語彙制約を間接的に用いた平易な要約の生成

森田 一[1]　飯塚 洸二郎[1]　久保 光証[1]
[1] 株式会社 Gunosy
{hajime.morita, kojiro.iizuka, mitsumasa.kubo}@gunosy.com

概要

ニュース配信はより多くの人に必要とする情報を届けることを目的としているが，必要な情報が過不足なく行き渡るまでには多くの課題が残されている．例えば，読解力や語彙力には個人により開きがあり，必要な情報であったとしても記事の難しさが読者に合っていなければ読むことが難しくなってしまう．必要な情報が届かない要因の一つが記事の文章が難しすぎることであるとすれば，記事の平易な要約を提供することがより多くの人へニュースを届けるために重要となる．

ここでは，ニュース記事を対象に平易な要約の生成を目的として，大規模言語モデルを利用した文章の平易化と要約生成の複合タスクを学習する方法を提案する．本稿では，要約に用いる語彙を制限することが，平易な言葉への言い換えを言語モデルから引き出せる一方で，幻覚とよばれる原文と異なる内容の要約生成が生成されやすくなることを示し，語彙を制限した要約候補をランキング学習に用いることで平易な要約生成を学習する方法を提案する．

1　はじめに

ニュース配信は最新の出来事や情報をすばやく，より多くのニュースを必要とする人々へ届けることを使命としており，推薦・パーソナライズに代表される多くの研究が行われている [1, 2]．しかし，ニュースが必要とする人すべてに行き渡る理想的な状態までには大きな隔たりが残っている．例えば，年金のような社会制度の変更はすべての人に関わるニュースであるが，令和 4 年の年金制度変更について知っていると回答した割合は 33%にとどまるとされている [3]．

ニュースが必要な人に届かない要因は，大きく分けて記事自体，配信プラットフォーム，ユーザのそれぞれに存在していると考えられる．ニュース記事自体が要因となるケースとしては長過ぎる記事や難解な記事などユーザの求める形とは記事が異なる場合，配信プラットフォームが要因となるケースとしては興味のあるユーザへ記事を届けられていない場合，ユーザが要因となるケースとしては可処分時間の減少など，様々な要因がありうる．本稿では記事自体が要因となるケースの中でも特に，必要としているにも関わらず読解の困難さからニュースを読むことができない，またはニュースを読むこと自体を避けてしまう場合について注目する．

ニュース記事の難しさがニュースを敬遠する要因となる場合，もちろん難解な語彙や複雑な文法はニュースの理解を妨げる要因となるが，記事の長さも同様にニュースを読解を妨げる要因となりうる．この両方の要因を取り除くためには，難解な語彙や文法で書かれた文を分かりやすいものに変えるだけでなく，要点を絞り記事の長さ自体を短くする必要がある．つまり，記事に対して 難しい語彙や文法を一定の意味を保ちつつ平易な文章に置き換える平易化と，文章の重要な情報を残したまま文章を短く書き換える要約の両方を行う必要がある．

本稿では，平易な要約の生成を目的として，大規模言語モデルを利用した文章の平易化と要約生成の複合タスクを学習する方法を提案する．生成型のモデルを用いて平易な要約生成を行う単純な方法としては，難しい語彙を生成しないように要約生成時に利用可能な語彙を制限し，平易な語彙のみで要約を生成することが考えられる．しかし語彙の制限は，言語モデルが言い換え表現として適切なものを知らなかったとしても原文通りに出力することができなくなるため，幻覚 [4] とよばれる，原文とは内容の異なる誤った要約文の生成を増加させる懸念がある．提案手法では，語彙の制限による幻覚の増加を回避するため，通常の要約候補に加えて語彙を制限して生成した要約候補を候補のランキング学習 [5] に用いる．

