分類モデル BERT による不整合生成文の検出について

大野瞬 [1]　森脇恵太 [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
{18aj031,18aj128}@ms.dendai.ac.jp　h.sugi@ieee.org
{shuzo,maeda.e}@mail.dendai.ac.jp

概要

ニューラル文章生成において, 文章としては自然であるが, 内容が事実と異なるという事実不整合（factual inconsistency）の発生が問題となっている. そこで本研究では, BERT（Bidirectional Encoder Representations from Transformers）の分類タスクを応用して事実不整合となる生成文を検出することを試みる. 実験対象として観光地案内をドメインとし, 客の質問に適切な知識を用いて答えるニューラル生成モデル Hobbyist の出力文を扱う. 料金, 営業時間, アクセス方法に関する知識情報源と正しい応答文との対 2,299 件をもとにし, そこから整合対 6 万件, 不整合対 6 万件からなる疑似データセットを作成した. この疑似データセットを用いて BERT の学習を行った結果, 事実不整合を含むニューラル生成文に対して recall 0.69 の結果を得た. 一方, 日本語 SNLI データセットを用いた学習では 0.50 であり, 不整合検出におけるドメイン適応の重要性が明らかになった.

1　はじめに

GPT-3（Generative Pre-trained Transformer - 3）[1] や BART（Bidirectional Auto-Regressive Transformer）[2] など, 現代のニューラルネットを使った文章生成モデルは人間の書いた文章と相違ない自然な文章生成が可能になっている. 生成モデルを用いた質問応答チャットシステムの実装などができればより人手応答に近い柔軟な応答ができるようになると考えられる. しかし, 生成モデルはしばしば図 1 のような事実と異なる内容を含んだ文章を出力してしまう. このような出力はユーザの信頼を損なうことにつながり, システムを実用レベルにするためにはこの問題に対策を講じる必要がある.

原文：営業9:30〜17:00 休業：無休
生成文章：朝7時45分に開店します。
理想生成：朝9時30分に開店します。

図 1　事実不整合を含んだ出力例

そこで我々は事実不整合を含む出力を検出し, 修正するため分類モデルと修正モデルを用いた事実不整合修正の為の機構を提案する. 生成モデルの出力を分類モデルにかけ不整合検出を行い, もし不整合が検出された場合さらに下に構えている修正モデルに引き渡し, 事後修正を行うことで文章を正しいものにすることが狙いである. 生成モデルの学習用に作られたデータセットを改変, 増強して学習を行うことで特定の生成モデルに特化した修正機構を構築することを試みる. 機構のイメージは図 2 に示した通りである. 本稿では修正モデル部分の話には触れず, 提案機構の中の分類モデル部分についてのみ記述する. 分類モデルには BERT（Bidirectional Encoder Representations from Transformers）[3] を用いた.

これまでの研究として, 例えば, [4] は既存のデータセット（ここでは CNN/DailyMail dataset[5]）から改変例を大量に用意し, BART に学習させ, 他のニューラル生成モデルの出力文章を入力として事実に即した内容に書き換えをするという手法であるが, 十分な成果を得られていない. 一方我々は生成モデルの学習に用いたデータをテンプレートとして改変例を作成し, 事実不整合検出の為の学習データとした. これにより特定の生成モデルに対して高い精度で事実不整合の検出, 修正を行うことを期待できる.

2　データセット

今回作成する学習用データセットは観光地情報に基づく旅行代理店対話コーパス [6] に含まれる基準対話データセットを利用したものである. これに含まれる知識情報と, それを使って作られた店員発話を改変することで疑似的に多くのデータを作成する. 知識と店員発話の整合性を保った疑似整合例と,

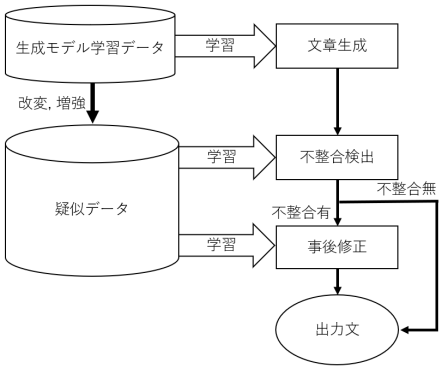

図 2　提案する機構のイメージ

元々の知識及び作成文章
知識：大人:200円子供:100円子供…小中学生平日のみ子供無料
店員発話：いいえ、入園料は大人200円、子供は100円で、平日子供料金は無料なんですよ。

↓

疑似整合例
知識：大人:230円子供:110円子供…小中学生平日のみ子供無料
店員発話：いいえ、入園料は大人230円、子供は110円で、平日子供料金は無料なんですよ。

