生成・分類言語モデルに基づく対話システムの構築

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

概要

生成言語モデルとは，テキストデータから単語単位で学習し，単語生成機能を持つ言語モデルである．近年，生成言語モデルに基づく対話システムが，人間らしい会話を模擬する上で優れた性能を示している．しかし，生成言語モデルは生成した単語を構成する文章が適切かどうかを判定することができないため，繰り返しや間違いを含む文章を生成する傾向がある．そこで，本研究では，単語生成と文章分類両方の機能を持つ生成・分類言語モデルを提案する．自動評価では，提案する生成・分類言語モデルに基づく対話システムにより，ベースライン対話システムに比べ良質な応答が生成できることを示した．

1　はじめに

近年，Microsoft の DialoGPT [1] や，Google の Meena [2] など，Transfomer ベースの言語モデルを用いて，人間同士の対話データを応答生成のタスクで学習し，学習した言語モデルを用いて応答を生成するという生成言語モデルに基づく対話システム（以下，生成対話システムと呼ぶ）が主流となっている．このような対話システムは入力した対話履歴に基づいて，流暢な応答文を生成できる利点はあるが，「猫が好き．猫が好き」と言った繰り返される応答，または「分からない」と言った一般化された応答を頻繁に出力してしまい，情報性が低い対話になる傾向があり，問題となっている．

本研究では，上に述べたような応答を生成しないことを目的とし，複数応答を生成した後に応答の適切性を判定できる生成・分類言語モデルに基づく対話システム（以下，生成・分類対話システムと呼ぶ）を提案する．図 1 に提案する対話システムの応答生成の流れを示す．従来手法との比較実験により，提案する対話システムは，多様性と正解性評価指標において最高値を示した．

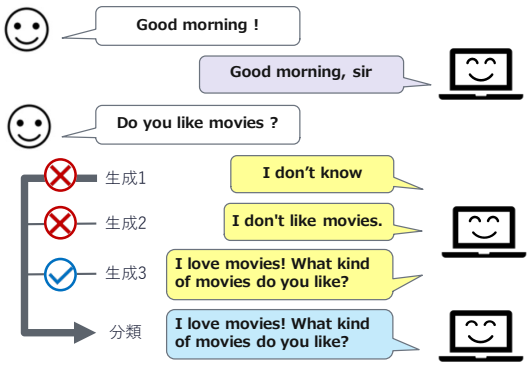

図 1　生成・分類言語モデルに基づく対話システムの応答生成の流れ．

本稿では，まず生成言語モデルと生成・分類言語モデルについて述べる．次に生成対話システムと，提案する生成・分類対話システムについて述べる．最後に提案する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

本章では，ベースラインで用いる生成言語モデルと本研究で用いる生成・分類言語モデルについて述べる．

2.1　生成言語モデル

Transformer ベースの生成言語モデルには，主に Encoder-Decoder と Decoder の二種類の構造がある．一つ目の Encoder-Decoder 構造とは，Encoder 部に対話履歴を入力し，Decoder 部において逐次的な単語選択により応答文を生成する構造である．二つ目の Decoder 構造とは，Decoder 部に対話履歴を入力し，対話履歴の尾部に接続して単語選択により応答文を生成する構造である．

近年の研究では，応答文の多様性を向上させ，話題の切り替えを可能とするために，<entity, relation, entity >形式で表現された対話に関連する構造化知識

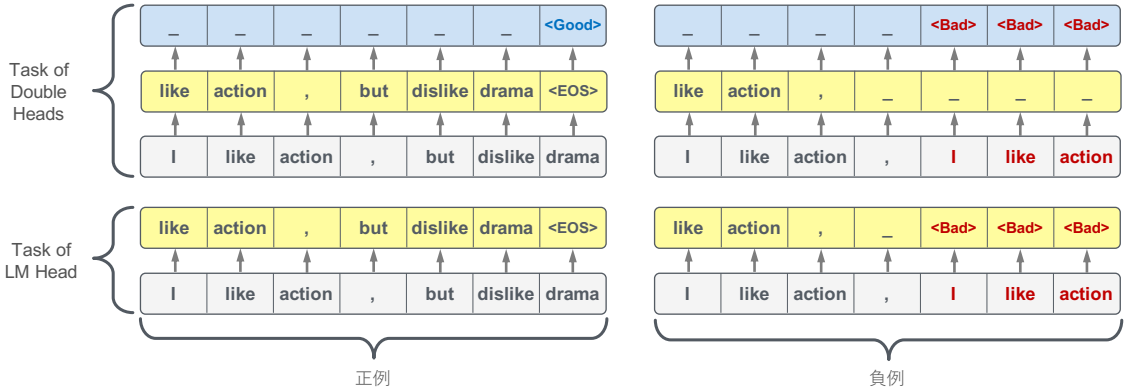

図 2 生成・分類言語モデルの学習タスク．白色はソース，黄色部分は生成器のターゲット，青色部分は分類器のターゲット．上部分は Double Heads モデルのタスク，下部分は LM Head モデルのタスク．

