学修番号 18860608

修士論文

天気予報原稿のニューラル誤り検出

白井 稔久

2020 年 2 月 21 日

首都大学東京システムデザイン研究科情報科学域

白井 稔久

審査委員：

小町 守 准教授　（主指導教員）

山口 亨 教授　（副指導教員）

高間 康史 教授　（副指導教員）

天気予報原稿のニューラル誤り検出*

白井 稔久

修論要旨

天気予報原稿は一般的に人手で記述されているため，誤りを含んでいる場合がある．それらの誤りは公開する前に校正する必要がある．通常これらの誤りは人手での多重チェックなどで公開前に校正されているが，この校正には大きなコストがかかっている．そこで，本研究ではこの校正のコストを削減するために日本語母語話者が記述したテキストの自動誤り検出器の作成を試みる．

日本語母語話者が記述した誤りがアノテーションされているコーパスには日本語書き言葉均衡コーパス（BCCWJ）などが存在する．しかし，日本語母語話者が記述したテキストの自動誤り検出を行うには，日本語母語話者のテキストに誤りがアノテーションされている数が少なく，教師信号が不足しているため教師あり学習の手法をとることは困難である．一方，日本語学習者が記述したテキストを対象とした研究は盛んに行われている．しかし，日本語母語話者が記述したテキストは，日本語学習者が記述したテキストに比べて誤りが少なく，加えて誤り傾向が異なる可能性があり，学習者の誤り訂正コーパスを今回の学習にそのまま使用することは不適切であると考える．

そこで我々はまず日本語母語話者の誤り傾向を分析するためウェザーニューズの天気予報原稿コーパスを用いた．このコーパスには 2 年分の天気予報原稿があり，編集前の原稿と編集後の原稿がペアになっているため，擬似的に校正コーパスとみなすことができる．このコーパスを用いた分析の結果，誤りの大部分は誤変換，助詞誤り，タイポの 3 つに分類できることが分かった．

そこで，我々は教師あり学習が適用可能な誤変換と助詞誤りに着目しそれらの検出を行うことにした．しかし，それらを教師ありの手法を用いて検出するにはアノテーションされているデータ数が少ない．そこで我々は小規模データでの教師あり

*首都大学東京大学院システムデザイン研究科情報科学域 修士論文, 学修番号 18860608, 2020 年 2 月 21 日.

学習に有用であることが分かっている擬似コーパスでコーパスを拡充するために，擬似誤りを生成した．

擬似誤りの生成は助詞の擬似誤りと誤変換の擬似誤りで分けて生成した．助詞の擬似誤りに関しては，助詞ごとに誤り傾向を分析し，その傾向に従って分布を作り擬似的な誤りを生成した．誤変換の擬似誤りに関しては，分析したコーパス全体で単語単位で誤変換をしている割合を基に，単語を別の同音の単語に変換して擬似誤りを生成した．

次に，我々は最も件数が多かったタイポに着目しそれらの検出を行った．タイポの検出は前述した助詞誤り，誤変換と異なり様々な誤り方が想定されると考え，周辺の単語から特定箇所の単語を予測することができる Bidirectional Encoder Representations from Transformers（BERT）を用いて誤り検出を行う．

本論文の貢献は以下の 3 つである：

- 天気予報原稿の誤り検出のために実際の天気予報原稿のアノテーションおよび分析を行った
- 擬似誤り生成によるコーパスの拡充が天気予報原稿の助詞誤り・誤変換のニューラル誤り検出に有用であることを示した
- 天気予報原稿におけるタイポに対して BERT による検出を行った

本論文では第 1 章に本論文の概要をまとめる．次に第 2 章に本研究に関連する研究を挙げ，本研究との違いについて述べる．第 3 章では分析の対象とした天気予報原稿とその分析結果について詳しく述べる．第 4 章では助詞・誤変換の擬似誤りを用いたコーパス拡充の手法について述べる．第 5 章では BERT による天気予報原稿のタイポ検出の手法について述べる．第 6 章では第 4 章，第 5 章で述べたそれぞれの手法についての実験方法や実験設定，またその実験の結果について述べる．第 7 章では第 6 章の実験結果に対しての考察を行う．第 8 章では本論文のまとめを行い，今後の展望について記述する．

Neural Error Detection for Weather Forecast Manuscript*

Naruhisa Shirai

Abstract

Weather forecast manuscripts are generally written manually and may contain errors. These errors need to be corrected before they can be published. Normally, these errors are corrected before publication by manual multiple checks, but this correction is costly. In this study, we try to create an automatic error detector for texts written by native speakers of Japanese to reduce the cost of correction.

The corpus with annotated errors written by native Japanese speakers includes the Balanced Corpus of Contemporary Written Japanese (BCCWJ). However, for automatic error detection of texts written by native speakers of Japanese, the number of errors annotated in the texts of native speakers of Japanese is small. It is difficult to take the supervised learning method because of the lack of labeled data. In addition, it is considered inappropriate to use the learner's corpus because it has more errors than the text written by the Japanese native speakers and may have a different error tendency.

Therefore, we first use the weather forecast manuscripts from Weathernews Inc. to analyze the error tendency. This corpus has weather forecast manuscripts of two years, and the manuscript before editing and after editing is paired. As a result of the analysis using this corpus, it is found that most of the errors could be classified into three types: mistransformation, particle error, and typo.

*Master's Thesis, Department of Information Science, Graduate School of System Design, Tokyo Metropolitan University Graduate School, Student ID18860608, February 21, 2020.

Therefore, we focus on mistransformations and particle errors to which supervised learning can be applied and detect them. However, the number of annotated data is small to detect them using supervised learning. Therefore, we generate pseudo-errors in order to expand the corpus with pseudo-corpora and find it is useful for supervised learning with small data.

