学修番号 17890510

修士論文

敵対的学習を用いた対話システムの自動評価

尾形　朋哉

2019 年 2 月 22 日

首都大学東京大学院
システムデザイン研究科 情報通信システム学域

尾形　朋哉

審査委員：

小町 守 准教授　（主指導教員）

山口 亨 教授　（副指導教員）

高間 康史 教授　（副指導教員）

敵対的学習を用いた対話システムの自動評価*

尾形　朋哉

修論要旨

近年，クラウド技術の発達に伴い大量のデータを安価に収集できるようになったことで，データドリブンな手法で問題を解くことが可能になり，幅広いタスクにおいてニューラルネットワークが用いられるようになった．ユーザの発話文に対して自動で応答することを目的にした対話システムについてもニューラルネットワークを用いた研究が盛んに行われており，サービスに対話システムを取り入れる企業も増えるなど，産業界においても対話システムの利用は注目を浴びている．

様々なニューラル対話システムが提案されている一方で，対話システムの性能を評価するための自動評価尺度は明確に定まっておらず，多くの対話システムの研究において発話文に対するシステム応答の妥当性を評価するために人手評価がなされている．しかし，人手評価は評価者の主観の影響を強く受けるためシステム間の相対的な評価には不向きであり，定量的評価が必要となる．そのため，人的コストを減らすだけでなく，システム間の性能を正しく評価するためにも，自動評価尺度の確立は重要である．

対話システムのための自動評価尺度には，機械翻訳と同様に BLEU や Perplexity などが伝統的に用いられている．これらは発話文に対する正解の応答文とシステム応答の単語の一致率を測るものである．しかし，ある発話に対して様々な応答文が正解となり得るので，これらの自動評価尺度は人手評価との相関がないという問題がある．一方で，deltaBLEU は発話文に対して応答文候補とその人手評価スコアを複数用意することで，多様な応答文を考慮できるように BLEU を改良した自動評価尺度であり，deltaBLEU を用いた評価が人手評価と弱い相関があることを示した．しかしながら，評価データにおける発話文に対して応答文を人手評価する必要があることや，ある発話文に対して複数の応答文候補を作成する手法が確立され

*首都大学東京大学院 システムデザイン研究科 情報通信システム学域 修士論文, 学修番号 17890510, 2019 年 2 月 22 日.

ていないことから，一般的な評価データに deltaBLEU を適用するのは実用的ではない．

また，対話システムの評価に関連して対話破綻検出の研究がある．これらの研究は対話における破綻は避けられないという前提で，対話破綻を検知することを目的としており，対話システムの自動評価を目指す本研究とは目的が異なる．

本研究では評価のための正解の応答文候補の作成やラベル付けされたデータを学習に用いずに，対話システムの性能を入力の発話文に対するシステム応答の妥当性に基づき，自動で評価することを目標とする．本研究では入力に対して妥当な応答文を識別するような識別器を対話システムに対して敵対的に学習することで，システム応答を評価するために用いる．まず，対話システムの生成器には Encoder Decoder モデルを用いる．このモデルは RNN により構成され，入力の発話文に対し，学習データにおいて生成確率が最大となる応答文を生成する．また，識別器は入力された発話文と応答文を RNN によりベクトル化し，応答文が生成されたものであるかどうかを識別するモデルである．そして，本研究では識別器の性能を向上させるために，お互いの出力をもとに双方のモデルのパラメータを動的に更新し，敵対的に学習させる．具体的には，対話システムでは識別器に識別されないような妥当性の高い応答文を出力するように学習が進み，識別器では妥当性が高くなる対話システムによる応答文を識別できるように学習が進む．本研究では，この識別器に対して発話文と応答文の対を入力として与えた時に予測される応答文が正解である確率をシステム応答のスコアとして利用する．

提案手法に対して対話破綻検出チャレンジの日本語と英語のデータを用いて実験を行う．このデータはシステム応答に対して，対話破綻の可能性が 3 段階でラベル付けされたユーザとシステムの対話ログであり，提案モデルの学習には発話文とそれに対する応答文のみを用いる．本研究では，データに付けられたラベルを得点化したものを人手評価スコアとして扱い，識別器の予測スコアと人手評価スコアの相関を計算することで識別器の性能を評価する．日本語と英語のそれぞれのデータセットで，ベースラインよりも人手評価スコアとの相関が高くなることを示す．また，日本語と英語による実験の結果を分析し，データセットや言語による識別器の学習への影響について考察を行う．

本研究の主要な貢献は以下の通りである．

- 本論文では応答文の質を自動評価するための手法を提案した．提案手法は学習においてラベル付けされたデータを用いず，評価データにおいて人手による正解の応答文候補の作成を必要としないため，低コストで対話システムの質を評価できる．
- 対話破綻検出チャレンジに付けられたラベルを人手評価スコアとして利用し，提案モデルによる評価がベースラインより人手評価スコアとの相関が高くなることを示した．
- 生成器と識別器を敵対的に学習することで，識別器による評価が人手評価スコアとの相関が高くなることを示した．
- 英語データにおいても，同様に識別器を学習し，応答文の評価に利用できることを示した．また，日本語データにおける結果と交えて分析を行い，識別器を応答文の評価に利用できる設定について考察した．

本論文の構成は以下のようになっている．第 1 章では概要と貢献を述べる．第 2 章では対話システムおよびその評価尺度に対する関連研究について述べる．第 3 章では提案手法である対話システムの自動評価のためのモデルについて述べる．第 4 章では実験設定と実験結果を示す．第 5 章では結果の考察を行う．最後に第 6 章では本研究のまとめについて述べる．

Adversarial Evaluation for Dialog system*

Tomoya Ogata

Abstract

In recent years, along with the development of cloud computing, it became possible to collect a large amount of data at low cost. Therefore, it became possible to solve the problem with a data-driven method, and the neural network has been applied to a wide range of tasks. Research on a dialogue system aiming to automatically respond to user utterance has also made much use of research using a neural network. In addition, the development of dialogue systems has attracted attention in the industry, and some companies adopt a dialogue system for service. While various neural dialogue systems have been proposed, automatic evaluation metrics for evaluating the performance of the dialogue system are not clearly defined, and in the many previous work, the validity of the system response to the user utterance is evaluated by human. However, because human evaluation is strongly influenced by the subjectivity of the evaluator, it is not suitable for relative evaluation between systems, thus quantitative evaluation is required. For that reason, it is important to establish an automatic evaluation metric not only to reduce human costs but also for evaluating the performance between systems.

BLEU and Perplexity etc. are traditionally used for automatic evaluation metric in dialogue system. These measure the matching rate of words between the gold response to the input utterance and the generated response. However, since multiple responses can be correct answers to a certain utterance, there is a problem that there is no correlation with human evaluation. On the other hand, deltaBLEU is BLEU that has been improved so as to consider various response

*Master's Thesis, Department of Information and Communication Systems, Graduate School of System Design, Tokyo Metropolitan University, Student ID 17890510, February 22, 2019.

sentences by creating multiple response candidates and their human evaluation scores. It is shown that the evaluation using deltaBLEU has a weak correlation with human evaluation. However, it is necessary to manually evaluate the response sentence in the evaluation data, and there is not established a method for creating multiple response candidates for an utterance, so It is not practical to apply deltaBLEU to general evaluation data.

In relation to the automatic evaluation of the dialogue system, there is research on dialogue breakdown detection. This line of researches is aimed at detecting dialogue collapse on the assumption that breakdown in dialogue can not be avoided, so its purpose is different from my research aiming at automatic evaluation of dialogue system.

In this paper, I aim to automatically evaluate the performance of the dialogue system based on the validity of the system response to the user utterance without creating the gold response candidates at the time of evaluation or using the labeled data for training. In my research, I train a discriminator, which discriminates valid response sentence to the input utterance, adversarially with the dialogue system and evaluate system responses. First, I use the Encoder Decoder model as a generator of dialogue system, which generate a response sentence that maximizes the probability of the response to the input sentence in the training data. In addition, the discriminator is a model which vectorizes the input sentence and response sentence with RNN and discriminates whether or not the response sentence is a correct answer. Then, in order to improve the performance of the discriminator, parameters of both models are dynamically updated based on their respective output and trained adversarially. Concretely, the parameters of the generator are learned so as to output highly valid response sentences which are not discriminated by the discriminator, and those of the discriminator are learned so that it can discriminate the response sentence by the dialogue system. In this paper, I use the probability that is predicted when giving a pair of an utterance and its response as input to this discriminator as the score of the system response.

