概要

本研究は日本語の文章に対して，BERT と最大エントロピー法を用いることにより，段落分割の自動推定を行い，どちらが優れているかを比較することを目指す．

段落とは，文章において段落とは読み手が文章を読む上で，また書き手も伝えたいことを表現する上で大切なものである．形式段落に関する研究では，飯倉ら [1] の Focal Loss を利用した BERT の小説の形式段落を求めた研究がある．また日本語以外では，Genzel[2] の英語のテキストに機械学習を利用して，段落開始文と非段落開始文を分けた研究がある．しかし，飯倉ら [1] の研究では BERT の形式段落推定に対しての有用性は求めたが，BERT 以外の他手法との比較，素性分析が行われていない．そこで本研究では，日本語の文章に対して，BERT と最大エントロピー法を用いて段落分割の自動推定を行い，どちらが優れているかを比較し，段落分割の結果から分割，非分割に関する素性を得る．

本研究の成果は 3 つある．1 つ目は，新聞記事と小説を用いて段落分割の推定を行い，BERT は最大エントロピー法の正解率に対して全て上回っており，BERT と最大エントロピー法の段落分割の正解率を比較することで，BERT による段落分割手法の優位性を示した．2 つ目は，BERT を用いた段落分割のテストデータを 3 単語ごとに分け，分けた 3 単語それぞれに対する分割，非分割の出力値から素性分析を行う新しい手法を提案し，分割に関する素性は 1 つしか得ることができなかったが，非分割に関する素性は得ることができた．3 つ目は，最大エントロピー法と BERT の素性分析を比較することで，素性分析において最大エントロピー法の方が有効な手法であることを示した．

3 つの成果から，更に推定精度を上げることで，書き手に対して文章を生成する際の段落作成，修正や読み手に対して文章理解の支援に役立つと考えている．

目 次

表 目 次

図 目 次

第1章　はじめに

文章において段落とは読み手が文章を読む上で，また書き手も伝えたいことを表現する上で大切なものである．段落の有無で文章の可読性が変わり，1つの段落で複数の内容が入っていると理解しにくく，とても読みにくいものである．内容や場面の転換に基づき段落分けがなされることは，読み手に対して十分な理解を促す．

段落には，形式段落と意味段落の2種類がある．一般的に形式段落は，形式上ひとまとまりになっている段落のことを指し，文頭を1字下げたところから改行までのまとまり．意味段落は，一つの文章を内容や意味に応じて分けたまとまりのことを指し，一つ以上の形式段落からなるまとまりのことを意味している．

文章の形式段落を推定するにあたって，文と文の間の箇所が段落であるか，非段落であるかといった2クラス分類問題とすることで，段落の推定を行うことができる．

形式段落に関する研究では，飯倉ら [1] の Focal Loss を利用した BERT の小説の形式段落を求めた研究がある．また日本語以外では，Genzel[2] の英語のテキストに機械学習を利用して，段落開始文と非段落開始文を分けた研究，Caroline ら [3] の英語，ドイツ語，ギリシャ語の段落の境界を自動的に予測する研究がある．しかし，日本語での形式段落の研究は飯倉ら [1] の BERT の研究のみで，BERT での段落分割に対する有用性について述べているが，他手法との比較を行っていない．

その問題を解決するため，本研究では BERT と最大エントロピー法を用いて形式段落を推定し，手法間での性能差の比較を行う．BERT とは自然言語処理のタスクにおいて高い精度が示されており，損失関数として2クラス分類問題を解く上で基本的な Softmax Cross Entropy Loss を使用することで段落分割を行う．また最大エントロピー法とは，入力したそれぞれの素性の判定における寄与率を数値化できる機械学習法である．分割，非分割に対する確率値を算出することで、段落分割の推定を行う．形式段落の推定を行うことで，文章を生成する際の段落の作成や修正の支援に役立つと考えている．

また，先行研究では段落の推定しか行っておらず，なぜ段落分割されたか分割されなかったかといった考察がない．そこで本研究では，最大エントロピー法では正規化

α 値を，BERT では Softmax 関数の出力値を用いることで，段落の分割，非分割に影響を与えた単語を取得し，素性分析を行う．BERT の結果を用いた素性分析の研究はなく，本研究では手法を新たに提案している．

本研究の主な主張点を以下に整理する．

- 段落分割の推定ために BERT と最大エントロピー法を使用し，新聞記事と小説に対して実験を行った結果，全ての正解率がベースラインの正解率を上回っている．
- 段落分割に BERT，最大エントロピー法を用い，それぞれの手法での正解率を比較することで，BERT での段落分割の手法の優位性を示した．
- BERT の結果を用いた素性分析の研究はなく，新規の研究である．
- BERT の結果に対して，新たに提案した 3 単語連続を入力とする手法で素性分析を行った結果，分割に関する素性は 1 つしか得ることができなかったが，非分割に関する素性を得ることができた．
- 素性分析を行う上で，最大エントロピー法は BERT よりも多くの素性を得ることができることから，素性分析においての最大エントロピー法での手法の優位性を示した．

本論文の構成は以下の通りである．第 2 章では，本研究に関連する研究としてどのような研究が行われてきたかを記述し，その研究と本研究との関連を説明する．第 3 章では，本研究が扱う問題の設定とそれを解決するために提案した手法について説明を行う．第 4 章では，本研究が行った実験についての説明と，その結果と考察について記述する．第 5 章では，素性分析について結果と考察について記述する．第 6 章では，今後の課題について記述する．第 7 章では，まとめを行う．

第2章　先行研究

本章では，先行研究について記述する．2.1 節では，飯倉ら [1] が行った Focal Loss を利用した BERT によって，小説の形式段落を推定した研究について記述する．2.2 節では，Genzel[2] が行った段落開始文と非段落開始文を分けた研究について記述する．2.3 節では，Caroline ら [3] の段落の境界を自動的に予測する研究について記述する．2.4 節では，中野ら [4] の話題境界判定モデルを提案し，意味段落を求めた研究について記述する．

2.1　Focal Loss を利用した BERT による小説の段落境界推定

飯倉らは，損失関数に Focal Loss を使用した BERT を用いて，小説の段落の境界推定を行った [1].

飯倉らは，BERT の損失関数に Binary Cross Entropy(BCE) Loss と Focal Loss を用い，2 つの手法で形式段落の推定精度を比較し，Focal Loss の形式段落の推定に対する分類器としての性能の向上を示した．実験で用いている BCE Loss とは，2 値分類の問題において使われる損失関数であり，データに含まれるクラスごとのインスタンスが不均一であるときに学習がうまくいかないことを防ぐ．また，Focal Loss とは，識別が容易な例からのエラーの寄与を減衰させることで，損失関数を圧倒することを防ぐ係数を導入し，困難な識別を可能とした損失関数である．

実験には，推定箇所の前の部分を sentence1，推定箇所の後の部分を sentence2 としたデータ A と推定箇所の前後の 1 文を sentence1，推定箇所の後ろの 2，3 文を sentence2 としたデータ B の 2 種類で実験を行った．データ B の方が精度は高く，BCE Loss を用いた結果は 0.8262 に対して Focal Loss は 0.8315 の F 値を示した．

2.2 英語の文章の形式段落の推定

Genzel は，英語の文章を段落開始文と非段落開始文を分ける研究を行った [2].

Genzel は，Penn Treebank と War and Peace を用いて別々の実験を行い，また様々な小説を無作為に選んで同様の実験を行った．Penn Treebank に含まれる情報を以下に示す.

- テキスト境界
  - 各文章の最初の文-1，それ以外の文 0 として，文章の最初の文は段落となるので最初の文を区別する
- 品詞
  - 各品詞について，現在の文とその前の文で発生した回数
- 文章の長さ
  - 現在の文の文章の長さ
- 最初の単語
  - 文中の最初の単語のみ
- 主題タイプ
  - 各対象タイプについて，現在の文の対象がこのタイプであるかどうか
- 内部ノード
  - 解析木の各内部ノードについて文中に出現する回数
- cosine
  - コサイン特徴

これらの素性を用いて Penn Treebank で実験を行った．段落開始文と非段落開始文を分ける推定精度は，ベースラインの正解率が 0.55 に対して，Genzel の提案手法の正解率が 0.67 であった.

また，War and Peace に含まれる情報を以下に示す.

- 語彙
  - 現在の文および前の文の語彙内の各単語についての出現回数
- 文章の長さ
- cosine
  - コサイン特徴
- 最初の単語
- 内部ノードの大きさ
  - 解析木の各内部ノードについて，その平均の大きさ

これらの素性を用いて War and Peace で実験を行った．段落開始文と非段落開始文を分ける推定精度は，ベースラインの正解率が 0.63 に対して，Genzel の提案手法の正解率が 0.78 であった．

また，プロジェクト・グーテンベルクから小説を無作為に 5 つ選んで，War and Peace を使用して段落開始文と非段落開始文を分ける実験を行った．「3 Musketeers」に対して, 段落開始文と非段落開始文を分けた時の推定精度が一番高く，推定精度はベースラインの正解率が 0.58 に対して，Genzel の提案手法の正解率が 0.75 であった．

2.3 機械学習を用いた英語, ドイツ語, ギリシャ語の文章の形式段落の推定

Caroline らは，テキスト，構文，談話のてがかりの 3 つに着目して，教師あり機械学習を用いて段落の境界を自動的に予測できるかどうかを調べた [3].

Caroline らは，BoosTexter を機械学習システムとして使用し，英語，ドイツ語，ギリシャ語の 3ヶ国語の小説，報道，議事についての 3 つの分野からそれぞれコーパスを作成して実験を行った．使用した機能は非構文機能，言語モデリング機能，構文機能の 3 つで構文機能は英語にしか適用されなかった．これらの機能を用いて，段落の境界を自動的に予測する研究を行った．

3ヶ国語の実験結果のうちベースラインとの正解率の差が大きかったものを挙げると，英語では報道に関するコーパスを用いた段落の境界の推定精度が高かった．段落の境界の推定精度はベースラインの正解率が 0.51 に対して，Caroline らの手法では正解率が 0.71 であった．またドイツ語では議事に関するコーパスを用いた段落の境界の推定精度が高かった．段落の境界の推定精度はベースラインの正解率が 0.66 に対して，Caroline らの手法では 0.79 であった．またギリシャ語では報道に関するコーパスを用いた段落の境界の推定精度が高かった．段落の境界の推定精度はベースラインの正解率が 0.53 に対して，Caroline らの手法では 0.76 であった．

2.4 反復語の文間距離を用いた意味段落の推定

中野らは，語が近接して反復する区間に話題形成ポテンシャルが生じ，それが話題結束力を形成するという話題境界判定モデルを考案した [4]．

中野らは，同一語が文をまたがって反復するときに，反復語が出現するすべての文間の組み合わせ，それらを反復セットと呼び話題形成の対象要素とした．反復セットにより，反復区間内の隣接する文と文の間に区間距離に反比例した話題を形成するポテンシャルが生じ，これを話題形成ポテンシャルと呼ぶ．すべての反復セットによる話題形成ポテンシャルを集計したものが隣接する文間の結束力を示すものとして，話題結束力と呼ぶ．話題結束力の高まりは話題の存在を示し，低下した箇所が話題境界となっているので，その位置が意味段落の境界であると推定した．社説記事を連結して，記事境界と段落分割点が一致したときを正解とする境界判定実験を行い，再現率 0.678，適合率 0.618 の精度を得た．

第3章　問題設定と提案手法

本章では，本研究で扱う問題と提案手法の説明を記述する．3.1 節では, 本研究で扱う問題設定について記述している．3.2 節では，本研究で使用する学習ツールについて説明を記述している．3.3 節では，提案手法について記述している．3.4 節では，素性分析について記述している．

3.1　問題設定

段落分割推定したい文章があるとする．その文章中の 1 文目以降の全ての文と文の間の箇所を「？」とし，その位置が段落箇所であるか否かを機械学習で推定する．図 3.1 のように，ある文とその文の直前の 1 文を入力として，それらの文の間の箇所が段落分割位置であるか否かを出力とする．

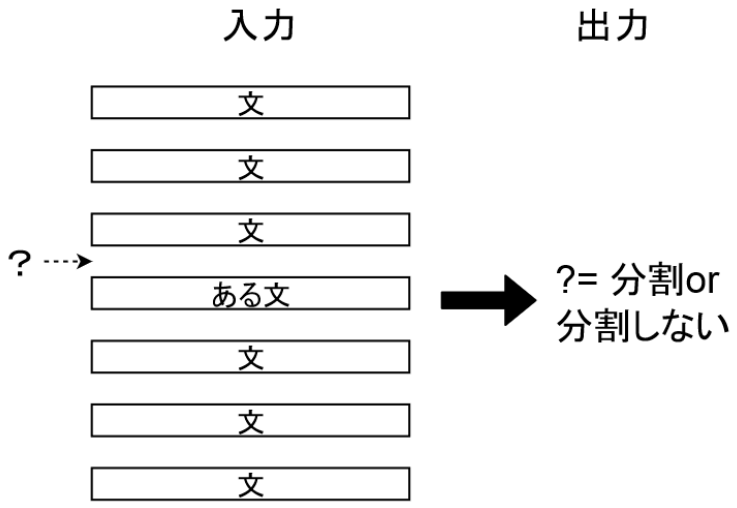

図 3.1: 段落分割

元々段落分けされていた文章の段落を取り除いて学習データとすることで，機械学習が段落だと決めた箇所が正しいかを判断することができ，そこから導き出される正

解率を求める．また本研究で推定する形式段落は，文頭を 1 文字下げた箇所を形式段落段落段落と定義する．

3.2 学習ツール

本研究で使用する学習ツールについて記述している．3.2.1 節では, 最大エントロピー法について記述している．3.2.2 節では，BERT について記述している．3.2.3 節では，サポートベクトルマシン法について記述している．

3.2.1 最大エントロピー法

本研究では，教師あり機械学習法に，最大エントロピー法を使用する．

最大エントロピー法は，あらかじめ設定しておいた素性 $f_j(1 \leq j \leq k)$ の集合を F とするとき，式 (3.1) を満足しながらエントロピーを意味する式 (3.1) を最大にするときの確率分布 $p(a,b)$ を求め，その確率分布にしたがって求まる各分類の確率のうち，もっとも大きい確率値を持つ分類を求める分類とする方法である [5, 6, 7, 8]．

$$\sum_{a \in A, b \in B} p(a,b) g_j(a,b) \;=\; \sum_{a \in A, b \in B} \tilde{p}(a,b) g_j(a,b) \;\; for \; \forall f_j \;\; (1 \leq j \leq k)$$

$$H(p) \;\; = \;\; - \sum_{a \in A, b \in B} p(a,b) \; log \, (p(a,b)) \tag{3.1}$$

