Transformer による hallucination error の事後修正

森脇恵太 [1]　大野瞬 [1]　杉山弘晃 [2] 酒造正樹 [1]　前田英作 [1]
[1] 東京電機大学システムデザイン工学部
[2]NTT コミュニケーション科学基礎研究所
{18aj128，18aj031}@ms.dendai.ac.jp
h.sugi@ieee.org　{shuzo，maeda.e}@mail.dendai.ac.jp

概要

大規模対話コーパスで学習された End-to-End 対話モデルは非常に自然な文章を生成できることが知られている。一方，文生成時に与えた外部知識と異なる内容を持つ発話文を生成してしまう，hallucination error(以降，HE) が課題となっている．そこで本研究では，HE を含むデータを疑似的に作成し，BART や transformer などのニューラルモデルを用いて，HE の事後修正 (postfix) を試みた．旅行対話タスクにおける自動発話生成において発生した HE の中で重要性が高く，且つ，比較的扱いが容易であると考えられる，3 つの情報源 (アクセス，営業時間，料金) を対象に検討を行った．その結果，transformer では実際にニューラル生成モデルで生成した HE を含む文章 52 件中 29 件を修正した．

1　はじめに

GPT-3（Generative Pre-trained Transformer-3）[1] など大規模ニューラル文章生成モデルの研究が進み，可読性が高く且つ自然な文章が生成できるようになりつつある．それによって，雑談対話のように非目的指向型の対話も実現可能になった。しかし，こうしたニューラル生成モデルによる文章生成では，生成した文章が事実と異なる HE が発生するという問題が指摘されている [2]．例えば，「新宿三丁目駅から徒歩 5 分」という事実があるのにも関わらず，生成モデルが「新宿駅から徒歩 5 分」など誤った文章を生成することがある．この HE の問題に対し，文章生成モデルと TruthfulQA を用いた検証が行われているが，モデルの規模が大きくなるほど事実と異なる生成を行う傾向がある [3].

文章要約の対象となる原文と要約文のいずれか人工的に書き換えることで HE を含む疑似不整合例を作成し，これを学習データとして用いることで postfix モデルを構築する．具体例として，要約文中のエンティティを書き換えることにより作成した疑似不整合例を BART（Bidirectional Auto-Regressive Transformer）[4] を用いて修正する [5]．また，要約文の代名詞や数字の入れ替え，要約文中にある単語を追加，削除することで疑似不整合例を作成し，スパン選択ヘッドを追加した BERT(Bidirectional Encoder Representations from Transformers)[6] を用いて HE の検出と修正の両方を行う手法も提案されている [7].

本論文では，HE を検出し修正する機構 [8] のうち，HE を修正する postfix モデルの検証を行った．旅行対話タスクにおける生成発話において発生した HE の中で重要性が高く，且つ，比較的 HE の修正を学習するためのデータセットの作成が容易であると考えられる，3 つの情報源 (アクセス，営業時間，料金) を対象に検討を行う．

知識源を参考に記述された文章にあるエンティティの書き換える手法と知識源とは無関係の文を追加する手法のそれぞれを用いて HE を含むデータを作成し，BART や trandformers を finetune することで postfix モデルを構築した．また，旅行対話データで finetune した大規模雑談対話モデル [9] を利用し，HE を含むデータセットで学習した postfix モデルがどの程度事実と異なる内容を正しく書き換えられるかを評価した．

2　データセット

2.1　データの収集

クラウドソーシングを利用して収集した観光地情報に基づく旅行代理店対話コーパス [10] を対象にして実験を行った．このコーパスは旅行代理店における客と店員の対話ログで構成されており，12 種類の知識情報を有する 600 箇所の観光地がある．店員の発話文は知識源にある内容を参考に記述されており，

