プロンプトと複数の音声認識候補による青空文庫振り仮名注釈付き音声コーパスの再構築

佐藤文一[1]　吉永直樹[2]　豊田正史[2]　喜連川優[3,4]
[1]国立国会図書館 [2]東京大学生産技術研究所 [3]大学共同利用機関法人情報・システム研究機構 [4]東京大学
f-sato@ndl.go.jp, {ynaga, toyoda, kitsure}@tkl.iis.u-tokyo.ac.jp

概要

著者らは漢字の読み推定の学習・評価用コーパスとして「青空文庫振り仮名注釈付き音声コーパス」を構築し，2024 年 1 月に国立国会図書館 NDL ラボから公開している．音声コーパスの構築においては音声と元テキストとの間の対応付けの成功率が最終的なコーパスのサイズに影響を及ぼす．本研究では対応付けの成功率を高めるため，OpenAI の Whisper において，専門用語等の未知語の音声認識を高める目的で使用されるプロンプトを追加し，認識精度の改善を図ることで，コーパスの拡張を試みた．結果として 7604 万文字，4617 時間の振り仮名注釈付き音声コーパスの構築に成功した．本コーパスは近日の公開を予定している．

1 はじめに

近年の情報アクセシビリティ関連の法律[1, 2]の施行により，従来の障害者に加えて高齢者，読み書きの困難者及び日本語の学習者に対しても配慮を求められるようになってきている．対応すべき情報アクセシビリティの中には，書かれている内容を読み上げソフトウェア等が正しく読めるようにすることが含まれる．こうした「読み」のアクセシビリティの向上のためには，より正確な読みを付与できる機械学習モデルを開発したり，読みの正確性を検証したりするための注釈付きデータセットが必要である．

筆者らは，OpenAI が 2022 年に公開した高性能な音声認識モデル Whisper [3]を使って，視覚障害者情報総合ネットワーク「サピエ」[4]が視覚障害者に提供する音声デイジーの xml データと青空文庫[5] の公開作品のテキストデータから振り仮名注釈付き音声コーパスを構築している．具体的には，音声データを文単位で分割して青空文庫のテキストデータ中の文への対応付けを行い，複数の音声認識候補と，単語の読み候補を相互に参照して読みの推定を行う．結果として，漢字の読みの推定情報を持つ，合計 5545 万文字，3520 時間の振り仮名注釈付き音声コーパスを構築した[6, 7].

本研究では，人名や専門用語等の未知語をプロンプトとして与えることにより音声認識精度を高める手法[8]を用いて，この青空文庫振り仮名注釈付き音声コーパスの作成元データの音声認識精度を高め，振り仮名注釈付き音声コーパスの拡張を試みる．音声認識精度を改善することで，より多くの文について青空文庫のテキストデータと対応づけることができるため，結果として読みを推定できる文を増やすことができ，コーパスの拡張に繋がる．

具体的には，まず，以前行った研究の後で公開された最新の音声認識モデル（large-v3）を用いて，音声認識を再実行し，音声認識を行った結果と青空文庫のテキストの対応を取り，読みの推定を行った．次に，読みが推定できなかった連続した音声認識結果をそれぞれまとめ，対応する青空文庫のテキストを求めた．Whisper のプロンプトとして与えられるトークンの数には制限があるため，テキストが短いときはそのまま与え，長いときは，青空文庫のテキストと音声認識結果を比較して誤りの多い形態素の上位をプロンプトとした．

これらの手順によってコーパスの再構築を行った結果，含まれる録音時間が 3520 時間から 4617 時間へと増加した（本 3686 冊，7604 万文字分）．これらを新バージョンの青空文庫振り仮名注釈付き音声コーパスとして国立国会図書館 NDL ラボ[i]から近日公開する予定である．

[i] https://github.com/ndl-lab

2 関連研究

漢字の読み推定を機械学習で扱うためには，漢字ごとに正しい読みの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求められている．

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，学習データの構築に焦点を当てて研究が行われてきた[9, 10, 11]．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，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ている[12, 13, 14]．近年では，音声コーパスや音声データを使用して，読みに関する研究が行われてきている．小林らは，「日本語話し言葉コーパス」の訓練データの不足を，疑似訓練データを追加することにより，同形異音語の読み推定での有効性を調査している[15]．増山らは，「国会審議映像検索システム」により，同期された音声とテキストから，同形異音語の出現数を分析している[16]．Yin らは，テレビの字幕と音声データから，約 3.5 万時間の大規模な音声コーパスを構築して公開している[17]．

