不要文削除によるデータ拡張に関する研究

大内 智仁[†]・田伏 正佳[†]

近年インターネット上の情報量は指数関数的に増加している。そのため、自動要約システム技術は必要不可欠なものとなってくると思われる。自動要約システムを構築するには要約コーパスが必要となる。しかし、多量の要約コーパスを作成するには人手が必要となりコストがかかってしまう。そこで、本研究では自動要約システムにおけるデータ拡張として、記事に対して、最も重要度の低い文を除去する手法を提案する。本研究では、Pointer-Generator モデルにおいて提案手法の効果を検証した。また、本研究の比較対象として、文書分類において用いられたデータ拡張手法である EDA (Easy Data Augmentation Techniques) や、Luhn, LexRank を用いた手法で実験を行った。Pointer-Generator モデルで用いたコーパスは CNN/Daily Mail dataset であり、トータルで、287,226 記事存在する。本研究では、287,226 記事の他に、57,000 記事、28,000 記事において比較実験した。結果は、EDA や Luhn, LexRank を用いた手法では拡張せずに元の記事だけで学習する手法（拡張なし手法）よりも悪くなることがあったが、提案手法は全ての記事数において拡張なし手法よりも良い結果となった。

キーワード：自動要約, data augmentation, Pointer-Generator モデル, EDA

The Effectiveness of Data Augmentation by Removing Unimportant sentence

TOMOHITO OUCHI[†] and MASAYOSHI TABUSE[†]

In recent years the amount of information on the Internet has increased exponentially. Consequently, automatic article summarisation technology will be indispensable. In this study, we propose a data augmentation method for an automatic summarisation system. The proposed method removes the least important sentence in an article. We used a topic model to determine the importance of sentences in articles. The Luhn and LexRank methods were used as comparative methods for determining the importance of sentences in articles. Additionally, we used Easy Data Augmentation (EDA) techniques as the comparison method for this study. EDA is a data augmentation method applied to document classification. A comparative experiment was performed using input datasets with 28,000, 57,000, and 287,226 articles. The Luhn and LexRank methods always produced the worst results, while EDA sometimes performed worse than the baseline method without a data augmentation. The proposed

[†] 京都府立大学, Kyoto Prefectural University

method performed the best in all cases.

Key Words: *Automatic Summarization, Data Augmentation, Pointer-Generator Model, EDA*

1 はじめに

近年、全世界のインターネット上の情報量は指数関数的に増加しており (Reinsel et al. 2020), 2010 年から 2024 年まで年平均 29% で成長し, 2024 年には 143ZB に達すると予想されている. また, テキストデータに関しても, 2018 年の全世界の Web サイト数は 16 億程度であったのに対して 2021 年現在においては 18 億程度 (Internet Live Stats 2021) と, 3 年で 2 億程度増えていることからこれからもどんどん増えていくことが予想される. このような状況のなかで, インターネットの情報を取捨選択する必要性は高まっており, 自動要約の技術が必須となってくる. しかし, 実用段階に至っている自動要約システムの多くが, 元の文章から要約文を抽出するだけに終わっている. これは抽出型要約といわれており, 問題点として, 指示語の指す情報の不明確さ, 不必要な接続詞の存在, 文同士の不自然なつながりなどがあげられる. しかし, 抽出型要約システムによっては, 要約をする元の文章だけで学習できる場合があり, 必ずしも要約の文章が必要とならないことがある. その場合, 抽出型要約は, データ整備のコストはあまりかからない. 一方で, 抽出型要約に対して生成型要約というものがある. 生成型要約の多くは, Encoder-Decoder モデルを基本として作られている. この方法でシステムを作る際, 学習には, 要約をする元の文章と要約の文章のペアが多く必要となる. さらに, 例えば, 日本語の論文の要約システムを作るとすれば, 日本語の論文とその要約文のデータセットが必要となる. このシステムの学習に, ニュース記事の要約データセットを用いることはできない. なぜなら, 文章の体裁が全く違うからである. このように, 特定のドメイン・言語で大規模データセットを作るのは困難である. 本研究では, 様々な長さの文章に対して複数の要約文を生成する要約システムに焦点を当て, 学習のためのデータセットに対してデータ拡張の効果を検証した. まず, Encoder で, 入力文章を固定長のベクトルに変換し, Decoder でその固定長のベクトルから, 一単語ずつ生成していく. 生成する際には, 一単語ずつ, 学習に使った語彙の中から選んで要約を作る. 初期のころは, Decoder 側で使われる語彙集合から, 単語を選んでしたが, のちに, Copy Mechanism (Gu et al. 2016) が提案され, Encoder 側の単語をコピーして生成する機構が提案された. これによって, 特定の文章にしかでてこない単語 (固有名詞など) を生成することを可能にした. また, 生成された要約文に同じフレーズが繰り返される問題を解決するために Coverage Mechanism (Tu et al. 2016) が提案された. 本研究では, データ拡張の検証としてこの Copy Mechanism と Coverage Mechanism を組み合わせた Pointer-Generator モデル (See et al. 2017) を使用した. 学習のためのデータセットとして, 日本語では, Livedoor News を用いた三

行要約がある (小平 2018) が, 本研究では, 様々な長さの文章に対して複数の要約文を生成する要約システムでのデータ拡張の検証を目指しており, 採用しなかった. また, 出力長制御を考慮した見出し生成モデルのための大規模コーパス (JNC, JAMUL) (人見 他 2019) では, 見出し生成を目的としており, これも採用しなかった. 英語では, CNN/Daily Mail Dataset (Hermann et al. 2015) や, Cornell Newsroom (Grusky et al. 2018), Gigaword (Napoles et al. 2012) などのデータセットが整備されている. このデータセットは, 人手の要約が付いており, 様々な長さの要約をする文章があり, 複数文要約にもなっている. そこで, 本研究では, データセットとして CNN/Daily Mail Dataset を用い, 少ないデータでもデータ拡張することで効果的に自動要約システムを作ることができないかということに着目した. 既存の自然言語処理におけるデータ拡張手法を以下に示す.

1. 同義語・類義語で置き換える
2. 類似度を計算して置き換える
3. 反意語で置き換える
4. 文章内の語と語を入れ替える
5. ランダムに削除する
6. Back Translation を用いて文章を水増しする

本研究においてデータ拡張の比較対象として用いた EDA (Easy Data Augmentation Techniques) (Wei and Zou 2019) は, 1 と 4 と 5 に該当する. 2 の類似度を計算するには, 教師なしクラスタリングや単語埋め込み行列などを使う. 単語埋め込み行列には, word2vec (Mikolov et al. 2013), GloVe (Pennington et al. 2014), fastText (Bojanowski et al. 2017) などがある. 6 の Back Translation (Edunov et al. 2018) は, 機械翻訳におけるデータ拡張手法である. しかし, いずれの手法も文書分類システムや機械翻訳システムで効果が確かめられているだけであり, 自動要約システムでは効果を検証されていない. 一方, 画像処理分野では, 少ないデータでも効果的に画像処理ができるデータ拡張の方法がある. 具体的には, 以下の 9 つがある.

1. 水平・垂直方向に画像をシフトする
2. 水平方向・垂直方向に画像を反転させる
3. 回転させる (回転角度はランダムの場合あり)
4. 明度を変える
5. ズームインする・ズームアウトする
6. 画像の一部をくり抜く, 削除する
7. 背景色を変える
8. 背景を置き換える
9. Mixup・CutMix

ここで説明する Mixup と CutMix では, カテゴリを表すラベルを 1 (あり) か 0 (なし) で

はなく、連続量で表す。例えば、犬であるかないか (1 か 0) ではなく、犬である可能性が高い (例えば, 0.8) という風に表す。Mixup は、二枚の画像とラベルを組み合わせて一つのデータに合成する (Zhang et al. 2018)。例えば、犬 0.4 猫 0.6 といったラベルの画像は各画素値を犬 0.4 猫 0.6 の比率で合成した画像になる。CutMix は、複数の画像の一部を切り取ってつなぎ合わせて 1 枚の入力画像にするデータ拡張である (Yun et al. 2019)。ラベルの表し方は、合成した画像中のそれぞれのラベルの付いた画像の面積比となる。例えば、犬と猫の面積比が 1:3 であるならば、犬 0.25 猫 0.75 というラベルとなる。Mixup や CutMix 以外の手法では、元画像だけでデータ拡張が行える。我々は、これらの手法の内、背景を置き換えるという拡張手法に着目し、データ拡張手法を提案する。具体的には、要約する文章において、不要文を取り除き、文章の大意を損なわない拡張文章を作ることでデータ拡張を行う。