ただし，A, B は分類と文脈の集合を意味し，$g_j(a,b)$ は 文脈 b に素性 f_j があってなおかつ分類が a の場合 1 となりそれ以外で 0 となる関数を意味する．また，$\tilde{p}(a,b)$ は，既知データでの (a,b) の出現の割合を意味する．また，k は素性の総数を意味する．

$p(a,b)$ を求め，そこから $p(a|b)$ を求める．$(p(a|b) = p(a : \text{解答} \, | b : \text{問題}))$

$$p(a|b) = \frac{p(a,b)}{\sum_i p(a_i, b)} \tag{3.2}$$

式 (3.1) は確率 p と出力と素性の組の出現を意味する関数 g をかけることで出力と素性の組の頻度の期待値を求めることになっており，右辺の既知データにおける期待

値と，左辺の求める確率分布に基づいて計算される期待値が等しいことを制約として，エントロピー最大化 (確率分布の平滑化) を行って，出力と文脈の確率分布を求めるものとなっている．

3.2.2 BERT

BERTとは，Bidirectional Encoder Representations from Transformersの略で,「Transformerによる双方向のエンコード表現」と訳され，2018年10月にGoogleのJacob Devlinらの論文で発表された自然言語処理モデルである[9]．従来の自然言語処理では，大量のラベルのついたデータを用意させ，処理を行うことで課題に取り組む．しかし従来の手法に対し，BERTは事前学習でラベルのないデータをはじめに大量に処理を行う．その後，ファインチューニングで少量のラベルのついたデータを使用することで課題に対応させる．

本研究では，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，京都大学の黒橋・村脇研究室で公開されている学習済みBERTモデル[10]を使用する．このモデルでは，JUMAN++[1]によって形態素解析を行い，Byte Pair Encoding(BPE)によってsubwordに分割する．また基本的なモデルの設定については，ベースのTransformerの層数L=12，隠れベクトルの次元数H=768，Multi-headのSelf-Attention機構のヘッド数A=12であり，この値は一般に配布されている英語の事前学習済みモデルと等しい．

またBERTのモデルでの損失関数は，クラス分類問題を解く基本的なSoftmax Cross Entropy Lossを用いる．Softmax Cross Entropy Lossは式(3.3)，式(3.4)で表される．

$$y_{ij} = \frac{exp(x_{ij})}{\sum_{k=1}^{n} exp(x_{ik})} \tag{3.3}$$

$$L = -\frac{1}{m}\sum_{ij} t_{ij} log y_{ij} \tag{3.4}$$

式(3.3)のSoftmax関数によって複数値からなるベクトルの入力に対して，正規化したベクトルを出力する．そのとき出力されるベクトルは合計値が1になる．本実験では出力されたベクトルを確率とみなし，段落の分割，非分割の2クラス分類問題にあてることで段落の推定を行う．

[1] https://github.com/ku-nlp/jumanpp

3.2.3 サポートベクトルマシン法

サポートベクトルマシン法は，空間を超平面で分割することにより 2 つの分類からなるデータを分類する手法である．このとき，2 つの分類が正例と負例からなるものとすると，学習データにおける正例と負例の間隔 (マージン) が大きいもの (図 3.2 参照[2]) ほどオープンデータで誤った分類をする可能性が低いと考えられ，このマージンを最大にする超平面を求めそれを用いて分類を行う．基本的には上記のとおりであるが，通常，学習データにおいてマージンの内部領域に少数の事例が含まれてもよいとする手法の拡張や，超平面の線形の部分を非線型にする拡張 (カーネル関数の導入) がなされたものが用いられる．この拡張された方法は，以下の識別関数を用いて分類することと等価であり，その識別関数の出力値が正か負かによって二つの分類を判別することができる [11, 12].

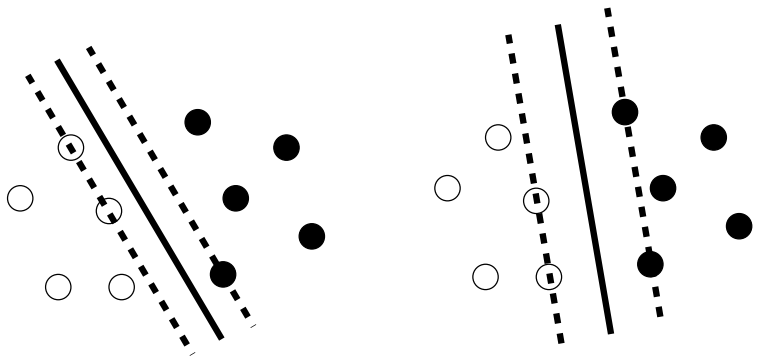

図 3.2: マージン最大化

$$
\begin{aligned}
f(\mathbf{x}) &= sgn\left(\sum_{i=1}^{l}\alpha_i y_i K(\mathbf{x}_i,\mathbf{x})+b\right) \\
b &= -\frac{max_{i,y_i=-1}b_i + min_{i,y_i=1}b_i}{2} \\
b_i &= \sum_{j=1}^{l}\alpha_j y_j K(\mathbf{x}_j,\mathbf{x}_i)
\end{aligned}
\tag{3.5}
$$

ただし，$\mathbf{x}$ は識別したい事例の文脈 (素性の集合) を，$\mathbf{x}_i$ と $y_i(i=1,...,l,y_i\in\{1,-1\})$ は学習データの文脈と分類先を意味し，関数 sgn は，式 3.6 である．

[2]図の白丸，黒丸は，正例，負例を意味し，実線は空間を分割する超平面を意味し，破線はマージン領域の境界を表す面を意味する．

$$sgn(x) = \begin{array}{ll} 1 & (x \geq 0) \\ -1 & (otherwise) \end{array} \tag{3.6}$$

また，各 α_i は式 (3.8) と式 (3.9) の制約のもと式 (3.7) の $L(\alpha)$ を最大にする場合のものである．

$$L(\alpha) = \sum_{i=1}^{l} \alpha_i - \frac{1}{2} \sum_{i,j=1}^{l} \alpha_i \alpha_j y_i y_j K(\mathbf{x_i}, \mathbf{x_j}) \tag{3.7}$$

$$0 \leq \alpha_i \leq C \ (i = 1, ..., l) \tag{3.8}$$

$$\sum_{i=1}^{l} \alpha_i y_i = 0 \tag{3.9}$$

また，関数 K はカーネル関数と呼ばれ，様々なものが用いられるが本稿では以下の多項式のものを用いる．

$$K(\mathbf{x}, \mathbf{y}) = (\mathbf{x} \cdot \mathbf{y} + 1)^d \tag{3.10}$$

C, d は実験的に設定される定数である．本稿では C,d はともにすべての実験を通して 1 に固定した．ここで，$\alpha_i > 0$ となる $\mathbf{x}_i$ は，サポートベクトルと呼ばれ，通常，式 (3.5) の和をとっている部分はこの事例のみを用いて計算される．つまり，実際の解析には学習データのうちサポートベクトルと呼ばれる事例のみしか用いられない．

サポートベクトルマシン法は分類の数が 2 個のデータを扱うもので，通常これにペアワイズ手法を組み合わせて用いることで，分類の数が 3 個以上のデータを扱うことになる [13].

ペアワイズ手法とは，N 個の分類を持つデータの場合，異なる二つの分類先のあらゆるペア (N(N-1)/2 個) を作り，各ペアごとにどちらがよいかを 2 値分類器 (ここではサポートベクトルマシン法[3]) で求め，最終的に N(N-1)/2 個の 2 値分類器の分類先の多数決により，分類先を求める方法である．

[3]本稿の 2 値分類器としてのサポートベクトルマシンは，工藤氏が作成した TinySVM[12] を利用している．

本稿のサポートベクトルマシン法は，上記のようにサポートベクトルマシン法とペアワイズ手法を組み合わせることによって実現される．

3.3 提案手法

本研究では，3.1 節の問題設定に対して 3.2 節で記述した 3 種類の学習ツールを用いることで，段落分割の推定を行う．3.3.1 節では，最大エントロピー法を用いた推定方法について記述している．3.3.2 節では，BERT を用いた推定方法について記述している．3.3.3 節では，サポートベクトルマシン法を用いた推定方法について記述している．3.4 節では，素性分析について記述している．

3.3.1 最大エントロピー法での推定方法

本研究では，最大エントロピー法を用いて，文章の段落推定を行うために様々な素性を用いて実験を行う．本研究で用いる素性を以下に示す．

① 文と文の間の箇所「？」(以下「文間箇所」) の直前の 1 文と直後の 1 文にある全単語

- 図 3.3 の「？」の箇所の直前の 1 文と直後の 1 文に対して，MeCab を用いることで形態素解析を行い推定した．以下に整形前の新聞記事を図 3.4 に，単語ごとに分け整形した文を図 3.5 に示す．図 3.5 の「文-1」は直前の 1 文，「文 1」は直後の 1 文のことを示している．

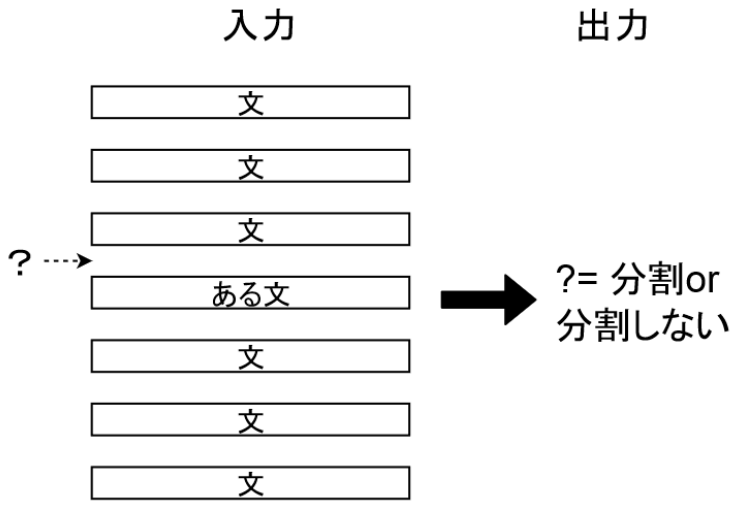

図 3.3: 文章例

文-１：久間章生元防衛相が自民党総務会長だった０６年１月、福井県敦賀市の知人男性（６４）から１億円を受領していたことが分かった。
文１：久間氏と男性は毎日新聞の取材に授受を認め「賃借関係」としているが、無担保で実行された形になっている。

図 3.4: 新聞記事

文-1:久間　文-1:章生　文-1:元　文-1:防衛　文-1:相　文-1:が　文-1:自民党　文-1:総務　文-1:会長　文-1:だっ　文-1:た　文-1:０　文-1:6　文-1:年　文-1:1月　文-1:、　文-1:福井　文-1:県　文-1:敦賀　文-1:市　文-1:の　文-1:知人　文-1:男性　文-1:（　文-1:6　文-1:4　文-1:）　文-1:から　文-1:1　文-1:億　文-1:円　文-1:を　文-1:受領　文-1:し　文-1:て　文-1:い　文-1:た　文-1:こと　文-1:が　文-1:分かっ　文-1:た　文-1:。
文1:久間　文1:氏　文1:と　文1:男性　文1:は　文1:毎日新聞　文1:の　文1:取材　文1:に　文1:授受　文1:を　文1:認め　文1:「　文1:貸借　文1:関係　文1:」　文1:と　文1:し　文1:て　文1:いる　文1:が　文1:、　文1:無　文1:担保　文1:で　文1:実行　文1:さ　文1:れ　文1:た　文1:形　文1:に　文1:なっ　文1:て　文1:いる　文1:。

図 3.5: 文間箇所の直前，直後の 1 文にある全単語

図 3.5 のように，直前の 1 文と直後の 1 文にある全単語を素性として追加する．また直前 2 文と直後 2 文，直前 3 文と直後 3 文も文章数が増えるだけで，MeCab を用いて形態素解析を行い同様に単語ごとに分ける．

② 文間箇所の直前の 2 文と直後の 2 文にある全単語

③ 文間箇所の直前の 3 文と直後の 3 文にある全単語

④ 文間箇所の直前の 1 文と直後の 1 文の文頭の単語

– 文頭の単語の情報を素性として推定した．接続詞など段落の分割に大きく影響を及ぼす品詞などを文頭の素性することで，推定精度の向上に繋がると推測した．以下に整形前の新聞記事を図 3.6 に，整形を行った文頭の素性を図 3.7 に示す．図 3.7 の「文頭-1」は直前の 1 文の文頭,「文頭 1」は直後の 1 文の文頭のことを示している．

文-１：そこで私たちの狩猟文化や食などの生活様式が築かれた。
文１：しかし、時代の変化が速く、（都市化の発展で）薬物中毒や暴力、自殺などの問題が表面化した。

図 3.6: 新聞記事

文頭-１：そこで　　文頭１：しかし

図 3.7: 文頭の単語

⑤ 文間箇所の直前, 直後の文中の同単語の出現数

– 2 文章に続けて同じ単語が出現する場合，その 2 つの文は同じ段落内である可能性が高いと考えられる．以下に同単語が出現する例文を図 3.8 に示す．

文-１：一つの爆弾に仕組まれた多数の小さな爆弾が広範囲に飛び散り、無差別に人を殺傷するクラスター爆弾。
文１：クラスター爆弾は、イラクやアフガニスタンをはじめ多くの戦争で使われ、不発のまま残った子爆弾は戦後も悲劇を生む。

図 3.8: 同単語例文

図 3.8 より「クラスター爆弾」という固有名詞が 2 文に続けて出現しており，例文の 2 文章は同段落内である．

素性として追加する際は，文-1 と文 1 の名詞の一致数を調べ，3 種類の情報を追加する．1 つ目の情報は，一致数が 1 個以上の場合 1，2 個以上の場合 2，3 個以上の場合 3，5 個以上の場合 5．2 つ目の情報は，一致数が 1 個以下の場合 1，2 個以下の場合 2，3 個以下の場合 3，5 個以下の場合 5．3 つ目の情報は，一致数が 0 個の場合 0，1 個の場合 1，2 個の場合 2，3 個の場合 3，4，5 個の場合 5，5 個より多い場合は many．この 3 種類の情報を追加することで同単語の出現回数を素性として追加する．