I experiment my proposed method in both Japanese and English data sets of dialog breakdown detection challenge. This data set is a conversation log whose system responses are labeled with three stages of possibility of dialogue breakdown, and I only use the input sentence and response sentence for learning of the proposed model, and I don't use the label. In my research, I treat scores of labels attached to data sets as human score and evaluate the performance of the discriminator by calculating correlation with human score. In both Japanese and English, it shows that the evaluation of the proposed models is higher correlation with the human score than that of baselines. In addition, I analyze the results of experiments in Japanese and English, and consider the influence to training discriminator by dataset and language. The main contribution of this research is as follows.

- In this paper, I proposed a method to automatically evaluate the quality of response sentences. The proposed method can evaluate the quality of the dialogue system at low cost because it does not need labeled data in the training and dose not need to create multiple gold responses manually in evaluation data.
- I use the label attached to the dialog breakdown detection challenge for the evaluation of the discriminator, and it showed that the evaluation of the proposed models is higher correlation with the human score than that of baselines.
- By adversarially training the discriminator with the dialogue system, it showed that the correlation with the human score becomes higher.
- I also trained the discriminator in English data and showed that it can be used for evaluating response sentences. In addition, I analyzed the results in both Japanese and English, and consider settings that the discriminator can be used as the evaluation of the response sentence.

The structure of this thesis is shown below. In Section, I show the abstract and contribution of this thesis. In Section 2, I describe related work on the dialogue

system and its evaluation metric. In Section 3, I describe a model for automatic evaluation of the dialogue system which is my proposed method. In Section 4, experiment setup and experiment results are shown. In Section 5, I consider the result. Finally, Section 6 describes the summary of this research.

目次

図目次

第 1 章　はじめに

近年，クラウド技術の発達に伴い大量のデータを安価に収集できるようになったことで，ニューラルネットワークを用いてデータドリブンに問題を解くことが可能になり，幅広いタスクにおいてニューラルネットワークが用いられている．ユーザの発話文に対して自動で応答することを目的にした対話システム*についてもニューラルネットワークを用いた研究が盛んに行われている．また，LINE Clova†や Amazon Alexa [1] など対話システムの技術を取り入れたサービスが提供されており，産業界においても対話システムの利用は注目を浴びている．

様々なニューラル対話システムが提案されている一方で，対話システムの性能を評価するための自動評価尺度は明確に定まっておらず，多くの対話システムの研究において発話文に対するシステム応答の妥当性を評価するために人手評価がなされている．しかし，人手評価は評価者の主観の影響を強く受けるためシステム間の相対的な評価には不向きであり，定量的評価が必要となる．そのため，人的コストを減らすだけでなく，システム間の性能を正しく評価するためにも，自動評価尺度の確立は重要である．

対話システムのための自動評価尺度には，機械翻訳と同様に BLEU [2] や Perplexity [3] などが伝統的に用いられている．これらは発話文に対する正解の応答文とシステム応答の単語の一致率を測るものである．しかし，ある発話に対して様々な応答文が正解となり得るので，これらの自動評価尺度は人手評価との相関がないという問題がある．一方で，deltaBLEU [4] は発話文に対して応答文候補とその人手評価スコアを複数用意することで，多様な応答文を考慮できるように BLEU を改良した自動評価尺度であり，deltaBLEU を用いた評価が人手評価と弱い相関があることを示した．しかしながら，評価データにおける発話文に対して応答文を人手評価する必要があることや，ある発話文に対して複数の応答文候補を作成する手法が確立されていないことから，一般的な評価データに deltaBLEU を適用するのは実用的ではない．

また，対話システムの評価に関連して対話破綻検出の研究がある．これらの研究

---

*本論文での対話システムはテキストベースで入出力が行われるものとする．

†https://clova.line.me/

は対話における破綻は避けられないという前提で，対話破綻を検知することを目的としており，対話システムの自動評価を目指す本研究とは目的が異なる．

本研究では評価のための正解データ候補の作成やラベル付けされたデータを学習に用いずに，対話システムの性能を入力の発話文に対するシステム応答の妥当性に基づき，自動で評価することを目標とする．本研究では入力に対して妥当な応答文を識別するような識別器を対話システムに対して敵対的に学習することで，システム応答を評価するために用いる．まず，対話システムの生成器として Encoder Decoder モデルを用いる．このモデルは RNN により構成され，入力の発話文に対し，学習データにおいて生成確率が最大となる応答文を生成する．また，識別器は入力された発話文と応答文を RNN によりベクトル化し，応答文が生成されたものであるかどうかを識別するモデルである．そして，本研究では識別器の性能を向上させるために，お互いの出力を基に双方のモデルのパラメータを動的に更新し，敵対的に学習させる．具体的には，対話システムでは識別器に識別されないような妥当性の高い応答文を出力するように学習が進み，識別器では妥当性が高くなる対話システムによる応答文を識別できるように学習が進む．本研究では，この識別器に対して発話文と応答文の対を入力として与えた時に予測される確率をシステム応答のスコアとして利用する．

提案手法に対して対話破綻検出チャレンジ [5, 6] の日本語と英語の対話破綻の可能性がラベル付けされたデータを用いて実験を行う．提案モデルの学習には発話文とそれに対する応答文のみを用いる．本研究では，データに付けられたラベルを得点化したものを人手評価スコアとして扱い，人手評価スコアとの相関を計算することで識別器を評価する．特に日本語における提案手法の実験において，提案モデルがベースラインよりも人手評価スコアとの相関が高くなることを示す．また，日本語と英語による実験の結果を分析し，データセットや言語による識別器の学習への影響について考察を行う．

本研究の主要な貢献は以下の通りである．

- 本論文では応答文の質を対話システムの妥当性に基づいて自動評価するための手法を提案した．提案手法は学習においてラベル付けされたデータを用いず，評価データにおいて人手による正解の応答文候補の作成を必要としないため，低コストで対話システムの性能を評価できる．

- 対話破綻検出チャレンジのデータに付けられたラベルを識別器の評価に利用し，提案モデルがベースラインより人手評価スコアとの相関が高くなることを示した．
- 対話システムと識別器を敵対的に学習することで，人手評価スコアとの相関が高くなることを示した．
- 英語データにおいても，同様に識別器を学習し，応答文の評価に利用できることを示した．また，日本語データにおける結果と交えて分析を行い，識別器を応答文の評価として利用できる設定について考察した．

第 2 章　関連研究

近年, 与えられた発話文に対して尤もらしい発話を出力するようにニューラルネットワークを学習する End-to-end な対話システムが注目を浴びている [7, 8, 9, 10]. 一方で，これらのシステムを評価するのは困難な問題であり，ほとんどの対話システムにおいて BLEU や Perplexity などの生成された応答文と実際の対話における正解の応答を比較する評価尺度が一般的に用いられている．しかし，ある発話文に対して尤もらしい応答文は複数存在する場合が多く，これらの手法では対話システムの性能を正確に評価することはできない.

評価データ中の発話文に対して応答文候補とその人手評価スコアを複数用意することで，多様な応答文を考慮できるように BLEU を改良した deltaBLEU が存在する [4]．deltaBLEU は人手評価と弱い相関があることが示されたが，評価データにおける発話文に対して応答文を人手評価する必要があることや，ある発話文に対して複数の応答文候補を作成する手法が確立されていないことから，一般的な評価データに適用するのは実用的ではない.