2 関連研究

自動要約システムに関連する研究は、大別して抽出型要約と生成型要約の二つの種類がある。以下の三つの節でそれぞれの要約とデータ拡張の関連研究について説明する。

2.1 抽出型要約

一つ目は、抽出型要約である。これは、要約元の文章から重要な部分を抽出して行う要約で、文単位のものから、フレーズ、単語単位まである。複数文書要約においては、同じような内容の部分抽出しないように、冗長性を考慮した要約にする必要がある。抽出型要約でよく使われる手法としては、Graph Base 手法と Feature Base 手法、Topic Base 手法などがある。

Graph Base 手法でよく用いられるアルゴリズムとして TextRank (Mihalcea and Tarau 2004) が挙げられる。これは、Google 検索で用いられている PageRank 手法 (Page et al. 1998) をもとに、自然言語に適用したものである。

PageRank では、あるページが人気の高いページから多くリンクされていればいるほどそのページは重要度が高いとしている。TextRank は相互に関連度の高い語句は重要度が高いという仮定に基づき語句を選択する手法である。この TextRank の派生に LexRank (Erkan and Radev 2004) がある。本研究において提案手法に対する比較手法として LexRank を用いたデータ拡張を行う。以降、この比較手法を DA-LexRank と呼ぶ (この DA は Data Augmentation からとった)。

Feature Base 手法は、文の特徴を定義して、その特徴による重みづけを行うことで文選択を行うものである。本研究では、単語の位置と頻度の情報を元に文の重要度を測る手法 (Luhn 1958) を用いてデータ拡張を行い提案手法に対する比較手法とする。以降、この比較手法を DA-Luhn と呼ぶ。

Topic Base 手法は、トピックモデル (Blei et al. 2003) を用いて、各文においてトピックの出

現確率が高いものから文を選択するという手法である。文におけるトピック分布は、文において TF-IDF 値を求め、それから、トピック分布の事前分布であるディリクレ分布を推定することで得られる。本研究では、重松らの研究(重松, 小林 2012)をもとにトピックモデルを用いて文の重要度を決めた。

2.2 生成型要約

二つ目は、生成型要約である。これは、要約元の文章と要約の対応関係を学習させ、未知の文章が入力されたときに、要約を生成するというものである。ここで、生成するというのは、入力文章から文、フレーズ、単語を抽出するのではなく、一から要約を生成するというものである。生成の仕方は、生成確率の高い単語をひとつずつ選んでいき、文章の終わりを示すトークンが選ばれたときに、生成を終了するというものである。抽出型要約に比べて自然な文が生成される可能性が高いが、二つの課題がある。

一つ目の課題は、要約元の文章と要約のペアになったデータ作成の困難さである。要約元の文章だけなら用意できるが、要約は、人による介入が必要となりコストがかかってしまう。人によっても要約の仕方は変わり、また、どれくらいの長さにするかも変わってくる。そのため、厳密にデータを作ろうとすると、要約の基準を定めなければならない。また、機械翻訳では、翻訳元の文数と翻訳先の文数は等しくなるが、自動要約では、要約の文数に対して、要約元の文数はその数倍の量になることが多い。したがって、できるかぎり少ないデータで効率よく自動要約システムを作ることが求められる。

二つ目の課題は、自動要約タスクの定義の曖昧性である。機械翻訳においては、入力された文の内容を過不足なく翻訳することがタスクの定義の正解として設定できるが、自動要約に関しては、どの程度要約するのがタスクの定義の正解なのかはタスクの目的による。実用段階となると、そのタスクで求められている程度の要約ができればいいということになる。このように、タスクによって求められる正解が変わってくるという意味で、データ作成の困難さがある。その難しさから、生成型要約の研究が始まった当初は、ニュース記事の1文目を入力してタイトルを生成するというタスクに限定して研究が行われた。このタスクは通称見出し生成タスクと呼ばれる。また、Encoder-Decoder モデルを初めて導入した論文に ABS (Attention Based Summarization) がある (Rush et al. 2015)。これを皮切りに、深層学習を用いた自動要約システムの研究が盛んになった。そして、2017 年には、CNN/Daily Mail Dataset を用いた Pointer-Generator モデル (See, Liu, and Manning 2017) が開発された。今までは、Encoder-Decoder モデルに Attention モデルを加えたものが主流であったが、記事特有の固有名詞などがうまく生成できないという問題があった。そこで、Encoder 側の単語をコピーするか、Decoder 側の語彙から選ぶかを確率で決めるという Copy Mechanism が開発された (Gu et al. 2016)。しかし、このメカニズムで生成された要約は、同じフレーズを繰り返してしまうという問題があった。これを解決するため

に, 以前選ばれた単語を選ばれにくくする Coverage Mechanism (Tu et al. 2016) というものが開発された.

2.3 既存のデータ拡張手法

自然言語処理におけるデータ拡張に関する研究はいくつかある. 多くは, 文書分類に適用したもので, そのうちの 하나가 EDA (Wei and Zou 2019) である. また, 機械翻訳におけるデータ拡張手法として, Back Translation (Edunov et al. 2018) がある. Back Translation とは, まず, 入力ソースと出力ソースのデータから, 翻訳モデルと逆翻訳モデルを作る. 次に, 入力ソースが存在しない出力ソースから, 逆翻訳モデルで, 入力ソース (疑似データ) を作るというものである. したがって, 本来のデータに加えて, 逆翻訳した入力ソースと出力ソースができ, データ拡張を行うものである. Edunov らの研究 (Edunov et al. 2018) によって機械翻訳のベンチマークを大幅に更新したことが確認された.

3 データ拡張

この章では, 本研究において用いたデータ拡張について説明する. 3.1 節では, 本研究で提案した手法を, 3.2 節では, それに対して比較する対象の手法を説明する.

3.1 提案手法

3.1.1 トピックモデル

潜在的意味解析は文書のもつ話題や分野など「潜在的な意味カテゴリ」を解析する手法 (Blei et al. 2003) である. この「潜在的な意味カテゴリ」を潜在的なトピック, または単にトピックと呼び, 文書内の複数の単語の共起性によって創発され则认为. トピックモデルは, 文書に隠れた潜在的なトピックを推定するモデルである. 特に LDA (Latent Dirichlet Allocation) は, 一つの文書に複数の潜在的なトピックが存在すると仮定し, そのトピック分布をディリクレ分布としてモデル化する. 単語などを表層的と表現するならば, トピックは, 単語と違って表面には現れないので潜在的と表現できる. 潜在的なトピック分布にディリクレ分布を仮定することから, 潜在的ディリクレ配分法と呼ばれる.

例えば, 「貴景勝 実戦稽古再開 カド番で臨む7月場所 自分の攻めを意識」という文があったとする. 人はこの文を見て相撲に関する文書であることを理解できる. 文書中に相撲という言葉は出てこないにもかかわらずである. よって, 単語は単に存在しているのではなく, 潜在的なトピックから生成され, それぞれの文書はその文書がもつ複数のトピックに応じて出現しやすい単語集合で構成されていると考える.

3.1.2 文の重要度の決め方

トピックモデルを用いた文の重要度の決め方は、重松らの研究 (重松, 小林 2012) を参考にした。今回用いたトピックモデルには、LDA 手法を用いた。また、本研究で用いたモジュールは gensim のバージョン 3.7.3 を利用した。トピックモデルは、一つの文書が複数のトピックからなることを仮定した言語モデルの一つである。また、それぞれのトピックは出現単語分布を持つ。文 s の重要度 i_s の決め方は次のようになる。

1. 文を構成している単語 w のトピック t における出現確率 $p(w|t)$ と文 s におけるトピック t の出現確率 $q(t|s)$ の積を算出
 2. 文を構成している全ての単語 W で総和
 3. 文長の平方根 $\sqrt{|s|}$ で割る
 4. 全てのトピック T において総和する
- 式で表すと以下の通りとなる。

$$i_s = \sum_{t \in T} \sum_{w \in W} \frac{p(w|t)q(t|s)}{\sqrt{|s|}} \quad (1)$$

この手法は、文を構成している主要なトピックにおいて出現確率の高い単語が多く含まれているときにスコアが高くなる。

3.1.3 処理の手順

我々は、画像処理におけるデータ拡張の考え方をもとに、文章から不要文を削除するデータ拡張の方法を用いてデータ拡張の効果を検証した。画像処理におけるデータ拡張の一種である「背景の差し替え」における背景を、自動要約システムにおける入力文章の不要な部分に置き換え、入力文章から不要な部分を取り除いたものを新たな入力文章とした。入力文章から不要文を取り除くことは、画像処理における背景の除去（無背景に置き換えること）に対応する。他の背景に置き換える、つまり不要文を他の文に置き換えることは、今後の課題として検討すべきである。また、この方法で生成される拡張データは、不要文を除くだけなので意味的に破綻した文章が生み出される可能性は少ないといえる。本研究の提案手法は、具体的に、以下の手順をとる。

