忠実性向上のために n-gram の抽出性を報酬とする強化学習を用いる抽象型要約

星野 智紀　上垣外 英剛　渡辺 太郎
奈良先端科学技術大学院大学
{hoshino.tomoki.ho6, kamigaito.h, taro}@is.naist.jp

概要

抽象型要約生成モデルの出力する要約には、要約元となる原文書に記述される事実とは異なる内容が出力されることがある。本研究では強化学習の報酬を要約の原文書に対する抽出性とし、それを向上させることで、要約生成モデルが出力する要約の忠実性向上に取り組んだ。抽象型要約のタスクで使用されることが多い xsum データセットを用いた実験の結果、出力された要約の抽出性が向上し、忠実性に対する自動評価尺度である FEQA のスコアと原文書に対する含意が改善されることが確認された。

1　はじめに

近年、事前学習済み系列変換モデルに対して fine-tune することで、要約生成モデルの出力する抽象型要約の精度は大きく向上し、流暢な要約生成が可能となった [1, 2]。一方、要約生成モデルの出力する要約に、要約元となる原文書には存在しない、異なる事実を含んでしまう問題が指摘されている [3]。この問題点は要約生成をアプリケーションとして使用する上では深刻であり、例えば、ニュース記事の要約が事実とは反する誤りを含む場合、誤情報を拡散する可能性があり、対処が必要である。この深刻な問題を解決するために、先行研究では、要約生成モデルの出力する要約の事実性を向上させる様々な研究がされてきた。

抽象型要約において、事実と異なる要約が生成される原因の一つとして、原文書には存在しない単語や意味関係が出現する問題（hallucination）が指摘されている [4]。これは、原文書に対して忠実ではない要約を生成することが事実とは異なる要約を生成する要因となり得るという考えである。

また、学習を行うデータセット自体に忠実性を低下させる問題があるとの指摘もなされている。主流な要約タスクのデータセットである xsum データセット [5] において、正解となる要約中に多くの hallucination を含むことが指摘されている [4, 6]。従って、このようなデータセットを用いて学習された要約生成モデルは忠実ではない要約を生成する可能性がある。

表 1　事実に誤りのある要約例

原文書	Sir Bruce hosted Strictly .... recover in time to co-host Strictlyś Children In Need special .... Actors Jenny Agutter, Laura Main, Stephen McGann and Jack Ashton will compete in the one-off contest as part of BBC Oneś Children In Need telethon. ....
要約	Former Strictly Come Dancing host Sir Bruce Forsyth is to return to the BBC for a one-off special.

上記で指摘されている問題点を考慮すると、抽象型要約における忠実性を向上させるためには、出力すべき要約をデータセットのみに基づいて学習しないような方法が求められる。

これらの背景を踏まえ、本研究では強化学習を通じて要約生成モデルの忠実性を向上させるための手法を提案する。強化学習では報酬に基づいた学習が行われるため、学習データのみに制約されることなく、与えられた報酬に沿う要約を学習することが可能である。また、強化学習の報酬として、我々は生成された要約の原文書に対する抽出性に着目した。

抽象型要約は、原文書に存在する文や表現を抜き出す（抽出する）のではなく、系列変換モデルのデコーダでトークン単位での生成を行うことで要約を出力している。これにより抽象型要約モデルは流暢な要約生成が可能である一方、その柔軟な表現力により原文書に存在しない文字列を出力することで hallucination を引き起こす可能性がある。