これらの伝統的な自動評価尺度が人手評価と相関がないことが示されている [11]．そのため，対話システムの性能を評価するために人手評価が用いられることがある [7]．これらの研究では複数の対話システムにより生成された応答文に対し，どれが良い応答文であるかを評価者に選択してもらうことで対話システムの質を評価する．しかし，人的コストが高いうえ，評価者の質や設定に応じて結果が変化するため，システム間の相対的な評価には向いていない．この人的コストが高い問題に対して，多様な応答に対する人間の評価スコアを学習データとして利用し，人が評価時に考慮する特徴を捉えてスコアを予測するようにニューラルネットワークを学習し，自動評価に用いた研究がある [12]．この研究において，人間の評価スコアを使って学習したモデルによる評価が人手評価と高い相関があることを示したが, 予測するスコアが学習データに影響を受けるという問題やデータにスコア付けする必要がある．一方，[13, 14] では対話システムにより生成された応答文と実際の対話における応答文を識別するように学習を行うことで，ラベル付けされたデータを必要とせず文が妥当かどうかを予測することで対話システムの評価を行なっているが，応答文に対してスコアの予測を行なっておらず，対話システムと識別器の双方

のモデルパラメータを更新する敵対的学習を行なっていない．

機械翻訳の研究において敵対的な設定で学習することで，BLEU によって正しく評価できない事例に対して正しく評価できるようになることが示されている [15]．本研究では松村ら [15] が機械翻訳の研究で用いた敵対的学習の枠組みを対話システムの自動評価のために適応することを提案する．本研究の提案手法は学習データに対するラベル付けや deltaBLEU のように複数の正解の応答文候補を作成するなどの人的コストを必要とせず，対話システムによって生成された応答文を妥当性に基づき自動評価する．

第 3 章　敵対的学習を用いた対話システムの評価

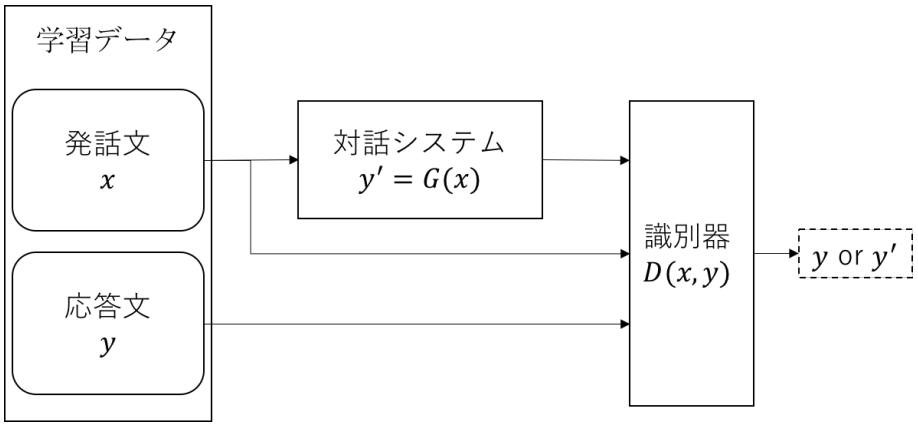


図 3.1: 敵対的学習の概要

この章では提案手法である対話システムを評価するための識別器とそれを敵対的学習を用いて訓練するための枠組みについて詳しく述べていく．

本研究で敵対的学習を行うための全体のネットワークは図 3.1 のように対話システムと識別器を組み合わせたものであり，対話システムは識別器の評価を基に学習を行い，識別器は入力された応答文が学習データのものか対話システムにより出力されたものかを識別できるように学習する．対話システムのモデルについては 3.1 節，識別器のモデルについては 3.2 節で述べる．

3.1　対話システム

発話文から応答文を予測する対話システムの生成器としては図 3.2 に示すような Encoder と Decoder から成る Encoder Decoder が一般的に用いられる．Encoder では入力された発話文の単語系列をリカレントニューラルネットワーク (RNN) に与え，発話文の情報を隠れ層のベクトルへ圧縮する．Decoder では隠れ層のベクトルの情報を基に単語を予測していき，応答文の生成を行う．この Encoder Decoder は学習データ中の発話文に対して応答文の単語の予測確率が最大となるようにパラメータを更新することで，入力の発話文に対する尤もらしい応答文を生成できるよ

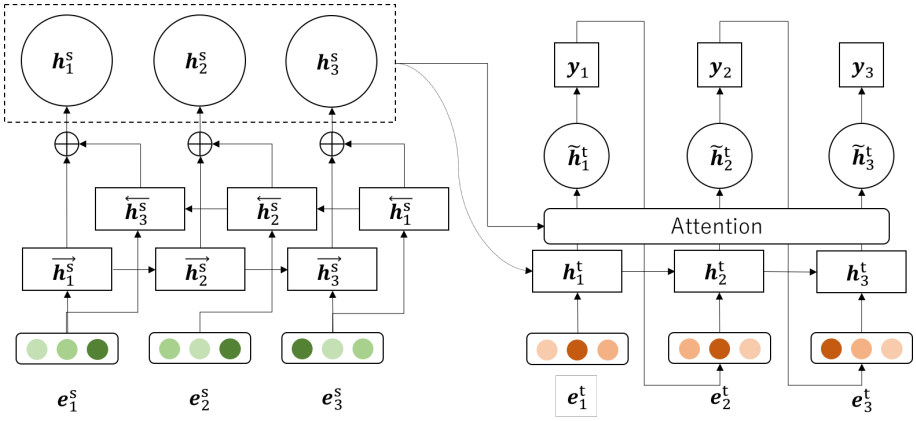


図 3.2: Encoder Decoder の概要

うに学習される．本研究では，Luong ら [16] が提案したアテンション機構付きの Encoder Decoder を用いる．

具体的な Encoder Decoder の処理を数式と共に示す．まず, Encoder Decoder に発話文が入力されると，発話文のそれぞれのトークンはそれぞれの次元が Encoder 側の語彙に対応するような one-hot ベクトルの系列 $\boldsymbol{X} = [\boldsymbol{x}_1, \cdots, \boldsymbol{x}_{|\boldsymbol{X}|}]$ へと変換される．その後，i 番目の one-hot ベクトル $\boldsymbol{x}_i$ は線形変換され，活性化関数 tanh にかけられることで埋め込み表現 $\boldsymbol{e}^{\mathrm{s}}_i$ に変換される．埋め込み表現はそれぞれのトークンの意味を表現するベクトルである．

$$\boldsymbol{e}^{\mathrm{s}}_i = \tanh(\boldsymbol{W}_{\mathrm{x}} \boldsymbol{x}_i) \tag{3.1}$$

ここで，$\boldsymbol{W}_{\mathrm{x}} \in \mathbb{R}^{q \times v_{\mathrm{x}}}$ は重み行列であり，q は埋め込み表現の次元数で，v_{x} が Encoder 側の語彙サイズを表している．単語系列を左から右に入力するのを順方向，右から左に入力するのを逆方向とした時，Encoder の隠れ層は順方向の LSTM [17] と逆方向の LSTM を組み合わせた BiLSTM を用いて次式のように計算される．なお，区別のため Encoder の隠れ層と Decoder の隠れ層をそれぞれ $\boldsymbol{h}^{\mathrm{s}}$, $\boldsymbol{h}^{\mathrm{t}}$ とする．

$$\boldsymbol{h}^{\mathrm{s}}_i = \mathrm{BiLSTM}(\boldsymbol{e}^{\mathrm{s}}_i) = \overrightarrow{\boldsymbol{h}^{\mathrm{s}}_i} + \overleftarrow{\boldsymbol{h}^{\mathrm{s}}_{|\boldsymbol{X}|+1-i}} \tag{3.2}$$

ここで $\overrightarrow{\boldsymbol{h}^{\mathrm{s}}_i}$ と $\overleftarrow{\boldsymbol{h}^{\mathrm{s}}_i}$ はそれぞれ次のように計算される．

$$\overrightarrow{\boldsymbol{h}^{\mathrm{s}}_i} = \overrightarrow{\mathrm{LSTM}}(\overrightarrow{\boldsymbol{h}^{\mathrm{s}}_{i-1}}, \boldsymbol{e}^{\mathrm{s}}_i) \tag{3.3}$$

$$\overleftarrow{\boldsymbol{h}_i^{\mathrm{s}}} = \overleftarrow{\mathrm{LSTM}}(\overleftarrow{\boldsymbol{h}_{i-1}^{\mathrm{s}}}, \boldsymbol{e}_i^{\mathrm{s}}) \tag{3.4}$$

