対話状態追跡における言語モデルのスキーマに基づく Hallucination の抑制

佐藤明智 [1] 南泰浩 [1]
[1] 電気通信大学大学院
akitomo-sato@uec.ac.jp minami.yasuhiro@is.uec.ac.jp

概要

近年では，ChatGPT をはじめとした大規模言語モデルを活用した対話システムが注目されている．これらのシステムは，日常会話や要約などの幅広い用途で利用されており，多種多様なユーザーのニーズに応えることができる．また，ユーザーの意図に応じて情報を取得したりデバイスを操作したりするタスク指向の対話システムに対するニーズも高まりつつある．しかし，大規模言語モデルには Hallucination の問題が存在しており，不正確なテキストの生成によって，誤ったアクションを引き起こすリスクがある．

これを解決するため，本研究ではタスク指向対話システムにおける Hallucination を抑制する新たな手法を提案する．具体的には，Schema Guided Dialogue データセットを用いて，スキーマに基づいて生成する語彙を制約するスキーマ名制約付きデコーディング (SNCD: **S**chema **N**ame **C**onstrained **D**ecoding) を提案する．この手法を用いることでスキーマ名を正確に生成し，エラーを防止することが期待できる．本稿では提案手法の有効性を検証し，Hallucination を抑制するための可能性について分析する．

1 関連研究

対話状態追跡 (DST:Dialogue State Tracking)

タスク指向対話システムとは，例えば，旅行予約や商品購入などのように，ユーザーが特定の目的やタスクを達成するために，自然言語でコンピューターと対話するシステムのことである．タスク指向対話システムは，ユーザーの要求を理解したり，必要な情報を問い合わせたり，得られた情報を処理をしたりして，最終的に目的を達成するためのサポートを行う．このようなシステムを実現するために，対話状態追跡 (DST:Dialogue State Tracking) というタスクが研究されている．

Jeffrey ら [1] は，複数のドメインに関する Schema Guided Dialogue (SGD)[2] 大規模対話データセットを用いる DST タスクの研究をしている．SGD データセットには図 1 の例に示すように，スキーマ情報，対話データ，対話状態情報が存在する．Jeffrey らは，スロットの説明，インテントの説明，対話履歴を入力とし，対話状態 (スロット名，スロット値，インテント名) を出力として事前学習済み言語モデル T5[3] をファインチューニングしている．

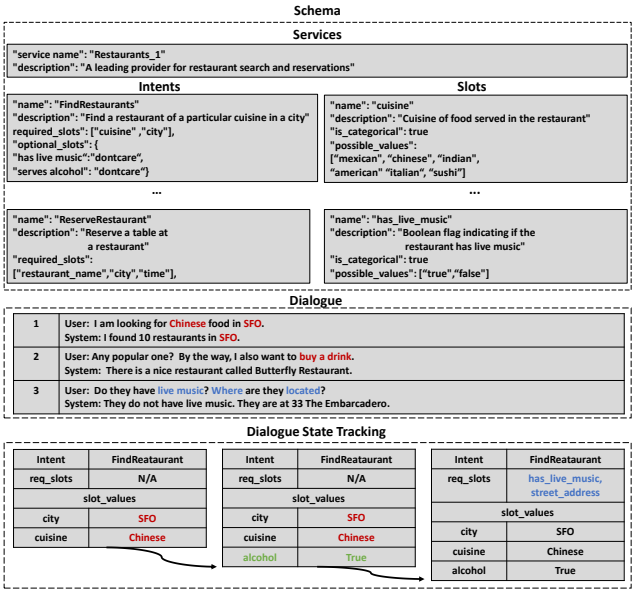

図 1 レストランに関するスキーマ，対話，対話状態の例

Jeffrey らの研究では，以下のことが示されている．

- インテントの説明，スロットの説明に番号を割り当て，インテント名，スロット名ではなく対応する番号を予測することで学習データに存在しないドメインでの精度が向上
- 対話状態を一括で予測可能

