教師あり要約モデルの研究においては、タイトルを本文の要約とみなし学習データとするのが一般的であるが、これらはノイズ、すなわち不適切な本文－要約ペアを多く含む。本研究では、カリキュラムラーニングを用いてこうしたノイズを含むデータから効率的に要約モデルを学習させる手法を検討する。カリキュラムラーニングは学習データを難易度やノイズの量などを表す指標に従ってソートし、段階的な学習を行うことで性能を向上させる手法であり、ノイズを含むデータの学習にも有効である。本研究の目的の1つは、これまで検証されてこなかったカリキュラムラーニングの要約タスクへの有効性を検証することである。翻訳タスクの先行研究では、ノイズの多いコーパスと少ないコーパスから学習されたモデルでノイズ定量化を行ったが、要約分野にようしたコーパスは存在しない。本研究のもう1つの目的は、単一コーパスからノイズを定量化してカリキュラムラーニングに適用する手法を提案することである。提案モデルである Appropriateness Estimator は、本文－要約の正しいペアと、ランダムに組み合わせたペアを分類するタスクによって学習され、本文－要約ペアの適切性を計算する。本研究では3つの要約モデルで実験を行い、カリキュラムラーニング及び提案手法が要約モデルの性能を向上させることを示す。

キーワード：要約、生成型要約、生成、ノイズ、カリキュラムラーニング、ニューラルネットワーク、Transformer、BART

Quantifying Appropriateness of Summarization Data for Curriculum Learning

RYUJI KANO†,††, TOMOKI TANIGUCHI† and TOMOKO OHKUMA†

Previous research of summarization models regards titles as summaries of source texts. However, much research has reported these training data are noisy. We propose an effective method of curriculum learning to train summarization models from noisy data. Curriculum learning is a method to improve performance by sorting training data based on difficulty or noisiness, and is effective to training models with noisy data. However, previous research never applied curriculum learning to summarization tasks. One aim of this research is to validate the effectiveness of curriculum learning to summarization tasks. In translation tasks, previous research quantified noise using two models trained with noisy and clean corpora. Because such corpora do not exist in summarization fields, it is difficult to apply this method to summarization tasks.

† 富士フィルムビジネスイノベーション株式会社, FUJIFILM Business Innovation Corp.
†† 東京工業大学工学院, School of Engineering, Tokyo Institute of Technology

(C) The Association for Natural Language Processing. Licensed under CC BY 4.0 (https://creativecommons.org/licenses/by/4.0/)
Another aim of this research is to propose a model that can quantify noise using a single noisy corpus. The training task of the proposed model, Appropriateness Estimator is to distinguish correct source-summary pairs of from randomly assigned pairs. Throughout the training, the model learns to compute the appropriateness of source-summary pairs. We conduct experiments on three summarization models and verify curriculum learning and our method improves the performance.

**Key Words:** Summarization, Abstractive Summarization, Generation, Noise, Curriculum Learning, Neural Network, Transformer, BART

1 はじめに

ニューラルネットワークを利用したSequence-to-sequenceモデルの発展により、生成型自動要約の性能は飛躍的に向上した。Sequence-to-sequence要約モデルの学習においては、新聞記事(Nallapati et al. 2016)であれば見出し、ソーシャルメディア(Kim et al. 2019)やレビュー(Li et al. 2019)であればタイトル、メール(Zhang and Tetreault 2019)であれば件名を要約とみなして使用する。これらの要約は本文に書かれた内容の重要な箇所を適切かつ簡潔に記述していことが望ましい。しかしながら、過去の多くの研究が要約モデルの学習データセットには不適切な本文一要約ペアが多く含まれることを報告している(Zhang and Tetreault 2019; Li et al. 2019; Kryscinski et al. 2019; Matsumaru et al. 2020)。具体例を表1に示す。例はReddit Titleデータ(Kim et al. 2019)、Enron Subjectデータ(Zhang and Tetreault 2019)から引用したものである。表の上段の例では本文にはタイトルの続きが書かれており、タイトルは本文に書かれている内容を反映していない。下段の例では、件名は簡潔すぎて情報不足であり、要約としての体裁を成していない。こうしたノイズを含むデータセットに対処する方法が求められている。

| データ種類   | タイトル／件名                                                                 | 本文                                                                 |
|------------|------------------------------------------------------------------------------|----------------------------------------------------------------------|
| Reddit Title | accidentally drinking 0 day old coffee w / milk that was sitting on my desk next to my new coffee | just happened will update with further details as they emerge       |
| Enron Subject | hey                                                                         | here is what vickie told me about capacity on your pipeline after you eliminate segmenting. example # 1 assumption : peco has num dt/day of telescoped capacity with a primary delivery point of peco in z6 |

表1 タイトル、件名が本文の要約として不適切な例
ノイズを含むデータから効率的にモデルを学習させる方法の1つとしてカリキュラムラーニング (Bengio et al. 2009) が用いられている。カリキュラムラーニングは元来、学習データの順序を変えることで、収束速度やモデルの性能を上げる手法であるが、ノイズを含むデータでモデルを学習させる際にも有効性が示されている (Wang et al. 2018, 2019; Kumar et al. 2019)。しかしながら、これまでカリキュラムラーニングは要約タスクに応用されてこなかった。本研究の目的の1つはカリキュラムラーニングの要約タスクへの有効性を検証することである。

カリキュラムラーニングにおける学習データの順序の変更には、ノイズの量や難易度を表す指標が通常用いられる。学習はノイズの多いデータ群あるいは難易度の低いデータ群から始まり、徐々にノイズの少ないものあるいは難易度の高いものに移行する。ソートの際に使用する指標として、文生成タスク (Cirik et al. 2016) や翻訳タスク (Kocmi and Bojar 2017; Platanios et al. 2019; Zhou et al. 2020) においては、出力文の長さや難易度の指標として用いられている。ノイズを表す指標として、翻訳タスクにおいて2つの生成モデルの尤度差を用いて、カリキュラムラーニングに適用した研究がある (Wang et al. 2018, 2019; Kumar et al. 2019)。2つの生成モデルはノイズの少ないコーパスとノイズの多いコーパスでそれぞれ学習したSequence-to-sequenceモデルである。ここではノイズは翻訳元の文章と翻訳先の文章で対応の取れない情報を指している。