（外部知識）と対話履歴を統合して対話背景とし，これを GPT-2 [3] のような生成言語モデルに入力して応答を生成するという知識ベース生成対話システム [4] が提案されている．本研究では，以上に述べた生成対話システムをベースラインとする．

2.2 生成・分類言語モデル

Encoder 構造を持つ BERT モデルでは，大量のテキストデータから文単位の学習を行うために，分類トークンの位置からモデル出力を一つの分類器（Class Head）に入力し，2 つの文章が連続しているかどうかを判定するという次文予測（Next Sentence Prediction）タスクを用いている [5]．Decoder 構造を持つ GPT-2 モデルでは，大量対話データから文単位の学習を行うために，次文予測方法と同じように，停止トークンの位置からモデル出力を一つの分類器に入力し、応答が正解応答かどうかを判定するという応答判定タスクを用いている [6]．本研究では，複数の応答に対する判定を学習するために，複数分類器を用いた応答判定タスクを提案する．

Arora 等らは，生成言語モデルの生成器（LM Head）と同じ数を持つ分類器 (Class Head) を各位置に追加し，推論時に生成器が辞書にある各単語の確率を生成すると同時に，分類器が辞書にある各単語が適切かどうかを判定するという Double Heads モデルを提案した [7]．本研究では，単語単位ではなく，文単位で判定する Double Heads モデルを提案するものであり，このモデルを生成・分類言語モデルと呼ぶ．

3 生成対話システム

学習データでは，対話履歴を $(x_1, \cdots, x_M)$，目標応答を $(x_{M+1}, \cdots, x_N)$ とすると，学習データの負の対数尤度は以下のように計算される．

$$L_{\mathrm{LM}} = -\sum_{t=M+1}^{N} \log P\left(x_t \mid x_1, \cdots, x_{t-1}\right) \qquad (1)$$

生成言語モデルの学習タスクでは，学習データの負の対数尤度を最小化する．生成対話システムは学習した生成言語モデルを用いて応答を生成する．

4 生成・分類対話システム

本節では，提案する二種類の生成・分類対話システム (Double Heads: 4.2 節と LM Heads: 4.3 節）について，学習データ，学習段階と推論段階に関して述べる．

4.1 学習データ

生成・分類対話システムが適切な応答を判定できるために，学習データに対する正例応答（目標応答）を用いて負例応答を作る．正例応答と負例応答の例を図 2 に示す．一つの学習データで，生成タスクと分類タスクの学習を同時に行うために，負例応答は，正例応答の冒頭からランダムの長さで抽出した部分 $(x_{M+1}, \cdots, x_R, M+1 < R < N)$ と，負例部分 $(x_{R+1}, \cdots, x_S, R+1 < S)$ を組み合わせて構成する．繰り返し応答の生成を防ぐために，抽出した正例応答部分のコピーを一番目の負例部分とする．間違えた応答の生成を防ぐために，正例応答以外の応答の任意部分を二番目の負例部分とする．

4.2 Double Heads モデル

Double Heads モデルの分類器は，生成言語モデルの出力を <Good>（良い応答）と <Bad>（悪い応答）に分類する．図 3 の左部分に，分類器の構造を示す．同じ学習データに対して Double Heads モデルの