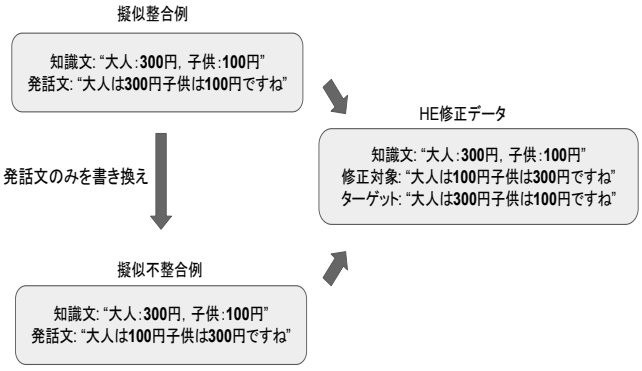

図 1　HE 修正データの作成手順

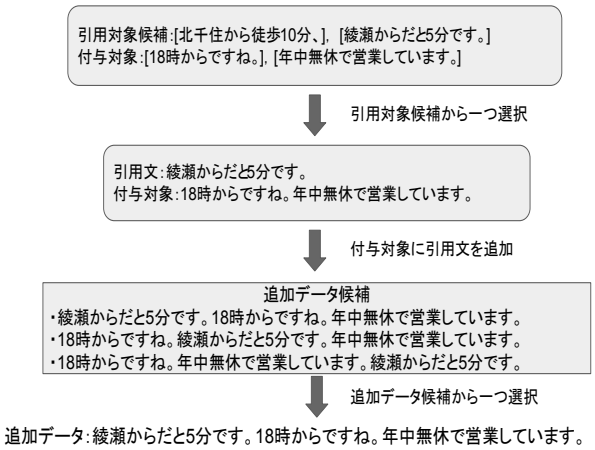

図 2　無関係な発話を含むデータの作成手順

表 1　知識カテゴリごとの書き換え対象エンティティ

知識カテゴリ	書き換え対象 (エンティティ)
アクセス	駅名，路線名，分
営業時間	時分，曜日，日付
値段	値段

12 種類の知識源のうちどの知識源を参考にしたかが明記されている．

2.2　事後修正のための学習用データセット

知識源に数値や固有名詞がある場合ニューラル生成モデルが HE を起こしやすいという傾向に基づき，コーパスにある 12 種類の知識源のうち，営業時間，アクセス，料金のいずれかを参考に記述された店員の発話文と知識源のペアを抽出した．収集した発話文と知識源のペアは営業時間が 468 件，アクセスが 1,600 件，料金が 224 件と知識カテゴリで数が異なるだけでなく，最も多いアクセス情報で約 1,600 件と数が少ないため，一つの発話文と知識源のペアをテンプレートとして複数のデータを疑似的に作成することで，各知識カテゴリごとに 40,000 件の発話文と知識源のペアを作成した．

抽出した発話文と知識源のペアに存在する同一のエンティティを書き換えることで疑似整合例を作成し，疑似整合例の発話文にあるいずれかのエンティティを書き換えることで疑似不整合例を作成した [8]．表 1 に示すように，書き換え対象となるエンティティは知識源に割り当てられているカテゴリによって異なる．図 1 に示すように疑似整合例と疑似不整合例を組み合わせることで，HE を含む発話文とその HE を修正した発話文で構成される HE 修正データを作成した．postfix モデルは HE を含む発話文と参考にした知識源を入力として受け付け，HE を修正した発話文を出力するように学習する．これにより，postfix モデルは知識源に則していないエンティティを書き換える．

2.3　無関係な発話を含むデータの追加

ニューラル生成モデルは事実と無関係な文章を生成する場合がある．これを考慮し、エンティティの書き換えだけでなく，無関係な発話を含んだデータを作成した．データを追加することで事実と不整合な文を削除し、HE を修正することを期待した．このデータセットは 2.2 で作成した HE 修正学習データセットに 40,000 件の無関係な発話を含むデータを追加したものである．エンティティの書き換えを行って集めた 90,000 件の HE 修正データから二つの異なる発話文を抽出した後に，抽出した一方の発話文を句読点で分解し，分解したうちの一つの文をもう一方の発話文に加えることで無関係な発話を含むデータを作成した．無関係な発話を含むデータの作成手順と具体例を図 2 に示す．

2.4　評価用データの収集

HE 修正学習データセットはルールベースで作成されており，同じデータを拡張することで件数を増やしているためデータに偏りが存在する．これを考慮し，作成した HE 修正学習データセット以外で評価用のデータを収集した．評価用データはニューラル生成モデルからの出力データ（ニューラル生成データセット）と人の手で作成した知識源と発話文のペアで構成される評価用データ（handmade データセット）の二種類を作成した．

ニューラル生成データセットはニューラル生成モ

表 2　各種データセットの件数と概要

データセット名	件数
HE 修正学習データセット (baseline)	120,000
HE 修正学習データセット (add_unrelated)	160,000
ニューラル生成データセット	52
handmade データセット	42