要約分野においては、新聞記事などのデータセットはソーシャルメディアやメールのデータセットに比べてノイズが少ないと考えられる。しかし、要約データは要約の長さ、Density（要約箇所が本文の全体か、一部分かを示す指標）、圧縮率、抽出率（要約の単語が本文に含まれる割合）などの性質がデータセットによって大きく異なる (Zhong et al. 2019)。異なるデータセットで学習したモデルは、ノイズのみでなく、こうした性質を考慮したモデルになってしまう問題がある。そのため、先行研究 (Wang et al. 2018, 2019; Kumar et al. 2019) を要約モデルに適用する場合、同じドメインでノイズの多寡のみが異なるデータセットが必要になるが、こうしたデータセットは存在しない。そこで本研究のもう1つの目的として、ノイズを含む単一コーパスからノイズを定量化してカリキュラムラーニングに適用する手法を提案する。

本研究では、ノイズを含む単一コーパスからノイズを定量化できるモデル Appropriateness Estimatorを提案する。本モデルは本文－要約の正しいペアと、ランダムに組み合わせたペアを分類する。ランダムに組み合わせたペアの要約は本文の内容を反映していない不適切なものである、不適切なペアと実際のペアを分類するように学習することで、Appropriateness Estimatorは本文－要約ペアの“適切性”が判別可能になる。この適切性をカリキュラムラーニングに適用する。すなわち、適切性をデータのソートに使用し、要約モデルの学習時に、学習データを不適切なペアから適切なペアへと徐々に変化させる。

本研究ではノイズを多く含む要約のデータセットとして、2つのデータセットで実験を行った。Enron Subject データセット (Zhang and Tetreault 2019) と Reddit Title データセット (Kim
狩野、谷口、大熊

要約データの適切性定量化を利用したカリキュラムラーニング

et al. 2019) である。両者とも学習データにはノイズが多く含まれるが、Enron Subject データセットの開発データセットと評価データセットは、人手により整理されたものである。一方 Reddit Title データセットの開発データセット、評価データセットはノイズを含む生のデータセットである。

本研究では、要約タスクに対するカリキュラムラーニングの有効性と、提案手法の効果を検証するため、3 つの要約モデルと 3 つのカリキュラムで実験を行った。要約モデルには、事前学習要約モデルと非事前学習要約モデルを用いる。事前学習モデルとして BART (Lewis et al. 2020), 非事前学習モデルとして Transformer (Vaswani et al. 2017) と Seq2seqWithAttention (Bahdanau et al. 2015) を採用する。実験において、カリキュラムラーニングおよび提案手法である Appropriateness Estimator は事前学習モデル、および非事前学習モデル両方の性能を改善した。

カリキュラムラーニングに用いられるカリキュラムにはいくつかの種類が存在する。学習データを徐々に変更するもの、学習データを徐々に増やしていくもの、学習データを徐々に減らしていくものなどがある。実験結果から、事前学習モデルに有効なカリキュラムと非事前学習モデルに有効なカリキュラムが異なることが判明した。事前学習モデルにとっては、終盤に少数のデータで Fine-tuning を行うカリキュラムが有効であり、非事前学習モデルにとっては序盤に多数のデータで汎化を行うことが有効であった。また、人手による評価を行い、提案手法である Appropriateness Estimator をカリキュラムラーニングに適用した方法が要約モデルの性能を向上させることを示した。

要約のデータの性質の評価に、抽出率（要約の単語が本文に含まれる割合） (Kim et al. 2019) や、含意判定確率 (Matsumaru et al. 2020) がこれまで用いられてきた。本研究で提案した適切性をこれらの性質や入力長、出力長などの統計量と比較し、適切性の性質を議論する。加えてこれまでカリキュラムラーニングに用いられてこなかった上記抽出率や含意判定確率が要約タスクにおけるカリキュラムラーニングに対して有効であることを示す。本論文の貢献は以下である。

- 3 つの要約モデルでカリキュラムラーニングの実験を行い、カリキュラムラーニングの要約タスクに対する有効性を示した。
- 単一のノイズを含む学習データから学習可能な、入力文と出力文の適切性を計算するモデル Appropriateness Estimator を提案し、実験により要約モデルの性能を向上させることを確認した。
- 異なるカリキュラムが事前学習モデル、非事前学習モデルの性能にどのような影響を与えるかを分析した。
2 関連研究

カリキュラムラーニングは学習データの順序を変更することで収束速度やモデル性能を上げる手法である (Bengio et al. 2009)。過去の研究 (Cirik et al. 2016) はこれを文生成タスクに応用し、2 種類のカリキュラムを提案した。Baby step カリキュラムと One-Pass カリキュラムである。後続の研究がカリキュラムラーニングを翻訳タスクに応用したが (Kocmi and Bojar 2017; Platanios et al. 2019; Wang et al. 2019; Zhou et al. 2020)，要約タスクに応用した研究はこれまで存在しなかった。

カリキュラムラーニングは元来、難易度を学習データをソートする手法であった。しかし、近年ではノイズの多さでソートを行う手法が提案されている。Wang et al. (2018) は、2 つのモデルを使って学習データのノイズを定量化する手法を提案した。同様のノイズ定量学手法を使い、Kumar et al. (2019) は強化学習を用いて学習に適している学習データのセグメントを適宜選択していく手法を提案した。要約分野においては、同じドメインでノイズの多寡が異なるコーパスは存在しないため、これらの手法を要約モデルに適用することはできない。そのため、本稿ではノイズを含む単一コーパスからノイズを定量化する手法を提案する。

Sequence-to-sequence 要約モデルの学習においては一般に、見出し、タイトル、件名などを要約とみなして学習する。しかしながら、これらのデータの内、メールやソーシャルメディアの件名およびタイトルは、非公式な場や、匿名性の高い場で書かれたものであるため、品質が担保されていない。Zhang et al. (2019) はメールの件名を生成するタスクを提案したが、元の Enron カーパスに含まれている件名がノイズを多く含んでいたため、新たにノイズの少ない評価データセットを構築している。Li et al. (2019) は、ルールや分類モデルを使ってレビューデータのノイズをフィルタリングした。新聞記事の見出しは、新聞社が不特定多数の読者に公開するものであるため、比較的要約として適切なものが多いものの、本文から推測が不可能な情報が含まれることが指摘されており、含意判定モデルによってそうしたデータをフィルターする手法が提案されている (Matsumaru et al. 2020)。

