D-010

校閲作業のための LSTM による確認箇所抽出
Extraction of Confirmation Points by LSTM for Revision Work

古田 朋也[1)]　鈴木 優[1)]
Tomoya Furuta　Yu Suzuki

1　はじめに

文書を校閲する際に，機械による校閲の補助は有効である．文書の校閲には，誤字脱字の訂正と内容の正しさを確認する事実確認の 2 種類の作業がある．このうち，事実確認作業についての研究は少ない．そこで，本研究では事実確認箇所列挙の自動化を試みる．本稿では，過去の事例に基づく LSTM による分類手法を提案する．評価実験によって，過去の修正結果から誤りの可能性がある文章の抽出が可能であることを明確にした．本手法によって誤りの可能性が高いと判定された文章は，確認作業が必要な文章となっており，目的に対して良い結果が得られた．ところが，事実確認作業において重要とされる数字や固有名詞を含む文章は，事実確認が必要かどうか判定できなかった．そのため今後，数字や固有名詞を含む文章は事実確認が重要だと判定されるような改善が必要であることが分かった．

2　関連研究

校閲作業の補助に関する研究が行われている．高橋ら [1] や今村ら [2] は，校閲作業のうち，誤字脱字の訂正に焦点を当てた手法を提案している．高橋らの研究では，Bidirectional LSTM とランダムフォレストを，今村らの研究では，形態素解析と条件付き確率場 (CRF) を組み合わせた手法が用いられている．一方で，本研究では事実確認作業に焦点を当てた．確認箇所の列挙を自動化し，利用者に提示する．本手順により，先行研究とは異なる作業の補助を行うことが可能であると考えた．

3　分類手法

我々は，事実確認の重要度は文章それぞれにおける修正の必要性から判断できると考えた．与られた文章に対して修正の必要性を確かめるためには，すでに修正されたことがある文章と類似しているかどうかを，教師データとして用意する必要がある．そこで，以下の手順で分類器を構築し，与えられた文に対して修正が必要かどうかを判定する．モデルの構築について図 1 右側に示す．

(a)Wikipedia の編集履歴から過去バージョンと最新バージョンを取得
(b)2 つのバージョンを比較して学習データを作成
(c)学習データで LSTM による分類器を構築・学習

3.1　学習データ

学習データには，2020 年 1 月 26 日時点の Wikipedia ダンプデータによる編集履歴から，過去バージョンの記事と最新バージョンの記事を用いる．過去の記事を選ぶ際に，2 つの条件を考える．一つは，修正箇所が多い古いバージョンであること，もう一つは，最新バージョンと同程度の文章数があることである．よって，次のように選出する．まず，全バージョンの平均文章数を求める．そして，平均文章数を超えた記事のうち，最も古い記事を過去の記事として使用する．過去の記事と最新の記事を比較することによって修正がされた文章を抽出する．

1)　岐阜大学工学部電気電子・情報工学科

表 1　分類項目の具体例

	訂正前	訂正後
1 に該当	主な旧国名は、「飛騨国」「美濃国」であるが僅かに「越前国」「信濃国」の区域も含む。	主な旧国名は、「飛騨国」「美濃国」。わずかに「越前国」「信濃国」**「伊勢国」「尾張国」**の区域も含む。
2 に該当	北部の飛騨地方の大部分は、標高 **3000 メートル**級の飛騨山脈をはじめとする山岳地帯**であり**、平地は高山盆地など**わずかである。**	北部の飛騨地方の大部分は、標高 **3,000m** 級の飛騨山脈をはじめとする山岳地帯**で**、平地は高山盆地など**わずかしかない。**

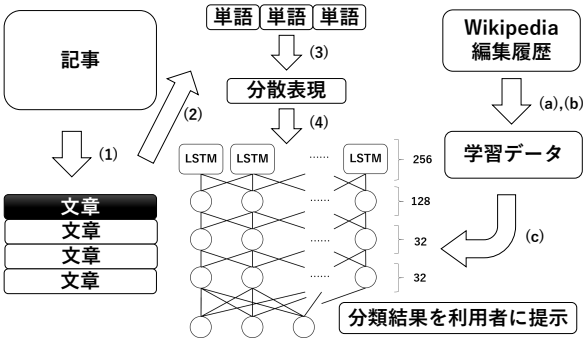

図 1　分類の流れ

ここで得られた文章対を用いて，過去の記事にラベル付けを行う．このとき，以下の 1 と 2 は人手で，3 は自動で分類する．

1. 削除または訂正がされている文章
2. 言い回しのみが変更され，意味が変わらない文章
3. 一字一句そのまま残っている文章

上記の 1 と 2 について，表 1 に具体例を示す．

この分類による事実確認の重要性の高さは上記の 3 種類において 1，2，3 の順となるため，それぞれ重要度大，中，および小と呼ぶ．

3.2　分類

分類の流れを図 1 左側に示す．

(1)与えられた記事を文章単位に分割
(2)形態素解析を行い，基本形で単語単位に分割
(3)単語それぞれを分散表現に変換
(4)構築した分類器へ文章単位で入力し，分類