1. 英語の Wikipedia のデータからトピックモデルを作る。Wikipedia のデータを選んだ理由は多くの項目を解説しており、内容に偏りが少ないためである。今回の実験では、トピック数を 100 と設定し、20 文書以上かつ全記事 (287,226 記事) の 1 割以下の文書で使われた語に限定した。トピック数を決める際に Chang らの研究 (Chang et al. 2009) を参考にした。この研究では、Wikipedia においてトピック数を 50, 100, 150 とおき LDA 手法を含む三つの手法で実験している。LDA 手法においてはトピック数が 150 のときが

```

1. make topic model
2.  $r \leftarrow \text{None}$ 
3.  $i \leftarrow 0$ 
4. For  $s \in S$ 
     $i_s \leftarrow \text{calculate score}(s, \text{topic model})$ 
    If  $i < i_s$ 
         $r \leftarrow s$ 
         $i \leftarrow i_s$ 
5.  $E \leftarrow S \text{ remove } r$ 
6. Output  $E$ 

```

図 1 提案手法の疑似コード

最もパフォーマンスが高かった。しかし、150 で実験すると各文が持つトピックの重なりが少なくなり、各文の重要度がうまく計算できないと思われたため本研究ではトピック数を 100 に減らした。

2. 1. で作成したトピックモデルを用いて、各文章 S における各文 s の重要度 i_s をスコアとして算出する

3. 重要度 i_s の最も低い文を取り除き、拡張データ E とする

アルゴリズムを図示すると、図 1 のようになる。

3.2 比較手法

本研究では、提案手法の比較対象として、EDA (Wei and Zou 2019) と、Feature Base 手法である、単語の位置と頻度の情報を元に文の重要度を測る手法 (Luhn 1958) を用いたデータ拡張手法 (DA-Luhn) と、Graph Base 手法である LexRank を用いたデータ拡張手法 (DA-LexRank) を用いた。以下では、EDA と、DA-Luhn, DA-LexRank について説明する。

3.2.1 EDA

EDA は極性分類やレビュー推定などの文書分類におけるタスクにおいて、データ拡張を行った方法である。中間ベクトルにノイズを加えるなどのデータ拡張ではなく、記事自体を変えるデータ拡張として、EDA を比較手法として選んだ。拡張手法の概要は以下となる。

- (1) Synonym Replacement (SR)
ある単語を同義語に置き換える
- (2) Random Insertion (RI)
ある単語の同義語を記事のランダムな位置に挿入する
- (3) Random Swap (RS)
ランダムな二つの単語を入れ替える

(4) Random Deletion (RD)

ランダムな単語を削除する

EDA では、上記の四つの手法を行う頻度を記事に使われている延べの単語数にハイパーパラメータ α を掛けて求めている。この研究では、訓練記事数が 5,000 以上だと α の値は 0.1 に設定しているので本研究でもそれに倣った。また、EDA の論文 (Wei and Zou 2019) によると、一つの記事に対する拡張する記事の数を 4 にしたときが最も良い結果が出たということで、本研究でもそれに倣った。EDA を Pointer-Generator モデル (See et al. 2017) に適用して、提案手法との比較検討を行った。本研究では、EDA の各手法 (SR, RI, RS, RD) と、文章ごとに各手法をランダムに選択した手法（以降、この手法を EDA と呼ぶ）を比較手法とした。

3.2.2 DA-Luhn

単語の位置と頻度の情報を元に文の重要度を測る手法 (Luhn 1958) では、まず、対象とする文章のうち、ストップワードでない頻出単語のトップ 100 単語を重要単語とし、重要単語の文章中の位置を調べる。次に、重要単語同士の距離が 4 以下のものを一つのクラスタとしてクラスタリングする。そして、クラスタ内の重要単語数の二乗の値を、クラスタの最初の単語と最後の単語の距離で割った値を、そのクラスタのスコアとする。最終的に各クラスタのスコアの最大値がその文のスコアになる。そのスコアを用いて、不要文（最小スコアの文）を除去するデータ拡張手法が DA-Luhn となる。

3.2.3 DA-LexRank

LexRank は、多くの文と類似する文や重要な文と類似する文は重要な文と考え、文章同士の TF-IDF ベクトルのコサイン類似度から遷移確率行列を求め、遷移確率行列に滞在確率ベクトルをかけたものが滞在確率ベクトルになるような滞在確率ベクトルを求める。このときの滞在確率ベクトルが文の重要度となる。そして、文の重要度を用いて、不要文（最小スコアの文）を除去するデータ拡張手法が DA-LexRank となる。今回の実験では、コサイン類似度が 0.1 以上のものは全て 1 とし、更新前の滞在確率ベクトルと更新後の滞在確率ベクトルの差のノルムが 0.1 以下になったときを収束条件とした。

4 Pointer-Generator モデル

この章では、本研究において使用した自動要約システムである Pointer-Generator モデル (See et al. 2017) について説明する。そのために必要な知識として、Encoder-Decoder モデル、Attention モデルについても説明する。

4.1 Encoder-Decoder+Attention モデル

図2がLSTM Encoder-Decoder+Attention モデルの概要図になる。LSTM (Long Short Term Memory) は (Hochreiter and Schmidhuber 1997) を参照されたい。

$\mathbf{x}(\tau)$ は τ 番目の単語を指す。Encoder は, T 個の単語から成り立っており, Decoder 側で t 番目の単語の隠れベクトル $\mathbf{h}_{dec}(t)$ を算出する場合を表している。 $\mathbf{h}_{dec}(t)$ の算出の仕方は, 以下の式による。

$$\mathbf{h}_{dec}(t) = f(\mathbf{h}_{dec}(t-1), \mathbf{y}(t-1), \mathbf{c}(t)) \quad (2)$$

$\mathbf{c}(t)$ は以下の式によって求められる。

$$\mathbf{c}(t) = \sum_{\tau=1}^T \alpha_{\tau}(t) \mathbf{h}_{enc}(\tau) \quad (3)$$

つまり, この $\alpha_{\tau}(t)$ が, Encoder 側のそれぞれの単語をどれくらいデコーダー側に伝えるかの割合を表している。

$$\mathbf{w}_{\tau}(t) = f(\mathbf{h}_{dec}(t-1), \mathbf{h}_{enc}(\tau)) \quad (4)$$

なる $\mathbf{w}_{\tau}(t)$ を用いて,

$$\alpha_{\tau}(t) = \frac{\exp(w_{\tau}(t))}{\sum_{\rho=1}^T \exp(w_{\rho}(t))} = \text{softmax}(w_{\tau}(t)) \quad (5)$$

と表される。この $\mathbf{w}_{\tau}(t)$ が最適化されるべきパラメータである。

図2 LSTM Encoder-Decoder+Attention モデルの概要図

4.2 Pointer-Generator モデル

Pointer-Generator モデルについて説明する. Pointer-Generator モデルは Encoder-Decoder + Attention モデルに, Copy Mechanism と Coverage Mechanism を追加したモデルである. 概要図は図 3 の通りとなる.

このモデルは以下の手順で処理を行う.

1. 入力テキスト (Source Text) を, 双方向 LSTM でエンコードする. 可変長の長さのデータを固定長のベクトルになおす. (赤色箇所)
2. EncoderLSTM で Attention Mechanism (青色箇所) を用いて, 単語生成確率 (黒色箇所) を求める.
3. デコード部分 (Partial Summary) における単語の生成確率分布 (Vocabulary Distribution) に, エンコード部分における Attention の値 (Attention Distribution) を加算する (Copy Mechanism) (Final Distribution). Copy Mechanism によって, モデルの出力語彙には含まれない単語を生成することが可能となる.
4. Final Distribution から最終的に生成する単語を選択
5. 2~4. を繰り返して, 逐次的に単語を生成. 文章の終わりを表すトークンが選ばれたら, そこで, デコード (要約) は終了する

図 3 Pointer-Generator モデルの概略図 (See et al. 2017)

Coverage Mechanism を使うときは, 2. の時点において以下の処理を行う.

以前 Attention の値が大きかったところは, Attention 値を減衰させる. これには, 現在注目している Encoder の単語以前の Attention を総和した Coverage Vector を導入する. また, それを Attention Distribution (Encoder に入力された全語彙の Attention 分布) の式に加える. そして, 最終的な損失に, 現在の Attention か Coverage Vector の低い方を加える. そうすることで, Attention の方が低いときは, 現在の単語の Attention が減衰するように学習され, Coverage Vector の方が低いときは, 現在の単語が重要ということなので, それまでの Attention が減衰するように学習する.

