# 機械学習を用いた分割表記文字の判別に関する研究

髙木 健斗　　指導教員：山村 毅

## 1　はじめに

インターネットの普及によりユーザはインターネットメディアでの発信が容易となった．特に SNS では手軽に情報を発信できるため，適時な情報を持つ投稿も存在する．そのため SNS で自然言語を分析することにより，流行の把握や犯罪検知など多岐にわたり応用が可能である．よって SNS 上の投稿を自然言語処理により解析することは重要であると言える．しかし「ググる」や「おいしー」など，流行に応じて新たな語や崩れた表記が含まれていることがある．このような語は自然言語処理による解析が難しいため，特別な処理をする必要がある．

本研究では**分割表記文字**に着目する．分割表記文字とは，崩れた表記の中でも「動」を「重力」と表記するような，1 つの文字を複数の文字に分割する文字を指す．乾ら [1] は文字 N-gram と確率的言語モデルを用いて分割表記文字を判別する**文字 N-gram** モデルを提案した．しかし，文字 N-gram モデルは前後 $N-1$ 文字のみを参照するだけなので，文脈を充分に考慮できないことがある．

そこで本研究では機械学習を用いた分割表記文字の判別手法を提案する．機械学習のモデルである RNN と BERT[2] を用いる．RNN は再帰的なモデルにより対象文字以前の文脈情報を保持し，BERT は双方向 Transformer[3] と呼ばれるアーキテクチャを用いて文全体の情報を保持することができる．広い文脈情報を保持した状態することにより，高精度な分割表記文字の判別を目的とする．

## 2　分割表記文字

分割表記文字は 1 つの文字（**分割可能文字**）を複数の文字で表記するものであり，人の視覚では 1 つの文字として認識できる．しかしコンピュータは複数の文字と認識するため書き手が意図した通り判別することができない．この性質は SNS や掲示板の投稿を検索されないことに利用できる．例えば「死」という文字を「タヒ」と表記し，他者への誹謗中傷にも用いられる場合がある．このような分割表記文字が含まれる文を本研究では分割表記文字文と呼ぶ．

一方，分割表記文字はパターンマッチを行うだけでは解決しない．「タヒチ」のような名詞に対して「死チ」と変換を行ってしまうと意味が通じなくなってしまうためである．

## 3　RNN を用いた判別手法

本研究で使用する RNN を用いたモデル（RNN 分類モデル）の構造を図 1,2 に示す．図 1 に示すように，RNN 分類モデルでは入力文を文字 uni-gram へ変換する．文字 uni-gram は Embedding 層により $H$ 次元のベクトルに変換され，単方向 RNN の場合は $H$ 次元，双方向 RNN の場合は $2H$ 次元のベクトルを出力する．また，図 2 に示すような FFN 層では入力と同じ次元へ，その後 Linear 層を用いて各入力をスカラー値へ変換する．スカラー値を用いてロジスティック回帰により閾値を定め，閾値以上なら 1，閾値未満なら 0 を出力する二値に分類を行う．各文字が分割表記文字なら 1，そうでなければ 0 を出力するモデルである．これにより分割表記文字の位置を特定し，判別

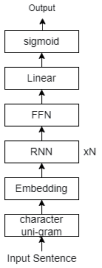


図 1　RNN 分類モデルの外観

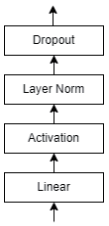


図 2　FFN の構造

も行うことができる．なお本研究では RNN をゲート付き・双方向の RNN を含めて指すものとする．

本研究で利用するモデルのハイパーパラメータは次の通りとする．エポック数は 500，$H$=200，$N$=1，バッチサイズは 128，Dropout の割合は 0.3，最適化関数は Adam，$lr$=1e-6，損失関数は MSE Loss を用いる．