2 関連研究

2.1 平易な要約生成

これまでにも平易な要約を生成するいくつかの取り組みが行われている．菅井ら [6] は平易化と要約の処理をそれぞれ適用し，短く平易なテキストの生成を行っている．平易化には統計的機械翻訳の手法を用いて，要約に用いる学習データとは別に「やさしい日本語コーパス」[7]，「やさしい日本語拡張コーパス」[8]，NNWE (NHK NEWS WEB EASY) [9] の一部を用いて学習している．本稿も同じく NNWE のデータで評価を行っているが，平易化の学習のための外部リソースは利用していない．

Zaman らは Pointer-generator network の損失関数を拡張し，平易な要約の生成を行っている [10]．この手法では平易な文生成を学習するため，平易な語か否かをアノテーションした辞書を持っており，生成した文の難しさを損失関数として表現している．本稿の提案手法では辞書のような外部リソースは用いておらず，事前学習済み言語モデルを用いることにより少量の学習データで平易化を学習している点が異なる．

2.2 BRIO

BRIO (Bringing order to Abstractive Summarization) は Liu らによって提案された生成型自動要約の学習手法である [5]．BRIO では訓練時の目的関数を拡張し，通常のクロスエントロピー損失関数に加えて複数候補のリランキングによるコントラスティブ損失関数を用いる．ここで，リランキングに用いる候補は事前学習済みの要約モデルによって予め生成されたものを利用する．コントラスティブ損失関数 $\mathscr{L}$ は S_i と S_j をそれぞれ i，j 番目の候補，S^* を参照要約として，$\mathrm{ROUGE}(S_i, S^*) > \mathrm{ROUGE}(S_j, S^*)$ とするとき，次の式で表される．

$$\mathscr{L}_{ctr} = \sum_i \sum_{j>i} \max\left(0, f(S_j) - f(S_i) + \lambda_{ij}\right) \quad (1)$$

ここで，$f(S_i)$，$f(S_j)$ は長さについて正規化された要約文の生成確率，λ_{ij} は候補間の順位差に基づいて計算されたマージンを表す．

本稿では BRIO の枠組みを平易な要約文のランキング学習のために利用するが，候補を並び替えるスコアとして ROUGE [11] の代わりに文の平易化を考慮した指標を用いるほか，平易化の学習をより効率的にするために候補生成時に語彙制約を利用する．

3 コーパス

3.1 NHK NEWS WEB EASY

NNWE (NHK NEWS WEB EASY) [9] は小中学生や日本語を母語としない人に向けて，NNW (NHK NEWS WEB)[12] の記事をやさしい日本語に書き直したニュースサイトである．各記事は元となった NNW の記事に対応付けられており，本稿では NNW-NNWE の記事対を原文と参照要約のペアとして利用する．平日ごとに 4–5 記事が配信されており，過去 1 年分の記事を参照することができる．

3.2 Livedoor 3 行要約データセット

livedoor ニュース [13] は Web ニュースサイトであり，一部の記事には編集者によって人手で作成された 3 行要約が追加されている．小平ら [14] はここから，2014 年から 2016 年までの 3 年分を収集し，記事 ID のリストを訓練，開発，テストに分けたデータセットとして公開している．本稿ではベースとなる日本語要約モデルを学習するため，このデータセットをもとにクロールしなおしたものを訓練データとして利用する．

4 提案手法

語彙の制限による幻覚の増加を回避するため，利用可能な語彙を制限した候補を通常の候補と合わせて候補のランキング学習 [5] に用いる．まず，候補生成に用いる事前学習モデルについて説明し，その後そのモデルを元にランキング学習を行う提案手法について説明する．

4.1 平易な要約生成のベースモデル

本稿で用いる手法は，ランキング学習に用いる候補を生成するために平易な要約を生成するベースモデルが必要となる．このベースモデルを構築するため，Livedoor 3 行要約のデータセットで学習した後，さらに平易な要約のデータセットで追加の学習を行った．以降，このモデルをベースモデルと呼ぶ．本稿では，要約のモデルとして T5 [15] を用いている．本稿で扱うデータに合わせ，日本語の事前学習済み言語モデルとして，**megagonlabs/t5-base-japanese-web**[16] を利用した．