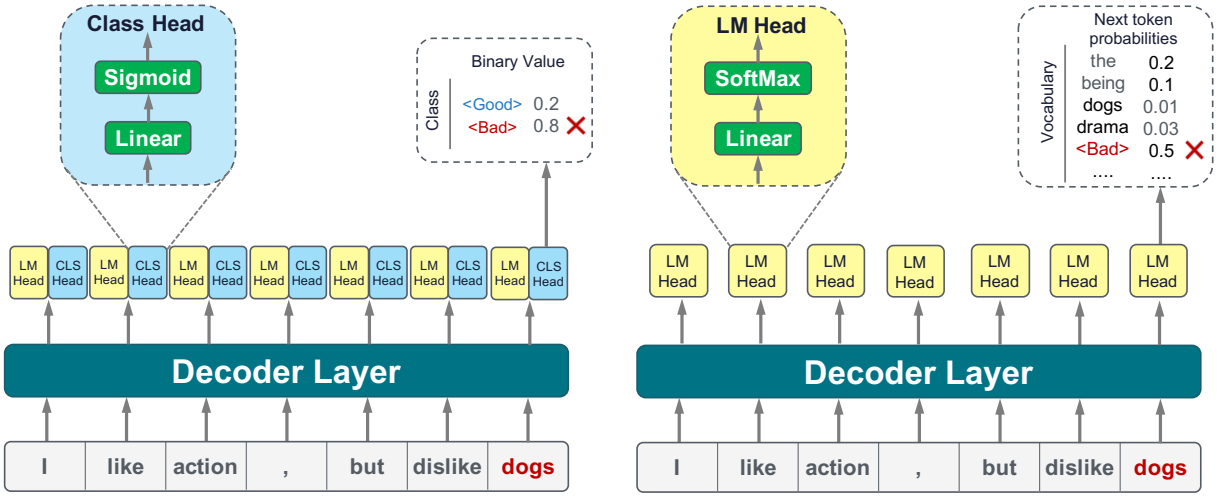

図 3　提案する二種類の生成・分類言語モデルの構造．左は Double Heads モデル，右は LM Heads モデル．

生成器と分類器は同時に学習を行う．

- **正例学習データ：**生成器の学習タスクでは，学習データの負の対数尤度を式 1 で計算する．分類器の学習タスクでは，学習データの負の対数尤度を以下の式で計算する．

$$L_{\text{class}} = -\log P\left(<Good> \mid x_1, \cdots, x_N\right)$$

- **負例学習データ：**生成器の学習タスクでは，学習データの負の対数尤度を以下の式で計算する．

$$L_{\text{LM}} = -\sum_{t=M+1}^{R} \log P\left(x_t \mid x_1, \cdots, x_{t-1}\right)$$

分類器の学習タスクでは，学習データの負の対数尤度を以下の式で計算する．

$$L_{\text{class}} = -\sum_{t=R+1}^{S} \log P\left(<Bad> \mid x_1, \cdots, x_t\right)$$

モデルの共同学習 Loss を以下の式で計算する．

$$L_{\text{train}} = L_{\text{LM}} + \gamma L_{\text{class}}$$

ここで，γ はハイパラメータである．

表 1　ハイパーパラメータ

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
Epochs	10
Beam Size	5
Learning Rate	6.0e-5
γ	1.0

4.3　LM Heads モデル

LM Heads モデルは分類器の代わりに，<Bad> をトークンとして辞書に登録し，生成器を用いて悪い応答を判定する．辞書にある <Bad> 以外のトークンは良い応答と判定できるため，<Good>を登録する必要がない．図 3 の右部分に，生成器の構造を示す．

- **正例学習データ：**生成器の学習タスクでは，学習データの負の対数尤度を式 1 で計算する．
- **負例学習データ：**生成器の学習タスクでは，学習データの負の対数尤度を以下の式で計算する．

$$L_{\text{LM}} = -\sum_{t=M+1}^{R} \log P\left(x_t \mid x_1, \cdots, x_{t-1}\right) - \sum_{t=R+1}^{S} \log P\left(<Bad> \mid x_1, \cdots, x_t\right)$$

4.4　推論段階

生成・分類対話システムは，まず生成器を用いて，Beam Search によりビーム幅サイズの応答を複数生成する．次に生成された応答を分類器に入力して<Bad>のスコアを計算し，最小値を持つ応答を出力する．

5　実験と評価

実験では，生成対話システム（3 章），Double Heads 構造を持つ生成・分類対話システム 1（4.2 節），LM Head 構造を持つ生成・分類対話システム 2（4.3 節）を実験対象とする．本章では，言語モデルの学習段階と推論段階で用いるデータセット，ハイパーパラ

表 2　各応答生成手法によって生成された応答文の評価結果.

