情報処理学会論文誌 Vol.62 No.3 891–902 (Mar. 2021)

不均衡分類問題としての小説の段落境界推定

飯倉 陸[1,a)] 岡田 真[1] 森 直樹[1]

受付日 2020年6月1日, 採録日 2020年12月1日

概要：小説の創作支援に関する研究は多岐にわたる．本研究では，読み手が文章の内容理解を深めるための重要な文章技法の 1 つである段落分けに焦点を当てた．この段落分けは，対象とする文どうしの間における，段落としての境界の存在の有無に関する分類問題としてとらえることが可能である．しかしその場合，一般に段落の数は文の数と比較して少ないため，データの不均衡性がボトルネックとなる．我々はこの問題に対処するため，BERT に不均衡データの分類問題に対して頑健な損失関数を導入した．そして本研究のために新たに作成したデータセットを対象とした実験を通して，Focal Loss および Dice Loss を導入した場合に，従来の BERT と比較して有意に高い精度が得られることを実験的に確認した．また，モデルに対する入力文の範囲を拡張することが段落境界を推定するために有効であることを明らかにした．

キーワード：小説の創作支援，段落分け，不均衡分類問題，BERT，コスト考慮型学習

Automatic Paragraph Segmentation of Novels as Imbalanced Classification

Riku Iikura[1,a)] Makoto Okada[1] Naoki Mori[1]

Received: June 1, 2020, Accepted: December 1, 2020

Abstract: There are various studies on creation support for writing novels. In this study, we focus on paragraph segmentation, which is one of the important writing techniques for readers to deepen their understanding of the texts. The paragraph segmentation can be considered as a classification problem regarding the presence or absence of a boundary as a paragraph between the target sentences. However, in that case, the data imbalance becomes a bottleneck because the number of paragraphs is generally smaller than the number of sentences. In order to deal with this problem, we have introduced several loss functions which is robust for the imbalanced classification in BERT. We confirmed experimentally that significantly higher accuracy is obtained when using the model with Focal Loss and Dice Loss compared to the conventional BERT through experiments on the dataset newly created for this study. In addition, it was clarified that expanding the range of input sentences to the model is effective for estimating paragraph boundaries.

Keywords: creation support, paragraph segmentation, imbalanced classification, BERT, cost-sensitive learning

1. はじめに

人工知能を芸術分野に適用することへの気運が高まりつつある中で，小説や漫画 [1], [2]，俳句 [3], [4] に代表される詩歌などの文学作品を扱う研究が多数なされるようになった．特に小説に関して，2012 年には星新一作品のようなショートショートを自動生成するグランドチャレンジとして「きまぐれ人工知能プロジェクト 作家ですのよ」が開始された [5]．それにともない，募集対象として人工知能により創作された作品を明示的に許可した新人文学賞である「星新一賞」[*1]が創設された．これは小説家としての人工知能が誕生することへの期待を象徴した事例であるといえる．

小説の創作に対する工学的な研究は，文章自動生成と創作支援に大別される．文章自動生成に関する研究では，機

[1] 大阪府立大学大学院工学研究科
Graduate School of Engineering, Osaka Prefecture Uniersity, Sakai, Osaka 599–8231 Japan
[a)] iikura@ss.cs.osakafu-u.ac.jp

[*1] https://hoshiaward.nikkei.co.jp

械翻訳や文章要約などの文生成タスクにおいて高い精度を記録した sequence-to-sequence モデル [6] を用いたものが多数を占める．これまでに，独立した説明文を連結し 1 つの物語を生成するシステム [7] や，物語の筋書きを生成する操作とそれを物語へと変換する操作を組み合わせた階層的な物語生成システム [8] などが提案されている．また人工知能を利用した小説の創作支援に関する研究には，読み手が物語における幸福度の推移としてどのようなものを好むかという観点を導入したプロット作成支援システム [9], [10] や，物語の結末は一意に定まるものでなく複数の可能性が考慮可能であることに着目した，物語の多様な結末を生成する研究 [11] などがある．そのほかにも，ABS モデルを用いて物語を生成する手法 [12], [13] などが提案されている．これらの研究は，小説の自動生成という壮大な目標を実現させるために欠かすことのできない段階的な研究として位置付けされる．

我々はこの小説の創作支援と関連して，文章を書くうえで重要な技法の 1 つである段落分けに着目した．この段落分けとは，文章の可読性を高めるために，文章を形式段落に分割する操作である．文章における段落の重要性について論じたものに，関らの研究 [14] がある．この研究では，文章のレイアウトとしての段落表示が読み手の内容理解に与える影響が調査された．具体的には，正しく段落分けされた文章，故意に誤って段落分けされた文章，段落分けされていない文章を読解後，それぞれに対する内容の理解度を比較する実験から，適切な段落設定が読み手の内容理解を促すうえで重要であることが結論づけられている．

この段落分けについて，学術論文などの説明的あるいは論理的な文章と，小説や随筆文などの文学的文章では，やや性質が異なることが一般的に知られている．これは，論理的文章では読み手に対して正確かつ簡潔に情報を開示することに重点が置かれるのに対して，文学的文章では読み手に感動や情緒を与えることに重点が置かれるという本質的な目的の相違によるものである [15]．前者には，1 つの段落では 1 つのトピックについてのみが語られ，それぞれがトピック・センテンスと称される内容の核となる文を中心に構成されるという特徴がある．文章を構成する各段落における内容や主張が一貫していることが強く要求され，科学技術文章を対象にした段落の一貫性を数値的に定義する研究 [16] なども存在する．一方で後者は，時間経過にともなう話題や情景の推移に基づいており，一般的にトピック・センテンスを有していない [15]．そのため小説の段落分けには書き手が一律に従うべき共通的な規則がなく，高い技量が必要とされる非常に難しい操作であるといえる．小説における段落分けを補助するシステムの需要は高い一方で，その難しさから先行研究はほとんど存在しない．

以上を背景として，本研究では小説の創作支援の観点から，既存の小説に対する高精度な段落分けを実現することを目的とする．我々はこのタスクを，対象とする文どうしが同一の形式段落に所属するかどうかという 2 クラス分類問題としてとらえた．しかしその場合，文の数に対する段落の数は小さいため，データ数における不均衡性を考慮する必要がある．そこで我々は，様々な自然言語処理のタスクにおいて高い精度が示されている汎用言語モデル Bidirectional Encoder Representations from Transformer（BERT）[17] の損失関数として，不均衡データの分類問題に対して頑健性が確認されている Focal Loss [18] や Dice Loss [19] を使用することで推定精度の向上を図った．さらに，入力文が含む情報を増加させることによる精度の向上を期待して，モデルに対する入力文の範囲の拡張についても試みた．

本研究における貢献は以下のとおりである．

- 小説文を自動的に形式段落に分割するモデルを新たに構築し，訓練データとして単一の作者により書かれた作品のみで構成されたデータよりも，複数の作者により書かれた作品からなるデータを用いた場合の方が，モデルの性能が向上することを実験的に確認した．
- 小説文の段落分けを任意の連続する 2 文間における段落境界の有無に関する不均衡分類問題としてとらえ，複数の異なる損失関数を導入した BERT を適用した．結果として，従来のテキストセグメンテーション手法を上回る精度で段落境界の推定に成功した．また，Focal Loss および Dice Loss を導入した場合に，従来の BERT と比較して有意に高い精度が得られることを実験的に確認した．
- 段落境界を推定するうえで，提案モデルに対する入力文の範囲を一定の範囲まで拡張することで精度が向上し，より多く文章の情報を与えることの有効性を示唆する結果が得られた．