この処理を行うことで, できるだけ入力文章全体の情報を用いて要約を生成するようにする.

5 実験条件

この章では, 本研究における実験条件について述べる. 5.1 節では, パラメータ設定と, デコードする (要約を生成する) ときに使うビームサーチについて説明する. 5.2 節では Early Stopping について説明する.

5.1 パラメータ設定

本研究では, データセットに CNN/Daily Mail Dataset (DeepMind Q&A Dataset) を用いた. このデータセットは, 訓練データが 287,226 記事, 検証データが 13,368 記事, テストデータが 11,490 記事存在する. データ拡張の効果を見るために, 287,226 記事の他に, 約 1/5 の記事数である 57,000 記事と, 約 1/10 の記事数である 28,000 記事で実験を行った. 隠れ層のベクトルサイズは 256 に設定し, 埋め込みベクトルサイズは 128 に設定した. ドロップアウトは使っていない. バッチサイズは 8 に設定した. Pointer-Generator モデル (See, Liu, and Manning 2017) (以降この論文を原論文と呼ぶ) では, バッチサイズが 16 となっているため, 同じ記事数学習するためには, 二倍の学習回数が必要となる. 要約を生成する部分には, ビームサーチを使用した. 貪欲法では, 単語を生成するときに, 最も生成確率の高い単語をひとつ選ぶのに対し, ビームサーチでは, 上位 K 個までの生成単語を保持しながら処理を進めていく. そして最終的にできた複数の要約に対して, それぞれの単語の生成確率を掛け, 最も高いものを最終的な生成要約とした. 今回の実験では, K の値は 4 に設定している. 語彙数は 50,000 に, 学習率は 0.15 に設定した. 原論文では, エンコードする際に, 入力記事の最初から 400 単語のみを用いて, 学習していたが, 提案手法の場合, 取り除かれる不要文が最初から 400 単語以内でない場合, 文を取り除いた効果が無くなってしまうので, 入力記事全ての単語を用いて学習した. 具体的には, 入力単語数に, 入力記事の中の最大単語数である 2,380 を用いた. また, 最適化アルゴリズムは AdaGrad を用いた. これらのパラメータを表 1 に示した.

隠れ層のベクトルのサイズ	256
埋め込みベクトルサイズ	128
バッチサイズ	8
ビームサイズ	4
語彙数	50,000
学習率	0.15
入力単語数	2,380
最適化アルゴリズム	AdaGrad

表 1 パラメータ設定

5.2 Early Stopping

Early Stopping というのは、通常、各エポックの検証誤差（学習したモデルを使って、学習データ以外の検証データを入力して得られた要約と参照要約との生成確率の誤差）をモニタリングし、ある一定のエポック数が進んでも、検証誤差が下がらなかったところで学習を終了するというものである。Ouchi らの研究 (Ouchi and Tabuse 2019) では 90,000 記事、45,000 記事、において提案手法と拡張なし手法の比較を行っているが、異なる記事数における学習回数の決定法について考察をしていなかった。具体的には、90,000 記事、45,000 記事が全記事に対して約 1/3、1/6 の記事数になるということから、学習回数も 287,226 記事のときの 1/3、1/6 回にしていた。それに対し本研究では、モデルや訓練記事数に応じた学習回数を決定する方法として Early Stopping を採用した。訓練データ数 287,226 記事、57,000 記事、28,000 記事それぞれで、各手法において検証誤差をモニタリングし、どれだけのエポック数をモニタリングすればよいかを決定した。モニタリングするエポック数については、最小の検証誤差を出したエポック数の 2 倍までモニタリングする（例えば、5 エポックで検証誤差が最小なら 10 エポックまでモニタリングしてそれでも最小値が更新されなければ、5 エポックのモデルを採用する）。デコード（要約）に使うモデルは、誤差が最小となったときのモデルを使うことにする。我々は、Coverage Mechanism においては、Coverage Loss を用いて、各ステップにおいてモニタリングした。

6 実験

6.1 節では、評価に用いた尺度を、6.2 節、6.3 節では拡張なし手法、提案手法、SR、RI、RS、RD について、Early Stopping の条件を変えた比較実験を述べ、6.4 節、6.5 節では 6.3 節の実験 2 に対して、記事数を変えて比較実験をした結果を載せた。6.6 節では、さらに比較対象として、DA-Luhn、DA-LexRank、EDA を加えて実験をした結果を載せた。6.7 節では、生成した要約例を載せた。

6.1 評価尺度

本研究では, 評価尺度に ROUGE を用いた (Lin 2004). ROUGE は, 参照要約と生成要約の単語の一致率を見るものである. 本研究における参照要約は, CNN/Daily Mail Dataset (DeepMind Q&A Dataset) において人が作った要約である. 本研究では, ROUGE-1, ROUGE-2, ROUGE-L を用いた. ROUGE-1, ROUGE-2 は, それぞれユニグラム, バイグラムの一致率, ROUGE-L は, 最長共通部分列の文長に対する割合で測る. 参照要約を A B C D とし, 生成要約を A B C X Y とすると, 例えば, ROUGE-1 の再現率は, $3/5$ で, 適合率は, $3/4$ である. ROUGE-2 では, 要約をバイグラムに直さなければならないので, 参照要約は, (A B), (B C), (C D), 生成要約は, (A B), (B C), (C X), (X Y) となるので, 再現率は, $2/4$ で, 適合率は, $2/3$ となる. また, それぞれの ROUGE の測り方として, 再現率と適合率と F 値を用いる. 再現率は, 参照要約がどれだけ生成要約に含まれているか, つまり, 参照要約の単語をどれだけ生成できたかをみるもので, 適合率は, 生成要約がどれだけ参照要約に含まれているかを考えるものである. F 値は再現率と適合率を両方考慮したもので, 式 (6) で表される.

$$F \text{ 値} = \frac{2 \times \text{Recall} \times \text{Precision}}{\text{Recall} + \text{Precision}} \quad (6)$$

ただし, *Recall* は再現率, *Precision* は適合率を表す. *Recall*, *Precision* の計算式はそれぞれ式 (7), 式 (8) となる.

$$\text{Recall} = \frac{\text{参照要約と生成要約の一致単語数もしくは一致単語長}}{\text{生成要約の単語数}} \quad (7)$$

$$\text{Precision} = \frac{\text{参照要約と生成要約の一致単語数もしくは一致単語長}}{\text{参照要約の単語数}} \quad (8)$$

6.2 原論文と Coverage Mechanism の条件を揃えた比較実験 (実験 1)

本実験では, 学習記事数を最大記事数である 287,226 記事にして実験を行った. Copy Mechanism に関しては, Early Stopping で, 検証誤差が最小となるようなエポックを見つけた. Coverage Mechanism については, 原論文に倣って, 6,000 回とした.

6.2.1 結果

表 2 に結果を示す.

原論文と Coverage Mechanism の条件を揃えた条件では, 提案手法が最も良くなり, EDA の RD (Random Deletion) 手法が最も悪い結果となった. 続いて, RI (Random Insertion), SR (Synonym Replacement), 拡張なし手法, RS (Random Swap), そして提案手法の順番に良くなっていった.

	拡張なし手法	提案手法	SR	RI	RS	RD
ROUGE-1	0.3893	0.3948	0.3869	0.3856	0.3911	0.3817
ROUGE-2	0.1676	0.1724	0.1683	0.1649	0.1696	0.1657
ROUGE-L	0.3567	0.3624	0.3551	0.3527	0.3596	0.3499

表 2 学習記事数 287,226 の F 値 (Copy Mechanism: Early Stopping, Coverage Mechanism: 6,000)

	拡張なし手法	提案手法	SR	RI	RS	RD
ROUGE-1	0.3841	0.3916	0.3865	0.3876	0.3760	0.3828
ROUGE-2	0.1631	0.1702	0.1702	0.1674	0.1633	0.1667
ROUGE-L	0.3516	0.3613	0.3554	0.3563	0.3448	0.3520

表 3 学習記事数 287,226 の F 値 (Copy Mechanism: Early Stopping, Coverage Mechanism: Early Stopping)

6.2.2 考察

Coverage Mechanism は 6,000 回と固定したが、あくまで拡張なし手法において最適な回数である可能性もあるので、実験 2 では、Coverage Mechanism に関して、Early Stopping を用いることで、手法に限らず、自動的に決定する手法をとった。実験 1 の条件下では、提案手法が最も良い結果となり、拡張なし手法に比べて、スコアが 0.055 良くなっている。本実験の拡張なし手法と原論文との違いは大きく二つある。まず入力単語数を原論文では 400 単語と限定していたのに対し、本研究では限定をなくした。二つ目は、Early Stopping を用いたところである。これは、他の記事数で実験したり、他の手法で比べる際に学習回数をヒューリスティックに決めるのを避けるために必要なことであった。