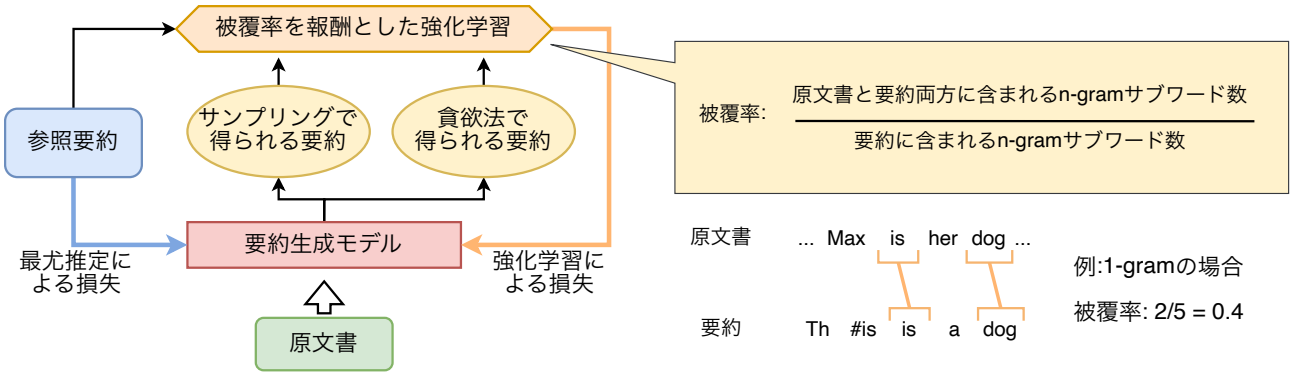

図 1　（左）提案手法の概要（右上）被覆率の計算式（右下）被覆率の計算例

抽出性への着目は、系列変換モデルの出力の抽出性を向上させることで、原文書の内容に沿う要約を生成しやすくなることで hallucination を低減させ、結果として要約の忠実性が向上するのではないかという考えに基づく。またこの方法は、参照要約ではなく原文書に対しての抽出性を向上させるため、参照要約に含まれる hallucination の影響を低減させることも期待できる。さらに抽出性の計算は後述するようにモデルでの予測を必要とせず軽量であるため、強化学習の報酬としても適している。

上記を実現するため、要約生成モデルから出力される要約がどの程度、要約元の原文書と一致しているかをサブワード単位の n-gram で算出し、その一致率（以下、被覆率）[7] を強化学習の手法である Self-critical Sequence Training (SCST) [8] の報酬として fine-tune 済みの要約生成モデルに対して学習を行った。

抽象型要約タスクで頻繁に使用される xsum データセットを用いた実験の結果、提案手法は最尤推定に基づく損失関数で fine-tune されたベースラインと比較し、正解の要約との一致を測る要約の評価指標である ROUGE は低下したものの、忠実性に対する評価指標である FEQA と原文書に対する含意の向上が確かめられた。

2　提案手法

提案手法は事前学習済み系列変換モデルを要約データにより fine-tune したモデルに対して、SCST に従った重みパラメータの更新を行う。モデルの概略を図 1（左）に示した。

2.1　要約生成に用いる系列変換モデル

本研究では、事前学習済み系列変換モデルの BART [1] を要約生成に用いる。BART は原文書を入力として要約を出力する Transformer エンコーダ・デコーダモデル [9] である。エンコーダへの入力系列を $\boldsymbol{x} = x_1 \cdots x_n$、デコーダからの出力系列を $\boldsymbol{y} = y_1 \cdots y_m$ とするとき、出力系列は次式に従い出力される。

$$\hat{\boldsymbol{y}} = \arg\max_{\boldsymbol{y}} \prod_{t=1}^{m} p(y_t | \boldsymbol{x}, y_1 \cdots y_{t-1}) \tag{1}$$

2.2　目的関数

提案手法の目的関数は最尤推定に基づく損失 $\mathcal{L}_{mle}$ と、強化学習に基づく損失 $\mathcal{L}_{rl}$ のハイパーパラメータ γ による重み付き和 $\mathcal{L}_{mixed} = \gamma\mathcal{L}_{rl} + (1-\gamma)\mathcal{L}_{mle}$（ただし、$0 \leq \gamma \leq 1$）で表される。

$\mathcal{L}_{mle}$ は訓練データ中の参照要約の系列 $\boldsymbol{y}^*$ を用いて次のように計算される

$$\mathcal{L}_{mle} = -\sum_{t=1}^{m} \log p(y_t^* | \boldsymbol{x}, y_1^* \cdots y_{t-1}^*) \tag{2}$$

