入出力文の関係を考慮した複数文要約でのデータ拡張

三沢翔太郎　三浦康秀
富士フイルム株式会社
{shotaro.misawa, yasuhide.a.miura}@fujifilm.com

概要

自然言語処理でのデータ拡張（Data Augmentation）は，単語の削除，挿入，置換や，学習済みモデルで文全体を言い換える手法などが一般的に用いられている．しかし複数文から成り立つ長い文章を要約する設定に対して，文章全体のトピック構成を変化させる方法は存在しない．本研究では，入出力文の関係性を担保したまま，文章全体のトピック構成を変えたデータを作成する方法を提案する．入力文章が複数文である文書要約を対象とした実験で本手法の有効性を確認した．

1　はじめに

データ拡張（Data Augmentation）は主に画像処理の領域で多用されており [1]，近年言語処理においてもデータ拡張する方法が研究されている．Wei ら [2] は文中の特定単語を同じカテゴリの別単語へ置換することで，異なる意味の文を作成する方法を提案した．Sennrich ら [3] は文章を機械翻訳で別言語に変換した後に逆翻訳で元の言語に戻すことで，同じ意味で元の表現とは異なる表現の文を作成する方法を提案した．

複数の文から成り立つ文章の要約を行う際には，個々の文を理解することに加えて，文章全体のトピックの構成や推移などを考慮する必要がある．従来のデータ拡張手法では単一の文を別の表現に書き換えることが主に想定されているため，文章が複数の小さなトピックで構成されているとみなした際に，それらのトピックの順番や数は変化させられない．そこで，このようなトピックの構成を変化させた学習データを追加することで，トピック構成の変化に柔軟なモデルを目指す．

文章内の小さなトピックは，少なくとも単一の文もしくは複数の文によって構成されていることが想定される．そのため，このような一部の文を文章全体から削除するなどの操作を行うことで，トピック構成の異なる新たなサンプルを生成する方法が考えられる．しかし，文書要約などのタスクでは入力文章と出力文章のトピック構成の一部または全てには対応関係がある．このような場合に，特別な条件付けをせずに一部の文を操作してしまうと，一部トピックの対応関係が取れない入出力文章がペアとなったサンプルが生成されてしまう．そこで，入出力文章のトピックの関係性を明確にし，関係性が強い入出力文章のトピックはその対応関係が崩れないように操作する．これにより，入力文章と出力文章で対応するトピックの関係は壊さずに，元のサンプルとはトピック構成が異なる新たなサンプルを作成することが可能となる．本研究では複数文要約を対象として，要約文章の 1 文ごとに単一のトピックを形成しているとみなした処理方法を提案する．

実験では 3 つの複数文要約タスクのデータセットを用い，学習データをダウンサンプリングした条件でデータ拡張を行った結果を評価する．longformer ベースの BART [4] に対して，トピック構成を変化させたデータ拡張を行うことで，データ拡張を行わない場合よりも性能が改善することが分かった．さらに，対応関係を考慮することで条件付けせずに文を抽出するよりも性能が高くなることが分かった．

2　関連研究

自然言語処理領域のデータ拡張は大別すると，ルールベース，中間表現ベース，モデルベースの方法が存在する．ルールベースの代表的な手法には，辞書を用いた単語の書き換えがある [2]．このほかにも，係り受け関係を考慮した手法 [5] など様々な方法が提案されている．中間表現ベースは Encoder-Decoder ベースのモデルに対して Encoder からの出力の中間表現に対して直接摂動を加える方法がある [6]．モデルベースは，機械翻訳を利用した方法が代表的で，日本語のデータを英語に変換してさらに日本語に変換し直すなどの処理を行い，元の表現とは異なる別の表現を獲得する [3]．これ以外

