Building Large-Scale Japanese Pronunciation-Annotated Corpora for Reading Heteronymous Logograms

Fumikazu Sato,[1,3] **Naoki Yoshinaga,**[2] **Masaru Kitsuregawa**[4,2]
[1] The University of Tokyo [2] Institute of Industrial Science, the University of Tokyo
[3] National Diet Library [4] National Institute of Informatics
{fsato0609, ynaga, kitsure}@tkl.iis.u-tokyo.ac.jp

Abstract

Although screen readers enable visually impaired people to read written text via speech, the ambiguities in pronunciations of heteronyms cause wrong reading, which has a serious impact on the text understanding. Especially in Japanese, there are many common heteronyms expressed by logograms (Chinese characters or *kanji*) that have totally different pronunciations (and meanings). In this study, to improve the accuracy of pronunciation prediction, we construct two large-scale Japanese corpora that annotate kanji characters with their pronunciations. Using existing language resources on i) book titles compiled by the National Diet Library and ii) the books in a Japanese digital library called Aozora Bunko and their Braille translations, we develop two large-scale pronunciation-annotated corpora for training pronunciation prediction models. We first extract sentence-level alignments between the Aozora Bunko text and its pronunciation converted from the Braille data. We then perform dictionary-based pattern matching based on morphological dictionaries to find word-level pronunciation alignments. We have ultimately obtained the Book Title corpus with 336M characters (16.4M book titles) and the Aozora Bunko corpus with 52M characters (1.6M sentences). We analyzed pronunciation distributions for 203 common heteronyms, and trained a BERT-based pronunciation prediction model for 93 heteronyms, which achieved an average accuracy of 0.939.

Keywords: pronunciation-annotated corpus, pronunciation prediction, pre-trained model

1. Introduction

The screen reader is a canonical tool not only for visually impaired people to understand written text, but also for children with reading difficulties to learn from textbooks. For example, in Japan, the government has enforced in 2019 a new law on act to further the improvement of reading environments for visually impaired persons and begun to develop multimedia DAISY (digital accessible information system) textbooks for children with reading difficulties. Although screen readers play a vital role in these kinds of activities, incorrect reading causes confusion in understanding the text, especially in languages with common heteronymous logograms (*e.g.*, Japanese and Chinese). For example, in Japanese, if '表' in a phrase '表に出る' is pronounced as '*hyou*' instead of '*omote*,' the meaning changes from 'go outside' to 'listed in a table'; in Chinese, if '好' in a phrase '这个人好说话' is pronounced as *hào* instead of *hǎo*, the meaning changes from 'this person is easy-going' to 'this person likes talking.'

Since pronunciations of heteronyms depend on individual contexts, we want to use a machine-learning classifier to predict a pronunciation of a heteronym for a given context. However, since there is no large-scale pronunciation-annotated corpus in Japanese, it is difficult to train an accurate pronunciation classifier for various heteronyms. Although the recent pretrain-finetune framework advocated by BERT (Devlin et al., 2019) provides resource-efficient training of neural models on natural language processing tasks, the low-resource problem remains to be resolved. This is because pronunciation disambiguation is analogous to word sense disambiguation, which inherently requires substantial training data for individual heteronyms. We therefore need a massive language resource that annotates text with pronunciation at word-level.

Aiming to facilitate corpus-based studies on pronunciation prediction in Japanese, we built two large-scale corpora for training pronunciation classifiers. We leverage existing corpora with sentence-level and corpus-level annotations: 1) book titles compiled by National Diet Library[1] and 2) fiction and non-fiction books compiled by "Aozora Bunko" digital library[2] and their Braille data translated by SAPIE,[3] a Japanese national online library services for persons with print disabilities. We first convert Aozora Bunko text and its Braille translation to sentence-level parallel data as book titles via chapter-level alignment, and then perform word-level alignment by using a pronunciation dictionary compiled from various morphological analyzer dictionaries. We finally obtained the Book Title corpus with 336M characters (16.4M titles) and the Aozora Bunko corpus with 52M characters (2044 books, 120 authors). In experiments, we analyzed distributions of pronunciations for 203 major heteronyms in the obtained corpora, and then evaluated the utility of our corpora on pronunciation prediction task. We finetuned a pretrained Japanese BERT model on the target task for 93 heteronyms with 223 pronunciations, and confirmed the utility of our corpora.

