Focal Loss を導入した BERT に基づく分類器による小説の段落境界推定

1 はじめに

計算機による小説の自動生成を実現するためには，小説特有の文章構造や技法を計算機が十分に理解できることが必須である．小説における文章の可読性を高めるための重要な技法の 1 つに，段落の分割がある．ここで段落を分割する位置には，人間が小説を創作あるいは鑑賞する際の感覚的な情報が含まれると考えられる．本稿では小説の自動生成を最終的な目標として見据えた段階的な研究として，計算機による人間の小説に対する理解の観点から形式段落境界の推定に取り組む．

文章の形式段落を推定する場合，対象とする 2 文が同一の形式段落に所属するかどうか，すなわち 2 文間における形式段落としての境界の有無に関する 2 クラス分類問題として捉えることが可能である．しかしその場合，文の数に対する段落の数は極めて小さいため，その不均衡性を考慮する必要がある．本稿で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 Bidirectional Encoder Representations from Transformer(BERT)[1] に基づく分類器に対して，損失関数として Focal Loss[2] を導入することで推定精度の向上を図る．

2 段落境界推定モデル

本章では，本稿における段落境界の推定に用いたモデルについて詳述する．

2.1 BERT

BERT は，複数の双方向 Transformer[3] に基づく汎用言語モデルであり，入力された文および，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．

本稿では，日本語 Wikipedia 全文 (約 1,800 万文) を用いて事前学習されたモデルを使用した．このモデルでは JUMAN++[1] により形態素解析された入力文を，Byte Pair Encoding を適用して subword に分割する．

BERT を極性判定や文書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，入力文に対して得られる BERT の出力を分類器への入力とする．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．これにより，The General Launguage Understanding Evaluation (GLUE) ベンチマーク問題や Stanford Question Answering Dataset (SQuAD) を用いた機械読解のためのベンチマーク問題などの複数のタスクで高い精度を示した．

2.2 損失関数

BERT がクラス分類問題を解く場合に使用する損失関数は Softmax Cross Entropy Loss である．本稿では 2 値分類問題を扱うため，便宜的に以下の式により表される Binary Cross Entropy Loss f_{BCE} を用いる．

$$f_{\mathrm{BCE}}(p_t) = -\log(p_t), \tag{1}$$

$$p_t = \begin{cases} p & \text{if } y = 1 \\ 1-p & \text{otherwise,} \end{cases} \tag{2}$$

ここで $y \in \{\pm 1\}$ は正解ラベルであり，$p \in [0,1]$ は sigmoid 関数により出力されたクラスラベル $y = 1$ である推定確率を表す．

一般的に，不均衡データを対象とする分類問題には，先の f_{BCE} に重み $\alpha \in [0,1]$ を導入することでバランス調整された (3) 式で表される α-balanced Binary Cross Entropy Loss $f_{\alpha\mathrm{BCE}}$ を用いることでモデルの汎用性を高めることが可能である．実用的な α 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．

$$f_{\alpha\mathrm{BCE}}(p_t) = -\alpha\log(p_t). \tag{3}$$

また，画像処理の分野における物体検出の問題に対して有効性が示されている損失関数に Focal Loss

[1] http://nlp.ist.i.kyoto-u.ac.jp/index.php?JUMAN++

表 1: 作成したデータセットの情報 (数値実験 1)

データ	作品	平均単語数 (標準偏差)	最小単語数	最大単語数	ラベル比 (0 : 1)
訓練	三四郎，それから，門，彼岸過迄，行人	36.50 (20.12)	3	220	23896 : 3927
テスト	こころ	35.52 (14.91)	5	124	4331 : 735

f_{FL} がある．物体検出の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ことは困難になる．上述の $f_{\alpha\mathrm{BCE}}$ 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とするが，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 f_{FL} は，識別が容易である多数派クラスの重みを小さくする．具体的には，(4) 式に表すようにチューニング可能な $\gamma \geq 0$ を含んだ項 $(1-p_t)^{\gamma}$ を f_{BCE} に導入している．ここで $\gamma = 0$ のとき，f_{FL} は f_{BCE} と同等である．

$$f_{\mathrm{FL}}(p_t) = -(1-p_t)^{\gamma}\log(p_t). \tag{4}$$

実際には，この Focal Loss に対して重み α を加えてバランス調整された α-balanced Focal Loss $f_{\alpha\mathrm{FL}}$ を用いることで性能が向上することが文献 [2] において確認されている．

$$f_{\alpha\mathrm{FL}}(p_t) = -\alpha(1-p_t)^{\gamma}\log(p_t). \tag{5}$$

3 数値実験

3.1 数値実験 1: Focal Loss の導入

BERT の転移学習を適用させることで小説の形式段落を推定する．その際，一般的な損失関数として Cross Entropy Loss を使用した場合と比較することで，Focal Loss を用いることの有効性を検討する．