2. 関連研究

本章では，本研究に関連するテキストセグメンテーションと，不均衡データに対する分類問題について述べる．

2.1 テキストセグメンテーション

文章をトピックなどを基準にした意味的なまとまりに分割する操作は，一般にテキストセグメンテーションとして知られている．このタスクは，計算機による自然言語の意味理解の観点から，文章要約や質問応答など自然言語処理分野の様々なタスクに応用される重要なタスクである．これまでに提案されたテキストセグメンテーション手法は，教師なしアルゴリズムと教師ありアルゴリズムに基づくものに大別される．

教師なしのテキストセグメンテーション手法の 1 つとして，TextTiling [20] がある．この手法は，特定の単語が同一のセグメントに頻出することを利用し，それらのベクト

ルから各セグメントの類似度を算出する．Glavaš ら [21] は，単語の埋め込み表現と短い文章の意味的な関連性の尺度を利用して文章の意味的な関連性グラフを構築する教師なしアルゴリズムを提案した．このグラフのノードは文を表し，2 文の間のエッジは文どうしが意味的に類似していることを示している．テキストのセグメンテーションは，隣接する文の最大クリークを見つけることで決定される．

また，教師あり学習による手法には Recurrent Neural Network（RNN）の一種である Long Short-Term Memory（LSTM）を用いるモデルがある [22]．このようなモデルは時間の経過にともなう情報の流れを制御することで，入力系列を効率的にモデル化することが可能である．Badjatiya ら [23] はこれらに対して Attention 機構を導入し，セグメンテーションのための文章中の各文の相対的な重要性を学習する Attention-based CNN-BiLSTM を提案した．

本研究では上述の先行研究に対して，テキストセグメンテーションを，セグメントの数と文の数における不均衡性を考慮すべき不均衡データの分類問題として扱う．

2.2 不均衡データに対する分類問題

各クラスのデータ数が不均衡な分類問題には，大別するとリサンプリング手法とコスト考慮型学習の 2 つのアプローチがある．

リサンプリング手法とは，不均衡なデータセットに変更を加えることでバランスのとれた分布を生成する手法である．基本的には多数派クラスをアンダサンプリング [24], [25] する，あるいは少数派クラスをオーバサンプリングすることでデータ数における不均衡性を解消する．最も単純なオーバサンプリング手法は少数派クラスのインスタンスを無作為に複製する方法であるが，データの分布が冗長になるため過学習を引き起こす原因になりうる．この問題に対してデータ合成を使用した基本的なアプローチである SMOTE [26] が提案された．SMOTE はデータセットのバランスをとるために無作為にシードサンプルを選択し，シードサンプルとその近隣の 1 つとの間に線形補間を適用して新しいサンプルを合成する．

一方で，コスト考慮型学習は訓練データの分布を変更する代わりに，各データサンプルに対して異なる重みを付与した損失関数を適用した学習を通じて分類器自体を改善する手法である．この手法は，特に画像処理分野における物体検知問題を扱う研究と関連づけられることが多い．これは，物体検知問題では画像の大部分を背景が占めるため，少数派クラスとなる特定の物体を識別するためにはそのラベルの不均衡性を解消する必要があるからである．これまでに，Dice 係数に基づく損失関数を用いて医療画像をセグメンテーションする研究 [27], [28] などがあり，3 章で詳述する Focal Loss や Dice Loss など，不均衡データに対して頑健な損失関数が複数提案されている．我々はこれまでに，Focal Loss を損失関数として採用した BERT を小説の段落境界推定問題に適用し，その手法の有効性について示した [29], [30]．また，Li ら [31] は自然言語処理分野における不均衡データの分類問題として，品詞タグ付けや固有表現抽出などのタスクに対して Dice Loss を導入した BERT を適用し，その有効性を明らかにした．本研究ではこの Dice Loss を導入した BERT を小説の段落境界推定問題に適用し，その有効性について検討する．

3. 段落境界推定モデル

本章では，本研究において提案する段落境界推定モデルに関して，そのベースモデルである BERT と分類において使用した損失関数について詳述する．

3.1 BERT

BERT は，複数の双方向 Transformer [32] に基づく汎用言語モデルであり，入力された単語系列および，含まれる各単語に対応する分散表現を出力する．BERT は大規模コーパスに対して事前学習を施すことで，言語モデルとしての性能を向上させている．事前学習には，トークン [MASK] で入力文の一部が置換された文に対してその元単語を予測するように訓練する Masked word prediction と，2 文を入力としてその連続性を正しく識別するように訓練する Next sentence prediction のタスクが用いられる．本研究で扱うタスクは，対象とする 2 文を入力してその 2 文間に段落としての境界が存在しているかどうかを判別するため，この Next sentence prediction タスクと強く関連する．

BERT を極性判定や文章分類などのクラス分類タスクに対して適用するためには，入力文の先頭に付与される [CLS] トークンに対して出力されるベクトルを分類器への入力とする．特に 2 文をモデルに入力する場合は，2 文の間に [SEP] トークンを挿入して結合し，単一のシーケンスとして扱う．BERT では，文あるいは文対を分散表現に変換したのち，それを入力として多層パーセプトロンによって分類や回帰などの応用タスクを解く．このとき，学習済みモデルを基に転移学習し解決すべきタスクに適用させることが可能である．

3.2 損失関数

本節では，本研究で扱うタスクのボトルネックであるデータの不均衡性に対処するために採用した，いくつかの損失関数について詳述する．以降では便宜的に，表記する損失関数は 2 クラス分類問題を想定したものとする．サンプル数 N の訓練データセット X に含まれる各サンプル $x_i \in X$ が属する正解クラスを表すバイナリラベルを $y_i = [y_{i0}, y_{i1}]$，それぞれに対する推定確率を $p_i = [p_{i0}, p_{i1}]$ と定める．ここで $y_{i0}, y_{i1} \in \{0, 1\}$，$p_{i0}, p_{i1} \in [0, 1]$ であり，$p_{i0} + p_{i1} = 1$ である．

3.2.1 Cross Entropy Loss

BERT がクラス分類問題を解く場合に使用する基本的な損失関数は以下の式 (1) で表される Cross Entropy Loss である．

$$\mathrm{CE} = -\frac{1}{N}\sum_{i}\sum_{j\in\{0,1\}} y_{ij}\log(p_{ij}). \tag{1}$$

一般的に，不均衡データを対象とする分類問題には先の Cross Entropy Loss に対して重み $\alpha = [\alpha_0, \alpha_1]$ を導入することで式 (2) のようにバランス調整し，各クラスのサイズに応じた重要性を考慮することが可能である．実用的な重みの値として各クラスに含まれるデータ数の逆数が採用される場合が多いが，本研究では $\alpha_0, \alpha_1 \in [0,1]$，$\alpha_0 + \alpha_1 = 1$ とし，ハイパーパラメータとして扱う．これは，重みの値として単にデータ数の逆数を採用する場合より詳細にパラメータを調整することで，モデルの推定精度を向上させるためである．

$$\text{weighted-CE} = -\frac{1}{N}\sum_{i}\sum_{j\in\{0,1\}} \alpha_j y_{ij}\log(p_{ij}). \tag{2}$$

Madabushi ら [33] は，BERT の最終層である全結合層における損失関数を重み α 付き Cross Entropy Loss に変更することでプロパガンダの識別に関する不均衡データの分類問題を解きその有効性を示した．

