Yinfei Yang, Daniel Cer, Amin Ahmad, Mandy Guo, Jax Law, Noah Constant, Gustavo Hernandez Abrego , Steve Yuan, Chris Tar, Yun-hsuan Sung, Ray Kurzweil, “Multilingual Universal Sentence Encoder for Semantic Retrieval,” Association for Computational Linguistics, 2019.
- [15] HIGASHINAKA RYUICHIRO, FUNAKOSHI KOTARO, ARAKI MASAHIRO, TSUKAHARA HIROSHI, KOBAYASHI YUKA, MIZUKAMI MASAHIRO, “Text Chat Dialogue Corpus Construction and Analysis of Dialogue Breakdown,” 自然言語処理 23(1), 59-86, 2016.
- [16] University Tohoku, “ <https://huggingface.co/>, ” [ オンライン ]. Available: <https://huggingface.co/cl-tohoku/bert-base-japanese-whole-word-masking>. [アクセス日: 2022 1 12].
- [17] e. a. Saizheng Zhang, “Personalizing Dialogue Agents: I have a dog, do you have pets too?,” 2018.
- [18] Sugiyama, Hiroaki, “Dialogue breakdown detection using BERT with traditional dialogue features,” Springer Nature, 2021.