applied sciences

MDPI

Article

Paragraph Boundary Recognition in Novels for Story Understanding

Riku Iikura *, Makoto Okada and Naoki Mori

Graduate School of Engineering, Osaka Prefecture University, 1-1 Gakuen-cho, Naka-ku, Sakai, Osaka 599-8231, Japan; okada@cs.osakafu-u.ac.jp (M.O.); mori@cs.osakafu-u.ac.jp (N.M.)
* Correspondence: iikura@ss.cs.osakafu-u.ac.jp

Abstract: The understanding of narrative stories by computer is an important task for their automatic generation. To date, high-performance neural-network technologies such as BERT have been applied to tasks such as the Story Cloze Test and Story Completion. In this study, we focus on the text segmentation of novels into paragraphs, which is an important writing technique for readers to deepen their understanding of the texts. This type of segmentation, which we call "paragraph boundary recognition", can be considered to be a binary classification problem in terms of the presence or absence of a boundary, such as a paragraph between target sentences. However, in this case, the data imbalance becomes a bottleneck because the number of paragraphs is generally smaller than the number of sentences. To deal with this problem, we introduced several cost-sensitive loss functions, namely. focal loss, dice loss, and anchor loss, which were robust for imbalanced classification in BERT. In addition, introducing the threshold-moving technique into the model was effective in estimating paragraph boundaries. As a result of the experiment on three newly created datasets, BERT with dice loss and threshold moving obtained a higher *F*1 than the original BERT had using cross-entropy loss as its loss function (76% to 80%, 50% to 54%, 59% to 63%).

Keywords: natural-language processing; story understanding; text segmentation; imbalanced classification; BERT; cost-sensitive loss

check for updates

Citation: Iikura, R.; Okada, M.; Mori, N. Paragraph Boundary Recognition in Novels for Story Understanding. *Appl. Sci.* **2021**, *11*, 5632. https://doi.org/10.3390/app11125632

Academic Editor: Juan Manuel Montero Martínez, Fernando Fernández-Martínez and Ascension Gallardo Antolín

Received: 1 May 2021
Accepted: 16 June 2021
Published: 18 June 2021

Publisher's Note: MDPI stays neutral with regard to jurisdictional claims in published maps and institutional affiliations.

Copyright: © 2021 by the authors. Licensee MDPI, Basel, Switzerland. This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution (CC BY) license (https://creativecommons.org/licenses/by/4.0/).

1. Introduction

With the improvement in natural-language processing, commercial expectations for the automatic generation of scenarios in novels, movies, games, etc. have been increasing. For example, in Japan, there is a literary award that explicitly allows for works written by artificial intelligence (https://hoshiaward.nikkei.co.jp/, accessed on 18 June 2021). Studies on story generation have been proposed since ancient times [1–3]; recently, various approaches based on neural networks were proposed [4–8]. However, it is very difficult to generate high-quality narrative sentences that satisfy readers. Therefore, it is important for the computer to understand the special sentence structure of the novel as a step-by-step study with the ultimate goal of the automatic generation of the novel by the computer. A computer understanding a story such as a novel is generally a task of modeling the consistency of sentences. Thus far, research has been conducted on the understanding and generation of narrative sentences from various perspectives such as the Story Cloze Test (SCT) [9,10] and Story Completion (SC) [11], ordering sentences into a story [12–14].

In relation to tasks aimed at story understanding by a computer as described above, we focused on segmenting text from novels into paragraphs, which is an important technique used in sentence writing in novels. It is also an operation that divides a document into paragraphs to improve readability. Seki et al. [15] discussed the importance of paragraphs in texts. In their study, the effect of a paragraph display as a document layout on the reader's understanding of content was investigated. Specifically, they conducted an experiment in which a collaborator read correctly paragraphed sentences, intentionally incorrectly paragraphed sentences, and unparagraphedsentences, and then compared

their comprehension of the contents. They concluded that a proper paragraph setting was important for facilitating the reader's understanding of the document.

The features of paragraphs are slightly different between descriptive or logical sentences, such as academic treatises, and literary sentences, such as in novels and essays. This is essential because logical texts focus on accurately and concisely disclosing information to the reader, whereas literary texts focus on emotionally impressing on the reader [16]. In the former, one paragraph focuses solely on one topic, centered on a core sentence, which is called a topic sentence. It is necessary for the content and assertions in each sentence composing a paragraph to be consistent, and studies numerically defined the consistency of paragraphs for scientific and technological sentences [17]. By contrast, the latter is based on the transition of topics and scenes over time, and generally does not apply a topic sentence [16]. Therefore, there are no common rules that the writer should uniformly follow in the text segmentation of a novel, and such segmentation is an extremely difficult operation that requires a high level of skill. Although the demand for systems that assist in text segmentation of novels into paragraphs is high, few studies have been conducted in this area owing to the difficulty.

Placing new paragraphs in the appropriate positions on the basis of the transitions of scenes and topics helps readers to fully understand the story. Therefore, the position where a new paragraph starts contains important sensory information for people who write or read novels. On the basis of this assumption, in this study, as a stepwise approach with the automatic creation of a novel as the ultimate objective, we estimate paragraph boundaries from the perspective of a computer's understanding of the story in the novel. We regarded this task as a binary classification problem regarding whether target sentences belong to the same paragraph, which we call "paragraph-boundary recognition". However, because the number of paragraphs is small relative to the number of sentences, it is necessary to consider the imbalance in the number of the data. Therefore, we used focal loss [18], dice loss [19], and anchor loss [20], which were confirmed to be robust to imbalanced classification, as a loss function of BERT, which is highly accurate in various natural-language processing (NLP) tasks. Through experiments, we confirmed that the novel could be divided into appropriate paragraphs by using our method. This result suggests that it is possible to divide narrative sentences generated by humans or computers into appropriate paragraphs and improve the readability of the texts. Our contributions are summarized as follows:

- We regarded the text segmentation of novels into paragraphs as imbalanced classification regarding the presence or absence of a paragraph boundary between two consecutive sentences, and applied BERT, which introduced multiple cost-sensitive loss functions.
- We confirmed that the accuracy of paragraph-boundary recognition by our approach can be improved by applying threshold moving, which is one of the methods for dealing with imbalanced classification.
- Our experiment using multiple author-specific datasets, newly created for this study, showed that the proposed method could recognize paragraph boundaries with higher accuracy than that of a conventional text-segmentation method.

2. Related Works

The tasks in this study are strongly related to text segmentation and studies on classification problems for imbalanced data. We describe the background of such research in this section.

2.1. Story Understanding

Thus far, research has been conducted on the understanding and generation of narrative sentences from various perspectives. For example, the Story Cloze Test (SCT) [9,10] selects the optimal ending following the input sentences, and evaluates the performance of the model on the basis of the correct answer rate. Various studies included a classifier

using recurrent neural networks (RNN) and a model based on the transfer learning of bidirectional encoder representations from transformers (BERT) [21–26].

Guan et al. [11] proposed Story Completion (SC), which is an extension of SCT, and is the task of generating a sentence that is missing in a given sentence. In addition, Gupta et al. [27] focused on the fact that the ending of a story is not uniquely determined and that multiple endings are possible, and proposed a method for generating various story endings. Mori et al. [28] added missing-position prediction, which is an operation used to estimate a missing sentence, to a conventional SC where the position of the complementary sentence is specified, assuming an actual writing-support scene.