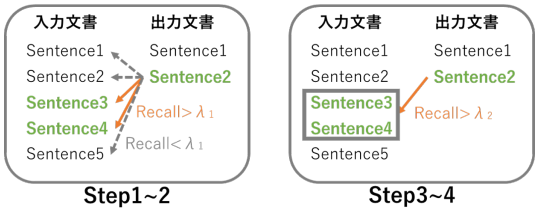


図 1　同一トピック文ペアの決定手順

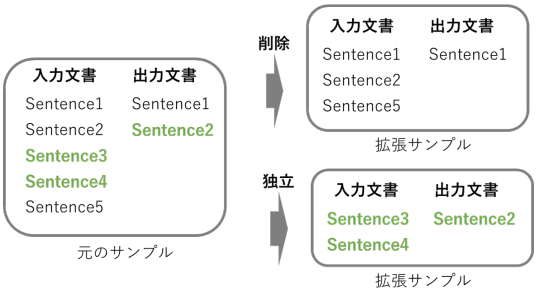


図 2　削除と独立による拡張サンプル作成の例

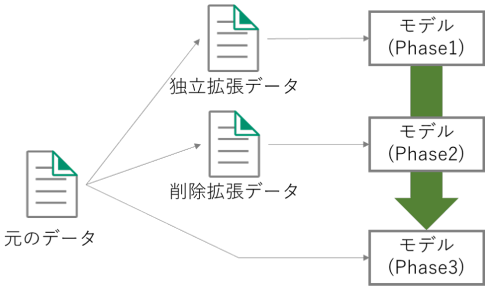


図 3　拡張データを用いた学習手順

にもパラフレーズ生成などを用いる方法も提案されている [7]．要約タスクに限定すると，Web から収集可能な CommonCrawl や Wikipedia データで学習した逆翻訳を用いるデータ拡張の方法で効果が確認されている [8]．いずれの方法も文や文章全体に対して均一な処理を行い，文章構成を大きく変える目的のものではない．

本研究と類似している手法として，対象サンプルの入力文章の変化量と出力文章の変化量を両方同時に考慮する手法がある [9]．この先行研究では，拡張したデータと元データの変化量の和を一定以内に抑えるために，入力文章と出力文章の変化量の合計値を損失関数に用いる．本研究では入出力文章のトピック構成の関係性を考慮し，入出力間の変化量が大きく変わらないように設計している点で異なる．

3　入出力文章の関係を考慮したデータ拡張

文章のトピック構成を変化させた拡張データの作成を行うために，一部の文に対して削除などの操作を行う．特別な条件付けをせずに入出力文章からそれぞれ一部の文を抽出してしまうと，入出力文章のトピックの対応関係が崩れたサンプルが生成されてしまう．そこで，入力文章と出力文章を構成する文ごとの関係性を考慮して，操作を行う入出力文のペアを決定する．以下に具体的な方法を示す．

3.1　入出力文の対応関係の取得

対応関係は既存のアライメント手法を採用することもできるが，本研究は要約タスクの特性を考慮した簡易な手法を採用した．要約文章の各文がそれぞれ単一のトピックを構成すると仮定して，要約文書の各文に類似する入力文章の文を探索する方法で同一トピック文を取得する．以下に具体的なステップを示し，図 1 に一部のステップの例を示す．

[Step1] 出力文章の 1 文ごとに，その文が正解と定義した際の，入力文章の各文の Recall を算出する

[Step2] 出力文章の 1 文ごとに，閾値 λ_1 以上の Recall を持つ入力文章の文をすべて紐づける

[Step3] Step2 で得られた出力文章の 1 文と入力文章の複数文に対し，出力文章の文を正解と定義した際の，入力文章の複数文を繋ぎ合わせた新たな文の Recall を算出する

[Step4] Step3 で算出された Recall が閾値 λ_2 以上の文の組み合わせを操作対象の候補である同一トピック文ペアとする

[Step5] Step4 で得られた同一トピック文ペアを複数組み合わせた新たなペアを作成する

以上のステップで得られた同一トピック文ペアのうち，トピック数が少ないものを優先として指定された個数まで操作する．また，実験において上記のステップで指定個数までバリエーションが用意できないサンプルに対しては，ランダムに文を選択してサンプルを増やした．

