解釈可能な内部表現を使用した タスク指向ニューラル対話システムの試作

村田　憲俊† 　酒井　哲也†

† 早稲田大学基幹理工学部情報理工学科　〒 169–8555 東京都新宿区大久保 3-4-1
E-mail: †muratacnttto@suou.waseda.jp, ††tetsuyasakai@acm.org

あらまし　古典的なタスク指向の対話システムは人手で作成された semantic frame に基づいており，ユーザーの発話を元にテンプレートや外部のデータベースの情報を利用して返答を作成していた．この観点において，ニューラルネットワークベースの End-to-End の対話モデルの出現には長所と短所が存在する．人手で返答の戦略を作成する必要はなくなったが，システム自体はブラックボックスになってしまった．これにより，ユーザーの発話に合わせて返答の戦略を切り替えることが困難になった．本研究において我々は、双方のアプローチの長所を享受するための試みを提案する．ニューラル対話モデルの内部表現を解釈可能なものにすることにより，Decoder の出力やテンプレートによる返答，外部のデータベースに基づいた情報の提示など，様々な戦略を選択することができるようになる．より具体的には，ユーザーの発話を人間が解釈可能な one-hot vector として内部表現に持つタスク指向の対話用 Sequence-to-Sequence モデルを作成した．提案したモデルのユーザー発話に対するクラスタリング能力の評価とシステム返答文の正当性の評価の結果，提案したモデルが従来のモデルに比べ著しい劣化なく，返答戦略の設計に有益な解釈可能な内部表現を獲得できることを示した．

キーワード　対話システム，解釈可能性, Seq2seq

1 はじめに

古典的な対話システムはユーザーのリクエストを semantic frame に変換している．これらの semantic frame の構造は人手で作成されることが多い．システムはユーザーの発話から抽出した semantic frame を元に，外部のデータベースや返答のテンプレートを活用して返答の生成する．これはシステム開発者が非常に制御しやすい反面，各々のシナリオに合わせた semantic frame の設計や抽出は非常にコストが高いという課題があった．

これに対し，近年研究が進んでいる End-to-End のニューラル対話モデルは明示的な返答戦略の設計を行うことなく，様々なドメインを含む対話システムの構築に成功しており [16]，設計コストの課題を解決しているように思われる．一方，End-to-End のニューラル対話モデルの欠点の 1 つは，システムが特定の返答をする理由を開発者が解釈することが困難という点である．ニューラル対話モデルは、ユーザー発話の解釈可能な内部表現を提供しないため，発話に応じて返答戦略を切り替えることが難しくなっている．モデルの学習をする際に，ユーザーの発話とシステムの返答だけでなく，外部リソースのラベル（例: 返答に必要なデータベース中のデータの id など）を提供することにより，外部リソースを利用可能なニューラル対話モデルを作るのは可能である [2]．しかし，そのようなシステムは新しい外部リソースを追加する度にラベルの付け直しや，モデルの再学習など維持するためのコストが非常に高くなる．

本研究において我々は、双方のアプローチの長所を享受するための試みを提案する．ニューラル対話モデルの内部表現を解釈可能なものにすることにより，Decoder の出力やテンプレートによる返答，外部のデータベースに基づいた情報の提示など，様々な戦略を設計することができるようになる．より具体的には，ユーザーの発話を人間が解釈可能な one-hot vector として内部表現として使用するタスク指向の対話用 Sequence-to-Sequence モデルを作成した．

本論文では初めに one-hot vector が解釈可能な内部表現として適している理由を説明する，次に，有益な内部表現を獲得するための提案モデルの学習手法を示す．その後，ユーザー発話を対話行為についてクラスタリングする実験を通して，他のクラスタリングアルゴリズムより優れていることを示す．これは，提案モデルの内部表現が効果的な返答戦略の切り替えに，実際に役立つ可能性があることを示唆する．また，システムの最終出力の比較を通して，内部表現の変更前後でのシステムの返答について *BLEU* [8] の差が大きくないことを示す．これにより，内部表現の変更に伴う情報の欠損が著しいものではないことを意味する．これらの実験の結果，提案したモデルが従来のモデルに比べ著しい劣化なく，返答戦略の設計に有益な解釈可能な内部表現を獲得できることを示した．

2 関連研究

我々の提案は VAE（Variational Autoencoder）の分野の研究と関連がある．VAE では発話を入力とし，入力をそのまま出力として再現することを目指す．一般的に中間層の次元数はベクトル化された発話の次元数より小さい．よって，再現に成功した場合，中間層のベクトルは発話のベクトルの次元圧縮

に成功していることになる．CVAE（Conditional Variational Autoencoder）[3] 等が一般的に発話の連続値ベクトルへの変換によく用いられている．Zhao らは [19]，これらのモデルをベースとし解釈可能な内部表現を獲得するために，Gumbel Softmax [6] と vector quantization を使用した．この結果，勾配の伝搬が困難であった発話の離散値ベクトルへの変換を可能とした．しかし，これらのモデルは発話の内部表現のみを獲得するため，実際の返答を生成するには内部表現を解釈し，返答を生成するためのモデルを別に作らなくてはいけない．

