テキスト穴埋めによる知識ベース対話システムの構築

薛 強　滝口 哲也　有木 康雄
[1] 神戸大学大学院 システム情報学研究科
{xueqiang,takigu,ariki}@stu.kobe-u.ac.jp

概要

近年，外部知識を用いる知識ベース対話システムでは，生成する応答文の一貫性と外部知識の利用能力を向上させるために，会話に関連する外部知識と対話履歴を連結して深層学習モデルに入力する方法が提案されている．しかし，知識ベース対話システムは検索された外部知識を利用することなく，応答文を生成してしまうという問題がある．本研究では，テキスト穴埋めタスクを用いることにより，知識ベース対話システムが常に外部知識を利用することができる方法を提案する．実験により，テキスト穴埋め手法を用いた知識ベース対話システムは，正確性評価指標において最高値を示した．

1　はじめに

自然な応答を生成することができる雑談対話システムの構築は，自然言語処理分野における挑戦的な研究領域である．近年，Microsoft の DialoGPT [1] や，Google の Meena [2] など，人間同士の対話データを大量に収集して深層学習を行う生成ベース対話システムが知られている．しかし，生成ベース対話システムの対話相手となる人間が，「今は何時ですか」と尋ねた場合，生成ベースシステムは学習データに含まれていた古い情報をもとに会話を展開することになる．こういった最新の事実に基づかない応答文を生成する「幻想問題」が報告されている [3]．「幻想問題」の改善を目指し，適切な外部知識を検索できる検索ベース対話システムと，生成ベース対話システムを統合した対話システムとして，外部知識を利用した知識ベース対話システムの研究が近年注目を集めている．

一方，多くの知識ベース対話システム [4, 5] では，学習段階において，検索された外部知識と対話履歴を連結した形で深層学習モデルに入力し，目標応答文を生成するように学習が行われている．しかし，推論段階において，検索された外部知識を入力しているにもかかわらず，これを無視して，入力した対話履歴のみに基づいて応答文を生成してしまうという問題が報告されている [6]．そこで，本研究では，知識ベース対話システムの推論段階において，検索された外部知識が応答文に含まれるように，テキスト穴埋め手法を用いた知識ベース対話システムを提案する．

本稿では，知識ベース対話システムに基づいて，テキスト穴埋め手法を用いた深層学習モデルの学習段階と推論段階を述べた後，構築した知識ベース対話システムの実験と評価について報告する．

2　関連研究

本節では本研究で用いる知識ベース対話システムのベースラインと，テキスト穴埋め手法について述べる．

2.1　知識ベース対話システム

近年，知識ベース対話システムが生成した応答文に関して，一貫性と外部知識の利用能力を向上させるために，対話に関連する外部知識と対話履歴を統合した会話背景を GRU [7] や GPT-2 [8] などの深層学習モデルに入力する．

しかし，深層学習モデルは最大入力長があるため，会話の進行により，徐々に増える会話背景も制限される必要がある．会話背景に必要な入力空間を節約するために，Galetzka らは異なる知識のエンティティと対応する関係を連結することにより，検索された知識グラフをより簡潔にエンコードできるエンコード手法を提案した [4]．本研究では，以上述べたエンコード手法を用いる知識ベース対話システムをベースラインとする．