⑥ 段落情報

– 文間箇所以外の段落の分割情報を，段落箇所の場合 1，段落箇所ではない場合 0 とすることで，素性として実験を行った．例えば，素性に直前の 1 文と直後の 1 文の全単語と段落情報を用いて実験を行うとき，以下に直前，直後の 1 文の段落情報の例文を図 3.9 に示す．

(文-１)人の世を作ったものは神でもなければ鬼でもない。(文１)やはり向う三軒両隣りにちらちらするただの人である。(文２)ただの人が作った人の世が住みにくいとて、越す国はあるまい。

図 3.9: 段落情報例文 1

このとき「文-1」が図 3.3 の「？」の直前の文,「文 1」が図 3.3 の「？」の直後の文とする．文-1 の文頭が段落であるので 1，文 1 の文末が段落ではないので 0 とすることで段落情報の素性として扱う．

また，素性に直前の 2 文と直後の 2 文の全単語と段落情報を用いて実験を行うとき，以下に直前，直後の 2 文の段落情報の例文を図 3.10 に示す．

(文-３)住みにくさが高じると、安い所へ引き越したくなる。(文-２)どこへ越しても住みにくいと悟った時、詩が生まれて、画が出来る。
(文-１)人の世を作ったものは神でもなければ鬼でもない。(文１)やはり向う三軒両隣りにちらちらするただの人である。(文２)ただの人が作った人の世が住みにくいとて、越す国はあるまい。(文３)あれば人でなしの国へ行くばかりだ。人でなしの国は人の世よりもなお住みにくかろう。

図 3.10: 段落情報例文 2

このとき，先ほどの文-1 の文頭の段落情報，文 1 の文末の段落情情報に加えて，文-2 の文頭の段落情報，文 2 の文末の段落情報を段落情報の素性として扱う．ここでは，文-2 の文頭は段落箇所ではないので 0，文 2 の文末は段落箇所ではないので 0 である．素性に直前の 3 文と直後の 3 文の全単語と段落情報の場合も同様に文-3 の文頭の段落情報，文 3 の文末の段落情報を加える．

上記の素性をそれぞれ組み合わして複数のパターンで段落分割の推定を行う．また本研究は段落情報のない文章に対して段落の推定を行い段落の付与を行うことを目標であるが，⑥段落情報の追加は実際の文章の，文間箇所前後の段落の正解の分割，非分割の情報を素性としている．文間箇所の前後の正解の段落情報は，本来扱うことのできない情報である．そのため素性の追加での実験では，推定結果の数値を確認するために行うが，考察では⑥段落情報の追加については取り扱わない．

3.3.2 BERT での推定方法

本研究では，3.2.2 節の BERT を用いて，文章の段落推定を行うために実験を行う．本研究で用いる BERT は事前学習済みのモデルであり，新規でラベルなしデータを用意する必要がない．また，本研究で使用した BERT では，3.3.1 節の最大エントロピー法と違って，入力データに空白部分があると実行できないなどの制約がある．そこで実験に用いるデータは，テキスト部分に図 3.1 のデータを，ラベルに整形前の文間箇所の段落，非段落情報を入れる．テキストとラベルでセットとした訓練データ，検証データ，テストデータの 3 種類を新聞記事，小説それぞれで用意し，段落分割の推定を行った．

入力データのテキスト部分は，2 文章をそのまま接続したデータ 1 と 2 文章間を「★」で挟んだデータ 2 の 2 種類のデータを用意し，実験を行った．入力データのラベル部分は，元の文章に段落があった場合「T」，なかった場合「F」とすることで用意した．以下に実験に使用したデータ例を図 3.1 に示す．

表 3.1: BERT で用いた入力データ

データ	ラベル	テキスト
データ 1	F	久間章生元防衛相が自民党総務会長だった０６年１月、福井県敦賀市の知人男性（６４）から１億円を受領していたことが分かった。久間氏と男性は毎日新聞の取材に授受を認め「賃借関係」としているが、無担保で実行された形になっている。
データ 2	F	久間章生元防衛相が自民党総務会長だった０６年１月、福井県敦賀市の知人男性（６４）から１億円を受領していたことが分かった。★久間氏と男性は毎日新聞の取材に授受を認め「賃借関係」としているが、無担保で実行された形になっている。

図 3.1 のデータを用いる．推定方法は，3.2.2 節の Softmax 関数を使うことで，文間箇所での分割，非分割に対する確率を算出する．その値が分割の値の方が大きいとき分割，非分割の値が大きいとき非分割とみなす．算出された値が，テストデータのラベルと一致するかで正解率を出力する．

3.3.3　サポートベクトルマシン法での推定方法

本研究では，サポートベクトルマシン法を用いて，文章の段落推定を行うために実験を行う．サポートベクトルマシン法の推定実験では，最大エントロピー法で使用したデータを使用するため新規でデータセットを用意する必要がない．

3.4 素性分析

段落分割の際にどのような素性が段落の分割，非分割かを判断しているかを調べた．BERT を用いた素性分析の研究は新規であり，本研究では 2 種類の手法を提案する．3.4.1 節では，最大エントロピー法での素性分析について記述している．3.4.2 節では，正規化 α 値について記述している．3.4.3 節では，3 単語連続を用いた BERT での素性分析について記述している．

3.4.1 最大エントロピー法での素性分析

最大エントロピー法を用いて文章の段落推定を行った際，正規化 α 値という数値を得ることができる．得られた正規化 α 値からどのような素性が段落の分割，非分割に影響しているかが分かる．この正規化 α 値を基に，最大エントロピー法の結果では素性分析を行う．

3.4.2 正規化 α 値

正規化 α 値とは，最大エントロピー法で求まる α 値を全分類先での合計が 1 となるように正規化した値である．また，素性 a と分類先 b の対によって定まる値であり，素性 a のみが適用される場合に分類先 b となる確率に相当する．各素性の，分類先ごとに与えられた正規化 α 値が高いほど，その分類先であることを推定するのに重要な素性であることを意味する．

例えば以下の素性の場合，分割位置である分割先に対する正規化 α 値が 0.84，分割位置ではない分類先に対する正規化 α 値が 0.15 であるので,「もっとも」のみで分類を行った場合，分割位置であると推定する確率が 0.84 となることを示す．

- 文 0：もっとも 0.156020(非分割箇所) 0.843980(分割箇所)

3.4.3　3 単語連続を用いた BERT の素性分析

3 単語連続を用いた BERT での素性分析の手法について説明する．最大エントロピー法での素性分析では，1 単語ごとに正規化 α 値を求めて分割，非分割に役立つ素性を調べている．BERT では正規化 α 値を得ることができないため，最大エントロピー法と同じ手法では素性分析をすることができない．

またサポートベクトルマシン法 (SVM) も同様に正規化 α 値を得ることができないので最大エントロピー法と同じ手法を用いることができない．この問題に対して，SVM で分類する際に得た，分離平面との距離を用いた素性分析がある [14]．元のデータで 1 つの文章に対して用いた複数の素性を，テストデータで 1 つの文章に対して用いた複数の素性を 1 つずつに分ける．1 つずつに分けることで，分けたそれぞれの素性に対して SVM で分類を行う．データ例を図 3.11 に示す．

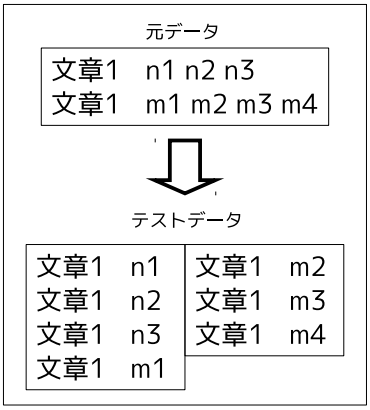

図 3.11: 先行研究 SVM データ例

図 3.11 のテストデータを SVM で分類する際，事例に対する分離平面との距離がそれぞれ算出される．算出された分離平面との距離により，分離平面との距離が大きい素性の事例を有用な素性とする素性分析である．

本研究では，[14] の手法を BERT に用いることで素性分析を行う．BERT を用いて段落分割を行った際のテストデータを 1 単語ずつに分ける．分けた 1 単語に対して BERT を用いることで，分けた 1 単語の分割，非分割に関するそれぞれ値が算出される．算出された値が分割の値が大きい場合，分割に関する素性，非分割の値が大きい場合，非分割に関する素性であると判断する．1 単語ずつに分けた例と算出された数値の例を図 3.12 に示す．

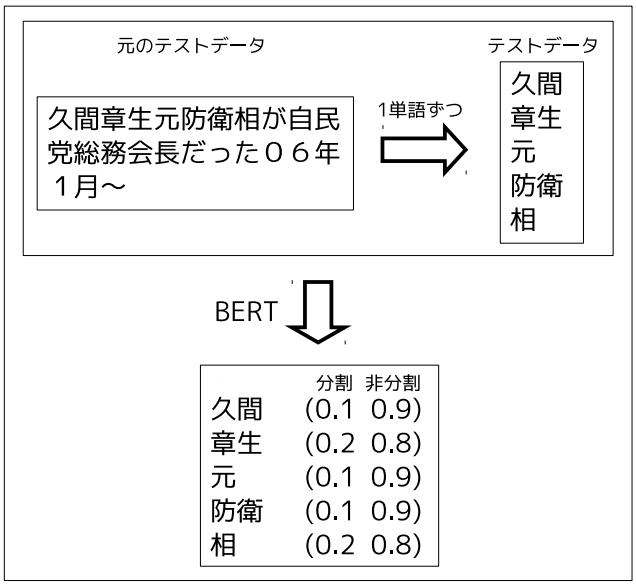

図 3.12: 1 単語ずつ分けた例

図 3.12 のように 1 単語ずつで素性分析を行った．しかし BERT を用いた 1 単語ずつでの素性分析では，分割，非分割に関する素性を得ることができなかった．

そこで，本研究ではテストデータを 3 単語連続にすることで，素性分析を行った．以下に 3 単語連続の例を図 3.13 に示す．

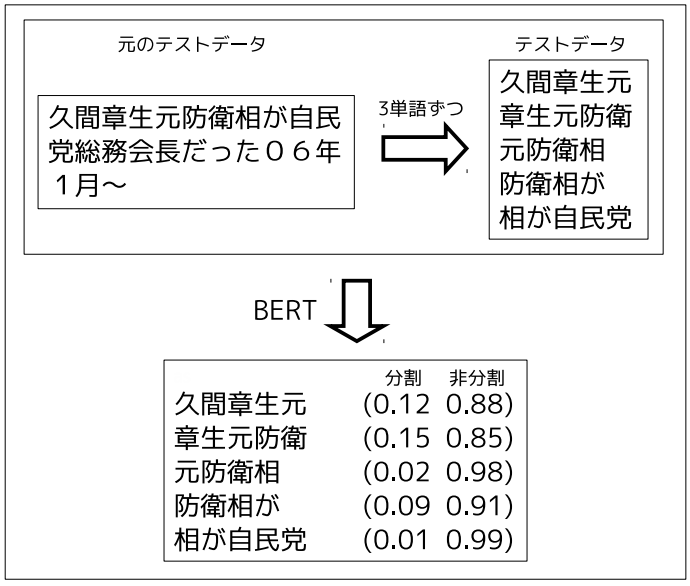

図 3.13: 3 単語連続例

図 3.13 のように，元のテストデータを 3 単語連続に分ける．テストデータを 3 単語連続に整形し素性分析を行うことで，分割，非分割に関する値がそれぞれの 3 単語連続に対して算出される．算出された値が分割の値が大きい場合，分割に関する素性，非分割の値が大きい場合，非分割に関する素性であると判断する．図 3.13 の「久間章生元」は，分割の値が 0.12 に対して非分割の値が 0.88 と非分割の方が大きい.「久間章生元」は非分割に関する素性であると判断できる．

「分割」,「非分割」の数値の上位の素性を最大エントロピー法から得た素性と比較することで，3 単語連続での素性分析が有用な手法であるかどうかを調べる．

第4章　実験結果

本章では，本研究の段落分割の実験で行った実験方法と実験結果を記述する．4.1 節では，最大エントロピー法での実験について記述している．4.2 節では，BERT での実験について記述している．4.3 節では，サポートベクトルマシン法での実験について記述している．4.4 節では，新聞記事を用いた推定精度の比較と考察について記述している．4.5 節では，小説を用いた推定精度の比較と考察について記述している．

4.1　実験 [最大エントロピー法]

4.1.1 節では，最大エントロピー法を用いた実験結果について記述している．4.1.2 節では，学習データ，テストデータともに新聞記事を使用した実験結果について記述している．4.1.3 節では，学習データ，テストデータともに小説を使用した実験結果について記述している．

4.1.1　実験方法

最大エントロピー法の推定実験では，3.3.1 節の素性を用い，それぞれを組み合わせて実験を行う．3.3.1 節で説明した素性を表 4.1 に示す．

表 4.1: 素性

①	文と文の間の箇所 (以下「文間箇所」) の直前の 1 文と直後の 1 文にある全単語
②	文間箇所の直前の 2 文と直後の 2 文にある全単語
③	文間箇所の直前の 3 文と直後の 3 文にある全単語
④	文間箇所の直前 1 文と直後の 1 文の文頭の単語
⑤	文間箇所の直前, 直後の 2 文中の同単語の出現数
⑥	段落情報

3.3.1 節でも記述したが，⑥段落情報の追加について，実際のデータでの段落情報の有無を素性としているため容易に推定精度が上がることが分かる．実験は推定結果の数値を確認するために行うが，考察では⑥段落情報の追加については取り扱わない．

入力文は，毎日新聞と小説を用いて，別々に実験を行う．実験のベースライン手法は，全ての文間箇所が分割位置ではない (段落ではない) と判断する方法とし，提案手法とベースライン手法の性能の比較を行う．また，文間箇所の数を文間数と表す．

4.1.2 実験結果 (新聞記事)