Another task for understanding stories is ordering sentences of narrative stories. Sentence ordering [29] is the task of rearranging sentences to maximize the evaluation value for the consistency of a document. A method for obtaining context information at the paragraph level using the Pointer Network [30] based on an RNN was also proposed [12,31]. Wang et al. [13], and Oh et al. [32] proposed models using the attention mechanism, and Cui et al. [14] proposed a model that utilizes the dependency between sentences acquired by BERT.

2.2. Text Segmentation

The operation of dividing a sentence into semantic groups based on topics is generally known as text segmentation [33–35]. This is an important task applied to various aspects in the field of NLP, such as sentence summarization and question answering, from the viewpoint of understanding the meaning of natural language by a computer. Text-segmentation methods proposed thus far are roughly divided into unsupervised and supervised algorithms.

TextTiling [33] is a type of unsupervised text segmentation that utilizes the fact that specific words frequently appear in the same segment, and calculates the similarity of each segment from such vectors. Glavaš et al. [36] proposed an unsupervised algorithm for constructing semantic-relevance graphs of sentences using the word-embedding expression and a measure of semantic relevance of short sentences. The nodes in this graph represent sentences, and the edges between the two sentences indicate that the sentences are semantically similar. Segmentation is then determined by finding the maximal cliques of adjacent sentences and heuristically completing the segmentation.

In addition, there is a model that uses long short-term memory (LSTM), which is a type of RNN, as a method of supervised learning [37]. Such a model can efficiently model the input sequence by controlling the flow of information over time. Badjatiya et al. [38] proposed an attention-based convolutional neural network bidirectional LSTM model that introduced the attention mechanism and learned the relative importance of each sentence in the text to achieve segmentation. Glavaš et al. [39] proposed a multitask learning model that couples the sentence-level segmentation objective with the coherence objective that differentiates correct sequences of sentences from corrupt ones.

In our study, we treat text segmentation as an imbalanced classification that should consider the imbalance in the number of segments and the number of sentences, as opposed to the above-mentioned studies.

2.3. Imbalanced Classification

There are two main approaches to the classification problem, in which the number of data in each class is imbalanced: a resampling method and cost-sensitive learning.

Resampling is a method of generating balanced distribution by making changes to imbalanced data. Imbalance in the number of data is eliminated by undersampling the majority class [40,41] or oversampling the minority class. The simplest oversampling method is to randomly duplicate a minority-class instance, but can cause overfitting owing to the redundant distribution of data. To solve this problem, Chawla et al. [42] proposed SMOTE, which is a basic approach using data synthesis. SMOTE randomly selects seed

samples to balance the dataset, and applies linear interpolation between the seed sample and one of its neighbors to synthesize a new sample.

By contrast, cost-sensitive learning is a method for improving the classifier itself through learning, which applies a loss function with different weights to each data sample, instead of changing the distribution of training data. This method is often associated with research dealing with object-detection problems, particularly in the field of image processing. This is because the background occupies most of the image in an object-detection problem, and it is necessary to eliminate the imbalance of the label to identify a specific object in the minority class. Thus far, studies segmented medical images using a loss function based on the Dice coefficient [43,44], and robust losses for imbalanced data, such as focal loss [18], dice loss [19], and anchor loss [20] were proposed. We applied BERT, adopting focal loss and dice loss as the loss functions, to the text segmentation of novels into paragraphs, and demonstrated the effectiveness of the approach [45–47]. Furthermore, Li et al. [48] also applied BERT, introducing dice loss to tasks such as part-of-speech tagging and named-entity recognition as a classification problem of imbalanced classification in the field of NLP, and clarified its effectiveness.

Threshold moving is an alternative technique that can deal with class imbalance. As the main difference between resampling and threshold-based methods, the former relies on data preprocessing before the learning phase, whereas the latter relies on manipulating the model output. This technique is utilized using some popular learning methods, including ensemble learning [49–52]. In this study, we apply BERT, which introduces a cost-sensitive loss function, to the paragraph-boundary recognition of the novel, and adjust the decision threshold as a hyperparameter to improve estimation accuracy.

3. Technical Background

In this section, we describe BERT in detail, which is the basis of our paragraph-boundary recognition model, and the loss functions used in the classification.

3.1. BERT

BERT [22] is a general-purpose language model based on a multiple bidirectional transformer [53] that outputs a distributed representation of an input sequence and words included in the sequence. In this work, we used BERT$_{BASE}$ ($L = 12$, $H = 768$, $A = 12$, Total Parameters = 110M), where L, H, and A are the number of transformer blocks, the hidden size, and the number of self-attention heads, respectively. BERT improves the performance of a language model by pretraining a large-scale corpus. For prelearning, masked word prediction was applied to predict the original word of a sentence, in which a portion of the input sentence had been replaced with a token `[MASK]`, and the next sentence prediction was used to correctly identify the continuity of the two sentences as the input.

BERT represents a single sentence or a pair of sentences (for example, pair $<$ *question*, *answer* $>$) as a sequence of tokens according to the following features: BERT uses WordPiece embeddings [54]. To apply BERT in classification tasks, such as polarity determination and document classification, the vector output for token `[CLS]` added to the head of the input sentence was input into the classifier. In particular, when inputting two sentences into the model, token `[SEP]` is inserted between the two sentences, which are combined and treated as a single sequence. Embedding is added to every token indicating whether it belongs to the first or the second sentence. For a given token, its input representation is constructed by summing the corresponding token, position, and segment embeddings. In BERT, after converting a sentence or sentence pair into a distributed representation, it is used as an input to solve applied tasks such as classification and regression using a multilayer perceptron. At this time, fine tuning using a pretrained model can be applied to the tasks to be solved.

3.2. Loss Functions

We detail some of the loss functions adopted to address the data imbalance, which is the bottleneck of the tasks on which this study focuses. In the following, the shown loss function assumes a binary-classification problem for convenience. Let X denote a set of training instances, and each instance $x_i \in X$ be associated with a golden binary label $y_i = [y_{i0}, y_{i1}]$ denotes the ground-truth class to which x_i belongs, and $p_i = [p_{i0}, p_{i1}]$ denote the predicted probabilities of the two classes, respectively, where $y_{i0}, y_{i1} \in \{0, 1\}$, $p_{i0}, p_{i1} \in [0, 1]$ and $p_{i1} + p_{i0} = 1$.

3.2.1. Cross Entropy Loss

The original BERT uses cross-entropy (CE) loss to solve classification problems, which is given as follows:

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

In general, for classification problems that target imbalanced data, weights $\alpha \in [0, 1]$ are introduced into the CE loss to adjust the balance, as shown in Equation (2), and the importance is considered on the basis of the size of each class. In many cases, the reciprocal of the number of data included in each class is adopted as a practical value of α.

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

Madabushi et al. [55] showed the effectiveness of changing the loss function in a fully connected layer, which is the final layer of BERT, into weighted CE loss for the classification problem of imbalanced data in identifying propaganda.

3.2.2. Focal Loss

Focal loss (FL) is a loss function proposed by Lin et al. [18] that dynamically scales CE loss. The above-mentioned weighted CE loss makes it possible to consider importance on the basis of the size of each class; however, it cannot distinguish the difficulty of identification for each class. However, FL introduces a modulation factor that attenuates the contribution of errors from easily identifiable examples and prevents overwhelming loss functions. This allows for the model to effectively focus on examples that are difficult to identify. Specifically, term $(1 - p_t)^\gamma$ containing $\gamma \geq 0$ is introduced into CE loss, which can be tuned as shown in Equation (3).

$$\mathrm{FL} = -\frac{1}{N}\sum_{i}\sum_{j\in\{0,1\}} y_{ij}(1 - p_{ij})^\gamma \log(p_{ij}). \tag{3}$$

When $\gamma = 0$, the FL is equivalent to the CE loss.