[1] https://www.ndl.go.jp/
[2] https://www.aozora.gr.jp/
[3] https://www.sapie.or.jp/

2. Related Work

This section first reviews existing language resources to predict pronunciation in Japanese, and then mentions the recent pretrain-finetune framework for resource-efficient training of neural models.

2.1. Pronunciation Prediction

In processing Japanese text, the pronunciation prediction is considered as a subtask of morphological analysis (Kudo et al., 2004; Neubig and Mori, 2010). The morphological analysis in Japanese consists of three subtasks, word segmentation, part-of-speech tagging, and lemmatization. Part-of-speech tagging and lemmatization is done by disambiguating lexical entries for the token, which include pronunciation information. The largest public corpus whose words are manually[4] annotated with their pronunciations is the core data of the Balanced Corpus of Contemporary Written Japanese (BCCWJ) (Maekawa, 2008),[5] consisting of only 60k sentences. Because there are few resources that manually annotate words with their pronunciations, researchers have explored methods of acquiring pronunciation-annotated text to train a pronunciation prediction model.

Existing studies on predicting pronunciations of words in contexts focus on predicting unknown words such as proper nouns. Sumita and Sugaya (2006) proposed a method of reading proper nouns with multiple pronunciations, using Web pages that include both proper nouns and their pronunciations. Kurata et al. (2007) and Sasada et al. (2008) exploit speech data to disambiguate new word pronunciation candidates. Hatori and Suzuki (2011b; 2011a) use natural annotations in Wikipedia articles to collect pairs of words and pronunciations. Takahashi et al. (2014) take advantage of kana-to-kanji conversion logs in input methods as noisy pronunciation-annotated data. Nishiyama et al. (2018) leverage contexts of synonyms for each pronunciation of the target heteronym as the pseudo training data.

Although the above studies partially address the lack of the training data in the pronunciation prediction task, these automatically-collected annotated data suffer from noises, and are used just as temporal resources to train a pronunciation classifier (not distributed for future evaluation).

2.2. Word Sense Disambiguation

Recently, researchers attempted to employ the pretrain-finetune framework initiated by BERT (Devlin et al., 2019) for word sense disambiguation, which is the same word-level classification task as pronunciation prediction, and obtained promising results (Huang et al., 2019; Hadiwinoto et al., 2019; Yap et al., 2020; Loureiro et al., 2021). The pronunciation prediction task will also benefit from contextualized word embeddings computed by BERT to capture the similarity between pronunciations of heteronyms in similar contexts. Through preliminary experiments on a limited size of annotated data, we have confirmed the impact of contextualized word embeddings in the pronunciation prediction task in Japanese. This motivates us to develop a large-scale pronunciation-annotated corpora to obtain an accurate pronunciation classifier using BERT. In this study, we semi-automatically build annotated corpora that are enough large to train a neural-based classifier for the pronunciation prediction task. We exploit the obtained corpora to train a BERT-based classifier, and evaluate the utility of the corpus via the high prediction accuracy obtained by BERT.

[4] Precisely speaking, the annotation is obtained by correcting results of automatic morphological analysis.

[5] http://www.ninjal.ac.jp/corpus_center/anno/

3. Preliminaries

This section provides a brief overview of the Japanese writing system for those who speak a different first language other than Japanese, and then introduces several types of heteronyms in Japanese.

