大規模振り仮名注釈付きコーパスを用いた同形異音語の読み分類

佐藤文一[1]　吉永直樹[2]　喜連川優[3,2]
[1]東京大学大学院情報理工学系研究科/国立国会図書館　[2]東京大学生産技術研究所　[3]国立情報学研究所
{fsato0609, ynaga, kitsure}@tkl.iis.u-tokyo.ac.jp

概要

視覚障害者は，漢字交じりの文書を音声で聞いているが，読み誤りが起きると理解が困難になる．このため，漢字の振り仮名の自動付与の精度向上が強く望まれている．本研究では，振り仮名が付与された国立国会図書館の書誌データのタイトルや校正済みの点字データなどを活用し，機械学習のための振り仮名付きコーパスを半自動構築し，書誌のタイトル数約 1650 万，約 3.4 億文字，青空文庫の本数 1944 冊，約 5200 万文字を公開した．このコーパスをもとに 203 語の同形異音語の出現頻度を分析し，その中から東北大学の事前学習済み BERT に含まれる語彙の 93 語，読みの総数 223 個に対して転移学習による読み推定を行い，コーパスの有効性を確認した．

1 はじめに

「読書バリアフリー法」が施行され，誰も取り残されない社会の実現に向けた具体策として，読むことに困難がある児童生徒向けにマルチメディアデイジー教科書が製作・提供されている[1]．これらを制作するために，漢字交じりの文書を正しく読み上げられるよう読みを付与する必要がある．例えば，「表に出る」を「ひょうにでる」と読み上げられると理解が困難になる．視覚障害者が使用する画面読み上げソフトウェアでも同様の現象が起きるが，誤った漢字の読みを繰り返し聞くことは望ましくない．

このような背景から，漢字の振り仮名の自動付与精度の更なる向上が強く望まれている．近年，自然言語処理では，BERT [2]に代表される汎用事前学習済みモデルを用いて，転移学習の枠組みで様々な自然言語処理タスクが低コストで解かれるようになった．しかしながら，同形異音語（読み方が複数ある単語）の読みを分類問題として機械学習で扱うためには，読みに曖昧性がある漢字ごとに正しい振り仮名が付いた学習事例が必要となる．そのため，機械学習の適用に耐えうる規模の言語資源を整備することが依然，求められている．

このような背景から，我々はこれまで校正済みの振り仮名データを利用して振り仮名注釈付き日本語コーパスを半自動構築する取組を進めている[3]．しかし，データサイズとしては，国立国会図書館の提供する書誌データ[4]に含まれる幅広い年代の図書・雑誌のタイトルとタイトルの分かち書きされた振り仮名から作成した「書誌コーパス」（タイトル数 約 1650 万，約 3.4 億文字）に大きく偏っており，多様なテキストに読みを自動付与する学習データとしては不十分であった．

そこで本研究では，これまでに構築した振り仮名付きコーパスのうち，青空文庫[5]の公開作品のテキストデータと視覚障害者情報総合ネットワーク「サピエ」[6]が視覚障害者に提供している点字データから作成した「青空文庫コーパス」について規模を改善する．具体的には，振り仮名を付与する本の数を増やすと共に，見出しを抽出し，テキストを章単位に分割することで，点字の途中にある長文の注解等と青空文庫の文のペアのマッチングミスが伝播することを防ぎ，収集率（注釈が付与できた文字の割合）を改善した．さらに，旧仮名使い，踊り字，大字の漢数字に対応して，漢字振り仮名の語彙数を大幅に増やすことで収集率を改善した．得られたコーパスのうち，収集率が 90%以上の本 1944 冊，約 5200 万文字を青空文庫コーパスとして前述の書誌コーパスと合わせて NDL Lab から公開した[7, 8]．

実験では，本研究で構築したコーパスを用いて BERT による振り仮名の自動付与を行い，コーパスの有用性の評価を行う．具体的に，東北大学の事前学習済み BERT[9]を利用した転移学習により，93 種類，読みの総数 223 個の同形異音語に対して，読み推定の実験を行い，コーパスの有用性を評価した．

2 関連研究

日本語の漢字の読み推定は，形態素解析，仮名漢字変換，音声合成などのタスクと関連して，主に機械学習を適用するための学習データをどう構築するかに焦点を当てて研究が行われている[10, 11 ,12].