2.2　テキスト穴埋め

テキスト穴埋めタスクとは，空白スペースを含む語句、文、または段落で構成されているテキストにおいて，空白スペースに欠落している語句を予測す

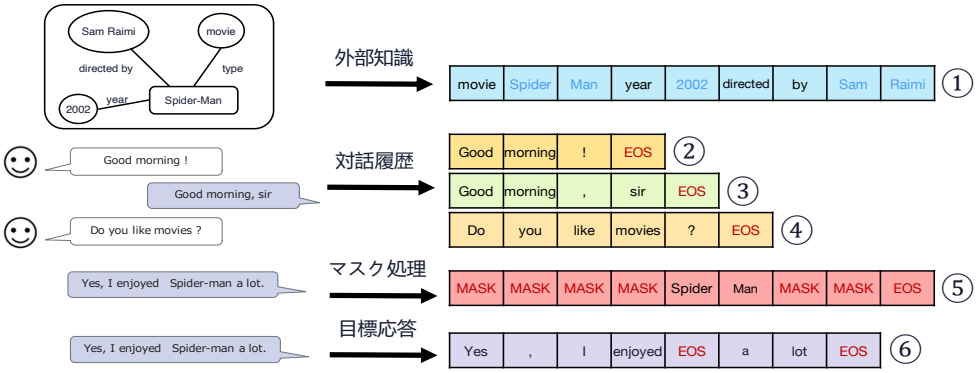

図 1　外部知識と対話データのエンコード．エンコードされた各種類の単語系列を異なる色で示す．

るタスクである．

Donahue らは，任意位置にある複数の空白スペースを含むテキスト入力に対して，欠落している目標テキストを空白スペースの位置に連結して，深層学習モデルの出力とするテキスト穴埋め手法を提案した [9]．このテキスト穴埋め手法を用いた深層学習モデルは，客観評価と主観評価で最高値に達している．本研究では，上に述べたテキスト穴埋め手法を知識ベース対話システムに応用する．

3　知識ベース対話システムの構築

本節では深層学習モデルの学習段階と推論段階において，テキスト穴埋め手法の応用により，提案する知識ベース対話システムの構築について述べる．

3.1　学習段階

本研究では，深層学習モデルとして，DialoGPT モデルを用いる．そのため，深層学習モデルの学習段階では，DialoGPT の学習タスクと Donahue らのエンコード手法 [4] を参照し，外部知識と対話データを以下のようにエンコードする．

- **外部知識のエンコード**（図 1・①）：検索された外部知識は，各知識のエンティティと関係を連結して知識系列とする．次に異なる知識系列をランダムに連結して単語系列①に変換する．
- **対話履歴のエンコード**（図 1・②③④）：対話履歴内の各発話文をトークンの列からなる単語系列②③④に変換する．変換された各単語系列の末尾に停止トークン<EOS>を追加する．
- **目標応答文のマスク処理**（図 1・⑤）：まず，目標応答文をトークンの列からなる単語系列⑤に変換する．次に，変換された単語系列⑤の長さを L とすると，整数 X と整数 Y（$1 < X < Y < L$）をランダムに選択する．X 番から Y 番までの単語系列を保留し（図 1 では，$X = 5, Y = 6$），それ以外の単語系列をマスクトークン<MASK>に入れ替える．最後に，変換された単語系列⑤の末尾に停止トークン<EOS>を追加する．
- **目標応答文のエンコード**（図 1・⑥）：まず，目標応答文のマスク処理において，マスクトークンに入れ替えられた二つの単語系列の末尾に停止トークン<EOS>を追加する．次に，二つの単語系列を順番に連結して単語系列⑥に変換する．

以上のようにエンコードされた単語系列を①〜⑥の順番に連結して，深層学習モデルの入力とする．深層学習モデルの学習タスクは，目標単語系列⑥の生成確率を最大化する．

3.2　推論段階

深層学習モデルの推論段階では，外部知識と対話履歴のエンコード，及び推論段階の出力を以下のように行う．

- **外部知識と対話履歴のエンコード**（図 1・①②③④）：推論段階の入力データに対して，3.1 節の学習段階と同じように単語系列①②③④に変換する．
- **知識のマスク処理**（図 2・⑤）：まず，検索された外部知識に関して，異なる知識のエンティティをランダムに一つ選択して，単語系列⑤に変換する．次に，整数 X と整数 Y（$0 < X, Y < 10$）をランダムに選択する．単語系列⑤の左方向と右方向に，X 個と Y 個のマスクトークン<MASK>を追加する．最後に，単語系列⑤の末尾に停止トークン<EOS>を追加する．

- **穴埋め生成**（図 2・⑥）：以上のようにエンコードされた単語系列を①〜⑤の順番に連結して，深層学習モデルに入力する．深層学習モデルはデコード戦略により，逐次的に単語系列⑥を生成する（穴埋め生成）．二つ目の停止トークン<EOS>を生成した時に，穴埋め生成を停止させる．
- **出力**（図 2・⑦）：停止トークン<EOS>により，単語系列⑥を二つ単語系列に分割し，単語系列⑤の左部分と右部分のマスクトークン<MASK>に入れ替えて単語系列⑦に変換する．単語系列⑦を推論段階の出力とする．