6.3 学習回数を全て Early Stopping で決定した比較実験（実験 2）

本実験では、実験 1 と同じく学習記事数を最大記事数である 287,226 記事にして実験を行った。Copy Mechanism に関しては、Early Stopping で行い、Coverage Mechanism についても Early Stopping で行った。

6.3.1 結果

表 3 に結果を示す。

提案手法が最も良い結果となり、EDA の RS (Random Swap) が最も悪い結果となり、続いて、RD (Random Deletion), SR (Synonym Replacement), RI (Random Insertion), 拡張なし手法、そして提案手法の順に良くなっていった。

6.3.2 考察

本実験では Coverage Mechanism においても Copy mechanism と同様に Early Stopping を用いた。結果からみると、提案手法が最も良い結果となった。提案手法は拡張なし手法に比べて、スコアが 0.016 程度良くなっている。実験 1 と実験 2 を比べてみると、拡張なし手法は若干であるが良くなっている。提案手法をみると、少し悪くなっている。

6.4 57,000 記事におけるデータ拡張効果の検証実験（実験 3）

本実験では、学習記事数を最大の約 1/5 の 57,000 記事にして、実験 2 と同様、Copy Mechanism, Coverage Mechanism の両方で Early Stopping を用いた。

6.4.1 結果

表 4 に結果を示す。

提案手法が最も良い結果となり RS (Random Swap) が最も悪い結果となった。

6.4.2 考察

RS (Random Swap)、拡張なし手法、RD (Random Deletion)、RI (Random Insertion)、SR (Synonym Replacement)、そして提案手法の順番に値が良くなっている。提案手法は拡張なし手法に比べて、スコアが 0.0251 良くなっている。RI (Random Insertion) と SR (Synonym Replacement) は拡張なし手法よりも良くなった。このことから、提案手法に関して 287,226 記事のときよりもデータ拡張の効果が大きくなったと言える。また、EDA のうち RS 以外は全て拡張なし手法よりも良い結果となった。少ない記事の方がデータ拡張の効果が出ることは、EDA の論文 (Wei and Zou 2019) でも報告されている。これは、多くの記事を学習した方が偏りが少なくなり、データ拡張の効果が出にくいことによると思われる。今回の例だと、57,000 記事は学習の偏りが多くなり、データ拡張によってデータの複雑性を増すことにより汎化性能が高まったものと思われる。さらに、本実験で言うと、提案手法が EDA の各手法よりも汎化性能を上げるのに有用であったと言える。拡張なし手法に比べて提案手法の方が拡張した分、データは増えているが、学習データを増やせることがデータ拡張手法のメリットであり、

	拡張なし手法	提案手法	SR	RI	RS	RD
ROUGE-1	0.3340	0.3591	0.3487	0.3402	0.3264	0.3390
ROUGE-2	0.1289	0.1454	0.1389	0.1289	0.1274	0.1308
ROUGE-L	0.3096	0.3321	0.3235	0.3147	0.2999	0.3138

表 4 学習記事数 57,000 の F 値 (Copy Mechanism: Early Stopping, Coverage Mechanism: Early Stopping)

学習回数が増える可能性については Early Stopping で自動化されているのでデータ拡張手法の効果を検証する上で問題ではないと考える。

6.5 28,000 記事におけるデータ拡張効果の検証実験（実験 4）

本実験では、学習記事数を最大記事数の約 1/10 である 28,000 記事にして、実験を行った。Copy Mechanism, Coverage Mechanism の学習回数の決め方は実験 2 と同じとした。

6.5.1 結果

表 5 に結果を示す。

本実験でも提案手法が最もよい結果となり、SR (Synonym Replacement) で最も悪い結果となった。

6.5.2 考察

RS (Random Swap), 拡張なし手法, RI (Random Insertion), RD (Random Deletion), 提案手法, SR (Synonym Replacement) の順番に良くなった。提案手法は、拡張なし手法に比べて、スコアが 0.0115 良くなった。EDA の手法は、拡張なし手法よりも良くなることもあれば、悪くなることもあった。しかし、提案手法は常に拡張なし手法よりも良い結果を出していた。今回の実験では、SR 手法が最も良い結果となっている。

実験 2 から実験 4 までをみると、どれも RS 手法が一番悪い結果となっている。一方で、実験 1 に関しては、提案手法を除いて最も良い結果となっている。RS 手法は、ランダムに二つの単語を入れ替えてデータ拡張する手法で、EDA の他の 3 手法より最も意味的に破綻した文章を作ってしまう可能性がある。そのため、実験 2 から実験 4 では RS 手法が最も悪い結果になったものと思われる。実験 1 に関しては、Coverage Mechanism の学習回数を 6,000 回として固定して学習しており、他の実験よりも多い回数になっている。そのため、他の実験に比べ全体の文脈をより考慮した形となり、多少文が破綻しても影響が少なかったのではないと思われる。逆に、SR 手法は、全体的に EDA 手法の中で最も良い結果を出している。その理由として、SR 手法は、同義語に入れ替えるものなので、意味的に最も影響の少ない手法となっているため

	拡張なし手法	提案手法	SR	RI	RS	RD
ROUGE-1	0.3358	0.3473	0.3506	0.3420	0.3298	0.3427
ROUGE-2	0.1303	0.1370	0.1399	0.1330	0.1214	0.1311
ROUGE-L	0.3122	0.3206	0.3247	0.3180	0.3071	0.3175

表 5 学習記事数 28,000 の F 値 (Copy Mechanism: Early Stopping, Coverage Mechanism: Early Stopping)

あると考えられる。残りの RI 手法と RD 手法は、前者の二つの手法の中間に位置する。RI 手法は、ある単語の同義語をランダムな位置に挿入する方法で、RD 手法は、ランダムな単語を削除する方法で、いずれも同じ単語数だけ文章全体に影響を与えるという意味で同程度の効果が表れたものと思われる。

提案手法よりも EDA 手法全体が悪くなった原因としては、EDA の場合は、特に意味のない単語を挿入したりや単語が入れ替わってしまい意味が取れなくなったり、重要な単語が消されてしまう可能性があることにより汎化性能を上げる効果よりも正しく学習できなかった影響のほうが大きかったと思われる。その点で、提案手法は、不要文を削除するだけなので文章全体の意味が取れなくなることもなく、さらに重要なところが削除されることもないのでデータ拡張による悪い影響は少なく効果が EDA より表れていたものと考えられる。

6.6 提案手法, DA-Luhn, DA-LexRank, EDA の比較実験 (実験 5)

本実験では、記事数 287,226, Copy Mechanism, Coverage Mechanism の学習回数の決め方は実験 2 と同じとした。そして、本実験ではさらに、DA-Luhn, DA-LexRank, EDA の方法でデータ拡張して学習させた結果も載せた。また、DA-Luhn, DA-LexRank 手法と提案手法における不要文抽出の評価実験も行った。

6.6.1 結果

表 6, 7 に結果を示す。

新たに加えた三手法 (DA-Luhn, DA-LexRank, EDA) のどの手法も拡張なし手法、提案手法、

	拡張なし手法	提案手法	SR	RI	RS
ROUGE-1	0.3841	0.3916	0.3865	0.3876	0.3760
ROUGE-2	0.1631	0.1702	0.1702	0.1674	0.1633
ROUGE-L	0.3516	0.3613	0.3554	0.3563	0.3448
	RD	DA-LexRank	DA-Luhn	EDA	
ROUGE-1	0.3828	0.3777	0.3739	0.3805	
ROUGE-2	0.1667	0.1616	0.1640	0.1628	
ROUGE-L	0.3520	0.3489	0.3449	0.3482	

表 6 学習記事数 287,226 の F 値 (Copy Mechanism: Early Stopping, Coverage Mechanism: Early Stopping). 全手法で比較。

	トピックモデル	LexRank	Luhn
不要文	134,144	79,846	2,509
最重要文	12,327	3,503	1,671

表 7 各手法と ROUGE を用いた手法で一致した文の数

EDA の各手法 (SR, RI, RS, RD) に及ばなかった. 表 8 においては, ROUGE を用いて, 要約と最も ROUGE-1 の F 値の高かった文を最重要文とし, 要約との ROUGE-1 の F 値の最も低かった文もしくは, ROUGE-1 の F 値が 0 になった文集合を不要文または不要文群とした. そして, 提案手法に用いたトピックモデルと DA-LexRank で用いた LexRank と DA-Luhn で用いた Luhn を使って最重要文と不要文を抽出し, 一致した文の数を調べた.