羽鳥らは Web から教師なし学習によって獲得した単語・読みのペアと，辞書を用いて，読み推定を行う手法を提案している[10]．高橋らは仮名漢字変換のログを，ノイズを含んだ注釈データとみなして単語分割・読み推定の学習データとして利用する手法を提案している[11]．西山らは同形異音語の各読みに対応する読みの曖昧性のない同義語に注目し，各読みの例文を収集する手法を提案している[12]．これら既存の手法を用いて得られた注釈データはラベルに少なからずノイズが含まれるため，学習ベースで読み推定の精度を改善する際に本質的な限界がある．

読み推定と同様に，単語単位の分類問題として定式化される語義曖昧性解消タスクでは，近年，事前学習済みモデルである BERT の利用による性能改善が報告され始めている[13, 14]．読み推定でも BERT により計算される文脈化単語埋め込みを用いて似た文脈で出てくる単語の類似性を捉えることが有効であると期待される．我々も，これまで少数の同形異音語に対して人手で読み推定の正解ラベルを付与したデータセットを用いて BERT で読み推定の予備実験[15]を行い，その効果を確認したが，学習データの構築コストが課題であった．

本研究では，既存の言語資源を組み合わせて深層学習の適用に耐えうる規模の振り仮名注釈コーパスを半自動構築する．得られたコーパスを用いて実際に BERT による読み推定を行い，その有効性を評価する．

3 振り仮名注釈付コーパスの構築

本節では，文（書）レベルで読みが付与されたテキストを利用して，振り仮名注釈付きコーパスを半自動構築する手法を説明する．具体的には，国立国会図書館の提供する書誌データと青空文庫を用いる．書誌データでは，書誌のタイトル（例えば「吾輩は猫である」）に対し，「わがはいわねこである」のように振り仮名が付与されている．一方，青空文庫では，本全体に対して「サピエ」により点字で読みが付与されている（以下，点字データ）．我々は，漢字仮名交じり文とその分かち書きされた振り仮名のペアを作成し，事前に収集した漢字に対する振り仮名候補に基づく文字レベルのマッチングを行い，本コーパスを構築している．

3.1 文字種毎の振り仮名候補の収集

3.4 節の振り仮名注釈の作成のために，各文字種（漢字，記号，英字，カタカナ，数字）の各単語に対して，辞書をもとに振り仮名候補を収集した．具体的には，下記の形態素解析辞書の「表層」と「読み」から，形態素を単語とみなして単語とその読みを収集した．

- MeCab [16]の IPA 辞書，MeCab-ipadic-neologd [17]，国語研究所の現代書き言葉 UniDic と現代話し言葉 UniDic [18]，sudachi [19]

また作成したコーパスに加えて，約 280 万件の著者名・団体名とその振り仮名のある書誌データからもコーパスを作成し，その振り仮名候補も追加している．結果として，漢字の語彙として 250 万語以上を収集することができた．

3.2 前処理

書誌データは，近代から現在までの幅広い年代に出版された書誌のタイトルを含むため，下記の前処理を行なっている．

- 英数字を半角に，カタカナを全角に正規化．
- 英文と繁体字・Hangul 文字を含むタイトルの除去（日本と中国で共通の漢字だけのタイトルは除去できていない）．
- 旧字体の漢字（472 文字）を新字体に変換．

青空文庫の漢字仮名交じりのテキストに対しては，下記の前処理を行っている．

- 英数字を半角に，カタカナを全角に正規化．
- ルビ，入力注を削除．
- JIS X 0213 の面区点番号を漢字の文字に変換．
- 見出し，ルビとその漢字のデータを収集．