3.1.1 実験設定

本稿における実験のために，電子図書館である「青空文庫」[2] において管理されている夏目漱石の小説からデータセットを作成した．具体的には，連続する 2 文が同一の段落に属する場合にはラベル 0 を付与し，属していない，すなわち 2 文間において改行が生じている場合にはラベル 1 を付与する．表 1 に訓練用およびテスト用データの各情報を示す．

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

3.1.2 評価指標

不均衡データに対する分類問題を扱う際には，モデルの評価指標の設定に注意する必要がある．本稿では，異なるパラメータで設定されたモデル同士の比較のための評価指標として F1 値と以下に詳述するマシューズ相関係数 (Matthews Correlation Coefficient: MCC) および Jaccard 係数を採用する．

マシューズ相関係数

マシューズ相関係数とは，機械学習における不均衡なデータに対する 2 クラス分類モデルの評価に適した性能指標であり，以下の式により算出される．ここで $-1 \leq \mathrm{MCC} \leq 1$ であり，$\mathrm{MCC} = 0$ の場合ランダムな予測を表し，$\mathrm{MCC} = 1$ の場合完全に真値を予測可能であることを表す．

$$\mathrm{MCC} = \frac{\mathrm{TP}\times\mathrm{TN} - \mathrm{FP}\times\mathrm{FN}}{\sqrt{(\mathrm{TP}+\mathrm{FP})(\mathrm{TP}+\mathrm{FN})(\mathrm{TN}+\mathrm{FP})(\mathrm{TN}+\mathrm{FN})}}. \tag{6}$$

ここで TP, TN はそれぞれ正しく分類された正と負のサンプル数であり，FN, FP はそれぞれ誤って分類された正と負のサンプル数である．

Jaccard 係数

Jaccard 係数とは 2 つの集合の類似度を表す指標の 1 つである．これを 2 クラス分類問題の評価指標として適用する場合，以下の式により算出される．

$$\mathrm{Jaccard} = \frac{\mathrm{TP}}{\mathrm{FP}+\mathrm{FN}+\mathrm{TP}}. \tag{7}$$

3.1.3 結果と考察

表 2 に，5 回の推定実験の結果得られた各評価指標の値の平均値および標準偏差について示す．なお，すべての予測値が少数派ラベル (ラベル 1) である場合の各評価指標の値をベースラインとして設定した．

表 2: 各評価指標の値 (数値実験 1)

		F1 値		MCC		Jaccard 係数	
α	γ	平均	標準偏差	平均	標準偏差	平均	標準偏差
0.25	**8.0**	**0.596**	0.0062	**0.541**	0.010	**0.425**	0.0063
0.50	0	0.569	0.0087	0.527	0.010	0.398	0.0085
ベースライン		0.253	-	0	-	0.145	-

表 3: 数値実験 1 における出力結果の例 (ラベル 0: 形式段落境界なし，ラベル 1: 形式段落境界あり)

例	真値	予測値	文 1	文 2
1	1	1	私が陸へ上がって雫の垂れる手を振りながら掛茶屋に入ると、先生はもうちゃんと着物を着て入れ違いに外へ出て行った。	私は次の日も同じ時刻に浜へ行って先生の顔を見た。
2	1	1	しかし私にはその意味がまるで解らなかった。	先生と私は通りへ出ようとして墓の間を抜けた。
3	0	1	そうして先生といっしょの方角に泳いで行った。	二丁ほど沖へ出ると、先生は後ろを振り返って私に話し掛けた。

Focal Loss を導入した BERT に基づく分類器は $\alpha = 0.25, \gamma = 8.0$ のとき F1 値と MCC，Jaccard 係数のそれぞれの評価指標において最大の値が得られ，Cross Entropy Loss ($\alpha = 0.50, \gamma = 0$ のとき) を用いる場合よりも高い性能を示した．この結果から，自然言語に関する不均衡データの分類問題を解く上で BERT に基づく分類器の損失関数として Focal Loss を採用することの有効性が確認できた．

表 3 に $\alpha = 0.25, \gamma = 8.0$ で設定した分類器を用いた 5 回すべての推定実験において，同一の推定結果を出力した例を示す．これらの例には，対象とする両方の文中に共通する単語が現れているという特徴がある．例 1, 2 については Cross Entropy Loss を用いた場合では正しい推定ができていなかった．一方で例 3 については，Cross Entropy Loss を用いた場合では正しくラベル 0 を予測値として出力していた．このことから，Focal Loss を用いる場合には同一の単語が含まれる 2 文間で改行される例が，より強く学習に寄与することが考察される．

