対話システムの応答選択のための BERT を用いたユーザー発話からの体験情報抽出の検討

A Study of Extracting Experience Information from User Utterances Using BERT for Response Selection in Dialogue Systems

市川　菜月
Ichikawa　Natsuki

岡山大学 阿部研究室
Abe Laboratory, Okayama University

概要 ユーザが対話システムに対して感じる親密度を向上させるアプローチの一つに，対話システムがユーザについて知ろうとする応答を返す方法がある．そこで本研究では，BERT を用いてユーザー発話からの体験情報抽出を行い，この情報を用いてユーザの体験をより詳細に聞き出すような応答を選択する雑談対話システムを提案する．本報告では提案方式の有効性を確認するために，BERT の体験情報抽出精度や応答選択方法ごとの自然性やスロットの得やすさを評価した．

1　はじめに

近年，雑談対話システムの研究が盛んに行われている．雑談対話システムとは，ユーザと雑談を行うことを目的とした対話システムであり，使用例としては，独居高齢者の話し相手や，教育支援ロボットでの雑談などが想定される．このように，一人のユーザが一対一で長期的に使用する状況においては，対話システムは，飽きを感じさせない多様な発話と，親密感を与える魅力的な発話を両立させる必要がある．

ユーザの親密度を向上させるために，様々なアプローチが考案されている．例えば，ユーザの発話を促すための「掘り下げ質問」を応答として返す対話システム [1] や，ユーザの意見に同意する対話システム [2] などが挙げられる．中でも傾聴対話システムは，ユーザ自身の話を聞いてほしいという欲求を満たし，なおかつ，「掘り下げ質問」で得たユーザ情報をその後の対話に活用できると期待されている．

本研究では，「掘り下げ質問」を通じてユーザの体験情報を深掘りし，システムに対する親密度を向上させることを目的としている．傾聴対話における「掘り下げ質問」を返す雑談対話システムの一つに，スロットフィリング型がある．スロットフィリング型雑談対話システムは，ユーザーとの対話を通じて特定の情報（スロット）を収集し，その情報を基に適切な応答を生成するシステムである．あらかじめ得るスロット項目を用意しておき，まだ得られていないスロット項目を埋めるためにユーザに特定の質問をすることで，ユーザ情報を雑談対話の中で得ることができる．このスロットフィリング型の利点は，ユーザについて知りたい情報を的確に聞き出せる点である．一方で，ユーザに質問する内容はあらかじめ用意しておく必要があり，応答の多様性が低いという欠点もある．応答の多様性という観点では，近年の雑談対話システムでは大規模な対話データを深層学習モデルに学習させ，ユーザ発話に対して応答文を生成する方法が多く取られている．これらの対話システムは応答の多様性が高いという利点があるが，必ずしもユーザについて知ろうとする応答を返すとは限らない．

そこで本研究では，スロットフィリング型の方式と生成ベース型の方式を組み合わせた雑談対話システムを提案する．具体的には，既存の応答生成器を用いて複数の応答文を生成し，その中からユーザの体験情報を深掘りする応答文を選択する．応答文の選択には，ユーザの発話から抽出した体験情報を用いる．

2　提案システム

2.1　提案システムの概要

提案システムの概要図を図 1 に示す．

まず，ユーザ発話とユーザ発話以前の対話履歴を応答生成器に入力し，複数の文とそれに対する予測スコアを出力する．システム応答文を一文ずつイベント情報付与部に入力し，システム応答文ごとにスロット情報とスロット確率を付与する．予測スコア，スロット情報，スロット確率を付与したシステム応答文の中から，現段階までの対話で得られたユーザスロットを用いて取捨選択を行う．残ったシステム応答群の中から応答選択方法１～３を用いてシステム応答文を選択し，出力する．

イベント情報付与部の詳細を図 2 に示す．システム応答文と対話履歴を応答生成器に入力し，複数の予測ユーザ発話文とそれに対する予測スコアを出力する．予測ユーザ発話文とはシステム応答対してユーザが返す発話を予測した文である．予測スコアは対話モデルが生成した応答文に対する Perplexity の対数である．出力された予測ユーザ発話文をスロット抽出器に入力し，文ごとにスロットを抽出する．抽出したスロットを集計した情報をスロット情報とし，予測ユーザ発話文の予測スコア，システム応答文の予測スコア，抽出したスロットの３つを用いて算出する「システム応答文を選んだ際に特定のスロットが得られる確率」をスロット確率とする．