点字データに対しては，次の前処理を行っている．

- 点字の BES，BSE，BET のバイナリーデータを仮名のテキストに変換．
- 旧仮名を新仮名に変換（例: 「くぁれ」->「かれ」）．
- 目次から，本のタイトルとページを抽出．
- インデントから見出しを抽出．
- 表紙，目次，注記，注解，奥付を削除．

3.3 パタンマッチングによる文ペア抽出

青空文庫については文単位で読みとの対応がついていないため，以下の手順で対応する文ペアを抽出した．

まず，前処理で得られた目次の情報から，青空文庫の本と点字の本のペアを，さらに見出しの情報から，章単位で文章と読みのペアを抽出した．最終的に，この章単位のペアから，文単位のペアを次の手順で抽出する．形態素解析を用いて青空文庫のテキストを読みに変換し，点字のテキスト（点字読み）とのパタンマッチングを行う．青空文庫と点字の本は,現代仮名遣いにした訳者や版が異なっていたり，片方だけに注解が含まれていたり，形態素解析器から得られる仮名等の不一致があるが，レーベンシュタイン距離の情報と句読点の位置を手掛かりにペアの文を抽出する．点字の文と対応づけられた青空文庫の読みを元の漢字仮名交じり文に戻すことにより，漢字仮名交じり文と点字の仮名の文単位でのペアが得られる．

3.4 振り仮名注釈の作成

前節までで得られた文と読みのペアについて，漢字仮名交じり文を文字種単位で分割し，各単語に対して振り仮名を付与する．漢字・記号・数字は 3.1 節で作成した振り仮名候補の中に該当の振り仮名があるかを調べ，さらに平仮名等は，1 文字ずつ振り仮名を割り当てる．書誌データでは「底力」が「そこじから」と記述され，点字データでは「京都へ」が「きょーとえ」と記述されるように，分かち書きされた振り仮名は，現代仮名遣いと一部異なっているため[20, 21, 22]，表層と読みが一致しない平仮名（例：「へ」と「え」，「は」と「わ」等）の対応付けを行っている．更に次の処理を行っている．

- 漢数字の大字（例:「弐萬」と複数読み（例：「一二」「十二三」）への対応．
- 繰り返し文字（ゝヽゞヾ）・踊り字に対応．
- 漢字(単語)の部分は，部分文字列の振り仮名候補辞書から生成したラティスのグラフ探索（深さ優先探索）で処理．
- 漢字の送り仮名で，「何にでも」の「何」が「な」「なに」のように漢字の読み候補だけから読みが確定できない場合に，後続する文字も考慮した読みの対応付け．

表 1 青空文庫コーパスの作家数等

	公開前[3]	公開時[7]	公開後
作家数	14	115	120
作品数	224	1944	2044
文字数	1784 万	5162 万	5453 万

以上により構築された振り仮名付きテキストに対して，コーパス特有の後処理を行い，収集率を改善した．具体的には，書誌コーパスに対しては，「コ-ヒ-」->「コーヒー」のように，'-'を長音'ー'に変換している．青空文庫コーパスの点字データの作成時期が古いデータは，データ構造が統一されていないため，目次や注記等の修正・削除の一部を手作業の後処理で行っている．

後処理で数箇所修正することにより，収集率が 100%になる点字の本, 約 450 冊に対して修正を行い，コーパスの注釈が正しく付与されていることを確認した．

コーパスの公開後もこの節の手法で拡張しており，青空文庫コーパスの 90%以上の収集率での公開時と公開前後での作家数等を示したのが表 1 である．次節の実験はこの公開後のデータを用いて行っている．更新したコーパスは今後，[8]で公開予定である．

4 読み分類実験

本節では，前節で得られた本コーパスから，同形異音語の出現数を調査した結果と，同形異音語の読みのクラス分類の実験結果を報告する．

4.1 同形異音語の出現数の調査

対象の同形異音語として，書誌データの「文字・読みの基準」[21]を参考にして 203 種類[付録]を選び，その読みの出現数を調査した．「国立駅」のように「国立」と「駅」からなる複合語は基本的に，読みが一つに確定するので，出現数から除外した．