3.2 数値実験 2: 入力文数の変更

数値実験 1 ではそれぞれ 2 文のみを対象として，同一の形式段落に属しているか否かを推定した．しかし小説という文章形態の特徴上，広範囲な文の情報を与えることでより適切な段落境界の推定が可能になると予想される．そこで本実験では，モデルに対する入力として用いる文の範囲の観点から形式段落境界の推定における性能を比較する．

表 4: データセットの変更点 (数値実験 2)

データ	平均単語数 (標準偏差)	最小単語数	最大単語数
訓練	73.98 (32.87)	11	297
テスト	72.04 (23.14)	19	195

3.2.1 実験設定

数値実験 1 で作成したデータセットにおける入力する文の範囲を拡張する．表 4 に，数値実験 1 におけるデータセットから変更した項目について示す．このデータセットに対して数値実験 1 と同様に形式段落の境界推定を試みる．

3.2.2 結果と考察

表 5 に，5 回の推定実験の結果得られた各評価指標の値の平均値および標準偏差について示す．

表 5: 各評価指標の値 (数値実験 2)

		F1 値		MCC		Jaccard 係数	
α	γ	平均	標準偏差	平均	標準偏差	平均	標準偏差
0.25	**8.0**	**0.612**	0.0053	**0.561**	0.0031	**0.441**	0.0056
0.50	0	0.577	0.0049	0.551	0.0069	0.405	0.0048
ベースライン		0.253	-	0	-	0.145	-

表 6: 数値実験 2 における出力結果の例 (ラベル 0: 形式段落境界なし，ラベル 1: 形式段落境界あり)

例	真値	予測値	文 1	文 2
1	0	0	だからここでもただ先生と書くだけで本名は打ち明けない。これは世間を憚かる遠慮というよりも、その方が私にとって自然だからである。	私はその人の記憶を呼び起すごとに、すぐ「先生」といいたくなる。筆を執っても心持は同じ事である。
4	1	1	私はその時この燈火が音のしない渦の中に、自然と捲き込まれている事に気が付かなかった。しばらくすれば、その灯もまたふっと消えてしまうべき運命を、眼の前に控えているのだとは固より気が付かなかった。	私は今度の事件について先生に手紙を書こうかと思って、筆を執りかけた。私はそれを十行ばかり書いて已めた。

表 5 に示す通り，$\alpha = 0.25, \gamma = 8.0$ のとき最も高い値が得られた．また，損失関数として Focal Loss あるいは Cross Entropy Loss を採用した場合ともに，数値実験 1 で得られた結果と比較して各評価指標の値について向上が見られた．このことから，より広範囲な文の情報を与えることは小説の形式段落としての境界推定に対して有効であると考察される．表 6 に，数値実験 1 における結果と比較し，入力文とする文の範囲を拡張することにより正しく段落境界の推定が可能となった例について示す．

4 おわりに

本稿では小説の形式段落推定を，段落境界の有無に関する 2 クラス分類問題として捉えた．そこで扱うデータの不均衡性を考慮し，BERT に基づく分類器における損失関数として Focal Loss を導入した．その結果 Cross Entropy Loss を用いる一般的な場合と比較して，形式段落境界の推定に対する分類器としての性能を向上させることができた．

また，入力文の数を増やすことにより段落境界の推定におけるモデルの性能を向上させることができた．この結果は，対象とする文章の範囲を拡大することが段落境界の推定において有効であることを示唆していると考えられる．

本稿で得られた成果は，例えば文章における段落分割の位置を書き手に推薦するような，計算機による人間の創作支援としても応用可能である．これは計算機による小説の自動生成の実現のためにも，非常に有意義であるといえる．

今後の展望としては，モデルに対してより広範囲な文の情報を与えるために文章の時系列性を利用することが挙げられる．具体的には，段落境界の推定を異常値検知問題として捉えることが有効であると考えられる．

参考文献

[1] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. BERT: pre-training of deep bidirectional transformers for language understanding. *CoRR*, abs/1810.04805, 2018.

[2] Tsung-Yi Lin, Priya Goyal, Ross B. Girshick, Kaiming He, and Piotr Dollár. Focal loss for dense object detection. *CoRR*, abs/1708.02002, 2017.

[3] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin. Attention is all you need. *CoRR*, abs/1706.03762, 2017.