Encoder で計算された隠れ層のベクトル $\boldsymbol{h}^{\mathrm{s}}$ は，Decoder で応答文の単語を予測するために利用される．

Decoder の隠れ層 $\boldsymbol{h}_j^{\mathrm{t}}$ は前のステップで出力された単語の埋め込みベクトル $\boldsymbol{e}_j^{\mathrm{t}}$ を用いて順方向の LSTM をかけて計算される．

$$\boldsymbol{h}_j^{\mathrm{t}} = \mathrm{LSTM}(\boldsymbol{h}_{j-1}^{\mathrm{t}}, \boldsymbol{e}_j^{\mathrm{t}}) \tag{3.5}$$

学習時において $\boldsymbol{e}_j^{\mathrm{t}}$ は，正解の応答文を one-hot ベクトルの系列に変換した $\boldsymbol{Y} = [\boldsymbol{y}_1, \cdots, \boldsymbol{y}_{|\boldsymbol{Y}|}]$ の j 番目の要素 $\boldsymbol{y}_j$ を線形変換し，活性化関数 tanh をかけることで得られる．また，評価時には前のステップで予測された単語を用いて埋め込みベクトル $\boldsymbol{e}_j^{\mathrm{t}}$ を計算する．

$$\boldsymbol{e}_j^{\mathrm{t}} = \tanh(\boldsymbol{W}_{\mathrm{y}} \boldsymbol{y}_j) \tag{3.6}$$

ここで，$\boldsymbol{W}_{\mathrm{y}} \in \mathbb{R}^{q \times v_{\mathrm{y}}}$ は重み行列であり，v_{y} は Decoder 側の語彙サイズを表している．隠れ層 $\boldsymbol{h}_j^{\mathrm{t}}$ に対してアテンションは Encoder の隠れ層 $\boldsymbol{h}_i^{\mathrm{s}}$ との内積に対し，ソフトマックス関数を取ることで計算される．

$$\boldsymbol{a}_{i,j} = \mathrm{softmax}(\boldsymbol{h}_j^{\mathrm{t}\,\mathrm{T}} \boldsymbol{h}_i^{\mathrm{s}}) = \frac{\exp(\boldsymbol{h}_j^{\mathrm{t}\,\mathrm{T}} \boldsymbol{h}_i^{\mathrm{s}})}{\sum_{i'}^{|\boldsymbol{X}|} \exp\left(\boldsymbol{h}_j^{\mathrm{t}\,\mathrm{T}} \boldsymbol{h}_{i'}^{\mathrm{s}}\right)} \tag{3.7}$$

入力の文脈情報 $\boldsymbol{c}_j$ はアテンションによる入力の隠れ層 $\boldsymbol{h}_i^{\mathrm{s}}$ の重み付き和で計算される．

$$\boldsymbol{c}_j = \sum_{i}^{|\boldsymbol{X}|} \boldsymbol{a}_{i,j} \boldsymbol{h}_i^{\mathrm{s}} \tag{3.8}$$

各ステップの単語の予測に用いられる最終的な隠れ層 $\tilde{\boldsymbol{h}}_j^{\mathrm{t}}$ は入力の文脈情報 $\boldsymbol{c}_j$ を考慮して次式で計算される．

$$\tilde{\boldsymbol{h}}_j^{t} = \tanh(\boldsymbol{W}_{\mathrm{c}}[\boldsymbol{c}_j; \boldsymbol{h}_j^{\mathrm{t}}] + \boldsymbol{b}_{\mathrm{c}}) \tag{3.9}$$

ここで，$\boldsymbol{W}_{\mathrm{c}} \in \mathbb{R}^{r \times 2r}$ は重み行列であり，r は隠れ層の次元数，$\boldsymbol{b}_{\mathrm{c}}$ はバイアスを表している．

応答文の j 番目の単語の生成確率は隠れ層 $\tilde{\boldsymbol{h}}_j^{\mathrm{t}}$ を用いて次の式で計算する．

$$\boldsymbol{P}_{\boldsymbol{\theta}}(\boldsymbol{y}_j|\boldsymbol{Y}_{<j}, \boldsymbol{X}) = \mathrm{softmax}(\boldsymbol{W}_{\mathrm{g}}\tilde{\boldsymbol{h}}_j^{\mathrm{t}} + \boldsymbol{b}_{\mathrm{g}}) \tag{3.10}$$

ここで，$\boldsymbol{W}_{\mathrm{g}} \in \mathbb{R}^{v_y \times r}$ は重み行列であり，r は隠れ層の次元数，$\boldsymbol{b}_{\mathrm{g}}$ はバイアスを表している．

これらのネットワークのパラメータ $\boldsymbol{\theta}$ は発話文と応答文が対となった N 件の学習データを用いて，次の対数尤度を最大化するように学習される．

$$\mathcal{L}_{\boldsymbol{\theta}} = \frac{1}{N}\sum_{n=1}^{N}\sum_{j=1}^{|\boldsymbol{Y}^{(n)}|} -\log \boldsymbol{P}_{\boldsymbol{\theta}}(\boldsymbol{y}_j^{(n)}|\boldsymbol{Y}_{<j}^{(n)}, \boldsymbol{X}^{(n)}) \tag{3.11}$$

3.2 識別器

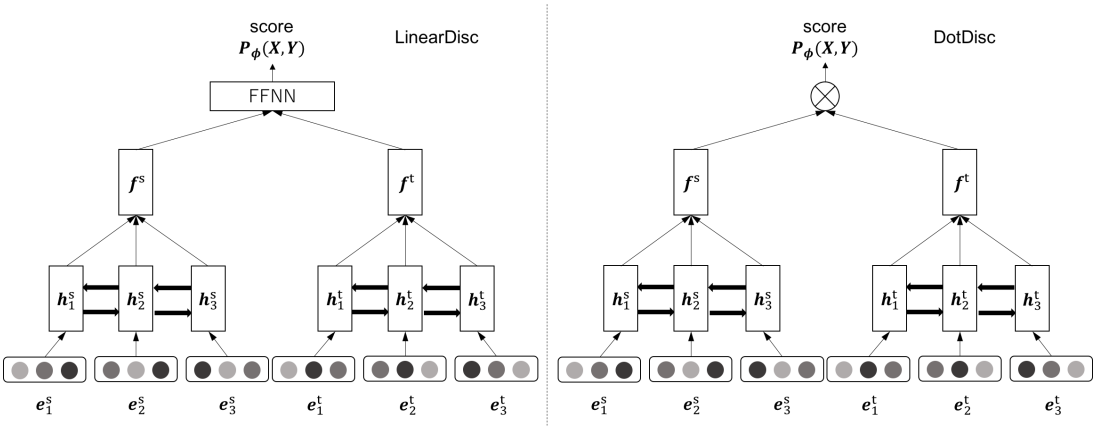


図 3.3: 識別器のネットワーク

識別器は松村ら [15] と同様のネットワークを用いる．入力された発話文と応答文は Encoder Decoder と同様にそれぞれ one-hot ベクトルの系列 $\boldsymbol{X} = [\boldsymbol{x}_1, \boldsymbol{x}_2, ..., \boldsymbol{x}_{|\boldsymbol{X}|}]$ と $\boldsymbol{Y} = [\boldsymbol{y}_1, \boldsymbol{y}_2, ..., \boldsymbol{y}_{|\boldsymbol{Y}|}]$ へと変換され，埋め込み表現が計算される．

$$\boldsymbol{e}_i^{\mathrm{s}} = \tanh(\boldsymbol{W}_{\mathrm{x}}\boldsymbol{x}_i),\ \boldsymbol{e}_i^{\mathrm{t}} = \tanh(\boldsymbol{W}_{\mathrm{y}}\boldsymbol{y}_i)$$

ここで計算された埋め込み表現を BiLSTM に入力することで隠れ層のベクトルが得られる．

$$\boldsymbol{h}_i^{\mathrm{s}} = \mathrm{BiLSTM}(\boldsymbol{e}_i^{\mathrm{s}}),\ \boldsymbol{h}_i^{\mathrm{t}} = \mathrm{BiLSTM}(\boldsymbol{e}_i^{\mathrm{t}})$$