## 4　BERT を用いた判別手法

本研究で使用する BERT を用いた分類モデル（BERT 分類モデル）の構造を図 3,4 に示す．

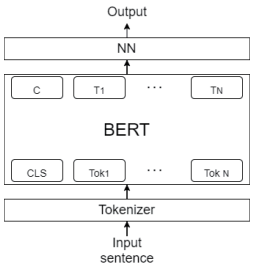


図 3　BERT 分類モデルの外観

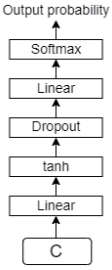


図 4　NN の構造

トークナイザは入力文をトークンと呼ばれる文字列の単位に分け，トークン毎に ID を割り振る．その後，各トークンの ID は BERT 内の埋め込み層によりサイズ $H$ のベクトルに変換される．BERT の出力は各トークン毎にサイズ $H$ の特徴ベクトルが出力され，Linear 層などを利用しタスクに応じた出力へ変換する．

また，BERT の学習は 2 段階で行い，最初に大量の教師なしデータを用いて事前学習を行い表現力を獲得する．次に少量の教師ありデータを用いてファインチューニングする．これにより，事前学習で得られた特徴量をそのまま様々なタスクに応用することができる．

BERT は入力トークンに対応して，その分散表現にあたる大きさ H のベクトル列と [CLS] と呼ばれる特殊な分類用トークンを出力するようになっている．これらのうち，[CLS] トークンの出力を分類モデルへの入力 $C$ として用いる．分類モデルでは出力 $C$ を Linear 層へ入力し，活性化関数として tanh 関数，過学習を防ぐ Dropout を行う．その後，2 つ目の Linear 層により $H$ から 2 次元へ変換し，softmax 関数により文が正解・不正解の確率へ変換し文に分割表記文字文が含まれているかどうかの判定を行う．

さて，図 5 は分割表記文字の判別手法を示したものである．ま

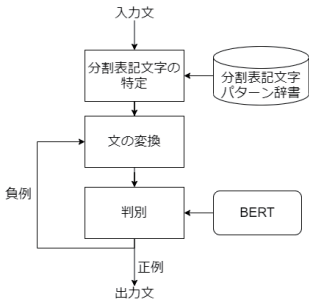


図 5　BERT 分類モデルを用いたシステムの全体像

ず，分割表記文字パターン辞書を用いて入力文内に含まれる分割表記文字を特定する．分割表記文字パターン辞書とは分割表記文字と分割することのできる文字（分割可能文字）が対となっている辞書である．次に分割表記文字パターン辞書の先頭から一致する分割表記文字を検索し分割可能文字へ置換する．再び入力文から分割表記文字パターン辞書を用いて分割可能文字へ置換を繰り返し，全ての分割表記文字文の組み合わせを作成する．BERT 分類モデルは入力された文が負例なら 1，正例なら 0 の値を出力するので分割表記文字文を BERT 分類モデルへ入力し，1 を出力した場合は別の分割表記文字文の入力を繰り返す．ただし，全て出力が 1 の場合は分割表記文字を全て分割可能文字へ変換した文を出力する．

本研究で利用するモデルのハイパーパラメータは次の通りとする．事前学習済みモデルは東北大学の乾研究室が作成したモデル[*1]を使用する．エポック数は 5，バッチサイズは 16，Dropout の割合は 0.3，最適化関数は AdamW を利用する．また，AdamW のパラメータは学習率 $lr$=1e-6, weight_decay=1e-2 である．損失関数は Binary Cross Entropy Loss を用いる．

