DEIM Forum 2021 G24-1

校閲時の事実確認作業における誤り箇所の自動推定

古田　朋也†　　鈴木　　優†

† 岐阜大学工学部電気電子・情報工学科　〒 501–1193 岐阜県岐阜市柳戸 1 番 1

あらまし　本研究では校閲作業のうち，事実確認の支援を目的とする．事実確認の必要性が高い文章と確認箇所を提示することによって，作業者の労力削減を図る．事実確認の必要性が高いとされる文章は内容に関する誤りを含む文章であり，文法上は正しい文章となっている．そのため，内容誤りを含む文章かどうかの判定基準を人手にて定義することは難しい．そこで，機械学習を使用することによって内容に関する誤りを含む文章かどうかの判定を行う．モデルの構築に必要となる内容に関する誤りを含む文章を訂正事例を利用することによって収集する．訂正前後の文書を比較することによってデータセットを作成した．作成したデータセットを用いてファインチューニングを行うことによって，BERT を使用した分類器を構築する．そして，構築した分類器によって文章ごとに内容に関する誤りを含む文章かどうかの判定を行う．また，事実確認の必要性が高いと判定された文章に対しては，分類時の Attention をそのまま可視化することによって確認箇所として作業者に提示する．上記手法を Web アプリケーションとして実装し，校閲作業における実用性について評価実験を行った．

キーワード　校閲作業，支援システム，文章分類，機械学習

1　はじめに

現在，インターネット上には校閲が行われていない記事がいくつか存在しているが，そのような記事は閲覧者に誤った情報を与えてしまう．そのため校閲作業は重要な作業となっているが，多くの時間と労力がかかる．文書の校閲作業は，誤字脱字の訂正と事実確認の二種類に大きく分けられる．そのうち，誤字脱字訂正の支援はいくつか行われているが，事実確認作業の支援はほとんど行われていない．そこで本研究では事実確認作業に着目し，確認が必要となる箇所の推定を自動で行い，結果を作業者に提示することによって，作業者の労力削減を図る．

事実確認作業において，すべての文章を確認した方が当然良いが，それを実行すると多くの時間と労力を消費してしまう．この作業において作業者が訂正すべき文章は，内容に関する誤りが発生している文章である．そのため，誤りを含んでいる可能性が低い文章に対する事実確認の必要性は低く，このような文章に対して確認作業を実施することは効率の良い作業とは言い難い．そこで，文章それぞれに対して誤りを含んでいるかどうかの判定を自動で行い，その判定結果に応じた確認必要性を作業者に提示する．それに対して，作業者は確認必要性の高い文章から順に事実確認することによって，効率よく作業を進めることができると考えた．その際に，どこまで事実確認を行うかの判断は作業者に委ねることになる．

確認の必要性が高いとされる内容誤りを含む文章は，文法上は正しい文章である．そのため，誤字脱字を含む文章と異なり，人手にて判定基準を定義することが困難である．そこで，機械学習を利用することによって文章の判定を行う．訂正事例を基にしたデータと機械学習を利用することによって文章中に内容誤りを含むかどうかの判定が可能になると考えた．本稿では，様々な自然言語処理タスクにおいて汎用性の高いモデルである BERT [1] による分類器を構築する．また，作業者には文章ごとの分類結果だけでなく，文章中の事実確認すべき箇所の提示も行う．確認必要性が高いとされた文章に対して確認箇所の提示を行うことによって確認作業の実施を促す．BERT にも存在している Attention [2] は分類時の判定根拠を示しているため，誤りを含むと判定された場合は誤り箇所を示していると考えられる．そこで，Attention を可視化することによって作業者に確認箇所として提示する．また，誤りを含んでいる可能性が低いとされた文章は確認作業を実施する必要がほとんどない．そのため，そのような文章に対して確認箇所の提示は行わない．

分類器を構築する際に，大量の内容に関する誤りを含む文章が必要となる．そのため，データセット作成時に訂正事例を利用する．著者が一人しかいない文書の訂正事例を用いても，特定の著者の誤りやすい箇所にしか対応できなくなってしまう．それでは，支援システムとしての汎用性に欠けてしまうため，著者が複数である文書を使用したい．また，大量に用意する必要があるのは内容誤りを含む文章であるが，訂正前の文書だけではどの文章に誤りが含まれているのか判断することが難しい．そのため，訂正前後の文章を比較することによって誤りを含む文章かどうかを判断する必要がある．そこで，複数人で編集が行われており，編集履歴を利用することによって訂正前後の記事が確保できる Wikipedia の記事をデータ源として使用することにした．

用意した訂正前後の文書を比較すると，主に四種類の文章が存在していた．一字一句そのままの文章，文章全体の意味は変わっていないが言い回しが訂正された文章，記述されている情報が追加・更新された文章，文章自体が削除または事実に誤りがあり訂正された文章である．この事実から，文章は訂正内容によって四種類の分類ができそうであると考えた．そのため，データセット作成時に訂正前の文章に対して四種類のラベルを振り分けた．ラベルを付与する際にすべてを人手による比較で

行うと，記事内の文章の約五割は一字一句そのまま残っていた文章であった．そのため，機械による比較を取り入れ，ラベルをできる限り自動で付与することによって，作成時間の短縮が可能であると考えた．まず，文書中から機械による比較にて，一字一句そのままの文章を抽出してラベルを付与する．これによって残った文章は何らかの訂正が行われており，訂正前と訂正後の文章が異なる．そして訂正された内容もそれぞれ異なるため，これ以上機械で判断することは困難である．そのため，残りの文章に対しては人手にて訂正前と訂正後の文章を比較することが適切であると考えた．人手による比較にて訂正内容を確認し，残りの文章にラベルを振り分けた．機械による比較を取り入れたことにより，すべて人手にて行った場合のおよそ半分の所要時間にてラベル付けを行うことが可能となった．