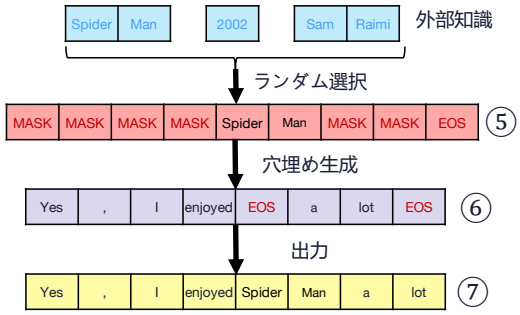

図 2　推論段階における外部知識のエンコードと出力．

4　実験と評価

実験では，知識なし対話システム（生成ベース対話システム），ベースライン（知識ベース対話システム），及びテキスト穴埋め手法を用いた知識ベース対話システム（提案対話システム）を実験対象とする．本節では，深層学習モデルの学習段階と推論段階で用いるデータセット，パラメータの設定，及び推論段階の実験評価について述べる．

4.1　データセット

OpenDialKG [10] は，本と映画についての推薦対話が含まれている雑談対話データセットである．話者は常に知識グラフに接続しているエンティティを含む発話を行い，知識グラフの事実に基づいた推薦対話を行う．また，知識グラフとして複数ドメインからなる Freebase の知識グラフを利用している．本研究では，OpenDialKG の各対話に対して，検索された関連外部知識と OpenDialKG を統合したデータセット [4] を利用する．

4.2　実験設定

本研究の深層学習モデルとして DialoGPT モデルを使用した．GPT-2 モデルをベースとした DialoGPT モデルは，Reddit から抽出した対話セッションの大規模なコーパスで学習され，自動評価と人間評価の両方で最先端の性能を達成している．表 1 に深層学習モデルのパラメータを示す．

表 1　深層学習モデルのパラメータ

モデル	DialoGPT
総パラメータ数	124M
最適化アルゴリズム	AdamW
最大対話履歴長	3 文
デコード戦略	Greedy
epochs	10
Batch Size	4
Learning Rate	6.0e-5

4.3　実験評価

実験評価では，応答文の正解性，多様性の二つの角度から応答文の質を評価する．多様性の評価指標として，応答文に含まれる単語数の平均を表す Avg Len，応答文に含まれる n-gram の種類数を表す DIST-n [11] を用いる．正解性の評価指標として，応答文と正解文の類似度を表す BLEU-n [12]，NIST-n [13] を用いる．ここで，NIST-n は BLEU-n をベースとした評価指標であるが，BLEU とは異なり，n-gram に対して情報量の大きさを用いて重み付けを行うため，機能語よりも内容語の一致が重視される．

表 2 に，各対話システムによって生成された応答文の評価結果を示す．表より，提案した対話システムは，正解性評価指標のスコアで全て最高値に達した．これより，テキスト穴埋めタスクの有効性が確認できる．一方，提案手法は，多様性評価指標である Avg Len スコアで最高値に達したが，多様性評価指標である DIST-1 スコアで最低値になった．提案手法による応答文の多様性の向上は，今後の研究課題である．

表 3 に，各対話システムによって生成された応答文の例を示す．表より，提案手法は検索された外部知識を正確に利用し，自然な応答文を生成できることが確認できる．一方，知識ベース対話システムは，外部知識が利用できていないが，対話履歴 1 に

表 2　各対話システムによって生成された応答文の評価結果

	多様性			正解性			
Dialogue System	Avg Len	DIST-1	DIST-2	BLEU-2	BLEU-4	NIST-2	NIST-4
生成ベース対話システム	10.86	**11.93**	**36.79**	15.74	8.71	1.39	1.43
知識ベース対話システム	10.45	10.77	31.84	17.77	10.47	1.62	**1.69**
提案対話システム	**13.11**	9.09	32.18	**18.79**	**10.64**	**1.64**	**1.69**