↓

疑似不整合例
知識：大人:230円子供:110円子供…小中学生平日のみ子供無料
店員発話：いいえ、入園料は大人110円、子供は230円で、平日子供料金は無料なんですよ。

図 3　疑似例の作成例（料金情報）

整合性をなくした疑似不整合例を作成し, モデルの学習に利用する. 知識情報はそれぞれ最大 12 件用意されていたが, 今回は特に数値に対してロバストなモデルになるよう学習を行うため, 知識に数値が含まれている料金情報, アクセス情報, 営業時間情報という 3 つのカテゴリに絞って学習に用いることとする. 数値の含まれているデータに絞る理由としては, 第一に整合性の判定が比較的容易と考えられることである. これら 3 つのカテゴリは観光地情報の中でもある程度決まったフォーマットになっており, 扱いやすいものと考える. 第二に我々が実際に生成モデルを動かした際, 数値周りに不整合が多く発見されたことである. 我々の扱うモデルは NTT コミュニケーション科学基礎研究所開発の雑談対話エンジン [7] を観光地情報に基づく旅行代理店対話コーパスでファインチューニングしたもの（以下, これを Hobbyist と呼ぶ）であり, これは入力に知識情報と質問文を入れると, 知識情報を用いて質問文への回答を生成する. Hobbyist に文章生成を行わせたところ, 特にこれら 3 つのカテゴリを用いて回答生成した際に数値を誤ったかたちの不整合が多くみられた. 以上の理由から, 我々の提案手法の有効性を見るために数値に特化することは有効であると考えた.

アクセス
知識：(1) ＪＲ原宿駅から徒歩5分(2)東京メトロ千代田線明治神宮前駅から徒歩5分
店員発話：はい、ＪＲ原宿駅から徒歩5分ですよ。

図 4　アクセス情報と店員発話の一例

営業時間
知識：営業： 10時30分～20時30分休業：第1・2・3水曜休
店員発話：営業時間は、10時30分～20時30分です。

図 5　営業時間情報と店員発話の一例

疑似例の作成は数値や日付, 駅名等を書き換えることで行うが, 料金, アクセス, 営業時間情報について含まれる書き換え対象が異なるため, それぞれについて異なる形でデータの改変を行う. 疑似例作成の一例を図 3 に示す. 疑似例のカテゴリによって改変方法が異なるため, 以下にそれぞれの手順を示す.

•料金情報の疑似例

疑似整合例　知識及び店員発話に含まれる料金を表す数値の十の位の書き換えを行う.

疑似不整合例　疑似整合例の内容から店員発話部分のみに対してスワッピングまたは数字の付け足しを行い知識と店員発話の整合性を消す.

•アクセス情報の疑似例

疑似整合例　知識及び店員発話に含まれる時間表記及び駅名を同様に書き換える.

疑似不整合例　疑似整合例の内容から店員発話部分のみに対して数値のスワッピングまたは書き換え, 及び駅名の書き換えを行い, 知識との整合性を消す.

•営業時間情報の疑似例

疑似整合例　曜日や日付, 時刻を示す表現に対し, 知識と店員発話の該当する部分を同様に書き換える.

疑似不整合例　疑似整合例の内容から店員発話のみに対して時刻の「分」部分の書き換えを行い知識との整合性を消す.

それぞれ疑似例の元とした知識と店員発話のペア（テンプレート）数として料金情報は 224 件, アクセス情報は 1,607 件, 営業時間情報は 468 件となっている.

疑似例を用いた学習用のデータセットとは別に, Hobbyist の出力を収集した評価用のデータセットも

表 1 事実整合性判定学習データセットの内訳

	整合例数	不整合例数	合計
料金	20,000	20,000	40,000
アクセス	20,000	20,000	40,000
営業時間	20,000	20,000	40,000
合計	60,000	60,000	120,000

表 2 ニューラル生成文データセットの内訳

	整合例数	不整合例数	合計
料金	77	14	91
アクセス	144	9	153
営業時間	99	55	154
合計	320	78	398

作成する. Hobbyist への入力は知識情報と質問文であり, Hobbyist の出力は知識情報を用いた質問文への回答である.

以下に今回作成した 2 つのデータセットについて述べる.

事実整合性判定学習データセット 料金, アクセス, 営業時間情報について作成した疑似整合例, 不整合例をそれぞれのカテゴリについて同じ件数ずつ集め, データセットを作成する.

実験においては, この件数を 20,000 件とする. 内訳を表 1 に示す.

ニューラル生成文データセット Hobbyist を用いて生成した文章に対して, 知識に即しているか否かを人手で判定してラベリングを行ったデータセットを作成する. 料金, アクセス, 営業時間情報のみを用いて 1, 000 件の応答を生成し, うち生成文にも知識にも数値が含まれているもののみを抽出し, まとめたデータセットとなっている. 内訳は表 2 を参照.