新聞記事を用いた実験では，学習データに 2007 年の毎日新聞 (文間数 2,500，5,000，10,000)，テストデータに 2008 年の毎日新聞 (文間数 10,000) のデータを用いて実験を行った．テストデータの各段落における文数を表したヒストグラムを図 4.1 に，テストデータの分類先の頻度を表 4.2 に，結果を表 4.3 に示す．

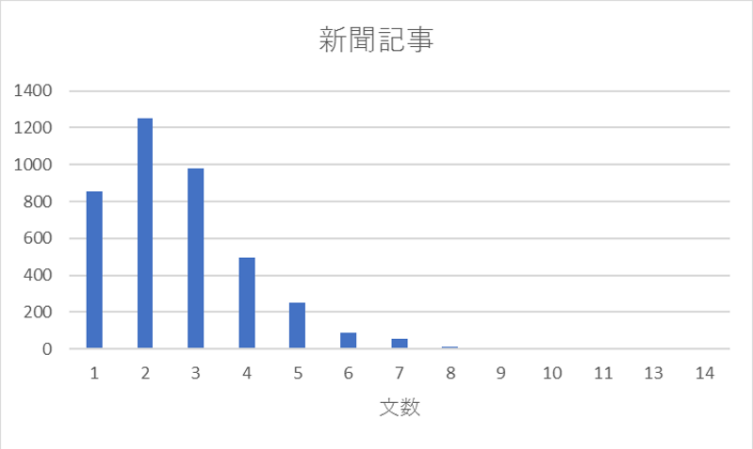

図 4.1: 新聞記事のヒストグラム

表 4.2: 分類先の頻度

分割位置	3,257
分割位置ではない箇所	6,743
合計	10,000

表 4.3: 最大エントロピー法 (新聞記事) の実験結果

	正解率		
素性	2,500	5,000	10,000
①	0.6866	0.6900	0.6905
①+④	0.6894	0.6913	0.6919
①+⑤	0.6907	0.6928	0.6935
①+⑥	0.7384	0.7428	0.7430
②	0.6857	0.6930	0.6959
②+⑥	0.7486	0.7594	0.7644
③	0.6813	0.6898	0.6956
③+⑥	0.7484	0.7556	0.7628
ベースライン	0.6743		

新聞記事で段落分割を行った結果，段落分割の推定精度はベースラインの正解率が 0.6743 に対して，⑥に関する素性以外で考えると，学習データが文間数 10,000 の②の正解率 0.6959 が最も高い数値であった．表 4.3 の結果をもとに,「ベースラインは正解であるが提案手法は不正解であった分割箇所」の数と,「ベースラインは正解であるが提案手法は不正解であった分割箇所」の数と「ベースラインは不正解であるが提案手法は正解であった分割箇所」の数の合計数を用い，二項分布に基づく有意水準 0.05 の符号検定 (片側検定) を行った．有意差検定で得た p 値を表 4.4 に示す.

表 4.4: ベースラインとの有意差検定 (MEM の新聞新聞)

素性	p 値
①	0.000322
①+④	0.000091
①+⑤	0.000022
①+⑥	3.005755×10^{-45}
②	0.000002
②+⑥	2.348028×10^{-81}
③	0.000002
③+⑥	7.329237×10^{-78}

表 4.4 より，それぞれの素性は全てベースラインの正解率と有意差があった.

また文間数 10,000 の①の正解率 0.6905 と，表 4.3 より①のみを除くそれぞれの文間

数 10,000 の正解率に対して,「①は正解であるが提案手法は不正解であった分割箇所」の数と,「①は正解であるが提案手法は不正解であった分割箇所」の数と「①は不正解であるが提案手法は正解であった分割箇所」の合計数を用い，二項分布に基づく有意水準 0.05 の符号検定 (片側検定) を行った．有意差検定で得た p 値を表 4.5 に示す.

表 4.5: ①との有意差検定 (MEM の新聞新聞)

素性	p 値
①+④	0.263935
①+⑤	0.059962
①+⑥	2.392929×10^{-43}
②	0.085289
②+⑥	1.214272×10^{-63}
③	0.114856
③+⑥	7.993192×10^{-56}

表 4.5 より，⑥の段落情報追加を含む素性以外有意差がなかった．しかし，①+⑤の p 値は有意差はないが，0.05 にかなり近い.

また表 4.3 より，ベースラインと比較すると最も数値の差が大きい②+⑥は，ベースラインより 0.0901 大きい．しかし，他に全単語と以外で追加した④と⑤は追加前と比較すると推定精度は向上しているが，数値はほとんど変わらない.

表 4.3 より，他の手法と比べて⑥の段落情報の追加での正解率がかなり高い理由は,直前直後の正確な段落有無が分かっているからである．図 4.1 より新聞記事の場合 2, 3 文に 1 つ段落があることが分かる．もし文間箇所の直前の文頭に段落がある場合，文間箇所は段落ではない可能性が高いと容易に考えられる.

4.1.3 実験結果 (小説)

小説を用いた実験では，実験の処理はコンピュータ上で扱うため，使用する小説は電子図書館青空文庫のテキストデータを用いた．学習データに夏目漱石の「草枕」(文間数 2,204),「坊ちゃん」(文間数 2,175),「吾輩は猫である」(文間数 4,928) の 3 作品全て (文間数 9,307) を，テストデータに夏目漱石の「それから」(文対数 4,643) を用いて行った．実験には，新聞記事で用いた素性の組み合わせを使用し実験を行った．テストデータの各段落における文数を表したヒストグラムを図 4.2 に，テストデータの分類先の頻度を表 4.6 に，結果を表 4.7 に示す．

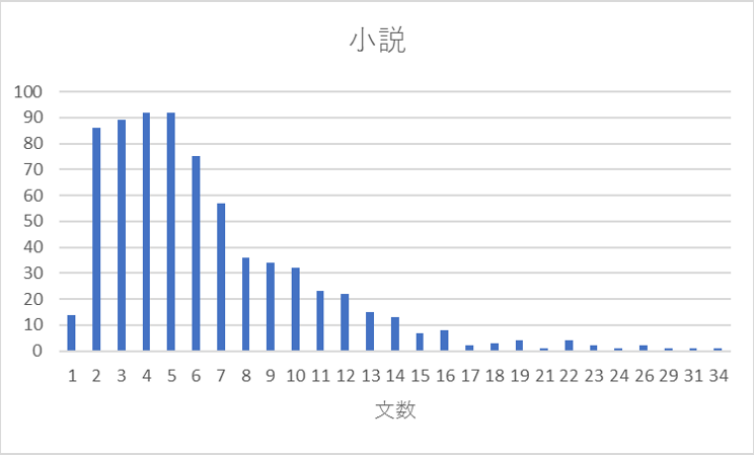

図 4.2: 小説のヒストグラム

表 4.6: 分類先の頻度

分割位置	700
分割位置ではない箇所	3,943
合計	4,643

表 4.7: 最大エントロピー法 (小説) の実験結果

素性	正解率
①	0.8645
①+④	0.8650
①+⑤	0.8658
①+⑥	0.8658
②	0.8652
②+⑥	0.8656
③	0.8639
③+⑥	0.8637
ベースライン	0.8492

小説で段落分割を行った結果，段落分割の推定精度はベースラインの正解率が 0.8492 に対して，⑥に関する正解率以外で，学習データが①+⑤の場合の正解率は 0.8658 で最も高い数値であった．表 4.7 の結果をもとに,「ベースラインは正解であるが提案手法は不正解であった分割箇所」の数と,「ベースラインは正解であるが提案手法は不正解であった分割箇所」の数と「ベースラインは不正解であるが提案手法は正解であった分割箇所」の数の合計数を用い，二項分布に基づく有意水準 0.05 の符号検定 (片側検定) で有意差検定を行った．有意差検定で得た p 値を表 4.8 に示す.

表 4.8: ベースラインとの有意差検定 (MEM の小説)

素性	p 値
①	1.112186×10^{-7}
①+④	3.471233×10^{-8}
①+⑤	1.042276×10^{-8}
①+⑥	3.005755×10^{-45}
②	2.095756×10^{-9}
②+⑥	4.672362×10^{-9}
③	1.351334×10^{-7}
③+⑥	0.000001

表 4.8 より，それぞれの素性は全てベースラインの正解率と有意差があった.

また，①の正解率 0.8645 と，表 4.7 より①のみを除くそれぞれの正解率に対して，有意差を検定した.「①は正解であるが提案手法は不正解であった分割箇所」の数と,「①

は正解であるが提案手法は不正解であった分割箇所」の数と「①は不正解であるが提案手法は正解であった分割箇所」の合計数を用い，二項分布に基づく有意水準 0.05 の符号検定 (片側検定) を行った．有意差検定で得た p 値を表 4.9 に示す．

表 4.9: ①との有意差検定 (MEM の小説)

素性	p 値
①+④	0.444942
①+⑤	0.153728
①+⑥	0.230696
②	0.422685
②+⑥	0.354661
③	0.637573
③+⑥	0.659391

表 4.9 より，全て有意差はなかった．したがってどの素性も，最大エントロピー法を用いて小説の段落分割の推定を行う際に，推定精度の大きな向上に有用な素性の追加でないことが分かる．

また，小説での推定で最も数値が大きい①+⑤，①+⑥の正解率 0.8658 は，ベースラインの正解率と比較すると 0.0166 しか変わらない．ベースラインの数値が高いため，素性を追加しても推定精度の数値の向上が難しいのではないかと考えられる．

4.2 実験 [BERT]

4.2.1 節では，BERT を用いた実験結果について記述している．4.2.2 節では，学習データ，テストデータともに新聞記事を使用した実験結果について記述している．4.2.3 節では，学習データ，テストデータともに小説を使用した実験結果について記述している．

4.2.1 実験方法

BERT を用いた実験では，3.3.2 節の 2 文間に「★」がないデータ 1 と「★」を入れたデータ 2 の 2 種類のデータを用いて実験を行う．また BERT は他の手法と異なり，訓練データ，検証データ，テストデータの 3 つのデータを用意する必要がある．そこで本研究では，訓練データと検証データは，最大エントロピー法の実験で使用した学習データを 3:1 で分けることで訓練データと検証データを用意した．またテストデータは，最大エントロピー法の実験で使用したテストデータと同じものを使用した．

BERT での実験のベースライン手法は，全ての文間箇所が分割位置ではない (段落ではない) と判断する方法とし，提案手法とベースライン手法の性能の比較を行う．

4.2.2 実験結果 (新聞記事)

新聞記事の推定実験で使用した訓練データ，検証データ，テストデータの内訳を表 4.10 に示す．BERT での学習の際，入力した文章は分散表現に変換されるため，結果が多少変化する．そこで本実験は，同じデータを用いて 10 回試行した．データ 1，データ 2 ともに 10 回試行した平均値を表 4.11 に示す．

表 4.10: 実験データ (新聞記事) の内訳

	訓練データ	検証データ	テストデータ
新聞記事	7,500	2,500	10,000

表 4.11: BERT(新聞記事) の実験結果

	正解率
データ 1	0.7547
データ 2	0.7564
ベースライン	0.6743

BERT を使用して，新聞記事に対して段落分割を行った結果，段落分割の推定精度はベースラインの正解率が 0.6743 に対して,「★」なしのデータ 1 の場合の正解率は 0.7547,「★」ありのデータ 2 の場合の正解率は 0.7564 であった．また，データ 1 とデータ 2 の正解率を比較すると，差はほとんどないが 0.0017 だけデータ 2 の方が正解率が大きいことが分かる．この結果をもとに有意差を検定した.「ベースラインは正解であるが提案手法は不正解であった分割箇所」の数と,「ベースラインは正解であるが提案手法は不正解であった分割箇所」の数と「ベースラインは不正解であるが提案手法は正解であった分割箇所」の数の合計数を用い，二項分布に基づく有意水準 0.05 の符号検定 (片側検定) を行った．表 4.11 の「★」なしのデータ 1 とベースラインとの有意差検定で得た p 値を表 4.12 に示す．表 4.11 の「★」ありのデータ 2 とベースラインとの有意差検定で得た p 値を表 4.13 に示す．表 4.12，表 4.13 は，表 4.11 で 10 回試行したデータ 1，データ 2 の 10 回分それぞれの結果と，ベースラインとの有意差検定を行った．

表 4.12: BERT(データ 1) とベースラインとの有意差 (新聞記事)

データ 1	p 値
0.7599	1.774087×10^{-61}
0.7590	4.196571×10^{-60}
0.7579	2.867386×10^{-57}
0.7555	5.933988×10^{-55}
0.7544	1.888493×10^{-53}
0.7515	5.523886×10^{-54}
0.7507	1.519219×10^{-47}
0.7505	1.244062×10^{-53}
0.7524	7.004474×10^{-50}
0.7556	1.184593×10^{-52}

表 4.13: BERT(データ 2) とベースラインとの有意差 (新聞記事)

データ 2	p 値
0.7600	1.723536×10^{-65}
0.7597	1.334621×10^{-61}
0.7592	7.462583×10^{-68}
0.7581	1.273909×10^{-58}
0.7569	1.397631×10^{-55}
0.7561	7.170220×10^{-56}
0.7547	5.730105×10^{-59}
0.7546	6.592184×10^{-57}
0.7527	1.228783×10^{-52}
0.7520	1.593564×10^{-49}

表 4.12，表 4.13 より，新聞記事に対して BERT での段落分割はデータ 1，データ 2 とも全てベースラインの正解率と有意差があった．

また，データ 1 とデータ 2 の正解率に対して有意差を検定した．「データ 1 は正解であるがデータ 2 は不正解であった分割箇所」の数と，「データ 1 は正解であるがデータ 2 は不正解であった分割箇所」の数と「データ 1 は不正解であるがデータ 2 は正解であった分割箇所」の数の合計数を用い，二項分布に基づく有意水準 0.05 の符号検定 (片側検定) を行った．データ 2 の 10 個の値の中上位 5 個での有意差検定で得た p 値を表 4.14 に、下位 5 個での有意差検定で得た p 値を表 4.15 に示す．データ 2 がデータ 1 の数値を上回っていない箇所には「×」と入力している．

表 4.14: BERT(データ 2 の上位 5 番目) とデータ 1 の有意差 (新聞記事)

	データ 2 の上位 5 番目				
データ 1	0.7600	0.7597	0.7592	0.7581	0.7569
0.7599	0.000598	×	×	×	×
0.7590	0.386196	0.428322	0.488438	×	×
0.7579	0.269232	0.305587	0.363037	0.487952	×
0.7556	0.099006	0.120407	0.157164	0.234149	0.357320
0.7555	0.083577	0.111694	0.140272	0.224256	0.346353
0.7544	0.002537	0.004802	0.007115	0.018442	0.032794
0.7524	0.011405	0.014984	0.021091	0.043649	0.078310
0.7515	0.005587	0.007688	0.010796	0.025924	0.054838
0.7507	0.042346	0.059683	0.087594	0.139622	0.229352
0.7505	0.001810	0.002755	0.003404	0.011468	0.027367

表 4.15: BERT(データ 2 の下位 5 番目) とデータ 1 の有意差 (新聞記事)

	データ 2 の下位 5 番目				
データ 1	0.7561	0.7547	0.7546	0.7527	0.7520
0.7599	×	×	×	×	×
0.7590	×	×	×	×	×
0.7579	×	×	×	×	×
0.7556	0.452748	×	×	×	×
0.7555	0.442091	×	×	×	×
0.7544	0.054838	0.129716	0.128826	×	×
0.7524	0.136546	0.254051	0.178072	0.371315	0.873844
0.7515	0.091097	0.169825	0.178072	0.371315	0.873844
0.7507	0.311158	0.477180	0.487886	0.701020	0.766241
0.7505	0.049517	0.105444	0.103732	0.263131	0.344104

表 4.14，表 4.15 より 68 個間の有意差検定を行い，そのうち 21 個間で有意差があった．しかし，全体で見ると有意差はないと思われるが，データ 2 は 100 個中 68 個データ 1 を上回っている．新聞記事に対して，BERT で段落分割の推定実験を行う際,「★」ありのデータ 2 を用いて実験を行うのが良いのではないかと考えられる．

4.2.3 実験結果 (小説)

小説の推定実験で使用した訓練データ，検証データ，テストデータの内訳を表 4.16 に示す．また新聞記事での実験と同様に，同じデータを用いて 10 回試行した．データ 1，データ 2 ともに 10 回試行した平均値を結果を表 4.17 に示す.