SCST に基づく $\mathcal{L}_{rl}$ は、系列変換モデルの出力するサンプリングされた系列 $\boldsymbol{y}^s$ と貪欲法によって得られる系列 $\hat{\boldsymbol{y}}$ を用いて、次のように計算される。

$$\mathcal{L}_{rl} = -(r(\boldsymbol{y}^s) - r(\hat{\boldsymbol{y}})) \sum_{t=1}^{m} \log p(y_t^s | \boldsymbol{x}, y_1^s \cdots y_{t-1}^s) \tag{3}$$

$r(\cdot)$ は与えられた系列に対する報酬を返す関数であり、次節で説明する。

2.3　報酬

提案手法の報酬 $r(\cdot)$ には要約の被覆率を用いた。被覆率は要約に出現する単語のうち、それらの単語がどの程度原文書でも出現しているのかを示す割合である。これは Grusky らが提案している coverage [7] の算出方法を元にしている。

本提案手法では、図 1（右下）のように単語単位ではなくサブワード単位で要約の被覆率を算出す

る。また、被覆率を 1 サブワードごとにだけ算出するのではなく、サブワードの n-gram に対して、被覆率を計算する。これによって、単語単位より細かい粒度から複数単語での被覆率までを報酬で考慮することができる。

繰り返しなどを含む生成結果に対して不当に高い被覆率を与えないよう、原文書に出現するあるサブワードの n-gram の回数以上に要約にそのサブワードの n-gram が出現した場合は、原文書での出現回数までしか被覆率の算出には含めないようにした。

なお、我々は適切な n-gram のサイズを知るため、後述する実験では、n-gram の範囲を変えて複数のモデルを学習し、それぞれのモデルが生成する要約の違いについて分析した。

3 実験設定

データセット 実験には xsum データセットを用いた。xsum はニュース記事を原文書とするデータセットで、一つのニュース記事に対して、一つの要約が対応している。xsum は CNN/DailyMail [10] などのデータセットにくらべ、抽象性の高い参照要約である。そのため、強化学習によって抽出性を高める本研究の目的に沿うデータセットと判断した。

評価指標 本研究では、提案手法の忠実性と流暢性を評価する。そのため、忠実性と流暢性に関する自動評価指標を用い、要約生成モデルの出力結果を評価した。

忠実性の評価には FEQA [11] と Entailment [12] を用いた。FEQA は要約から質問と回答を生成し、その質問に対して、質問応答モデルが原文書から回答を抽出する。回答があらかじめ生成した回答と一致していれば、忠実性があるとみなす。Entailment は自然言語推論タスクのデータセットで fine-tune した事前学習済言語モデル RoBERTa [13] を用いている。自然言語推論タスクとは、前提文と仮説文とが与えられた時に、その 2 文の関係が含意（entailment）・中立（neutral）・矛盾（contradiction）のいずれであるかを分類するタスクである [14]。前提文・仮説文をそれぞれ原文書・要約とし、含意に分類される要約が増加すれば忠実性が向上したとみなす。

流暢性の評価には、事前学習済み言語モデルである BERT [15] を用いる、擬似対数尤度スコア（以下、PLL）[16] を用いた。なお、Salazar らによって提案された PLL は系列長によってスコアが正規化されていないが、本稿の実験結果で記載している PLL の値は要約長で正規化された値である。

要約の情報性を測る指標として、ROUGE [17] を用いた。ROUGE は参照要約との語彙の一致を測ることで、要約として重要であると考えられる情報をどれだけ含んでいるかを示す。

また、抽出性を報酬とした強化学習によって、要約の被覆率が向上しているかの評価も行なった。

これらの評価には xsum データセットの評価データを用いた。

実装・学習 提案手法を実装し、最尤推定に基づく $\mathcal{L}_{mle}$ の損失関数のみで fine-tune した BART モデルをベースラインとし、比較を行った。