Next, we focus on the typos with the largest number of cases and detect them. We suppose that various types of errors are assumed for typo detection, unlike the particle errors and erroneous conversions described above. Therefore, we tried to detect typos by using Bidirectional Encoder Representations from Transformers (BERT), which can predict specific words from surrounding words.

The contribution of this paper is the three-fold:

- We annotate and analyze actual weather forecast manuscript for error detection of weather forecast manuscript
- This study shows the corpus expansion by pseudo error generation is useful for neural error detection of particle errors and mistransformation of weather forecast manuscripts
- We detect the typo in the weather forecast manuscript using BERT

In this thesis, Chapter 1 summarizes the outline of this paper. Next, Chapter 2 lists the thesis related to this research and describes the differences from this thesis. Chapter 3 details the weather forecast manuscripts and the analysis results. Chapter 4 describes a method of expanding the corpus using pseudo-errors in particle and mistransformations. Chapter 5 describes a typo detection method for weather forecast manuscripts using BERT. Chapter 6 describes the experimental methods and settings for each method described in Chapters 4 and 5, and the results of the experiments. Chapter 7 considers the experimental results in Chapter 6. Chapter 8 summarizes this paper and describes its future works.

目次

図目次

第 1 章　はじめに

誤りが含まれる文章対
編集前
今日は雲優勢のスッキリしない空。
髪が乱れるほどの風が強いのでご注意下さい。
日差しが少なくてもムシ暑くなります。
編集後
今日は雲優勢のスッキリしない空。
髪が乱れるほど風が強いのでご注意下さい。
日差しが少なくてもムシ暑くなります。

誤りが含まれない文章対
編集前
今日も雨が降り続きます。
激しく降る可能性があるので大きめの傘やレインコート・ブーツが良さそうです。
河川の増水・道路冠水・土砂災害にご注意下さい。
編集後
今夜～今日も雨が降り続きます。
激しく降る可能性があるので大きめの傘やレインコート・ブーツが良さそうです。
河川の増水・道路冠水・土砂災害にご注意下さい。

図 1.1　編集前後の文章対の例．下線部が編集された部分である．

天気予報原稿は一般的に人手で記述されているため，誤りを含んでいる場合がある．それらの誤りは公開する前に校正する必要がある．通常これらの誤りは人手での多重チェックなどで公開前に校正されているが，この校正には大きなコストがかかっている．

日本語母語話者が記述した誤りがアノテーションされているコーパスには日本語書き言葉均衡コーパス（BCCWJ）[1] などが存在する．しかし，日本語母語話者が記述したテキストの自動誤り検出を行うには，日本語母語話者のテキストに誤りがアノテーションされている数が少なく，教師信号が不足しているため教師あり学習の手法をとることは困難である．一方，日本語学習者が記述したテキストを対象とした研究は盛んに行われている [2][3][4]．しかし，日本語母語話者が記述したテキストは，日本語学習者が記述したテキストに比べて誤りが少なく，加えて誤り傾向が異なる可能性があり，学習者の誤り訂正コーパスを今回の学習にそのまま使用することは不適切であると考える．

そこで我々はまず誤り傾向を分析するためウェザーニューズ*の天気予報原稿コーパスを用いた．このコーパスには 2 年分の天気予報原稿があり，編集前の原稿と編集後の原稿がペアになっているため，擬似的に校正コーパスとみなすことができる．実際の編集前後の文章対を図 1.1 に示す．このコーパスを分析した結果，誤りの大部分は誤変換，助詞誤り，タイプミスによる誤字であるタイポの 3 つに分類できることが分かった．

そのため我々は教師あり学習が有効であると判断した誤変換と助詞誤りに着目し，それらの検出を行うことにした．しかし，それらを教師ありの手法を用いて検出するにはアノテーションされているデータ数が少ない．そこで我々は小規模データでの教師あり学習に有用である擬似コーパス [2] でコーパスを拡充するために，擬似誤りを生成した．

擬似誤りの生成は助詞の擬似誤りと誤変換の擬似誤りをそれぞれ生成した．助詞の擬似誤りに関しては，助詞ごとに誤り傾向を分析し，その傾向に従って分布を作ることで擬似的な誤りを生成した．誤変換に関しては分析したコーパス全体で単語単位で誤変換をしている割合を基に，単語を別の同音の単語に変換して擬似誤りを生成した．

次に，我々は最も件数が多かったタイポに着目しそれらの検出を行った．タイポの検出は前述した助詞誤りや誤変換と異なり様々な誤り方が想定されると考えた．入力文中の以前の単語・文字から次の単語・文字を予測することができる言語モデルを用いることで．タイポの検出を試みた．しかし，Recurrent Neural Network 言語モデル（RNNLM）を用いた誤りの自動検出 [5] では，誤警報率が非常に高く日本語母語話者の誤り検出が難しいことが報告されている．誤警報率はシステムが陽性であると判断したサンプルのうち，陰性であるものの割合である．そこで本研究では様々な自然言語処理タスクで優れた結果を残している Bidirectional Encoder Representations from Transformers（BERT）[6] を用いて誤り検出を行う．

本論文の貢献は以下の 3 つである：

- 天気予報原稿の誤り検出のために実際の天気予報原稿のアノテーションおよび分析を行った

*https://weathernews.jp

- 擬似誤り生成によるコーパスの拡充が天気予報原稿の助詞誤り・誤変換のニューラル誤り検出に有用であることを示した
- 天気予報原稿におけるタイポに対して BERT による検出を行った