作成した訂正事例によるデータセットを用いて，BERT による分類器を構築した．内容に関する誤りの有無についての分類を行い，確認必要性が高い文章に対しては Attention を可視化した確認箇所を提示する．そして，上記手法を Web アプリケーションとして実装した．提案手法の評価実験として，評価値に基づくモデルの分類性能の評価と，実例に基づく校閲作業における実用性の評価を行った．モデルの分類性能は Accuracy だけに注目した場合，高いもので 0.73 となっており，実用に値する数値が出ていると考える．また，判定結果を校閲作業の観点で確認したところ，本手法によって確認すべき文章と判定された文章の多くが確認作業の必要となる文章であった．Attention をそのまま可視化した確認箇所についても，適切な箇所が提示されることが多かった．内容に関する誤りを含む文章は，誤字脱字を含む文章と比較すると判定基準の定義が難しいため，事実確認作業の支援を行うことは困難であった．しかし評価実験の結果から，訂正事例を基にしたデータと機械学習を利用した内容誤り箇所の推定を行うことによって，事実確認作業の支援が可能であることが明らかとなった．一方で，誤りを含む可能性が高い文章の取りこぼし，不適切と思われる確認箇所がいくつか提示されている，など問題点もいくつか見つかった．校閲作業における実用性を向上させるために，上記の問題点について改善策を講じる必要がある．

本論文における貢献は以下の通りである．

- 訂正事例を用いた内容誤り箇所の推定を行うことにより，事実確認作業の支援が可能であることが明らかとなった．
- 事実確認すべき箇所の提示に，分類時の Attention が利用可能であることが明らかとなった．

2 関 連 研 究

校閲作業に関連する研究は，既にいくつか行われている．高橋ら [3] は，誤字脱字箇所推定と訂正候補文字の提示についての手法を提案している．Bidirectional LSTM を用いて，文字毎に誤字かどうかの判定を行う確率モデルと周辺単語から単語間に入る文字を予測する言語モデルを構築し，入力文に対して，それぞれのモデルで判定を行う．これら二種のモデルの判定結果から得られた情報をランダムフォレストの入力とすることによって誤字脱字の有無を判定している．そして，誤字脱字が存在していた場合，確率モデルと言語モデルの出力から得られた情報を利用して訂正を行っている．

今村ら [4] は，日本語助詞誤りの訂正についての手法を提案している．条件付き確率場 (CRF) を用いることによって，誤り箇所に対して訂正候補文字を決定している．その際に，マッピング素性とリンク素性の二種類の素性を用いている．さらに，リンク素性については，n-gram 素性と言語モデル確率を併用することにより，訂正結果が日本語文として適切であるかどうかを測っている．

鈴井ら [5] は，日本語文章の不自然箇所検知についての手法を提案している．Yahoo!知恵袋と Wikipedia から，自然な日本語文章のみをデータとして使用することによって，ニューラルネットワークによる言語モデルを構築している．このモデルに対して判定したい単語の周辺単語のみを入力として，その単語の出現確率の推定を行う．この自然な日本語文章における出現確率の推定によって確率が低いとされた単語を不自然箇所として決定している．

校閲作業は大きく二種類の作業に分けられる．誤字脱字訂正を行う作業と事実関係の確認を行う作業である．上記の先行研究は二種類の作業のうち，誤字脱字訂正に関連する研究となっている．先行研究では，誤字脱字訂正についての研究は多く行われているが，事実確認作業についての研究はあまり行われていない．そこで本研究では，二種類の作業のうち，事実確認作業に焦点を当てる．なお，事実確認作業に関連する研究として，ファクトチェックについての研究が行われている．

内山ら [6] はファクトチェックにおける要検証記事の探索支援についての手法を提案している．Twitter におけるニュース記事に対する言及を利用することによって，ツイートごとに検証の必要性を示唆する端緒情報とである確率を推定している．そして，端緒情報である確率によってスコア付けを行い，記事ごとにまとめる．まとめたスコアを用いて，それぞれの記事について検証必要度のランク付けを行い，作業者へ提示している．これにより，作業者はランクの高い記事から検証していけば良いことになる．

上記研究では，ファクトチェック作業を行うかどうかの判断支援をしている．Twitter の情報を基にした確認作業を行う前段階の支援となっている．本研究では，訂正事例のみを用いて事実確認が必要な箇所の推定を行う．この推定を行うことによって，確認作業を行う際の支援を試みる．先行研究とは異なる場面での作業支援を行うことによって，労力の更なる削減が可能になると考えた．

3 提 案 手 法

本手法では，誤りの発生が見込まれる箇所の推定を分類問題として扱う．提案手法の概要を図 1 に示す．提案手法によるモデル構築は Step.A，モデル構築後の誤り箇所推定は Step.B の流れで行う．それぞれの詳細については，括弧内に示す節にて述べる．