ここで，入力文の系列単位に対する全ての隠れ層ベクトルに対し平均をとって得られるベクトルを文のベクトル表現として扱う．

$$\boldsymbol{f}^{\mathrm{s}} = \mathrm{average}([\boldsymbol{h}_1^{\mathrm{s}}, \boldsymbol{h}_2^{\mathrm{s}}, ..., \boldsymbol{h}_{|\boldsymbol{X}|}^{\mathrm{s}}]),\ \boldsymbol{f}^{\mathrm{t}} = \mathrm{average}([\boldsymbol{h}_1^{\mathrm{t}}, \boldsymbol{h}_2^{\mathrm{t}}, ..., \boldsymbol{h}_{|\boldsymbol{Y}|}^{\mathrm{t}}]) \quad (3.12)$$

$\boldsymbol{f}^{\mathrm{s}}, \boldsymbol{f}^{\mathrm{t}}$ から発話文に対して応答文が正解である確率 $\boldsymbol{P}_{\boldsymbol{\phi}}$ を計算する．本研究では $\boldsymbol{P}_{\boldsymbol{\phi}}$ の計算方法により図 3.3 に示すように DotDisc と LinearDisc の 2 つのネットワークを構築した．DotDisc は $\boldsymbol{f}^{\mathrm{s}}$ と $\boldsymbol{f}^{\mathrm{t}}$ のドット積を用いて，発話文と応答文の正解である確率 $\boldsymbol{P}_{\boldsymbol{\phi}}$ を計算する．

$$\boldsymbol{P}_{\boldsymbol{\phi}}(\boldsymbol{X}, \boldsymbol{Y}) = \mathrm{sigmoid}(\boldsymbol{f}^{\mathrm{s}} \cdot \boldsymbol{f}^{\mathrm{t}}) \quad (3.13)$$

一方で，LinearDisc はドット積で確率を計算する代わりに重み行列 $\boldsymbol{W}_{\mathrm{p}} \in \mathbb{R}^{2r \times 1}$ を用いて計算する．

$$\boldsymbol{P}_{\boldsymbol{\phi}}(\boldsymbol{X}, \boldsymbol{Y}) = \mathrm{sigmoid}(\boldsymbol{W}_{\mathrm{p}}[\boldsymbol{f}^{\mathrm{s}}; \boldsymbol{f}^{\mathrm{t}}]) \quad (3.14)$$

識別器は式 3.11 により学習した対話システムの生成文を識別するように，3.3 節で示す式 3.16 に従い学習される．

3.3 敵対的学習における応答文生成器と識別器の目的関数

敵対学習において対話システムの応答文生成器は出力した応答文に対する識別器の予測スコアを用いて，目的関数 $\mathcal{L}_G$ が最大となるようにパラメータを更新することで，識別器が正解の応答文と区別することができないような応答文を生成するように学習される．

$$\mathcal{L}_G(\boldsymbol{\theta}, \boldsymbol{\phi}) = \frac{1}{N} \sum_{n=1}^{N} \Big\{ \sum_{j=1}^{|\boldsymbol{Y}|} \log \boldsymbol{P}_{\boldsymbol{\theta}}(\boldsymbol{y}_j^{(n)} | \boldsymbol{Y}_{<j}^{(n)}, \boldsymbol{X}^{(n)}) + \log \boldsymbol{P}_{\boldsymbol{\phi}}(\boldsymbol{X}^{(n)}, \boldsymbol{Y}^{(n)}) \Big\} \quad (3.15)$$

識別器は目的関数 $\mathcal{L}_D$ を最大化するようにモデルのパラメータ $\boldsymbol{\phi}$ を更新し，応答文生成器によって出力された応答文 $\tilde{\boldsymbol{Y}}$ と正解の応答文 $\boldsymbol{Y}$ を識別するように学習される．

$$\mathcal{L}_D(\boldsymbol{\phi}) = \frac{1}{N}\sum_{n=1}^{N}\Big\{\log \boldsymbol{P}_{\boldsymbol{\phi}}(\boldsymbol{X}^{(n)}, \boldsymbol{Y}^{(n)}) + \log\big\{1 - \boldsymbol{P}_{\boldsymbol{\phi}}(\boldsymbol{X}^{(n)}, \tilde{\boldsymbol{Y}}^{(n)})\big\}\Big\} \tag{3.16}$$

ここで，$\boldsymbol{\theta}$ は応答文生成器の全てのパラメータであり，$\boldsymbol{\phi}$ は識別器の全てのパラメータである．

第 4 章　識別器による対話応答文の評価

4.1　実験設定

本研究では日本語と英語において入力の発話文に対して応答文を出力する対話システムと，入力された応答文が発話文に対して正解かどうかを識別する識別器を学習する．また，日本語においては対話システムと識別器の入力として入力の発話文のさらに 1 つ前のターンの発話文を対話履歴として，入力の発話文と特殊記号“<eot>”で結合して与えることで対話履歴を考慮した実験も行う．

3 章で説明した生成器と識別器を学習するための日本語コーパスとして，対話破綻検出チャレンジのデータ [5, 6] から発話文と応答文の対データを作成する．このデータはシステムとユーザの会話による対話ログであり，対話履歴ありの場合には直前の 2 発話を発話文，それに続く発話文を応答文としデータを作成し，対話履歴なしの場合には連続する 2 発話を発話文と応答文としてデータを作成した．また，開発データと評価データについてはシステムの発話が応答文となるようにデータを作成した．これにより作成したデータは，学習データが 22,920 文対，開発データは 1,500 文対，評価データは 3,000 文対である．また，人手評価スコアについての詳細は後に述べるが，対話破綻検出チャレンジのシステムの応答文に対する人手評価スコアが 1.0 の発話文と応答文の対を抽出し，15,860 文対から成る Filter データを作成し，学習データとして用いることでノイズによる影響を調べるための実験を行う．さらに，NTCIR Short Text Conversation 日本語タスクで公開されている Twitter ID から発話文と応答文の 270,599 文対から成る Twitter データ*を作成し，応答文生成器の学習データとして用いることで，違うドメインのデータで生成器と識別器を学習した時の影響を調べる実験も行う．

また，英語コーパスとしては，2 人のクラウドワーカー同士の対話ログである PersonaChat データ [18] から抽出した発話文と応答文の 131,438 文対を生成器の学習として用いる†．また，識別器の学習及び，開発，評価データは対話破綻検出チャレンジのデータを用いて作成し，学習データが 2,150 文対，開発データと評価

*2018 年 7 月の段階でクロールしたものを利用する．

†対話破綻検出チャレンジの英語データは生成器の学習データとしては文対数が少ないため．

データはそれぞれ 1,000 文対である．

識別器の学習や評価時の入力文に未知語が複数含まれると，未知語を手掛かりとして生成文と正解文を識別するように学習してしまう場合や，適切な評価ができなくなる場合がある．そこで本研究では，入出力における未知語の割合を減少させるために学習データ中のそれぞれの文を Byte-Pair Encoding [19] を用いて subword 単位で分割する．なお，日本語と英語の語彙サイズはそれぞれ 2,000 と 3,000 とし[‡]，学習データに含まれない単語を未知語として扱い，単語ベクトルの初期値として学習済みのベクトルである bpemb を用いる [20]．bpemb は，Wikipedia のデータに対して GloVe [21] を用いて学習した subword のための単語エンベディングである．対話システムは 3.1 節で述べた Encoder Decoder モデルを用いて実験を行う．埋め込み層の次元数は 300，隠れ層の次元数は 512 とした．そして，ADAM アルゴリズム [22] で，初期学習率を 0.001 とし，バッチサイズは 128 で最適化した．

4.2 識別器のための評価尺度

対話破綻検出チャレンジのデータ [5, 6] におけるシステムの応答文について X（あきらかにおかしい），T（破綻とは言い切れないが，違和感がある），O（破綻ではない）が 30 名のアノテータによってラベル付けされている．本研究では X，T，O をそれぞれ 0, 1, 2 点とみなして，合計し正規化した値をシステムの応答文に対する人手評価スコアとして扱う．そして，ユーザの発話文とシステムの応答文を識別器に入力して出力される予測スコアと人手評価スコアとの Spearman の相関係数と Pearson の相関係数を用いて識別器の性能を評価する．発話文と応答文中の単語の一致率が高い場合，応答文は発話文と関連度の高い文となるため，妥当な応答文の自動評価に用いることができる．そこで，本研究では Simpson 係数[§]と発話文と応答文の文ベクトル[¶]の cosine 類似度を計算する CosSim をベースラインとして用いる．

