2W-02

# 中国人日本語学習者による日中同形異義語誤り検出方法について

崔　可欣†　　須子　統太‡

早稲田大学社会科学研究科†　　早稲田大学社会科学総合学術院‡

## １．はじめに

外務省所管の独立行政法人・国際交流基金の調査[1]によると，現在中国での日本語学習者は約 100 万人いる．その数は二位のインドネシア 70 万人を大きく差し置いて，世界第一位となっている．その理由の一つとして日中同形語により日本語単語に馴染みがある点が考えられる．中国語と日本語は同じ漢字を使用する言語であるため同形語が多数存在する．その中には，同じ意味を表している同形同義語もあるが，意味の異なる同形異義語も多数存在する．そのため，中国人日本語学習者が日本語文章を作成する際，同形異義語を中国語の意味として用いる誤用が発生する．このような文章は日本人が読むと違和感を感じるが，文法上の誤りがない場合一見して誤っていることが分かりにくい．

本研究では，中国人学習者による日中同形異義語誤りを自動的に検出・訂正を行うシステムを提案する．これにより，中国人日本語学習者の日本語文章作成支援を行う.具体的には事前学習済みの BERT を用いた誤り検出アルゴリズムを提案する．また，評価実験により有効性を検証する.

## ２．関連研究

日本語入力誤り修正システムの研究分野においては，田中らは日本語 Wikipedia の編集履歴から日本語入力誤りデータセットを収集し，BERT を seq2seq の形にした事前学習モデルに基づく訂正システムを構築した[2]．本研究でも入力誤り修正は日中同形異義語誤り修正に近いタスクと考え，ベースラインとして同じくデータセットをファインチューニングし，BERT を用いることで文脈を考慮した日中同形異義語誤り修正実験を行った.

A Method for Detecting Chinese-Japanese Homographs Mistakes among the Chinese learning Japanese language
†Kexin Cui Graduate School of Social Sciences,Waseda University
‡Tota Suko Faculty of Social Sciences,Waseda University

## ３．日中同形異義語誤り検出システム

### ３．１　日中同形異義語誤り混入データの作成

本研究では，よく間違われる日中同形異義語に着目し，東京大学松下達彦研究室が開発した日中対照漢字語データベース［3］に基づいて，よく間違える日中同形異義語 692 個を抽出し，日中同形異義語辞書を作成した．さらに関連研究で作成した日本語入力誤りデータに，日中同形異義語辞書を利用して人工的に日中同型異義語誤りの混入させたオリジナルデータセットを作成する．日中同形異義語誤り検出システムでは，人工的に作成した同形異義語誤り文と修正文のペアを表１のようにトレーニングデータとして用いた.

表１．データセットの具体例

| pre_text | post_test |
|---|---|
| フェリー運航会社や旅行代理店で予定する。 | フェリー運航会社や旅行代理店で予約する。 |
| ウェブ新聞サイトに取り上げがあった。 | ウェブニュースサイトに取り上げがあった。 |
| そのため汽車は運転しない。 | そのため自動車は運転しない。 |

### ３．２　BERT によるファインチューニング

本手法では，上記の誤用文と修正文２種類のテキストデータでファインチューニングを行い，同形異義語の学習を行う．事前学習モデルとして，BERT 日本語 Pretrained モデルを用いる[4]．

### ３．３　誤変換の検出手順

ファインチューニングで得られたモデルに対して，文章の検出・修正手順は図１に示す.

まず，文章を入力として受け，出力された同形異義語は適切な単語に置き換える．置き換える単語候補は，検出された単語をマスク処理し，その単語の周辺情報から，出現確率の高い順に求める．求めた単語候補の中から，最も出現確率の高い単語に置き換える．検出されなかった場合は，そのまま出力する.