表 4.16: 実験データ (小説) の内訳

	訓練データ	検証データ	テストデータ
小説	7,000	2,307	4,643

表 4.17: BERT(小説) の実験結果

	正解率
データ 1	0.8720
データ 2	0.8772
ベースライン	0.8492

BERT を使用して，小説に対して段落分割を行った結果，段落分割の推定精度はベースラインの正解率が 0.8492 に対して,「★」なしのデータ 1 の場合の正解率は 0.8720,「★」ありのデータ 2 の場合の正解率は 0.8778 であった．また，データ 1 とデータ 2 の正解率を比較すると，差はほとんどないが 0.0052 だけデータ 2 の方が正解率が大きいことが分かる．この結果をもとに有意差を検定した.「ベースラインは正解であるが提案手法は不正解であった分割箇所」の数と,「ベースラインは正解であるが提案手法は不正解であった分割箇所」の数と「ベースラインは不正解であるが提案手法は正解であった分割箇所」の数の合計数を用い，二項分布に基づく有意水準 0.05 の符号検定 (片側検定) を行った．表 4.17 の「★」なしのデータ 1 とベースラインとの有意差検定で得た p 値を表 4.18 に示す．表 4.17 の「★」ありのデータ 2 とベースラインとの有意差検定で得た p 値を表 4.19 に示す．表 4.18，表 4.19 は，表 4.17 で 10 回試行したデータ 1，データ 2 の 10 回分それぞれの結果と，ベースラインとの有意差検定を行った.

表 4.18: BERT(データ 1) とベースラインとの有意差 (小説)

データ 2	p 値
0.8770	1.536759×10^{-13}
0.8766	1.535316×10^{-16}
0.8759	1.993093×10^{-11}
0.8753	3.203538×10^{-13}
0.8716	9.076090×10^{-9}
0.8710	2.567372×10^{-11}
0.8703	5.684129×10^{-9}
0.8680	6.896507×10^{-8}
0.8680	0.000001
0.8660	0.000018

表 4.19: BERT(データ 2) とベースラインとの有意差 (小説)

データ 2	p 値
0.8828	1.074424×10^{-26}
0.8815	1.988576×10^{-15}
0.8808	3.203390×10^{-18}
0.8802	1.963167×10^{-18}
0.8787	9.656685×10^{-14}
0.8783	8.602681×10^{-15}
0.8774	1.876691×10^{-13}
0.8748	3.167651×10^{-11}
0.8718	7.126215×10^{-7}
0.8658	5.555714×10^{-5}

表 4.18，表 4.19 より，小説に対して BERT での段落分割はデータ 1，データ 2 とも全てベースラインの正解率と有意差があった．また，データ 1 とデータ 2 の正解率に対して有意差を検定した.「データ 1 は正解であるがデータ 2 は不正解であった分割箇所」の数と,「データ 1 は正解であるがデータ 2 は不正解であった分割箇所」の数と「データ 1 は不正解であるがデータ 2 は正解であった分割箇所」の数の合計数を用い，二項分布に基づく有意水準 0.05 の符号検定 (片側検定) を行った．データ 2 の 10 個の値の

中上位 5 個での有意差検定で得た p 値を表 4.20 に、下位 5 個での有意差検定で得た p 値を表 4.21 に示す．データ 2 がデータ 1 の数値を上回っていない箇所には「×」と入力している．

表 4.20: BERT(データ 2 の上位 5 番目) とデータ 1 との有意差 (小説)

	データ 2 の上位 5 番目				
データ 1	0.8828	0.8815	0.8808	0.8802	0.8787
0.8770	0.021317	0.076998	0.092091	0.122442	0.303943
0.8766	0.008618	0.069364	0.071227	0.110506	0.265361
0.8759	0.011754	0.039902	0.049916	0.073598	0.195104
0.8753	0.003776	0.026255	0.025293	0.039396	0.127706
0.8716	0.000107	0.001225	0.001407	0.002269	0.011881
0.8710	0.000010	0.001007	0.000188	0.000968	0.007516
0.8703	0.000003	0.000246	0.000092	0.000260	0.003749
0.8680	4.733506×10^{-8}	0.000037	0.000002	0.000014	0.000322
0.8680	0.000001	0.000028	0.000016	0.000035	0.000535
0.8660	2.430588×10^{-8}	0.000001	1.340262×10^{-7}	0.000001	0.000019

表 4.21: BERT(データ 2 の上位 5 番目) とデータ 1 との有意差 (小説)

	データ 2 の下位 5 番目				
データ 1	0.8783	0.8774	0.8748	0.8718	0.8658
0.8770	0.346381	0.468704	×	×	×
0.8766	0.300984	0.411083	×	×	×
0.8759	0.223589	0.320281	×	×	×
0.8753	0.148967	0.243783	×	×	×
0.8716	0.012742	0.027156	0.165266	0.500000	×
0.8710	0.008218	0.015651	0.123666	0.431264	×
0.8703	0.004327	0.008263	0.089191	0.363006	×
0.8680	0.000267	0.001026	0.020797	0.166361	×
0.8680	0.000737	0.000869	0.018617	0.146791	×
0.8660	0.000017	0.000071	0.003369	0.053004	×

表 4.20，表 4.21 より 82 個間の有意差検定を行い，そのうち 54 個間で有意差があった．新聞記事での有意差検定より有意差のある個数が多いが，全体の 100 個に対して

考えると有意差はないと思われる．しかし，データ 2 は 82 個がデータ 1 を上回っており，新聞記事と同様に小説に対して，BERT で段落分割の推定実験を行う際,「★」ありのデータ 2 を用いて実験を行うのが良いのではないかと考えられる．

4.3 実験 [サポートベクトルマシン法]

4.3.1 節では，サポートベクトルマシン法を用いた実験結果について記述している．4.3.2 節では，学習データ，テストデータともに新聞記事を使用した実験結果について記述している．4.3.3 節では，学習データ，テストデータともに小説を使用した実験結果について記述している．

4.3.1 実験方法

サポートベクトルマシン法の推定実験では，4.1 節の最大エントロピー法の推定実験で使用した素性の組み合わせで実験を行う．また実験のベースライン手法は，全ての文間箇所が分割位置ではない (段落ではない) と判断する方法とし，提案手法とベースライン手法の性能の比較を行う．

4.3.2 実験結果 (新聞記事)

新聞記事を使用した推定実験の結果を表 4.22 に示す．

表 4.22: SVM(新聞記事) の実験結果

	正解率		
素性	2,500	5,000	10,000
①	0.6683	0.6642	0.6651
①+④	0.6737	0.6693	0.6684
①+⑤	0.6703	0.6645	0.6679
①+⑥	0.7216	0.7185	0.7134
②	0.6716	0.6763	0.6737
②+⑥	0.7391	0.7457	0.7424
③	0.6702	0.6777	0.6788
③+⑥	0.7411	0.7447	0.7474
ベースライン	0.6743		

サポートベクトルマシン法を用いて新聞記事で段落分割を行った結果，段落分割の推定精度はベースラインの正解率が 0.6743 に対して，⑥に関する正解率以外で，学習

データが文間数 10,000 の③の正解率 0.6788 が最も高い数値であり，唯一ベースラインの正解率より高い．

また表 4.22 の結果をもとに，ベースラインとベースラインの数値を上回る③と⑥に関する素性の文間数 10,000 の正解率との有意差を検定した．「ベースラインは正解であるが提案手法は不正解であった分割箇所」の数と，「ベースラインは正解であるが提案手法は不正解であった分割箇所」の数と「ベースラインは不正解であるが提案手法は正解であった分割箇所」の数の合計数を用い，二項分布に基づく有意水準 0.05 の符号検定 (片側検定) を行った．有意差検定で得た p 値を表 4.23 に示す．

表 4.23: ベースラインとの有意差検定 (SVM の新聞記事)

素性	p 値
①+⑥	3.546496×10^{-15}
②+⑥	5.046621×10^{-46}
③	0.177926
③+⑥	1.766036×10^{-53}

表 4.23 より，6 に関する素性は全てベースラインと有意差があったが，③の正解率はベースラインに対して有意差はなかった．

またベースラインを下回る素性の文間数 10,000 の正解率とベースラインの有意差を検定した．「提案手法は正解であるがベースラインは不正解であった分割箇所」の数と，「提案手法は正解であるがベースラインは不正解であった分割箇所」の数と「提案手法は不正解であるがベースラインは正解であった分割箇所」の数の合計数を用い，二項分布に基づく有意水準 0.05 の符号検定 (片側検定) を行った．有意差検定で得た p 値を表 4.24 に示す．

表 4.24: ベースラインを下回る素性の有意差検定 (SVM の新聞記事)

素性	p 値
①	0.035423
①+④	0.122406
①+⑤	0.104193
②	0.459117

表 4.24 より，ベースラインは①の正解率に対して有意差があり，①以外に対しては有意差がなかった．

また，文間数 10,000 の①の正解率 0.6651 と表 4.22 より①のみを除く文間数 10,000 の正解率に対して，有意差を検定した.「①は正解であるが提案手法は不正解であった分割箇所」の数と,「①は正解であるが提案手法は不正解であった分割箇所」の数と「①は不正解であるが提案手法は正解であった分割箇所」の合計数を用い，二項分布に基づく有意水準 0.05 の符号検定 (片側検定) を行った．有意差検定で得た p 値を表 4.25 に示す.

表 4.25: ①との有意差検定 (SVM の新聞記事)

素性	p 値
①+④	0.103111
①+⑤	0.086291
①+⑥	1.446824×10^{-37}
②	0.025572
②+⑥	9.801338×10^{-60}
③	0.001844
③+⑥	5.159875×10^{-62}

表 4.25 より，①+④と①+⑤は①に対して有意差がなく，他は有意差があった.

⑥に関する素性ベースラインとの有意差はなかったが，サポートベクトルマシン法で文間箇所の直前直後 2 文，直前直後 3 文の全単語を追加することは，①と比べて正解率は上がっており，推定精度の向上に役立っていることが分かる.

4.3.3 実験結果 (小説)

小説を使用した推定実験の結果を表 4.26 に示す.

表 4.26: SVM(小説) の実験結果

素性	正解率
①	0.8533
①+④	0.8566
①+⑤	0.8542
①+⑥	0.8570
②	0.8652
②+⑥	0.8643
③	0.8568
③+⑥	0.8579
ベースライン	0.8492

サポートベクトルマシン法を用いて新聞記事で段落分割を行った結果，段落分割の推定精度はベースラインの正解率が 0.8492 に対して，学習データが②の場合の正解率は 0.8652 で最も高い数値であった.

また表 4.26 の結果をもとに,「ベースラインは正解であるが提案手法は不正解であった分割箇所」の数と,「ベースラインは正解であるが提案手法は不正解であった分割箇所」の数と「ベースラインは不正解であるが提案手法は正解であった分割箇所」の数の合計数を用い，二項分布に基づく有意水準 0.05 の符号検定 (片側検定) を行った．有意差検定で得た p 値を表 4.27 に示す.