我々の提案は従来のアプローチと比較して次の点で異なる．(1) 内部表現とシステム返答の生成を同時に学習している．これは従来のアプローチでは内部表現の獲得とそれに対応する返答の生成をそれぞれ別モデルとして学習をしていたが，我々の提案ではこれらを同時に学習させる．その結果，複数のモデルを作成する必要がなく，実装と学習のコストが軽減される．(2) 内部表現として実際に表現可能なベクトルの集合を少なく制限している．結果，学習後に獲得された各内部表現を解釈し、それに応じて返答戦略を手動で設計することを現実的にしている．

3 提 案 手 法

この章では，Seq2seq（Sequence-to-Sequence）[13] が解釈可能な内部表現を持ち，返答戦略の設計に有益な内部表現を学習させるための手法の提案を行う．3.1 では内部表現が満たすべき条件を提案するとともに，Seq2seq が解釈可能な内部表現，one-hot ベクトルを中間層に使用できるようにするための手法の説明する．3.2 ではその条件の元，有益な内部表現と返答を得るために効果的な損失関数の提案，3.3 では性能向上のために効果的な入力データへの処理を説明する．3.4 では提案モデルを返答の検索モデルとして扱う拡張の提案を行う．ここでは，$\mathbf{x}$ をユーザの発話，$\mathbf{y}$ を返答，内部表現を $\mathbf{z}$ とする．

3.1 解釈可能な内部表現

通常，Seq2seq の内部表現は連続値で構成されるベクトルであり，人間には解釈が難しい．我々は以下の 2 つの機能を Seq2seq に導入することを提案する．

（1） 内部表現が人間に解釈可能

（2） 個々の内部表現が唯一つの解釈を持つ

これらの要件は，ユーザー発話の解釈に従って返答戦略を設計できるようにするためのものである．(1) は内部表現が取りうるベクトルの集合の数を比較的小さい数（例: $\leqq 100$）に制限する必要がある．あまりに多くの値を取りうる内部表現は現実的に管理することが困難であるためである．(2) では，解釈に基づいて特定の 1 つの返答戦略を割り当てることができることを意味している．これは，内部表現を解釈する際に曖昧さが存在する場合，それらを考慮した返答戦略を構築する必要があり，システムの実装と保守が複雑になるためである．

上記の考慮事項に基づいて，我々は内部表現として上記の条件を満たす one-hot vector を使用する．Encoder はユーザー発話の解釈を one-hot vector の形式で出力する．その後，Decoder は解釈を元に返答のための文を生成する．従って，提案モデルの Seq2seq のシステム全体は単純には以下のように表すことができる．

$$\hat{\mathbf{y}} = Decoder(\mathbf{z}) \tag{1}$$

$$\mathbf{z} = Encoder(\mathbf{x}) \tag{2}$$

$$\mathbf{z} = \{z_1, z_2, ..., z_K\} \tag{3}$$

$$z_i = \begin{cases} 1 & \text{iff } index\ of\ selected\ category \\ 0 & otherwise \end{cases} \tag{4}$$

ここで，K は内部表現の次元数である．また，one-hot の各次元を返答戦略の category と呼ぶこととする．我々の目標は学習後，各 category が学習済みの Decoder だけでなく人間にも解釈可能であり，どのようなユーザー発話が同じ category に割り当てられ，どのような返答が生成されるかを分析できるようにすることである．これによって，一部の category では Decoder の出力以外を使用する返答戦略の選択をすることができる．例えば，ハードコーディングされた返答や外部のデータベースへの接続を行い，得られた情報を元にしたテンプレートでの返答の使用である．

本研究では Encoder と Decoder に Transformer [15] ベースの元としたものを使用した．ここで使用する Encoder，Decoder に制限はなく，RNN（Recurrent Neural Network）[7] など任意のアーキテクチャを使用することができる．ユーザー発話に対する Encoder の出力を h_x としたとき，$\mathbf{z}$ についての事後確率は $p(\mathbf{z}|\mathbf{x}) = \text{Softmax}(W_E h_x + b_E)$ となる．この分布に対して Gumbel-Softmax を使用することにより，中間表現である one-hot ベクトルを得る．関連研究では $\mathbf{z}$ を $K \times M$ の行列で表現し，中間表現の取りうる表現を K^M にしていた [14] [19]．しかし，予備実験の結果，これらは取りうる表現の内の一部のみを使用していた．そのため，全ての表現を有効に活用するために我々の提案では $K \times 1$（i.e., K 次元ベクトル）とした．

3.2 対話の学習

一般的に対話モデルに使用される Seq2seq は以下の目的関数を最大化する．

$$\mathcal{L}(\theta) = p_{p(\mathbf{z}|\mathbf{x})}(\mathbf{y}|\mathbf{z}) \tag{5}$$