3.2.3. Dice Loss

One of the indicators used to evaluate the classification model of imbalanced data is $F1$. Dice coefficient (Sørensen-Dice coefficient: DSC) is the $F1$-oriented statistical index. Although the DSC is generally an index for measuring similarity between two sets, it may also be used in the segmentation of affected images in the medical field in connection with the imbalanced classification problem [56]. Li et al. [48] showed the relationship between the Dice coefficient and $F1$ as follows: First, given two sets A and B, the DSC is given as follows:

$$\mathrm{DSC} = \frac{2|A \cap B|}{|A| + |B|}. \tag{4}$$

In this study, set A is the set of samples determined to be positive by the model, and B is the set of ground-truth samples. Here, using the true-positive (TP), false-positive

(FP), and false-negative (FN) rates, the relationship between the Dice coefficient and $F1$ is expressed as follows:

$$\begin{aligned} \mathrm{DSC} &= \frac{2\mathrm{TP}}{2\mathrm{TP}+\mathrm{FN}+\mathrm{FP}} \\ &= \frac{2\frac{\mathrm{TP}}{\mathrm{TP}+\mathrm{FN}}\frac{\mathrm{TP}}{\mathrm{TP}+\mathrm{FP}}}{\frac{\mathrm{TP}}{\mathrm{TP}+\mathrm{FN}}+\frac{\mathrm{TP}}{\mathrm{TP}+\mathrm{FP}}} \\ &= \frac{2\mathrm{Precision}\times\mathrm{Recall}}{\mathrm{Precision}+\mathrm{Recall}} \\ &= F1. \end{aligned} \tag{5}$$

On the basis of the above definition, the value of the DSC for each sample x_i is given as follows:

$$\mathrm{DSC} = \frac{\sum_i 2y_{i1}p_{i1}+\epsilon}{\sum_i y_{i1}+\sum_i p_{i1}+\epsilon}, \tag{6}$$

where ϵ provides numerical stability to prevent division by zero.

Milletari et al. [19] proposed an objective function that squares each term of the denominator in the Dice coefficient, and defined the dice loss as a loss function to maximize it as follows:

$$\mathrm{DL} = \frac{1}{N}\left(1-\frac{\sum_i 2y_{i1}p_{i1}+\epsilon}{\sum_i y_{i1}^2+\sum_i p_{i1}^2+\epsilon}\right). \tag{7}$$

3.2.4. Anchor Loss

Motivated by focal loss, anchor loss (AL) [20] is a loss function that dynamically scales CE loss on the basis of the difficulty of predicting the sample. Similar to focal loss, AL was proposed for use in an object-detection task where the imbalance between the number of pixels of the background and the target object is a bottleneck. Focal loss addresses the class-imbalance issue by avoiding updating the main gradients for samples that are easy to predict. AL, by contrast, takes advantage of the difference in probabilities for targeted and nontargeted objects, and adjusts the scale of loss for the sample during training. At this time, on the basis of the difficulty of the prediction defined using the reference value, called anchor probability p^*, obtained from the network prediction, the loss value is dynamically reweighted. Penalties larger than or equal to CE loss are imposed when the predicted probabilities for nontargets are higher than anchor probabilities. We set the target class-prediction score as the anchor probability on the basis of the report by Ryou et al. Anchor loss is given as follows using hyperparameter $\gamma \geq 0$:

$$\mathrm{AL} = -\frac{1}{N}\sum_i y_{i0}(1+p_{i1}-p_i^*)^\gamma \log p_{i0} + y_{i1}\log p_{i1}. \tag{8}$$

4. Evaluational Experiment

This section details the conducted experiments to confirm the effectiveness of the proposed method.

4.1. Paragraph-Boundary Recognition Dataset

A paragraph is a unit of semantically divided sentences, as shown in Figure 1. We set the task called "paragraph-boundary recognition" to identify whether any two consecutive sentences in a novel belonged to the same paragraph, that is, whether there was a paragraph boundary between any two consecutive sentences in a novel (Figure 2).

In my younger and more vulnerable years my father gave me some advice that I've been turning over in my mind ever since.
"Whenever you feel like criticizing anyone," he told me, "just remember that all the people in this world haven't had the advantages that you've had."

He didn't say any more, but we've always been unusually communicative in a reserved way, and I understood that he meant a great deal more than that. In consequence, I'm inclined to reserve all judgements, a habit that has opened up many curious natures to me and also made me the victim of not a few veteran bores. The abnormal mind is quick to detect and attach itself to this quality when it appears in a normal person, and so it came about that in college I was unjustly accused of being a politician, because I was privy to the secret griefs of wild, unknown men. Most of the confidences were unsought—frequently I have feigned sleep, preoccupation, or a hostile levity when I realized by some unmistakable sign that an intimate revelation was quivering on the horizon; for the intimate revelations of young men, or at least the terms in which they express them, are usually plagiaristic and marred by obvious suppressions. Reserving judgements is a matter of infinite hope. I am still a little afraid of missing something if I forget that, as my father snobbishly suggested, and I snobbishly repeat, a sense of the fundamental decencies is parcelled out unequally at birth.

And, after boasting this way of my tolerance, I come to the admission that it has a limit. Conduct may be founded on the hard rock or the wet marshes, but after a certain point I don't care what it's founded on. When I came back from the East last autumn I felt that I wanted the world to be in uniform and at a sort of moral attention forever; I wanted no more riotous excursions with privileged glimpses into the human heart. Only Gatsby, the man who gives his name to this book, was exempt from my reaction—Gatsby, who represented everything for which I have an unaffected scorn. If personality is an unbroken series of successful gestures, then there was something gorgeous about him, some heightened sensitivity to the promises of life, as if he were related to one of those intricate machines that register earthquakes ten thousand miles away. This responsiveness had nothing to do with that flabby impressionability which is dignified under the name of the "creative temperament"—it was an extraordinary gift for hope, a romantic readiness such as I have never found in any other person and which it is not likely I shall ever find again. No—Gatsby turned out all right at the end; it is what preyed on Gatsby, what foul dust floated in the wake of his dreams that temporarily closed out my interest in the abortive sorrows and short-winded elations of men.

Figure 1. An illustration of a paragraph extracted from a sample novel (*The Great Gatsby* by F. Scott Fitzgerald). The part surrounded by the purple line is a paragraph. Conversational sentences were not counted as independent paragraphs.

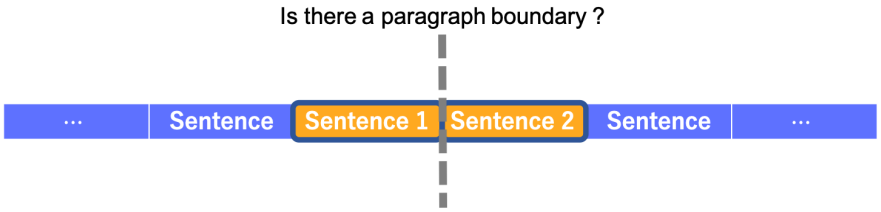

Figure 2. An illustration of paragraph-boundary recognition, which is a task to identify whether a paragraph boundary exists between any two consecutive sentences in a novel. Let two target sentences be `Sentence1` and `Sentence2`.