4.2 語彙制約モデル

平易な要約を生成しようとした時に，単純かつ外部データの必要のない方法として，言語モデルが生成する語彙を制限し，要約の生成時に難解な語彙の利用を禁止する方法がある．文生成時に語彙を制限する手法は，Hu ら [17] により提案された手法を元にしたものが Transformers ライブラリ [18] に実装されている．この実装を用い，難解な語彙を制限した要約生成モデルを候補生成に用いる．

この語彙制約モデルでは，語彙の難易度は語彙に含まれる漢字の難易度で近似することができると仮定する．これは荒い近似ではあるものの，Sato ら [19] によりテキスト中に用いられている文字 1-gram によりテキストの難易度を判定するモデルが提案されているなど，近い仮定はテキスト難易度の判定でも利用されている．この仮定を元に，常用漢字のうち，より難しいと考えられる中学校で習う 1,110 字を含んでいる語彙を言語モデルが出力する語彙から制限する．

常用漢字に含まれない，より難しい漢字は語彙制約として用いない．これは，ニューステキストで使われる漢字は基本的には常用漢字，小中学校義務教育で習う 2,136 字からなっており，ニュース記事では常用漢字外の漢字は固有名詞など特別な場合に限られ，人名や地名など要約上重要かつ平易に言い換える意義の薄い場合が多いと考えられるためである．

4.3 BRIO を用いた平易な要約の生成

語彙制約モデルでは，言語モデルが難解な語彙を平易な語彙で置き換えて表現することを期待しているが，これには 2 つの問題がある．1 つ目は，置き換えた表現が元の表現より平易な表現になっている保証はないことである．例えば，「**蛍**」を「光ることで有名な 15mm ぐらいの虫」と言い換えても，読者が蛍を知らなければ理解することは難しく，かえって理解を困難にしてしまう．語彙としては平易になっていたとしても，表現としては難しくなる可能性を語彙を制限するだけでは取り除くことができない．2 つ目の問題は，誤った表現への言い換えにより，幻覚と呼ばれるような原文に書かれていない内容の要約 [4] を生成してしまうことである．語彙制約モデルでは，言語モデルが言い換え表現として適切なものを知らなかったとしても原文通りに出力することができなくなるため，モデルには不可能な言い換えを強制してしまう可能性がある．

提案手法では，このような弊害を避けつつ平易な表現への言い換えを学習するため，ランキングの学習に利用する要約候補に語彙制約モデルの生成した要約を含めることにより，語彙を制限した要約を間接的に学習に利用する．具体的には，語彙制約モデルの生成した要約と，ベースモデルの生成した要約を同じ件数ずつ要約候補として利用し，BRIO の枠組みを用いて学習を行う．

Liu らは式 1 で要約候補の並び順を決めるスコアとして ROUGE を用いているが，提案手法ではスコアとして SARI [20] を組み合わせて用いる．SARI とは文章平易化の自動評価指標であり，原文と参照文，生成文の 3 つを入力として受け取り，原文から参照文に残された語，原文から削除された語，原文から追加された語のそれぞれを 1–4 gram で評価したスコアを $\mathrm{SARI}_{\mathrm{keep}}$，$\mathrm{SARI}_{\mathrm{del}}$，$\mathrm{SARI}_{\mathrm{add}}$ とし，最終的なスコアはこの 3 つを平均したものとなる．原文との差分を積極的にスコアへ反映することにより，平易な表現への言い換えをより鋭敏に捉えるように設計されている．しかし，要約候補を並べ替えるためのスコアとして見た場合には事情が異なり，要約ではほとんどの語が削除されるため $\mathrm{SARI}_{\mathrm{del}}$ は言い換えを捉える上で有用ではなく，$\mathrm{SARI}_{\mathrm{keep}}$ は ROUGE との重複が大きい．そこで提案手法では，平易な言葉への言い換えを評価するスコアとして $\mathrm{SARI}_{\mathrm{add}}$ のみを選び，これを元々 BRIO で用いられる ROUGE のスコアと平均して学習に用いた．