表 3　各対話システムによって生成された応答文の例．検索された外部知識のエンティティを青色で示す．

対話履歴 1	ユーザ A	what about tonto fistfight?
	ユーザ B	it was written by sherman alexie
	ユーザ A	was he a poet?
外部知識 1		poet: sherman alexie film producer: sherman alexie written by: sherman alexie
生成ベース対話システム		yes he was a poet
知識ベース対話システム		yes, he was a poet
提案対話システム		yes, he was a poet. the sherman alexie wrote the poems and drawings of a rose.
対話履歴 2	ユーザ A	they just got lebron james so that is a big benefit.
	ユーザ B	yes i was so happy that he signed. he is one of the greatest, right?
	ユーザ A	i'd say so. plus the lakes already have a good record, several nba finals championships under their belt.
外部知識 2		team: utah jazz
生成ベース対話システム		i am sure the fans would love it. since lebron james is your favorite player, you must like him?
知識ベース対話システム		that's awesome! i think he's a great player too.
提案対話システム		i agree. i like the utah jazz. do you know who won that year?

対して，自然な応答文を生成するには，外部知識 1 が必要ではないことが考えられる．外部知識の必要性を判断できるモジュールの開発は今後の研究課題である．

5　おわりに

本研究では，知識ベース対話システムは検索された外部知識を利用せずに，応答文を生成するという問題を改善することを目的として，テキスト穴埋め手法を用いた知識ベース対話システムを提案した．提案した対話システムは常に外部知識を利用することができる．実験より，提案した対話システムは正解性評価指標において最高値を得た．

参考文献

[1] Yizhe Zhang, Siqi Sun, Michel Galley, Yen-Chun Chen, Chris Brockett, Xiang Gao, Jianfeng Gao, Jingjing Liu and Bill Dolan. DialoGPT: Large-Scale Generative Pre-training for Conversational Response Generation, arXiv:1911.00536 (2019)

[2] Daniel Adiwardana, Minh-Thang Luong, David R. So, Jamie Hall, Noah Fiedel, Romal Thoppilan, Zi Yang, Apoorv Kulshreshtha, Gaurav Nemade, Yifeng Lu and Quoc V. Le. Towards a Human-like Open-Domain Chatbot (2020)

[3] Mojtaba Komeili, Kurt Shuster and Jason Weston. : Internet-Augmented Dialogue Generation. arXiv:2107.07566. (2021)

[4] Fabian Galetzka, Jewgeni Rose, David Schlangen, Jens Lehmann Space Efficient Context Encoding for Non-Task-Oriented Dialogue Generation with Graph Attention Transformer Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers), 7028–7041 (2021)

[5] Emily Dinan, Stephen Roller, Kurt Shuster, Angela Fan, Michael Auli, and Jason Weston. Of Wikipedia: Knowledge-powered conversational agents. 7th International Conference on Learning Representations, ICLR 2019, pages 1 − 18.(2019)

[6] Jason Weston, Emily Dinan and Alexander H. Miller. : Retrieve and Refine: Improved Sequence Generation Models For Dialogue, 2018; arXiv:1808.04776. (2018)

[7] Kyunghyun Cho, Bart Van Merrienboer, Caglar Gul- ¨ cehre, Dzmitry Bahdanau, Fethi Bougares, Holger Schwenk, and Yoshua Bengio. Learning phrase representations using RNN encoder-decoder for statistical machine translation. EMNLP 2014 - 2014 Conference on Empirical Methods in Natural Language Processing, Proceedings of the Conference, pages 1724 − 1734 (2014)

[8] Alec Radford, Jeffrey Wu, Rewon Child, David Luan, Dario Amodei, and Ilya Sutskever. Language models are unsupervised multitask learners. OpenAI Blog, 1(8):9.(2019)

[9] Chris Donahue, Mina Lee and Percy Liang. : Enabling Language Models to Fill in the Blanks, arXiv:2005.05339 (2020)

[10] Seungwhan Moon, Pararth Shah, Anuj Kumar,and Rajen Subba. OpenDialKG: Explainable conversational reasoning with attention-based walks over knowledge graphs. In Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics, pages 845 − 854, Florence, Italy,(2019)

[11] Li, J., Galley, M., Brockett, C., Gao, J., and Dolan, B. : A Diversity-Promoting Objective Function for Neural Conversation Models, Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, pp. 110 − 119 (2016)

[12] Kishore Papineni*et al.*,: Bleu: a Method for Automatic Evaluation of Machine Translation, Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics, p311-318 (2002)

[13] George Doddington. Automatic evaluation of machine translation quality using n-gram cooccurrence statistics. In Proceedings of HLT (2002)