3 手法

本節ではまずカリキュラムラーニングの説明を行い、その後、提案手法である Appropriateness Estimator の説明を行う。カリキュラムラーニングは、学習データにある指標に基づいてソートして学習する手法であるが、その指標には難易度を表す指標やノイズ量を表す指標が使われる。提案手法である Appropriateness Estimator は後者のノイズ量計算のために使われる。

カリキュラムラーニング カリキュラムラーニングの概要を図 1 に示す。カリキュラムラーニングではまず、ある指標（e.g. ノイズの少なさ、出力長など）に基づいて学習データを昇順に
ソートする。次に、学習データをセグメントごとに分割する。分割したセグメントを使ってどのように学習を進めるかに応じて複数のカリキュラムが提案されている。

One-Pass カリキュラム (Cirik et al. 2016) は最も簡単な、あるいは最もノイズが多いセグメントから学習を開始し、モデルが収束すると、学習データとして次のセグメントのデータを使用する。Baby step カリキュラム (Cirik et al. 2016) は最も簡単な、あるいは最もノイズが多いセグメントから学習を開始し、徐々に学習データを増やしていく。これら2つのカリキュラムは少量のデータから学習を開始する。そのため、過学習を引き起こすリスクが存在する。過学習への対処のため、最初に全てのデータで学習を行い、徐々にデータを減らしていくカリキュラムでも実験を行う。これをNoise-Annealing カリキュラムと呼称する。いずれのカリキュラムにおいても、セグメント単位での学習を終える度、モデルのパラメータを保存し、開発データでの評価値が下がった場合には、最後に保存したパラメータに戻った後、次のセグメントでの学習を開始する。

Appropriateness Estimator 本稿では Appropriateness Estimator を提案する。本モデルはノイズを含む単一コーパスから学習可能な、ノイズを定量化するモデルである。図2にモデルの概要図を載せる。1節で述べたように、要約モデルの学習データには要約として不適切なもののが含まれる。ここでの不適切とは、本文からの推測が難しい情報や本文と無関係な情報が含まれているものや、記述が漠然としていて情報量に欠けているものを指す。要約モデルは本文と要約が実際の対であるかを学習することによって、本文と要約の適切性を定量化する。要約
図 2 Appropriateness Estimator の概要図と要約モデル学習への適用。

モデルの学習データに存在する本文 $s_i$ と要約 $t_k$ のペアを正例とし、ランダムにサンプリングされた本文 $s_i$ と要約 $t_k$ のペアを負例とする。ラベルを $c$ とし、ラベルが正の時 $c = 1$ であり、負の時 $c = 0$ とする。本モデルの学習タスクは本文-要約ペアのラベル $c$ の正負を予測することである。要約モデルの学習データに含まれるペアは全て正例とみなすが、1節で説明したように、それらは不適切なペアを含む。Early Stopping は一般に過学習を防ぐ目的で使われている手法であるが、ノイズデータに対する過学習を防ぐことも有効であることが示されている (Li et al. 2020)。本研究でも、同様に Early Stopping を利用することでモデルがノイズデータに過学習することを防ぐ。モデルの出力確率を $p(c|s_i, t_k)$ とし、損失関数 $L$ は以下のようにエントロピーを用いて計算する。

$$L_{rep} = -c \log p(c|s_i, t_k) - (1-c) \log (1 - p(c|s_i, t_k))$$  \hspace{1cm} (1)

学習後のモデルの出力確率 $p(c = 1|s_i, t_k)$ をペアの適切性と呼称する。ここで、$p(c = 1|s_i, t_k)$ が高いペアは低ノイズであり、$p(c = 1|s_i, t_k)$ が低いペアは高ノイズであることを仮定している。要約の学習データ全てを Appropriateness Estimator に入力し、適切性を計算する。ここで $c = 0$ のデータ、すなわちネガティブサンプリングにより得られたペアは含めず、元々の要約データに存在するペアのみを利用し、適切性に基づいて学習データをソートし、カリキュラムラーニングを適用する。
4 実験

4.1 データセット

Enron Subject データセット Enron データセット (Klimt and Yang 2004) は Enron Corporation の従業員のメールを集めたデータセットである。Zhang et al. (2019) はこれらのデータを件名生成タスク用に整理した。元のデータセットは評価に使用するにはノイズが多かったため、彼らは適切な件名を人手で再度アノテーションし、開発データセットと評価データセットを構築した。学習データとして 14,436 の件名とメールの本文を使用する。開発データセットとして 1,906 件、評価データセットとして 1,960 件のデータを使用する。

Reddit Title データセット Reddit Title データセット (Kim et al. 2019) はソーシャルメディアである Reddit のデータセットである。データは、Reddit の TIFU と呼ばれる掲示板 (Subreddit) から収集されたものである。TIFU は“today i f*** up”の略であり、投稿される文書は投稿者の失敗経験についてのものである。各投稿のタイトルを要約とみなし、要約の学習データとして使用する。学習データとして 71,113 件、開発データおよび評価データセットとして、各 3,951 件のデータを使用する。

4.2 モデル

Appropriateness Estimator 本稿では Appropriateness Estimator として Decomposable Attention (Parikh et al. 2016) を利用する。単語 Embedding の初期値として GloVe¹を使用する。Embedding と隠れ層の次元はそれぞれ 300 と 200 とする。学習に使うエポック数は 20 とする。

1 節と 3 節で説明したように、ランダムに割り当てた本文と出力文のペアを負例とし、元々の要約データセットに含まれるペアを正例とした。負例の数は正例の数と同数である。すなわち、Appropriateness Estimator の学習データ、開発データの数は要約モデルのそれぞれのデータ数の 2 倍となる。開発データセットにおいて最も F1-Score が高くなるエポックのモデルを要約モデルのカリキュラムラーニングに使用した。開発データセットにおける、最も高い F1-Score は Enron データセットで 0.94、Reddit データセットで 0.92 であった。