表 2 のドメインの「書誌」「青空」は，該当のコーパスを示している．「東北大 v1・v2」は，東北大学の 2 種類の事前学習済み BERT のどの語彙に含まれているかを示している．

最大の出現数の単語は，「変化」の 88,322 個で，最小は「日供」の 0 個，出現数が 30 個以上の単語は 197 個であった．頻度にばらつきはあるが，全体的には機械学習のデータ数としては十分と思われる．

表 2 同形異音語の出現数

同形異音語	東北大 V1・v2	合計	読み0	書誌	青空	読み1	書誌	青空
変化	v1v2	88322	へんか	86365	1612	へんげ	281	64
市場	v1v2	85723	いちば	592	179	しじょう	84899	53
国立	v1v2	19445	こくりつ	19178	24	くにたち	243	0
口腔	v1v2	12051	こうこう	6459	16	こうくう	5573	3
表	v1v2	6052	おもて	544	2829	ひょう	2679	0
大分	v1v2	4421	だいぶ	7	1079	おおいた	3318	17
競売		1253	きょうばい	305	8	けいばい	938	2
礼拝	v1v2	944	れいはい	780	85	らいはい	12	67
後世	v1v2	743	こうせい	486	226	ごせ	4	27
日供		0	にちぐ	0	0	にっく	0	0

また，各読みの出現数はドメインによって大きな差があることが確認された．青空文庫コーパスの出現数の少ない読みとしては，国立（こくりつ，くにたち），表（ひょう）がある．ちなみに「国立駅は大正 15 年開業」である．書誌コーパスの出現数の少ない読みとしては，大分（だいぶ）であった．一方，書誌コーパスは，「競売（けいばい: 法律用語）」，「口腔（こうくう: 慣用読み,医学）」，「礼拝（らいはい: 仏教・神道）」，「現世（げんせ: 仏教）」の出現頻度から，用途別の読みもある程度カバーしていることを確認した．用途については[21]を参考にした．

4.2 同形異音語の読みのクラス分類の実験

東北大学の事前学習済み BERT（v2）を使用して，同形異音語の読み推定をクラス分類として定式化して以下の方法で分類器の学習を行った．

BERT の転移学習を使う手法である．token 数は学習最大 128 に制限した．学習，推論は，huggingface_transformers[23]の例を参考に，固有表現，品詞分類で使われる token classification で，複数の同形異音語の読みのクラス分類を行った．各単語の読み毎にラベルを割り当てている．例えば，表（ひょう），表（おもて），角（かく），角（つの）を，それぞれ 1, 2, 3, 4 のラベルを割り当てる．該当しない単語は，ラベル 0 を割り当てる．BERT に token 列とラベル列を入力し，推論の出力値の最大のインデックス（ラベル）を得ることにより，同形異音語の読みの予測値が得られる．書誌と青空文庫コーパスを結合し，得られたコーパスを学習・開発・テストデータとして 6:2:2 の割合で分割し，かつ，各読みも同じ割合で分割して実験を行った．

表 3 書誌・青空文庫コーパスでの読み分類結果

同形異音語	読み0	読み 1	出現数 0	出現数 1	正解数 0	正解数 1	外れ	accuracy	macro f値
大分	だいぶ	おおいた	218	664	216	663	0	0.997	0.995
身体	しんたい	からだ	4016	847	3998	770	0	0.98	0.965
一目	ひとめ	いちもく	335	49	332	36	0	0.958	0.897
心中	しんちゅう	しんじゅう	59	345	51	336	0	0.958	0.916
表	おもて	ひょう	662	526	603	522	5	0.947	0.947
玩具	おもちゃ	がんぐ	52	280	47	266	0	0.943	0.899
博士	はくし	はかせ	3585	535	3374	479	3	0.935	0.872
礼拝	れいはい	らいはい	174	17	168	9	0	0.927	0.761
故郷	こきょう	ふるさと	784	106	755	28	0	0.88	0.639
今日	きょう	こんにち	3682	1471	3403	1045	0	0.863	0.827
現世	げんせい	げんせ	36	49	25	48	0	0.859	0.848
金色	きんいろ	こんじき	200	104	197	57	0	0.836	0.791
上方	かみがた	じょうほう	291	128	238	112	2	0.835	0.819
口腔	こうこう	こうくう	1300	1113	1000	873	5	0.776	0.776