デルから出力された HE を含む生成文と文章生成に用いた知識源で構成される．データセットを作成するために，旅行対話データで fine-tuning したニューラル生成モデルを用いて発話文を生成し，人の目で HE を含むかどうかを判定した．

handmade データセットの一部は表 1 の書き換え対象を踏まえて，知識源と HE を含む発話文を作成した．これは HE 修正データがルールベースの書き換えで作成されたことを考慮している．また，表 1 の書き換え対象とは異なるデータを加えている．これは，モデルのロバスト性を評価する意図がある．

handmade データセットは作成時にターゲットとなる教師データを加えているが，ニューラル生成データセットにはターゲットとなる教師データは存在しない．

3　実験

3.1　実験設定

事前学習済みの Transformer，BART を今回作成した HE 修正学習データセットでファインチューニングすることで postfix モデルを構築した．事前学習済みモデルとして NTT コミュニケーション科学基礎研究所が公開している japanese-dualog-transformers[1)]，黒橋研究室が公開している BART 日本語 Pretrained モデル [11][2)] を使用した．それぞれのパラメータ数は japanese-dialog-transformers は 1.6B，BART は 0.12B である．HE 修正学習データセットは 2.3 で作成した無関係な発話を含むデータセット（add_unrelated）と含まないデータセット（baseline）の 2 種類を用意した．また，ニューラル生成データセットと handmade データセットを用いて、それぞれのモデルを評価した．学習，評価に使用したデータセットの件数を表 2 に示す．

モデルの学習，評価にはシーケンスモデリングツールキットである fairseq[3)] を利用した．

1）　https://github.com/nttcslab/japanese-dialog-transformers
2）　https://nlp.ist.i.kyoto-u.ac.jp/?BART 日本語 Pretrained モデル
3）　https://github.com/pytorch/fairseq

表 3　ニューラル生成データセットによる評価結果

モデル	学習用データ	Faithfulness(%)
transformer	baseline	53.8
	add_unrelated	55.8
BART	baseline	42.3
	add_unrelated	44.2

表 4　handmade データセットによる評価結果

モデル	学習用データ	BLEU-4	Faithfulness(%)
transformer	baseline	88.7	83.3
	add_unrelated	78.5	92.8
BART	baseline	88.7	45.2
	add_unrelated	67.4	59.5

モデルへの入力は知識源と HE を含む発話文をセパレーショントークン（[SEP]）で繋げたものである。しかし，事前学習済みの BART ではこのトークンに対応していない．これを回避するために Fusion-in-Decoder[12] を BART に実装し，知識源と修正対象を入力した．

3.2　評価指標

修正結果の評価指標として，修正結果の内容が知識源の内容と一致しているかを一手において評価した Faithfulness[13] と BLEU-4 score[14] を用いた．モデルの生成文に知識源と無関係な文，または知識源と矛盾した文のいずれかが一つも含まれない場合、モデルは HE を修正したと判定する．

モデルの出力を得るために fairseq の Command-line Tools を利用し，beam size を 10 に設定した．これにより一つの入力に対して 10 個の出力候補を得られる．Faithfulness では，10 個の出力候補の中に一つでも HE を修正したと判定されるものがあればその入力を修正できたとして評価している．

3.3　結果・考察

表 3，4 より，BLEU-4 score と Faithfulness の結果が比例するとは限らないことが分かる．入力に用いた発話文とターゲットが似ているため BLEU-4 score が高くなりやすいと考えられる．この結果は，BLEU-4 score では数値が異なるといった単純な HE を正しく評価することができないことを示唆している．BART と transformer で結果に大きな差がでたのはパラメータ数と事前学習に用いたデータセットの差によるものだと考えられる．

評価用データを用いて得られたいくつかの出力例

表 5 transformer におけるデータセットごとの出力例

知識源	修正対象	データセット	出力
料金: 大人:1800 円中学生以下:350 円	中学生以下は 180 円，大人は 350 円ですね	baseline	中学生以下は 1800 円，大人は 350 円ですね
		add_unrelated	大人は 1800 円ですね
大人:900 円，大学生:600 円，高校生:500 円，中学生以下:200 円	大人は 800 円，大学生は 900 円，高校生は 500 円，中学生以下は 500 円です	baseline	大人は 900 円，大学生は 500 円，高校生以下は 200 円です，中学生以下は 200 円です
		add_unrelated	大人は 900 円，大学生は 600 円，中学生以下は 200 円です
東京駅から徒歩 20 分、新宿駅から徒歩 15 分	新宿駅から歩いて 15 分、東京駅からなら 12 分ですね	baseline	新宿駅から歩いて 20 分、東京駅からなら 15 分ですね
		add_unrelated	東京駅から歩いて 20 分、新宿駅から歩いて 15 分ですね
営業時間:7:45〜18:00	18 時閉店です。開いているのは 18 時からです	baseline	7 時 45 分から閉店です。開いているのは 18 時からです。
		add_unrelated	7 時 45 分閉店です。