5 実験

5.1 実験設定

モデルの学習には日本語要約の学習データとして Livedoor 3 行要約データ，平易な要約の学習データとして NNWE を用いた．NNWE は 2021 年 10 月 4 日 から　2022 年 12 月 20 日 までの 1,201 記事を収集し，961 件を訓練データ，各 120 件を開発，評価データとして利用した．

要約長は最大で 200 トークンとし，学習時のパラメータは訓練エポック数を除き，Liu ら [5] のパラメータを利用した．学習で利用できる事例が少ないことから，訓練エポック数は最大 300 として実験を行っている．

表 1 実験結果

手法	ROUGE–1	ROUGE–2	ROUGE–L_{sum}	SARI
ベースモデル	53.0	25.0	51.9	42.2
+BRIO (ROUGE)	**57.6**	**31.1**	**57.2**	46.4
+BRIO (SARI)	56.3	28.9	55.1	48.2
語彙制約モデル: +BRIO (SARI) +語彙制約	53.4	28.2	53.1	47.7
提案手法:	56.4	29.9	56.1	**48.5**

表 2 生成例と対応する原文の一部: 太字は語彙制約で利用できない文字を表す

原文（対応箇所のみ）	…4 回目の接種を行う方針で、対象は当面、▼ 60 **歳**以上の人のほか、▼ 18 **歳**以上の基**礎疾患**のある人か医師が重**症**化リスクが高いと判断した人としています。
語彙制約モデル	4 回目の注射は、18 以上の病気がある人か医師が病気になるリスクが高いと考えています。
提案手法	4 回目の注射は、60 **歳**以上の人や、18 **歳**以上の病気がある人か、医師が重**症**になる危険が高いと判断した人だけです。

5.2 実験結果

表 1 に実験の結果を示す．BRIO の枠組みを利用してベースモデルに加えて学習を行うことで，ROUGE，SARI を用いた場合のどちらも，ベースモデルと比べ大きく性能が向上している．BRIO と ROUGE，SARI を組み合わせた場合を比較すると，学習に用いた指標についてより性能が向上しており，最適化したい指標で学習することが有効と分かる．また，語彙制約モデルでは ROUGE と SARI のどちらも性能が悪化してしまっており，単純な手法で語彙を制限することの弊害が明らかになった．一方，提案手法では BRIO (ROUGE) とくらべると ROUGE ではやや劣るものの，BRIO (SARI) と比べて SARI，ROUGE ともにやや性能が向上しており，要約としての質を保ちつつ平易化の質を高めることができている．

5.3 議論

語彙制約モデルでは，ROUGE，SARI ともに語彙の制限を行わないモデルに比べて性能の低下が見られた．性能低下の具体例を検討するため，語彙制約モデルと提案手法の出力の一部を表 2 に示す．語彙制約モデルの出力では「**歳**」や「**疾患**」などの文字の出力が制限されているため，省略あるいは「病気」などの言葉へ置き換えられている．提案手法や原文と比較すると，「重症化リスク」を「病気になるリスク」に置き換えた直後で語彙制約モデルでは原文との対応を見失い，「注射は…リスクが高いと考えています。」と幻覚が生じてしまっていることがわかる．この例に限らず，語彙制約モデルでは妥当な言い換えであっても直後に幻覚を生成してしまう現象が多く見られた．可能性として，言語モデルにとって確度の低い言い換えが強制されることにより，原文との対応を見失いやすくなることが考えられる．

語彙制約によりどの程度幻覚が生じるかを検証するために，手法ごとに 10 例を抜き出して生成された要約の各文に幻覚が含まれているかどうかを検証した．結果を表 3 に示す．表 1 と比較すると，ROUGE などの指標の変化以上に，語彙制約モデルでは幻覚が大きく増えている事がわかる．