3.2.2 Dice Loss

不均衡データの分類モデルを評価する際に用いられる指標の 1 つに $F1$ 値がある．Dice 係数（Sørensen-Dice coefficient: DSC）は，この $F1$ 値指向の統計指標である．Dice 係数は一般に，2 つの集合の類似性を測定する指標であるが，不均衡分類問題と関連して医療分野における患部画像のセグメンテーションなどで使用される場合もある [34]．Li ら [31] は Dice 係数と $F1$ の関係について，以下のように示した．まず 2 つの集合 A，B が与えられたとき，Dice 係数は式 (3) で与えられる．

$$\mathrm{DSC} = \frac{2|A\cap B|}{|A| + |B|}, \tag{3}$$

本研究では，集合 A をモデルによって正例であると判定されたサンプルの集合とし，集合 B を真の正例サンプルの集合とする．ここで真陽性（True Positive: TP），偽陽性（False Positive: FP），偽陰性（False Negative: FN）を用いて，Dice 係数と $F1$ 値の関係は式 (4) のようになる．

$$\begin{aligned}\mathrm{DSC} &= \frac{2\mathrm{TP}}{2\mathrm{TP} + \mathrm{FN} + \mathrm{FP}} \\ &= \frac{2\frac{\mathrm{TP}}{\mathrm{TP}+\mathrm{FN}}\frac{\mathrm{TP}}{\mathrm{TP}+\mathrm{FP}}}{\frac{\mathrm{TP}}{\mathrm{TP}+\mathrm{FN}} + \frac{\mathrm{TP}}{\mathrm{TP}+\mathrm{FP}}} \\ &= \frac{2\mathrm{Precision}\times\mathrm{Recall}}{\mathrm{Precision} + \mathrm{Recall}} \\ &= F1,\end{aligned} \tag{4}$$

以上のような定義に基づいて，各サンプル x_i に対する Dice 係数の値は以下の式 (5) で与えられる．

$$\mathrm{DSC} = \frac{\sum_i 2y_{i1}p_{i1} + \epsilon}{\sum_i y_{i1} + \sum_i p_{i1} + \epsilon}, \tag{5}$$

ここで ϵ はゼロ除算を防ぐための定数であり，本研究では $\epsilon = 10^{-5}$ とした．

Milletari ら [19] はこの Dice 係数における分母の各項を 2 乗した目的関数を提案し，それを最大化するための損失関数として Dice Loss を以下の式 (6) のように定めた．

$$\mathrm{DL} = \frac{1}{N}\left(1 - \frac{\sum_i 2y_{i1}p_{i1} + \epsilon}{\sum_i y_{i1}^2 + \sum_i p_{i1}^2 + \epsilon}\right). \tag{6}$$

3.2.3 Focal Loss

Focal Loss は Lin ら [18] により提案された，Cross Entropy Loss を動的にスケーリングする損失関数である．上述の重み α 付き Cross Entropy Loss は各クラスのサイズに応じた重要性を考慮することを可能にするが，各クラスに対する識別の難易度を区別することはできない．それに対して Focal Loss は，識別が容易な例からのエラーの寄与を減衰させるための係数を導入する．これによりモデルは識別が難しい例に効果的に焦点を合わせることが可能になる．具体的には式 (7) に表すように，チューニング可能な $\gamma \geq 0$ を含んだ項 $(1-p_{ij})^\gamma$ を式 (2) に導入している．

$$\mathrm{FL} = -\frac{1}{N}\sum_{i}\sum_{j\in\{0,1\}} \alpha_j y_{ij}(1-p_{ij})^\gamma \log(p_{ij}), \tag{7}$$

ここで $\gamma = 0$ のとき，Focal Loss は重み α 付き Cross Entropy Loss と同等である．

4. データセット構築

本研究における実験のために，電子図書館の「青空文庫」[*2]において管理されている小説から新たにデータセットを作成した．モデルの学習において使用する作品の作者の多様性が，段落境界の推定に与える影響を確認するという目的のもと，訓練データとして単一の作者の作品のみで構成されたもの（訓練 A）と，複数の作者の作品で構成されたもの（訓練 B）の 2 種類を作成した．訓練 A では夏目漱石の作品のみを使用し，訓練 B では夏目漱石と芥川龍之介，太宰治の 3 人の作者の作品を使用した．それぞれに使用した作品群については，後述する．検証データとしては，夏目漱石の『それから』を使用して作成した．またテストデータとしては，異なる作者ごとに精度を検証しモデルの汎用性について検討するため，夏目漱石の『こころ』，芥川龍之介の『歯車』，太宰治の『女生徒』をそれぞれ用いて 3 種類作成した．以降では，それぞれのテストデータを便宜的に，テスト A，B，C と呼称する．

図 1 に，夏目漱石の作品『坊っちゃん』より引用した形式段落の例を示す．本研究では，改行されたのち文頭にお

[*2] https://www.aozora.gr.jp

表 1 データセットに含まれるサンプルの例
Table 1 Examples of samples included in the datasets.

#	ラベル	Segment 1	Segment 2
1	0	親譲りの無鉄砲で小供の時から損ばかりしている。	小学校に居る時分学校の二階から飛び降りて一週間ほど腰を抜かした事がある。
2	0	小学校に居る時分学校の二階から飛び降りて一週間ほど腰を抜かした事がある。	なぜそんな無闇をしたと聞く人があるかも知れぬ。
3	1	小使に負ぶさって帰って来た時、おやじが大きな眼をして二階ぐらいから飛び降りて腰を抜かす奴があるかと云ったから、この次は抜かさずに飛んで見せますと答えた。	親類のものから西洋製のナイフを貰って奇麗な刃を日に翳して、友達に見せていたら、一人が光る事は光るが切れそうもないと云った。
4	0	寝巻のまま腕まくりをして談判を始めた。	「なんでバッタなんか、おれの床の中へ入れた」
5	0	「なんでバッタなんか、おれの床の中へ入れた」	「バッタた何ぞな」と真先の一人がいった。

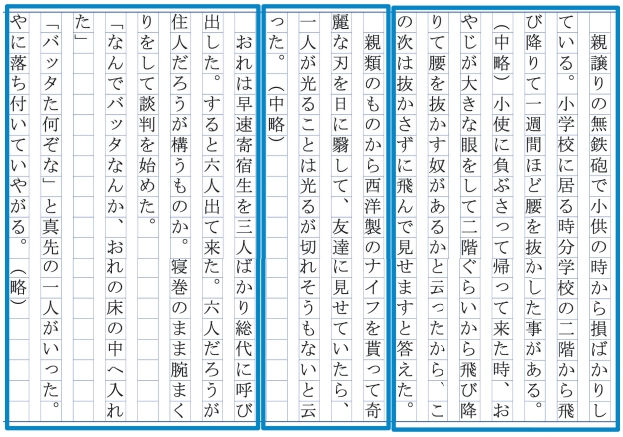

図 1 夏目漱石『坊っちゃん』における形式段落の例
Fig. 1 Examples of paragraphs from “Botchan” written by Soseki Natsume.

表 2 データセットの統計
Table 2 Statistics for each dataset.

データセット	ラベル数（0：1）	ラベル比（0：1）
訓練 A（単一の作者）	29078：5309	5.477：1.0
訓練 B（複数の作者）	28760：5048	5.697：1.0
検証（夏目漱石『それから』）	5308：716	7.413：1.0
テスト A（夏目漱石『こころ』）	4325：733	5.900：1.0
テスト B（芥川龍之介『歯車』）	737：122	6.041：1.0
テスト C（太宰治『女生徒』）	854：71	12.03：1.0

いて 1 文字の字下げがなされた箇所を形式段落の境界として定義する．これをふまえて，対象とする 2 文の間に形式段落としての境界が存在する場合にはラベル 1（正例）を付与し，存在しない場合にはラベル 0（負例）を付与した．

表 1 に，定義に基づいて作成したデータセットに含まれるサンプルの例を示す．一般的に会話文は直前の文から改行され，字下げされずに鉤括弧（「）から開始されるため，上記の定義に基づくと例 4，5 のような会話文は形式段落として計数されないことに注意する．ここで Segment 1 と Segment 2 の間に [SEP] トークンを挿入して結合した単一のシーケンスを，モデルへの入力形式とする．

文を単語単位に分割するための形態素解析には MeCab [35] を使用した．また**表 2** および**表 3** に，作成し

表 3 訓練データセットに含まれる作品
Table 3 Works included in the training dataset.