[‡]日本語の Twitter データを用いるモデルについては語彙サイズ 2000 では不十分であったため，3,000 として実験を行なっている．また，マージ操作の回数はそれぞれ 5,000 と 3,000 である．

[§]分割単位が subword の場合，意味のない単位での一致が多くなり，適切な評価が行えないため MeCab（辞書 : IPADIC）による単語単位で計算した．

[¶]入力系列に対する単語ベクトルの平均を文ベクトルとする．

4.3 実験結果

日本語における敵対的学習の前後の予測スコアと人手評価スコアとの Spearman の相関係数及び Pearson の相関係数を表 4.1 に示す．敵対的学習後の DotDisc は Spearman の相関係数と Pearson の相関係数の両方でベースラインを上回っており，その他の敵対的学習後のモデルも Spearman の相関係数においてベースラインを上回っている．対話履歴なしの場合には DotDisc が LinearDisc よりも良い結果を示すが，対話履歴ありの場合には LinearDisc の方が良い結果を示している．生成器の学習データとして Twitter データを用いたモデルは，対話破綻検出チャレンジのデータで学習したモデルより人手評価スコアとの相関は低くなるが，ベースラインよりも高い相関を示している．学習データを人手評価スコアでフィルタリングしたモデルは LinearDisc では Pearson の相関係数が低下するが，DotDisc では Spearman の相関係数と Pearson の相関係数の双方で最も高い値を示した．

敵対的学習後のすべてのモデルについて予測スコアと人手評価スコアとの Spearman の相関係数が敵対的学習前より高くなっており，敵対的学習を行うことにより識別性能が向上することが分かる．さらに，日本語における敵対的学習前後の DotDisc の評価例を人手評価スコアとともに表 4.4 に示す．また，敵対的学習を行った時の日本語における対話システムの出力を表 4.3 に示す．それぞれの出力の下に敵対的学習前の識別器による予測スコアを示している．敵対的学習を行うことで，対話システムの出力のスコアが高くなっており，妥当性の高い応答文を出力できるようになっていることが分かる．

英語データにおける Spearman の相関係数と Pearson の相関係数の結果を表 4.2 に示す．敵対的学習後のモデルは Spearman の相関係数と Pearson の相関係数の両方で，ベースラインを上回っている．DotDisc と LinearDisc の比較では日本語の場合と同様に DotDisc の方が良い結果を示し，DotDisc と LinearDisc の両方で敵対的学習を行うことで識別性能が向上している．

表 4.1: 日本語データにおける識別器の評価

モデル	敵対的学習前		敵対的学習後	
	Spearman	Pearson	Spearman	Pearson
CosSim	-	-	−0.021	−0.023
Simpson 係数	-	-	0.109	0.132
DotDisc	−0.054	−0.063	0.160	0.143
LinearDisc	0.094	−0.126	0.124	0.110
DotDisc + 対話履歴	0.094	0.094	0.137	0.132
LinearDisc + 対話履歴	−0.050	−0.066	0.156	0.131
DotDisc w/ Filter	0.017	0.012	**0.166**	**0.170**
LinearDisc w/ Filter	−0.126	−0.112	0.133	0.094
DotDisc w/ Twitter	0.058	−0.021	0.143	0.138
LinearDisc w/ Twitter	0.077	−0.007	0.129	0.122

表 4.2: 英語データにおける識別器の評価

モデル	敵対的学習前		敵対的学習後	
	Spearman	Pearson	Spearman	Pearson
CosSim	-	-	2.04e-05	0.008
Simpson 係数	-	-	0.023	0.031
DotDisc	0.097	0.002	**0.146**	**0.109**
LinearDisc	−0.001	0.026	0.121	0.070

表 4.3: 発話文に対する敵対的学習前と敵対的学習後の対話システムの出力例

発話文	敵対的学習前	敵対的学習後
テレビ東京はよく見ますか	テレビは好きですか 0.023	見ないよ 0.979
洋書を言語で読むのですか	洋画が好きですよね。 0.279	英語が読めるのです 0.959
夏は特に気をつけたいですね。	海は好きですね 0.471	予防が大切ですね 0.841
ももはいかがですか？	いえいえ 0.323	いいですね！ 0.860
ドライブです。	クラケはかわゆいですね 0.135	車は得意なんですよねえ 0.838

表 4.4: 発話文に対する敵対的学習前と敵対的学習後の識別器の評価例

正しく高い値をつけることができた例	
発話文	美味しいですよね。何が好きですか？
応答文	ラーメンがいいですね
人手評価：0.917, 敵対的学習前：0.0549, 敵対的学習後：0.905	

正しく低い値をつけることができた例	
発話文	北海道は夏休みが短いんですか？
応答文	北海道に住んでいて吹雪になると家から出るのも苦労しますからね。
人手評価：0.183, 敵対的学習前：0.949, 敵対的学習後：0.019	

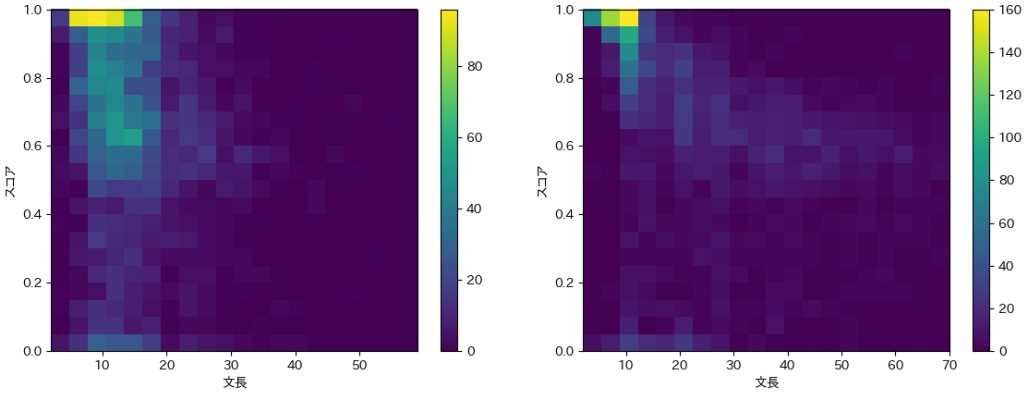
間違えて低い値をつけてしまった例	
発話文	京都も好きです
応答文	その気持ちよくわかります。あの土産屋さんがならんでいるところで、八つ橋の試食をしまくりながらただでお茶を飲んだり。
人手評価：0.917, 敵対的学習前：0.977, 敵対的学習後：0.000	

第 5 章　考察

表 4.1 や表 4.2 で示されたように敵対的学習をすることにより識別器が人手評価スコアとの相関が高い評価ができるようになっている．対話システムは表 4.3 のように，敵対的な学習を行うことで発話文に対して妥当性の高い応答文を出力できるようになっていく．一方，識別器は生成された応答を識別できるように学習を進めるので，敵対的学習の過程で妥当性が高くなる応答文の識別を可能にするために，識別に役立つより良い特徴量を捉えることができるようになる．その結果，敵対的学習を行うことで人手評価スコアとの相関が高い評価ができるようになったと考えられる．

Filter データを学習に用いたモデルは DotDisc では人手評価スコアとの相関が最も高くなった．このことから，ノイズの少ないデータを学習に用いることでより識別性能の高い識別器を作成できると考えられる．一方で，LinearDisc では相関が向上しなかったが，これは学習データが減少するために入力のベクトルとスコアとの関係を重みが上手く学習できなかったためであると推察できる．通常のデータセットにおいて DotDisc より LinearDisc の相関が低くなるのも，学習データが十分でないことが原因として考えられる．Twitter データで学習したモデルによる人手評価スコアとの相関が低くなったのは，生成器の学習と識別器の学習のドメインが離れていることや，Twitter データがノイズを多く含んでいることが原因であると考えられる．