本論文の構成を以下に示す．第 2 章に本研究に関連する研究を挙げ，本研究との違いについて述べる．第 3 章では分析の対象とした天気予報原稿とその分析結果について詳しく述べる．第 4 章では助詞・誤変換の擬似誤りを用いたコーパス拡充の手法について述べる．第 5 章では言語モデルによる天気予報原稿のタイポ検出の手法について述べる．第 6 章では第 4 章，第 5 章で述べたそれぞれの手法についての実験方法や実験設定，またその実験の結果について述べる．第 7 章では第 6 章の実験結果に対しての考察を行う．第 8 章では本論文のまとめを行い，今後の展望について記述する．

第 2 章　関連研究

日本語母語話者が記述したテキストの誤り訂正の研究として，新納ら [7] は平仮名 n-gram を用いて誤りを検出し，訂正する手法を提案した．また，南保ら [3] は文節内の特徴からルールを自動作成し，ルールベースで，日本語の助詞誤りを検出し，校正する手法を提案した．これら 2 つの手法は我々と同じく，日本語母語話者が記述したテキストを対象にしおり，特に南保らの研究とは助詞誤りに着目した点で我々の研究と共通するが，我々の研究では助詞誤りの検出に教師あり学習を用いた点，誤変換も対象とした点が異なる．

また，日本語学習者が記述したテキストに対する自動訂正の研究も広く進められている．今村ら [2] は日本語学習者が助詞を間違えやすいことを指摘し，その助詞を，間違えやすい助詞の単語テーブルを用いることによって修正する手法を提案した．また，今村らは小規模の誤りデータから擬似誤りを生成し，コーパスを拡充した．この研究は助詞に着目した点，擬似誤りを生成した点で我々と共通するが，我々の提案手法では助詞だけではなく，誤変換も対象としている．また我々は日本語母語話者が記述したテキストの誤りについての研究は日本語学習者のものと比較して困難であると判断したため，検出のみで訂正はしない．また，今村らは訂正に Condition Random Field を用いているが，本研究では Bidirectional Long Short-Term Memory（Bi-LSTM）を用いた RNN を使った．

水本ら [4] は，語学学習 SNS である Lang-8*から添削ログを抽出しコーパスを作成し，そのコーパスを用いて，文字単位での修正と，文字-単語間での修正の二つの手法を提案した．彼らが提案した手法は，統計的機械翻訳モデルを用いて誤りを修正するものである．我々の研究とコーパスを作成した点で共通するが，我々の提案手法では擬似誤りを生成してコーパスの拡充を図っている．また，本研究では助詞誤り，誤変換の検出には RNN を用い，タイポの検出には BERT を用いた．

*https://lang-8.com

第 3 章　天気予報原稿コーパスの分析

3.1　天気予報原稿の誤り傾向分析

誤りの傾向を分析するために本研究ではウェザーニューズの天気予報原稿コーパスの分析を行った．このコーパスには 2014 年と 2015 年の 2 年分の天気予報原稿の，編集前の文章と編集後の文章対が入力されている．それらの文章対の総数は 100,931 対である．

これらの文章対における編集の内容は，文法誤りなどの校正だけでなく，原稿内容や表現を大きく書き換えるようなものも含まれる．それらを除外するために，編集前後の文章間の文字単位での編集距離が 1 以上 5 以下の文章対を抽出した．本研究で学習データとして用いた 2014 年のデータを対象に抽出された文章対は 2,575 対で，文対数は 7,765 文対だった．また，天気予報原稿内に長さが 5 文字以下の文が含まれている可能性は非常に低いため，1 文字以上 5 文字以下の編集距離を用いて抽出した文章対に編集前後で文数が異なるものは含まれていないと判断した．そのため，抽出した文章対中の全ての文が，編集前後で 1 対 1 で対応しているものとする．

抽出された文章対に対して，日本語形態素解析システム JUMAN7.0*を用いて文を形態素解析し，編集距離を用いた動的計画法で形態素単位でアライメントをとった．その後，異なる形態素対およびその異なりが含まれる文を人手で確認し，実際に誤りであると判断したものの編集前後の形態素対を記録し分析した．その結果，誤りは大きく分けて誤変換，助詞誤り，タイポの 3 つに分類できることが分かった．

3.2　誤りタグ付きコーパスの作成

3.1 節と同様にして誤りであると判断した形態素対に人手でアノテーションを付与し，学習データとしてコーパスを作成した．誤りにアノテーションする際，以下のルールに基づきアノテーションした．下記のルールに基づきアノテーションの例を図 3.1 に示す．

*http://nlp.ist.i.kyoto-u.ac.jp/index.php?JUMAN

置換
編集前：こまめに 水分 を 取って 熱中 症 対策 を 万全に して ください。
編集後：こまめに 水分 を 摂って 熱中 症 対策 を 万全に して ください。

文の不成立
編集前：室内 でも 熱中 症 に なること が あり 体調 管理 は 万全に。
編集後：室内 でも 熱中 症 に なること が ある 体調 管理 は 万全に。

余字
編集前：今日 の 今日 の 朝 は 雨 の 可能性 が ありますが 〜
編集後：今日 の 朝 朝 朝 は 雨 の 可能性 が ありますが 〜

脱字
編集前：今日 は 夏 空 広がり ます が 急な 雨 も あり ます。
編集後：今日 は 夏 空 が ます が 急な 雨 も あり ます。

図 3.1　ルールに基づいてタグを付与した例．下線部の単語に誤りのタグを付与する．編集後の文は編集前の文中における単語にアライメントを取っている単語の列であるため，実際の文ではない．また，編集前の文中における 1 単語に対して編集後の文中における複数の単語がアライメントを取っている場合，最も文頭に近い単語のみを出力している．