3.1. Japanese Writing System

Japanese sentences are basically composed of three types of characters; kanji, hiragana, and katakana; for example in a sentence ‘パリに立ち寄る (*pari ni tachi yoru*, I stop off at Paris),’ ‘立 (*ta*)’ and ‘寄 (*yo*)’ are kanji, ‘に (*ni*),’ ‘ち (*chi*),’ and ‘る (*ru*)’ are hiragana, and ‘パ (pa)’ and ‘リ (*ri*)’ are katakana.

Kanji characters are logograms mainly used for nouns and stems of verbs and adjectives. Japanese kanji were originally imported from China more than 1500 years ago, and Joyo Kanji, regular-use kanji characters officially announced by the Japanese Ministry of Education, now includes 2136 kanji characters. Each kanji character has two types of pronunciation, On-yomi, which derives from the Chinese pronunciations for that kanji (*e.g.*, ‘麦 (*baku*)’), and Kun-yomi, which derives from Japanese words associated with that kanji (*e.g.*, ‘麦 (*mugi*)’). We can explain pronunciations of kanji tokens (*e.g.*, ‘東京 (*toukyou*)’) by concatenations of pronunciations of individual kanji characters in the token (‘東 (*tou*)’ and ‘京 (*kyou*)’), although there are some idiomatic pronunciations only used for specific kanji tokens (*e.g*, ‘東風 (*kochi*)’ and ‘麦酒 (*biiru*, beer)’); in particular, proper nouns for places and person names have many idiomatic pronunciations.

The hiragana and katakana are phonograms (like alphabet in European languages); Hiragana is mainly used for function words and inflectional endings of verbs and adjective. Katakana is mainly used to transcribe foreign words and basically has a corresponding hiragana character (*e.g.*, あ ↔ ア). The number of hiragana and katakana characters is 169 if half-width variants of katakana (*e.g.*, ｶﾒﾗ for カメラ (*kamera*, camera)) are ignored. Hiragana and katakana have basically a one-to-one correspondence with their pronunciations; few exceptions includes ‘は (*ha*)’ which is pronounced as

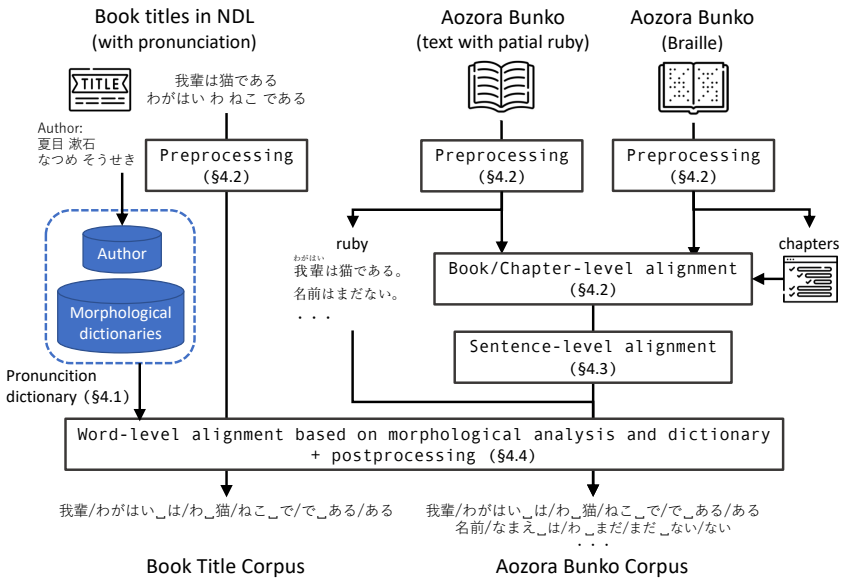

Figure 1: Building corpora with word-level pronunciation annotation.

'wa' when it appears as a particle. Kanji characters are sometimes replaced with their pronunciations represented by hiragana characters (*e.g.*, 'よ (*yo*)' instead of '寄 (*yo*)' in '立ち寄る').