より自然な日本語文章を得るため，文章を符号化し，類似度を表す BERT の分類スコアを用いて求める．添削前の文章と添削後の文章間の分類スコアを比較し，分類スコアが基準値より高い場合，一貫性があると判断し出力する．また，

分類スコアが基準値より低い場合，添削前の文に戻し，再添削を行う．初回の添削で求めた適切な単語候補の中から，使用していない単語を選択する．最後に，添削後の文章のテキストデータを出力する．

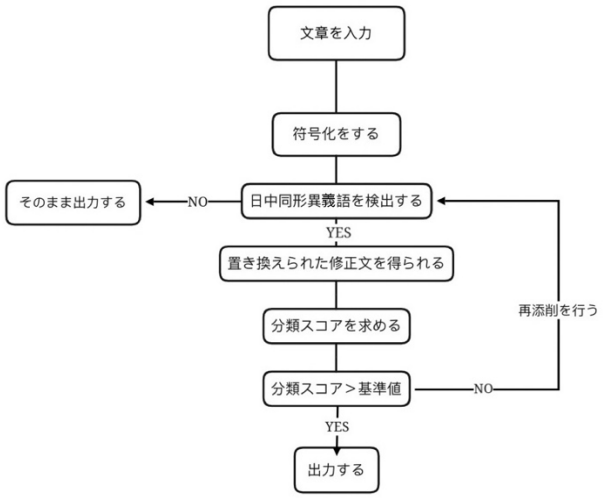


図１．誤変換の検出手順

## ４．評価実験

システムを評価するには日中同形異義語誤りの検出精度の検証と添削後の置き換えた文章の正解率の検証の２種類の実験を行う．テストデータとして誤りなし文混入率２５％と混入率５０％の２パターンのテストデータを使用して検証する．具体的なサンプルサイズは表２に示す．

実験１では，誤り文と誤りなし文を入力し，日中同形異義語誤りの検出再現率と検出適合率で評価する．実験２では，添削後の文章と添削前の文章を比較することで修正成功率を確認する．システム判定結果は表３～４のように示す．実験結果による得られた検出再現率，検出適合率，修正成功率は表５にまとめる．

表２．サンプルサイズ

| | 誤り文 | 誤りなし文 |
|---|---|---|
| 誤りなし文混入率25％ | 150 | 50 |
| 誤りなし文混入率50％ | 100 | 100 |

表３．誤りなし文混入率 25％，システム判定の結果

| | 誤り文 | 誤りなし文 |
|---|---|---|
| 誤りありと判定 | 108 | 4 |
| 誤りなしと判定 | 42 | 44 |

表４．誤りなし文混入率 50％，システム判定の結果

| | 誤り文 | 誤りなし文 |
|---|---|---|
| 誤りありと判定 | 60 | 7 |
| 誤りなしと判定 | 40 | 93 |

表５．実験結果による得られた
検出再現率，検出適合率，修正成功率

| | 検出再現率 | 検出適合率 | 修正成功率 |
|---|---|---|---|
| 正解文混入率25％ | 0.720 | 0.964 | 0.2222 |
| 正解文混入率50％ | 0.600 | 0.896 | 0.2000 |

## ５．考察

まず実験１においては，誤りなし文混入率２５％と混入率５０％のどちらのパターンでも検出再現率が６～７割であるため，提案した検出方法は有効であることを確認した．実験２では，修正成功率が１～２割しかない．修正成功率を上げるためには置き換え候補文字の決め方を工夫しなければならない．つまり，さらにデータセットの拡張が必要である．

## ６．おわりに

本研究では，同形異義語の検出および修正方法について提案した．また，添削後の文章の意味の一貫性を確認し、システムの精度を評価した．評価実験では検出再現率が７割という結果が得られた．また，修正成功率が低かったため，今後についてはデータセットをさらに拡張し，精度の高いシステムを目指す．

## 謝辞

本研究の一部は，日本学術振興会科学研究費基盤研究（C）一般（No.21K11796）により行われた．

## 参考文献

[1]『海外の日本語教育の現状　2018 年度日本語教育機関調査より』
https://www.jpf.go.jp/j/project/japanese/survey/result/survey18.html（参照 2021-12-15）
[2] 田中佑，村脇有吾，河原大輔，黒橋禎夫“日本語 Wikipedia の編集履歴に基づく入力誤りデータセットと訂正システムの改良”，言語処理学会 第 27 回年次大会, pp.1540-1545，2021.
[3]“東京大学 松下達彦研究室，日中対照漢字語データベース”
http://www17408ui.sakura.ne.jp/tatsum/database.html - cvd（参照 2021-11-10）
[4] 柴田知秀，河原大輔，黒橋禎夫“BERT による日本語構文解析の精度向上”，言語処理学会第 25 回年次大会，pp.205-208，2019.