For the experiments conducted in this study, we created new datasets from the novels in Project Gutenberg (https://www.gutenberg.org/, accessed on 1 April 2021). The data were divided into sentences using the PUNKT tokenizer from NLTK [57]. We defined a set of sentences from an indentation at the beginning of a sentence to line breaks as a paragraph. On the basis of the above definition, a conversational sentence was defined as a single independent paragraph. Generally speaking, a conversational sentence starts with a quotation mark; therefore, it is easier to discriminate on the basis of the surface or symbolic grounds compared to a paragraph among descriptive sentences. Therefore, we did not count conversational sentences as a single paragraph.

Table 1 shows examples in the datasets. The input format for the model is based on the BERT prelearning format, that is, `[CLS] Sentence1 [SEP] Sentence2 [SEP]`. We used samples where `Sentence 1` and `Sentence 2` were in different paragraphs, that is, samples with paragraph boundaries between two target sentences as positive samples. On the other hand, we used samples where `Sentence 1` and `Sentence 2` were in the same paragraphs, that is, samples with *NO* paragraph boundaries between two target sentences as negative samples. We constructed the datasets for our experiment using the following works:

Fitzgerald dataset: "Head and Shoulders," "Dalyrimple Goes Wrong," "Benediction," "The Cut-Glass Bowl," "The Ice Palace," "The Four Fists," "This Side of Paradise," "The Beautiful and the Damned," "The Jelly-Bean," "Bernice Bobs Her Hair," "The Offshore Pirate," "The Great Gatsby"

Stevenson dataset: "The Strange Case of Dr. Jekyll and Mr. Hyde," "The Master of Ballantrae," "The Black Arrow," "Kidnapped," "Weir of Hermiston," "Treasure Island"

Twain dataset: "The Man That Corrupted Hadleyburg and Other Stories," "Tom Sawyer, Detective," "A Connecticut Yankee in King Arthur's Court," "The Mysterious Stranger and Other Stories," "Adventures of Huckleberry Finn," "The Tragedy of Pudd'nhead Wilson," "The Adventures of Tom Sawyer"

Table 1. Examples of samples in the Fitzgerald dataset. Positive samples in which a paragraph boundary exists between `Sentence 1` and `Sentence 2` are labeled '1', and negative samples in which a paragraph boundary does not exist between `Sentence 1` and `Sentence 2` are labeled '0'.

#	Label	Sentence 1	Sentence 2
1	0	In my younger and more vulnerable years my father gave me some advice that I've been turning over in my mind ever since.	"Whenever you feel like criticizing anyone," he told me, "just remember that all the people in this world haven't had the advantages that you've had."
2	1	"Whenever you feel like criticizing anyone," he told me, "just remember that all the people in this world haven't had the advantages that you've had.	He did n't say any more, but we've always been unusually communicative in a reserved way, and I understood that he meant a great deal more than that.
3	0	He didn't say any more, but we've always been unusually communicative in a reserved way, and I understood that he meant a great deal more than that.	In consequence, I'm inclined to reserve all judgements, a habit that has opened up many curious natures to me and also made me the victim of not a few veteran bores.
4	0	In consequence, I'm inclined to reserve all judgements, a habit that has opened up many curious natures to me and also made me the victim of not a few veteran bores.	The abnormal mind is quick to detect and attach itself to this quality when it appears in a normal person, and so it came about that in college I was unjustly accused of being a politician, because I was privy to the secret griefs of wild, unknown men.

Figure 3 shows how to split text for each novel into training, validation, and test data. In this study, 90% of the text from the beginning of each work was used for the training and validation data, and the last 10% of the text was used as the test data. This is based on the assumption that, when actually writing a novel, part of the novel is manually divided into paragraphs, and the rest of the text is automatically divided into paragraphs on the basis of the tendency. Table 2 shows the statistical information on the number of labels for each dataset.

Table 2. Statistics of each dataset showing the number of positive and negative samples, and their ratios in each dataset. Numbers of tokens per each sample were also included.

Dataset		#Negative	#Positive	Ratio	#Token/Sample	#Token/Negative	#Token/Positive
Fitzgerald	Training	14,115	3596	3.93	34.46	33.29	39.07
	Validation	1815	398	4.56	36.61	35.41	42.15
	Test	1816	404	4.50	33.98	32.83	39.24
Stevenson	Training	13,879	2550	5.44	45.35	43.34	56.34
	Validation	1734	319	5.44	46.64	44.23	59.87
	Test	1596	462	3.45	44.72	46.17	39.81
Twain	Training	19,215	2478	7.75	41.77	40.53	51.39
	Validation	2349	364	6.45	40.09	39.35	45.02
	Test	2277	437	5.21	38.19	38.89	34.66

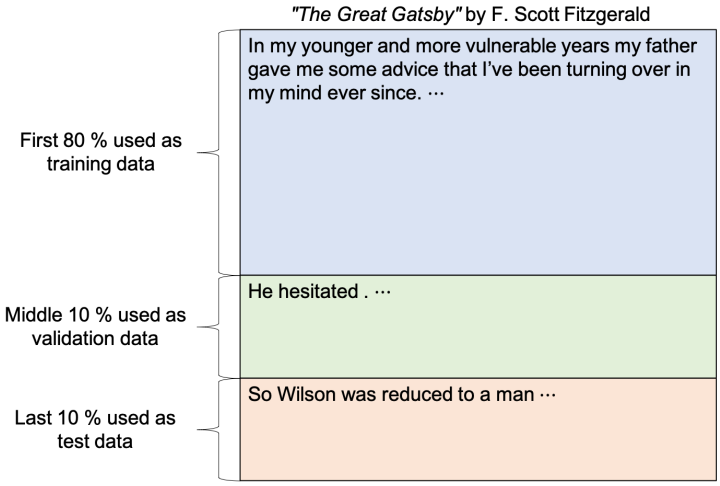

Figure 3. How to split text for each novel into training, validation, and test data.

4.2. Setup

We set the BERT parameters in the proposed model as follows: a maximal sequence length of 256, training batch size of 32, learning rate of 5×10^{-6}, and 5 training epochs. We used a standard three-layer perceptron as the final classifier to compare the estimation accuracy of the task owing to the difference in the loss function used for classification. We used a pretrained model (uncased_L-24_H-1024_A-16) publicly available from Google Research (https://github.com/google-research/bert, accessed on 1st October 2020). The compared models in the experiment are as follows:

TextTiling [33]: Baseline model. This is one of the first unsupervised algorithms for linear-text segmentation that uses the fact that words tend to be repeated in coherent segments, and measures the similarity between paragraphs by comparing their sparse term vectors.

Koshorek et al. Model [37]: Baseline model. This is a text-segmentation method based on LSTM. The distributed expression for words contained in a sentence obtained by Word2Vec [58] is input using bidirectional LSTM, and the output is used as the distributed expression of the sentence. For word embeddings, we used the Google News word2vec pretrained model (https://code.google.com/archive/p/word2vec/, accessed on 25 November 2020).

BERT + CE, BERT + FL, BERT + DL, BERT + AL: This model adopts cross entropy loss, focal loss, dice loss, and anchor loss as the loss functions of BERT, respectively. The values of the γ hyperparameters of focal loss and anchor loss were determined through a grid search on the verification data.

BERT + CE + TM, BERT + FL + TM, BERT + DL + TM, BERT + AL + TM: This is a model in which threshold moving is applied to the above-mentioned BERT + CE, BERT + FL, BERT + DL, and BERT + AL. Decision threshold τ was determined through a grid search on the verification data.