形態素解析には MeCab を，分散表現への変換には word2vec を使用した．使用した word2vec は日本語 Wikipedia の全文データで学習を行った．

重要度大または中に分類された文章を事実確認の必要性が高い文章として利用者に提示する．

3.2.1 LSTM による分類

LSTM を 1 層含んだ隠れ層 4 層のニューラルネットワークにより分類を行う．日本語の単語は文章によって主語，述語など文章内の役割が異なる．このような単語間関係は分類には重要な情報だと考える．LSTM を用いることで，主語，述語，修飾関係など，文章の文構造を特徴として保持でき，分類に利用できると考えた．

3.2.2 パラメータ調整

ニューラルネットワークの構築において，単語の分散表現における次元数と隠れ層のユニット数についての調整を行った．以下の括弧内のパラメータでグリッドサーチを行い，それぞれの Accuracy と F 値を比較した．

- 分散表現次元数 (300,250,200)
- LSTM 層 (256,128)
- 全結合層 1 (128,64,32)
- 全結合層 2 (128,64,32,16)
- 全結合層 3 (32,16)

グリッドサーチの結果，単語ベクトル 300 次元，LSTM 層 256，全結合層 128-32-32 を採用した．このとき，Accuracy および F 値はそれぞれ 0.85，0.85 となった．

4 実験

上記分類手法の校閲作業における実用性を確かめるため，評価実験を行った．

4.1 データ数

学習用データとして，Wikipedia ダンプデータより，無作為に選出した過去バージョンの記事 25 件，記事内の文章総数 7123 を用いた．これに対し，第一著者が 3.1 節で示した手順に沿って，ラベル付けを行った．その結果，重要度大が 2142，中が 1435，小が 3546 とラベル間の偏りが発生した．偏りを解消するため，重要度大と中に同じ文章を追加することで文章数を揃えた．その結果，文章総数は 10638 となった．このうち 8 割を学習用データ，残り 2 割を評価用データとして使用した．

4.2 分類結果

パラメータの調整後に，改めて分類器の構築を行った．モデルの学習には 60 エポックの学習を実行した．モデル学習後の評価用データによる分類結果を表 2 に示す．表 2 により，Accuracy と F 値はどちらも 0.84 となった．定量的な評価における精度から，過去の修正結果による分類で誤りの可能性がある文章を抽出することができていると考える．

4.3 実例

モデルの評価として，学習用，評価用データとは別の記事を分類器に通して，分類をした．今回は，Wikipedia の「岐阜駅」というタイトルの記事を分類した．その分類結果の一部を表 3，4，5 に示す．

重要度大への分類は，確認作業が必要な文章となっており，期待通りの分類がされている．例えば，表 3 では，本当に利用者が多くなっているのか，本当に現在行われているのか，などの確認が必要となる．ところが，重要度中や小に分類された文章の中には，事実確認が必要なものも存在している．例えば，地名など固有名詞や，年代など数字の入っている文章である．これらは，校閲作業においては確認作業が必要であり，重要度大に分類されるべきである．過去の修正結果を基に行った分類のため，偶然学習データの中に誤りがなかったことなどが原因として考えられる．事実確認作業の補助として，実用性を高めるためには，この点を改善する必要がある．

表 2 評価用データ分類時の混同行列

		分類結果		
		大	中	小
正解ラベル	大	605	43	60
	中	14	646	34
	小	63	124	539

5 おわりに

本稿では，校閲作業における事実確認作業の補助を目的とした確認箇所の抽出を，LSTM を用いて実現できるかどうか評価実験を行った．過去の訂正結果による分類で誤りの可能性がある文章を抽出することができた．また，誤りの可能性が高いと判定された文章は，確認作業が必要な文章となっており，目的に対して良い結果が得られた．現時点の課題として，本来事実確認の必要性が高いとされるべき，固有名詞や数字を含む文章が事実確認が不要なものとして分類されてしまうことが挙げられる．課題解決のため，学習データを増やすほか，固有名詞や数字を含む文章を分類するための手法を組み合わせるなどの改善策を考える必要がある．

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．

参考文献

[1] 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.

[2] 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.

表 3 重要度大と判断された文章

1. 国鉄分割民営化後は名鉄岐阜駅より利用者が多くなっている
2. 高架化により現駅舎の北側にあった旧駅舎が最近取り壊されたことや、周辺ビルの老朽化が著しいことから、現在「岐阜駅北口駅前広場整備計画」に基づいて、北口では大規模な駅前再開発が行われている

表 4 重要度中と判断された文章

1. 今なお東南隣の木曽川駅（愛知県一宮市）までは 7.7km 離れており、この周辺のほかの各駅間よりも距離が開いている
2. 東側コンコースの北口は長良口（ながらぐち）、南口は加納口（かのうぐち）と呼ばれている

表 5 重要度小と判断された文章

1. 岐阜駅（ぎふえき）は、岐阜県岐阜市橋本町一丁目にある、東海旅客鉄道（JR 東海）の駅である
2. 当駅周辺では、高山本線が比較的まっすぐ東進する一方、東海道本線上り側には半径 600m のカーブが存在する