要約モデル 本稿では 3 つの要約モデルで実験を行った。その内の 1 つは事前学習モデルである BART (Lewis et al. 2020) である。残りの 2 つは、非事前学習モデルである Transformer (Vaswani et al. 2017) と Seq2seqwithAttention (Rush et al. 2015) である。

Seq2seqwithAttention および Transformer の隠れ層の次元は共に 256 とする。単語 Embedding の次元は Seq2seqwithAttention が 300、Transformer が 256 である。Seq2seqwithAttention の単語 Embedding の初期値には Appropriateness Estimator と同様に GloVe を使用する。ミニバッチ数

¹ https://nlp.stanford.edu/projects/glove/
は3つの要約モデルで共に64とする。Beam Searchのサイズは8とする。Seq2seqwithAttentionとTransformerの最適化にAdamを使用した。学習率は0.0007とした。BARTの最適化にはAdamW (Loshchilov and Hutter 2019)を使用した。学習率は3e-5、$\beta_1$ は0.9、$\beta_2$ は0.999、epsは1e-8とした。

ランダムシードを変えつつ、同じ実験を5回繰り返し、平均値を結果に使用する。評価にはROUGEのF1値 (Lin 2004) を使用する。各セグメントのエポック毎にValidationを行い、Validationの評価指標にはROUGE-1のF1値を使用する。

### 4.3 カリキュラムラーニング

カリキュラム 本稿では4つの条件下で実験を行う。3節で説明した3つのカリキュラムラーニングを使ったものと、カリキュラムラーニングを使わないものである。セグメント数として、5と10で実験を行った。学習データの各セグメント内のデータはシャッフルする。

データのソートに使用する指標 学習データをソートする指標として以下の2つを使用する。提案指標である適切性と、出力長である。出力長は難易度を表す指標として、カリキュラムラーニングで一般的に使われる指標である (Cirik et al. 2016; Platanios et al. 2019; Wang et al. 2019; Zhou et al. 2020)。本研究では、出力長は生成対象の要約文の単語数を指す。

### 5 結果

本節では、1節で述べた本研究の目的であるカリキュラムラーニングの要約タスクへの有効性と、提案手法であるAppropriateness Estimatorの有効性の検証結果について述べる。

#### 5.1 カリキュラムラーニングの要約タスクへの有効性

自動評価 表2に、3種のカリキュラムラーニングを要約モデルの学習に用いた場合と、用いなかった場合の実験結果を載せる。評価は自動評価指標であるROUGEを用いて行った。各モデル各データの最も高い性能（モデル1）は太字で記されている。†は同モデル同データ内で2番目に性能が高いもの（モデル2）との有意差を示している（p < 0.05）。有意差検定には、Bootstrap Resampling (Koehn 2004) を使用した。テスト時のモデルの出力の件数の半分に当たる回数、ランダムに結果を複数でサンプリングし、モデル1とモデル2の結果を比較する。これを1000回繰り返し、950回以上モデル1がモデル2の性能を上回った場合、有意とみなす。いずれのモデル、データにおいても、カリキュラムラーニングを使ったモデルが最も高い性能を示しており、要約タスクにもカリキュラムラーニングが有効であることが確認された。

人手評価 本稿では2つのBARTモデルで人手評価を行った。1つはNoise-Annealingカリキュラムおよび適切性で学習されたものであり、もう1つはカリキュラムラーニングなしで学習さ
| カリキュラムラーニング | トピック | 内容 |
|-------------------|---------|------|
| フィードバックと入力 | 幼児の認知 | 情報の取り入れ |
| フィードバックと入力 | 幼児の認知 | 情報の取り入れ |
| フィードバックと入力 | 幼児の認知 | 情報の取り入れ |

表 2 カリキュラムラーニングを使った要約モデルの実験結果。適切性は Appropriateness Estimator の出力確率を表す。R-1-F, R-2-F, R-L-F はそれぞれ ROUGE-1, ROUGE-2, and ROUGE-L の F1 値である。各モデル各データの最も高い性能は太字で記されており、†は同モデル同データ内で 2 番目に性能が高いものとの有意差を示している (p < 0.05)。有意差検定には、Bootstrap Resampling (Koehn 2004) を使用した。

過去に生成された要約が同じである場合、それを除去し、計 90 の生成要約を Enron データセット、Reddit データセットそれぞれに対して得た。アノテーターはどちらの要約が良いかを情報量、流暢性の観点から評価する。情報量は生成要約が本文の重要な情報をどの程度含んでいるかを表しており、流暢性は文法の観点から生成された文が正しいかを判断する。表 3 に結果を示す。表は、Noise-Annealing と適切性で学習されたモデルは情報量、流暢性両方の観点でより高い評価を得たことを示している。統計的な有意性を評価するため、“良い”と “どちらかといえば良い” の票数を統合し χ²二乗検定を行った。

153
表3 人手評価の結果。CL ありはNoise-Annealingカリキュラム使用時の結果を指し、CL なしはカリキュラムラーニング無しの結果を指す。† and ‡ は、BART（CL あり）が“良い”と“どちらかといえば良い”の票を多く得る確率をχ2二乗検定で検定したものである（† p < 0.01, ‡ p < 0.05）。

5.2 Appropriateness Estimatorの有効性

 Appropriateness Estimator の適切性を利用したカリキュラムラーニングは、3つのモデル全てで、カリキュラムラーニングなしのモデルと比べ、性能を向上させた。出力長を利用したカリキュラムラーニングと比較した場合、適切性を利用した場合は Enron Subject データにおいてより上げ幅が大きかった。これは、Reddit Titleデータセットでは、テストデータが学習データ同様未整備であるのに対し、Enron Subjectデータのテストデータが、人手で整備されたノイズの少ないデータであることに関係していると考えられる。適切性を用いたカリキュラムラーニングは、Noise-AnnealingカリキュラムとOne-Passカリキュラムの場合、学習の終盤においてノイズの少ないデータで学習を行う。これにより、学習モデルが低ノイズの高品質なデータに Fine-tuning されているため、より低ノイズのテストデータセットにおいて性能を発揮すると考えられる。