We used $F1$ and P_k as evaluation metrics. P_k is an evaluation metric for text segmentation proposed by Beeferman et al. [59]. This metric calculates whether two sentences separated by a distance of k belong to the same segment from both the system-output result and the correct-answer data. The unmatched ratio of both is the score of P_k; the smaller the value is, the better the model performance. According to Koshorek et al. [37], k was set to half the average size of the correct segment.

4.3. Results and Analysis

We adjusted the γ hyperparameters of the loss function in BERT + FL and BERT + AL, and the decision threshold through the experiment for the validation data. Hyperparameter

values shown in Table 3 were set in the models and evaluated on the test data. Table 4 shows the experiment results for each model. The model based on BERT outperformed the estimation accuracy of TextTiling and of the model of Koshorek et al., which is the baseline model.

Table 3. Hyperparameters tuned through grid search on validation data. Value of γ is a hyperparameter for focal loss and anchor loss. Value of τ is the decision threshold. Value of τ for models with threshold moving could be tuned between 0 and 0.5 in 0.01 increments. However, values for BERT + FL and BERT + AL were fixed at 0.5 because threshold moving was not introduced in these models.

	Fitzgerald		Stevenson		Twain	
Model	τ	γ	τ	γ	τ	γ
BERT + FL	0.5	3.0	0.5	2.0	0.5	5.0
BERT + AL	0.5	2.0	0.5	0.5	0.5	1.0
BERT + CE + TM	0.13	-	0.44	-	0.42	-
BERT + DL + TM	0.20	-	0.29	-	0.33	-
BERT + FL + TM	0.14	0.5	0.42	2.0	0.43	5.0
BERT + AL + TM	0.14	2.0	0.39	2.0	0.24	1.0

Table 4. Experiment results for each model. The higher $F1$ is, the better the outcome. By contrast, the lower P_k is, the better the results. Bold values represent the best evaluation results for each dataset.

	Fitzgerald		Stevenson		Twain	
Model	$F1(\uparrow)$	$P_k(\downarrow)$	$F1(\uparrow)$	$P_k(\downarrow)$	$F1(\uparrow)$	$P_k(\downarrow)$
TextTiling	0.494	0.452	0.484	0.441	0.490	0.495
Koshorek et al. model	0.598	0.234	0.420	0.275	0.275	0.315
BERT + CE	0.759	0.158	0.498	0.226	0.591	0.186
BERT + DL	0.783	0.144	0.518	0.219	0.613	0.173
BERT + FL	0.767	0.153	0.519	0.219	0.614	0.170
BERT + AL	0.752	0.161	0.508	0.220	0.607	0.176
BERT + CE + TM	0.773	0.132	0.504	0.223	0.605	0.178
BERT + DL + TM	**0.797**	**0.125**	**0.536**	**0.217**	**0.627**	**0.164**
BERT + FL + TM	0.780	0.133	0.521	0.219	**0.627**	0.165
BERT + AL + TM	0.786	0.128	0.521	**0.217**	0.622	**0.164**

We confirmed that each model using dice loss, focal loss, and anchor loss as the loss functions estimates paragraph boundaries with higher accuracy than that of BERT + CE, which is used in conventional BERT classification tasks. In particular, BERT + DL marked the highest $F1$ and P_k in all datasets. Introducing threshold moving to the models also improved their performance in each dataset. This may have been because the properties of the validation data and test data were similar, and the determination threshold as a hyperparameter adjusted using validation data could identify paragraph boundaries with higher accuracy, even in the test data. These results suggest that the introduction of the cost-sensitive loss function and threshold moving into BERT improves the accuracy of paragraph-boundary estimation as an imbalanced-classification problem.

Figure 4 shows the probability of being a positive example (positive probability) of each sample output by BERT + CE and BERT + DL for the part of the Fitzgerald test dataset. Samples that sunk at the bottom of the figure were predicted to be negative samples, while samples located at the top of the figure were predicted to be positive samples. If the positive probability for a sample was higher than threshold τ, then the sample was determined to be a positive sample. Therefore, the positive probability output for the sample on the red vertical line, which is the actual position of the paragraph boundary, should exceed the threshold, whereas that for the sample not on the red vertical line should not. Since the number of samples judged as positive samples by threshold moving increases, the

number of samples that are correctly identified as positive samples increases, but the number of actually negative samples that are mistakenly identified as positive samples also increases. However, the negative sample, which was mistakenly judged as a positive sample, accounted for a small proportion of the total negative sample. As a result, the value of each evaluation metric indicating the performance of the model improved.

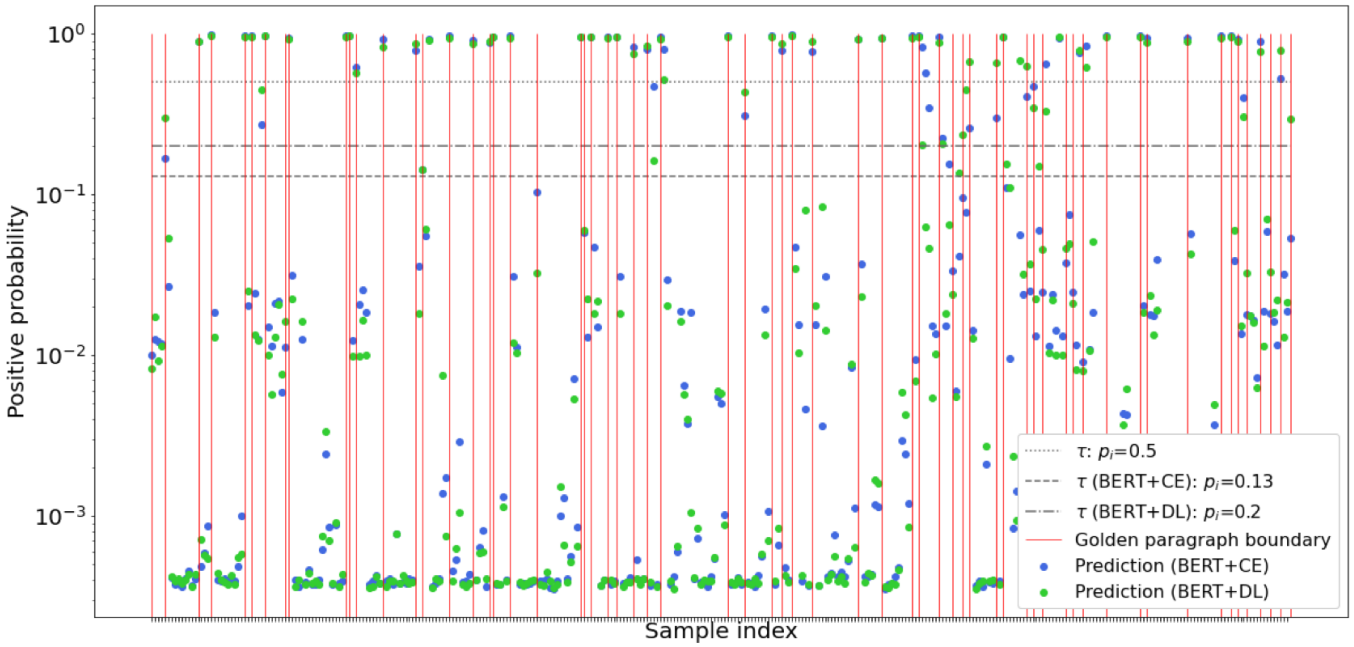

Figure 4. Positive probability in paragraph-boundary recognition for BERT + CE and BERT + DL in the Fitzgerald dataset. Samples are from test data based on the novel "The Great Gatsby". Vertical axis represents the positive probability for each sample $x_i : p_i = p(y_i = 1|x_i)$. Vertical red lines represent golden paragraph boundaries that were actual positions of paragraph breaks. Each dotted line represents the threshold positive probability. Blue points are BERT + CE predictions, and lime green points are BERT + DL predictions.