制約付きデコーディング

佐良らは知識応答生成モデルにおいて，誤った知識を生成しないようにエンティティ名制約付きデコーディング (ENCD: **E**ntity **N**ame **C**onstrained **D**ecoding) を提案した [4]．これは，知識グラフを外部知識として用いる応答生成モデルが，検索してきた知識をより正確に利用するよう促す手法である．これにより，間違ったエンティティ名の出力を防いでいる．

Jeffrey らの研究の予測精度には依然改善の余地がある．これは，予測対象を番号にしたことで，単語の意味的特徴を捉えられていないという点である．そこで，本研究では予測対象を番号ではなく，スロット名，スロット値，インテント名とし，さらに佐良らによる制約付きデコーディングを応用し，誤った名称を生成しないように，スキーマ名制約付きデコーディング (SNCD: **S**chema **N**ame **C**onstrained **D**ecoding) を提案する．

2 提案手法

章 1 で述べた制御手法で，タスク指向対話における言語モデルによる Hallucination を抑制する．本手法では学習時と推論時で処理が異なる．

2.1 学習時

図 2 に学習の方法を示す．入力は SGD データセットのスロット名，スロットの説明，スロット候補値，インテント名，インテントの説明，対話履歴を連結した文字列とし，出力はスロット名，スロット値，インテント名を連結した文字列とする．スロット名，スロット値，インテント名の間には [SLOT], [SEP], [INTENT] タグを挿入する．

2.2 推論時

図 3 に推論時のデコーディング制御機構を示す．推論時は学習時の構成にデコーディング制御モジュールを追加する．モデルが語彙を生成する場合，単語を生成する確率分布を制約し，該当するスキーマに含まれる語彙のみを生成する．図 4 に推論時のデコーディング制御の具体例を示す．始めは語彙確率を制約せずに生成する．出力の先頭の [SLOT] タグが生成された後は，スロット名を生成するため，参照するスキーマ内の全スロット名に含まれる語彙のみを生成する．スロット名生成後に「:」トークンが出現した後は具体的なスロット値を生成する．ここで，直前に生成したスロット名でスキーマを参照し，“is_categorical”が true の場合は “possible_values”に含まれる語彙のみを生成する．“is_categorical”が false の場合は，単語を生成する確率分布を制約せずに生成する．[SEP] タグが生成された後は上記と同様にスロット名，スロット値を生成する．[INTENT] タグが生成された後は，インテント名を生成するため，スキーマ内の全インテント名に含まれる語彙のみを生成する．<\s>トークンが生成された場合は生成を終了する．

3 データ作成

本稿では 1 つの対話中に複数のドメインが存在する SGD データセット (all-domain) を用いて，図 2 の形式で入出力データを作成した．各データ数を表 1 に示す．

表 1 データ数

	number of data
train	16142
dev	2482
test	4201

4 実験内容

本稿では Jeffrey らの実験と同じ t5-v1_1-base(248M parameters)[1)]，t5-v1_1-large(770M parameters)[2)] を使用した．章 3 で述べたデータでファインチューニングし，dev データでの loss が最小となるステップ数のモデルを選択した．その後，Greedy Search を用いて推論を行い，評価した．以下に評価対象手法を示す．

- D3ST：Jeffrey らの手法
- SNCD：提案手法
- SNCD(w/o constrained)：SNCD の予測時に制約を設けない手法
- SNCD-add-tag2vocab：提案手法＋タグをトークンとして語彙に追加する手法
- SNCD-add-tag2vocab(w/o constrained)：SNCD-add-tag2vocab の予測時に制約を設けない手法

ここで，実験を行う前に提案手法ではタグ生成が上手くいかないことが分かった．そのため，

1） https://huggingface.co/google/t5-v1_1-base
2） https://huggingface.co/google/t5-v1_1-large