6 考察

本節では、本実験で用いた3つのカリキュラムが与える性能への影響、翻訳タスクとの相違点、適切性の性質について議論する。

6.1 カリキュラムごとの相違点

今回本研究では、3種類のカリキュラムで実験を行った。One-PassカリキュラムおよびNoise-Annealingカリキュラムは学習の終盤で少数のデータで Fine-tuning を行い、Noise-Annealingカリキュラムは学習の初期に多様なデータでの汎化を行うという特徴がある。こうしたカリキュラムの特徴と、カリキュラムラーニングなしと比べた際の要約モデルの性能に与える影響のまとめを表4に載せる。One-PassおよびNoise-AnnealingカリキュラムはBARTモデルの性能を向上させたが、Baby stepカリキュラムは性能を悪化させた。考えられることは、BARTは事前
表 4 3つのカリキュラムの特徴とカリキュラムラーニングなしと比べた場合の性能差。

学習モデルであるため、改めて汎化を行う必要はなく、Fine-tuning が重要であるという点である。非事前学習モデルにおいては Noise-Annealing カリキュラムのみが要約モデルの性能を向上させた。これは、事前学習モデルとは対照的に、非事前学習モデルにとっては汎化が重要であることを示唆している。

6.2 翻訳タスクとの相違点

Baby step カリキュラムは、徐々に学習データを増やしていくカリキュラムである。過去の翻訳タスクにおける研究では出力長を同様のカリキュラムに使用し、翻訳モデルの性能を向上させた (Kocmi and Bojar 2017; Platanios et al. 2019; Zhou et al. 2020) が、本稿の実験設定では要約モデルの性能は低下した。

要約タスクと翻訳タスクの出力文の違いとして、翻訳タスクでは、出力文の質と出力文の文長に相関はないが、要約タスクではその限りではないということである。翻訳タスクにおいては、出力文は入力文の長さに比例するため、入力文が短い場合は必然的に出力文も短くなり、翻訳の正確性には関係がない。しかしながら、要約タスクにおいては、出力文と入力文は比例せず、短すぎる要約は十分な情報を含んでいないと考えられる。そのため、少数のデータで学習をはじめると Baby step カリキュラムおよび One-Pass カリキュラムを非事前学習モデルに適用した場合、低品質データに対する過学習を引き起こし、性能が低下すると考えられる。

6.3 適切性が高い／低い本文－要約ペアの具体例

表 5, 6 に適切性が高い、あるいは低い入力文 (本文) と出力文 (要約) のペアの例を載せる。表にある適切性の低い例の出力文は本文における情報と関係がないことがわかる。表 6 の下段の例では、適切性が低いペアにおいて、出力文の続きが入力文に書かれており、出力文の情報が入力文に書かれていない。対照的に、適切性が高いペアでは出力文は本文の内容を反映している。

6.4 適切性の表す性質

本節では、適切性が表す性質について議論する。6.2 節で述べたように、要約タスクにおける出力長はノイズの量と相関すると考えられるが、ノイズの量そのものではない。表 5 や表 6 で
### Table 5 適切性が高い／低い本文－要約 (Subject) ペアの例 (Enron Subject データセット).

| Subject | 本文 | 適切性 |
|---------|------|------|
| TW/ Lonestar Ward and Pecos Counties interconnect bi-directional | The following is a level “A” cost estimate to make TW/ Lonestar existing interconnects bi-directional. TW/ Lonestar at Ward County (50 to 60 mmcf/d) According to Operations this is already bi-directional. The only things are required on this one is to take the flapper out of the check-valve and blow down the gas in 5.33 miles of 12”. Cost of gas loss & labor = $8,000 TW/ Lonestar at Pecos County (100 mmcf/d) A): TW/ Lonestar interconnect Scope: On this one we need a bi-directional valve skid using the existing meter run. Cost of material & labor = $195,000 B): Pecos Compressor Station In order to make this interconnect bi-directional we also need to make the station (two-compressor units) bi-directional. Scope: Install outlet from Lonestar I/C to inlet filter with 12” piping & valves. Unit discharge would be modified to tie in to West Texas 20” Cost for material & labor = $330,000. If you need more accurate costs (B -release) please let me know. | 0.99 |

### Table 6 適切性が高い／低い本文－要約 (Title) ペアの例 (Reddit Title データセット).

| Title | 本文 | 適切性 |
|-------|------|------|
| Asking if my roommate had any plans for mother’s day. | Yesterday, technically, I was at home making myself a nice meal because I couldn’t be with my family for mother’s day due to distance. As I’m preparing my dinner, my roommate came into the kitchen. Thinking I would be a good roommate and strike up some passing conversation, I asked him if he had any plans for mother’s day, to which he replied that his mom had died just last month. He hasn’t exactly made this super well known in the house, but I had seen a fb post of his last month mentioning this. I felt like the most insensitive asshole ever and apologized as well as I could. but I’ll always feel bad about that one. | 0.97 |
| Leaving a 12-pack of beer in the bottom of a shopping cart in the grocery store parking lot. | I went back to get it 30 minutes later and it was still there : ) | 0.01 |
示した例のように、出力長が長くてもノイズとみなせる要約も存在する。本節では適切性を、他にノイズと相関すると考えられる4つの性質と比較して議論する。その内の2つは、入力長と出力長である。3つ目は一致する単語の割合である。Appropriateness Estimatorは、本文－要約ペアの適切さを一致する単語によって判定している可能性が高い。要約に含まれる単語が本文に含まれる割合を抽出率と定義し、抽出率と適切性との関係性について議論する。4つ目の性質は合意関係である。松丸ら（Matsumaru et al. 2020）は、合意判定器を用いて要約データのフィルタリングを行った。合意判定器の確率値と適切性の違いについて議論する。上記4つの性質と適切性とのピアソノの相関係数を表7に載せる。

また、上記4つ内の、入力長、抽出率、合意確率を用いてカリキュラムラーニングを行った際の結果を表8に載せる。この時、カリキュラムとして用いたのは、表2の実験で有効であったOne-PassカリキュラムとNoise-Annealingカリキュラムである。要約モデルには最も性能の高いBARTを使用した。表2の実験と同様に、実験はランダムシードを変えて5回行った際の平均値を載せる。有意差検定は、表2の実験と同様Bootstrap Resampling（Koehn 2004）を用いて行う（p < 0.05）。