6.6.2 考察

DA-Luhn が最も悪い結果となり, 続いて, DA-LexRank, EDA, RS (Random Swap), RD (Random Deletion), SR (Synonym Replacement), RI (Random Insertion), 拡張なし手法, そして提案手法の順に良くなっていった. この結果から, データ拡張をするために行う文の重要度決め用いるのは, DA-Luhn や DA-LexRank よりもトピックモデルを用いた提案手法の方が効果があることを確認した. 実際, 表 8 を見てもわかるとおり, 提案手法を用いた場合の方が, 不要文, 最重要文とも一致度は高い. ROUGE を使って不要文削除を行えばいいのではないかということも考えられるが, ROUGE はあくまで単語の一致率を見ているので単語がひとつも一致しない文はスコアが 0 になる. したがって, スコアが 0 の文同士の重要度の差は生まれない. 本研究では, 不要文として一文を削除することになっているが, ROUGE だとスコアがどちらも 0 で差が生まれない場合に一文を選ぶことができないため本研究では扱わなかった. また, EDA は, 組み合わせて使うよりもそれぞれの手法を用いた方がデータ拡張の効果が表れやすいことが確認できた. DA-Luhn がトピックモデルを用いたときよりも拡張の効果が表れなかった原因として, DA-Luhn では, 単語の位置と頻度だけを考慮しており, 一方で, トピックモデルは TF-IDF 行列からトピック分布の事前分布であるディリクレ分布を推定するため, 各文章のトピック分布を評価することができる点にあると思われる. 次に DA-LexRank がトピックモデルを用いたときよりも拡張の効果が表れなかった原因として考えられるのは, トピックモデルでは, TF-IDF 行列から推定したトピックを用いているので, car と automobile といった同じ意味でも異なる単語を捉えられる. そのようなトピックの観点が DA-LexRank にはなかったことが文の重要度を決めるときに影響を及ぼし, トピックモデルより悪い結果になったものと思われる.

6.7 生成した要約例

本文と生成した要約例を付録に載せる. 要約例は, 拡張なし手法で生成した要約と提案手法で生成した要約とを比較するため, 参照要約との ROUGE-1 の F 値をもとに, 実験 1 において提案手法で生成した要約と参照要約との ROUGE-1 の F 値が最大になる場合 (付録 1), 最小になる場合 (付録 2), 中央値になる場合 (付録 3) の三つの場合の要約元文, 参照要約, 拡張なし手法での要約, 提案手法での要約を示した. また, 図 4~6 は, そのうち要約例だけ示す.

参照要約

this page includes the show transcript .

use the transcript to help students with reading comprehension and vocabulary .

at the bottom of the page , comment for a chance to be mentioned on cnn student news .

you must be a teacher or a student age 13 or older to request a mention on the cnn student news roll call .

生成要約（拡張なし手法 coverage 6,000 回）

this page includes the show transcript .

use the transcript to help students with reading comprehension and vocabulary .

use the weekly newsquiz to test your knowledge of stories you saw on cnn student news .

生成要約（提案手法 coverage 6,000 回）

this page includes the show transcript .

use the transcript to help students with reading comprehension and vocabulary .

at the bottom of the page , comment for a chance to be mentioned on cnn student news .

you must be a teacher or a student age 13 or older to request a mention on the cnn student news roll call .

図 4 ROUGE-1 の F 値が最大になる場合（要約のみ）

参照要約

gray 's family asked there be no protests ; they condemned violence .

community leaders and brave residents got in between rioters and police .

生成要約（拡張なし手法 coverage 6,000 回）

“ i want them all to go back home , ” rev. jamal bryant tells cnn 's rev. jamal bryant .

the gangs have signed a peace deal and are uniting to push against police lines .

the gangs are consistently pursued by the fbi .

生成要約（提案手法 coverage 6,000 回）

“ i want them all to go back home , ” rev. jamal bryant says .

“ i want y' all to get justice for my son , ” “ i 'm just a soldier , ” stepfather says .

“ i want y' all to get justice for my son , but do n't do it like this is for freddie , ” stepfather says .

図 5 ROUGE-1 の F 値が最小になる場合（要約のみ）

ROUGE-1 の F 値が最大になる場合（図 4）は、参照要約が 4 文あるが、拡張なし手法モデルはそのうち最初の 2 文だけ正解している。提案手法は 4 文とも正解している。ROUGE-1 の F 値が最小になる場合（図 5）は、どちらの手法も正解には程遠いが、“jamal bryant”というこの文章にしか出てこない固有名詞に対して、うまく生成できているのがわかる。提案手法の方は、文が繰り返されており、Coverage Mechanism がうまく働いていないことがわかる。ROUGE-1 の F 値が中央値をとる場合（図 6）は、どちらの手法も最初の一文だけうまく生成できているのがわかる。総合すると、若干ではあるが提案手法の方が全体として良い要約になっているものと思われる。

参照要約

more than half of members of group of industrial nations reduced aid .
 but britain failed to follow suit - thanks to target imposed by mr cameron .
 its overseas spending is double the average level of other nations in group .

生成要約（拡張なし手法 coverage 6,000 回）

more than half the members of a group of 28 industrialised nations reduced their aid budget between 2013 and 2014 , a report found .

britain 's failure to follow suit means that its overseas aid spending of 0.7 per cent of national income is double the average level of the other nations .

overall across these 28 countries aid budgets fell by 0.5 per cent in real terms .

生成要約（提案手法 coverage 6,000 回）

more than half the members of a group of 28 industrialised nations reduced their aid budget between 2013 and 2014 .

out of every # 100 spent by these western nations on overseas aid , no less than # 14 now comes from the uk .

despite this , the uk increased its spend by 1.2 per cent to # 11.7 billion in 2014 .

図 6 ROUGE-1 の F 値が中央値をとる場合（要約のみ）

7 おわりに

本研究では、自然言語処理におけるデータ拡張の手法を提案し、自動要約システム Pointer-Generator モデル (See et al. 2017) の場合で、データ拡張の効果があることを示した。また、比較実験により、自動要約システムにおいて汎化性能を出すためには、EDA の各手法およびそれらを組み合わせた手法よりも提案手法の方が効果があることが確認できた。さらに、本研究では、重要度の低い文を取り出す方法として、抽出型要約の 3 手法（トピックモデル、DA-Luhn, DA-LexRank）で比較実験したが、実験 5 の結果からトピックモデルの場合が最も不要文抽出の一致度は高かった。したがって、データ拡張の効果もトピックモデルの場合が最も良かった。より少ない学習記事においてデータ拡張の効果が大きく現れるのではないかと考えていたが、実験結果からは、57,000 記事のときに一番効果が表れていた。しかし、少なすぎるとデータ拡張の効果が少なくなることがわかった。今回は、一番重要度の低い文を取り除いた文章を拡張記事データとして水増しすることでデータ拡張を行った。しかし、元の文章の長短に関わらず、一文しか取り除いていないので、今後の課題として、文章の長さに応じて取り除く文の数を変えろといった拡張が考えられる。また、重要度の低い文を取り除くのではなく、重要度の高い文を残すというやり方も考えられる。最近の研究では、BERT を使った自動要約システムも考えられている (Liu and Lapata 2019)。このように、Pointer-Generator モデルだけでなく最先端のモデルにも提案手法が有効であるか試していく必要がある。

参考文献

- Blei, D. M., Ng, A. Y., and Jordan, M. I. (2003). “Latent Dirichlet Allocation.” *The Journal of Machine Learning Research*, **3**, pp. 993–1022.
- Bojanowski, P., Grave, E., Joulin, A., and Mikolov, T. (2017). “Enriching Word Vectors with Subword Information.” In *Transactions of the Association for Computational Linguistics (TACL)*, Vol. 5, pp. 135–146.
- Chang, J., Gerrish, S., Wang, C., Boyd-graber, J., and Blei, D. (2009). “Reading Tea Leaves: How Humans Interpret Topic Models.” In *Proceedings of the 22nd International Conference on Neural Information Processing Systems*, pp. 288–296.
- Edunov, S., Ott, M., Auli, M., and Grangier, D. (2018). “Understanding Back-Translation at Scale.” In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pp. 489–500.
- Erkan, G. and Radev, D. R. (2004). “LexRank : Graph-based Lexical Centrality as Saliency as Text Summarization.” *Journal of Artificial Intelligence Research*, **22**, pp. 457–479.
- Grusky, M., Naaman, M., and Artzi, Y. (2018). “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*, Vol. 1, pp. 708–719.
- Gu, J., Lu, Z., Li, H., and Li, V. O. (2016). “Incorporating Copying Mechanism in Sequence-to-sequence Learning.” In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (ACL)*, Vol. 1, pp. 1631–1640.
- Hermann, K. M., Kocisky, T., Grefenstette, E., Espeholt, L., Kay, W., and Mustafa Suleyman, P. B. (2015). “Teaching Machines to Read and Comprehend.” In *Proceedings of the 28th International Conference on Neural Information Processing Systems (NIPS)*, pp. 1693–1701.
- 人見雄太, 田口雄哉, 田森秀明, 菊田洸, 西鳥羽二郎, 岡崎直観, 乾健太郎, 奥村学 (2019). 出力長制御を考慮した見出し生成モデルのための大規模コーパス. 言語処理学会第 25 回年次大会発表論文集, pp. 1225–1228. 言語処理学会. [Y. Hitomi et al. (2019). Shutsuryokucho Seigyo wo Kouryo shita Midashi Seisei Moderu no tameno Daikibo Kopasu. Proceedings of the 25th Annual Meeting of the Association for Natural Language Processing, pp. 1225–1228.].
- Hochreiter, S. and Schmidhuber, J. (1997). “Long Short-term Memory.” In *Neural Computation*, pp. 1735–1780.
- Internet Live Stats. “Total number of Websites.” <https://www.internetlivestats.com/total-number-of-websites/>.