データ	著者	作品	ラベル数（0：1）	ラベル比（0：1）
A	夏目漱石	『坊っちゃん』	2513：199	12.63：1.0
		『門』	3292：586	5.618：1.0
		『三四郎』	6193：755	8.203：1.0
		『彼岸過迄』	4016：550	7.302：1.0
		『道草』	3791：1222	3.102：1.0
		『明暗』	9273：1997	4.643：1.0
B	夏目漱石	『坊っちゃん』	2513：199	12.63：1.0
		『門』	3292：586	5.618：1.0
		『三四郎』	6193：755	8.203：1.0
	芥川龍之介	『あばばばば』	209：37	5.649：1.0
		『アグニの神』	206：67	3.075：1.0
		『秋』	263：47	5.596：1.0
		『舞踏会』	95：27	5.407：1.0
		『玄鶴山房』	319：59	5.649：1.0
		『鼻』	118：42	2.810：1.0
		『手巾』	128：59	2.169：1.0
		『芋粥』	335：53	6.321：1.0
		『犬と笛』	137：56	2.446：1.0
		『邪宗門』	559：134	4.172：1.0
		『地獄変』	380：96	3.958：1.0
		『神々の微笑』	254：43	5.907：1.0
		『河童』	1038：163	6.368：1.0
		『蜘蛛の糸』	33：13	2.538：1.0
		『魔術』	133：43	3.093：1.0
		『蜜柑』	50：7	7.143：1.0
		『魔術』	133：43	3.093：1.0
		『南京の基督』	173：52	3.327：1.0
		『おぎん』	116：23	5.043：1.0
		『羅生門』	123：28	4.393：1.0
		『蜃気楼』	154：38	4.053：1.0
		『トロッコ』	107：30	3.567：1.0
		『杜子春』	178：67	2.657：1.0
	太宰治	『眉山』	202：54	3.7407：1.0
		『竹青』	195：24	8.125：1.0
		『富嶽百景』	379：63	6.016：1.0
		『グッドバイ』	471：141	3.340：1.0
		『八十八夜』	502：43	11.67：1.0
		『恥』	253：15	16.87：1.0
		『走れメロス』	431：25	17.24：1.0
		『皮膚と心』	344：23	14.96：1.0
		『乞食学生』	1154：78	14.79：1.0
		『古典風』	296：54	5.481：1.0
		『女神』	142：62	2.290：1.0
		『人間失格』	1140：448	2.545：1.0
		『律子と貞子』	160：20	8.0：1.0
		『佐渡』	492：31	15.87：1.0
		『斜陽』	1918：849	2.259：1.0
		『水仙』	371：48	7.729：1.0
		『雀』	187：35	5.343：1.0
		『津軽』	2656：203	13.08：1.0
		『ヴィヨンの妻』	361：108	3.343：1.0

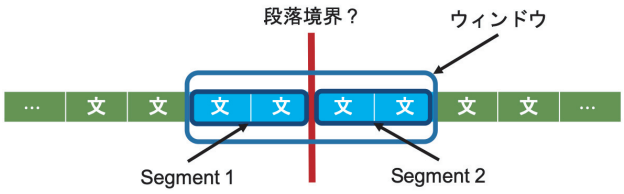

図 2 ウィンドウサイズの説明図（ウィンドウサイズ 4 の場合）
Fig. 2 Illustration of window size 4.

表 4 各ウィンドウサイズのデータセットに含まれるサンプルの単語系列長
Table 4 Sequence length of samples included in each dataset with different window size.

ウィンドウサイズ	データ	単語系列長		
		最小	最大	平均 (標準偏差)
2	訓練 A	2	252	39.88 (22.1)
	訓練 B	3	560	41.35 (31.2)
	検証	5	217	38.59 (18.6)
	テスト A	4	139	40.44 (16.6)
	テスト B	5	93	36.62 (14.2)
	テスト C	3	235	39.32 (26.4)
4	訓練 B	7	660	83.64 (51.0)
	検証	14	268	78.16 (29.1)
	テスト A	20	227	81.86 (25.9)
	テスト B	18	147	74.20 (21.7)
	テスト C	9	298	79.50 (40.3)
6	訓練 B	12	784	125.87 (69.0)
	検証	31	311	117.74 (37.9)
	テスト A	41	281	123.29 (34.0)
	テスト B	28	201	111.74 (27.8)
	テスト C	19	331	119.54 (50.8)
8	訓練 B	17	1163	168.04 (86.0)
	検証	53	375	157.31 (45.9)
	テスト A	57	321	164.70 (41.4)
	テスト B	55	258	149.24 (33.5)
	テスト C	40	378	159.57 (59.9)

た各データセットと使用した作品における各ラベルの数とその比率を示す．これらの表から，各作品ごとに 1 つの段落あたりの文数が大きく異なることが分かる．つまり，同一作者であっても作品ごとに段落分けの性質が大きく異なることが推察される．

本研究では提案モデルに対する入力として適切な文の範囲を検討するため，入力文の範囲の異なる複数のデータセットを作成した．以降では便宜的に，入力文の範囲をウィンドウサイズと呼称する．ここでウィンドウサイズとは各入力サンプルの前半部分（Segment 1）の文数と後半部分（Segment 2）の文数の和を表す．**図 2** にウィンドウサイズ 4 の場合の入力文の範囲を図示する．この例では，入力文 Segment 1，2 としてそれぞれ 2 文ずつ，合計 4 文を使用している．**表 4** に各ウィンドウサイズのデータセットに含まれるサンプルの単語系列長の最小値および最大値，平均値と標準偏差を示す．また，各ウィンドウサイズのデータセットにおける各ラベルの数および比率はすべて同一である．

表 5 に，表 1 における例 1，2，3 について，異なるウィンドウサイズで表現したサンプルを示す．ここで各作品の冒頭文を含むサンプルおよび最終文を含むサンプルでは，Segment 1 あるいは Segment 2 に含まれる文数が他のサンプルより少なくなることに注意する．

5. 数値実験

本章では，1) 異なる訓練データを用いてモデルを学習させた場合の精度比較，2) 適切な損失関数の選定，3) モデルに対する入力文の範囲の検討という 3 つの観点から，提案モデルの有効性を実験的に検討する．

5.1 使用する訓練データの比較検討

本節では，4 章で説明した，構成する作品群の作者が単一である訓練データ（訓練 A）と，複数の作者の作品群により構成された訓練データ（訓練 B）を使用した場合の段落境界推定実験における精度について比較する．ここで使用する各データのウィンドウサイズは 2 である．

5.1.1 実験設定

本実験で使用したモデルは，損失関数として Cross Entropy Loss を採用した従来の BERT に基づく分類器（BERT+CE）である．使用した BERT のパラメータ設定値はそれぞれ，最大単語系列長 512，訓練バッチサイズ 16，学習率 2×10^{-5}，訓練エポック数 3 である．また，本研究では分類する際に使用する損失関数の相違による段落境界の推定精度を比較するという目的から，最終的な分類器には標準的な 3 層の多層パーセプトロンを採用した．事前学習済みモデルとしては，一般に公開されている日本語 Wikipedia 全文を学習した日本語 BERT モデル（BERT-base_mecab-ipadic-bpe-32k_whole-word-mask）[*3] を使用した．

また，各モデルの推定精度を比較するための評価指標として，$F1$ 値と P_k を採用した．P_k とは Beeferman ら [37] が提案したテキストセグメンテーションの評価指標である．距離 k だけ離れた 2 つの文が同一のセグメントに属しているかどうかをシステム出力結果と正解データの両方で計算する．そして両方の一致しない割合が P_k のスコアとなり，値が小さいほどモデルの性能が高いことを示す．Koshorek ら [22] に従って，k を正解セグメントのサイズの平均値の半分に設定した．

5.1.2 結果と考察

表 6 に，10 回の推定実験の結果得られた各評価指標の平均値とその標準偏差を示す．テスト A に関して，訓練 B で学習したモデルの方が，訓練 A で学習したモデル

[*3] https://github.com/cl-tohoku/bert-japanese

表 5 ウィンドウサイズの異なる各データセットに含まれるサンプルの例（‘/’ は文の区切りを表す）

Table 5 Examples of samples included in each dataset with different window sizes (‘/’ represents a sentence break).