提案手法の実装には huggingface が公開する transformers を用いて実装を行った。細かな実験設定については付録 A に記した。

提案手法である SCST による学習の前に、BART モデルに対して最尤推定に基づく損失 $\mathcal{L}_{mle}$ による学習を 3 エポック行った。これは、SCST による学習の際に収束を速めるためである。また、この学習を行ったモデルを本実験のベースラインとしている。SCST に基づく学習は 1 エポック行った。

これらの学習はともに、xsum データセットの訓練データを用いた。

4 実験結果と考察

忠実性と流暢性 表 2 に忠実性と流暢性に関する結果を示した。忠実性に関しては報酬なしのベースラインと比べて、FEQA の場合は 2, 4-gram 報酬の場合には向上が見られた。また、Entailment に関しては、全ての n-gram 報酬の場合で向上が見られた。流暢性に関しては PLL の値は 4-gram 報酬で低下が見られるものの、他の設定では参照要約と同等であり大きく低下してはいない。この結果から、我々が提案する n-gram に基づく抽出性を報酬に用いることで参照要約と同等程度に流暢で、かつより高い忠実性を持つ要約が生成可能であることが分かった。

情報性 表 3 に情報性に関する結果を示した。報酬を使用しないベースライン手法が最も高い ROUGE-1,2,L を達成している。また、提案手法に関しては、3,4-gram 報酬を与えたモデルは指標の値が低くなった一方、それらより少ない、あるいは、多い n-gram 報酬を与えたモデルは、指標の値が相対的に高くなった。

これは強化学習ではモデルが参照要約とは異なる系列を学習しようとする特性と、ROUGE スコアが

表 2 FEQA・Entailment・PLL

	FEQA	Entailment	PLL
参照要約	0.2636	0.1147	-1.585
報酬なし	0.2679	0.1571	**-1.233**
1-gram 報酬	0.2579	0.1697	-1.677
2-gram 報酬	0.2739	0.1757	-1.639
3-gram 報酬	0.2525	**0.1833**	-1.548
4-gram 報酬	**0.2823**	0.1809	-1.992
5-gram 報酬	0.2668	0.1808	-1.654
6-gram 報酬	0.2654	0.1689	-1.640

表 3 情報性

	ROUGE-1	ROUGE-2	ROUGE-L
報酬なし	**0.4099**	**0.1812**	**0.3311**
1-gram 報酬	0.3755	0.1505	0.2934
2-gram 報酬	0.3733	0.1506	0.2937
3-gram 報酬	0.3562	0.1323	0.2688
4-gram 報酬	0.3525	0.1357	0.2727
5-gram 報酬	0.3705	0.1490	0.2923
6-gram 報酬	0.3769	0.1527	0.2977

単語単位に基づく表層上の一致のみに限定されることに起因することが要因の一つであると考えられる。その一方で、提案手法が抽出性を高める過程で、要約上重要である、キーワードなどの情報を用いていないため、抽出対象の重要度を考慮することは今後の改善点であると考えられる。

抽出性 表 4 に参照要約、ベースラインと提案モデルの被覆率の平均値の比較を示した。提案モデルの値は、それぞれ n-gram サブワード被覆率報酬を与えたモデルの n-gram サブワード被覆率を示している。この結果から、提案手法の強化学習によって、実際に n-gram サブワード被覆率は向上することがわかった。

考察 実験結果から本提案手法は参照要約と同等程度に流暢でかつより忠実性が高い要約が生成可能であることが判明した。その一方で情報性については低下が見られ、抽出性を考慮する際に対象とする n-gram の重要度も同時に考慮すべきであることを示唆する結果が得られた。

なお、要約性の評価指標である ROUGE は、基本的には参照要約との単語単位の n-gram の一致を測る指標であるため、強化学習の影響によって ROUGE の値が下がったということは、参照要約に現れない単語が出現するようになったと考えることもできる。この結果は、参照要約に存在する hallucinaiton の影響を低減させることを期待する本提案手法の動機に沿ったものであるといえる。