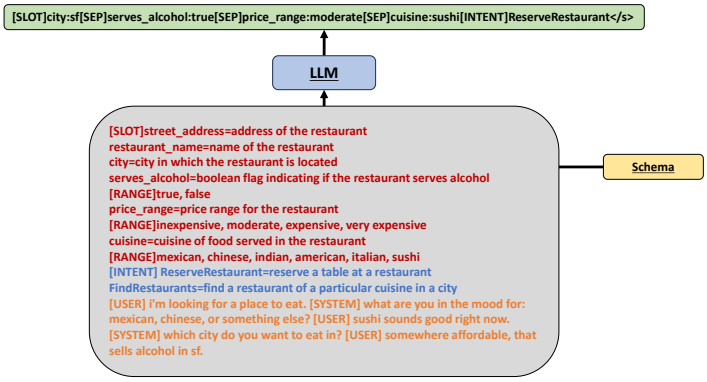

図 2　学習入出力フォーマットの例：赤のテキストはスロット記述，青のテキストはインテント記述，黄色のテキストは会話内容

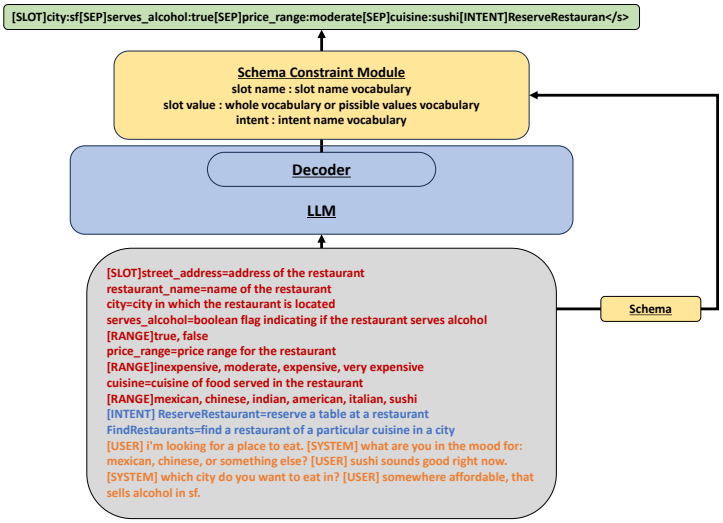

図 3　推論時のデコーディング制御機構

SNCD-add-tag2vocab では，タグの出力ミスを防ぐことを期待し，タグを語彙に追加して学習，評価した．

また，評価項目を以下に示す．

- JGA(Joint Goal Accuracy)：対話状態を完全一致で予測できた割合
- Slot Accuracy：正しくスロット名とスロット値を予測できた割合
- Intent Accuracy：正しくインテント名を予測できた割合
- Output Accuracy：正しい形式で出力できた割合 (タグの配置，タグ名が崩れずに出力できたか)

5　結果と考察

評価結果を表 2 に示す．結果より，Jeffrey らの手法である D3ST が全ての評価項目で高い評価となった．また，提案手法 (SNCD, SNCD-add-tag2vocab) では，単語を生成する確率分布を制約したほうが Output Accuracy 以外の項目では高い評価となった．これより，Schema Constraint Module の有効性が分かった．SNCD-add-tag2vocab ではタグの出力ミスを防ぐためにタグを語彙に追加したが，Output Accuracy は他の手法と比べて低い評価となった．これは新たに語彙を追加したため，生成モデルが扱うべき語彙の数が増え，生成モデルが選択肢の中から正しい単語を選び出す難しさが増したためだと考えられる．全体として，提案手法では先行研究を上回る結果は得られなかった．先行研究では数字を予測しているのに対し，提案手法では単語を予測しおり，無数の組み合わせが存在する．この予測トークン数の差が原因だと考えられる．また，本稿で使用した Greedy Search が結果に影響を与えていると考える．実際に，誤ったトークン予測の約 92%では，正解トークンが上位 4 つ以内に含まれていることが示されている [5]．そのため，今後は Beam Search の導入と評価を行う．