表 4.27: ベースラインとの有意差検定 (SVM の小説)

素性	p 値
①	0.148940
①+④	0.023799
①+⑤	0.100837
①+⑥	0.017474
②	1.035975×10^{-7}
②+⑥	0.000002
③	0.014153
③+⑥	0.007510

表 4.27 より，①と①+⑤の正解率はベースラインに対して有意差がなく，他は有意差があった．

また，①の正解率と表 4.26 より①のみを除くそれぞれの正解率に対して，有意差を検定した．「①は正解であるが提案手法は不正解であった分割箇所」の数と，「①は正解であるが提案手法は不正解であった分割箇所」の数と「①は不正解であるが提案手法は正解であった分割箇所」の合計数を用い，二項分布に基づく有意水準 0.05 の符号検定 (片側検定) を行った．有意差検定で得た p 値を表 4.28 に示す．

表 4.28: ①との有意差検定 (SVM の小説)

素性	p 値
①+④	0.110558
①+⑤	0.365504
①+⑥	0.062502
②	0.000131
②+⑥	0.000676
③	0.184170
③+⑥	0.128359

表 4.28 より，②と②+⑥は①に対して有意差があった．新聞記事と違って他の段落情報の追加が①に対して有意差がないことから，小説に対してサポートベクトルマシン法を用いて段落分割をする際，文間箇所の直前直後 2 文の全単語を追加することは，推定精度の向上に役立っていることが分かる．

新聞記事に対しての段落分割で②のみの正解率はベースラインを下回っているが，新聞記事，小説に共通して，文間箇所の直前直後 2 文の全単語を追加することは，推定精度の向上に役立つと考えられる．

4.4 新聞記事の推定精度の比較と考察

BERT での推定実験で文間箇所の直前，直後の 1 文を用いているため，比較する最大エントロピー法とサポートベクトルマシン法も同様に直前，直後 1 文に関する素性の推定結果 (①，①+④，①+⑤) を用いて比較する．新聞記事に対して「★」を含むデータを使用した BERT の推定精度を表 4.29 に示す．新聞記事に対して，最大エントロピー法 (MEM) とサポートベクトルマシン法 (SVM) の文間数 10,000 の推定精度を表 4.30 に示す．

表 4.29: BERT の推定精度

	BERT
正解率	0.7564
ベースライン	0.6743

表 4.30: MEM と SVM の推定精度

	MEM	SVM
①	0.6905	0.6651
①+④	0.6919	0.6684
①+⑤	0.6935	0.6679
①+⑥	0.7430	0.7134
②	0.6959	0.6737
②+⑥	0.7644	0.7424
③	0.6956	0.6788
③+⑥	0.7628	0.7474
ベースライン	0.6743	

3 つの手法の推定精度を比較すると，明らかに BERT が優れていることが分かる．他の 2 つの手法では，+④，+⑤として素性を追加することで少し精度の向上は見られるが，BERT に比べれば低い．サポートベクトルマシン法の推定精度に至っては，表 4.30 の素性ではベースライン手法の正解率を全て下回っている．

4.2.2 節の 10 回の BERT の実験，表 4.30 の結果をもとに，有意差を検定した.「MEM(SVM) は正解であるが BERT は不正解であった分割箇所」の数と,「MEM(SVM) は正解であ

るが BERT は不正解であった分割箇所」の数と「MEM(SVM) は不正解であるが BERT は正解であった分割箇所」の数の合計数を用い，二項分布に基づく有意水準 0.05 の符号検定 (片側検定) を行った．BERT と MEM との有意差検定から得た p 値を表 4.31 に示す．BERT と SVM との有意差検定から得た p 値を表 4.32 に示す．BERT の正解率は高い順から並べた．

表 4.31: BERT と MEM の有意差 (新聞記事)

BERT	MEM			
正解率	①	①+④	①+⑤	①+⑥
0.7600	9.784885×10^{-42}	8.086561×10^{-44}	1.869735×10^{-42}	0.000333
0.7597	5.219217×10^{-43}	1.556236×10^{-41}	5.695576×10^{-40}	0.000441
0.7592	3.313563×10^{-44}	2.218098×10^{-42}	5.642767×10^{-41}	0.000530
0.7581	5.314831×10^{-42}	2.610601×10^{-40}	7.587332×10^{-39}	0.001286
0.7569	8.748531×10^{-39}	2.138146×10^{-37}	3.687393×10^{-36}	0.003028
0.7561	3.576113×10^{-38}	8.624363×10^{-37}	3.058490×10^{-35}	0.004861
0.7547	1.909356×10^{-38}	5.718660×10^{-37}	1.715539×10^{-35}	0.009505
0.7546	3.876496×10^{-38}	2.467767×10^{-36}	2.595151×10^{-35}	0.010689
0.7527	4.505665×10^{-35}	1.236687×10^{-33}	1.864842×10^{-32}	0.027248
0.7520	1.234700×10^{-33}	3.584358×10^{-32}	3.004267×10^{-31}	0.039394

表 4.32: BERT と SVM の有意差 (新聞記事)

BERT	SVM			
正解率	①	①+④	①+⑤	①+⑥
0.7600	9.517977×10^{-71}	3.692238×10^{-68}	1.814817×10^{-69}	1.114632×10^{-19}
0.7597	9.028848×10^{-71}	3.868065×10^{-66}	6.405102×10^{-68}	3.402280×10^{-19}
0.7592	2.496741×10^{-72}	7.342014×10^{-68}	3.539634×10^{-69}	2.558431×10^{-19}
0.7581	3.939067×10^{-75}	3.453684×10^{-65}	7.003671×10^{-67}	3.062743×10^{-18}
0.7569	1.087028×10^{-65}	6.635255×10^{-62}	8.661066×10^{-63}	5.878538×10^{-17}
0.7561	2.643251×10^{-64}	1.466825×10^{-60}	1.630576×10^{-61}	2.056838×10^{-16}
0.7547	3.011686×10^{-64}	1.131126×10^{-60}	4.392182×10^{-61}	8.358095×10^{-16}
0.7546	5.397737×10^{-65}	9.946261×10^{-61}	2.739091×10^{-62}	1.169556×10^{-15}
0.7527	1.075303×10^{-60}	5.596489×10^{-57}	6.056939×10^{-58}	2.636923×10^{-14}
0.7520	1.019836×10^{-58}	4.033717×10^{-55}	4.710915×10^{-56}	1.380748×10^{-13}

表 4.31，表 4.32 より，10 回試行した BERT の正解率は MEM，SVM の正解率に対して全て有意差があった．また表 4.30 より，MEM と SVM で同じ素性を用いた時の正解率の有意差を検定した.「SVM は正解であるが MEM は不正解であった分割箇所」の数と,「SVM は正解であるが MEM は不正解であった分割箇所」の数と「SVM は不正解であるが MEM は正解であった分割箇所」の数の合計数を用い，二項分布に基づく有意水準 0.05 の符号検定 (片側検定) を行った．有意差検定から得た p 値を表 4.33 に示す.

表 4.33: MEM と SVM の有意差 (新聞記事)

SVM	MEM
①	4.079042×10^{-16}
①+④	2.100663×10^{-15}
①+⑤	1.478956×10^{-16}
①+⑥	2.847059×10^{-26}
②	8.866361×10^{-16}
②+⑥	7.655651×10^{-21}
③	9.442747×10^{-12}
③+⑥	8.201363×10^{-15}

表 4.33 より，MEM は SVM に対して全ての素性で有意差があった.

4.5 小説の推定精度の比較と考察

小説に対して「★」を含むデータを使用したBERTの推定精度を表4.34に示す．小説に対して段落情報以外の直前，直後の1単語の素性 (①，①+④，①+⑤) を使用したときの最大エントロピー法 (MEM) とサポートベクトルマシン法 (SVM) の推定精度を表4.35に示す．

表 4.34: BERTの推定精度

	BERT
正解率	0.8772
ベースライン	0.8492

表 4.35: MEMとSVMの推定精度

	MEM	SVM
①	0.8645	0.8533
①+④	0.8650	0.8566
①+⑤	0.8658	0.8542
①+⑥	0.8658	0.8570
②	0.8652	0.8652
②+⑥	0.8656	0.8643
③	0.8639	0.8568
③+⑥	0.8637	0.8579
ベースライン	0.8492	

3つの手法の推定精度を比較すると，明らかにBERTの推定精度が一番高いことが分かるが，新聞記事ほどの数値の差はない．サポートベクトルマシン法は新聞記事での正解率とは異なり，ベースラインを上回っているが，3つの手法の中では一番劣っている．

表4.34，表4.35の結果をもとに，有意差を検定した．「MEM(SVM)は正解であるがBERTは不正解であった分割箇所」の数と，「MEM(SVM)は正解であるがBERTは不正解であった分割箇所」の数と「MEM(SVM)は不正解であるがBERTは正解であった分割箇所」の数の合計数を用い，二項分布に基づく有意水準0.05の符号検定 (片側検定) を行った．有意差検定から得たp値を表4.36に示す．BERTとSVMとの有意差検定から得たp値を表4.37に示す．BERTの正解率は高い順から並べた．

表 4.36: BERT と MEM の有意差 (小説)

BERT	MEM			
正解率	①	①+④	①+⑤	①+⑥
0.8828	5.324240×10^{-9}	1.675412×10^{-10}	3.751062×10^{-8}	3.751062×10^{-8}
0.8815	0.000006	0.000009	0.000021	0.000021
0.8808	0.000002	0.000002	0.000008	0.000008
0.8802	0.000003	0.000004	0.000013	0.000013
0.8787	0.000116	0.000173	0.000389	0.000389
0.8783	0.000070	0.000102	0.000262	0.000262
0.8774	0.000331	0.000428	0.001008	0.001008
0.8748	0.003653	0.004869	0.009082	0.009082
0.8718	0.051519	0.062835	0.089551	0.089551
0.8658	0.394362	0.435951	0.521620	0.521620

表 4.37: BERT と SVM の有意差 (小説)

BERT	SVM			
正解率	①	①+④	①+⑤	①+⑥
0.8828	1.146417×10^{-15}	3.310440×10^{-13}	9.586034×10^{-15}	4.568805×10^{-13}
0.8815	1.814551×10^{-11}	9.259268×10^{-10}	3.634554×10^{-11}	2.158623×10^{-9}
0.8808	1.582604×10^{-12}	1.745165×10^{-10}	7.634098×10^{-12}	2.810156×10^{-10}
0.8802	2.638708×10^{-12}	2.260352×10^{-10}	9.559595×10^{-12}	6.099582×10^{-10}
0.8787	8.878181×10^{-10}	4.947080×10^{-8}	2.158623×10^{-9}	7.525441×10^{-8}
0.8783	2.184967×10^{-10}	1.855115×10^{-8}	8.769981×10^{-10}	2.884471×10^{-8}
0.8774	4.454613×10^{-9}	2.336214×10^{-7}	1.511549×10^{-8}	3.034303×10^{-7}
0.8748	2.029667×10^{-7}	0.000006	0.000001	0.000009
0.8718	0.000041	0.000532	0.000079	0.000669
0.8658	0.002621	0.018612	0.004389	0.022818

表 4.36 より，10 回試行した BERT の正解率は MEM に対して，10 回の中で下から 2 個の正解率では有意差はなかったが，残りの 8 個全て有意差があった．また，表 4.37 より，10 回試行した BERT の正解率は SVM に対して全て有意差があった．

また表 4.30 より，MEM と SVM で同じ素性を用いた時の正解率の有意差を検定した.「SVM は正解であるが MEM は不正解であった分割箇所」の数と,「SVM は正解で

あるがMEMは不正解であった分割箇所」の数と「SVMは不正解であるがMEMは正解であった分割箇所」の数の合計数を用い，二項分布に基づく有意水準0.05の符号検定(片側検定)を行った．有意差検定から得たp値を表4.33に示す．

表 4.38: MEMとSVMの有意差(小説)

SVM	MEM
①	0.000004
①+④	0.000175
①+⑤	0.000003
①+⑥	0.000138
②	0.549673
②+⑥	0.285793
③	0.000189
③+⑥	0.002020

表4.33より，②，②+⑥は有意差がなかったが，その他の素性に関しては有意差があった．全て有意差があったわけではないが，表4.35より推定精度はMEMが上回っている．MEM，SVMの2手法で比較すると，段落分割を行う上で，MEMの手法の方が有用な手法だと考えられる．

新聞記事，小説ともに正解率はBERT＞最大エントロピー法＞サポートベクトルマシン法という結果となった．小説に対する推定精度は差があまりないが，新聞記事に対する推定精度の差は大きい．今回の実験で使用したBERTの損失関数はSoftmax Cross Entropy Lossを使用しており，Softmax Cross Entropy Lossより2クラス問題を扱うことに優れている損失関数にすることや入力のデータの整形で，更に精度が上がるのではないかと考えられる．またMEMの実験で素性をいくつか追加したが，いずれもBERTの推定精度を上回っていない．最大エントロピー法は素性分析など行うことに適しているが，単純な段落の推定精度ではBERTに勝ることはできないのではないかと考えられる．

また今回の実験では，BERTは段落情報を用いていないのでMEM，SVMとの段落情報の正解率なしで手法の比較を行った．BERTでの入力文に段落情報を表す記号などを付与することで，同じ条件下での実験を行えるのではないかと考えている．段落情報付与での性能の比較を今後したい．

第5章　素性分析

段落分割の際にどのような素性が段落の分割，非分割かを判断しているかを調べた．5.1 節では，正規化 α 値を参考にした最大エントロピー法での素性分析について記述している．5.2 節では，BERT での素性分析について記述している．

5.1　最大エントロピー法での素性分析

最大エントロピー法での段落分割の際に得られた正規化 α 値 (3.4.2 節) をもとに素性分析を行った．なお最大エントロピー法を用いた結果は「文と文の間の箇所 (以下「文間箇所」) の直前の 1 文と直後の 1 文にある全単語」を素性として推定実験を行った結果から素性分析を行った．

5.1.1 節では，新聞記事の段落分割での素性分析について記述している．5.1.2 節では，小説の段落分割での素性分析について記述している．

5.1.1　新聞記事の素性

新聞記事で段落分割を行った際に，段落分割するか否かの判断に用いられた素性が得られた．文間箇所の直前の文にある素性によって段落の分割，非分割に影響を与えた素性と，その正規化 α 値と素性が入っている文と，その後の文を表 5.1 に示す．文間箇所の直後の文にある素性によって段落の分割，非分割に影響を与えた素性と，その正規化 α 値と，素性が入っている文とその前の文を表 5.2 に示す．また，表に示す正規化 α 値は，値が大きいほど段落分割の推定に役立ち，値が小さいほど非段落分割の推定に役立つ．

表 5.1: 新聞記事の素性

単語	正規化α値	文
直後が段落分割位置であると考えられた素性		
そして	0.7754	そして、わが日本は世界一の長寿国である。 ２００７年の年頭に当たって私たちは、この世界一のリストをどんどん増やしていこう、と提案する。
だから	0.7380	だから 犯人は自首して罪だけは償ってほしい。 昨年１２月１０日、喜代治さんは１カ月早く一周忌の法要を営んだ。
そのうえで	0.7037	そのうえ で当面の間、全国での洋菓子販売を全面的に休止すると発表した。 不二家によると、昨年１１月８日に埼玉工場で製造したシュークリーム２０００個に前日が消費期限となっていた牛乳を使用していた。
直後が段落分割位置ではないと考えられた素性		
例えば	0.2757	例えば 米ウォールストリート・ジャーナル紙は、ネットは事件事故の第一報や発表記事、紙は独自の調査報道記事や論説・解説とすみ分け、生き残ろうとしています。 紙の新聞から記者をネットによる報道の部門に移す新聞社も目立っています。
伝え	0.1717	ロイター通信は５日、議長派の正規部隊を強化するため、ブッシュ米政権が議長護衛隊の訓練費用や車両などの購入費として８６００万ドル（約１０２億円）を供与する方針だと 伝え た。 欧米はハマスを「テロ組織」とみなしており、議長の指導力強化を目指している。
供述	0.1623	自分もがんを患っており、認知症気味の母を一人にするのがふびんだったと 供述 しているという。 東容疑者は約５年前、兵庫県内に住む母が腰にけがをしたため、隣家に引き取り世話をしてきたという。