置換　編集前の単語と編集後の単語を置換して文が成立する場合，その単語に誤りタグを付与

文の不成立　編集前の文が成立していない場合に原因と思われる単語に誤りタグを付与

余字　編集前の単語を削除すると文が成立する場合その単語に誤りタグを付与

脱字　編集前の文に明らかに単語が不足している場合，不足していると思われる位置の直後の単語に誤りタグを付与

上記に該当しない単語は誤っていないものとして誤りタグを付与しない

また，上記とほぼ同様の手順で誤りがアノテーションされているテストデータも作成した．相違点は，2014 年のデータではなく 2015 年のデータを対象にしたこと，編集距離が 1 以上 5 以下の文章対ではなく 0 以上 5 以下の文章対を用いたこと，抽出された文章対の中から，季節や時期による文章の内容の偏りを防ぐために各月ごとに 200 文章対ずつランダムサンプリングしたことである．その結果 2,400

表 3.1　学習・開発・評価データの詳細

	学習	開発	評価
誤変換	90	5	2
助詞誤り	76	5	6
タイポ	190	8	20
誤りの総数	356	18	28
総文数	7,765 文	3,842 文	2,971 文

文章対からなる 6,813 文のアノテーション付きのデータを作成した．さらに作成したコーパスを各月から 100 対ずつ，合計 1,200 対ずつ開発データと評価データに分割した．学習データ，開発データ，評価データ誤りの種類毎の件数，総文数を表 3.1 に示す．

第 4 章　擬似誤りコーパスを用いた誤変換・助詞誤り検出

4.1　擬似誤りコーパスの作成

我々は教師信号が少ないコーパスを拡充するために，擬似誤りコーパスを作成した．擬似誤りコーパス内に含まれる擬似誤りは，助詞誤りと誤変換をそれぞれ異なる方法で作成した．擬似的な助詞誤りは実際に天気予報原稿コーパスで誤りのあった助詞を基に作成した．擬似的な誤変換は天気予報原稿コーパスに含まれる誤変換の割合で元の単語を誤変換させて作成した．また，それらの誤変換は元の単語を平仮名にした後に再変換し，元の単語と異なるものに置換することで作成した．

一般的に学習者が記述したテキストの擬似誤りの生成は，実際に誤ったような誤りの生成割合で行われる [2]．しかし，6.1.3 節に示す通り，予備実験によって本コーパスにおいてはこの手法はあまり有用でないことが分かった．そこで，本研究では 3.2 項で作成したコーパスの誤った単語対を基に，一様分布を用いて抽出前の全ての編集後の文章を誤らせることで擬似誤りを生成した．

我々は誤変換はどのような単語に対しても起こりうると考えた．そこで，作成したコーパスの全単語を一定の割合に基づき再変換し，擬似誤変換を生成した．誤変換の作成は単語を京都テキスト解析ツールキット KyTea [8] を用いてひらがなに直し，その後 Google 日本語入力 API *を用いて変換された単語のうちランキング上位 5 件から無作為に元の単語と異なる単語を選び，擬似誤変換を作成した．擬似誤変換の生成の例を図 4.1 に示す．

助詞誤りに関しては，2 つの方法で擬似誤りを生成した．1 つは 3.2 項で作成した学習コーパス中で出現した編集後の各助詞 w について，編集前では誤っている各単語 $w_i = \{w_1, w_2, ...w_L\}$ と元の単語に対して一様に確率を割り当てる一様分布に従い擬似助詞誤りを生成した．また，元の単語を選んだ場合誤りタグは付与しない．擬似助詞誤り生成の例を図 4.2 に示す．

2 つ目は擬似誤変換の生成と同様に一定の割合で助詞誤りを発生させる方法である．全ての助詞に対して一定の割合に基づきその助詞を誤らせるか決め，誤らせる場合学習コーパス内で出現した誤り方から無作為に 1 つを選択し，誤らせる．ま