ウィンドウサイズ	#	ラベル	Segment 1	Segment 2
4	1	0	親譲りの無鉄砲で小供の時から損ばかりしている。/	小学校に居る時分学校の二階から飛び降りて一週間ほど腰を抜かした事がある。/ なぜそんな無闇をしたと聞く人があるかも知れぬ。/
	2	0	親譲りの無鉄砲で小供の時から損ばかりしている。/ 小学校に居る時分学校の二階から飛び降りて一週間ほど腰を抜かした事がある。/	なぜそんな無闇をしたと聞く人があるかも知れぬ。/ 別段深い理由でもない。/
	3	1	と囃したからである。/ 小使に負ぶさって帰って来た時、おやじが大きな眼をして二階ぐらいから飛び降りて腰を抜かす奴があるかと云ったから、この次は抜かさずに飛んで見せますと答えた。/	親類のものから西洋製のナイフを貰って奇麗な刃を日に翳して、友達に見せていたら、一人が光る事は光るが切れそうもないと云った。/ 切れぬ事があるか、何でも切ってみせると受け合った。/
6	1	0	親譲りの無鉄砲で小供の時から損ばかりしている。/	小学校に居る時分学校の二階から飛び降りて一週間ほど腰を抜かした事がある。/ なぜそんな無闇をしたと聞く人があるかも知れぬ。/ 別段深い理由でもない。/
	2	0	親譲りの無鉄砲で小供の時から損ばかりしている。/ 小学校に居る時分学校の二階から飛び降りて一週間ほど腰を抜かした事がある。/	なぜそんな無闇をしたと聞く人があるかも知れぬ。/ 別段深い理由でもない。/ 新築の二階から首を出していたら、同級生の一人が冗談に、いくら威張っても、そこから飛び降りる事は出来まい。/
	3	1	弱虫やーい。/ と囃したからである。/ 小使に負ぶさって帰って来た時、おやじが大きな眼をして二階ぐらいから飛び降りて腰を抜かす奴があるかと云ったから、この次は抜かさずに飛んで見せますと答えた。/	親類のものから西洋製のナイフを貰って奇麗な刃を日に翳して、友達に見せていたら、一人が光る事は光るが切れそうもないと云った。/ 切れぬ事があるか、何でも切ってみせると受け合った。/ そんなら君の指を切ってみろと注文したから、何だ指ぐらいこの通りだと右の手の親指の甲をはすに切り込んだ。/
8	1	0	親譲りの無鉄砲で小供の時から損ばかりしている。/	小学校に居る時分学校の二階から飛び降りて一週間ほど腰を抜かした事がある。/ なぜそんな無闇をしたと聞く人があるかも知れぬ。/ 別段深い理由でもない。/ 新築の二階から首を出していたら、同級生の一人が冗談に、いくら威張っても、そこから飛び降りる事は出来まい。/
	2	0	親譲りの無鉄砲で小供の時から損ばかりしている。/ 小学校に居る時分学校の二階から飛び降りて一週間ほど腰を抜かした事がある。/	なぜそんな無闇をしたと聞く人があるかも知れぬ。/ 別段深い理由でもない。/ 新築の二階から首を出していたら、同級生の一人が冗談に、いくら威張っても、そこから飛び降りる事は出来まい。/ 弱虫やーい。/
	3	1	新築の二階から首を出していたら、同級生の一人が冗談に、いくら威張っても、そこから飛び降りる事は出来まい。/ 弱虫やーい。/ と囃したからである。/ 小使に負ぶさって帰って来た時、おやじが大きな眼をして二階ぐらいから飛び降りて腰を抜かす奴があるかと云ったから、この次は抜かさずに飛んで見せますと答えた。/	親類のものから西洋製のナイフを貰って奇麗な刃を日に翳して、友達に見せていたら、一人が光る事は光るが切れそうもないと云った。/ 切れぬ事があるか、何でも切ってみせると受け合った。/ そんなら君の指を切ってみろと注文したから、何だ指ぐらいこの通りだと右の手の親指の甲をはすに切り込んだ。/ 幸ナイフが小さいのと、親指の骨が堅かったので、今だに親指は手に付いている。/

よりも $F1$ 値の平均値は高く，P_k の平均値は低くなった．t 検定により有意差を確認すると，$F1$ 値に関しては p-value $= 6.1 \times 10^{-9}$，P_k に関しては p-value $= 5.7 \times 10^{-7}$ であり，両方の評価指標について有意水準 0.05 で有意差が認められた．またテスト B に関しても同様に，訓練 B で学習したモデルの方が，訓練 A で学習したモデルより

表 6 各訓練データを使用した場合の評価指標の値（平均と標準偏差）
Table 6 Experimental results for each training data (mean and std.).

	テスト A（夏目漱石『こころ』）		テスト B（芥川龍之介『歯車』）		テスト C（太宰治『女生徒』）	
データ	$F1$ 値	P_k	$F1$ 値	P_k	$F1$ 値	P_k
訓練 A（単一の作者）	0.522 (0.010)	0.271 (0.0042)	0.732 (0.016)	0.183 (0.0094)	**0.349**(**0.049**)	**0.354**(**0.031**)
訓練 B（複数の作者）	**0.569**(**0.0087**)	**0.254**(**0.0048**)	**0.755**(**0.011**)	**0.166**(**0.0089**)	0.337 (0.020)	0.355 (0.0092)

も $F1$ 値の平均値は高く，P_k の平均値は低いという結果が得られた．t 検定により確認したそれぞれの評価指標に対する p-value は，$F1$ 値では p-value $= 0.0027$，P_k では p-value $= 0.00080$ であり，両方の評価指標について有意水準 0.05 で有意差が認められた．

一方でテスト C に関しては，訓練 B で学習したモデルよりも，訓練 A で学習したモデルの方が $F1$ 値の平均値は高く，P_k の平均値は低くなった．t 検定により有意差を確認すると，$F1$ 値に関しては p-value $= 0.51$，P_k に関しては p-value $= 0.91$ であり，いずれの評価指標についても有意水準 0.05 で有意差は認められなかった．

本実験における結果は，モデルが学習するための訓練データとして単一の作家の作品のみを使用するよりも，複数の作家の作品を使用することで，段落境界の推定における精度の向上が可能であることを示唆しているといえる．

5.2 適切な損失関数の選定

本節では，4 章で説明したウィンドウサイズ 2 のデータセットを対象とした段落境界推定実験について詳述する．この実験では，提案モデルとベースラインモデルの推定精度を比較し，導入する損失関数として最適なものを検討することを目的とする．なお，本実験では訓練データとして複数の作者の作品により構成されたものを使用する．

5.2.1 実験設定

実験に使用した各モデルは以下のとおりである．

Koshorek's method ベースラインモデル．Koshorek ら [22] により提案された，LSTM に基づくテキストセグメンテーション手法である．Word2Vec [36] により得られた，文に含まれる単語に対する分散表現を双方向 LSTM 入力し，出力を文の分散表現として用いる．本研究では，一般に公開されている Word2Vec の事前学習済みモデルとして，日本語 Wikipedia [*4] 全文から学習した日本語 Wikipedia エンティティベクトル [*5] を使用した．

BERT+CE 損失関数として Cross Entropy Loss を採用した従来の BERT に基づく分類器である．使用した BERT のパラメータ設定値はそれぞれ，5.1 節における設定値と同一である．

BERT+DL BERT の損失関数として Dice Loss を採用したモデルである．BERT のパラメータ設定値は BERT+CE と同一である．

BERT+FL BERT の損失関数として Focal Loss を採用したモデルである．BERT のパラメータ設定値は BERT+CE と同一である．また，Focal Loss のパラメータ α，γ の値は検証データに対するグリッドサーチにより決定した．

また，各モデルの推定精度を比較するための評価指標としては 5.1 節と同様，$F1$ 値と P_k を採用した．

表 7 α を変動させた場合の検証データにおける各評価指標の値
Table 7 Value of each evaluation index in the verification data when α is changed.