適切性と入力長および出力長との関係 表7より、適切性と、入力長及び出力長との相関係数は0.2未満であった。これは、適切性が入力長や出力長とは異なるテキストの性質を表していることを示している。表5および表6に示したように、要約のデータセットには、本文や要約が短くない場合にも要約として不適切なものが含まれる。また、表8より、入力長をカリキュラムラーニングの指標として用いた場合の性能は、他の指標と比べて悪くなった。これは入力長と要約データとしての適切さが比例しないことを示唆している。

適切性と抽出率との関係 Appropriateness Estimatorは、本文と要約のペアの正しさを学習している。そのため、要約に含まれる単語が本文に含まれるかを元に判定している可能性が高い。そこで、要約に含まれる単語が本文に含まれる割合（これを抽出率と呼称する）と適切性の関係性を議論する。抽出率と適切性のピアソノ相関係数は、表7にあるように高い。これは、適切性が単語の一致率を反映していることを示す。ただし、適切性が抽出率に比べ有利な点は単語の一致だけではなく、関連する単語に対応できる点と、一般的な視点の重要度を相対的に下げられる点である。Appropriateness Estimatorは他の異なる要約を対象として判断するよう学習するため、要約の内容が一般的で情報量が少ない場合、他の本文にも含まれる可能性が

| データセット  | 出力長 | 入力長 | 抽出率 | 含意確率 |
|------------|--------|--------|--------|----------|
| Enron Subject | 0.151 | 0.079 | 0.711 | 0.244 |
| Reddit Title | 0.156 | 0.018 | 0.572 | 0.174 |

表7 適切性と各統計量との相関係数（ピアソノ）
表8 入力長、抽出率、含意確率をカリキュラムラーニングを適用した要約モデルの実験結果。各モデルとデータの最も高い性能は太字で記されており、†は同モデル同データ内で適切性以外で最も性能が高いものとの有意差を示している（p < 0.05）。有意差検定には、Bootstrap Resampling（Koehn 2004）を使用した。

| 要約モデル | カリキュラム | ソート指標 | Reddit Title | Enron Subject |
| --- | --- | --- | --- | --- |
| BART | 無し | ー | R-1-F | R-1-F |
| | One-P | 适切性 | 0.276† | 0.137 |
| | | 入力長 | 0.241 | 0.112 |
| | | 抽出率 | 0.251 | 0.222 |
| | | 含意確率 | 0.243 | 0.210 |
| Noise-Annealing | 适切性 | 0.271 | 0.132 |
| | 入力長 | 0.260 | 0.129 |
| | 抽出率 | 0.269 | 0.236 |
| | 含意確率 | 0.267 | 0.235 |

表9 抽出率が高いが、適切性が相対的に低い本文−要約ペアの例（Reddit Title データセット）。

高くなり、正しい要約と誤った要約の区別がつきにくくなる。そのため、こうした本文−要約ペアには相対的に低い適切性が与えられる。表9にある例の抽出率は高いが、書かれている内容が一般的で情報量を欠いているため、適切性が相対的に低くなっている。適切性は全体的に高い数値であるため、この数値は全体の下位23.2%に該当する。

抽出率をカリキュラムラーニングに適用した場合の結果を表8に載せる。抽出率はNoise-Annealingカリキュラムにおいては、適切性と同等の性能であったが、One-Passカリキュラムにおいては適切性が有意に上回った。これは前述した点で適切性が抽出率に比べ有利であるからと考えられる。また、抽出率を使った結果は、カリキュラムラーニングを使わない場合の性能を上回っており、この実験結果はこれまでカリキュラムラーニングに用いられて来なかった抽出率が要約タスクにおいて有効であることを示している。
適切性と含意確率との関係 松丸ら (Matsumaru et al. 2020) は、含意判定器を用いて要約学習データセットから不適切な本文と要約のペアを除去した。含意確率と適切性の関係性についても本節で議論する。含意判定器には、松丸ら (Matsumaru et al. 2020) と同様に、RoBERTa (Liu et al. 2019) を Multi-Genre Natural Language Inference (MultiNLI) データセット (Williams et al. 2018) で学習させたものを使用し、本文が要約を含意する確率を計算する。なお、松丸らは、MultiNLI で学習させたモデルを、要約データセットごとに独自にアノテーションしたデータセットで更に Fine-tuning したモデルを用いているが、本研究では対象要約データセットが異なるため、MultiNLI で学習させたモデルを用い、Fine-tuning は行わない。

適切な要約の要件には、要約の内容が本文にも書かれているという含意性の他に、本文の重要な箇所が要約に反映されているべきという情報量の観点がある。Appropriateness Estimator は一般的に情報量が少ない要約には相対的に低い適切性を与える。これは要約の記述内容に情報量が少ない場合、本文に対する正しい要約であるか、他の本文に対する要約であるかを判定するのが難しくなるからである。一方含意判定器は含意性のみを判定するため、こうした例を除去することはできない。表 10 の上段の例では、含意確率は高くなってしまうが、要約として適切な情報が含まれているとは言い難い。一方 Appropriateness Estimator は、主語と目的語の逆転や否定肯定の反転などを区別することはできないため含意性に関して含意判定器が高性能に行うことができる。

また、既存の含意判定器は要約データセット向けに作られたものでは無いため、要約データセットに対しては適切な含意判定ができないという問題も存在する。一般的に含意判定器は同程度の長いの文同士を比較するが、要約の場合は複数の文にまたがる情報を含意することがある。既存の含意判定学習データセット MultiNLI を用いて学習された含意判定器で含意確率が高いと判定された本文−要約のペアは、表 10 の中段の例に例のように、同程度の長さで要約の内容が本文に存在していることが多い。これに対し、下段の例のように複数文にまたがる内容の要約は含意判定器では含意確率が低いと判定される。含意判定器の確率値と本文の単語数（入力長）とのピアソン相関係数は、Enron Subject データセットにおいて -0.30、Reddit Title データセットにおいて -0.42 であり、いずれも負の相関性を示していた。この相関関係は本文が長すぎる時には、既存の含意判定器が上手く機能しないという仮説を支持している。

先行研究 (Matsumaru et al. 2020) では、含意判定器を要約データセットのフィルタリングに使用したが、この時含意判定器を要約データセット向けに新たにアノテーションされたものを用いて Fine-tuning されており、正確に要約データの含意性を判定するためには要約データセットに対する含意関係のアノテーションが必要となる。提案手法である Appropriateness Estimator は、このような要約データセットのアノテーションを必要とせず、含意判定器やその学習データセットが無い言語に対しても有効である。

表 8 の結果は、含意確率をカリキュラムラーニングに用いた場合の性能が、Enron Subject
I woke up this morning, went out to the barn to discover that one of my kittens had died :-( then I was milking a cow and she kicked me so hard in the hand I thought she broke it, then I went to jump on my boyfriends back so he could give me a piggy back ride to the house, and smashed my bad knee into an air tank in the milk house. Now I can’t bend my leg, and to top everything off, as I was cutting my breakfast sausage with a fork, it slipped and poured sausage covered with syrup all over me... I should have just gone back to bed...