3 実験

3.1 実験設定

分類モデルとして学習を行うのは Laboro 社の日本語 BERT（base）である.[8]

黒橋研究室が公開している日本語 SNLI データセット [9] をベースラインデータセットとして我々の提案データセットである事実整合性判定学習データセットと比較を行う. このデータセットからはフィルタリングされた学習データからのみ抽出し, 元についていたラベルが entailment（含意）であるものを整合例, ラベルが contradiction（矛盾）であるものを不整合例として, 表 3 の形でデータセットを整形する.

表 3 日本語 SNLI データセットの内訳

	整合例数	不整合例数	合計
件数	60,000	60,000	120,000

我々の作成したデータと異なりアクセスや料金などのカテゴリ分けはできないため, 単に学習に用いたデータ数を整合, 不整合で 60,000 件ずつに合わせる形とする.

モデルの学習には huggingface transformers[10] を利用してファインチューニングする. このとき, 学習に用いるデータセットを 9：1 に分割し, 前者を学習データ, 後者を検証データとする. ファインチューニングのパラメータ設定は学習率を 5e-5, loss を cross entropy, バッチサイズを 64 とし, 10 エポックの学習を行う. 評価結果には最良モデルの混同行列を載せるが, 最良モデルの基準としては評価データであるニューラル生成文データセットに対する不整合例の recall が最大であるものとする. 本来であれば検証データに対する loss が低いエポックについてデータを載せることが一般的であると考える. このようにした理由として, 学習の際 loss の収束は早く完了してしまうことが多く, しかし評価データに対しての精度は loss 収束後のエポック同士でもかなり異なっていることから loss のみを見ても本来の分類対象に対しての性能が測れないと考えたからである. 加えて, 今回の目的は不整合を含む生成モデル出力の検出であるため評価データの不整合例に対する recall が低いモデルは大量の不整合を見逃していることになり, 目的を果たせていないと考えたことによる.

3.2 実験結果・考察

結果はそれぞれ図 6,7 に示すとおりである. (positive, negative はそれぞれ整合例, 不整合例を示す) 加えて提案データセットについて学習曲線の上から評価データの recall をプロットしたものを図 8 に示す. (train, validation はそれぞれ学習, 検証データである) 提案手法では不整合例, 整合例の recall ともにベースラインよりも高い結果を記録した.

学習曲線について, 学習, 検証データの accuracy は単調増加をしているような動きで順調に学習が進んでいるように見える. そのうえで学習曲線に重ねた評価データの recall を見ると, 特に負例について学習が進むごとに上がっているように見られ, 提案手法に事実不整合検出に対しての有効性があることを示

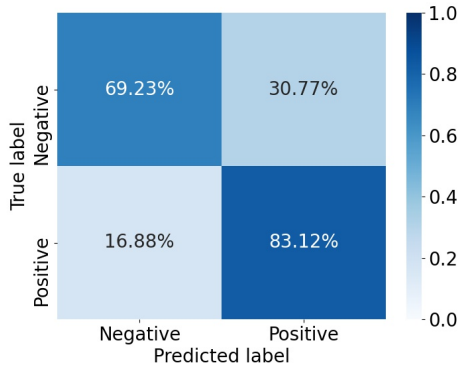

図 6 提案データセットで学習したモデルを用いた評価結果

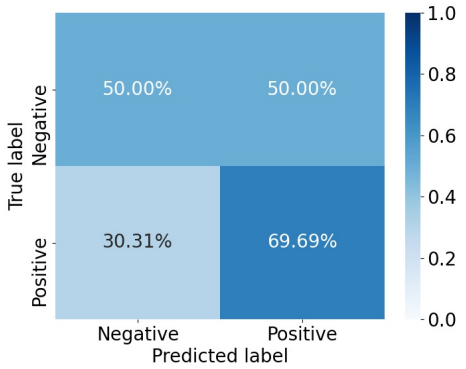

図 7 ベースラインデータセットで学習したモデルを用いた評価結果

している可能性がある.

今回正解できなかった不整合例の内訳は料金 7 件, アクセス 1 件, 営業時間 16 件となっている. 表 2 と見比べると料金, 営業時間, アクセスの順で recall が低い. 今回用意できたテンプレートは料金 224 件, 営業時間 468 件, アクセス 1,607 件であり, テンプレートの件数とカテゴリごとの recall の順は同一である. 精度向上にはそれぞれのテンプレート数を増やす必要が想定され, 今後この手法で精度を上げていくためにはテンプレートの拡充が必須であると考えられる.