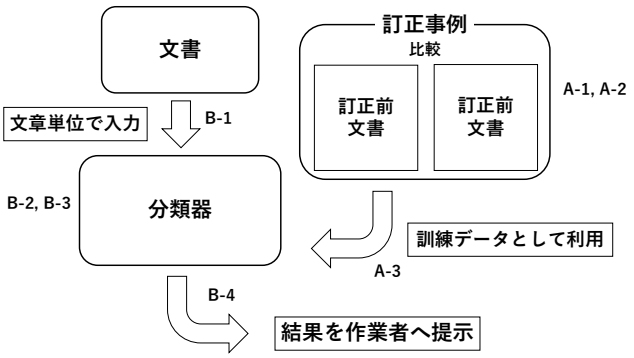

図 1　提案手法の概要

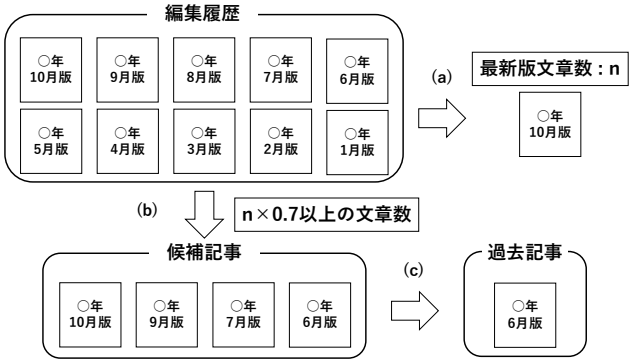

図 2　過去記事の選出方法

Step　A-1　過去の訂正事例から，訂正前と訂正後の二種類の文書を用意する．(3.1.1 節)

A-2　二種類の文書を比較することによって，四種類のラベルを付与する．(3.1.2 節)

A-3　過去の訂正事例を利用したデータセットにて，分類器として用いる BERT のファインチューニングを行う．(3.2 節)

Step　B-1　事実確認作業を行う文書を入力する．その際，文章単位に分割して，各文章ごとに判定を行う．(3.2.1 節)

B-2　構築した分類器を用いて四種類に分類する．(3.2.1 節)

B-3　レベル 3，2，1 と判定された文章については，確認箇所として Attention の可視化を行う．(3.3.1 節)

B-4　分類器による判定結果と，Attention による確認箇所を作業者へ提示する．

3.1　データセット作成

本手法では，与えられた文章に対して誤りが発生しているかどうかを機械学習を利用して判定する．そのため，訓練データとして内容に関する誤りが発生している文章を大量に用意して，分類器の訓練を行う必要がある．そこで，過去の訂正事例を利用することにより内容に関する誤りを含む文章を収集する．データとして大量に用意したいのは内容誤りを含む文章であるため，文書としては訂正前のものを使用することが適切であると考えた．しかし，訂正前の文書だけではどの文章に誤りが含まれているのか判断することが困難である．そこで，訂正前と訂正後の文書を用意して比較することによって内容誤りが発生していた文章を抽出する．訂正前の文章に対して，表 1 にて示すラベルを文書の比較時に付与することによってデータセットを作成する．

3.1.1　使用データ

本手法では訂正事例として Wikipedia の記事を利用する．著者が一人しかいない文書の訂正事例を用いても，その著者の誤りやすい箇所にしか対応できなくなってしまう．また，大量に用意する必要があるのは内容誤りを含む文章であるが，訂正前の文書だけではどの文章に誤りが含まれているのか判断することが難しいため，訂正前後の文書が必要となる．そこで，複数人で編集が行われており，編集履歴に訂正前後の記事がそのまま残存している Wikipedia の記事をデータ源として使用することにした．

Wikipedia の編集履歴から，訂正前の文書として過去バージョンの記事を，訂正後の文書として最新バージョンの記事を使用する．最新バージョンの記事と同程度の文章数を持っており，できる限り訂正箇所が多い記事を過去バージョンの記事として使用したい．古いバージョンの記事であればあるほど，訂正された箇所は多くなる．しかし，あまりに古すぎると記事として不十分，不適切なものとなってしまう．そのため，図 2 の流れで過去の記事の選出を行う．

(a)　Wikipedia の編集履歴から最新バージョン記事の文章数を求めることによって，選出の基準とする文章数 n を決定する．

(b)　基準とする文章数 n の七割以上の文章数を持つ記事を編集履歴から抽出して，候補記事とする．

(c)　候補記事のうち，最も古いバージョンの記事を過去の記事として採用する．

3.1.2　作成方法

選出した訂正前の文書と訂正後の文書の比較を行う．文章比較時の具体例を表 1 に示す．訂正前後の文章対を比較すると，主に四種類の文章が存在していた．そのため本手法では，入力した文章を表 1 に示す四種類の項目に分類する．訂正前後の文書を比較することによって，訂正前の文章に対してラベル付けを行う．表 1 に示す基準に従って，二種類の文章の比較結果に応じたラベルを付与していく．

まず，訂正前の文章を基準として，一字一句完全一致している文章が最新の記事に存在しているか否かの判定を機械によって自動で行う．それによって，抽出した文章に対してレベル 0 のラベルを付与する．残った文章は何らかの訂正が行われており，訂正前と訂正後の文章が異なる．そして，訂正された内容もそれぞれ異なるため，これ以上機械で判断することは困難で