また，表 4.1 で示されるように，対話履歴を入れても識別性能が向上しなかった．これは，学習データとして用いた対話破綻検出チャレンジのデータが人とシステムの対話ログであり，システム応答による応答文が多く含まれていることが原因として考えられる．つまり，システム応答を応答文として学習する際には対話履歴として前のシステムの応答が用いられるが，システムの応答は履歴を踏まえた応答になっていないことが多く，履歴を捉えた応答文を生成するように生成器を上手く学習できなかったためである．対話履歴ありの場合に DotDisc が LinearDisc より人手評価スコアとの相関が低くなった原因としては，DotDisc では発話文と応答文に単語の一致があるとスコアが高くなりやすく，対話履歴の単語だけを見て高いスコアをつけることがあったためである．



(a) 発話文　　(b) 応答文

図 5.1: 日本語データにおける発話文または応答文の文長と DotDisc の予測スコアの関係

次に DotDisc の自動評価の結果について分析する．表 4.4 の 1 番目と 2 番目の例は敵対的学習を行うことで適切な評価ができるようになった例である．1 番目の例は敵対的学習前は内容語の単語の一致が少なく，文全体の意味を考慮して評価しなくてはいけないが，敵対的学習後の識別器は高いスコアを正しく予測できている．2 番目の例は発話文と応答文に “北海道” が含まれており，敵対的学習前の識別器は高いスコアを予測してしまっているが，敵対的学習により文全体の妥当性に基づき低いスコアを出すことができている．一方で，3 番目の例は敵対的学習により適切な評価ができていない例である．この例のように応答文の文長が長い場合に，敵対的学習後の識別器は低いスコアを予測することが多かった．

そこで，識別器による予測スコアと入力される発話文及び応答文の文長の関係に着目した．図 5.1 に DotDisc における発話文または応答文の文長と識別器の予測スコアの関係性を示す．図 5.1a のように発話文の文長によらず低いスコアと高いスコアを予測できているが，図 5.1b のように応答文の文長が長くなるにつれ，高いスコアを予測できなくなり，低いスコアを予測する傾向が見られる．このことから，識別器の予測スコアは応答文の文長を特徴量として考慮していることが分かる．また，応答文の文長と人手評価スコアの関係を図 5.2 に示す．これによると，DotDisc

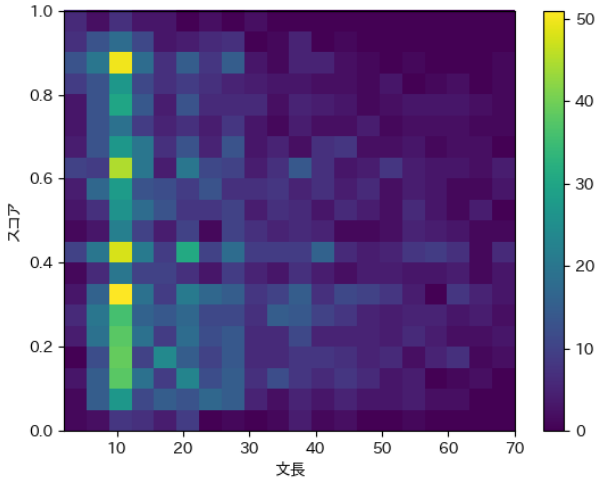


図 5.2: 日本語データにおける応答文の文長と人手評価スコアの関係

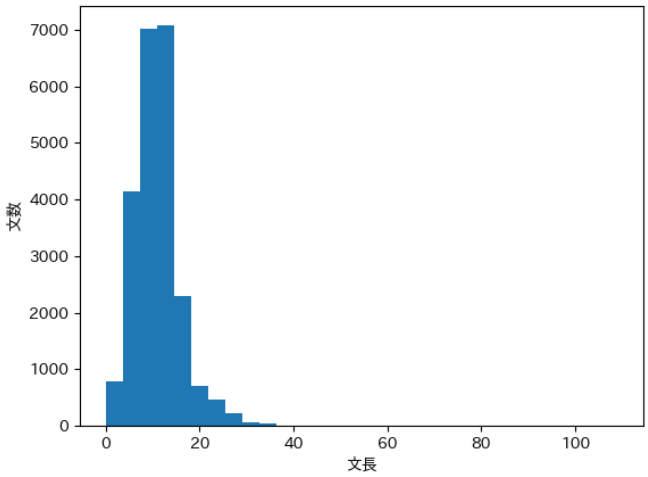
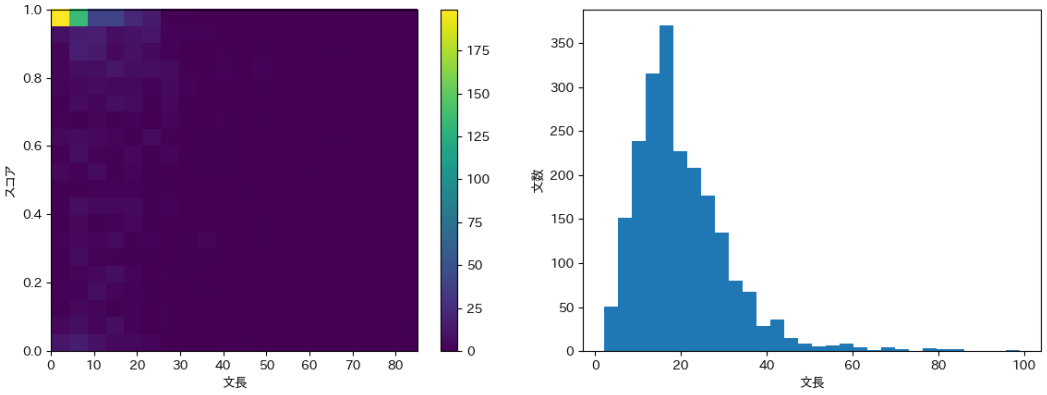


図 5.3: 識別器の日本語の学習データ中の応答文の文長と文数の関係

の予測スコアと同様に，人手評価スコアでも応答文の文長が長くなるにつれ，スコアが下がる傾向にあり，DotDisc は妥当な応答文の特徴を捉えるよう学習されていると言える．一方で，文長の短い応答文に対しては人手評価スコアが均等に分布しているのに対し，DotDisc の予測スコアは高いスコアを予測する傾向があり，短



(a) 英語の評価データにおける応答文の文長と DotDisc の予測スコアの関係

(b) 識別器の英語の学習データ中の応答文の文長と文数の関係

図 5.4: 英語データにおける文長とスコアの関係

い文長の応答文の評価にバイアスがかかることが分かる．ここで，学習データにおける発話文の文長とその文数についてのヒストグラムを図 5.3 に示す．これによると，識別器の学習データにおいて文長の短い応答文の文数が多く，文長が長い応答文はほとんど存在しないため，文長の短い応答文が正解となることが多く，応答文の評価にバイアスが発生したと考えられる．英語データにおける DotDisc の予測スコアの関係を図 5.4a に，識別器の学習データにおける文長とその文数の関係を図 5.4b に示す．英語においても日本語と同様に文長の短い応答文にバイアスがあることが言える．これらの結果から，学習時の入力文長によって，評価したい対話システムに識別器を適用できるかどうかが決まると考えられる．したがって，実際に識別器を用いて対話システムを評価する際には識別器の学習データの入力文長を調べて，適用先の対話システムが生成する応答文の文長との差が離れ過ぎていないことが望ましい．

また，図 5.1b と図 5.4a を比較すると英語データ学習したモデルは高いスコアを予測する傾向が顕著に見られる．英語の学習データと評価データには “no” や “i don’t know” などの汎用的な応答文が多く含まれる傾向にあり，評価時にはこれらの応答文に対し，不適切な場合でも高いスコアを予測してしまっていた．このことから，識別器の正解の学習データにおいて表現の出現頻度に偏りがある場合，識別

器はその表現に対して，高いスコアを出すようにバイアスがかかる．特に対話の応答文においては汎用的な表現の出現頻度が多くなる傾向があるため，多様な応答文を出力するような対話システムに対して高い評価を与えたい場合には前処理が重要になる．これは日本語に対しても同様のことが言えるが，表 4.2 において，単語の一致率を測るベースラインの値が特に低くなっていることから分かるように，英語では指示語や代名詞などが多く抽象的な表現が多くなる傾向があり，バイアスが発生しやすいと考えられるので，特に注意が必要である．