予備実験の結果，上記の式のみの利用は我々の目指すモデルにおいては不適切であることがわかった．内部表現 $\mathbf{z}$ を学習により適切に獲得することができないためである．この式には $\mathbf{z}$ について制限が存在しないため，category を一つでだけ使用する状態で学習が収束してしまった．しかしながら，我々の提案モデルでは内部表現の次元数を少ない数に制限しているため，理想的には全ての category が返答を生成するために有効に活用されることが望まれる．結果，我々の提案モデルでは以下の式を最大化する．

$$\mathcal{L}(\theta) = p_{p(\mathbf{z}|\mathbf{x})}(\mathbf{y}|\mathbf{z}) - \alpha(\mathrm{D_{KL}}(\bar{p}(\mathbf{z})||p(\mathbf{z}))) \tag{6}$$

α はスケール因子（i.e., ハイパーパラメータ）である．新たに導入した項について以下で説明する．

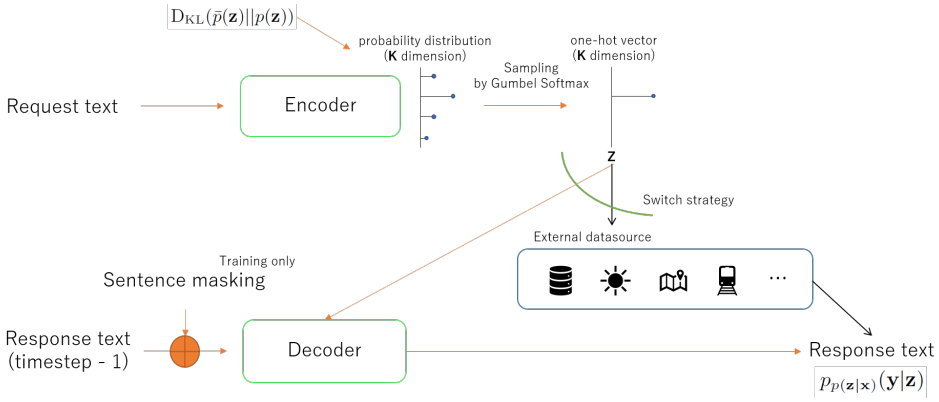

図 1　提案モデルの概略図

3.2.1　内部表現の正則化

以下の項は全ての category を Encoder が使用することを目的に導入する．

$$D_{KL}(\bar{p}(\mathbf{z})||p(\mathbf{z})) \tag{7}$$

この項がない場合，唯一つの category のみを使用するように学習が収束し，どのようなユーザーリクエストに対しても最もありふれた単一の返答をしてしまう．しかし，上記の式の計算に必要な実際の中間表現の確率分布を学習前に知ることはできない．そのため，提案手法では学習中は以下のように計算を行う．

$$p(\mathbf{z}) = \frac{1}{|\mathbf{B_x}|}\sum_{i}^{|\mathbf{B_x}|} p(\mathbf{z}|\mathbf{x}_i) \tag{8}$$

$$\mathbf{B_x} = \{\mathbf{x}_0, \mathbf{x}_1, \mathbf{x}_2, ...\} \tag{9}$$

$\mathbf{B_x}$ はミニバッチ学習をする時の一回分のユーザー発話の集合を意味する．$p(\mathbf{z})$ はミニバッチ内の各ユーザー発話 $\mathbf{x}_i$ を元に算出された内部表現の事後分布である $p(\mathbf{z}|\mathbf{x}_i)$ の平均として算出する．この項では $\bar{p}(\mathbf{z})$ に何らかの確率分布を仮定し，それにより各 category の使用頻度を制御する．本研究では $\bar{p}(\mathbf{z}) = \text{uniform}$（一様分布）とする．これによりミニバッチごとの確率分布の平均が uniform になるようになり，結果，各 category の利用頻度が均等になるように学習が行われる．実用上大きな問題にはならないが，この項は本来偏りがあるであろう確率分布を何らかの確率分布に制限してしまう．今回のクラスタリング実験の場合，この項は長期的には性能を悪化させてしまうため，学習初期段階のみ使用した．

3.3　教師データの返答文への Mask

上記の目的関数の変更とは別に，Encoder の性能向上のために学習時に返答文の入力の一部に mask をかけた．Seq2seq の学習には一般的に Teacher Forcing [5] と呼ばれる手法が利用される．Seq2seq ではまず Encoder へユーザー発話を入力する．その後，Encoder の出力を Decoder に入力し，Decoder は教師データの返答を再現できるように学習を行う．Teacher Forcing では学習の高速化や性能の向上のために Decoder に対して Encoder の出力に加えて，返答の 1 timestep 前の単語列を入力として加える．つまり，返答の N 単語目を予測する際に $p(y_N|\mathbf{z}, y_0, ..., y_{N-1})$ という条件で予測をしていることになる．言い換えればこの手法は N 単語目の予測の際に $N-1$ 単語目までが正しく予測できている前提で学習を行うということである．しかし，この手法は我々が目指すモデルにおいて一部不適切である．提案モデルでは Encoder の出力は非常に疎なベクトルとなっており，学習の初期段階では情報を持っていない事が多い．よって，正しく予想できているという前提の $N-1$ 単語目までの情報が強く優先されてしまい，Decoder が Encoder の出力をあまり考慮せずに返答を学習してしまう．これは，解釈可能な内部表現そのものの性能の劣化に繋がるため避けるべき状態である．