α_0	α_1	$F1$ 値	P_k
0.10	0.90	0.573	0.267
0.20	0.80	0.578	0.268
0.30	0.70	**0.591**	**0.250**
0.40	0.60	0.564	0.267

表 8 $\alpha_0 = 0.30$，$\alpha_1 = 0.70$ と定め，γ を変動させた場合の検証データにおける各評価指標の値
Table 8 When $\alpha_0 = 0.30$ and $\alpha_1 = 0.70$ are set, the value of each evaluation index in the verification data when γ is changed.

γ	$F1$ 値	P_k
0	0.591	0.250
1.0	0.590	0.259
2.0	**0.597**	**0.244**
3.0	0.589	0.250
4.0	0.574	0.271

5.2.2 結果と考察

まず，**表 7** および**表 8** に BERT+FL におけるハイパーパラメータ α，γ の影響について示す．表 7 のように各クラスに付与する重みを表す α を変更した結果，$\alpha_0 = 0.30$ のとき検証データにおいて最も高い精度を示した．また，この結果をふまえて $\alpha_0 = 0.30$ として γ を調整した結果，$\gamma = 2.0$ のとき最も高い推定精度となった．

続いて**表 9** に，10 回の推定実験の結果得られた各評価指標の平均値とその標準偏差を示す．実験の結果，BERT に基づくモデルはベースラインモデルである Koshorek らの手法による推定精度を上回る推定精度を示した．また BERT+DL あるいは BERT+FL を用いた場合，各テストデータに対して BERT+CE よりも高い推定精度が得られ

*4 https://ja.wikipedia.org
*5 https://github.com/singletongue/WikiEntVec

表 9 各手法に対する評価指標の値（平均と標準偏差）
Table 9 Experimental results for each model (mean and std.).

	テスト A（夏目漱石『こころ』）		テスト B（芥川龍之介『歯車』）		テスト C（太宰治『女生徒』）	
モデル	$F1$ 値	P_k	$F1$ 値	P_k	$F1$ 値	P_k
Koehorek's method	0.419 (0.14)	0.333 (0.10)	0.506 (0.17)	0.276 (0.087)	0.162 (0.060)	0.425 (0.13)
BERT+CE	0.569 (0.0087)	0.254 (0.0048)	0.755 (0.011)	0.166 (0.0089)	0.337 (0.020)	0.355 (0.0092)
BERT+DL	0.571 (0.0059)	**0.253**(**0.0027**)	**0.778**(**0.011**)	**0.150**(**0.0075**)	0.353 (0.035)	**0.345**(**0.016**)
BERT+FL	**0.578**(**0.0080**)	**0.253**(**0.0058**)	0.769 (0.011)	0.160 (0.0076)	**0.379**(**0.033**)	0.347 (0.016)

表 10 各ウィンドウサイズに対する評価指標の値（平均と標準偏差）
Table 10 Experimental results for each window size (mean and std.).

	テスト A（夏目漱石『こころ』）		テスト B（芥川龍之介『歯車』）		テスト C（太宰治『女生徒』）	
ウィンドウサイズ	$F1$ 値	P_k	$F1$ 値	P_k	$F1$ 値	P_k
2	0.571 (0.0059)	0.253 (0.0027)	0.778 (0.011)	0.150 (0.0075)	0.353 (0.035)	0.345 (0.016)
4	0.600 (0.013)	0.240 (0.0057)	0.783 (0.024)	0.148 (0.012)	0.438 (0.0042)	**0.322**(**0.024**)
6	**0.613**(**0.0075**)	**0.232**(**0.0051**)	**0.784**(**0.012**)	**0.141**(**0.0083**)	**0.471**(**0.023**)	0.330 (0.021)
8	0.601 (0.014)	0.247 (0.010)	0.722 (0.032)	0.192 (0.023)	0.439 (0.015)	0.357 (0.017)

る傾向があることも確認された．

ここで BERT+CE と BERT+DL がそれぞれ各テストデータに対して得た各評価指標の値について比較する．すべてのテストデータに対して BERT+DL により得られた $F1$ 値の平均値は高く，P_k の平均値は低くなった．t 検定により有意差を確認すると，テスト B における $F1$ 値に関しては p-value $= 2.6 \times 10^{-4}$，P_k に関しては p-value $= 6.1 \times 10^{-4}$ であり，両方の評価指標について有意水準 0.05 で有意差が認められた．しかし，他のテスト A およびテスト C における各評価指標に関して，有意差は認められなかった．

また BERT+CE と BERT+FL の比較においても同様に，すべてのテストデータに対して BERT+DL により得られた $F1$ 値の平均値は高く，P_k の平均値は低くなった．t 検定の結果，テスト A における $F1$ 値について p-value $= 0.025$，テスト B における $F1$ 値について p-value $= 0.011$ であり，有意水準 0.05 で有意差が認められたが，他のテストデータにおける各評価指標に関して有意差は認められなかった．

BERT+DL と BERT+FL により得られた推定精度について，テスト A における $F1$ 値とテスト C における $F1$ 値に関しては BERT+FL の方が BERT+DL よりも高い平均値をとり，テスト B における P_k およびテスト C における P_k の平均値はいずれも BERT+DL の方が BERT+FL よりも低い値をとった．ここでテスト A の $F1$ 値について p-value $= 0.035$，テスト B における P_k について p-value $= 0.014$ であり，それぞれ有意水準 0.05 で有意差が認められた．しかし他の各評価指標の値については，有意差は認められなかった．

上述の結果から，BERT に対して損失関数として Dice Loss あるいは Focal Loss を導入した場合に，段落境界推定において同程度に高い精度が得られることを確認した．したがって，サンプル数の少ないラベル 1 に対して重みを大きく付与する，つまり段落境界が存在するサンプルについてより重点的に学習を進めることの有効性について確認できたといえる．

5.3 モデルに対する入力文の範囲の検討

本節では，ウィンドウサイズの異なるデータセットを対象とした段落境界推定実験について詳述する．この実験では，モデルに対する入力文の範囲を拡張することの，段落境界の推定に対する有効性について検討する．

5.3.1 実験設定

本実験では，4 章で説明したウィンドウサイズ 2，4，6，8 のデータセットを対象とする．使用したモデルは，前節の適切な損失関数の選定実験においてベースラインモデルおよび BERT+CE より高い精度を示し，平均値における比較で BERT+FL と同等の推定精度を記録した BERT+DL である．また，$F1$ 値と P_k を評価指標とする．

5.3.2 結果と考察

表 10 に，10 回の推定実験の結果得られた各評価指標の平均値とその標準偏差を示す．表 10 から，ウィンドウサイズが大きいほど推定精度が高い傾向にあることが確認できる．テスト C における P_k の平均値はウィンドウサイズ 4 の場合が最も低く，他のテストデータにおける各評価指標の値は，ウィンドウサイズ 6 の場合が最も良い精度を示した．ここでウィンドウサイズ 2 の場合とウィンドウサイズ 6 の場合の各評価指標の値について，t 検定によるとテスト B における $F1$ 値およびテスト C における P_k 以外の結果に関して，有意水準 0.05 で有意差が認められた．

表 11 に，ウィンドウサイズ 2，6 の各テストデータに対して 10 回すべての試行において同一の推定結果を出力したサンプルの例を，ウィンドウサイズ 6 のデータ形式で示す．例 1 および例 2 はいずれも，ウィンドウサイズ 2 の場合では誤って推定されたが，ウィンドウサイズ 6 の場合では正しく推定されたサンプルである．これらの 2 つのサ

表 11 ウィンドウサイズ 2，6 の各テストデータに対して 10 回すべての試行で同一の推定結果を出力したサンプルの例
Table 11 Examples of samples that outputs the same estimation results in all 10 trials for each test data of window size 2 and 6.