表 5.2: 新聞記事の素性

単語	正規化α値	文
直前が段落分割位置であると考えられた素性		
まずは	0.8549	「公（おおやけ）」の感覚の喪失とも言えるだろう。 まずは米国。
こうした	0.8270	米アップル・コンピュータの大ヒット商品「ｉＰｏｄ」も日本製の部品や素材がなければ製造できないのだ。 こうした日本製の「世界一」が、私たちの豊かな暮らしのモトだ。
さて	0.7894	植物としての効用をＰＲする動きも活発だ。 さて、あなたは春の七草を全部言えるだろうか。
直前が段落分割位置ではないと考えられた素性		
ようやく	0.3025	午後２時、約２０人の若者が「ワッショイ、ワッショイ」と声を上げながら三段の人やぐらを作り始めたが、花笠までもう少しというところで何度も崩壊。 ようやく花笠に飛びついて引き落とすと、見物客たちは歓声を上げ、我先にと花笠に群がっていた＝写真・山口政宣。
そういう	0.2224	地中海クロマグロの場合は、３分の２以上が日本市場向けだとみられます。 そういう意味で、日本の消費者にも責任の一端があるのは否定できません。
ほとんど	0.2194	悪くいえば、イノベーションとはＴＬＯ（技術移転機関）などへのお金のばらまきのように感じる。 米国のまれにしかない成功例だけを見た大学発ベンチャーはほとんどが失敗している

新聞記事に対して，最大エントロピー法での分割に関する素性分析では，表 5.1 より添加の意味をもつ接続詞「そして」,「そのうえで」が文間箇所の前の文章にあるとき分割と判断している．添加で前の文の付け加えをして，文間箇所の後の文章で話題転換している文章が多いのではないかと考えられる．順接の意味を持つ接続詞「だから」もあり，文間箇所で話題転換をしていることが多いのでないかと考えられる．また表 5.2 より，列挙の意味をもつ接続詞「まずは」が文間箇所の後の文章にあるとき分割と判断している．

非分割では，表 5.1 より例示の意味をもつ接続詞「例えば」が文間箇所の前にあるとき，後の文章でも同じ話をしている文章が多いのではないかと考えられる.「伝え」,「供述」といった似た意味の単語が文の終わりに付くと，次の文章も同じ内容のことを話すことが多く，非分割であると考えていると思われる．また表 5.2 より「そういう」などの前の文章の補足的な意味を示す単語が，文間箇所の後の文書にあるとき，非分割と判断していることが分かった．

5.1.2 小説の素性

小説で段落分割を行った際に，段落分割するか否かの判断に用いられた素性が得られた．文間箇所の直前の文にある素性によって段落の分割，非分割に影響を与えた素性と，その正規化α値と素性が入っている文と，その後の文を表5.3に示す．文間箇所の直後の文にある素性によって段落の分割，非分割に影響を与えた素性と，その正規化α値と，素性が入っている文とその前の文を表5.4に示す．表に示す正規化α値は，値が大きいほど段落分割の推定に役立ち，値が小さいほど非段落分割の推定に役立つ．

表 5.3: 小説の素性

単語	正規化α値	文
直後が段落分割位置であると考えられた素性		
ついに	0.7646	代助は必竟何しに新聞社まで出掛て来たのか、帰るまで ついに 問い詰めずに済んでしまった。 代助は翌日になって独り書斎で、昨夕の有様を何遍となく頭の中で繰り返した。
――	0.7425	あの声はと、耳の走る見当を見破ると ―― 向うにいた。 花ならば海棠かと思わるる幹を背に、よそよそしくも月の光りを忍んで朦朧たる影法師がいた。
どうしても	0.7166	どうしても 表情に一致がない。 悟りと迷が一軒の家に喧嘩をしながらも同居している体だ。
直後が段落分割位置ではないと考えられた素性		
あるいは	0.3627	あるいは この詩の意味をわれらの身の上に引きつけて解釈しても愉快だ。 二人の間には、ある因果の細い糸で、この詩にあらわれた境遇の一部分が、事実となって、括りつけられている。
たしかに	0.3281	たしかに 誰かうたっている。 細くかつ低い声には相違ないが、眠らんとする春の夜に一縷の脈をかすかに搏たせつつある。
ところが	0.3026	ところが この男がある芸妓と関係って、何時の間にか会計に穴を明けた。 それが曝露したので、本人は無論解雇しなければならないが、ある事情からして、放って置くと、支店長にまで多少の煩が及んで来そうだったから、其所で自分が責を引いて辞職を申し出た。

表 5.4: 小説の素性

単語	正規化α値	文
直前が段落分割位置であると考えられた素性		
やがて	0.8268	が三千代の顔は陰になって見えなかった。 やがて、平岡は筆を机の上へ抛げ付ける様にして、座を直した。
それから	0.7731	奇心のあるにも拘わらず、取り合う事を敢てしなかった。 それから 約四十分程して、老人は着物を着換えて、袴を穿いて、俥に乗って、何処かへ出て行った。
たちまち	0.6878	固より急ぐ旅でないから、ぶらぶらと七曲りへかかる。 たちまち 足の下で雲雀の声がし出した。
直前が段落分割位置ではないと考えられた素性		
「	0.2516	茶の色の黒く焦げている底に、一筆がきの梅の花が三輪無雑作に焼き付けられている。 「御菓子を」と今度は鶏の踏みつけた胡麻ねじと微塵棒を持ってくる。
とうとう	0.2434	三千代は看護の為附添として一所に病院に移った。 病人の経過は、一時稍佳良であったが、中途からぶり返して、とうとう 死んでしまった。
しかし	0.2343	子供のうちは心魂に徹して困却した事がある。 しかし 成人の今日では、それにも別段辟易する必要を認めない。

小説に対して，最大エントロピー法での分割に関する素性分析では，表 5.3 より文間箇所の前の文章中に「どうしても」や「——」がある時，文間箇所は分割位置だと判断している．また表 5.4 より，文間箇所の後の文章の文頭に「やがて」,「それから」などが出現すると分割と判断している．

非分割では表 5.3，表 5.4 より文間箇所の前の文章に逆接の意味をもつ接続詞「ところが」があると 2 文章同じ内容が続くことが多く，文間箇所は非分割である考える．また「ところが」と同じ逆接の意味の「しかし」が後の文章にあると,「しかし」から始まる文章は 1 文で終わることが多く,「しかし」の前の文間箇所が非分割であると判断している．

新聞記事での直前の文章中の「ところが」，直後の文章中の「しかし」について調べた．表 5.5 に示す．

表 5.5: 「ところが」,「しかし」の新聞記事の頻度

	新聞記事での頻度	
単語	分割	非分割
ところが	2	15
しかし	35	93

表 5.5 より，直前の文章中の「ところが」があるときの文間箇所，直後の文章中の「しかし」があるときの文間箇所はともに非分割の頻度が高い.

5.2　3単語連続を用いたBERTの素性分析

3.4.3節の3単語連続を入力とすることでBERTでの素性分析を行った．なお用いたデータは,「★」の入ったデータを3単語連続区切りに整形し入力した．5.2.1節では,新聞記事の段落分割での素性分析について記述している．5.2.2節では，小説の段落分割での素性分析について記述している．

5.2.1　新聞記事の素性

新聞記事を用いたBERTでの実験の素性分析を行った．素性分析から得た分割に関する単語とその数値の上位30個を表5.6，非分割に関する単語とその数値の上位30個を表5.7に示す．表5.6，表5.7は「★」が含まれる3単語連続での数値の順に示した．

また，表5.6の上位30個の他に素性分析から得た素性の中で，有用であると判断できる分割に関する単語とその数値を表5.8に示す．表5.7の上位30個の他に素性分析から得た素性の中で，有用であると判素できる非分割に関する単語とその数値を表5.9に示す.

表 5.6: 3 単語連続での分割に関する上位 30 個素性 (新聞記事)

	数値	
単語	分割	分割しない
目指したい——。★	0.9984	0.0016
べきだ——。★	0.9984	0.0016
ね」——。★	0.9979	0.0021
★こうした取り組み	0.9978	0.0022
もいる——。★	0.9975	0.0025
ない」——。★	0.9975	0.0025
」——。★毎日	0.9974	0.0026
や」——。★	0.9973	0.0027
マフィア」——。★	0.9965	0.0035
★こうした動き	0.9963	0.0037
なぜか——。★	0.9962	0.0038
ならず——。★	0.9959	0.0041
たい——。★東京	0.9959	0.0041
」——。★昨年	0.9957	0.0043
は無理——。★	0.9956	0.0044
★こうしたこと	0.9956	0.0044
★イラン生まれ	0.9954	0.0046
ず——。★東大阪	0.9952	0.0048
★イランは	0.9952	0.0048
★イランについて	0.9948	0.0052
無理——。★この	0.9947	0.0053
★猷サブプライムローン	0.9945	0.0055
」——。★石油	0.9943	0.0057
★こうした問題	0.9941	0.0059
★イノベーションは	0.9939	0.0061
」——。★0	0.9939	0.0061
★事件は	0.9938	0.0062
★この忙しい	0.9937	0.0063
★イランにとって	0.9932	0.0068
★かつては	0.9927	0.0073

表 5.7: 3 単語連続での非分割に関する上位 30 個素性 (新聞記事)

	数値	
単語	分割	分割しない
下り。★	0.0011	0.9989
中身。★	0.0012	0.9988
。★後で	0.0012	0.9988
今年。★	0.0012	0.9988
欠席。★	0.0012	0.9988
持参。★	0.0012	0.9988
画。★	0.0012	0.9988
行き。★	0.0012	0.9988
。★それで	0.0012	0.9988
男の子。すばらしい	0.0012	0.9988
前後。★	0.0012	0.9988
銭。★	0.0012	0.9988
刻み。★	0.0012	0.9988
失点。★	0.0012	0.9988
円。★	0.0012	0.9988
。★没後	0.0012	0.9988
補給。★	0.0012	0.9988
四球。★	0.0012	0.9988
未定。★	0.0012	0.9988
入り。★	0.0012	0.9988
毎日。★	0.0012	0.9988
罰金。★	0.0013	0.9987
。★不偏不党	0.0013	0.9987
クリア。★	0.0013	0.9987
変化。年末年始	0.0013	0.9987
クリケット。★	0.0013	0.9987
。★読み返す	0.0013	0.9987
。★喪主	0.0013	0.9987
。★それから	0.0013	0.9987
味。★	0.0013	0.9987

表 5.8: 分割に関する素性 (新聞記事)

	数値	
単語	分割	分割しない
目指したい——。★	0.9984	0.0016
べきだ——。★	0.9984	0.0016
★こうした取り組み	0.9978	0.0022
★こうしたこと	0.9956	0.0044
★かつては	0.9927	0.0073
★この時期	0.9924	0.0076
★もともとは	0.9918	0.0082

表 5.9: 非分割に関する素性 (新聞記事)

	数値	
単語	分割	分割しない
下り。★	0.0011	0.9989
中身。★	0.0012	0.9988
。★後で	0.0012	0.9988
欠席。★	0.0012	0.9988
持参。★	0.0012	0.9988
。★それで	0.0012	0.9988
。★それから	0.0013	0.9987

表5.8，表5.9の有用であると判断した素性の実データ (訓練データの新聞記事7,500) に対する実際の分割，非分割の頻度について調べた．表5.8の「——。★」,「★こうした」,「★かつて」,「★この」,「★もともと」の5つに対する実データの頻度を表5.10に示す．表5.9の「下り。」「中身。」などの文末の体言止め,「★それで」,「★それから」の3つに対する実データの頻度を表5.11に示す．

表 5.10: 分割に関する素性の頻度 (新聞記事)

	実データの頻度	
単語	分割	非分割
——。★	3	5
★こうした	4	2
★かつて	1	1
★この	26	53
★もともと	1	0

表 5.11: 非分割に関する素性の頻度 (新聞記事)

	実データの頻度	
単語	分割	非分割
体言止め	110	1,167
★それで	8	14
★それから	0	3

表 5.10 より,「★こうした」,「★もともと」は実データの分割の頻度が非分割より多かった.「★こうした」は，MEM の素性分析にも出現しており，分割に関して有用な素性であると言える.「★もともと」は実データの頻度が 1 回と少なく，有用な素性であるとは言い切れない．他の 3 単語は，実データでは非分割の方が多い，または同数である.

また表 5.11 より，頻度を調べた 3 つは全て実データで非分割の方が多く,「★それから」は 3 回と頻度が少ないが，いずれも非分割に関する素性として有用な素性であると言える.

「★こうした」,「★それで」,「★それから」に対して，MEM での素性分析の正規化 α 値について調査した．得た値を表 5.12 に示す.

表 5.12: BERT で得た素性の MEM の正規化 α 値 (新聞記事)

	正規化 α 値	
単語	分割	非分割
こうした	0.8270	0.1730
それで	0.3831	0.6169
それから	0.4890	0.5110

表 5.12 より,「こうした」は分割の値が 0.8270 と高く，最大エントロピー法においても分割に関する素性であることが分かる.「それで」は非分割の値が 0.6169 であり，非分割に関する素性であるが,「それから」は非分割の値が 0.5110 であり、最大エントロピー法では非分割に関する素性であるとは言い切れない.

有用な素性と考えられる「★こうした」,「★それで」,「★それから」の 3 単語の小説の訓練データでの頻度について調べた.「★こうした」は小説に存在しなかったので、他 2 単語について調べた頻度を表 5.13 に示す.

表 5.13: 小説での表 5.11 の単語の頻度

	小説の頻度	
単語	分割	非分割
★それで	1	20
★それから	15	27

表 5.13 より，2 単語とも非分割の頻度の方が多いことが分かる.