我々はこの問題に対処するために，我々は学習時に Teacher Forcing に使用される返答文の入力の一部に mask をかけた．これにより，Teacher Forcing による性能向上の恩恵を一部得ながら，Encoder の出力を Decoder が活用するようにした．具体的には教師データの返答の入力について一部の文全体を $\langle msk \rangle$（mask）に置き換えて学習を行った．$\langle msk \rangle$ は mask したことを表すために導入した特殊単語である．mask された返答の予測の際には使用できる情報が Encoder の出力のみとなるので，Encoder の出力を考慮した Decoder の作成を行うことができる．本実験ではこの処理をランダムにデータ全体の 7 割程度に対して行った．以上のことを踏まえ，提案モデルの概略図を図 3.1 に示す．

3.4　検索モデルとしての活用

提案モデルは Encoder の出力による有益な解釈可能な内部表現の獲得がなされている場合，学習用のデータセット中から適切な出力を選択する検索モデルとして活用することもできる．これは，解釈可能な内部表現を元に自動で擬似的に人間による返答の設計することになる．具体的な方法は以下のように示される．

$$\hat{\mathbf{y}} = \arg\max_{\hat{\mathbf{y}}} \left(\sum_{c \in C} Score(\hat{y_c}, \mathbf{y}_c) \right) \tag{10}$$

ここで，C は category の集合（i.e., 表現しうる全ての one-hot vector）であり，$\hat{\mathbf{y}}$ は検索モデルの出力として使用する返答．$\hat{y_c}$，$\mathbf{y}_c$ はそれぞれ学習用のデータセットに含まれるカテゴリごとの，ある返答と返答全体の集合である．$Score$ 関数は 2 つ

の文の一致度を表す任意の指標を扱うことができ，ある返答文とカテゴリに含まれる全ての返答文の一致度の平均を算出する関数である．つまり，検索モデルでは返答として各カテゴリ内で一番他の返答との一致度が高いものを採用する．本実験では *Score* 関数には *BLEU*（詳細は 4.3.2 にて記述）を使用した．

4 評価・実験

この章では提案モデルの内部表現の解釈可能性と内部表現の変更に伴う返答文生成への影響について実験を行う．4.1 では実験に使用したデータや実験での前処理等の設定，4.2 では使用したニューラルネットのアーキテクチャとハイパーパラメータ，4.3 では実験で使用する評価指標，4.4 では分析に使用する指標の説明をする．4.5 では提案モデルの評価を行う．まず，4.5.1 では提案モデルの内部表現の解釈可能性についてユーザー発話のクラスタリングを通して評価を行う．これは提案モデルの Encoder が，データセット中の複数のユーザー発話を同一の内部表現にマッピングしているため，クラスタリングのモジュールとして扱うことができるためである．評価の際には各発話ごとの対話行為（dialogue act）を使用する正解データとして利用する．次に，4.5.2 では提案モデルのユーザー発話に対する返答文の性能の評価を行う．これにより提案モデルで行われている内部表現の変更による，返答文の生成に関する性能への影響を確かめる．

4.1 実験設定

4.1.1 データセット

本研究では実験に MultiWOZ [2] を使用した．このデータセットはタスク指向の対話コーパスであり，旅行のサポートサービスを行うための 6 つの異なるドメインを含んでいる．また，各発話は 13 の対話行為 (“inform”, “request”, “select”, “recommend”, “not found”, “request booking info”, “offer booking”, “inform booked”, “decline booking”, “welcome”, “greet”, “bye”, “reqmore”) を 1 つ以上含んでいる．multi-turn の会話が 10,438 件含まれており，8,438 件の会話が train 用，validation 用と test 用に 1,000 件ずつに分割されている．

データセットに対して以下の前処理を行った．提案モデルは単独の発話に対して 1 つの category を割り当てるため，multi-turn の対話を分割し，個々の発話を独立の対話として扱った．また，クラスタリングの実験の際に，時間や固有名詞などの slot をより一般的なタグへの変換を行った (例: “18:30” から “$\langle time \rangle$”)．これは住所や電話番号といったユニークな値が多いデータによって語彙数が過剰に増え，ベースラインの性能の悪化を防ぐためである．

4.2 ベースモデル

提案モデルはニューラル言語モデルであり，Encoder と Decoder をモデルのアーキテクチャに含むモデルを想定している．本実験では Transformer [15] をベースとしたモデルで実験を行った．以下の設定は Transformer のモデルを構築する上での設定である．input と output の単語辞書は train のデータセットに含まれる文を元に作成し，各単語の embedding の次元数は 300 とした．この際，特定の事前学習の影響をなくすために pre-trained の word embedding は使用しなかった．positional encoding をした後に Encoder と Decoder への入力とし，Encoder block は 2 層，Decoder block は 1 層とした．また，それぞれの block 内の feed forward layer の次元数を 500，multi-head attention の head 数を 6，Encoder の dropout rate を 0.2 とした．