表 4 抽出性

	参照要約	ベースライン	提案手法
1-gram	0.7250	0.7128	**0.7524**
2-gram	0.2983	0.2859	**0.3402**
3-gram	0.1426	0.1373	**0.1658**
4-gram	0.07455	0.07220	**0.1526**
5-gram	0.04085	0.03985	**0.06183**
6-gram	0.02385	0.02345	**0.03598**

また、3,4-gram 付近の報酬を与えたモデルの忠実性がベースラインに比べて向上し、流暢性・情報性は低下した。それらに比べ、1-gram や 6-gram 報酬を与えたモデルは忠実性の向上が小さく、流暢性・情報性の低下も小さかった。これは、1-gram サブワードの被覆率はベースラインモデルの出力自体が高く、報酬を与えても出力への変化が小さく、また、高い n-gram サブワードの被覆率は、ベースラインモデルの出力がかなり低く、被覆率を向上させる学習がうまくできなかったことからと考えられる。

このことから、本実験設定において、数単語程度にわたる n-gram サブワードを利用できる 3,4-gram を使用することが最も効果的な報酬となることがわかった。

5 まとめ

本研究では、要約生成モデルが出力する要約の忠実性向上を目的として、要約元となる原文書と要約生成モデルが出力した要約との n-gram サブワード被覆率を報酬とした SCST に基づく強化学習モデルを提案した。

xsum データセットを用いた実験の結果、提案手法は最尤推定に基づく損失関数で fine-tune されたベースラインと比較し、参照要約に含まれる情報をどれだけカバーしたかを示す評価指標である ROUGE は低下したものの、流暢性に関する評価指標である PLL は参照要約と同等程度の値を維持し、忠実性に対する評価指標である FEQA と原文書に対する含意の向上が確かめられた。

参考文献

[1] Mike Lewis, Yinhan Liu, Naman Goyal, Marjan Ghazvininejad, Abdelrahman Mohamed, Omer Levy, Veselin Stoyanov, and Luke Zettlemoyer. BART: Denoising Sequence-to-Sequence pre-training for natural language generation, translation, and comprehension. In **Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics**, pp. 7871–7880, Online, July 2020. Association for Computational Linguistics.

[2] Jingqing Zhang, Yao Zhao, Mohammad Saleh, and Peter J Liu. PEGASUS: pre-training with extracted gap-sentences for abstractive summarization. In **Proceedings of the 37th International Conference on Machine Learning**, No. Article 1051 in ICML'20, pp. 11328–11339. JMLR.org, July 2020.

[3] Ziqiang Cao, Furu Wei, Wenjie Li, and Sujian Li. Faithful to the original: fact-aware neural abstractive summarization. In **Proceedings of the Thirty-Second AAAI Conference on Artificial Intelligence and Thirtieth Innovative Applications of Artificial Intelligence Conference and Eighth AAAI Symposium on Educational Advances in Artificial Intelligence**, No. Article 586 in AAAI'18/IAAI'18/EAAI'18, pp. 4784–4791. AAAI Press, February 2018.

[4] Joshua Maynez, Shashi Narayan, Bernd Bohnet, and Ryan McDonald. On faithfulness and factuality in abstractive summarization. In **Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics**, pp. 1906–1919, Online, July 2020. Association for Computational Linguistics.

[5] Shashi Narayan, Shay B Cohen, and Mirella Lapata. Don't give me the details, just the summary! Topic-Aware convolutional neural networks for extreme summarization. In **Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing**, pp. 1797–1807, Brussels, Belgium, 2018. Association for Computational Linguistics.

[6] Feng Nan, Ramesh Nallapati, Zhiguo Wang, Cicero Nogueira dos Santos, Henghui Zhu, Dejiao Zhang, Kathleen McKeown, and Bing Xiang. Entity-level factual consistency of abstractive text summarization. In **Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume**, pp. 2727–2733, Online, April 2021. Association for Computational Linguistics.

[7] Max Grusky, Mor Naaman, and Yoav Artzi. Newsroom: A dataset of 1.3 million summaries with diverse extractive strategies. In **Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)**, pp. 708–719, New Orleans, Louisiana, June 2018. Association for Computational Linguistics.