3.2　同一トピック文ペアに着目したサンプル生成

上記の方法で獲得した同一トピック文ペアに含まれる文を同時に操作することで入出力文章のトピックの対応関係を保つことができる．本手法では，同一トピック文ペアに含まれるすべての文に対して以下のいずれかの操作を行うことで新たなサンプルを獲得する．図 2 に例を示す．

[削除]：同一トピック文ペアに含まれるすべての文を，元の入出力文章から削除する

[独立]：同一トピック文ペアに含まれる入力文と出力文を，新たなサンプルとして独立させる

表 1　各データセットの平均単語数

	PubMed	BillSum	BigPatent
入力文章	3,063	1,818	6,527
出力文章	205	289	97

表 2　同一トピック文ペアの特徴

条件	統計量
単一サンプルにおける平均のペア数	5.5
サマリ 1 文に対応する入力文の平均数	15
ペアが 1 つ以上あるサンプルの割合	95%
サマリの文でいずれかの入力文と関係する確率	70%
入力文章の文うち，いずれかのペアに入る確率	40%

3.3　拡張データを利用した学習

拡張したデータは，元データと混ぜて学習する方法と，事前学習として使う方法が考えられる．事前実験の結果，本研究では事前実験として用いる方法を採用した．ここで，同一トピック文ペアを独立させたサンプルは入出力ともに文数が短く，入力文章は出力文章と同一のトピックに限定されており，タスクの難易度が低い．一方ペアを削除したサンプルは文数が多い傾向にあるためタスクの難易度が高くなる．これらを考慮して，事前学習のステップをさらに分割し，独立で学習した後に削除で学習する方法を採用した．学習の概略を図 3 に示す．

4　実験

4.1　データ

入力文章の系列が長い文書要約タスクとして，医療分野の論文本文を要約する PubMed [10] を用いた．さらに複数のデータセットでの効果を確認するため一部の実験条件に限定して，米国議会及びカリフォルニア州法案の要約である BillSum [11] と米国の特許要約の BigPatent [12] を用いた．データ拡張の効果を明らかにするために，いずれのデータセットも学習データはランダムサンプリングで 100 件に，検証用データとテストデータをそれぞれ 200 件に削減した．各データセットの平均単語数を表 1 に示し，$\lambda_1 = 0.3, \lambda_2 = 0.7$ とした場合の PubMed における同一トピック文ペアに関する統計値を表 2 示す．

4.2　比較手法

実験では，以下の手法の性能を比較した．

Original：データ拡張を行わない方法

RandDel： ランダムな確率 (p) で選択した文を文章から削除したサンプルで事前学習する．事前実験より p は 0.1 とした．なお，ランダムの独立と削除は p の値で決まるため，独立の比較は行わない．

PairDel： 同一トピック文ペアを元のサンプルから削除した新たなサンプルで事前学習する

PairInd： 同一トピック文ペアを独立して新たなサンプルとして事前学習する

PairInd-RandDel： PairInd の事前学習を行った後に，RandDel で追加学習する

PairInd-PairDel： PairInd の事前学習を行った後に，PairDel で追加学習する

表 3　実験結果 PubMed

Model	ROUGE-1-F	ROUGE-L-F
Original	26.15	23.61
RandDel	33.47	30.23
PairInd	32.93	29.57
PairDel	32.90	29.57
PairInd-RandDel	34.36	30.84
PairInd-PairDel	**34.62**	**31.14**

表 4　実験結果 BillSum

Model	ROUGE-1-F	ROUGE-L-F
Original	29.94	25.69
RandDel	41.77	36.65
PairInd-PairDel	**45.53**	**40.55**

4.3　条件

事前学習済みモデルとして longformer ベースの BART [4] を利用した．評価指標は ROUGE-1-F と ROUGE-L-F を用いた．ハイパパラメータは検証用データに対する ROUGE-1-F が最良となるものを選択した．また，最適なエポック数も検証用データを利用している．エポック数は 100 とし，検証用データに対する損失が 5 エポック改善しない場合は早期終了した．文章の文分割には nltk [13] を用いた．