本実験で作成した提案モデルの内部表現に用いている one-hot vector の次元数 K はクラスタリングの実験において 50，返答の出力に関する実験においては 200 とした．

4.3 評価指標

本実験で使用した評価指標はクラスタリングでは *NMI* [17]，*homogeneity* [11]，*completeness* [11]，*BCubed* [1] である．*homogeneity* は各クラスタに単一のクラスのメンバーのみがどれだけ含まれているのかを表す指標である．*completeness* は特定のクラスのすべてのメンバーが同じクラスタにどれだけ含まれているかを表す指標である．この 2 つの指標は対応する指標であり，これらの指標を元に総合的なクラスタリングの評価指標として *NMI* が算出される．システムの返答文の作成では *BLEU* を評価指標とした．ここでは *BCubed* と *BLEU* の詳細について記述する．

4.3.1 *BCubed*

ユーザー発話のクラスタリングの評価指標の一つとして，*BCubed* を使用する．ここでは，e がユーザー発話，$L(e)$ と $C(e)$ がそれぞれ正解と予測されたクラスタである．また，(e, e') は $e' \neq e$ となる発話のペア全体である．この定義のもと *BCubed* は以下のようになる．

$$Correctness(e, e') = \begin{cases} 1 & \text{iff } L(e) = L(e') \ \ \& \ \ C(e) = C(e') \\ 0 & otherwise \end{cases}$$

$$PrecisionBCubed_e = \mathrm{Mean}_{e' \text{ s.t. } C(e) \cap C(e') \neq \emptyset}[Correctness(e, e')]$$

$$PrecisionBCubed = \mathrm{Mean}_e[PrecisionBCubed_e]$$

$$RecallBCubed_e = \mathrm{Mean}_{e' \text{ s.t. } L(e) \cap L(e') \neq \emptyset}[Correctness(e, e')]$$

$$RecallBCubed = \mathrm{Mean}_e[RecallBCubed_e]$$

$$F1BCubed = \mathrm{Mean}_e[\frac{2 \times PrecisionBCubed_e \times RecallBCubed_e}{PrecisionBCubed_e + RecallBCubed_e}]$$

4.3.2 *BLEU*

システムが生成した返答文の正当性を評価するために *BLEU* [8] [20] を使用した．定義は以下の式である．

$$BLEU = BP * PREC \tag{11}$$

$$BP = \exp(\min(0, 1 - \frac{SBML}{SYSL})) \tag{12}$$

$$SBML = \sum_{s} BML(s) \tag{13}$$

$$= \sum_{s} \arg\min_{len(s^*)} |len(s) - len(s^*)| \tag{14}$$

$$SYSL = \sum_{s} len(s) \tag{15}$$

$$len(s) = \text{文 } s \text{ の長さ} \tag{16}$$

$$PREC = \exp(\frac{1}{2} \sum_{N \in \{1,2\}} \ln Prec_N) \tag{17}$$

$$Prec_N = \frac{\sum_s \sum_{e \in gram_N(s)} Clip(c, s)}{\sum_s \sum_{e \in gram_N(s)} C(e, s)} \tag{18}$$

$$Clip(e, s) = \min(\max_{s^*} C(e, s^*), C(e, s)) \tag{19}$$

$$C(e, s) = s \text{ の中に含まれている単語 } e \text{ の個数} \tag{20}$$

今回の場合はそれぞれ，s はシステムの返答文，s^* はテストデータの返答文，$gram_N(s)$ は文 s から得られる N-gram の集合，e は文 s を形態素解析をした単語である．また，$N \in \{1, 2\}$ とした．すなわちユニグラムとバイグラムのみを考慮した．これは $BLEU$ が一般的に使用される翻訳のタスクと異なり，3 グラム，4 グラムの一致は正解となる返答が異なる趣旨で複数パターン考えられることの多い対話のタスクには求められないためである．

4.4 分析手法

ユーザー発話のクラスタリングの結果を分析をするために OW（offer weight）[10] を使用した．OW により各クラスタの特徴語の抽出を行い，提案手法と従来手法のクラスタリングの特性の違いを分析する．OW の定義を以下に示す．

$$RW(i) = \log[\frac{(r + 0.5)(N - n - R + r + 0.5)}{(n - r + 0.5)(R - r + 0.5)}] \tag{21}$$

$$OW(i) = r * RW(i) \tag{22}$$

n は単語 $t(i)$ が含まれる文の数，N はデータセットに含まれる文の数，r は単語 $t(i)$ が含まれる関連文の数，R は関連文の数である．ここでの関連文は同一クラスタに割り当てられた文である．