[8] Steven J Rennie, Etienne Marcheret, Youssef Mroueh, Jerret Ross, and Vaibhava Goel. Self-critical sequence training for image captioning. In **2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)**, pp. 7008–7024. IEEE, July 2017.

[9] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Lukasz Kaiser, and Illia Polosukhin. Attention is all you need. In Von Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, I Guyon R. Garnett, and U, editors, **Advances in Neural Information Processing Systems**, Vol. 30. proceedings.neurips.cc, June 2017.

[10] Ramesh Nallapati, Bowen Zhou, Cicero dos Santos, Çağlar Gulçehre, and Bing Xiang. Abstractive text summarization using sequence-to-sequence RNNs and beyond. In **Proceedings of the 20th SIGNLL Conference on Computational Natural Language Learning**, pp. 280–290, Berlin, Germany, August 2016. Association for Computational Linguistics.

[11] Esin Durmus, He He, and Mona Diab. FEQA: A question answering evaluation framework for faithfulness assessment in abstractive summarization. In **Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics**, pp. 5055–5070, Online, July 2020. Association for Computational Linguistics.

[12] Tobias Falke, Leonardo F R Ribeiro, Prasetya Ajie Utama, Ido Dagan, and Iryna Gurevych. Ranking generated summaries by correctness: An interesting but challenging application for natural language inference. In **Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics**, pp. 2214–2220, Florence, Italy, July 2019. Association for Computational Linguistics.

[13] Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. RoBERTa: A robustly optimized BERT pretraining approach. July 2019.

[14] Adina Williams, Nikita Nangia, and Samuel Bowman. A Broad-Coverage challenge corpus for sentence understanding through inference. In **Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)**, pp. 1112–1122, New Orleans, Louisiana, June 2018. Association for Computational Linguistics.

[15] 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: Human Language Technologies, Volume 1 (Long and Short Papers)**, pp. 4171–4186, Minneapolis, Minnesota, June 2019. Association for Computational Linguistics.

[16] Julian Salazar, Davis Liang, Toan Q Nguyen, and Katrin Kirchhoff. Masked language model scoring. In **Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics**, pp. 2699–2712, Online, July 2020. Association for Computational Linguistics.

[17] Chin-Yew Lin. ROUGE: A package for automatic evaluation of summaries. In **Text Summarization Branches Out**, pp. 74–81, Barcelona, Spain, July 2004. Association for Computational Linguistics.

A　実装の詳細

本研究のベースライン・提案手法ともに huggingface にで公開されている事前学習済み系列変換モデルのパラメータ “facebook/bart-base“[1)] を用いた。

ベースラインモデルは、BART に対して xsum データセットを用いて fine-tune を行った。学習率は 5×10^{-5}、ミニバッチサイズは 8、Gradient Accumulation Step は 1 とした。

提案手法は、ベースラインモデルに対して、xsum データセットを用いて SCST に基づく強化学習を行った。学習率は 1×10^{-4}、バッチサイズは 8、Gradient Accumulation Step は 8 とした。また、目的関数のハイパーパラメータ γ は 0.5 とした。

強化学習のための要約生成時は beam サイズを 4 に設定して生成を行い、最もスコアの高かった出力系列を報酬の計算と学習に用いた。また、強化学習のためのサンプリングされた要約を生成するために、Top-K, Top-P サンプリングを用いた。K, P はそれぞれ 50, 0.8 に設定した。

強化学習は初期値の影響を受けやすいため、提案手法では初期値を変更した 3 つのモデルを学習し、その中で開発データに対し最も ROUGE-1 の値が高いものを忠実性と流暢性との評価対象とした。これは、要約として最も相応しい文書に対して忠実性と流暢性とを確認するべきと考えたからである。

評価分析に用いる要約生成は、beam サイズを 4 に設定して生成を行い、最もスコアの高かった出力系列を出力結果とした。

1)　https://huggingface.co/facebook/bart-base