4.4　結果と考察

PubMed の実験結果を表 3 に示す．RandDel でも Original より性能が高いことから，入出力文章間のトピックの対応関係が崩れる可能性を考慮しても，トピック構成を変化させる効果があるといえる．PairInd と PairDel は Original よりは性能が高いものの，RandDel よりは性能が低い．RandDel の p が 0.1 であったために入出力文章にあるトピックの対応関係を崩すほどの変更は生じず，PairInd や PairDel

表 5　実験結果 BigPatent

Model	ROUGE-1-F	ROUGE-L-F
Original	22.65	19.34
RandDel	33.88	29.75
PairInd-PairDel	**35.97**	**31.63**

表 6　学習データ数が変化した際の PubMed 実験結果

Data Size	Original	PairInd-PairDel
50	24.49	33.90
100	26.15	34.85
500	30.53	35.08

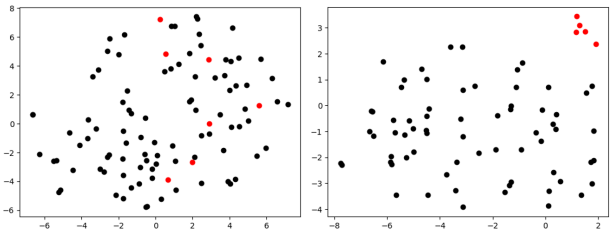
よりも RandDel の方がトピックを限定せず多くのパターンを学習できた効果が強く出たためと考えられる．PairInd-PairDel は RandDel よりも高い性能を示した．これは PairInd-PairDel が RandDel と比較して入出力文章の対応関係を考慮しつつ，PairInd や PairDel などの単独の作成方法と比較してトピックを限定せず学習できた効果があったと考えられる．さらに PairInd-PairDel は後段をランダムに置き換えた類似手法である PairInd-RandDel よりも性能が高い．どちらも 1 段階目でトピックを限定した難易度が低いデータで学習し，2 段階目で元データに近いデータで学習しているが，PairDel の方が PairInd と相乗効果が大きかった考えられる．

BillSum と BigPatent の結果を表 4 と 5 に示す．いずれのデータセットでも PairInd-PairDel の性能が最も高いことが分かる．このことから，入力文書が長い複数の文書要約のデータセットに対して提案するデータ拡張手法の効果があることが確認できた．

4.5　分析

[データ数に対する影響]：実験では学習データ数を 100 として評価したが，表 6 にデータ数を 50 と 500 に変化させた場合の結果を示す．なお本実験では計算時間の都合で拡張データ数を減らした． この結果から，データ数 50 および 500 の場合でもデータ拡張の効果があることが分かった．一方で，データ数 500 の場合は他と比較して効果が小さい．このことから本手法は，一般的なデータ拡張と同様に，データ数が少ない場合に効果が大きいと推測できる．

[同一トピック文ペアの傾向]：ランダムに選択したある 1 つのサンプルの入力文章について，いくつかの文をランダムに選択した場合 (Random) と，出力文章のある 1 つの文と同一トピックである入力文を選択した場合 (SingleTopic) とで分布の違いを可視化



(a) Random (p=0.2)　(b) SingleTopic

図 4　あるサンプルで選択された文の分布の比較

表 7　実験結果 選択データ

Model	ROUGE-1-F	ROUGE-L-F
Random (p=0.2)	12.97	11.09
SingleTopic	31.09	27.92

した．ランダムに選択する割合は，同一トピック文ペアに含まれる文の数とほぼ同数になるように 2 割に設定した．各文を SentenceBERT [14] でベクトル化し，t-SNE [15] で次元圧縮した結果を図 4 に示す．図において赤でプロットした点は選択された文を示し，黒でプロットした点は選択されていない文を示す．この結果より Random は全体から均一に選択されており，選択された文，選択されていない文，入力文章全体で大きな傾向の違いはない．一方，SingleTopic は文章全体の中で一部に集中し，トピックが限定的であると考えられる．結果として，選択されていない文も一部のトピックが一切含まれていない状態であり，選択された文，選択されていない文，入力文章全体で傾向が異なると考えらえる．