4.5 実験結果・考察

4.5.1 ユーザー発話のクラスタリング評価

ここでは提案モデルの内部表現の解釈可能性について，ユーザー発話のクラスタリングを通して評価を行う．提案モデルの Encoder は複数のユーザー発話を同一の内部表現にマッピングしている．これは提案モデルが対話における発話と返答に関して，多対一の関係にあることを前提としているためである．よって，提案モデルの Encoder のみに着目するとユーザー発話を限定された数の内部表現にマッピングするクラスタリングのモジュールとして扱うことができる．

実験で使用しているデータセットは各発話について対応する対話行為が設定されている．対話行為は実際にシステムが返答として行うべき行動と紐づくため，対話行為を正解データとしてクラスタリングの整合性を評価することは提案モデルが特定のユーザー発話を適切な返答戦略にマッピングできているかを定量的に図ることができる．

提案モデルのユーザー発話におけるクラスタリングの有用性を確認するために，各種一般的なクラスタリングアルゴリズムや短文用のクラスタリングアルゴリズムと性能の比較を行う．baseline として使用するモデルは k-means，matrix factorization ベースのモデルとして NMF，短文用のクラスタリング手法として近年の研究も含め比較的安定して高性能を示している GSDMM [18] [9] を使用した．

教師データにおける対話行為の種類数（13 種類）を鑑み，各モデルのクラスタサイズを 50 としてクラスタリングを行った．評価指標として *NMI*，*homogeneity*，*completeness*，*BCubed* を使用した．実験の結果を表 1 に示す．

実験の結果，我々が提案したモデルはクラスタリングの総合的な評価指標である *NMI* と *F1BCubed* において他のモデルより優れた性能を示した．より詳細には baseline と比較して特に *completeness*，*RecallBCubed* が優れている．*BCubed* について Tukey HSD 検定 [12] をしたところ最も優れたモデルと他のモデルでの差は統計的に有意であることが確認された．

クラスタリング結果の分析のために提案モデルと GSDMM のクラスタリング例と特徴語を示す．test のデータセットよりランダムにユーザー発話をサンプリングし，該当する発話と同一クラスタの発話例と OW によって抽出された特徴語の上位 20 件を表 2 に示す．ここでは，2 つのサンプリングされた発話について示した．1 つ目の例は何らかの施設の住所を求める発話である．提案モデルでは同一クラスタの例に住所やその他情報を要求している文が確認でき，特徴語においても電話番号や住所，料金など情報の要求対象に用いられる単語が見られる．GSDMM ではシステムの利用を終了する際の別れの挨拶等が確認でき，特徴語でも “thanks”，“goodbye” といった単語が見られる．別れの挨拶と伴って頻繁に使用される発話の冒頭の “thanks” にクラスタリングをする上で強く注目してしまい，住所の要求については無視されていると考えられる．このことから，提案モデルは発話内の要求を判断するために重要な単語に適切に注目できていると考えられる．2 つ目の例は情報の要求，特に宿泊先の種類に関する情報を求めている発話である．提案モデルでは宿泊先に限らずレストランなどの検索や推薦を求める発話が同一クラスタとしてまとめられている．特徴語からも様々な施設の情報に関することがクラスタに含まれていることが確認できる．GSDMM では宿泊場所に関連する情報を求める発話がまとめられており，特徴語についても宿泊場所の詳細に関連する単語が並んでいる．以上より，提案モデルでは対話行為に対応する施設の検索，推薦について宿泊場所等のドメインに縛られずクラスタリングできている．しかし，GSDMM のような宿泊場所といったドメインに厳密なクラスタリングは一つ前の例も含め行われていない．これが提案モデルにおける *completeness*，*RecallBCubed* の向上に繋がっている一方，*homogeneity*，*PrecisionBCubed* の課題点である

model	*homogeneity*	*completeness*	*NMI*	*BCubed*		
				Precision	*Recall*	*F*1
NMF	0.3045	0.1630	0.2124	0.3833	0.0822	0.1353
k-means	0.3699	0.2135	0.2708	0.4239	0.1177	0.1843
GSDMM	0.4143	0.2433	0.3066	0.4543*	0.1772	0.2549
proposed model	0.3286	0.3285	**0.3285**	0.3922	0.4140*	**0.4028***

表 1　ユーザー発話の対話行為についてのクラスタリング評価（対話行為数: 13）
∗ は α =0.01 にて他のモデルとの差が統計的に有意 (*BCubed* のみ Tukey HSD 検定を実施)