#	正解ラベル	Segment 1	Segment 2
1	1	頁の上に眼は着けていながら、お嬢さんの呼びに来るのを待っているくらいなものでした。/ 待っていて来ないと、仕方がないから私の方で立ち上がるのです。/ そうして向うの室の前へ行って、こっちから「ご勉強ですか」と聞くのです。/	お嬢さんの部屋は茶の間と続いた六畳でした。/ 奥さんはその茶の間にいる事もあるし、またお嬢さんの部屋にいる事もありました。/ つまりこの二つの部屋は仕切があっても、ないと同じ事で、親子二人が往ったり来たりして、どっち付かずに占領していたのです。/
2	0	蒲団が冷いので、背中がほどよくひんやりして、ついうっとりなる。/ 幸福は一夜おくれて来る。/ ぼんやり、そんな言葉を思い出す。/	福を待って待って、とうとう堪え切れずに家を飛び出してしまって、そのあくる日に、素晴らしい幸福の知らせが、捨てた家を訪れたが、もうおそかった。/ 幸福は一夜おくれて来る。/ 幸福は、――/
3	0	僕はいつか憂鬱の中に反抗的精神の起るのを感じ、やぶれかぶれになった賭博狂のようにいろいろの本を開いて行った。/ が、なぜかどの本も必ず文章か挿し画かの中に多少の針を隠していた。/ どの本も？/	僕は何度も読み返した「マダム・ボヴァリイ」を手にとった時さえ、畢竟僕自身も中産階級のムッシウ・ボヴァリイに外ならないのを感じた。/ 日の暮に近い丸善の二階には僕の外に客もないらしかった。/ 僕は電燈の光の中に書棚の間をさまよって行った。/

ンプルは，Segment 1 および Segment 2 が正解データにおいてそれぞれ同一の形式段落に属する文で構成されているという特徴がある．また一方で，例 3 は先の 2 つの例とは異なり，ウィンドウサイズ 2 の場合では正しく推定されたが，ウィンドウサイズ 6 の場合では誤って推定されたサンプルである．このサンプルについては，Segment 2 の『僕は何度も読み返した「マダム・ボヴァリイ」を手にとった時さえ、畢竟僕自身も中産階級のムッシウ・ボヴァリイに外ならないのを感じた。』と『日の暮に近い丸善の二階には僕の外に客もないらしかった。僕は電燈の光の中に書棚の間をさまよって行った。』がそれぞれ異なる段落に属している．このような例から，Segment 1 と Segment 2 のそれぞれに，同一の形式段落に属する文のみが含まれているサンプルについては推定精度が向上し，一方で Segment 1 あるいは Segment 2 に異なる形式段落に所属する文が混在しているサンプルについては推定が難しくなる例が存在していることが推察される．

また，モデルに対する入力系列のウィンドウサイズを拡張することで段落境界の推定精度は向上する傾向が確認されたが，ウィンドウサイズ 8 で設定したとき，ウィンドウサイズ 6 で設定した場合よりも精度は低くなった．これに関して，本研究において作成したデータセットにおけるラベル 1 に対するラベル 0 の比率は，テスト C を除いてすべて 8 未満であった．このラベル比は 1 つの段落に含まれる文の数の平均値を表しているため，ウィンドウサイズを 8 以上で設定することで，先に述べたような Segment 1 あるいは Segment 2 に異なる形式段落に所属する文が混在しているサンプルが増加し，結果としてモデルの性能低下を引き起こしたと考えられる．

6. まとめと今後の展望

本研究では小説の創作支援の観点から，既存の小説に対する段落分けの精度向上を目的とした．我々はこの問題について，対象とする文間に段落としての境界が存在しているか否かという 2 クラス分類問題としてとらえた．しかしこの場合，段落の数が文の数と比較してきわめて少ないことから，データとしての不均衡性がボトルネックとなる．そこで我々は，損失関数として不均衡データの分類問題に対して頑健な Dice Loss および Focal Loss を BERT に導入することで，モデルの推定精度向上を図った．そして実験的に，Focal Loss および Dice Loss を導入した段落境界推定モデルが他のモデルと比較して有意に高い推定精度を示すことを確認した．

さらに我々は，段落境界推定モデルに対する入力文の範囲を 2 文から 4 文，6 文，8 文と拡張することで，推定精度の向上を図った．実験の結果，6 文つまり前後 3 文を入力文範囲とした場合に最も高い推定精度が得られ，入力文の範囲を拡張することの有効性を確認した．このことから，より広範囲な文の情報を付与することが，段落境界の推定において有効に働くことが考察される．

今後の展望としては，たとえば段落分けの位置を異常値とした異常値検知のような，文章の時系列的な性質を利用したアプローチの採用があげられる．この手法では，前後の数文だけでなく，さらに過去の文の情報を考慮することができるという利点がある．

また本研究では，段落境界推定モデルの性能評価について定量的な評価をするにとどまった．しかし小説としての読みやすさを論ずるうえで，出力結果に対する定性的な評価も必要である．そこで今後は，本研究における定量的な評価と，実際にモデルが段落分けを施した文章を読後，ど

のような印象をいだくか，あるいはどれほど内容を理解したかというようなアンケート実験による定性的な評価をあわせた，複眼的なモデル評価に取り組みたいと考えている．

謝辞 なお，本研究は一部，日本学術振興会科学研究補助金基盤研究（B）（課題番号 19H04184）の補助を得て行われたものである．また，本研究は一部，日本学術振興会科学研究補助金基盤研究（C）（課題番号 20K11958）の補助を得て行われたものである．

参考文献

[1] Matsui, Y., Ito, K., Aramaki, Y., Fujimoto, A., Ogawa, T., Yamasaki, T. and Aizawa, K.: Sketch-based manga retrieval using manga109 dataset, *Multimedia Tools and Applications*, Vol.76, No.20, pp.125–134 (2017).

[2] Ogawa, T., Otsubo, A., Narita, R., Matsui, Y., Yamasaki, T. and Aizawa, K.: Object Detection for Comics using Manga109 Annotations, *CoRR*, 1bs/1803.08670 (2018).

[3] 米田航紀，横山想一郎，山下倫央，川村秀憲：深層学習を用いたモチーフ画像に基づく俳句生成，*SIG-SAI*, Vol.31, No.3, pp.1–8 (2018).

[4] Ito, T. and Ogata, T.: A Framework for Haiku Generation from a Narrative, *Journal of Robotics, Networking and Artificial Life*, Vol.6, No.1, pp.23–26 (2019).

[5] 松原 仁，佐藤理史，赤石美奈，角 薫，迎山和司，中島秀之，瀬名秀明，村井 源，大塚裕子：コンピュータに星新一のようなショートショートを創作させる試み，人工知能学会全国大会論文集，JSAI2013, pp.2D11–2D11 (2013).

[6] Sutskever, I., Vinyals, O. and Le, Q.V.: Sequence to Sequence Learning with Neural Networks, *CoRR*, abs/1409.3215 (2014).

[7] Jain, P., Agrawal, P., Mishra, A., Sukhwani, M., Laha, A. and Sankaranarayanan, K.: Story Generation from Sequence of Independent Short Descriptions, *CoRR*, abs/1707.05501 (2017).

[8] Fan, A., Lewis, M. and Dauphin, Y.: Hierarchical Neural Story Generation, *Proc. 56th Annual Meeting of the Association for Computational Linguistics* (*Volume 1: Long Papers*), pp.889–898 (2018).