今回整合例に対して不整合例の精度が落ちているのは, 整合例は単に知識の内容が生成内容に含まれていれば良いのに対し, 不整合例は整合例以上に広い分布をするからと考える. 今回正解できなかった不整合例をさらうと, 知識内容の一部のみ矛盾している, 知識に一切含まれない内容が含まれているというもので占められていた. どちらも今回の作成疑似例でカバーすることができると考えていたが, 実際にはそういったことを学習させることはできておらず, よりバリエーションに富んだ疑似例作成手法を考える必要がある.

4 おわりに

本稿では, 生成モデルに含まれる事実不整合を BERT の分類タスクを応用して検出することを試みた. 分類モデルの学習に生成モデルの学習に用いたデータを改変したデータを用いることで, ベースラインの手法に対する提案手法の有効性を示すことができた. 今後, 他の生成モデルに対して同様の手法を用いた事実不整合検出の可能性についても検討を進める. また, 実際に現れる事実不整合のパターンは無

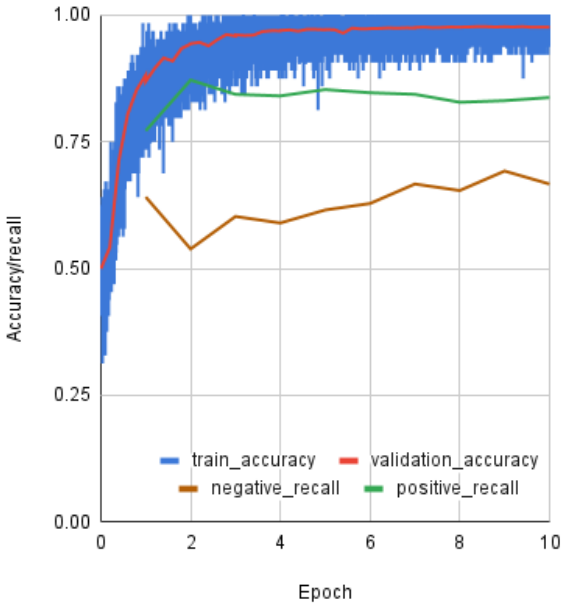

図 8 提案データセットを用いたモデルの学習曲線, 評価データの recall

数にあり, それらを限られたデータから効率的に検出する手法の調査も必要となるだろう.

謝辞

本研究は JSPS 新学術研究 JP19H05693 の助成を受けたものである.

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, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel M. Ziegler, Jeffrey Wu, Clemens Winter, Christopher Hesse, Mark Chen, Eric Sigler, Mateusz Litwin, Scott Gray, Benjamin Chess, Jack Clark, Christopher Berner, Sam

McCandlish, Alec Radford, Ilya Sutskever, and Dario Amodei. Language models are few-shot learners. In **Advances in Neural Information Processing Systems 33: Annual Conference on Neural Information Processing Systems 2020**, Online, December 2020.

[2] 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:1910.13461 [cs, stat]**, October 2019.

[3] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. BERT: Pre-training of deep bidirectional transformers for language understanding. In **Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics (NAACL)**, Vol. 1, pp. 4171–4186, Minneapolis, MN,USA, June 2019. Association for Computational Linguistics.

[4] 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.

[5] Karl Moritz Hermann, Tomáš Kočiský, Edward Grefenstette, Lasse Espeholt, Will Kay, Mustafa Suleyman, and Phil Blunsom. Teaching machines to read and comprehend. In **Proceedings of the 28th International Conference on Neural Information Processing Systems (NIPS)**, Vol. 1, pp. 1693–1701, Cambridge, MA, USA, 2015. MIT Press.

[6] 金田龍平, 芳賀大地, 杉山弘晃, 酒造正樹, 前田英作. 知識源との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.

[7] 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. **arXiv:2109.05217 [cs]**, September 2021.

[8] Xinyi Zhao, Masafumi Hamamoto, and Hiromasa Fujihara. Laboro BERT Japanese: Japanese BERT pre-trained with web-corpus. **https://github.com/laboroai/Laboro-BERT-Japanese**, 2020.

[9] 吉越卓見, 河原大輔, 黒橋禎夫. 機械翻訳を用いた自然言語推論データセットの多言語化. 情報処理学会研究報告, Vol. 2020-NL-244, No. 6, pp. 1–8, 2020.

[10] Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Remi Louf, Morgan Funtowicz, Joe Davison, Sam Shleifer, Patrick von Platen, Clara Ma, Yacine Jernite, Julien Plu, Canwen Xu, Teven Le Scao, Sylvain Gugger, Mariama Drame, Quentin Lhoest, and Alexander Rush. Transformers: State-of-the-art natural language processing. In **Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP): System Demonstrations**, pp. 38–45, Online, October 2020. Association for Computational Linguistics.