- 小平知範 (2018). 文書構造に着目したニューラル文書要約. Master's thesis, 首都大学東京大学院システムデザイン研究科情報通信システム学域. [T. Kodaira (2018). Incorporating Document Structure into Neural Abstractive Summarization. Master's thesis, Graduate School of System Design, Tokyo Metropolitan University.].
- Lin, C.-Y. (2004). "ROUGE: A Package for Automatic Evaluation of Summaries." In *Text Summarization Branches Out*, pp. 74–81.
- Liu, Y. and Lapata, M. (2019). "Text Summarization with Pretrained Encoders." In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing*, pp. 3730–3740.
- Luhn, H. P. (1958). "The Automatic Creation of Literature of Abstracts." *IBM Journal of Research and Development*, **2** (2), pp. 159–165.
- Mihalcea, R. and Tarau, P. (2004). "TextRank: Bringing Order into Texts." In *Proceedings of the 2004 Conference on Empirical Methods in Natural Language Processing (EMNLP-WS)*, pp. 404–411.
- Mikolov, T., Chen, K., Corrado, G. S., and Dean, J. (2013). "Efficient Estimation of Word Representations in Vector Space." In *International Conference on Learning Representations (ICLR2013)*.
- Napoles, C., Gormley, M., and Durme, B. V. (2012). "Annotated Gigaword." In *Proceedings of the Joint Workshop on Automatic Knowledge Base Construction & Web-scale Knowledge Extraction (AKBC-WEKEX)*, pp. 95–100.
- Ouchi, T. and Tabuse, M. (2019). "Effectiveness of Data Augmentation in Automatic Summarization System." In *The Proceedings of the 2019 International Conference on Artificial Life and Robotics*, pp. 177–180.
- Page, L., Brin, S., Motwani, R., and Winograd, T. (1998). "The PageRank Citation Ranking: Bringing Order to the Web." In *Proceedings of the 7th International World Wide Web Conference*, pp. 161–172.
- Pennington, J., Socher, R., and Manning, C. (2014). "GloVe: Global Vectors for Word Representation." In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pp. 1532–1543.
- Reinsel, D., Rydning, J., and Gantz, J. F. . "Worldwide Global DataSphere Forecast, 2020–2024: The COVID-19 Data Bump and the Future of Data Growth." <https://www.idc.com/getdoc.jsp?containerId=US44797920>.
- Rush, A. M., Chopra, S., and Weston, J. (2015). "A Neural Attention Model for Abstractive Sentence Summarization." In *Proceedings of the 2015 Conference on Empirical Methods in*

Natural Language Processing, pp. 379–389.

See, A., Liu, P. J., and Manning, C. D. (2017). “Get To The Point: Summarization with Pointer-Generator Networks.” In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (ACL)*, pp. 1073–1083.

重松遥, 小林一郎 (2012). 潜在的トピックの重要度を考慮した要約文の生成. 言語処理学会第 18 回年次大会発表論文集, pp. 3–26. 言語処理学会. [H. Shigematsu and I. Kobayashi (2012). Senzaiteki Topikku no Juyodo wo Kouryo shita Yoyakubun no Sakusei. Proceedings of the 18th Annual Meeting of the Association for Natural Language Processing, pp. 3–26.].

Tu, Z., Lu, Z., Liu, Y., Liu, X., and Li, H. (2016). “Modeling Coverage for Neural Machine Translation.” In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (ACL)*, Vol. 1, pp. 76–85.

Wei, J. and Zou, K. (2019). “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.

Yun, S., Han, D., Oh, S. J., Chun, S., Choe, J., and Yoo, Y. (2019). “CutMix: Regularization Strategy to Train Strong Classifiers with Localizable Features.” In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 6023–6032.

Zhang, H., Cisse, M., Dauphin, Y. N., and Lopez-Paz, D. (2018). “mixup: Beyond Empirical Risk Minimization.” In *International Conference on Learning Representations (ICLR)*.

付録 1

ROUGE-1 の F 値が最大になる場合

要約元の文章

“april 8 , 2015 . after a report on headlines concerning the u.s. and cuba , cnn student news gives some perspective on california ’s historic drought : its effects are becoming more likely to ripple across the nation . there ’s been a flu outbreak in chicago – but not one that affects humans . and a proposed tunnel would connect denmark and germany via an undersea road and rail line . on this page you will find today ’s show transcript and a place for you to request to be on the cnn student news roll call . transcript . click here to access the transcript of today ’s cnn student news program . please note that there may be a delay between the time when the video is available and when the transcript is published . cnn student news is created by a team of journalists who consider the common core state standards , national standards in

different subject areas , and state standards when producing the show . roll call . for a chance to be mentioned on the next cnn student news , comment on the bottom of this page with your school name , mascot , city and state . we will be selecting schools from the comments of the previous show . you must be a teacher or a student age 13 or older to request a mention on the cnn student news roll call ! thank you for using cnn student news !”

参照要約

this page includes the show transcript .

use the transcript to help students with reading comprehension and vocabulary .

at the bottom of the page , comment for a chance to be mentioned on cnn student news .

you must be a teacher or a student age 13 or older to request a mention on the cnn student news roll call .

生成要約（拡張なし手法 coverage 6000 回）

this page includes the show transcript .

use the transcript to help students with reading comprehension and vocabulary .

use the weekly newsquiz to test your knowledge of stories you saw on cnn student news .

生成要約（提案手法 coverage 6000 回）

this page includes the show transcript .

use the transcript to help students with reading comprehension and vocabulary .

at the bottom of the page , comment for a chance to be mentioned on cnn student news .

you must be a teacher or a student age 13 or older to request a mention on the cnn student news roll call .

付録 2

ROUGE-1 の F 値が最小になる場合

要約元の文章

“please , not even a demonstration . freddie gray ’s family had asked there be quiet on baltimore ’s streets the day they laid him to rest . and above all , no violence . raging hordes turned a deaf ear to that on monday . but a handful of people repeated the family ’s message . they became criers in the desert against countless young people flinging rocks at police , breaking windows , looting and setting fires . the peacemakers – clergy , gray ’s family and brave residents – placed themselves in the rioters ’ way . their message was the same . “ i want them all to go back home , ” said rev. jamal bryant . “ it ’s disrespect to the family . the family was very clear –

we 've been saying it all along – today there was absolutely no protest , no demonstration , ” he said . but the messengers were a finger in a dam that quickly crumbled , as rowdy groups swelled into a full urban riot . it overshadowed the message peaceful protesters delivered on prior days – justice for gray . the 25-year-old african-american man died from spinal injuries after being arrested earlier this month . the early fits of violence came in the afternoon , about the time mourners left gray 's memorial services blocks away , rev. bryant said . they bumped right into it . “ for us to come out of the burial and walk into this is absolutely inexcusable , ” he said . he did not want to see it spread to downtown baltimore , where some rioters said it would , and organized people to stand in the way . “ we have a line of gentlemen from the nation of islam to build a human wall , as well as men from the christian church making that human wall , ” he said . but as crowds turned into multitudes , the intervention became a drop in the bucket by compare , and police lines were also no match . as officers in riot gear receded , flames engulfed cars and stores and roared out of apartment buildings into the night sky . a senior living facility under construction by a baptist church burned to the ground . the blazes stretched the fire departments ' resources , as at least 30 trucks deployed . looters streamed into a cvs , bodegas and liquor stores and walked out with what they could carry . a young man in a blue sweatshirt tried to talk people down by himself . he walked up to cnn correspondent miguel martinez , as a store nearby was being looted . it later went up in flames . the man , who did n't say his name , was disgusted at what was happening in his neighborhood and disappointed in the police response to rioting . there was a line of police down the street , not far away . “ they could have moved down here to stop it , ” he told martinez . the gray family 's lawyers , again , put the family 's wish out to the public that there be no protests that day , let alone violence . it 's marring the cause and hope for change that may have come out of the investigation into freddie gray 's death , said family attorney mary koch . “ that 's just disintegrated into just looking at baltimore city and thinking that the city is the city of violence , ” she said . against all odds , a handful of individuals kept trying to stop it . a tall , adult man walked up to a young man who was confronting riot police. he slung an arm over his shoulder , turned him back around in the other direction and marched him away from police lines . but as they strolled past a crowd , a young man behind them hurled a stone at police , putting his whole body into the throw . at least one young man paid the price for his participation , when his mother turned up to spank him home . before running cameras , she slapped him in the head again and again , driving him away from the crowd , as she cursed . police commissioner anthony batts later thanked her . “ i wish i had more parents that took charge of their kids out there tonight , ” he said . after night fell , giving way to a