I got too drunk and bought nascar ’15 on xbox 360. seriously, who the fuck does that?

Obligatory “not actually today”: this happened last summer. I (then 23) was bouncing around on a trampoline with my two brothers, then 10 and 6, and my girlfriend. My brothers asked if I knew how to backflip, and I jokingly bragged “yes” although I’d never done it before. Now, I mastered (trampoline) front flips pretty quickly many years ago, but always found backflips difficult. basically, I never had the guts to go through with it and never had enough spin to go past landing on my back. not this time. I figured I’d at least make an honorable effort, and I kicked myself into a spin as I jumped. （後略）

表 10 含意確率が低い、あるいは高い本文－要約ペアの例（Reddit Title データセット）。

データセット、Reddit Title データセット両方において、適切性と One-Pass カリキュラムを用いた場合の性能を有意に下回っていることを示している。これは、前述したように含意確率が情報量を考慮できない点と、既存の含意確率モデルが要約データセットに対して必ずしも有効でないことによると考えられる。一方カリキュラムラーニングを用いない場合に比べ性能は向上しており、これまでカリキュラムラーニングに用いられて来なかった含意確率という指標が、要約データセットに対して有効であることを示している。

6.5 どのセグメントで最も高性能となるか？

One-Pass カリキュラム、および Noise-Annealing カリキュラムは両方とも少量のデータで Fine-tuning を行う。どのセグメントで最も要約モデルの開発データの評価性能がよくなるかを調査することで、どのセグメントが最も Fine-tuning に適しているかを明らかにする。

表 11 に開発データセットの評価性能 (ROUGE-1-F) が最も高くなるセグメントの番号を記す。モデルには BART を使用する。表には 5 回行った実験の平均値と標準偏差を記してある。
表11 開発データセットにおける評価指標（ROUGE-1-F）が最大になるセグメントを表示している。ランダムシードを変えた時の5回の実験の平均値と標準偏差を記してある。セグメントの数は10である。

セグメント数は10であり、10番目のセグメントが最も出力長が長い、あるいは最も適切性が高いセグメントである。出力長を用いてカリキュラムラーニングを行った際、要約モデルはより早いセグメントで最も開発データセットにおける評価性能が高くなっているが、提案した適切性を用いた時には、より後段のセグメントで最も評価性能が高くなっている。これは出力が長すぎるデータは要約モデルのFine-tuningに適切でないことを示唆している。適切な要約は、情報の取捨選択が適切に行われていないからである。一方、本稿で提案した適切性は、高いものがよりFine-tuningに適した、要約モデルの学習に適したデータになっていると考えられる。

7 結論

本稿では、これまで要約タスクにおいて取り組まれてこなかったカリキュラムラーニングを要約タスクに適用し、その有効性を検証した。本稿では、実験設定として、3つのカリキュラム（One-Pass, Baby step, Noise-Annealing）とカリキュラムを用いない場合で実験を行った。また、実験対象の要約モデルとして1種類の事前学習モデルBARTと2種類の非事前学習モデルTransformerとSeq2seqWithAttentionを用いた。結果、事前学習モデルにおいてはNoise-Annealingカリキュラムのみが要約モデルの性能向上を示し、また、事前学習モデルにおいてはOne-PassカリキュラムとNoise-Annealingが性能を向上させた。これらの実験結果から、少量のデータでFine-tuningを行うことが事前学習要約モデルに重要であり、汎化が非事前学習要約モデルにとって重要であることを結論付けた。より効率的にFine-tuning対象のデータを探す手法を考案することが今後の課題である。

また本稿では既存の翻訳タスクで取り組まれてきたノイズを指標にしたカリキュラムラーニングを要約タスクに適用することを試みた。その際、要約の学習データには翻訳タスクにあるようなノイズの少ないあるいは多いコーパスが存在しないため、単一のノイズを含む学習データから学習可能なノイズ定量化モデルAppropriatenessEstimatorを提案し、カリキュラムラーニングの実験を行った。実験においては、テストデータが人手で整備されたEnronSubjectデータ
タセットにおいてより効力を発揮することを明らかにした。本稿で提案した疑似負例により学習したモデルをカリキュラムラーニングに応用する手法は、翻訳タスクや対話タスクに応用可能だと考えられる。

また、要約タスクにカリキュラムラーニングを応用する際、適切性と出力長の他に、抽出率や含意判定確率も有効であることを示した。今後の課題はこれらの指標を効果的に組み合わせる手法を開発することである。今回、ノイズの多い要約タセットに対するカリキュラムラーニングの有効性を調査したが、新聞記事などのノイズの比較的少ない要約タセットに対する調査は今後の課題となる。

謝辞

東京工業大学奥村・船越研の奥村先生、船越先生、上垣外先生、学生の皆様及び、産業技術総合研究所の高村先生には研究の方針に関して様々なアドバイスを頂きました。ここに感謝の意を表します。本研究は、The 16th Conference of the European Chapter of the Association for Computational Linguistics (EACL 2021) の発表“Quantifying Appropriateness of Summarization Data for Curriculum Learning”で報告した研究に、適切性の性質についての追加実験の結果や考察を加筆し、修正したものです。

参考文献

Bahdanau, D., Cho, K., and Bengio, Y. (2015). “Neural Machine Translation by Jointly Learning to Align and Translate.” In Bengio, Y. and LeCun, Y. (Eds.), Proceedings of the 3rd International Conference on Learning Representations, (ICLR 2015).

Bengio, Y., Louradour, J., Collobert, R., and Weston, J. (2009). “Curriculum Learning.” In Proceedings of the 26th Annual International Conference on Machine Learning (ICML 2009), pp. 41-48.