表 1　分類ラベルと付与時の比較例

ラベル	分類項目	データ作成時の付与基準	訂正前	訂正後
レベル 3	内容についての誤り発生の可能性あり	文章自体の削除または内容について訂正が行われた	ドットハック セカイの向こうに（**2012 年 1 月 21 日公開**） - 岡野智彦 役（声の出演）[37]	ドットハック セカイの向こうに（**2012 年 1 月 26 日公開**） - 岡野智彦 役 [66]
レベル 2	情報の追加が見込める可能性あり	情報の追加・更新が行われた	徳島県立徳島商業高等学校を卒業後、大阪の美術専門学校に通いながらバンド活動を行う	徳島県立徳島商業高等学校を卒業後、大阪の美術専門学校に通いながら**「Ernst Eckmann」で**バンド活動を行う
レベル 1	言い回しや表記についての誤り発生の可能性あり	言い回しや表記のみの変更，また分割や統合が行われた	米津が創造した架空のかいじゅうの**イラスト**を描き、その特徴、習性を紹介するという内容だった	米津が創造した架空のかいじゅうの**イラストレーション**を描き、その特徴と習性を紹介するという内容だった
レベル 0	誤りの発生ほとんどなし	一字一句そのまま残っている	2014 年 6 月に初ライブを行った（ただし、2014 年 5 月にシークレットでライブを行っている）	2014 年 6 月に初ライブを行った（ただし、2014 年 5 月にシークレットでライブを行っている）

ある．そのため，残りの文章に対しては人手にて訂正前と訂正後の文章を比較することが適切であると考えた．人手による比較にて訂正内容を確認することによって，残った文章に対してレベル 3，2，1 のラベルを付与する．ここでは，過去の記事には誤った内容が含まれており，最新の記事には誤りが一切含まれていないという前提の下での比較作業にてラベルを付与している．また，この分類は文章中の内容に関する誤り箇所の推定を目的としているため，文章の比較時に誤字脱字についての訂正は考慮しないことにしている．

3.2　モデル構築

BERT はあらゆる自然言語処理タスクにおいて，汎用性の高いモデルである．モデルの構造を修正せずとも転移学習することによって，様々なタスクに応用でき，高い精度を発揮している．

内容に関する誤りを含む文章は誤字脱字を含む文章と比べて，誤りを含む文章かどうかの判定基準の定義が難しい．それは，事実確認作業にて訂正される文章は内容には誤りがあるが，文法上は正しい文章となっているためである．そこで，分類器として BERT を利用することにより，内容誤りを含む文章の判定を行う．3.1 節のように，内容誤りに関する訂正事例を用いて，内容誤りがある文章を収集する．BERT による分類器を構築する際に，収集した誤り文章を用いてファインチューニングを行うことによって，分類を行うことが可能になると考えた．

本稿では，訓練済み BERT モデルとして，東北大学の乾・鈴木研究室の訓練済み日本語 BERT モデル [1]を使用した．訓練済みモデルは日本語版 Wikipedia にて，事前学習が行われており，語彙数は 32,000 となっている．訓練済み BERT モデル 12 層，これに出力層を 1 層加えた 13 層のモデルを構築する．3.1 節に従って作成したデータセットを用いて，ファインチューニングを行う．ファインチューニング時には，BERT モデルのパラメータは，最終層のみ更新するように設定する．

3.2.1　モデルによる分類

構築したモデルによる分類の流れを示す．まず，与えられた文章を形態素解析して，単語ごとに BERT モデルの語彙に対応した ID 化を行うことにより数値へ変換する．形態素解析には，辞書として mecab-ipadic-NEologd を用いた MeCab を使用している．単語の ID 化は BERT の訓練済みモデルのトークナイザにて行う．ID 化した単語を w_i として，入力文章の単語列 $S = w_1, w_2, \cdots, w_n$ を構築したモデルに入力する．入力後，Embedding レイヤーによって，w_i を対応した単語ベクトルに変換する．ここでは単語ベクトルとして，BERT の事前学習時に得られた 768 次元の分散表現を使用している．その後，BERT を使用した 12+1 層のモデルにて，計算を行う．その結果，4 次元のベクトルである出力 $O = (o_0, o_1, o_2, o_3)$ が得られる．この出力 O に対して，(1) 式のように定義される Softmax 関数を適用することにより．$O' = Softmax(O) = (o'_0, o'_1, o'_2, o'_3)$ を得る．O' の要素 o'_i は，とり得る値の範囲が $0 < o'_i < 1$ となっており，O' のすべての要素を足し合わせると 1 になる．そのため，O' の要素 o'_i は，入力文章 S がラベル (i) に属する確率を表している．よって，入力文章に対する判定は O' の要素のうち，最大値となる要素に対応したラベルを選択することによって決定する．

$$Softmax(x_i) = \frac{\exp(x_i)}{\exp(x_1) + \exp(x_2) + \cdots + \exp(x_n)} \quad (1)$$

3.3　可　視　化

分類時の Attention を可視化することによって，確認箇所として利用する．Attention は，文章などの系列データを扱う際に要素ごとの関係性や注意を向ける箇所を学習する機構であり，BERT にも存在している．この Attention を可視化することによって，入力データのどの部分に注目して分類を行ったのか，という分類時の判定根拠を確認することができる．そのため，文章中に誤りを含むと判定された場合の Attention は，誤りの原因箇所を示していると考える．誤りを含むと判定した根拠や原因を示す箇所の確認作業は重点的に行うべきである．そこで，分類時の Attention の可視化を行い，確認箇所として作業者へ提示することによって，確認作業の実施を促す．

本手法では，確認必要性が高いとされた文章に対して確認箇所の提示を行う．そのため，四種類への分類後にレベル 3，2，1 と判定された文章に対して確認箇所として Attention の可視化をする．内容誤りを含む可能性が高いと判定されたレベル 3