表 3 幻覚の出現数

手法	誤りを含む文の割合
BRIO (SARI)	29.8 %
+語彙制約	39.2 %
提案手法	22.8 %

6 おわりに

本稿では，平易な要約の生成のため，BRIO の枠組みを用いて平易化評価指標の SARI について学習を最適化するとともに，語彙を制限した要約を候補に含めて平易化の性能向上を試みた．提案手法では SARI と BRIO を組み合わせた場合と比べ，SARI で 0.3 ポイントの向上が見られた．また，語彙を直接制限する場合と比べ，幻覚の出現が抑えられていることを確認した．今後の課題として，日本語平易化で用いられるコーパスや辞書などのリソースを活用した要約生成を行い，さらなる平易化の質の向上・安定化をはかることが必要と考えている．

参考文献

[1] Shiwen Wu, Fei Sun, Wentao Zhang, Xu Xie, and Bin Cui. Graph neural networks in recommender systems: A survey. **ACM Comput. Surv.**, Vol. 55, No. 5, dec 2022.

[2] Mohammad Mehdi Afsar, Trafford Crump, and Behrouz Far. Reinforcement learning based recommender systems: A survey. **ACM Comput. Surv.**, Vol. 55, No. 7, dec 2022.

[3] 2022 年 4 月から年金制度が改正。66.27 %が知らなかったという結果に。| 株式会社日本マーケティングリサーチ機構のプレスリリース. https://prtimes.jp/main/html/rd/p/000001194.000033417.html.

[4] Joshua Maynez, Shashi Narayan, Bernd Bohnet, and Ryan McDonald. On faithfulness and factuality in abstractive summarization. In **Proceedings of ACL 2020**, pp. 1906–1919, July 2020.

[5] Yixin Liu, Pengfei Liu, Dragomir Radev, and Graham Neubig. BRIO: Bringing order to abstractive summarization. In **Proceedings of ACL 2022**, pp. 2890–2903, May 2022.

[6] 菅井内音, 西川仁, 徳永健伸. ニューステキストの要約及び平易化. NLP 2020 発表論文集, 2020.

[7] Takumi Maruyama and Kazuhide Yamamoto. Simplified corpus with core vocabulary. In **Proceedings of LREC 2018**, May 2018.

[8] Akihiro Katsuta and Kazuhide Yamamoto. Crowdsourced corpus of sentence simplification with core vocabulary. In **Proceedings of LREC 2018**, May 2018.

[9] NHK NEWS WEB EASY. https://www3.nhk.or.jp/news/easy/.

[10] Farooq Zaman, Matthew Shardlow, Saeed-Ul Hassan, Naif Radi Aljohani, and Raheel Nawaz. HTSS: A novel hybrid text summarisation and simplification architecture. **Information Processing & Management**, Vol. 57, No. 6, p. 102351, 2020.

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

[12] NHK NEWS WEB. https://www3.nhk.or.jp/news/.

[13] ライブドアニュース（livedoor ニュース）. https://news.livedoor.com/.

[14] 知範小平, 守小町. TL;DR 3 行要約に着目したニューラル文書要約. 電子情報通信学会技術研究報告, Vol. 117, No. 212, pp. 193–198, 09 2017.

[15] Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J. Liu. Exploring the limits of transfer learning with a unified text-to-text transformer. **Journal of Machine Learning Research**, Vol. 21, No. 140, pp. 1–67, 2020.

[16] megagonlabs/t5-base-japanese-web · Hugging Face. https://huggingface.co/megagonlabs/t5-base-japanese-web.

[17] J. Edward Hu, Huda Khayrallah, Ryan Culkin, Patrick Xia, Tongfei Chen, Matt Post, and Benjamin Van Durme. Improved lexically constrained decoding for translation and monolingual rewriting. In **Proceedings of NAACL 2019**, pp. 839–850, June 2019.

[18] 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 EMNLP 2020**, pp. 38–45, October 2020.

[19] Satoshi Sato, Suguru Matsuyoshi, and Yohsuke Kondoh. Automatic assessment of Japanese text readability based on a textbook corpus. In **Proceedings of LREC 2008**, May 2008.

[20] Optimizing statistical machine translation for text simplification. **Transactions of the Association for Computational Linguistics**, Vol. 4, pp. 401–415, 2016.