*https://www.google.com/inputtools/try/

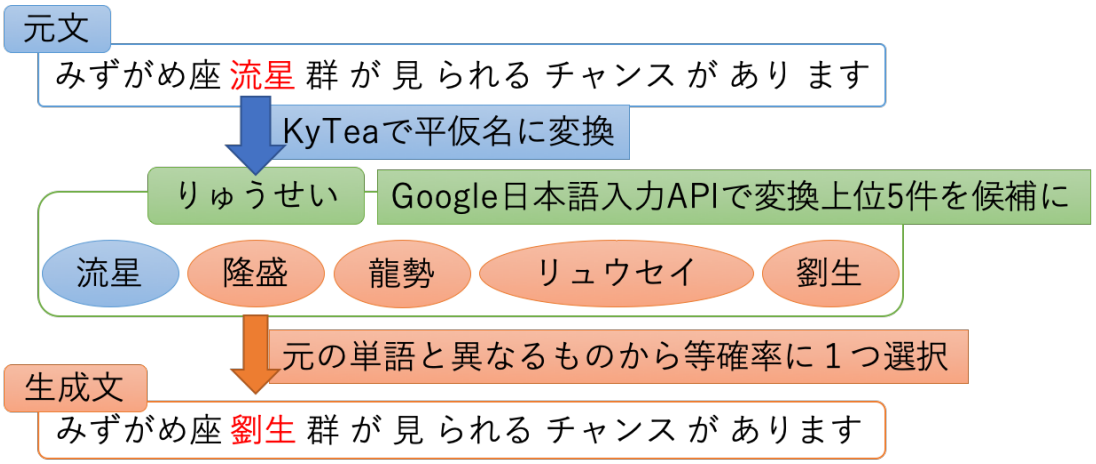

図 4.1　擬似誤変換生成の例

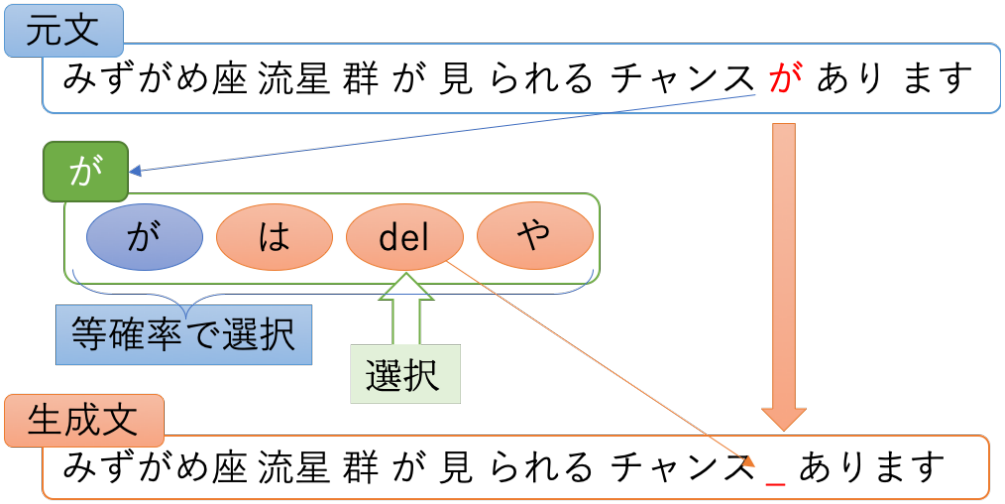

図 4.2　擬似助詞誤り生成の例

た，学習コーパス内で誤らなかった助詞を誤らせることが選択された場合，その助詞を削除することで擬似助詞誤りを生成した．

4.2　助詞誤りと誤変換の検出実験に用いる誤り検出器

本実験では Bi-LSTM [9] を用いた誤り検出器で実験を行う．

入力文 $S = (w_1, w_2, \cdots, w_n)$ の各単語 w_t は単語ベクトル $e_t \in \mathbb{R}^{de \times 1}$ に変換される．n は文長であり，de は単語ベクトルの次元である．単語ベクトルから LSTM により順方向の隠れ層 $\overrightarrow{h_t} \in \mathbb{R}^{dh \times 1}$ と逆方向の隠れ層 $\overleftarrow{h_t} \in \mathbb{R}^{dh \times 1}$ を作成する．dh は隠れ層の次元とする．$\overrightarrow{h_t}$ と $\overleftarrow{h_t}$ を連結することで最終的な隠れ層 $h_t^{(lstm)} \in \mathbb{R}^{2dh \times 1}$ を獲得する．隠れ層 $h_t^{(lstm)}$ を以下のように線形変換しソフトマックス関数を使い正誤タグの確率分布 $p_t \in \mathbb{R}^{tag \times 1}$ を獲得する．tag はタグのサイズであり，正誤のどちらかを予測するためサイズは 2 である．

$$p_t = \mathrm{softmax}(W_h h_t^{(lstm)} + b_h) \tag{4.2.1}$$

$W_h \in \mathbb{R}^{v \times dh}$ は重み行列であり，$b_h \in \mathbb{R}^{v \times 1}$ はバイアスである．v は語彙サイズの次元数である．

誤差関数である $loss$ は交差エントロピーによって以下の式を用いて計算される．

$$loss = -\sum y_t \log p_t \tag{4.2.2}$$

y_t は正解のタグであり学習データ内で正誤のどちらかが付与されている．

第 5 章　BERT を用いたタイポ検出

日本語母語話者が記述した日本語テキストにおける誤り検出において，RNNLM を用いた検出は非常に困難であることが報告されている [5]．RNN 言語モデルは入力文中の t 番目の単語 w_t をそれ以前の単語 $w_{<t}$ を用いて出現確率 $p(w_t|w_{<t})$ の予測をするが，以前の入力 $w_{<t}$ しか用いないため，それ以降の単語 $w_{t<}$ の情報を用いることができない．そのため，文中のある単語に誤りがあった場合，正しくない単語のつながりを考慮して次の単語を予測してしまう．そのため，直後の単語を誤りとして誤検出してしまうことが考えられる．そこで本研究では文中の t 番目の単語 w_t 以外の全ての単語の情報を扱うことができる BERT [6] を用いてタイポの検出を行う．

5.1　Bidirectional Encoder Representations from Transformers（BERT）

BERT は双方向の Transformer [10] を用いたネットワークを用いて，単語予測と隣接文推定の 2 つの事前学習を行う．単語予測では学習データ内におけるトークンの 15% を [MASK] トークンに置き換え，その [MASK] トークンに入る単語を前後の文脈から予測することによって学習を行う．これにより文中におけるトークンの関係を学習することができる．隣接文推定では学習データの一部の文を無作為に他の文に置換し，それらの文が隣接しているか否かを学習する．これにより 2 文間の関係を学習することができる．

5.2　BERT を用いたタイポ検出法

5.1 項で記述した通り BERT の事前学習は [MASK] トークンを文中の他の単語から予測することで行われる．BERT を用いた研究では一般的に事前学習後，他のコーパスを用いて各タスクに適応するための fine-tuning が行われる．しかし，本研究では BERT の単語予測を行う言語モデルの一面に着目し，事前学習モデルをそのまま検出器として用いる．

本研究では入力文における t 番目の単語 w_t を [MASK] トークンに置換し，その [MASK] トークン部の単語の出現確率分布を用いて誤り検出を行う．そうして得た確率分布から [MASK] トークンが元の単語 w_t である確率 $p(w_t|w_{\neq t})$ と閾値 θ を比較し，閾値より低いものを誤りとして検出する．そして全ての単語に対して同様の処理を行い，入力文中の全てのタイポを検出する．

第 6 章　天気予報原稿の誤り検出実験

6.1　擬似誤りコーパスを用いた助詞誤りと誤変換の検出実験

6.1.1　実験設定

表 6.1　実験に用いた各学習データの文数.「orig」は 3.2 節で作成したコーパス,「pp」は助詞の擬似誤りを生成したコーパス,「conv」は誤変換を擬似生成したコーパスを示す.