音声認識の精度向上の一つの手法として，人名や専門用語等の未知語をプロンプトに追加する手法が提案されている．Whisper では話者分離に利用する研究[18]もあり，また OpenAI からプロンプト利用のためのガイドも公開されている．

筆者らは，振り仮名注釈コーパスを構築し，複数の同形異音語の読みの推定を行っている[19, 20]，また音声データから音声認識により，振り仮名注釈コーパスの構築を試みている[6]．

3 予備実験

Whisper には 1 つ前のセグメントの音声認識結果であるテキストを文脈情報としてプロンプトで考慮して音声認識を行う機能があるが，このプロンプトには実際は任意のテキストを追加することができる（以下，「ユーザープロンプト」という）．本研究においては，音声に対応するテキストを青空文庫から取得することができるため，取得した情報をユーザープロンプトとして入力することで音声認識精度をさらに改善することができる．

音声認識精度の改善の読み推定への効果を確かめるため，下記の予備実験を行った．予備実験用音声データとして，レアゾン・ヒューマンインタラクション研究所が公開している字幕放送から音声と字幕を抽出した音声コーパスである ReazonSpeech [17]を利用した．音声認識の精度評価は一般的には文字誤り率や単語誤り率で行われるが，本研究では音声認識の改善が読み推定にもたらす影響に関心がある．そこで，このコーパスの先頭 1000 文に対して，ユーザープロンプト有りとなしで音声認識を行い，先行研究[6]の読みの推定の手法を用いて，結果を漢字の読みの推定ができた行数でその効果を評価した．例えば，「金」を「鐘」と音声認識しても読みの「かね」が推定できるため，必ずしも音声認識の誤りが読み推定の成功を妨げないことに注意されたい．

音声認識に利用した Whisper はバージョン: 20240930，モデル: large-v3-turbo，オプション: beam_size=5 である．なお，本研究では Whisper の decording.py を修正し，出力過程の認識候補を複数取得できるように変更している．ユーザープロンプトとしては，字幕をそのまま与えた．

このようにして音声認識を行い，漢字を含む行 851 行に対して読みの推定を行えた行数を数えると，

ユーザープロンプト有り: 463 行

ユーザープロンプト無し: 375 行

でありユーザープロンプト有りの方が優れた結果を示した．

例えば，「心」は「しん」と「こころ」の読みがあり，音声認識で，もしどちらの読みかを推定できないときは，上記の例のように「こころ」の読みをプロンプトで与えることにより，読みの推定精度の向上が期待できる．実際，音声認識結果が「夏目漱石著 心」となる，音声データに対して，音声認識を行う際，「夏目漱石著 こころ」とユーザープロンプトを与えると，「心」が「こころ」と音声認識された．ユーザープロンプトに読みの情報を付与することで，漢字の読みの推定精度の改善に効果があると推察される．

4 振り仮名注釈付き音声コーパスの再構築

本節では，まず既に構築・公開した振り仮名注釈付き音声コーパス[6]における構築手法を概説し，次に本研究の検討手法であるプロンプトを用いて音声認識を行い，漢字の読みを推定する構築手法を説明する．

4.1 前回の青空振り仮名注釈付き音声コーパスの構築手法

「サピエ」の音声デイジーの音声データと，青空文庫のテキストデータを使用して，読みの推定を行った．手順は以下の通りである．

- 音声デイジー[21]の xml を解析し，目次を取得し，文単位の音声データの開始・終了時刻を収集する
- 音声データの分単位での分割を行う
- Whisper による音声認識と認識候補の収集を行う
- 音声認識結果のテキストから注記等を削除する
- 青空文庫テキストの前処理で，ルビ，入力注の削除し，見出し，ルビのデータを収集する
- 目次と見出しの情報から，青空文庫と音声デイジーの本と章の対応を取る
- 編集距離等の情報から，音声データと青空文庫テキストの文の対応を取る
- 読み辞書と音声認識結果から得られた複数候補を組み合わせることで読みの推定を行う

4.2 プロンプトを用いた青空振り仮名注釈付き音声コーパスの構築手法

ユーザープロンプトを用いて音声認識を行う際は，その音声データに対応した青空文庫のテキストをユーザープロンプトにするのが最も効果的であると期待できる．このため，まず音声認識を行い，音声データと青空文庫の文の対応を取る．前回構築したコーパスに対して，Whisper のバージョン 20231117 で，モデルを medium から，最も認識精度の高い large-v3 に変更して，音声認識と読みの推定を行った．このとき，形態素解析器は MeCab-ipadic-neologd [22, 23] から，sudachi [24]に変更した．sudachi のバージョンは 0.6.8 で，辞書を full (20241021)で，分割モードは人間に自然なレベルの B で行った．