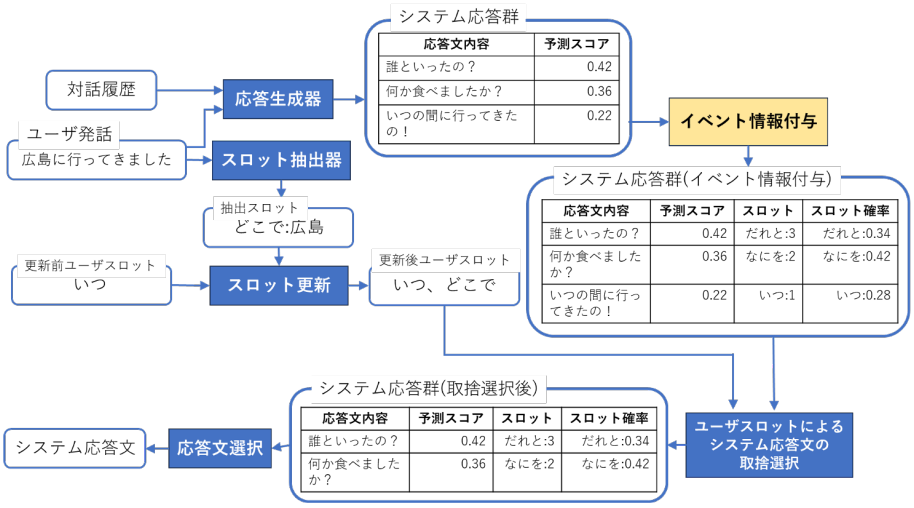

図 1: システム概要

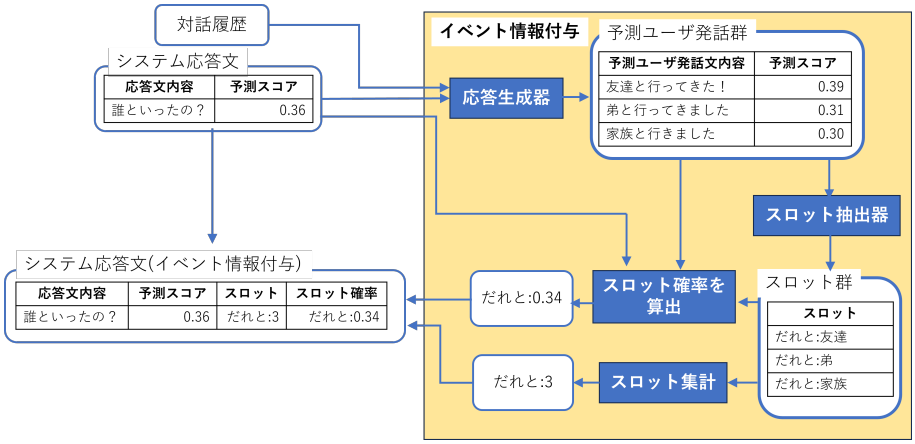

図 2: イベント情報付与部

2.2 応答生成器

本研究では応答生成器に事前学習済み日本語 Transformer Encoder-decoder 対話モデル [3] を JEmpatheticDialogues データセット [3] を用いてファインチューニングしたモデルを用いた．入力に対話履歴を入れることで一定区間の文脈を考慮した応答を生成することが可能となっている．

2.3 スロット抽出器

スロット抽出器では，ユーザ発話から「いつ」「どこで」「誰と」「何をした」の４つのスロットを抽出する．例として，「昨日友達と広島に旅行に行った」という文章が入力された時，スロット抽出器の出力は「いつ：昨日」「誰と：友達」「どこで：広島」「何をした：旅行」となる．このスロット抽出器はユーザの体験情報を含む文からのみスロットを抽出するように学習してある（学習の詳細は 3.1 で延べる）．例えば，「広島に行ってきた」「広島が好きです」「あなたは広島が好きですか？」という 3 つの文章があったとき，ユーザの体験情報が含まれているのは１つ目の文のみなので，同じ広島という場所の情報を含む文でも抽出されるのは最初の文からのみとなる．