Table 5 shows examples of the test samples and outputs of BERT + CE and BERT + DL for the samples. Example 1 is a sample in which both BERT + CE and BERT + DL were judged as positive with high probability. In this sample, `Sentence1` and `Sentence2` belonged to different sections, and it is clear that the scenes and topics were different; thus, they could easily be identified as positive examples. By contrast, Examples 2 and 3 are samples in which BERT + CE and BERT + DL were both presumed to differ from the correct label with high probability. The dataset also included samples that were difficult for humans to discriminate because the sentences that they contained were short, and little information was given. Example 4 is a sample that was correctly identified as a positive sample by introducing threshold moving. On the other hand, Example 5 is an actually negative sample mistakenly identified as a positive sample by BERT + CE + TM and BERT + DL + TM. Example 6 was correctly determined to be a positive sample by BERT + DL + TM, but was determined to be a negative sample by BERT + CE, BERT + CE + TM, and BERT + DL.

Table 5. Examples of samples in the Fitzgerald dataset, and positive probability of BERT + CE and BERT + DL for these samples. If the positive probability was higher than the threshold, the sample was recognized as a paragraph boundary.

	Sample Information			Positive Probability p_i	
#	**Label**	**Sentence 1**	**Sentence 2**	**BERT + CE**	**BERT + DL**
1	1	Then he went into the jewellery store to buy a pearl necklace—or perhaps only a pair of cuff buttons—rid of my provincial squeamishness forever.	Gatsby's house was still empty when I left—the grass on his lawn had grown as long as mine.	0.924	0.898
2	0	A puzzled look passed across her face.	Back aft the negroes had begun to sing, and the cool lake, fresh with dawn, echoed serenely to their low voices.	0.955	0.943
3	1	It was the man with owl-eyed glasses whom I had found marvelling over Gatsby's books in the library one night three months before.	I'd never seen him since then.	0.0338	0.0239
4	1	At first I was surprised and confused; then, as he lay in his house and didn't move or breathe or speak, hour upon hour, it grew upon me that I was responsible, because no one else was interested—interested, I mean, with that intense personal interest to which everyone has some vague right at the end.	I called up Daisy half an hour after we found him, called her instinctively and without hesitation.	0.169	0.298
5	0	I was sure he'd start when he saw the newspapers, just as I was sure there'd be a wire from Daisy before noon—but neither a wire nor Mr. Wolfshiem arrived; no one arrived except more police and photographers and newspaper men.	When the butler brought back Wolfshiem's answer I began to have a feeling of defiance, of scornful solidarity between Gatsby and me against them all.	0.270	0.447
6	1	He did not know that it was already behind him, somewhere back in that vast obscurity beyond the city, where the dark fields of the republic rolled on under the night.	So we beat on, boats against the current, borne back ceaselessly into the past.	0.0538	0.294

5. Discussion and Conclusions

In this study, we proposed a method for paragraph-boundary recognition to divide an existing novel into paragraphs from the viewpoint of story understanding by a computer. We regarded paragraph-boundary recognition as a binary classification problem of whether a paragraph boundary existed between two consecutive targeted sentences. However, in this case, the number of paragraphs was extremely small compared to the number of sentences; thus, the data imbalance became a bottleneck. Therefore, we improved the estimation accuracy of the model by introducing cost-sensitive loss functions, namely, focal loss, dice loss, and anchor loss, which are robust against imbalanced classification, into BERT as a loss function. We experimentally confirmed that our approach showed high estimation accuracy compared to that of the conventional text-segmentation method and the original BERT.

Furthermore, we improved estimation accuracy by introducing threshold moving, which adjusts the threshold value when the model determines the presence or absence of paragraph boundaries. It was also experimentally confirmed that paragraph boundaries can be recognized with higher accuracy by setting the determination threshold value as a hyperparameter to a value smaller than the conventional 0.5 using validation data. Threshold moving is a simple idea, but it is expected to improve the performance of the classifier in cases where it is assumed that the properties of validation data and test data are similar, such as the task dealt with in this work.

From the above results, t our work brings to the community of story understanding a new perspective of solving the operation of dividing the text of a novel into paragraphs as an imbalanced classification problem. Our work is also related to the research of creation support. As related research on creative support, there is a plot-creating support system that considers the reader's preference for the transition of happiness in the story [60], and a system that supports efficient story generation using the similarity between sentences and templates [61]. The results obtained in this work could be applied as a system to support the creation of novels by humans in the technique of paragraph division. Specifically, we

envision a system in which a model that learns the paragraph division of existing novels recommends the appropriate paragraph-division position to the writer.

In this study, the effectiveness of the method was confirmed for works written by multiple writers. However, it is necessary to confirm the difference in model performance depending on the characteristics of the work. Therefore, we aim to evaluate the model by cross-validation with each novel used as training or validation data in further work.

Future studies also include the adoption of approaches utilizing the time-series nature of sentences, such as anomaly detection with paragraph breaks as outliers. This method has the advantage of being able to consider not only the information of the preceding and following sentences, but also the information of past sentences.

We evaluated the performance of the paragraph-boundary recognition model by focusing on quantitative indicators. However, in discussing readability as a novel, it is also necessary to qualitatively evaluate the output results. Therefore, in the future, we would like to evaluate the model from both the quantitative evaluation described in this study and qualitative evaluation, such as a questionnaire-based experiment regarding the impression of the collaborator after reading the sentences divided into paragraphs when applying the proposed model.

Author Contributions: Methodology, Validation, Writing—original draft, R.I.; Writing—review and editing, M.O. and N.M. All authors have read and agreed to the published version of the manuscript.

Funding: This work was funded by JSPS KAKENHI Grant, Grant-in-Aid for Scientific Research(B), 19H04184. This work was also funded by JSPS KAKENHI Grant, Grant-in-Aid for Scientific Research(C), 20K11958.

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: Data sharing not applicable.

Acknowledgments: This work was supported by JSPS KAKENHI Grant, Grant-in-Aid for Scientific Research(B), 19H04184. This work was also supported by JSPS KAKENHI Grant, Grant-in-Aid for Scientific Research(C), 20K11958.

Conflicts of Interest: The authors declare no conflict of interest.

References

1. Meehan, J.R. TALE-SPIN, an Interactive Program That Writes Stories, In Proceedings of the 5th International Joint Conference on Artificial Intelligence, Cambridge, MA, USA, 22–25 August 1977; Volume 1, pp. 91–98.
2. Turner, S.R. Minstrel: A Computer Model of Creativity and Storytelling. Ph.D. Thesis, University of California, Los Angeles, CA, USA, 1993.
3. Liu, H.; Singh, P. MAKEBELIEVE: Using Commonsense Knowledge to Generate Stories. In Proceedings of the Eighteenth National Conference on Artificial Intelligence and Fourteenth Conference on Innovative Applications of Artificial Intelligence, Edmonton, AB, Canada, 28 July–1 August 2002; pp. 957–958.
4. Fan, A.; Lewis, M.; Dauphin, Y. Hierarchical Neural Story Generation. In Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), Melbourne, Australia, 15–20 July 2018; pp. 889–898.
5. Martin, L.; Ammanabrolu, P.; Wang, X.; Hancock, W.; Singh, S.; Harrison, B.; Riedl, M. Event Representations for Automated Story Generation with Deep Neural Nets, In Proceedings of the AAAI Conference on Artificial Intelligence, New Orleans, LA, USA, 2–7 February 2018; Volume 32, No. 1.
6. Xu, J.; Ren, X.; Zhang, Y.; Zeng, Q.; Cai, X.; Sun, X. A Skeleton-Based Model for Promoting Coherence Among Sentences in Narrative Story Generation. In Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing, Brussels, Belgium, 31 October–4 November 2018; pp. 4306–4315.
7. Yao, L.; Peng, N.; Weischedel, R.; Knight, K.; Zhao, D.; Yan, R. Plan-and-Write: Towards Better Automatic Storytelling. In Proceedings of the AAAI Conference on Artificial Intelligence, Honolulu, HI, USA, 27 January–1 February 2019; Volume 33, pp. 7378–7385.
8. Brahman, F.; Chaturvedi, S. Modeling Protagonist Emotions for Emotion-Aware Storytelling. In Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing, Online, 16–20 November 2020; pp. 5277–5294.