5.2.2 小説の素性

小説を用いた BERT での実験の素性分析を行った．素性分析から得た分割に関する単語とその数値の上位 30 個を表 5.14，非分割に関する単語とその数値の上位 30 個を表 5.15 に示す．

また，表 5.14 の上位 30 個の他に素性分析から得た素性の中で，有用であると判断できる分割に関する単語とその数値を表 5.16 に示す．表 5.15 の上位 30 個の他に素性分析から得た素性の中で，有用であると判素できる非分割に関する単語とその数値を表 5.17 に示す．表 5.14，表 5.15 は「★」のついた 3 単語連続を用いた．

表 5.14: BERT での分割に関する上位 30 個素性 (小説)

	数値	
単語	分割	分割しない
、一一」★代	0.9970	0.0030
もの一一」★代	0.9969	0.0031
一一」★代助	0.9967	0.0033
」★彼	0.9952	0.0048
」★まだ	0.9950	0.0050
ゃい」★	0.9947	0.0053
」★来	0.9944	0.0056
★朝は	0.9943	0.0057
」★嫂	0.9942	0.0058
」★疳	0.9942	0.0058
★翌日、	0.9939	0.0061
」★今度	0.9939	0.0061
」★誠	0.9935	0.0065
」★また	0.9933	0.0067
★夜は	0.9932	0.0068
」★こう	0.9929	0.0071
」★雨	0.9926	0.0074
★朝の	0.9925	0.0075
」★まず	0.9924	0.0076
★しばらくは	0.9924	0.0076
★朝飯は	0.9924	0.0076
★翌日は	0.9917	0.0083
★夜が	0.9914	0.0086
」★寺尾	0.9913	0.0087
★夜の	0.9912	0.0088
」★兄	0.9911	0.0089
」★縫子	0.9909	0.0091
」★親爺	0.9907	0.0093
★翌日朝	0.9906	0.0094
★しばらくする	0.9905	0.0095

表 5.15: BERT での非分割に関する上位 30 個素性 (小説)

	数値	
単語	分割	分割しない
★「遣る	0.0014	0.9986
★「でしょ	0.0015	0.9985
★「そいつ	0.0015	0.9985
★「妙	0.0015	0.9985
★「庭	0.0015	0.9985
★「門野	0.0015	0.9985
★「焦る	0.0015	0.9985
★「ありゃ	0.0015	0.9985
★「学校	0.0015	0.9985
★「支店	0.0015	0.9985
★「愚図	0.0015	0.9985
★「菓子	0.0015	0.9985
★「分ら	0.0015	0.9985
★「如何なる	0.0015	0.9985
★「何で	0.0015	0.9985
★「落ち	0.0015	0.9985
★「偉い	0.0015	0.9985
★「姉さん	0.0015	0.9985
★「奇麗	0.0015	0.9985
★「兄さん	0.0015	0.9985
★「此奴	0.0015	0.9985
★「うん	0.0015	0.9985
★「本当	0.0015	0.9985
★「好い	0.0015	0.9985
★「貴様	0.0015	0.9985
★「なるほど	0.0015	0.9985
★「先刻	0.0015	0.9985
★「覚え	0.0015	0.9985
★「不断	0.0016	0.9984
★「代	0.0016	0.9984

表 5.16: BERT での分割に関する有用素性 (小説)

	数値	
単語	分割	分割しない
」★彼	0.9952	0.0048
」★まだ	0.9950	0.0050
★しばらくは	0.9924	0.0076
★その後から	0.9895	0.0105
★その後の	0.9872	0.0128
★やがて、	0.9821	0.0179

表 5.17: BERT での非分割に関する有用素性 (小説)

	数値	
単語	分割	分割しない
★「遣る	0.0014	0.9986
★「でしょ	0.0015	0.9985
。★それでいて	0.0050	0.9953
。★ただし	0.0050	0.9950
★けれども機嫌	0.0071	0.9929

表 5.16，表 5.17 の有用であると判断した素性の実データ (訓練データ) に対する実際の分割，非分割の頻度について調べた．表 5.16 の「」★彼」,「」★まだ」などの会話後の「」★」,「★しばらく」,「★その後」,「★やがて」の 4 つに対する実データの頻度を表 5.18 に示す．表 5.17 の「★「遣る」「★「でしょ」などの文間箇所のすぐ後に会話の「★「」,「★ただし」,「★けれども」の 3 つに対する実データの頻度を表 5.19 に示す．

表 5.18: 分割に関する素性の頻度 (小説)

	実データの頻度	
単語	分割	非分割
」★	10	136
★しばらく	8	17
★その後	3	3
★やがて	6	15

表 5.19: 非分割に関する素性の頻度 (小説)

	実データの頻度	
単語	分割	非分割
★「	0	842
★ただし	0	4
★けれども	0	5

表 5.18 より,「★その後」は分割と非分割の頻度は同数であり，他の 3 つは非分割の方が頻度が多い．小説では分割に関する素性を得ることができていない．また表 5.19 より，頻度を調べた 3 つは全て実データでは非分割の方が多く，全て有用な素性であると言える．

表 5.18，表 5.19 の結果より，分割に関する素性を得ることができなかったが，非分割に関する素性を得ることができた．

有用な素性と考えられる「★「」,「★ただし」,「★けれども」に対して，MEM での素性分析の正規化 α 値について調査した．得た値を表 5.20 に示す．

表 5.20: BERT で得た素性の MEM の正規化 α 値 (小説)

	正規化 α 値	
単語	分割	非分割
「	0.2516	0.7484
ただし	0.4738	0.5262
けれども	0.4879	0.5121

表 5.20 より,「「」は非分割の値が 0.7484 と高く，最大エントロピー法においても非分割に関する素性であることが分かる.「ただし」,「けれども」は非分割の値が 0.5 に近く，最大エントロピー法では非分割に関する素性であるとは言い切れない.

表 5.12 より,「こうした」は分割の値が 0.8270 と高く，最大エントロピー法においても分割に関する素性であることが分かる.「それで」は非分割の値が 0.6169 であり，非分割に関する素性であるが,「それから」は非分割の値が 0.5110 であり、最大エントロピー法では非分割に関する素性であるとは言い切れない.

有用な素性と考えられる表 5.19 の単語の新聞記事での正解数について調べた.「★けれども」は新聞記事になかったため残り 2 つの調べた結果を表 5.21 に示す.

表 5.21: 新聞記事での表 5.19 の単語の頻度

	新聞記事の頻度	
単語	分割	非分割
★「	98	324
★ただし	2	5

表 5.21 より，2 単語とも非分割の方が多く，新聞記事でも有用な素性であると言える.

新聞記事，小説の BERT での素性分析の結果から，文章を 3 単語連続にすることで，文間箇所 (「★」) 付近の箇所での非分割に関する素性を得ることができた．非分割を判断する文間箇所 (「★」) の位置を含まない 3 単語連続の分析をすることが難しい．しかし，文間箇所付近の単語 (例えば接続詞) が分割，非分割に関して影響が高く，文間箇所付近の素性を以外を得ることができないことは大きな問題ではない.

最大エントロピー法を用いた素性分析は，分割，非分割に関する素性を得ることができた．また，BERT を用いた素性分析は，新聞記事と小説に対して分割に関する素性を 1 つしか得ることができなかった．1 つしか得ることができていないことから，BERT での素性素性では，分割に関する素性を得ることは難しいと考えられる．非分割に関する素性は得ることができたが，最大エントロピー法での素性分析と比べると有用な素性の数は少なかった．段落分割の推定において BERT は最大エントロピー法の推定精度を上回るが，素性分析は最大エントロピー法を用いた手法の方が有効であると考えられる.

また今回の実験では 3 単語連続で行ったが，4 単語，5 単語連続に単語を増やすこと

で入力データが 3 単語連続より情報量が多いため，分割に関する素性を得ることができるのではないかと考えられる．今後，単語数を増やして分割に関する素性を得ることができるかを調査したい．

第6章　今後の課題

本研究では，BERT と最大エントロピー法を用いて日本語文章の段落分割の推定を行い，どちらが優れているかを比較した．また最大エントロピー法では正規化 α 値を用いて分割，非分割に関する素性を，BERT では 3 単語連続それぞれに対する出力値から素性分析を行い，非分割に関する素性を得たが，いくつかの問題が残っている．本章では，残っている問題を今後の課題として以下にまとめる．

- BERT の損失関数などを変更や入力データの改良を行い，更なる段落分割の推定精度の向上をはかる．
- MEM での推定精度を向上させるため，更なる素性の追加を検討する．
- MEM，SVM での段落情報の素性追加は，BERT を用いた分割推定での条件と公平ではなかったため考察を行わなかったが，BERT での入力データに段落情報を付与して推定を行い，手法間での段落情報付与の性能比較をしたい．
- 3 単語連続に対して素性分析で分割に関する素性を得ることができなかったので，4 単語，5 単語連続に単語数を増やすことや新しい入力データを提案することで分割に関する素性を調査したい．
- 段落分割の正解率は BERT が MEM を上回るため，BERT が正解かつ MEM が不正解だと判断した文章内にある単語が，BERT での素性分析に役立つのではないかと考えた．そこで「BERT が正解，最大エントロピー法が不正解」と「BERT が不正解，最大エントロピー法が正解」の文章の単語の数を用いて，有意差検定を行い，有意差のある単語を得ることで，素性分析を行いたい．

第7章　おわりに

本研究では，BERT と最大エントロピー法を用いて日本語文章の段落分割の推定を行い，どちらが優れているかを比較した．また最大エントロピー法では正規化 α 値を用いて分割，非分割に関する素性を，BERT では 3 単語連続それぞれに対する出力値から素性分析を行い，非分割に関する素性を得た．

新聞記事と小説を学習データとテストデータとして実験を行った結果，新聞記事ではベースラインの正解率 0.6743 に対して，BERT の正解率は 0.7564，最大エントロピー法の正解率は 0.6959 であった．小説ではベースラインの正解率 0.8492 に対して，BERT の正解率は 0.8772，最大エントロピー法の正解率は 0.8658 であった．どちらの手法もベースラインを上回っており，新聞記事に対しての手法間の推定精度の差は 0.0605 と大きい．BERT のモデルを変更することで推定精度の向上が見込めるため，最大エントロピー法の精度が上回ることは難しい．今後は BERT の損失関数の変更，入力データの整形の変更など行い推定精度を向上させたい．

また，最大エントロピー法と BERT の結果から素性分析を行った．最大エントロピー法では正規化 α 値を使用することで素性分析を行い，分割，非分割に関する有用な素性を得ることができた．BERT では，分割の推定で用いたテストデータを 3 単語ごとに分け，分けた 3 単語それぞれに対する分割，非分割の出力値から素性分析を行う新しい手法を提案した．非分割に関する素性を得ることはできたが，分割に関する素性は新聞記事で 1 個の素性しか得ることができなかった．BERT の素性分析の結果から，最大エントロピー法の素性分析には劣るが，BERT でも素性分析を行うことができると言える．しかし，BERT では分割に関する素性を十分に得ることができていないことから，最大エントロピー法の方が有効な手法であると考えられる．今後は BERT での素性分析において，テストデータの単語数を 3 単語から増やすことで分割に関する素性を得ることができるかを調査したい．

謝辞

本研究を進めるにあたり，研究の進め方や本論文の書き方など，細部にわたる御指導を頂きました，鳥取大学工学部電気情報系学科自然言語処理研究室の村田真樹教授に心から御礼申し上げます．また，本研究を進めるにあたり，御指導，御助言を頂きました，村上仁一准教授に心から御礼申し上げます．また，ご多忙の中，助言をいただきました木村周平教授に厚く御礼申し上げます．その他様々な場面で御助言を頂いた自然言語処理研究室の皆様に感謝の意を表します．

参考文献

[1] 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第 34 回全国大会, Vol. 4, No. 34, pp. 1–4, 2020.

[2] Dmitriy Genzel. A Paragraph Boundary Detection System. *CICLing 2005*, pp. 816–826, 2005.

[3] Caroline Sporleder and Mirella Lapata. Automatic Paragraph Identification: A Study across Languages and Domains. *held in conjunction with ACL 2004*, pp. 72–79, 2004.

[4] 中野滋徳, 足立顕, 牧野武則. 語の反復距離に基づく段落境界の認定. 自然言語処理, Vol. 13, No. 2, pp. 3–26, 2006.

[5] Eric Sven Ristad. Maximum Entropy Modeling for Natural Language. In *ACL/EACL Tutorial Program, Madrid*, 1997.

[6] 村田真樹, 内山将夫, 内元清貴, 馬青, 井佐原均. 種々の機械学習手法を用いた多義解消実験. 電子情報通信学会言語理解とコミュニケーション研究会, NLC2001-2, pp. 7–14, 2001.

[7] Masao Utiyama. Maximum Entropy Modeling Packagen: http://www.nict.go.jp/x/x161/members/mutiyama/software.htmlmaxent, 2006.

[8] Masaki Murata and Kiyotaka Uchimoto and Masao Utiyama and Qing Ma and Ryo Nishimura and Yasuhiko Watanabe and Kouichi Doi and Kentaro Torisawa. Using the maximum entropy method for natural language processing: Category estimation, feature extraction, and error correction. *Cognitive Computation*, Vol. 2, pp. 272–279, 2010.

[9] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*, 2018.

[10] 黒橋 禎夫 and Chenhui Chu and 村脇 有吾. BERT: http://nlp.ist.i.kyoto-u.ac.jp/?ku_bert_japanese#c8ee1db1, 2019.

[11] Nello Cristianini and John Shawe-Taylor. An Introduction to Support Vector Machines and Other Kernel-based Learning Methods. *Cambridge University Press*, 2000.

[12] Taku Kudoh. TinySVM: http://cl.aist-nara.ac.jp/ taku-ku// software/TinySVM/index.html, 2000.

[13] 工藤拓, 松本裕治. Support Vector Machine を用いた Chunk 同定. 自然言語処理, Vol. 9, No. 5, pp. 3–21, 2002.

[14] Masaki Murata and Satoshi Ito and Masato Tokuhisa and Qing Ma. Order Estimation of Japanese Paragraphs by Supervised Machine Learning and Various Textual Features. *Journal of Artificial Intelligence and Soft Computing Research*, Vol. 5, No. 4, pp. 247–255, 2015.