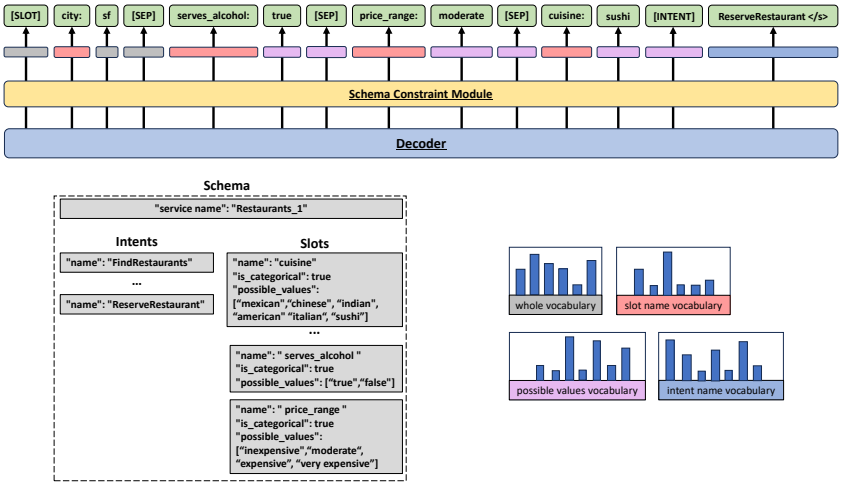

図 4　推論時のデコーディング制御例

表 2　評価結果

	JGA	Slot Accuracy	Intent Accuracy	Output Accuracy
D3ST(t5-base)	0.5544	0.8096	0.8651	0.9952
D3ST(t5-large)	0.5925	0.8372	0.9129	0.9949
SNCD(t5-base)	0.4025	0.6622	0.7972	0.9648
SNCD(t5-base, w/o constrained)	0.3559	0.5665	0.7940	0.9976
SNCD(t5-large)	0.4140	0.6309	0.8559	0.9233
SNCD(t5-large, w/o constrained)	0.3770	0.6039	0.8983	0.9910
SNCD-add-tag2vocab(t5-base)	0.3957	0.6301	0.8480	0.9193
SNCD-add-tag2vocab(t5-base, w/o constrained)	0.3490	0.5557	0.8213	0.9357
SNCD-add-tag2vocab(t5-large)	0.4031	0.6485	0.8507	0.9065
SNCD-add-tag2vocab(t5-large, w/o constrained)	0.3548	0.5621	0.8371	0.9391

6　おわりに

本稿ではタスク指向対話における言語モデルの Hallucination 抑制手法である SNCD を提案し，評価を行った．その結果，先行研究を上回る結果は得られなかったが，提案手法の有効性が確認できた．本手法を用いれば ChatGPT のような汎用的な言語モデルでも，追加の学習を必要とせず，より正確なテキストを生成することが可能となる．これにより，誤ったアクションを引き起こす可能性が低くなると期待される．

謝辞

本研究は，電気通信大学人工知能先端研究センター (AIX) の計算機を利用して実施したものです．

参考文献

[1] Jeffrey Zhao, Raghav Gupta, Yuan Cao, Dian Yu, Mingqiu Wang, Harrison Lee, Abhinav Rastogi, Izhak Shafran, and Yonghui Wu. Description-driven task-oriented dialog modeling, 2022.

[2] Abhinav Rastogi, Xiaoxue Zang, Srinivas Sunkara, Raghav Gupta, and Pranav Khaitan. Towards scalable multi-domain conversational agents: The schema-guided dialogue dataset, 2020.

[3] Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J. Liu. Exploring the limits of transfer learning with a unified text-to-text transformer, 2020.

[4] 佐良和孝 滝口哲也有木康雄. 知識グラフに基づく応答文生成におけるエンティティ名 制約付きデコーディング, 2022.

[5] Seungpil Won, Heeyoung Kwak, Joongbo Shin, Janghoon Han, and Kyomin Jung. BREAK: Breaking the dialogue state tracking barrier with beam search and re-ranking. pp. 2832–2846, Toronto, Canada, July 2023. Association for Computational Linguistics.