## 5　実験

### 5.1　実験方法

1991 年から 2010 年の毎日新聞記事を用いてデータセットを作成する．毎日新聞記事に対してクリーニング処理を行う．すなわち本文を抜き出し，文頭に墨付き括弧や大なり・小なり内に見出しが含まれる文は見出しを削除，128 文字以上の文章の場合は句点によって文章長が 128 文字以下になるよう区切る，40 文字未満の文は削除することを順に行い 11837241 文を抽出する．その後，分割表記文字パターン辞書に含まれる分割表記文字 1010 文字から，分割表記文字が 10 回以上出現する分割表記文字の合計 204 文字を抽出する．しかし，分割表記文字が出現する個数は異なるため，「重力」や「メリ」など出現しやすい分割表記文字に対して 100 回以上は出現しづらいようダウンサンプリングを行った．前処理を終えた毎日新聞記事を 26887 文を抽出し，学習：検証：テストを 8：1：1 へ変換する．原文を正解文とし，文中に含まれる分割可能文字を分割表記文字へ全て変換し学習を行う．

RNN はゲート付き RNN である LSTM を用いて，順方向 LSTM と双方向 LSTM の 2 種類で比較する．また，BERT 分類モデルはデータ拡張（Data Augmentation：DA）を行う．学習データ中の文に含まれる分割表記文字の一部を分割可能文字へ変換し 20509 文から 404951 文へ拡張する．その後，本研究で用いる BERT のアーキテクチャ，パラメータは同一で，データ拡張のみ施したモデルを BERT+DA とし比較する．

また，テストデータは毎日新聞だけではなくクリーニング処理を施した 1997 年の読売新聞データ 16833 文，SNS の一種である Twitter[*2]を用いて人手でアノテーションされた 158 文を利用しそれぞれ評価を行う．

### 5.2　実験結果・考察

実験結果は次のとおりである．

表 1　実験結果

| model | 毎日新聞 | 読売新聞 | Twitter |
|---|---|---|---|
| bi-gram | 0.928 | 0.903 | **0.923** |
| 順方向 LSTM | 0.485 | 0.400 | 0.611 |
| 双方向 LSTM | 0.922 | 0.898 | 0.853 |
| BERT | 0.918 | 0.919 | 0.878 |
| BERT+DA | **0.981** | **0.975** | **0.923** |

指標は各テストデータが正解文と一致している正解率である．毎日新聞の判別精度は BERT+DA が最も高く，Twitter データの判別精度は bi-gram が最も高い結果となった．

また，RNN 分類モデルに関しては順方向と双方向で大きな差があることが分かった．そのため双方向が効果的であり，分割表記文字は前後の文脈に依存することが考えられる．しかし，読売新聞や Twitter など文体が異なるものに対しては正解率が低くなっている．

BERT 分類モデルに関しては一部が分割可能文字だったときに判別できないことがあったため，学習データのパターンを増やすことにより結果向上に繋がったと考えられる．Twitter データにおける正解率が低い理由は全ての文に対して 1 を出力した点が挙げられる．学習を正書法に従っている文章に対して 0 を出力するような学習をしたため，崩れた表記に対する文章に 1 を出力してしまったことが考えられる．全ての組み合わせに対し 1 と出力した文が 26 文表れた．

## 6　おわりに

本研究では機械学習を用いた分割表記文字の判別手法を提案した．RNN 分類モデルを用いることにより分割表記文字の位置を特定，検出をすることができ，BERT 分類モデルを用いることでより高い精度で判別することができた．

今後の課題としては本研究で提案した BERT 分類モデルが他の崩れた表記や新語に対してどの程度有効か，他の Transformer や RNN を利用した系列変換による文章生成を利用した場合の有効性，分割表記文字を中心とした周囲の文字列のみ入力をした場合の効果検証などが挙げられる．

## 参考文献

[1] 乾亮, 山村毅: “視覚的「読み」を用いた分割表記文字の処理”, 言語処理学会第 25 回年次大会発表論文集, 名古屋大学, P1-33, pp. 434-437, 2019.

[2] J. Devlin, M. Chang, K. Lee and K. Toutanova: “BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding”, *arXiv preprint arXiv:1810.04805*, 2019.

[3] A. Vaswani et al.: ”Attention Is All You Need”, *arXiv preprint arXiv:1706.03762*, 2017.

[*1] https://huggingface.co/cl-tohoku/bert-base-japanese-v2

[*2] https://twitter.com/