1：https://github.com/cl-tohoku/bert-japanese

の文章は確認作業の必要性が最も高く，誤りがほとんど存在しないと判定されたレベル 0 の文章は必要性が最も低い．確認の必要性が低い文章に対して確認箇所を提示することの意味があまりない．また，レベル 0 においては，Attention が文章中に誤りが存在していない根拠を示すと考えられるため，確認箇所として使用するには不適切である．これらの理由から，レベル 0 とされた文章に対しては確認箇所の提示を行わないこととしている．

3.3.1 可視化方法

Attention の可視化方法について示す．可視化する Attention は，ファインチューニング時に，パラメータ更新を行った BERT の最終層の数値を使用する．BERT の Self Attention は 12 個の Multi head Attention で構成されており，一単語につき 12 個の数値が得られる．これら 12 個の数値は，それぞれが異なる特徴空間にて得られた判定根拠となっている．そのため，得られた 12 個の数値に対して単語ごとに加算平均を計算することによって，確認箇所として利用する．判定結果を表示する際に，Attention の数値に応じて，背景色を単語ごとに変化させる．ある単語の Attention の加算平均結果を A として，背景色を構成する RGB の値をそれぞれ (2) 式のように決定する．

$$\begin{cases} R = 255 \times (1 - A) \\ G = 255 \\ B = 255 \times (1 - A) \end{cases} \tag{2}$$

決定した RGB の値を各単語の背景色に設定して表示する．Attention が強く掛かっている単語については背景色が濃い緑色に，Attention があまり掛かっていない単語については背景色が白色に近くなる．この可視化方法に従って Attention を確認箇所として提示することによって，作業者が確認すべき箇所や内容を直感的に把握できるようにする．

4 評価実験

提案手法について，以下二つの評価実験を行った．

実験 1　提案手法による分類モデルの性能評価
- 目的 1　提案手法によって，どの程度の分類性能を発揮するのか確認する．
- 目的 2　不均衡データの扱い方によって分類性能に変動が起きるのか確認する．

実験 2　提案手法による推定の校閲作業における実用性評価
- 目的 1　提案手法によるモデルにて，適切な判定が行われているか校閲作業の観点で確認する．
- 目的 2　事実確認作業における確認箇所として，適切な箇所の提示が行われているのか確認する．

データセット作成時に，2020 年 10 月 16 日時点を最新記事とした日本語版 Wikipedia 内の記事をデータ源として使用した．Wikipedia 内の記事において，人物記事は閲覧数が多い記事となっており，そのような記事は優先して確認作業を行うべきであると考える．また，使用する記事はできるだけ訂正された箇所が多いものを採用したい．そのため，使用する記事のカテゴリを更新が頻繁に行われている「日本の芸能人」に限定した．カテゴリ「日本の芸能人」に属する記事のうち、閲覧数が多い記事 27 件を使用した．使用した記事内の文章 6,328 文に対し，3.1 節にて示した手順に従って第一著者がラベルを付与し，データセットを作成した．ラベルを付与した結果，ラベルの割合はレベル 3 から順に，9%，6%，28%，56%となっており，ラベルの内訳はレベル 3 から順に，591 件，402 件，1,794 件，3,541 件となった．

4.1 実　験　1

本節では，提案手法による分類モデルの性能評価を評価値に従って行う．また，データの使用方法を変えて実行し，それぞれの評価値を比較することによって，不均衡データの扱い方による分類性能の変動具合の確認をする．

4.1.1 実 験 内 容

まず，作成した前述のデータセットをラベルの割合を保った状態で 10 分割する．その後，訓練用データに対してラベル間の偏りを解消するための処理を行う．基準となるデータ数より多いラベルに対しては余剰データの削除を，少ないラベルに対しては同じデータを追加を行うことによってラベル間のデータ数の偏りを解消した．ラベル間の偏りの解消方法として，以下四種類の方法でそれぞれ 10 分割交差検証を実行し，精度を比較する．

(1)　ラベル間の偏りは考えずにデータをそのまま使用
(2)　データ数を最も少ないラベルに揃えて使用
(3)　データ数を最も多いラベルの半数に揃えて使用
(4)　データ数を最も多いラベルに揃えて使用

(1) から (4) までの訓練データの総数は (1) から順に，5,696 件，1,448 件，6,376 件，12,748 件となり，ラベルの内訳は表 2 のようになった．BERT のファインチューニングのエポック数は，30 エポックを基準として行い，Validation Loss が下がりきっていない場合はエポック数を増やして再度実行するという方針をとった．その結果，(1) から (4) まですべて 30 エポックでの実行となった．それぞれの学習曲線は図 3 のようになった．

4.1.2 結果・考察

交差検証後のそれぞれの評価値を表 3 に示す．Recall，Precision，F 値については，それぞれレベル 3，レベル 2，レベル 1，レベル 0 を正例とした場合の数値を示している．

Accuracy だけに注目した場合，低いものは 0.47，高いものは 0.73 という結果となった．データの使用方法ごとの数値のばらつきは，データ数の違いによる影響であると考えられるた

表 2　訓練データのラベル内訳

使用方法	レベル 3	レベル 2	レベル 1	レベル 0	総数
(1)	531	362	1,616	3,187	5,696
(2)	362	362	362	362	1,448
(3)	1,594	1,594	1,594	1,594	6,376
(4)	3,187	3,187	3,187	3,187	12,748