9. Mostafazadeh, N.; Chambers, N.; He, X.; Parikh, D.; Batra, D.; Vanderwende, L.; Kohli, P.; Allen, J. A Corpus and Cloze Evaluation for Deeper Understanding of Commonsense Stories. In Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, San Diego, CA, USA, 12–17 June 2016; pp. 839–849.
10. Mostafazadeh, N.; Roth, M.; Louis, A.; Chambers, N.; Allen, J. LSDSem 2017 Shared Task: The Story Cloze Test. In Proceedings of the 2nd Workshop on Linking Models of Lexical, Sentential and Discourse-level Semantics, Valencia, Spain, 3 April 2017; pp. 46–51.
11. Guan, J.; Wang, Y.; Huang, M. Story Ending Generation with Incremental Encoding and Commonsense Knowledge. In Proceedings of the AAAI Conference on Artificial Intelligence, Honolulu, HI, USA, 27 January–1 February 2019; Volume 33, No. 1, pp. 6473–6480.
12. Gong, J.; Chen, X.; Qiu, X.; Huang, X. End-to-end Neural Sentence Ordering Using Pointer nNetwork. *arXiv* **2016**, arXiv:1611.04953.
13. Wang, T.; Wan, X. Hierarchical Attention Networks for Sentence Ordering. In Proceedings of the AAAI Conference on Artificial Intelligence, Honolulu, HI, USA, 27 January–1 February 2019; Volume 33, pp. 7184–7191.
14. Cui, B.; Li, Y.; Zhang, Z. BERT-enhanced Relational Sentence Ordering Network. In Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing, Online, 16–20 November 2020; pp. 6310–6320.
15. Seki, Y.; Akahori, K. Effects of Paragraphing Text on Reader Comprehension. *Jpn. J. Educ. Technol.* **1996**, *20*, 97–108.
16. Murakoshi, Y. On The Structure and Method of Paragraph. *Commun. Cult.* **2015**, *9*, 1–27.
17. Itakura, Y.; Shirai, H.; Kuroiwa, J.; Odaka, T.; Ogura, H. Analysis of Coherency of Paragraph in Several Documents. SIG Technical Reports on Information Processing Society of Japan, 2009; Volume 2009-NL-192, pp. 1–6. Available online: https://ipsj.ixsq.nii.ac.jp/ej/?action=pages_view_main&active_action=repository_view_main_item_detail&item_id=62647&item_no=1&page_id=13&block_id=8 (accessed on 17 June 2021)
18. Lin, T.-Y.; Goyal, P.; Girshick, R.; He, K.; Dollar, P. Focal Loss for Dense Object Detection. In Proceedings of the 2017 IEEE International Conference on Computer Vision, Venice, Italy, 22–29 October 2017; pp. 2999–3007.
19. Milletari, F.; Navab, N.; Ahmadi, S.-A. V-Net: Fully Convolutional Neural Networks for Volumetric Medical Image Segmentation. In Proceedings of the 2016 Fourth International Conference on 3D Vision, Stanford, CA, USA, 25–28 October 2016; pp. 565–571.
20. Ryou, S.; Jeong, S.-G.; Perona, P. Anchor Loss: Modulating Loss Scale based on Prediction Difficulty. In Proceedings of the 2019 IEEE International Conference on Computer Vision, Seoul, Korea, 27 October–2 November 2019.
21. Roemmele, M.; Kobayashi, S.; Inoue, N.; Gordon, A. An RNN-based Binary Classifier for the Story Cloze Test. In Proceedings of the 2nd Workshop on Linking Models of Lexical, Sentential and Discourse-level Semantics, Valencia, Spain, 3 April 2017; pp. 74–80.
22. Devlin, J.; Chang, M.-W.; Lee, K.; Toutanova, K. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. In Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers), Minneapolis, MN, USA, 2–7 June 2019; pp. 4171–4186.
23. Li, Z.; Ding, X.; Liu, T. TransBERT: A Three-Stage Pre-training Technology for Story-Ending Prediction. *ACM Trans. Asian Low Resour. Lang. Inf. Process.* **2021**, *20*, 1–20.
24. Cui, Y.; Che, W.; Zhang, W.-N.; Liu, T.; Wang, S.; Hu, G. Discriminative Sentence Modeling for Story Ending Prediction. In Proceedings of the AAAI Conference on Artificial Intelligence, New York, NY, USA, 7–12 February 2020; Volume 34, pp. 7602–7609.
25. Sharma, R.; Allen, J.; Bakhshandeh, O.; Mostafazadeh, N. Tackling the Story Ending Biases in The Story Cloze Test. In Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers), Melbourne, Australia, 15–20 July 2018; pp. 752–757.
26. Tian, Z.; Zhang, Y.; Liu, K.; Zhao, J.; Jia, Y.; Sheng, Z. Scene Restoring for Narrative Machine Reading Comprehension. In Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing, Online, 16–20 November 2020; pp. 3063–3073.
27. Gupta, P.; Bannihatti Kumar, V.; Bhutani, M.; Black, A.W. WriterForcing: Generating More Interesting Story Endings, In Proceedings of the Second Workshop on Storytelling, Florence, Italy, 1 August 2019; pp. 117–126.
28. Mori, Y.; Yamane, H.; Mukuta, Y.; Harada, T. Finding and Generating a Missing Part for Story Completion. In Proceedings of the 4th Joint SIGHUM Workshop on Computational Linguistics for Cultural Heritage, Social Sciences, Humanities and Literature, Online, 12 December 2020; pp. 156–166.
29. Barzilay, R.; Lapata, M. Modeling Local Coherence: An Entity-based Approach. *Comput. Linguist.* **2008**, *34*, 1–34. [CrossRef]
30. Vinyals, O.; Bengio, S.; Kudlur, M. Order Matters: Sequence to Sequence for Sets. *arXiv* **2016**, arXiv:1511.06391.
31. Logeswaran, L.; Lee, H.; Radev, D. Sentence Ordering and Coherence Modeling Using Recurrent Neural Networks. In Proceedings of the AAAI Conference on Artificial Intelligence, New Orleans, LA, USA, 2–7 February 2018; Volume 32, No. 1.
32. Oh, B.; Seo, S.; Shin, C.; Jo, E.; Lee, K.H. Topic-Guided Coherence Modeling for Sentence Ordering by Preserving Global and Local Information. 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), Hong Kong, China, 3–7 November 2019; pp. 2273–2283.
33. Hearst, M.A. Multi-Paragraph Segmentation Expository Text. In Proceedings of the 32nd Annual Meeting of the Association for Computational Linguistics, Las Cruces, NM, USA, 27–30 June 1994; pp. 9–16.