request	no , thanks . i just need to now the postcode .
proposed cluster example	can i get the postcode for that ? i also need to book a taxi to the ... i need to get the address please . can i please get the phone number , postcode and entrance fee ? i actually do not need reservations i just need the phone number ... could i please get the phone number for that ?
proposed keywords	‘phone’, ‘postcode’, ‘address’, ‘the’, ‘number’, ‘?’, ‘and’, ‘their’, ‘please’, ‘get’, ‘can’, ‘fee’, ‘entrance’, ‘what’, ‘give’, ‘me’, ‘could’, ‘,’, ‘just’, ‘sounds’
GSDMM cluster example	no , i think that s going to be all i needed . thanks . have a good day . whew , thanks , sorry for all of the confusion . i think that covers ... i think it might be a good time to end the conversation . i have ... that was all the questions i had , thanks very much for helping me . no , thanks . i just need to now the postcode .
GSDMM keywords	‘.’, ‘i’, ‘thanks’, ‘set’, ‘all’, ‘thank’, ‘not’, ‘no’, ‘goodbye’, ‘bye’, ‘this’, ‘you’, ‘else’, ‘your’, ‘everything’, ‘,’, ‘now’, ‘need’, ‘time’, ‘do’
request	i am just looking for information . what kind of hotel ? hotel or guest house ?
proposed cluster example	no hold off on booking for now . can you help me find ... hello , i am looking for a restaurant in [value_place] . ... great , can you also get me information or architecture in the area what do you recommend ? does it have a star rating of [value_count] ?
proposed keywords	‘called’, ‘?’, ‘[hotel_name]’, ‘am’, ‘looking’, ‘[restaurant_name]’, ‘a’, ‘information’, ‘restaurant’, ‘about’, ‘hotel’, ‘area’, ‘[attraction_name]’, ‘what’, ‘attraction’, ‘do’, ‘can’, ‘particular’, ‘me’, ‘find’
GSDMM cluster example	before we do that , what is the name of the guesthouse ? ... maybe . is either [value_count] a [value_count] star hotel ? ... i am interesting in info about [value_count] star hotel -s and ... do any of them have free parking ? do those both have [value_count] star rating -s and are [value_pricerange] ?
GSDMM keywords	‘free’, ‘?’, ‘parking’, ‘star’, ‘hotel’, ‘guesthouse’, ‘[value_count]’, ‘with’, ‘-s’, ‘wifi’, ‘do’, ‘have’, ‘any’, ‘a’, ‘are’, ‘[value_pricerange]’, ‘there’, ‘has’, ‘rating’, ‘stars’

表 2　各モデルの同一クラスタの文例と特徴語上位 20 件

と考えられる．

以上の結果から，提案モデルは対話行為に対する優れたクラスタリング性能が示された．よって，提案モデルが他のモデルより適切にユーザー発話を該当する返答戦略にマッピングする性能を有することを表し，このモデルの内部表現の解釈を使った返答戦略の活用の可能性を示唆している．

4.5.2　システムの返答文の評価

提案モデルのユーザー発話に対する返答文の性能の評価を行う．提案モデルは内部表現の解釈可能性を高めるために，従来の手法に比べ Encoder の出力を疎なベクトルとしている．これに伴い Encoder で本来得られるであろう情報が著しく損なわれることが危惧される．ここでは，我々の提案する変換を内部表現に適用した前後でのシステムの返答文の性能を比較することにより，変更の影響について確認を行う．返答文の性能の評価についてはデータセットで示されている返答を正解データとし，システムの返答を *BLEU* を評価指標として評価を行う．比較対象は 4.2 で示した Transformer ベースの対話モデルとこれに提案手法を適用したモデル，3.4 で示した検索モデルとし

て利用したモデルである．また，内部表現の次元数 K は 200 とした．実験の結果を表 3 に示す．

model	type	mean *BLEU*
baseline (Transformer)	generate	0.2148*
proposed model	generate	0.1736
	retrieval	0.2080

表 3　内部表現の変更前後での *BLEU* によるシステム返答文の評価（generate: Decoder の出力を利用，retrieval: 検索モデルでの出力を利用（詳細は 3.4 に記述））$*$ は α=0.01 にて提案モデルの retrieval との差が統計的に有意（Tukey HSD 検定を実施）

model pair	effect size
baseline vs proposed (generate)	0.315
baseline vs proposed (retrieval)	0.071
proposed (retrieval) vs proposed (generate)	0.244

表 4　各モデル対の mean *BLEU* についての効果量（generate: Decoder の出力を利用，retrieval: 検索モデルでの出力を利用）

type として “generate” で示されているものはシステムの返答として Decoder の出力をそのまま使用したものであり，“retrieval” で示されているものが検索モデルとして提案モデルを使用したものである．

実験の結果，Decoder の出力をそのまま使用したシステム同士の比較では性能の悪化が見られる．また，提案手法の検索モデルとしての利用についても baseline と比較した際，Tukey HSD 検定において差が統計的に有意であることが確認された．しかし，各モデル対について効果量 [12] を確認したところ（表 4），baseline と提案手法の検索モデルの差についての効果量は 0.071 程度であった．効果量が小さいとされる 0.2 より十分に小さいため [4]，統計的には有意であるが実質的な差は極めて小さいことが確認された．検索モデルとしての利用は Decoder を推論の際には使用をせず，人手での返答戦略の設計を擬似的に train のデータセットを用いて自動で行っていることを意味する．この手法において，内部表現の変更前と著しい性能の差がないことから，少なくとも提案モデルの Encoder は返答を作成する上で内部表現の変更前と同程度の性能であることが分かる．加えて，現状の提案モデルのアーキテクチャを用いてそのままの出力を返答として活用するには，Decoder が性能のボトルネックであることがわかる．