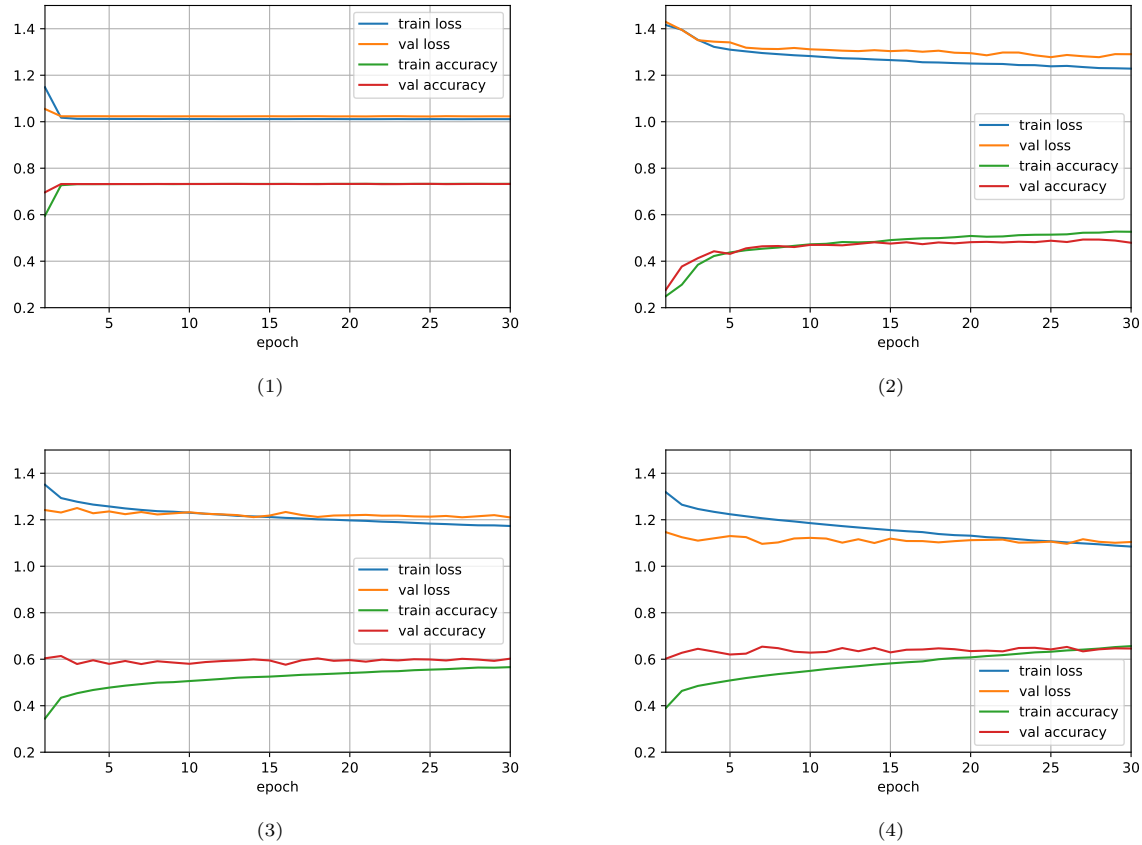

図 3　交差検証時の学習曲線

め，実用に値する結果が出ていると考える．確認作業において，誤りの可能性が低いものが高いと判定されることは問題ではないが，高いものが低いと判定されることは問題である．このような誤りの可能性が高い文章の取りこぼしを少なくしたい．そのため，この分類タスクにおいて重要視するべきは Recall であると考える．確認の必要性が高いレベル 3 やレベル 2 を正例とした Recall は，(2)，(3)，(4) においていずれも 0.3 から 0.4 程度の数値であった．この結果から，誤りの可能性が高い文章の取りこぼしがまだまだ多く発生していることがわかる．

表 3　データ使用方法別の評価値

使用方法	評価値	レベル 3	レベル 2	レベル 1	レベル 0
(1)	Accuracy		0.732		
	Recall	0.000	0.000	0.676	0.966
	Precision	0.000	0.000	0.735	0.731
	F 値	0.000	0.000	0.704	0.832
(2)	Accuracy		0.479		
	Recall	0.248	0.323	0.629	0.716
	Precision	0.494	0.539	0.517	0.452
	F 値	0.322	0.392	0.566	0.550
(3)	Accuracy		0.602		
	Recall	0.301	0.407	0.627	0.721
	Precision	0.305	0.288	0.782	0.655
	F 値	0.299	0.334	0.695	0.683
(4)	Accuracy		0.646		
	Recall	0.306	0.390	0.605	0.752
	Precision	0.239	0.219	0.732	0.803
	F 値	0.257	0.273	0.662	0.774

四種類のデータ使用方法を比較し，不均衡データの扱い方による分類性能の変動具合の確認をする．(1) はデータの偏りが結果に顕著に表れており，特にレベル 3 とレベル 2 を正例とした場合の評価値が極端に低い値になっている．そのため，良い結果とは言い難い．一方，(2)，(3)，(4) におけるレベル 3 を正例とした Recall を比較すると，(2) は相対的に低く，(3) と (4) は同程度の値となっていた．また，Accuracy においては訓練データ数が最も多い (4) が最も高い数値を得られる結果となった．以上の結果より，(4) の最も多いラベルにデータ数を揃えて使用する方法を採用することによって，この四種類の中では総合的に良い精度が得られることが確認できた．

4.2　実　験　2

本節では，校閲作業における実用性を確かめる．提案手法による文章の判定と確認箇所の提示を行い，結果を確認・分析することによって事実確認作業に適用可能かどうかの評価をする．

4.2.1　実 験 内 容

4.1 節の結果から，(4) のデータ数を最も多いラベルに揃えて使用する方法を採用してモデルを作成した．訓練データの総数は 12,748 件，ファインチューニングの最大エポック数は 50 エポックで訓練を行った．また，訓練時に Validation Loss が 15 エポック改善しなければ，Early Stopping を実行して訓練を停止するように設定した．そのため，訓練後のネットワーク