を表 5 に示す．HE 修正データの書き換え対象であるエンティティに着目すると，一部のデータでは正しく書き換えられていない．特に知識源に出現するエンティティ順序と発話文に出現するエンティティの順序が異なる場合，postfix モデルは正しく書き換え出来ていない．モデルは"大人"と"500 円"のようなエンティティの関連付けが出来ていないと言える．また，知識源に出現するエンティティの順序とモデルの出力に出現するエンティティの順序が同じであることから，エンティティが出現する順序に注目して対象となるエンティティの書き換えを行っていると考えられる．

baseline では発話文にあるエンティティのみ書き換えるのに対し，add_unrelated では発話文にある一文を削除して出力する傾向が見られた．モデルは入力にある発話文の一部を削除することを学習したと考えられる．しかし，無関係な一文に限らず知識源に則っている一文についても削除するものもある．これらから，発話文の削除とエンティティの書き換えを行うことで知識源との整合性を取るように学習したと考えられる．発話文の削除は無関係な一文に関わらず，知識源と矛盾，または含意の一文にも行うことで生成文全体が知識源に則るように文章を生成すると考えられる．

4 おわりに

本研究では，特定のエンティティのみを書き換えたデータ，無関係な発話を追加したデータなどで構成される HE 修正学習データセットを用いて，知識源を使用する文章生成における HE を修正する postfix モデルの学習を行った．

学習データとは独立に用意したデータセットにより評価実験を行い，学習時に使用する HE 修正データを変えることで HE 修正データに沿ってモデルの挙動が変化することを示した．また，HE 修正学習データセットと似たデータがあるにも関わらず HE を修正できない例も存在した．．今後は，HE 修正学習データセットの基となるデータの収集，書き換えルールなど作成手法の拡張が必要である．また，拡張した結果モデルの出力が変化し，HE の修正に貢献したかを手法ごとに検証すべきである．

謝辞

本研究は JSPS 新学術研究 JP19H05693 の助成を受けた．

参考文献

[1] Tom B Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, et al. Language models are few-shot learners. **arXiv preprint arXiv:2005.14165**, 2020.

[2] Ashish Agarwal Clara Fannjiang David Sussillo Katherine Lee, Orhan Firat. Hallucinations in neural machine translation. **Accepted in Workshop of Interpretability and Robustness in Audio, Speech, and Language**, 2018.

[3] Stephanie Lin, Jacob Hilton, and Owain Evans. Truthfulqa: Measuring how models mimic human falsehoods. **arXiv preprint arXiv:2109.07958**, 2021.

[4] Mike Lewis, Yinhan Liu, Naman Goyal, Marjan

Ghazvininejad, Abdelrahman Mohamed, Omer Levy, Ves Stoyanov, and Luke Zettlemoyer. Bart: Denoising sequence-to-sequence pre-training for natural language generation, translation, and comprehension. **arXiv preprint arXiv:1910.13461**, 2019.

[5] Meng Cao, Yue Dong, Jiapeng Wu, and Jackie Chi Kit Cheung. Factual error correction for abstractive summarization models. In **Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)**, pp. 6251–6258, Online, November 2020. Association for Computational Linguistics.

[6] 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.

[7] Wojciech Kryściński, Bryan McCann, Caiming Xiong, and Richard Socher. Evaluating the factual consistency of abstractive text summarization. **arXiv preprint arXiv:1910.12840**, 2019.

[8] 大野瞬, 森脇恵太, 杉山弘晃, 酒造正樹, 前田英作. 分類モデル bert による不整合生成文の検出について. "言語処理学会第 28 回年次大会", 2022(発表予定).

[9] Hiroaki Sugiyama, Masahiro Mizukami, Tsunehiro Arimoto, Hiromi Narimatsu, Yuya Chiba, Hideharu Nakajima, and Toyomi Meguro. Empirical analysis of training strategies of transformer-based japanese chit-chat systems, 2021.

[10] 金田龍平, 芳賀大地, 杉山弘晃, 酒造正樹, 前田英作. 知識源と一対多関係を有する対話コーパスによる発話生成. 言語処理学会第 28 回年次大会, 2022(発表予定).

[11] 田中佑, 村脇有吾, 河原大輔, 黒橋禎夫. 日本語 wikipedia の編集履歴に基づく入力誤りデータセットと訂正システムの改良. 言語処理学会第 27 回年次大会, 2021.

[12] Gautier Izacard and Edouard Grave. Leveraging passage retrieval with generative models for open domain question answering, 2020.

[13] Ankur P Parikh, Xuezhi Wang, Sebastian Gehrmann, Manaal Faruqui, Bhuwan Dhingra, Diyi Yang, and Dipanjan Das. Totto: A controlled table-to-text generation dataset. **arXiv preprint arXiv:2004.14373**, 2020.

[14] 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.