5 ま と め

本研究において我々は、人手で返答戦略を設計できる古典的な対話システムの手法と End-to-End で人手を介さずに返答戦略を構築できるニューラル対話システムの双方のアプローチの長所を享受するための試みを提案した．ニューラル対話モデルの内部表現を解釈可能なものにすることにより，Decoder の出力やテンプレートによる返答，外部のデータベースに基づいた情報の提示など，様々な戦略を選択することができるようになった．具体的には，ユーザーの発話を人間が解釈可能な one-hot vector として内部表現として使用するタスク指向の対話用 Sequence-to-Sequence モデルを作成した．また，提案したモデルのユーザー発話に対するクラスタリング能力の評価とシステム返答文についての実験を行った．これにより提案したモデルが従来のモデルに比べ著しい劣化なく，返答戦略の設計に有益な解釈可能な内部表現を獲得できることを示した．

文　献

[1] E. Amigó, J. Gonzalo, J. Artiles, and F. Verdejo. A comparison of extrinsic clustering evaluation metrics based on formal constraints. *Inf. Retr.*, 12(4):461–486, Aug. 2009.

[2] P. Budzianowski, T.-H. Wen, B.-H. Tseng, I. Casanueva, S. Ultes, O. Ramadan, and M. Gašić. MultiWOZ - a large-scale multi-domain wizard-of-Oz dataset for task-oriented dialogue modelling. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 5016–5026, Brussels, Belgium, Oct.-Nov. 2018. Association for Computational Linguistics.

[3] K. Cao and S. Clark. Latent Variable Dialogue Models and their Diversity. *arXiv:1702.05962*, 2017.

[4] J. Cohen. *Statistical Power Analysis for the Bahavioral Sciences (Second Edition)*. Psychology Press, 1988.

[5] A. Goyal, A. Lamb, Y. Zhang, S. Zhang, A. C. Courville, and Y. Bengio. Professor forcing: A new algorithm for training recurrent networks. In *NIPS*, 2016.

[6] E. Jang, S. Gu, and B. Poole. CATEGORICAL REPARAMETERIZATION WITH GUMBEL-SOFTMAX. *arXiv:1611.01144v5*, 2016.

[7] M. C. Mozer. *A Focused Backpropagation Algorithm for Temporal Pattern Recognition*, page 137–169. L. Erlbaum Associates Inc., USA, 1995.

[8] K. Papineni, S. Roukos, T. Ward, and W.-J. Zhu. Bleu: a method for automatic evaluation of machine translation. In *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics*, pages 311–318, Philadelphia, Pennsylvania, USA, July 2002. Association for Computational Linguistics.

[9] J. Qiang, Z. Qian, Y. Li, Y. Yuan, and X. Wu. Short text topic modeling techniques, applications, and performance: A survey. *CoRR*, abs/1904.07695, 2019.

[10] S. Robertson and K. S. Jones. Simple, proven approaches to text retrieval. Technical Report No. 356, Cambridge: Computer Laboratory, January 1994.

[11] A. Rosenberg and J. Hirschberg. V-measure: A conditional entropy-based external cluster evaluation measure. *Machine Learning Research*, pages 410–420, 2007.

[12] T. Sakai. *Laboratory Experiments in Information Retrieval: Sample Sizes, Effect Sizes, and Statistical Power*. Springer, 2018.

[13] I. Sutskever, O. Vinyals, and Q. V. Le. Sequence to Sequence Learning with Neural Networks. *NIPS*, 2014.

[14] M. E. Tiancheng Zhao, Kaige Xie. Rethinking Action Spaces for Reinforcement Learning in End-to-end Dialog Agents with Latent Variable Models. *arXiv:1902.08858*, 2019.

[15] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones,

A. N. Gomez, L. u. Kaiser, and I. Polosukhin. Attention is all you need. In I. Guyon, U. V. Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett, editors, *Advances in Neural Information Processing Systems 30*, pages 5998–6008. Curran Associates, Inc., 2017.

[16] P. Xu and Q. Hu. An End-to-end Approach for Handling Unknown Slot Values in Dialogue State Tracking. *arXiv:1805.01555v1*, 2018.

[17] N. Xuan Vinh, J. Epps, and J. Bailey. Information Theoretic Measures for Clusterings Comparison: Variants, Properties, Normalization and Correction for Chance. *Machine Learning Research*, 11:2837–2854, 2010.

[18] J. Yin and J. Wang. A Dirichlet Multinomial Mixture Model-based Approach for Short Text Clustering. *SIGKDD*, pages 233–242, 2014.

[19] T. Zhao, K. Lee, and M. Eskenazi. Unsupervised Discrete Sentence Representation Learning for Interpretable Neural Dialog Generation. *arXiv:1804.08069v1*, 2018.

[20] 酒井哲也. **情報アクセス評価方法論 検索エンジンの進歩のために**. コロナ社, 2015.