[9] Ashida, A. and Kojiri, T.: Plot-creation support with plot-construction model for writing novels, *Journal of Information and Telecommunication*, Vol.3, No.1, pp.57–73 (2019).

[10] 芦田 淳，小尻智子：小説読者の感情変化パターンに基づいた登場人物の感情設定支援，*SIG-ALST*, Vol.B5, No.3, pp.76–79 (2019).

[11] Gupta, P., Kumar, V.B., Bhutani, M. and Black A.W.: WriterForcing: Generating more interesting story endings, *Proc. Second Workshop on Storytelling*, pp.117–126 (2019).

[12] Fukuda, K., Fujino, S., Mori, N. and Matsumoto, K.: Semi-automatic Picture Book Generation Based on Story Model and Agent-Based Simulation, *Intelligent and Evolutionary Systems*, pp.117–132 (2017).

[13] 福田清人，森 直樹，松本啓之亮：既存小説に依存しない創発的なストーリーの自動生成に関する考察，人工知能学会全国大会論文集，Vol.29, pp.1–4 (2015).

[14] 関 友作，赤堀侃司：テキストにおける段落表示が内容理解に与える影響，日本教育工学雑誌，Vol.20, No.2, pp.97–108 (1996).

[15] 村越行雄：段落とパラグラフの構造と方法について，コミュニケーション文化 = *Communication in Culture*, Vol.9, pp.1–27 (2015).

[16] 板倉由知，白井治彦，黒岩丈介，小高知宏，小倉久和：様々な文書を対象とした段落一貫性の解析，研究報告自然言語処理（NL），Vol.192, No.9, pp.1–6 (2009).

[17] Devlin, J., Chang, M-W., Lee, K. and Toutanova, K.: BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding, arXiv preprint arXiv:1810.04805 (2018).

[18] Lin, T-Y., Goyal, P., Girshick, R., He, K. and Dollar, P.: Focal Loss for Dense Object Detection, *2017 IEEE International Conference on Computer Vision* (*ICCV*), pp.2999–3007 (2017).

[19] Milletari, F., Navab, N. and Ahmadi, S-A.: V-Net: Fully Convolutional Neural Networks for Volumetric Medical Image Segmentation, *2016 4th International Conference on 3D Vision* (*3DV*), pp.565–571 (2016).

[20] Hearst, M.A.: Multi-Paragraph Segmentation Expository Text, *32nd Annual Meeting of the Association for Computational Linguistics*, pp.9–16 (1994).

[21] Glavaš, G., Nanni, F. and Ponzetto, S.P.: Unsupervised Text Segmentation Using Semantic Relatedness Graphs, *Proc. 5th Joint Conference on Lexical and Computational Semantics*, pp.125–130 (2016).

[22] Koshorek, O., Cohen, A., Mor, N., Rotman, M. and Berant, J.: Text Segmentation as a Supervised Learning Task, *Proc. 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 2* (*Short Papers*), pp.469–473 (2018).

[23] Badjatiya, P., Kurisinkel, L.J., Gupta, M. and Varma, V.: *CoRR*, abs/1808.09935 (2018).

[24] Liu, X-Y., Wu, J. and Zhou, Z-H.: Exploratory Undersampling for Class-Imbalance Learning, *IEEE Trans. Systems, Man, and Cybernetics, Part B* (*Cybernetics*), Vol.39, No.2, pp.539–550 (2009).

[25] Zhang, J. and Mani, I.: KNN Approach to Unbalanced Data Distributions: A Case Study Involving Information Extraction, *Proc. ICML 2003 Workshop on Learning from Imbalanced Datasets* (2003).

[26] Chawla, N.V., Bowyer, K.W., Hall, L.O. and Kegelmeyer, W.P.: SMOTE: Synthetic Minority Oversampling Technique, *Journal of Artificial Intelligence Research*, Vol.16, pp.321–357 (2002).

[27] Shen, C., Roth, H.R., Oda, H., Oda, M., Hayashi, Y., Misawa, K. and Mori, K.: On the influence of Dice loss function in multi-class organ segmentation of abdominal CT using 3D fully convolutional networks, *CoRR*, abs/1801.05912 (2018).

[28] Kodym, O., Spanel, M. and Herout, A.: Segmentation of Head and Neck Organs at Risk Using CNN with Batch Dice Loss, *CoRR*, abs/1812.02427 (2018).

[29] 飯倉 陸，岡田 真，森 直樹：Focal Loss を利用した BERT による小説の段落境界推定，人工知能学会全国大会論文集, JSAI2020, pp.3D1OS22a02–3D1OS22a02 (2020).

[30] Iikura, R., Okada, M. and Mori, N.: Improving BERT with Focal Loss for Paragraph Segmentation of Novels, *Distributed Computing and Artificial Intelligence, 17th International Conference*, pp.21–30 (2021).

[31] Li, X., Sun, X., Meng, Y., Liang, J., Wu, F. and Li, J.: Dice Loss for Data-imbalanced NLP Tasks, *Proc. 58th Annual Meeting of the Association for Computational Linguistics*, pp.465–476 (2020).

[32] Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J.,

Jones, L., Gomez, A.N., Łukasz, K. and Polosukhin, I.: Attention is All you Need, *Advances in Neural Information Processing Systems 30*, pp.5998–6008 (2017).
[33] Madabushi, H.T., Kochkina, E. and Castelle, M.: Cost-Sensitive BERT for Generalisable Sentence Classification on Imbalanced Data, *Proc. 2nd Workshop on Natural Language Processing for Internet Freedom: Censorship, Disinformation, and Propaganda*, pp.125–134 (2019).
[34] Abraham, N. and Khan, N.M.: A Novel Focal Tversky Loss Function With Improved Attention U-Net for Lesion Segmentation, *2019 IEEE 16th International Symposium on Biomedical Imaging* (*ISBI 2019*), pp.683–687 (2019).
[35] 工藤 拓，山本 薫，松本裕治：Conditional Random Fields を用いた日本語形態素解析，情報処理学会研究報告自然言語処理（NL），Vol.2004, No.47, pp.89–96 (2004).
[36] Mikolov, T., Sutskever, I., Chen, K., Corrado, G.S. and Dean, J.: Distributed Representations of Words and Phrases and their Compositionality, *Advances in Neural Information Processing Systems 26*, pp.3111–3119 (2013).
[37] Beeferman, D., Berger, A. and Lafferty, J.: Statistical Models for Text Segmentation, *Machine Learning*, Vol.34, pp.177–210 (1999).

飯倉 陸 （学生会員）

2020 年大阪府立大学工学域電気電子系学類情報工学課程卒業．現在，同大学院工学研究科電気・情報系専攻知能情報工学分野博士前期課程在学中．主に機械学習，自然言語処理の研究に従事．

岡田 真 （正会員）

2001 年徳島大学大学院工学研究科知能情報工学専攻修了・博士（工学）．同年大阪府立大学総合科学部数理・情報科学科助手，同大学理学系研究科情報数理科学専攻助教を経て，2012 年より同大学大学院工学研究科電気・情報系専攻知能情報工学分野助教．所属学会は電気学会，人工知能学会，言語処理学会．専門分野は自然言語処理，知識処理，機械学習，人工知能等．

森 直樹

1968 年 9 月 2 日生．1992 年京都大学理学部物理学科卒業，1994 年同大学大学院工学研究科原子核工学専攻修士課程修了，1997 年同大学院工学研究科電気工学専攻博士後期課単位取得退学．同年大阪府立大学工学部情報工学科助手．2005 年大阪府立大学工学研究科講師，2007 年同大学准教授．2019 年より大阪府立大学大学院工学研究科教授．博士（工学）．主に進化型計算，機械学習，人工知能，人工市場の研究に従事．電気学会，システム制御情報学会，人工知能学会等の各会員．