学習データ	文数
orig	7,765 文
orig+pp	115,744 文
orig+conv	115,744 文
orig+pp+conv	225,305 文

入力は事前学習されている朝日新聞単語ベクトル [11] を用いてベクトル化した．埋め込み層は 300 次元である．本研究で用いる検出器は PyTorch 1.0 で実装し，出力層の値を基に誤っている確率を出力する．この確率が 0.5 を超えているものを誤りとして検出する．隠れ層は開発データを用いた実験により 1 層で 1024 次元に定めた．また，出力層は 200 次元，パラメータの初期化は -0.1 から 0.1 の間でランダムに初期化した．バッチサイズは 64，最適化手法には ADADELTA [12] を用いて学習した．

学習には 3.2 節で作成したコーパスとそれに擬似誤りを生成したコーパスを用いる．また，本章における誤り検出では擬似誤りを加えたことによる有用性を確かめるために助詞誤りと誤変換のみを誤りとして検出する．そのため学習データのタイポの誤りは誤りでないものとして学習した．開発データを用いて，最大で 30 エポック学習し，各エポックで再現率が最大のときに適合率が最も高いエポックのものを評価データの実験に用いた．また各データの文量を表 6.1 に示す．

評価には適合率と再現率を用いた．適合率はシステムが誤りだと判断した単語のうち，実際に誤りである割合である．再現率はコーパス内の検出の対象である全ての誤りである単語のうち，システムが検出した誤りの割合である．また本稿の実験

結果の表では可読性を重視し，適合率，再現率共に 100 倍した数値を表に記載している．

6.1.2 実験結果

表 6.2 誤変換，助詞誤り検出実験の結果．pp*は元の誤り分布に従い助詞誤りを生成したコーパスで学習したものである．@以下の数字は擬似誤りの生成割合である．

学習データ	分割	適合率	再現率
orig	開発	0.00	0.00
	評価	0.00	0.00
orig+pp*	開発	0.00	00.0
	評価	0.00	00.0
orig+pp	開発	0.21	20.0
	評価	0.26	25.0
orig+conv	開発	1.75	20.0
	評価	0.00	0.00
orig+pp*+conv	開発	0.16	20.0
	評価	0.00	0.00
orig+pp+conv	開発	1.38	50.0
	評価	0.29	37.5
orig+pp@50.0%+conv@50%	開発	1.03	70.0
	評価	0.71	50.0

実験結果を表 6.2 に示す．表を見ると擬似誤りコーパスを学習データとして追加すると適合率，再現率ともに上昇していることがわかる．ただ，再現率は全体的に低く，誤りをほとんど検出できていないことがわかる．また，3.2 節で作成したコーパスだけでは学習のための文数が非常に少ないため誤りを検出することができていないことが分かる．

表 6.3　擬似誤りの生成割合を変えた実験結果

		orig+conv		orig+pp	
生成割合	分割	適合率	再現率	適合率	再現率
0.01%	開発	0.00	0.00	0.00	0.00
	評価	0.00	0.00	0.00	0.00
0.1%	開発	3.30	30.0	0.00	0.00
	評価	0.00	0.00	0.00	0.00
1.0%	開発	0.79	20.0	13.3	20.0
	評価	0.00	0.00	16.6	12.5
10.0%	開発	0.90	30.0	8.33	10.0
	評価	0.31	12.5	7.14	12.5
20.0%	開発	0.71	30.0	3.03	30.0
	評価	0.52	25.0	3.09	37.5
30.0%	開発	0.69	30.0	0.77	40.0
	評価	0.49	25.0	1.13	62.5
40.0%	開発	0.74	30.0	1.26	50.0
	評価	0.51	25.0	1.47	62.5
50.0%	開発	1.50	30.0	1.24	50.0
	評価	1.03	25.0	1.52	62.5

また，コーパス内で擬似誤りを生成する割合を変えて実験した結果を表 6.3 に示す．この実験のデータ量は orig+conv，orig+pp と同一である．表を見ると擬似誤りを生成する割合を増やすと再現率が上昇していることが分かる．特に助詞誤りは再現率が著しく向上している．

6.2 BERT を用いたタイポ検出実験

6.2.1 実験設定

本研究では京都大学黒橋・河原研究室が公開している日本語事前学習モデル，pytorch 版 BERT 通常版*を用いた．この事前学習モデルは日本語 Wikipedia コーパス約 1,800 文を JUMAN++2.0.0†を用いて形態素解析し，Byte Pair Encoding [13] を用いて形態素をさらに細かく分割する subword 化したもので学習されている．隠れ層は 12 層 768 次元で 30 エポックで学習し，語彙数は 32,000 である．また実装は PyTorch 1.3.1 で行った．

本実験における比較用のベースラインとして RNNLM を用いた．学習データは 3.1 節で説明した天気予報原稿コーパスの 2014 年のデータを使用し，総文数は 115,744 文である．埋め込み層と隠れ層は 650 次元でパラメータは-0.1 から 0.1 の間でランダムに初期化した. また，dropout の係数は 0.5 で，勾配の大きさが 5 を超えた場合重みを更新せず，バッチサイズは 128 で学習した．最適化には SGD を用いた．語彙数は 2,362 であり，未知語は<unk>として処理した．評価データには 3.2 節で記述したデータのうち開発・評価データの 2 つを使用した．また，本実験ではタイポの検出を目標としているため，データ内のタイポに該当するもののみを誤りとする．また，評価指標には再現率と誤警報率を使用した．誤警報率はシステムが誤りだと検出したもののうち誤りではなかった割合である．