次にユーザープロンプトによる処理を行う．既に読みの推定ができたところは，そのまま採用する．読みの推定ができなかったところは，それらが連続した箇所を一つにまとめ，一つの領域とする．この領域の前後は読みの推定ができているため，この領域全体では音声データと青空文庫が対応しているとみなすことができるので，この領域に対してユーザープロンプトのデータを作成する．この青空文庫の文字数が短いとき，具体的には 100 文字以下のとき，それをそのままユーザープロンプトとする．100 文字以上のときは，形態素に分解し，形態素の青空文庫の出現数から音声認識結果の出現数の差を取り，最大 50 個までをユーザープロンプトとする．Whisper のプロンプトとして与えられる最大数は 224 トークンである．なお，ひらがなだけの形態素と助詞は無視している．Whisper のバージョンは 20240930 で，新たにサポートされたモデルの large-v3-turbo で行った．このモデルは，large-v3 に比べて，デコーダ層を 32 から 4 に削減することで，精度は若干犠牲にして処理速度を改善したものである．以上により，読み推定ができなかった領域で，先ほど求めたユーザープロンプト付きで音声認識を行い，音声認識結果と青空文庫の文の対応付けを行い，新たに対応がついた文について音声認識の複数候補を参照しながら読みの推定を行う．

4.3 構築したコーパスの統計情報

個々の作品に対し，収集した文の文字数の全体の文字数に対する割合を収集率として定義したとき，収集率が 50%以上の作品は，作家数 128 人，作品数 (重複タイトル有り) 3686 冊，文字数 7604 万文字であった．コーパスの再構築により，録音時間 3520 時間が 4617 時間になった．

表 1 は全コーパス，表 2 は二人の作家について，先行研究と本研究のコーパス構築結果を比較したものである．表からプロンプトの追加により収集率が大きく改善したことがわかる．また Whisper のモデルや形態素解析器の変更も収集率の改善に寄与した．

表 1 収集した全部の文字数と録音時間

	作家数	作品数(重複タイトル有り)	文字数	録音時間
2024年1月	118人	3252冊	5545万文字	3520時間
2025年1月 プロンプト無し	126人	3641冊	6793万文字	4212時間
2025年1月 プロンプト追加	128人	3686冊	7604万文字	4617時間

表 2 作家別の文字数と録音時間

著者名		作品数 (重複あり)	作品数 (重複無し)	合計の 青空文字数	文字収集率	収集した 文字数	収集した 録音時間
夏目漱石	2024年1月	73	50	5075712	0.712	3616410	242
	2025年1月 プロンプト無し	83	51	5465052	0.762	4165207	282
	2025年1月 プロンプト追加	83	51	5465052	0.85	4644174	308
江戸川乱歩	2024年1月	251	94	13121992	0.701	9195783	607
	2025年1月 プロンプト無し	270	97	13870266	0.752	10432939	680
	2025年1月 プロンプト追加	273	97	14094706	0.811	11426523	720

5 考察

本節では，提案手法が振り仮名付き注釈付き音声コーパスの拡張に貢献した要因を考察し，さらに ReazonSpeech での読み推定の課題を報告する．今回，青空文庫の新規に公開された約 1 年分の作品を取り込んでいるが，対応する音声デイジーが少ないため，収録時間の増加への寄与は少ない．収録時間が増えた主な理由は，下記の 3 点によると考えられる．

1 点目は，Whisper のモデルを medium からより高精度の音声認識モデルに変更したことにより音声認識結果の品質が改善したことにあると考えられる．音声デイジーと青空文庫のテキストの文の対応付けの成功率が改善したことで，読みの推定精度も向上したと考えられる．例えば，文の途中で誤っていた単語の 2 文字が正解になることにより，その前後を含めてより長い文字列が一致することになり，文の対応付けの成功率が向上する

2 点目は，ユーザープロンプトの導入により音声認識精度が改善したことで，1 点目と同様に音声デイジーと青空文庫のテキストの文の対応付けの成功率が改善したためと考えられる．

3 点目は，読みが推定できなかった領域に対して再度音声認識を行うことにより，幻覚の影響を低減できたことにあると考えられる．幻覚は色々なタイプがあるが，たとえば，同じ単語を大量に繰り返す幻覚があり，「こんにちは」が「こんにちはこんにちはこんにちは...」と大量に繰り返すと文の対応が困難になる．このように稀に発生する幻覚によって関係ない文字列が大量に出力された場合，正解テキストデータとの対応付けに失敗するが，複数回音声認識を行ったことで，幻覚を含まない出力が得られた可能性がある．