Cirik, V., Hovy, E., and Morency, L.-P. (2016). “Visualizing and Understanding Curriculum Learning for Long Short-Term Memory Networks.” arXiv preprint arXiv:1611.06204.

Kim, B., Kim, H., and Kim, G. (2019). “Abstractive Summarization of Reddit Posts with Multi-level Memory Networks.” In Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (NAACL 2019), pp. 2519-2531.

Klimt, B. and Yang, Y. (2004). “The Enron Corpus: A New Dataset for Email Classification Research.” In Proceedings of the 15th European Conference on Machine Learning, pp. 217-226.
Kocmi, T. and Bojar, O. (2017). “Curriculum Learning and Minibatch Bucketing in Neural Machine Translation.” In Proceedings of the International Conference Recent Advances in Natural Language Processing, (RANLP 2017), pp. 379–386.

Koehn, P. (2004). “Statistical Significance Tests for Machine Translation Evaluation.” In Proceedings of the 2004 Conference on Empirical Methods in Natural Language Processing (EMNLP 2004), pp. 388–395.

Kryscinski, W., Keskar, N. S., McCann, B., Xiong, C., and Socher, R. (2019). “Neural Text Summarization: A Critical Evaluation.” 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 2019), pp. 540–551.

Kumar, G., Foster, G., Cherry, C., and Krikun, M. (2019). “Reinforcement Learning based Curriculum Optimization for Neural Machine Translation.” In Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (NAACL 2019), pp. 2054–2061.

Lewis, M., Liu, Y., Goyal, N., Ghazvininejad, M., Mohamed, A., Levy, O., Stoyanov, V., and Zettlemoyer, L. (2020). “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 (ACL 2020), pp. 7871–7880.

Li, J., Li, H., and Zong, C. (2019). “Towards Personalized Review Summarization via User-Aware Sequence Network.” In the 33rd AAAI Conference on Artificial Intelligence, (AAAI 2019), pp. 6690–6697.

Li, M., Soltanolkotabi, M., and Oymak, S. (2020). “Gradient Descent with Early Stopping is Provably Robust to Label Noise for Overparameterized Neural Networks.”. Vol. 108 of Proceedings of Machine Learning Research, pp. 4313–4324.

Lin, C.-Y. (2004). “ROUGE: A Package for Automatic Evaluation of Summaries.” In Text Summarization Branches Out, pp. 74–81.

Liu, Y., Ott, M., Goyal, N., Du, J., Joshi, M., Chen, D., Levy, O., Lewis, M., Zettlemoyer, L., and Stoyanov, V. (2019). “RoBERTa: A Robustly Optimized BERT Pretraining Approach.” arXiv preprint arXiv:1907.11692.

Loshchilov, I. and Hutter, F. (2019). “Decoupled Weight Decay Regularization.” In Proceedings of the 7th International Conference on Learning Representations (ICLR 2019).

Matsumaru, K., Takase, S., and Okazaki, N. (2020). “Improving Truthfulness of Headline Generation.” In Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics (ACL 2020), pp. 1335–1346.
Nallapati, R., Zhou, B., dos Santos, C., Gulçehre, m., and Xiang, B. (2016). “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.

Parikh, A., Täckström, O., Das, D., and Uszkoreit, J. (2016). “A Decomposable Attention Model for Natural Language Inference.” In Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing (EMNLP 2016), pp. 2249–2255.

Platanios, E. A., Stre tcu, O., Neubig, G., Poczos, B., and Mitchell, T. (2019). “Competence-based Curriculum Learning for Neural Machine Translation.” In Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (NAACL 2019), pp. 1162–1172.

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 (EMNLP 2015), pp. 379–389.

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, L., and Polosukhin, I. (2017). “Attention Is All You Need.” In Guyon, I., Luxburg, U. V., Bengio, S., Wallach, H., Fergus, R., Vishwanathan, S., and Garnett, R. (Eds.), Advances in Neural Information Processing Systems 30 (NIPS 2017), pp. 5998–6008.

Wang, W., Caswell, I., and Chelba, C. (2019). “Dynamically Composing Domain-Data Selection with Clean-Data Selection by “Co-Curricular Learning” for Neural Machine Translation.” In Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics (ACL 2019), pp. 1282–1292.

Wang, W., Watanabe, T., Hughes, M., Nakagawa, T., and Chelba, C. (2018). “Denoising Neural Machine Translation Training with Trusted Data and Online Data Selection.” In Proceedings of the 3rd Conference on Machine Translation: Research Papers (WMT 18), pp. 133–143.

Williams, A., Nangia, N., and Bowman, S. (2018). “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 (NAACL 2018), pp. 1112–1122.

Zhang, R. and Tetreault, J. (2019). “This Email Could Save Your Life: Introducing the Task of Email Subject Line Generation.” In Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics (ACL 2019), pp. 446–456.

Zhong, M., Wang, D., Liu, P., Qiu, X., and Huang, X. (2019). “A Closer Look at Data Bias in Neural Extractive Summarization Models.” In Proceedings of the 2nd Workshop on New Frontiers in Summarization, pp. 80–89.
Zhou, Y., Yang, B., Wong, D. F., Wan, Y., and Chao, L. S. (2020). “Uncertainty-Aware Curriculum Learning for Neural Machine Translation.” In Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics (ACL 2020), pp. 6934–6944.

略歴

狩野 竜示：2013年東京大学大学院情報理工学系研究科修了。同年富士ゼロックス株式会社入社。2019年より東京工業大学工学部博士課程に在籍。現在、富士フィルム株式会社において、自動択訳の研究に従事。言語処理学会会員。

谷口 友紀：2004年神戸大学大学院自然科学研究科修士課程修了。同年、富士ゼロックス株式会社入社。現在、富士フィルム株式会社において、医療言語処理、情報抽出の研究に従事。言語処理学会会員。

大熊 智子：1996年慶應義塾大学政策・メディア研究科修士課程修了。同年、富士ゼロックス株式会社入社。現在、富士フィルム株式会社において、医療言語処理、情報抽出の研究に従事。言語処理学会理事、文部科学省学会理事、博士（学術）。

（2021年7月1日受理）
（2021年10月1日再受付）
（2021年11月9日採録）