6.2.2 実験結果

BERT を用いた天気予報原稿のタイポ検出実験結果を図 6.1 に示す．図を見ると BERT，RNNLM 共に誤警報率が非常に高く，タイポと通常の単語を区別できていないことがわかる．また BERT を用いた検出結果が RNNLM よりも再現率が低くなっていることが分かる．

*http://nlp.ist.i.kyoto-u.ac.jp/index.php?BERT 日本語 Pretrained モデル
†http://nlp.ist.i.kyoto-u.ac.jp/index.php?JUMAN++

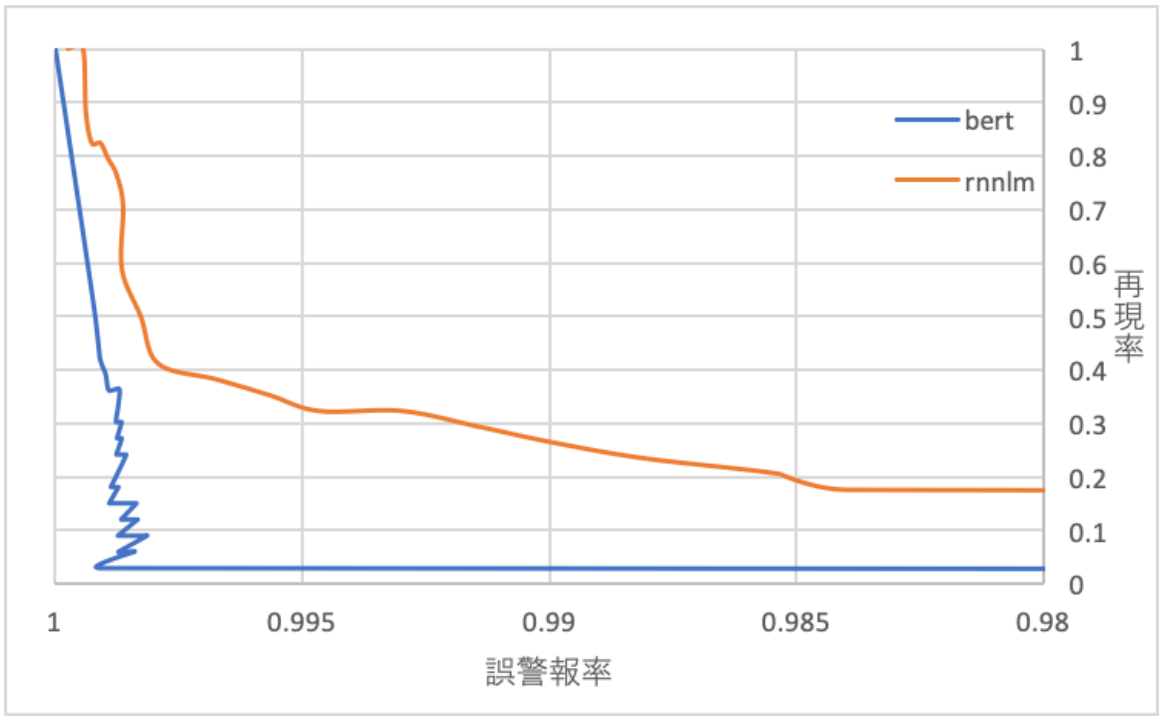

図 6.1　BERT を用いた天気予報原稿のタイポ検出実験結果

第 7 章　天気予報原稿の誤り検出実験結果の分析

7.1　擬似誤りコーパスを用いた助詞誤り・誤変換検出の分析

助詞誤りに関しては，生成割合を一様分布に変えた結果，適合率は変わらず非常に低いが再現率は上がり，一定の生成割合で助詞誤りを生成した結果，適合率と再現率両方の上昇が見られた．このことから助詞の誤り検出に関しては母語話者の誤る割合よりも大きい割合で擬似誤りを生成した方が，適合率と再現率が上昇することが分かった．また，一定の割合で助詞誤りを生成したコーパスで学習したモデルは，“は”や“が”の助詞誤りは検出できる傾向にあった．これは“は”と“が”は元のデータ内でも誤り件数と誤り方の種類も多いため，ほとんどの誤り方を再現できたからではないかと考える．

誤変換に関しては生成割合を上昇させた結果，評価データ内の誤変換は 2 件とも検出できているが，開発データ内の誤変換は 5 件のうち，2 件検出できていない．この検出できなかった誤変換は“うだるような”が“うだる様な”に変換されてしまっているものであった．同様の変換は学習データ内には存在したが，単語分割が開発データでは“よう_な”と分割されていたのに対し，学習データでは“ような”と分割されていたため，異なる誤りになってしまったのが原因であると考える．

助詞誤りの誤検出に関しては“は”，“の”，“を”の誤検出が多く見られた．この 3 つの助詞は今回提案した全ての擬似助詞誤り生成法で誤りとして生成する割合が他の助詞よりも高いため，誤りとして誤検出してしまう傾向にあるのではないかと考える．

誤変換の誤検出に関しては“熱”という単語を多く誤検出してしまう傾向にあった．これは学習データ内の“熱”という単語のほとんどに誤りのアノテーションが付与されており，出現してしまうこと自体が誤りだと認識したためだと考える．

7.2　BERT を用いたタイポ検出実験結果の分析

BERT による検出は RNNLM による検出よりも結果として再現率がより低くなっていたが，これは 6.2.2 節で記述したように RNNLM がより天気予報原稿につ