ReazonSpeech の音声コーパスのデータセット all は 3.5 万時間，219,32,215 文から構成されており，大規模振り仮名注釈コーパス構築の有力な候補となりうる．予備実験で振り仮名推定を行った結果では，読み推定で正解とみなすことができた文の割合は 54%であり，更なる読みの推定の向上が必要である．今後は，漢字以外の部分の音声認識誤りを無視し，またユーザープロンプトの与え方を工夫して，この収集率を増やそうと思っている．

6 おわりに

本論文では，OpenAI の Whisper にプロンプトデータを与えて音声認識を行うことにより，音声デイジーの音声データと青空文庫のテキストを元にした音声コーパスが 7604 万文字，4617 時間に拡張できた．今後も，単語の読みの推定の精度の改善の手法の検討と，コーパスの拡充を行いたいと考えている．

参考文献

1. 視覚障害者等の読書環境の整備の推進に関する法律. e-Gov. https://elaws.e-gov.go.jp/document?lawid=501AC0100000049
2. 「障害のある児童及び生徒のための教科用特定図書等の普及の促進等に関する法律」（通称：教科書バリアフリー法）について：文部科学省 https://www.mext.go.jp/a_menu/shotou/kyoukasho/1378183.htm
3. RADFORD, Alec, et al. Robust speech recognition via large-scale weak supervision. In: International Conference on Machine Learning. PMLR, 2023. p. 28492-28518.
4. サピエとは https://www.sapie.or.jp/sapie.shtml
5. 青空文庫　Aozora Bunko https://www.aozora.gr.jp
6. 音声認識を用いた青空文庫振り仮名注釈付き音声コーパスの構築の試み. 佐藤文一 ，吉永直樹 ,豊田正史 ， 喜連川優.言語処理学会第 30 回年次大会講演論文集, 2024.
7. 青空文庫振り仮名注釈付き音声コーパス https://github.com/ndl-lab/hurigana-speech-corpus-aozora
8. Wei, Xizi, and Stephen McGregor. "Prompt Tuning for Speech Recognition on Unknown Spoken Name Entities." Proc. Interspeech 2024. 2024.
9. 機械翻訳手法に基づいた日本語の読み推定. 羽鳥潤, 鈴木久美. 言語処理学会第 17 回年次大会, 2011. p.579-582.
10. 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. 高橋文彦, 森信介. 情報処理学会研究報告, 2014. p.1-10.
11. 読み曖昧性解消のためのデータセット構築手法. 西山浩気, 山本和英, 中嶋秀治. 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.
12. BERT を利用した教師あり学習による語義曖昧性解消. 曹鋭, 田中 裕隆, 白 静, 馬 ブン, 新納 浩幸. 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.
13. 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）. 佐藤文一, 喜連川優. 電子情報通信学会技術研究報告= IEICE technical report: 信学技報 119.478（2020）, 2020, p17-21.
14. BERT の Masked Language Model を用いた教師なし語義曖昧性解消. 新納浩幸, 馬ブン. 言語処理学会第 27 回年次大会発表論文集, 2021, p.1039-1042.
15. 疑似訓練データを用いた BERT による同形異音語の読み推定. 小林汰一郎, 古宮嘉那子, and 新納浩幸. 研究報告自然言語処理 (NL) 2022.3, 2022: p.1-5.
16. 国会審議における同形異音語の分析. 増山幹高, 松田謙次郎. 法學研究：法律・政治・社会. Vol.96 No.2. 2023, p.444-464
17. Yue Yin, Daijiro Mori, Seiji Fujimoto. ReazonSpeech: A Free and Massive Corpus for Japanese ASR. 言語処理学会第 29 回年次大会講演論文集, 2023. p.1134-1139
18. Ma, Hao, et al. "Extending Whisper with prompt tuning to target-speaker ASR." ICASSP 2024-2024 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). IEEE, 2024.
19. 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 ，吉永直樹 ，喜連川優. 言語処理学会第 28 回年次大会講演論文集, 2022.
20. SATO, Fumikazu, et al. Building Large-Scale Japanese Pronunciation-Annotated Corpora for Reading Heteronymous Logograms. In: Proceedings of the Thirteenth Language Resources and Evaluation Conference. 2022. p.7113-7121.
21. DAISY2.02，DAISY3 等の仕様の日本語訳を公開します. | 日本 DAISY コンソーシアム https://blog.normanet.ne.jp/jdc/?q=node/6
22. MeCab: Yet Another Part-of-Speech and Morphological Analyzer https://taku910.github.io/mecab/
23. 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
24. GitHub - WorksApplications_Sudachi_ A Japanese Tokenizer for Business https://github.com/WorksApplications/Sudachi