対話システム	DIST-1	DIST-2	BLEU-1	BLEU-2	NIST-2	NIST-4	Avg Len
生成対話システム	7.95	22.72	16.75	10.67	1.62	1.70	9.36
生成・分類対話システム 1	8.26	23.14	**18.85**	**12.22**	**1.99**	**2.08**	10.53
生成・分類対話システム 2	**8.68**	**25.13**	18.39	11.98	1.93	2.02	10.07

表 3　各応答生成手法による対話例．赤字は応答中にある問題単語

入力 1	User A	Could you recommend a film directed by Martin McDonagh?
	User B	Most definitely! He wrote Seven Psychopaths, In Bruges, and Six Shooter. Have you seen any of those?
	User A	No I have not. Who was in Six Shooter?
生成対話システム		Seven Psychopaths stars Michael Caine and Michael Caine. It's a crime fiction.
生成・分類対話システム 1		Seven Psychopaths stars Robert Duvall and Michael Pena. Do you like those actors?
生成・分類対話システム 2		Two of the most well-known stars in the film are Will Ferrell and Sam Rockwell.
入力 2	User A	I like Bob Peck. Can you tell me some interesting facts on him?
	User B	Bob Peck is an actor, some of his movies are Surviving Picasso, Smilla's Sense of Snow, Slipstream, Lord of the Flies
生成対話システム		What genre is Slipperystream?
生成・分類対話システム 1		What genre is Surviving Picasso?
生成・分類対話システム 2		Can you tell me more about Lord of the Flies?

メーの設定，及び推論段階の実験評価について述べる.

5.1　データセット

OpenDialKG [8] 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．話者は常に構造化知識中のエンティティを含む発話を行い，関連知識に基づいた推薦対話が行われている．学習データは 11041 文, 負例学習データは 5520 文である. 本研究では，ターゲット応答文に埋め込む知識を用いて実験を行う. 表 1 に言語モデルのハイパラメータの設定を示す.

5.2　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 [9] を用いる．正解性の評価指標としては，応答文と正解文の類似度を表す BLEU-n [10], NIST-n [11] を用いる.

表 2 に，各応答生成手法によって生成された応答文の評価結果を示す．表より，提案した生成・分類対話システム 1 は正解性評価全指標のスコアで最高値に達した．提案した生成・分類対話システム 2 は，多様性評価全指標のスコアで最高値に達した．これより，提案した対話システムの有効性が確認できる.

表 3 に，各応答生成手法によって生成された応答文の例を示す．表の生成例 1 と生成例 2 より，生成・分類対話システムは繰り返し応答と間違え応答を判定できたため，誤った単語「Michael Caine」と「Slipperystream」の生成を防ぐことができている.

6　おわりに

生成対話システムを単語単位だけではなく，文単位で応答の適切性を学習できるようにするため，Double Heads 構造と LM Head 構造を持つ二種類の生成・分類対話システムを提案した．従来手法との比較実験により，提案する対話システムは，多様性と正解性評価指標において最高値を示した.

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 preprint 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, et al. Towards a human-like open-domain chatbot. **arXiv**

preprint arXiv:2001.09977, 2020.

[3] Alec Radford, Jeffrey Wu, Rewon Child, David Luan, Dario Amodei, Ilya Sutskever, et al. Language models are unsupervised multitask learners. **OpenAI blog**, Vol. 1, No. 8, p. 9, 2019.

[4] Shuman Liu, Hongshen Chen, Zhaochun Ren, Yang Feng, Qun Liu, and Dawei Yin. Knowledge diffusion for neural dialogue generation. In **Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)**, pp. 1489–1498, 2018.

[5] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. Bert: Pre-training of deep bidirectional transformers for language understanding. **arXiv preprint arXiv:1810.04805**, 2018.

[6] Thomas Wolf. How to build a state-of-the-art conversational ai with transfer learning, 2019. https://medium.com/huggingface/.

[7] Kushal Arora, Kurt Shuster, Sainbayar Sukhbaatar, and Jason Weston. Director: Generator-classifiers for supervised language modeling. **arXiv preprint arXiv:2206.07694**, 2022.

[8] 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**, pp. 845–854, 2019.

[9] Jiwei Li, Michel Galley, Chris Brockett, Jianfeng Gao, and Bill Dolan. A diversity-promoting objective function for neural conversation models. **arXiv preprint arXiv:1510.03055**, 2015.

[10] Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. Bleu: a method for automatic evaluation of machine translation. In **Proceedings of the 40th annual meeting of the Association for Computational Linguistics**, pp. 311–318, 2002.

[11] George Doddington. Automatic evaluation of machine translation quality using n-gram co-occurrence statistics. In **Proceedings of the second international conference on Human Language Technology Research**, pp. 138–145, 2002.