2.4 応答選択方法

本研究ではユーザ発話および予測ユーザ発話から抽出したスロットを用いた応答選択方法を３つ提案する．応答選択方法１～３はいずれもすでに得られているユーザスロット以外のスロット情報を含むシステム応答文を選択する．

2.4.1 応答選択方法 1：応答の自然性を優先

システム応答群の中から最も予測スコアが高いシステム応答文を選択する．

2.4.2 応答選択方法 2：スロットの得やすさを優先

システム応答群の中から特定のスロット数が最も多いシステム応答文を選択する．

2.4.3 応答選択方法 3：応答の自然性とスロットの得やすさの両方を考慮

システム応答群の中から特定のスロットのスロット確率が最も高いシステム応答文を選択する．スロット

表 1: 使用データセット

	train	test
データセット（イベント文）	5000	1000
データセット（ペルソナ文）	2000	500
データセット（質問文）	5000	1000
データセット（その他の文）	2000	500

確率の算出式を式 (1) に示す.

$$P(x, S1) = \sum_{U2} P(x|S1, U2)P(U2)P(S1) \quad (1)$$

$P(S1)$，$P(U2)$ はそれぞれ出力した発話文の数 N 個の予測スコアの合計が 1 になるように softmax 関数で処理した後のシステム応答文 S1, 予測ユーザ発話文 U2 予測スコアである．また，$P(x|S1, U2)$ は S1 に対する予測ユーザ発話文の一つである U2 にスロット x が含まれている確率であり，スロット x が含まれていれば 1，含まれていなければ 0 とする．

3 評価実験

3.1 BERT によるスロット抽出精度評価

本研究ではスロット抽出器に事前学習済み BERT モデル [4] をラベル付けしたデータセットによってファインチューニングしたモデルを使用している．本節では使用データセットや学習条件，評価方法，結果について述べる．

3.1.1 使用データセット

ファインチューニングに用いたデータセットはテンプレートの文に単語を挿入して生成した．データセットは体験情報を含むイベント文，プロフィール情報を含むペルソナ文，質問文を生成し，イベント文の体験情報である単語に，「いつ」「どこで」「誰と」「何をしたか」のいずれかのラベルを付与し，ペルソナ文と質問文にはラベルを付けなかった．また，テンプレート文を用いて自動生成したデータセットのほかに，JPersonaChat データセット [3] から体験情報を含まないような文を 2500 文ピックアップし，体験情報のラベル付けをせずにその他の文としてデータセットに加えた．使用データセットの内訳を示す.

3.1.2 学習条件

事前学習済み BERT モデルをファインチューニングする際の学習条件を表 2 に示す

3.1.3 評価方法

BERT の精度評価には BERT によって予測されたラベルと正解ラベルとの F-measure を用いる．F-mesure

表 2: 学習条件

学習率	0.00002
最適化関数	Adam
学習回数	3

の算出式を式 2 に示す.

$$\begin{aligned} Precision &= \frac{num\_correct}{num\_predictions} \\ Recall &= \frac{num\_correct}{num\_predictions} \\ F-measure &= \frac{2 \times Precision \times Recall}{Precision + Recall} \end{aligned} \quad (2)$$

$num\_entities$ は正解ラベルの個数, $num\_predictions$ は予測されたラベル数，$num\_correct$ は予測されたラベルのうち正解だったラベルの個数を表す．また，$Precision$ は適合率，$Recall$ は再現率を表す.

3.1.4 実験結果

実験結果を表 3 に示す.

結果より，スロットごとの F-measure を見ると，「いつ」「誰と」のスロットに比べ，「どこで」「何を」のスロットの推定精度が低いことが分かる．これは，学習データを作成する際に用いたテンプレート文の内容や，挿入した単語の種類に関係するのではないかと考えられる.

3.2 対話シミュレーションによる応答選択方法の比較実験

提案システムの応答選択方法 1 ～ 3 を比較するために，日本語対話言語モデルがユーザの代わりとなって対話を行う対話シミュレーションを行い，対話中に得られたスロット数やシステム応答文の予測スコアから応答の自然性とスロットの得やすさを比較する.

3.2.1 日本語対話言語モデル

評価実験では日本語対話言語モデルの youri-7b-chat[5] にプロンプト文を入力し，ユーザの代わりとして提案システムと対話を行う.