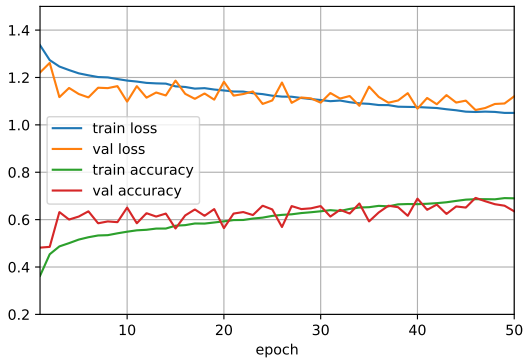

図 4　採用モデルの学習曲線

の重みは Validation Loss が最も低かった時の重みとなる．確認の必要性が高い文章の取りこぼしを少なくするため，いくつか作成したモデルの中で，レベル 3 を正例とした Recall ができる限り高い値となったものを採用した．採用されたモデルは，最大エポックである 50 エポックまで訓練が実行され，訓練後のネットワークの重みは 46 エポック目のものとなった．採用したモデルの学習曲線は図 4 のようになった．採用したモデルによる評価用データ分類時の混同行列を表 4 に，それに基づく評価値を表 5 に示す．表 5 における，Recall，Precision，F 値は，それぞれレベル 3，レベル 2，レベル 1，レベル 0 を正例とした場合の数値を示している．

このモデルを Ruby on Rails にて作成した Web アプリケーションに導入して，文書の判定を行った．レベル 3，2，1 と判定された文章に対しては分類結果だけでなく，確認箇所として Attention も可視化して提示している．ここでは，データセットに含まれていない文書の判定を行い，結果を確認することによって評価を行う．判定を行う文書には、データセットと同様の日本語版 Wikipedia のカテゴリ「日本の芸能人」に属する記事を使用した．

4. 2. 2　結果・考察

判定結果の一部を図 5 に示す．レベル 3 と判定されていた文章の多くが，確認が必要となる文章となっていた．また，レベル 2 と判定されていた文章の多くが，情報の追加・更新を行う

表 4　採用モデルの評価用データ分類時の混同行列

		分類結果			
		レベル 3	レベル 2	レベル 1	レベル 0
ラベル	レベル 3	26	5	9	19
	レベル 2	7	15	9	9
	レベル 1	21	8	119	32
	レベル 0	36	27	13	278

表 5　採用モデルの評価値

評価値	レベル 3	レベル 2	レベル 1	レベル 0
Accuracy		0.691		
Recall	0.440	0.375	0.661	0.785
Precision	0.288	0.272	0.793	0.822
F 値	0.348	0.315	0.721	0.803

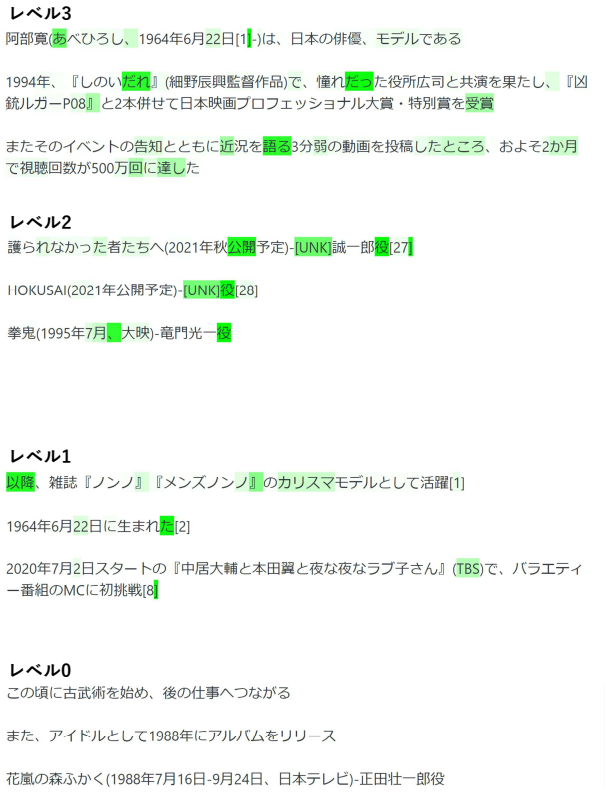

図 5　判定結果の一部

ことができそうな文章となっており，概ね期待通りの結果が得られている．例えば，図 5 のレベル 3 については，生年月日が正しいか，本当にこの賞を受賞したのか，というような事実確認が必要となる．レベル 2 については，日付や役名の追加，「予定」と記述してある箇所の更新などを行うことができそうである．

提示した確認箇所について，レベル 2 においては，多くが日付や役名などの追加や更新ができそうな箇所の前後に存在する結果となっていた．レベル 3 においては，Attention そのままでも確認箇所と見なすことができる箇所も提示されていたが，中には不適切と感じる箇所がいくつか提示されていた．Attention を可視化して，そのまま確認箇所とする方法によって，適切な箇所が提示されることが多かった．しかし，Attention をそのまま提示するとレベル 3 にて見られたような確認箇所として不適切な箇所も中には存在してしまう．不適切な確認箇所となる箇所は，Attention が独立した部分のみに掛かっており，確認内容について断片的な情報しか持たない．そのため，この箇所のみ提示されてもどんな事実について確認をするべきなのか把握することが困難である．改善案の一つとして，Attention とその周辺単語を含めた一定のまとまりを作り，確認箇所として提示することが考えられる．例えば，図 6 の赤枠内の部分において，Attention のみを提示した場合では「22」「受賞」などの単語しか提示されない．このような箇所に対して，Attention とその周辺単語を含めた赤枠で囲ったようなまとまりを作り提