10 p.m. curfew for juveniles , robert valentine stood alone with his back to a line of police in riot gear . he shoed away young people tempted to approach them . “ go ! step your – ss away ! ” “ i ’m just a soldier , ” said valentine . he told cnn ’s joe johns that he was a veteran of the vietnam war . young people had no business on the streets , he said . “ they need to be in their home units studying and doing something with their lives . ” even baltimore members of the crips and bloods , two street gangs renowned for drug dealing and extensive violent crime – and for killing each other – came together with others who condemned the rage that swept through their neighborhoods . “ the guys who pulled me aside are self-identifying as crips and say they do n’t approve of whats happening . ‘ this is our community , ’ ” wrote baltimore sun reporter justin fenton on his confirmed twitter account . gangs members joined community leaders and gray ’s family for a press conference monday night on the stage at new shiloh baptist church , which had held gray ’s funeral . an announcer thanked them for coming to the church . the gangs have signed a peace deal and are uniting to push against police lines in protests , according to a report by the daily beast . rev. bryant also mentioned their peace treaty . but police say the gangs ’ purpose goes much further – that they and another gang called the black guerilla family plan to ‘ take out ’ law enforcement officers , police said . “ this is a credible threat . ” the gangs are consistently pursued by the fbi . at the end of the day , gray ’s family had the last word on the violence at the press conference . it was n’t good . “ to see that it turned into all this violence and destruction , i am appalled , ” said richard shipley , gray ’s stepfather . “ i want y’ all to get justice for my son , but do n’t do it like this here , ” said gray ’s mother gloria darden , who wore a t-shirt with her son ’s photo . “ i do n’t think that ’s for freddie , ” his twin sister fredericka gray said . “ i think the violence is wrong . ” after had been bitterly marred by the rioting . murphy asked for a show of hands in the church audience of people who had experienced police brutality or personally knew someone who did . all but a few hands went up .”

参照要約

gray ’s family asked there be no protests ; they condemned violence .
community leaders and brave residents got in between rioters and police .

生成要約（拡張なし手法 coverage 6000 回）

“ i want them all to go back home , ” rev. jamal bryant tells cnn ’s rev. jamal bryant .
the gangs have signed a peace deal and are uniting to push against police lines .
the gangs are consistently pursued by the fbi .

生成要約（提案手法 coverage 6000 回）

“ i want them all to go back home , ” rev. jamal bryant says .

“ i want y’ all to get justice for my son , ” “ i ’m just a soldier , ” stepfather says .
 “ i want y’ all to get justice for my son , but do n’t do it like this is for freddie , ” stepfather says .

付録 3

ROUGE-1 の F 値が中央値をとる場合

要約元の文章

“britain bucked the trend of western nations by increasing spending on foreign aid last year while others imposed savage cuts . more than half the members of a group of 28 industrialised nations reduced their aid budget between 2013 and 2014 , a report found . britain ’s failure to follow suit – thanks to a target imposed by david cameron – means that its overseas aid spending of 0.7 per cent of national income is double the average level of the other nations in the group known as the development assistance committee . a target set by david cameron -lrpb- pictured at a rally this week -rrb- has left britain unable to follow other countries in reducing their aid budget between 2013 and 2014 . out of every # 100 spent by these western nations on overseas aid , no less than # 14 now comes from the uk . overall across these 28 countries aid budgets fell by 0.5 per cent in real terms , according to the report by the organisation for economic co-operation and development . despite this , the uk increased its spend by 1.2 per cent to # 11.7 billion in 2014 . it comes on top of even larger increases in previous years . france slashed its aid budget by 9.2 per cent to # 6.2 billion , the fourth successive decrease . japan ’s was cut by 15.3 per cent to # 5.5 billion . australia cut its foreign aid by 7.2 per cent to # 2.5 billion , while cash-strapped spain cut its funding by 20.3 per cent to # 1.1 billion . the uk ’s aid budget is now nearly double france ’s and ten times the level in spain . in total , 15 members of the dac cut foreign aid , including belgium , italy , norway , poland and portugal . britain was one of 13 countries which increased its budget despite the straitened times . these included germany , the us and new zealand . trade , not aid : conservative mp peter bone . surprisingly , greece , which is dependent on bailouts from the eu and the imf , put up its aid budget by 6.8 per cent . last night tory peter bone , who is standing for re-election as mp for wellingborough , said : ‘ when we have seen cuts at home , people find it very strange that we can give away so many billions of pounds a year , half of which we do n’t even control who it goes to . ‘ people are telling me on the doorstep during this campaign that we should cut the amount we spend on overseas aid , and increase opportunities for poorer countries to

trade with the eu – trade , not aid , is what we need . what this report shows is that other governments recognise the fact that we have got to look after our own country first , and ensure aid is better targeted . ’ britain ’s aid budget may need to rise by at least # 500million in future years , because changes in eu accounting rules make it harder to meet the 0.7 per cent target . the development assistance committee was set up by the oecd which represents richer countries . the oecd report , 2014 global outlook on aid , shows that britain is one of only five members of the dac to meet the target of spending 0.7 per cent of national income on aid along with sweden , luxembourg , norway and denmark . the average oecd member spends just 0.39 per cent . in monetary terms , the uk now has the second highest aid budget in the world behind the us which spends # 19.6 billion . but that is just 0.19 per cent of its national income . france spends 0.36 per cent , and germany 0.41 per cent . the 28 dac members spent a total of # 81billion on aid last year – a record figure , but down in real terms . it means the uk ’s budget of # 11.7 billion is a seventh of the entire total for the western world . max lawson , of oxfam , praised britain and criticised countries which have cut their aid budgets . ‘ aid saves lives , ’ he said . ‘ what we ’re seeing is shameful indifference on the part of many of the world ’s richest nations . ‘ their leaders are breaking with impunity the solemn promise they made to help make poverty history . ’ ”

参照要約

more than half of members of group of industrial nations reduced aid .

but britain failed to follow suit - thanks to target imposed by mr cameron .

its overseas spending is double the average level of other nations in group .

生成要約（拡張なし手法 coverage 6000 回）

more than half the members of a group of 28 industrialised nations reduced their aid budget between 2013 and 2014 , a report found .

britain ’s failure to follow suit means that its overseas aid spending of 0.7 per cent of national income is double the average level of the other nations .

overall across these 28 countries aid budgets fell by 0.5 per cent in real terms .

生成要約（提案手法 coverage 6000 回）

more than half the members of a group of 28 industrialised nations reduced their aid budget between 2013 and 2014 .

out of every # 100 spent by these western nations on overseas aid , no less than # 14 now comes from the uk .

despite this , the uk increased its spend by 1.2 per cent to # 11.7 billion in 2014 .

略歴

大内 智仁：2017 年京都府立大学生命環境学部環境・情報科学科卒業。2019 年同大学大学院生命環境科学研究科環境科学専攻修了。同年同大学大学院生命環境科学研究科環境科学専攻博士後期課程に進学。

田伏 正佳：1983 年神戸大学理学部物理学科卒業。1985 年同大学大学院理学研究科修士課程修了。1988 年同大学大学院自然科学研究科博士課程修了。学術博士。1992 年宮崎大学工学部助手。1999 年同大学助教授。2003 年京都府立大学人間環境学部助教授。2019 年同大学教授。機械学習，コンピュータビジョン，自然言語処理に興味を持つ。

(2020 年 8 月 3 日 受付)

(2020 年 12 月 1 日 再受付)

(2021 年 1 月 17 日 採録)