また，上記の方法で選択したサンプルで構築したデータがタスクとして成立するかを検証する．学習データとテストデータ共に図 4 の赤点の方法でサンプリングし，学習と評価を行った結果を表 7 に示す．Random では性能が著しく低く，PairInd の事前学習と類似した SingleTopic では性能が比較的高い．これらのことから PairInd はトピック構成が異なるサンプルの学習を実現しつつ，学習が成立しており，事前学習の効果があると考えられる．

5　おわりに

入出力文章の対応関係を担保しつつ，文章のトピック構成を変えるデータ拡張手法を提案した．入力文書の系列長が長い 3 つの文書要約のデータセットで手法の有効性を確認した．本研究では既存のデータ拡張手法と比較していないが，手法の組み合わせも可能なので，今後は既存手法との比較に加えて組み合わせた場合の効果検証などを行う．

参考文献

[1] Connor Shorten and Taghi M. Khoshgoftaar, *"A survey on Image Data Augmentation for Deep Learning"*, *Journal of Big Data*, Vol. 6, no. 60, 2019.

[2] Jason Wei and Kai Zou, *"EDA: Easy Data Augmentation Techniques for Boosting Performance on Text Classification Tasks"*, In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pp. 6382–6388, 2019.

[3] Rico Sennrich, Barry Haddow, and Alexandra Birch, *"Improving Neural Machine Translation Models with Monolingual Data"*, In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 86-96, 2016.

[4] Available online: https://huggingface.co/hyesunyun/update-summarization-bart-large-longformer

[5] Gözde GüL Şahin and Mark Steedman, *"Data Augmentation via Dependency Tree Morphing for Low-Resource Languages"*, In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pp. 5004–5009, 2018.

[6] Jiaao Chen, Zichao Yang, and Diyi Yang, *"MixText: Linguistically-Informed Interpolation of Hidden Space for Semi-Supervised Text Classification"*, In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pp. 2147–2157, 2020.

[7] Ashutosh Kumar, Satwik Bhattamishra, Manik Bhandari, and Partha Talukdar, *"Submodular Optimization-based Diverse Paraphrasing and its Effectiveness in Data Augmentation"*, 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. 3609-3619, 2019.

[8] Alexander Fabbri, Simeng Han, Haoyuan Li, Haoran Li, Marjan Ghazvininejad, Shafiq Joty, Dragomir Radev, and Yashar Mehdad. *"Improving Zero and Few-Shot Abstractive Summarization with Intermediate Fine-tuning and Data Augmentation"*. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pp.704–717, 2021.

[9] Xinyi Wang, Hieu Pham, Zihang Dai, and Graham Neubig, *"SwitchOut: an Efficient Data Augmentation Algorithm for Neural Machine Translation"*. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pp. 856–861, 2018.

[10] Arman Cohan, Franck Dernoncourt, Doo Soon Kim, Trung Bui, Seokhwan Kim, Walter Chang and Nazli Goharian, *"A Discourse-Aware Attention Model for Abstractive Summarization of Long Documents"*, In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 2 (Short Papers)*, pp. 615–621, 2018.

[11] Anastassia Kornilova and Vlad Eidelman, *"BillSum: A Corpus for Automatic Summarization of US Legislation"*. arXiv preprint arXiv:1910.00523 [cs.CL], 2019.

[12] Eva Sharma, Chen Li and Lu Wang, *"BIGPATENT: A Large-Scale Dataset for Abstractive and Coherent Summarization"*, In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pp. 2204-2213, 2019.

[13] Steven Bird and Edward Loper, *"NLTK: The Natural Language Toolkit"*, In *Proceedings of the ACL Interactive Poster and Demonstration Sessions*, pp. 214-217, 2004.

[14] Nils Reimers and Iryna Gurevych, *"Sentence-BERT: Sentence Embeddings using Siamese BERT-Networks"*, In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pp. 3982–3992, 2019.

[15] Laurens van der Maaten and Geoffrey Hinton, *"Visualizing Data using t-SNE"*, *Journal of Machine Learning Research*, Vol. 9, no. 86, pp. 2579–2605, 2008.