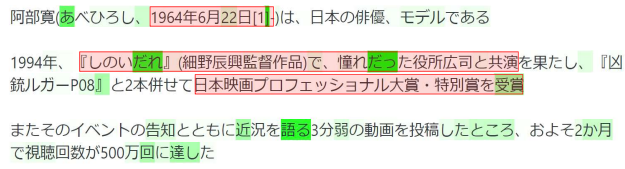

図 6　確認箇所の改善案

示する．周辺単語を含めると情報が補足され，どんな事実について確認をすればよいかが把握しやすくなる．このように，提示方法に一工夫加えることによって，確認箇所として，より良い提示ができるようになるのではないかと考える．

一方で，レベル 1 や 0 には年代などの数字を含む文章がいくつか分類されていた．年代などの数字を含む文章は，事実確認作業において重要視したい文章の一つである．このような結果となった原因の一つとして，訂正事例にて，数字に関する誤りが少なかったことが考えられる．本稿では訂正事例内の誤りの有無を基にして判定を行っているため，このような取りこぼしが発生しているが，数字を含む文章のような一般に誤りやすいとされる文章は重要視したい．そのため，重要視したい文章のうち判定基準がある程度定義できそうな文章に対しては別手法にて判定することも検討するべきと考える．

5　おわりに

本稿では事実確認作業の支援を目的として，訂正事例を用いることによる内容に関する誤り箇所の自動推定手法を提案した．事実確認作業において確認が必要となる文章は内容に関する誤りの発生が見込まれる文章であり，文法上は正しい文章となっている．そのため，誤字脱字を含む文章と異なり，人手にて判定基準を定義することが困難である．そこで，機械学習を利用することにより文章の判定を行うことにした．内容に関する誤りを含む文章が訓練データとして必要となるため，データセット作成に訂正事例を利用した．訂正事例として，Wikipedia の編集履歴から訂正前と訂正後の記事を用意した．二つの文書を比較すると訂正内容ごとに四種類の文章が存在していたため，それぞれの文章に対応したラベルを付与することによってデータセットを作成した．作成したデータセットによって，様々な自然言語処理タスクにおいて汎用性の高いモデルである BERT を用いた分類器を構築した．また，事実確認の必要性が高いと判定された文章に対しては，分類時の Attention をそのまま可視化することによって確認箇所として作業者に提示した．そして，上記手法を Web アプリケーションとして実装し，評価実験を行った．

モデルの分類性能は Accuracy だけに注目した場合，実用に値する性能を発揮している結果となった．また，判定結果を校閲作業の観点で確認したところ，確認すべき文章と判定された文章の多くが確認作業の必要な文章となっていた．Attention をそのまま可視化した確認箇所についても，適切な箇所が提示されることが多かった．これらの結果から，訂正事例を用いた内容誤り箇所の推定を行うことによって，事実確認作業の支援が可能であることが明らかとなった．

しかし，誤りを含む可能性が高い文章の取りこぼし，不適切と思われる確認箇所がいくつか提示されている，数字を含む文章のような重要視したい文章の取りこぼしの発生，など問題点もいくつか見つかった．今後，これらの課題を解決していき，校閲作業における実用性を高めていく．そのために，Attention とその周辺単語を確認箇所とする，重要視したい文章のうち判定基準が定義できそうな文章に対しては別手法にて判定する，などの改善案を適用していく．

また，本稿ではカテゴリ「日本の芸能人」に属する記事に対して提案手法を適用したが，それ以外のカテゴリの文書に対しても適用することによって提案手法の汎用性を確認する必要がある．そして，本稿での実用性の評価は著者の主観によるものが大きいため，第三者がこのシステムを使用した際の評価も取り入れていかなければならない．

謝辞　本研究の一部は JSPS 科研費 18H03342，19H04221，19H04218，および大川情報通信基金の助成を受けたものです．

文　献

[1] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 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)*, pp. 4171–4186, Minneapolis, Minnesota, June 2019. Association for Computational Linguistics.

[2] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. In *Advances in neural information processing systems*, pp. 5998–6008, 2017.

[3] 高橋諒, 蓑田和麻, 舛田明寛, 石川信行. Bidirectional lstm を用いた誤字脱字検出システム. 人工知能学会全国大会論文集, Vol. JSAI2019, pp. 3C4J903–3C4J903, 2019.

[4] 今村賢治, 齋藤邦子, 貞光九月, 西川仁. 識別的系列変換を用いた日本語助詞誤りの訂正. 言語処理学会第 18 回年次大会, pp. 18–21, 2012.

[5] 鈴井克徳, 若林啓. ニューラルネットワークを用いた日本語学習者の文章における不自然箇所検知. 第 10 回データ工学と情報マネジメントに関するフォーラム（DEIM2018）, 2018. online(G3-4).

[6] 内山香, 鈴木海渡, 田上翼, 塙一晃, 乾健太郎, 小宮篤史, 藤村厚夫, 町野明徳, 楊井人文, 山下亮. ファクトチェックのための要検証記事探索の支援. 人工知能学会全国大会論文集 第 32 回全国大会 (2018), pp. 4Pin126–4Pin126. 一般社団法人 人工知能学会, 2018.