表 3 では，書誌と青空文庫の両方のコーパスを結合して使用している．同形異音語の数は，93 種類[付録]で，読みのラベルの総数は 223 個である．表 3 は，その抜粋である．表 3 の「正解数 0」は，「読み 0」の「出現数」の中で，予測が正解になった数が「正解数 0」である．表 3 の「外れ」は，予測した読みのラベルが，ラベル 0 になった個数である．この「はずれ」の数が少ないので，全体からは除外した．

表 3 の結果から，「大分」「表」のように意味の異なる読みに対しては，読みのクラス分類ができていることを確認した．

5 おわりに

本論文では，振り仮名注釈付きコーパスを用いて，203 種類の同形異音語の出現数を調査し，その中の 93 種類に対して読みのクラス分類を BERT により転移学習により行った．得られた結果から，本コーパスは意味の異なる読みを区別するのに対して有効であり，一度に複数の同形異音語を学習・推論できるのは大きな利点であることを確認できた．今後は，読みの分類の精度の改善，事前学習済み BERT に含まれない語彙に対しても読みの分類ができる手法の検討，また，ドメインを拡張も行いたいと考えている．今後も視覚障害当事者の観点から，視覚障害者の情報障害の課題に取り組んでいきたい．

謝辞

青空文庫コーパスの作成と公開にあたっては，全国視覚障害者情報提供施設協会及び日本点字図書館のご理解とご協力を賜りました．この場を借りて御礼申し上げます．

参考文献

1. 令和 3 年版障害者白書（全体版）- 内閣府 https://www8.cao.go.jp/shougai/whitepaper/r03hakusho/zenbun/index-pdf.html
2. Devlin, Jacob, et al. "Bert: Pre-training of deep bidirectional transformers for language understanding." arXiv preprint arXiv:1810.04805（2018）.
3. 佐藤文一, 吉永直樹, 喜連川優. "書誌データ・青空文庫・点字データを用いた振り仮名注釈付き日本語コーパスの構築." 研究報告アクセシビリティ（AAC）2021.13（2021）: 1-8.
4. 国立国会図書館サーチが提供する OAI-PMH https://iss.ndl.go.jp/information/api/api-lists/oai-pmh_info/
5. 青空文庫　Aozora Bunko https://www.aozora.gr.jp
6. サピエとは https://www.sapie.or.jp/sapie.shtml
7. 国立国会図書館の書誌データから作成した振り仮名のデータセット https://github.com/ndl-lab/huriganacorpus-ndlbib
8. 青空文庫及びサピエの点字データから作成した振り仮名のデータセット https://github.com/ndl-lab/huriganacorpus-aozora
9. GitHub - cl-tohoku_bert-japanese BERT models for Japanese text https://github.com/cl-tohoku/bert-japanese
10. 羽鳥潤, 鈴木久美. "機械翻訳手法に基づいた日本語の読み推定." （2011）.
11. 高橋文彦, 森信介. "仮名漢字変換ログを用いた単語分割・読み推定の精度向上." 研究報告自然言語処理（NL）2014.15（2014）: 1-10.
12. 西山浩気, 山本和英, 中嶋秀治. "読み曖昧性解消のためのデータセット構築手法." 人工知能学会全国大会論文集 第 32 回全国大会 (2018). 一般社団法人 人工知能学会, 2018.
13. 曹鋭, et al. "BERT を利用した教師あり学習による語義曖昧性解消." 言語資源活用ワークショップ発表論文集= Proceedings of Language Resources Workshop. No. 4. 国立国語研究所, 2019.
14. 新納浩幸, 馬ブン. "BERT の Masked Language Model を用いた教師なし語義曖昧性解消." 言語処理学会第 27 回年次大会発表論文集（2021）: 1039-1042.
15. 佐藤文一, 喜連川優. "事前学習済み BERT の単語埋め込みベクトルによる同形異音語の読み誤りの改善（福祉情報工学）." 電子情報通信学会技術研究報告= IEICE technical report: 信学技報 119.478（2020）: 17-21.
16. MeCab: Yet Another Part-of-Speech and Morphological Analyzer https://taku910.github.io/mecab/
17. GitHub - neologd/mecab-ipadic-neologd: Neologism dictionary based on the language resources on the Web for mecab-ipadic https://github.com/neologd/mecab-ipadic-neologd
18. 「UniDic」国語研短単位自動解析用辞書 最新版ダウンロード https://ccd.ninjal.ac.jp/unidic/download
19. GitHub - WorksApplications_Sudachi A Japanese Tokenizer for Business https://github.com/WorksApplications/Sudachi
20. 書誌データ Q&A｜国立国会図書館―National Diet Library https://www.ndl.go.jp/jp/data/faq/index.html
21. 文字・読みの基準｜国立国会図書館―National Diet Library https://www.ndl.go.jp/jp/data/catstandards/characters/index.html#yomi
22. 『点訳のてびき 第 4 版』（特定非営利活動法人 全国視覚障害者情報提供施設協会，2019 年 2 月発行）
23. GitHub - huggingface_transformers　Transformers State-of-the-art Machine Learning for Pytorch, TensorFlow, and JAX https://github.com/huggingface/transformers