34. Utiyama, M.; Isahara, H. A Statistical Model for Domain-Independent Text Segmentation. In Proceedings of the 39th Annual Meeting of the Association for Computational Linguistics, Toulouse, France, 9–11 July 2001; pp. 499–506.
35. Brants, T.; Chen, F.; Tsochantaridis, I. Topic-based Document Segmentation with Probabilistic Latent Semantic Analysis. In Proceedings of the Eleventh International Conference on Information and Knowledge Management, McLean, VA, USA, 4–9 November 2002; pp. 211–218.
36. Glavaš, G.; Nanni, F.; Ponzetto, S.P. Unsupervised Text Segmentation Using Semantic Relatedness Graphs. In Proceedings of the Fifth Joint Conference on Lexical and Computational Semantics, Berlin, Germany, 11–12 August 2016; pp. 125–130.
37. Koshorek, O.; Cohen, A.; Mor, N. ;Rotman, M.; Berant, J. Text Segmentation as a Supervised Learning Task. In Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 2 (Short Papers), New Orleans, LA, USA, 1–6 June 2018; pp. 469–473.
38. Badjatiya, P.; Kurisinkel, L.J.; Gupta, M.; Varma, V. Attention-based Neural Text Segmentation. In *Advances in Information Retrieval*; Springer: Cham, Switzerland, 2018; pp. 180–193.
39. Glavaš, G.; Somasundaran, S. Two-Level Transformer and Auxiliary Coherence Modeling for Improved Text Segmentation. In Proceedings of the AAAI Conference on Artificial Intelligence, New York, NY, USA, 7–12 February 2020; Volume 34, pp. 7797–7804.
40. Liu, X.-Y.; Wu, J.; Zhou, Z.-H. Exploratory Undersampling for Class-Imbalance Learning. *IEEE Trans. Syst. Man Cybern.* **2009**, *39*, 539–550.
41. Zhang J.; Mani, I. KNN Approach to Unbalanced Data Distributions: A Case Study Involving Information Extraction. In Proceedings of the ICML'2003 Workshop on Learning from Imbalanced Datasets, Washington, DC, USA, 21 August 2003.
42. Chawla, N.V.; Bowyer, K.W.; Hall, L.O.; Kegelmeyer, W.P. SMOTE: Synthetic Minority Over-sampling Technique. *J. Artif. Intell. Res.* **2002**, *16*, 321–357. [CrossRef]
43. Shen, C.; Roth, H.R.; Oda, H.; Oda, M.; Hayashi, Y.; Misawa, K.; Mori, K. On the Influence of Dice Loss Function in Multi-class Organ Segmentation of Abdominal CT Using 3D Fully Convolutional Networks. *arXiv* **2018**, arXiv:1801.05912.
44. Kodym, O.; Španěl, M.; Herout, A. Segmentation of Head and Neck Organs at Risk Using CNN with Batch Dice Loss. *arXiv* **2018**, arXiv:1812.02427.
45. Iikura, R.; Okada, M.; Mori, N. Paragraph Segmentation for Novels using BERT with Focal Loss. In Proceedings of the 34th Annual Conference of the Japanese Society for Artificial Intelligence, Kumamoto, Japan, 9–12 June 2020.
46. Iikura, R.; Okada, M.; Mori, N. Improving BERT with Focal Loss for Paragraph Segmentation of Novels. In Proceedings of the 17th International Conference on Distributed Computing and Artificial Intelligence, L'Aquila, Italy, 16–19 June 2020; pp. 21–30.
47. Iikura, R.; Okada, M.; Mori, N. Automatic Paragraph Segmentation of Novels as Imbalanced Classification. *J. Inf. Process. Soc. Jpn.* **2021**, *62*, 891–902.
48. Li, X.; Sun, X.; Meng, Y.; Liang, J.; Wu, F.; Li, J. Dice Loss for Data-imbalanced NLP Tasks. *arXiv* **2019**, arXiv:1911.02855.
49. Provest, F. Machine Learning from Imbalanced Datasets 101. Invited Paper for the AAAI, Workshop on Imbalanced Data Sets; 2000. Available online: https://archive.nyu.edu/bitstream/2451/27763/2/CPP-02-00.pdf (accessed on 17 June 2021)
50. Maloof, M.A. Learning When Data Sets are Imbalanced and When Costs are Unequal and Unknown. In Proceedings of the Workshop on Learning from Imbalanced Data Sets II, ICML, Washington, DC, USA, 21 August 2003.
51. Zhou, Z.-H.; Liu, X.-Y. Training Cost-sensitive Neural Networks with Methods Addressing the Class Imbalance Problem. *IEEE Trans. Knowl. Data Eng.* **2006**, *18*, 63–77. [CrossRef]
52. Collell, G.; Prelec, D.; Patil, K. Reviving Threshold-Moving: A Simple Plug-in Bagging Ensemble for Binary and Multiclass Imbalanced Data. *arXiv* **2017**, arXiv:1606.08698.
53. Vaswani, A.; Shazeer, N.; Parmar, N.; Uszkoreit, J.; Jones, L.; Gomez, A.N.; Łukasz, K.; Polosukhin, I. Attention is All you Need. In Proceedings of the 31st International Conference on Neural Information Processing Systems, Long Beach, CA, USA, 4–9 December 2017; pp. 6000–6010.
54. Wu, Y.; Schuster, M.; Chen, Z.; Le, Q.V.; Norouzi, M.; Macherey, W.; Krikun, M.; Cao, Y.; Gao, Q.; Macherey, K.; et al. Google's Neural Machine Translation System: Bridging the Gap Between Human and Machine Translation. *arXiv* **2016**, arXiv:1609.08144.
55. Madabushi, H.T.; Kochkina, E.; Castelle, M. Cost-Sensitive BERT for Generalisable Sentence Classification on Imbalanced Data. In Proceedings of the Second Workshop on Natural Language Processing for Internet Freedom: Censorship, Disinformation, and Propaganda, Hong Kong, China, 4 November 2019; pp. 125–134.
56. Abraham, N.; Khan, N.M. A Novel Focal Tversky Loss Function With Improved Attention U-Net for Lesion Segmentation. In Proceedings of the 2019 IEEE 16th International Symposium on Biomedical Imaging, Venice, Italy, 8–11 April 2019; pp. 683–687.
57. Bird, S.; Loper, E.; Klein, E. *Natural Language Processing with Python*; O'Reilly Media Inc.: Sebastopol, CA, USA, 2009.
58. Mikolov, T.; Sutskever, I.; Chen, K.; Corrado, G.S.; Dean, J. Distributed Representations of Words and Phrases and their Compositionality. In Proceedings of the 26th International Conference on Neural Information Processing Systems, Lake Tahoe, NV, USA, 5–10 December 2013; pp. 3111–3119.
59. Beeferman, D.; Berger, A.; Lafferty, J. Statistical Models for Text Segmentation. *Mach. Learn.* **1999**, *34*, 177–210. [CrossRef]
60. Ashida, A.; Kojiri, T. Plot-creation Support with Plot-construction Model for Writing Novels. *J. Inf. Telecommun.* **2019**, *3*, 57–73. [CrossRef]
61. Katsui, T.; Ueno, M.; Isahara, H. Search for Similar Story Sentences based on Role of Characters in order to Support and Analyze Contents Creator's Ideas, In Proceedings of the 33rd Annual Conference of the Japanese Society for Artificial Intelligence, Niigata, Japan, 4–7 June 2019.