第 6 章　おわりに

近年，ニューラルネットワークを用いた研究が盛んになり，対話システムの研究も盛んになった．数々のモデルが対話システムの研究において提案されてきているが，対話システムを評価するための研究はほとんど行われておらず，自動評価はいまだ確立されていない．伝統的には BLEU スコアや Perplexity を用いて対話システムの性能を評価しているが，複数の応答文が正解になる可能性があり，人手評価との相関はほとんどない．deltaBLEU は，応答文候補とその人手評価を複数用意することにより，人手評価との相関が見られることを示したが，一般的な評価データに適用するためには，発話文に対して応答文候補を人手で作成する必要があり，実用的ではない．

本研究では評価時に正解応答文候補の作成やラベル付けされたデータを学習に用いることなく，対話システムの応答を自動で評価するための手法を提案した．本手法では入力に対して妥当な応答文を識別するような識別器を対話システムに対して敵対的に学習することで，システム応答を評価するために用いた．まず，対話システムの生成器としては Encoder Decoder モデルを用い，発話文に対して正解の応答文の生成確率を最大化するようにニューラルネットワークのパラメータを学習した．一方，識別器もニューラルネットワークを用いて構築され，発話文と応答文が入力された際に，入力の応答文が正解であるかどうかを識別できるように学習した．そして，この識別器に対して発話文と応答文の対を入力として与えた時に予測される確率をシステム応答のスコアとして利用することを提案した．

さらに，本研究では識別器の性能を向上させるために，双方のモデルのパラメータを動的に更新し，敵対的な学習を行うことで識別器の性能が向上することを示した．提案手法に対して対話破綻検出チャレンジの日本語と英語のデータセットを用いて実験を行った結果，どちらの言語においても Spearman の相関係数でベースラインよりも人手評価スコアの相関が高くなった．また，コーパスの分析を通して識別器を適用する際の設定について考察した．

提案手法を用いることで，評価時に正解応答文候補の作成やラベル付けされたデータを必要とせずに自動的に応答文の評価を行うことができるため，対話システムの評価における人手コストの削減が期待できる．本研究の提案手法が応答文の自

動評価の研究に役立ち，今後の対話研究全体の発展に対する一助となることを願っている．

発表リスト

1. 尾形朋哉, 叶内晨, 高谷智哉, 小町守. **キーワードに基づくニューラル文生成のためのリランキング.** 言語処理学会第 23 回年次大会. pp.679-682. March 15, 2017.
2. 尾形朋哉, 小町守, 高谷智哉. **修飾節付与による複文のニューラル生成.** 人工知能学会全国大会. 4Pin1-11. 4 pages. June 8, 2018.
3. Tomoya Ogata, Mamoru Komachi, Tomoya Takatani. **Divide and Generate: Neural Generation of Complex Sentences.** In arXiv e-prints, 1901.10196. 5 pages. January 30, 2019.

謝辞

研究に関して様々な相談に乗り，支えてくれた小町守先生に深く感謝します．また，先生には共同研究やインターンシップの機会を与えていただくなど，貴重な体験をすることができました．自分の成長に繋がる機会を作ってくださり，ありがとうございました．

叶内さんには研究について何も分からなかった時に，研究の仕方や論文の書き方などを直接指導していただき，深く感謝しています．先輩方や同期のみなさんには様々な場面で助けていただき，ありがとうございました．そして，山口先生と高間先生には副査を引き受けてくださり感謝しております．

参考文献

[1] A. Ram, R. Prasad, C. Khatri, A. Venkatesh, R. Gabriel, Q. Liu, J. Nunn, B. Hedayatnia, M. Cheng, A. Nagar, E. King, K. Bland, A. Wartick, Y. Pan, H. Song, S. Jayadevan, G. Hwang, and A. Pettigrue, "Conversational AI: The science behind the Alexa prize," arXiv:1801.03604, 2017.

[2] K. Papineni, S. Roukos, T. Ward, and W.-J. Zhu, "BLEU: a method for automatic evaluation of machine translation," ACL, pp.311–318, 2002.

[3] F. Jelinek, R.L. Mercer, L.R. Bahl, and J.K. Baker, "Perplexity – a measure of the difficulty of speech recognition tasks," JASA, vol.62, p.S63, 1977.

[4] M. Galley, C. Brockett, A. Sordoni, Y. Ji, M. Auli, C. Quirk, M. Mitchell, J. Gao, and B. Dolan, "deltaBLEU: A discriminative metric for generation tasks with intrinsically diverse targets," ACL, pp.445–450, 2015.

[5] 東中竜一郎，船越孝太郎，"Project next NLP 対話タスクにおける雑談対話データの収集と対話破綻アノテーション，" SIG-SLUD，vol.B4，no.02，pp.45–50，2014.

[6] 東中竜一郎，船越孝太郎，稲葉通将，荒瀬由紀，角森唯子，"対話破綻検出チャレンジ2，" SIG-SLUD，vol.B5，no.05，pp.64–69，2016.

[7] O. Vinyals and Q. Le, "A neural conversational model," ICML Deep Learning Workshop, 2015.

[8] A. Sordoni, M. Galley, M. Auli, C. Brockett, Y. Ji, M. Mitchell, J.-Y. Nie, J. Gao, and B. Dolan, "A neural network approach to context-sensitive generation of conversational responses," NAACL, pp.196–205, 2015.

[9] I.V. Serban, A. Sordoni, Y. Bengio, A.C. Courville, and J. Pineau, "Building end-to-end dialogue systems using generative hierarchical neural network models.," AAAI, pp.3776–3784, 2016.

[10] J. Li, W. Monroe, A. Ritter, D. Jurafsky, M. Galley, and J. Gao, "Deep reinforcement learning for dialogue generation," EMNLP, pp.1192–1202, 2016.

[11] C.-W. Liu, R. Lowe, I. Serban, M. Noseworthy, L. Charlin, and J. Pineau, "How not to evaluate your dialogue system: An empirical study of unsupervised evaluation metrics for dialogue response generation," EMNLP, pp.2122–2132, 2016.

[12] R. Lowe, M. Noseworthy, I. Serban, N. Gontier, Y. Bengio, and J. Pineau, "Towards an automatic Turing test: Learning to evaluate dialogue responses," ACL, pp.1116–1126, 2017.

[13] E. Bruni and R. Fernández, "Adversarial evaluation for open-domain dialogue generation," SIGDIAL, pp.284–288, 2017.

[14] J. Li, W. Monroe, T. Shi, S. Jean, A. Ritter, and D. Jurafsky, "Adversarial learning for neural dialogue generation," EMNLP, pp.2157–2169, 2017.

[15] 松村雪桜，小町　守，“敵対的生成ネットワークを用いた機械翻訳評価手法,” 言語処理学会 年次大会，pp.568–571，2018.
[16] M.-T. Luong, H. Pham, and C.D. Manning, “Effective approaches to attention-based neural machine translation.,” EMNLP, pp.1412–1421, 2015.
[17] S. Hochreiter and J. Schmidhuber, “Long short-term memory,” Neural computation, vol.9, no.8, pp.1735–1780, 1997.
[18] S. Zhang, E. Dinan, J. Urbanek, A. Szlam, D. Kiela, and J. Weston, “Personalizing dialogue agents: I have a dog, do you have pets too?,” ACL, pp.2204–2213, 2018.
[19] R. Sennrich, B. Haddow, and A. Birch, “Neural machine translation of rare words with subword units,” ACL, pp.1715–1725, 2016.
[20] B. Heinzerling and M. Strube, “BPEmb: Tokenization-free pre-trained subword embeddings in 275 languages,” LREC, pp.2989–2993, 2018.
[21] J. Pennington, R. Socher, and C.D. Manning, “GloVe: Global vectors for word representation.,” EMNLP, pp.1532–1543, 2014.
[22] D. Kingma and J. Ba, “Adam: A method for stochastic optimization,” ICLR, 2015.