プロンプト文の例を表 4 に示す．設定の文にあるユーザのペルソナ（プロフィール）の部分を 5 パターン用意した．システムの初発話は 4 パターン用意した．選択応答の内容はそれ以前の 2 文を提案システムに入力した際に返ってきた応答を入力する．シミュレーションではペルソナ部分とユーザの初発話の部分のほかに，対話モデルの出力する応答の多様性を調整する値である temperature を 0.7 0.9 まで 0.05 ずつ変化させる.

3.2.2 評価方法

3.2.1 節で述べたようにプロンプト文の内容やモデルの temperature を変えながら 8 発話の対話シミュレー

表 3: 学習条件

	いつ	どこで	誰と	何を	ALL
num_entities	232	628	257	560	1677
num_predictions	232	629	259	562	1682
num_correct	232	611	256	542	1641
Precisions	1.0000	0.9714	0.9884	0.9644	0.9756
Recall	1.0000	0.9729	0.9961	0.9679	0.9785
F-measure	1.0000	0.9722	0.9922	0.9661	0.9771

表 4: プロンプト文

設定: あなたは次のようなプロフィールを持つ人です.
私は北海道で生まれました．私は東京に住んで
みたいです．私は一軒家に住んでいます.
私は介護福祉士です．私は彼氏いない歴が長いです.
このプロフィールの人間としてユーザーと対話して
ください．また，なるべく対話は続けようと
心がけて対話してください．返答は 1 文までです.
ユーザー: お疲れ様です！最近何かありましたか？
システム: 昨日友達と旅行に行ってきました
ユーザー：(選択応答の内容)
システム：

表 5: 実験結果

	平均スロット種類数	平均予測スコア
方法 1	0.32	0.2096
方法 2	0.67	0.1623
方法 3	0.72	0.1968

ションを 100 回行う．100 回のシミュレーションを行い得られたスロット数の平均と選択した応答文の平均予測スコアから，応答選択方法同士の応答の自然性とスロットの得やすさを比較する.

3.2.3 実験結果

実験結果を表 5 に示す．結果より，応答の自然性を優先して選択する方法 1 の平均予測スコアが最も高くなっていることが分かる．一方で，対話中に得られたスロット数の平均が最も多いのは方法 3 であった.

4 まとめ

本報告では，BERT を用いたユーザ発話からの体験情報抽出と，抽出した体験情報と予測スコアをもちいた応答選択方法について検討した．BERT の精度評価実験の結果より，テストデータでは「どこで」「何を」のスロットが他のスロットよりも F-mesure が低いことが分かった．これは学習データを生成する際に使用したテンプレート文の内容が偏っていることが関係していると考えられる.

また，対話シミュレーションによる応答選択方法の比較実験では，方法 1 が平均予測スコアが最も高く，方法 3 が対話中に得られたスロット種類数の平均が最も多かった．応答の自然性を優先する方法 1 の平均予測スコアが最も高いのは妥当であるのに対し，スロットの得やすさを優先した方法 2 よりも方法 3 の方が得られたスロット数が多かったのは一考の余地がある．考えられる理由として，スロットの得やすさを優先して選択した応答よりも，スロットの得やすさと応答の自然性の両方を考慮して選択した応答の方が，ユーザにとって自己開示しやすかったのではないかということが挙げられる．しかし，本報告ではあくまでシミュレーション実験による結果に基づいているため，実際のユーザの対話システムに対する親密度の評価をするためには，人手評価を行い再度検討する必要がある.

そのため今後は BERT の推定精度を向上させることや，人手評価を行うために対話の評価基準を正確に定めることが課題となる.

参考文献

[1] 石田他, “傾聴対話システムのための発話を促す聞き手応答の生成” 人工知能学会研究会資料, Vol.77, Aug. 2016.

[2] 早瀬他, “好意の返報性を表出するエージェントがユーザの親密度に与える効果” 人工知能学会全国大会論文集, Vol.JSAI2018,pp. 2K205– 2K205, 2018.

[3] Hiroaki Sugiyama, *et al.*, “Empirical Analysis of Training Strategies of Transformer-based Japanese Chit-chat Systems,” ,2021.

[4] https://huggingface.co/tohoku-nlp/bert-base-japanese-v3

[5] https://huggingface.co/rinna/youri-7b-chat