付録

出現数を調査した同形異音語(203 種類)

1.**東北大事前学習済み** BERT v1・v2 **両方の語彙**(87)

'表', '角', '大分', '国立', '人気', '市場', '気質', '役所', '上方', '上手', '下手', '人事', '金星', '仮名', '内面', '礼拝', '遺言', '口腔', '後世', '骨', '一途', '一言', '最中', '一目', '係', '足跡', '今日', '明日', '生物', '変化', '大事', '水車', '一見', '一端', '大家', '心中', '書物', '一角', '一行', '一時', '一定', '一方', '一夜', '下野', '化学', '火口', '花弁', '玩具', '強力', '金色', '経緯', '故郷', '紅葉', '行方', '根本', '左右', '山陰', '十分', '上下', '身体', '水面', '世論', '清水', '大手', '大人', '大勢', '中間', '日向', '日時', '夫婦', '牧場', '末期', '利益', '工夫', '一味', '魚', '区分', '施行', '施工', '転生', '博士', '法華', '真面目', '眼鏡', '文字', '文書', '律令'

2.**東北大事前学習済み** BERT v1 **の語彙**(10)

'教化', '見物', '清浄', '谷間', '追従', '墓石', '大文字', '漢書', '作法', '兵法'

3.**東北大事前学習済み** BERT v2 **の語彙**(6)

'現世', '日中', '夜中', '前世', '二人', '立像'

4.**東北大事前学習済み** BERT **に含まれない語彙**(102)

'大人気', '半月', '黒子', '外面', '競売', '開眼', '求道', '血脈', '施業', '借家', '頭蓋骨', '法衣', '昨日', '氷柱', '風車', '寒気', '背筋', '逆手', '色紙', '生花', '白髪', '貼付', '一回', '一期', '一月', '一所', '一寸', '一声', '一石', '一日', '一分', '一文', '一片', '何時', '何分', '火煙', '火傷', '火床', '火先', '火筒', '芥子', '気骨', '銀杏', '元金', '五分', '後々', '後生', '御供', '細々', '細目', '三位', '疾風', '菖蒲', '世人', '世路', '船底', '早急', '相乗', '造作', '他言', '東雲', '頭数', '二重', '日供', '日次', '日暮', '日来', '梅雨', '風穴', '仏語', '分別', '面子', '木目', '目下', '夜直', '夜来', '夜話', '野兎', '野馬', '野分', '野辺', '野面', '野立', '冷水', '連中', '飛沫', '翡翠', '餃子', '一足', '意気地', '一昨日', '一昨年', '十八番', '十六夜', '明後日', '石綿', '公文', '読本', '仏国', '古本', '町家', '遊行'