いて特化した学習データだったためと思われる．BERT による検出では脱字の周辺の単語の出現確率が低くなり，誤検出してしまう傾向にあった．これは BERT の言語モデルとしての学習は，[MASK] トークン部の予測を入力文から行うため，脱字を含む正しい文でない入力文から [MASK] トークンを予測したときに，予測することが困難であったためだと思われる．また，検出できなかったタイポについては隣接する数トークンだけで見たときに違和感はないが，天気予報原稿としてみたときには明確に誤りであるものが多く見られた．

第 8 章　総括

本研究では天気予報原稿の誤り検出において教師あり学習を行うためのコーパスの作成，そのコーパスを拡充するための擬似誤りの生成，BERT を用いたタイポの検出の 3 つを行なった．擬似誤りを生成する提案手法は結果として日本語母語話者が記述した日本語テキストのニューラルネットワークを用いた誤り検出に有用であることが分かった．加えてコーパス中の誤る割合，および誤る分布に従って擬似誤りを生成するよりも擬似誤りの生成割合を上げることによりモデルの性能が向上することが分かった．Bi-LSTM を用いたモデルでの検出は未知の誤りの検出が非常に困難であることが分かった．

また，本論文の擬似誤りを用いた助詞誤り・誤変換の検出法では適合率が低いため，適合率を上げる手法を検討する必要がある．BERT を用いたタイポ検出法においては誤警報率が非常に高いため，BERT や RNNLM などの言語モデルによる検出結果から，さらにタイポであると考えられる単語を抜き出す処理が必要だと考える．

謝辞

学部の頃から手厚くご指導してくださり，困ったときに親切に対応していただいた小町先生，同じく 3 年間共同研究として天気予報原稿コーパスの提供や様々な知見・アドバイスをくださった株式会社ウェザーニューズの萩行様，今回修士論文の審査をしていただける山口先生と高間先生，楽しく接しながら様々なことを教えてくださった同期の皆様，多種多様な催し物や研究会などで共に盛り上がりながら困ったときは様々なアドバイスをくださった研究室の皆様、その中でも学部の頃から面倒を見ていただいた金子さん，大学院まで通わせてくださった家族，様々な人にご協力いただき修士論文を執筆することができました．皆様に深い感謝の意を示し，謝辞とさせていただきます．

参考文献

[1] K. Maekawa, M. Yamazaki, T. Ogiso, T. Maruyama, H. Ogura, W. Kashino, H. Koiso, M. Yamaguchi, M. Tanaka, and Y. Den, “Balanced corpus of contemporary written Japanese,” Language resources and evaluation, vol.48, no.2, pp.345–371, 2014.

[2] 今村賢治，齋藤邦子，貞光九月，西川仁，“小規模誤りデータからの日本語学習者作文の助詞誤り訂正，” 言語処理学会論文誌，vol.19，no.5，pp.381–400，2012． https://ci.nii.ac.jp/naid/10031134262/

[3] 南保亮太，乙武北斗，荒木健治，“文節内の特徴を用いた日本語助詞誤りの自動検出・校正，” 情報処理学会第 181 回自然言語処理研究会，vol.2007，no.94，pp.107–112，2007． https://ci.nii.ac.jp/naid/110006402907/

[4] 水本智也，小町守，永田昌明，松本裕治，“日本語学習者の作文自動誤り訂正のための語学学習 SNS の添削ログからの知識獲得，” 人工知能学会論文誌，vol.28，no.5，pp.420–432，2013． https://ci.nii.ac.jp/naid/130003362344/

[5] 白井稔久，“RNNLM を用いた日本語テキストの誤字・脱字検出および再変換を用いた誤変換検出，” 首都大学東京システムデザイン学部卒業論文，2018．

[6] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, “BERT: Pre-training of deep bidirectional transformers for language understanding,” 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, Association for Computational Linguistics, Minneapolis, Minnesota, June 2019. https://www.aclweb.org/anthology/N19-1423

[7] 新納浩幸，“平仮名 n-gram による平仮名文字列の誤り検出とその修正，” 情報処理学会論文誌，vol.40，no.6，pp.2690–2698，1999．

[8] G. Neubig and S. Mori, “Word-based partial annotation for efficient corpus construction,” Language Resouces Evaluation Conference, pp.2723–2727, 2010.

[9] A. Graves and J. Schmidhuber, “Framewise phoneme classification with

bidirectional LSTM and other neural network architectures," Neural networks, vol.18, pp.602–10, 07 2005.

[10] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A.N. Gomez, L.u. Kaiser, and I. Polosukhin, "Attention is all you need," Advances in Neural Information Processing Systems 30, eds. by I. Guyon, U.V. Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett, pp.5998–6008, Curran Associates, Inc., 2017. http://papers.nips.cc/paper/7181-attention-is-all-you-need.pdf

[11] 田口雄哉，田森秀明，人見雄太，西鳥羽二郎，菊田洸，"同義語を考慮した日本語単語分散表現の学習，" 情報処理学会第 233 回自然言語処理研究会，vol.2017-NL-233，pp.1–5，2017.

[12] M.D. Zeiler, "ADADELTA: an adaptive learning rate method," , 2012. http://arxiv.org/abs/1212.5701

[13] R. Sennrich, B. Haddow, and A. Birch, "Neural machine translation of rare words with subword units," Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), pp.1715–1725, Association for Computational Linguistics, Berlin, Germany, Aug. 2016. https://www.aclweb.org/anthology/P16-1162

発表リスト

[1] 白井稔久，萩行正嗣，小町守，擬似誤りを用いた天気予報原稿のニューラル誤り検出，第 33 回人工知能学会全国大会，2019