Given the above writing system in Japanese, this paper formulates a pronunciation prediction task that maps kanji tokens (*e.g.*, '東京') in a given sentence with their pronunciations represented by hiragana characters (*e.g.*, 'とうきょう').

3.2. Difficulties in Reading Logograms in Japanese

There are two challenges in reading logograms (kanji) in Japanese with screen readers. One is logograms with idiomatic pronunciations and the other is heteronymous logograms.[6] In what follows, we use the term logograms to refer to one *or more* consecutive kanji characters in Japanese.

The idiomatic logograms such as '東風 (*kochi*)' has been a major concern in pronunciation prediction since their pronunciations cannot be obtained from pronunciations of individual characters in the token. Although the dictionary-based approach, which uses a dictionary to obtain the target classes (pronunciations) for classification, suffers from an unknown pronunciation problem, the problem becomes less severe thanks to recent language resources such as those automatically derived from Wikipedia (Toshinori Sato and Okumura, 2017). This is because it is often the case that idiomatic logograms for proper nouns often have a unique pronunciation, and we do not need disambiguation for those logograms.[7]

On the other hand, heterogeneous logograms such as '表' (*hyou* vs. *omote*) remains to be solved, since we need a massive language resource to disambiguate pronunciations for these logograms. Analogously to word senses for polysemous words, the degree of differences in meanings of individual pronunciations for heteronymous logograms varies from one logogram to another. Some heteronymous logograms have different pronunciations associated with different meanings (*e.g.*, *hyou* (table) and *omote* (outside) for '表,' and *kokuritsu* (national) and *kunitachi* (city name) for '国立'), other heteronymous logograms have similar pronunciations associated with similar or almost identical meanings (*e.g.*, *reihai* (Christian worship) and raihai (Buddhism worship) for '市場,' and *koukou* and *koukuu* (mouth orifice) for '口腔').

4. Construction of Pronunciation-Annotated Corpora

This section explains a method that semi-automatically builds large-scale corpora with word-level pronunciation annotations, by combining existing language resources (Figure 1). We exploit two sets of language resources in this paper: 1) book titles compiled in the National Diet Library, and 2) a collection of fiction and non-fiction books in Aozora Bunko and their Braille translation provided by SAPIE, a Japanese national online library services for persons with print disabilities. The former provides the titles of all the books published in Japan (*e.g.*, '吾輩は猫である' ('I Am a Cat' written by Soseki Natsume) and their pronun-

[6] In addition, since a token consisting of more than one characters can be split by a line break, if the kanji token (*e.g.*, '導入 (*dounyuu*)') is split by a line break ('導\n入'), the screen reader will read (*e.g.*, '導 (*sirube*)') and (*e.g.*, '入 (*nyuu*)'). This problem can be easily solved by removing line breaks between kanji characters.

[7] Proper nouns are named to distinguish with each other, especially when they are in the same named entity category. We can exploit named entity recognition or entity linking when proper nouns cause ambiguities in pronunciation.

ciations in hiragana (*e.g.*, 'わがはいわねこである'). Aozora Bunko compiles more than thousands of books in Japan, and SAPIE provides their Braille translations. Using these sentence-level (title-level) and document-level annotated corpora, one may think of casting the pronunciation prediction task as a text generation task, and applying a neural encoder-decoder model such as Transformer (Vaswani et al., 2017) to generate pronunciation from text. However, document-level generation is still difficult due to the impact of the exposure bias (Ranzato et al., 2016). In addition, even with sentence-level text generation, the encoder-decoder models sometimes suffer from hallucinations. To help visually impaired people read, we want to localize prediction errors within a single word, allowing the user to recover with some effort.

We therefore convert these sentence- and document-level annotations into word-level annotations, using pattern matching based on word-level pronunciation dictionaries. The resulting corpora can be used to evaluate the world-level pronunciation prediction task.

The basic procedure to convert the document-level pronunciation annotation into word-level annotation is as follows; we

- compile a pronunciation dictionary from dictionaries of various morphological analyzer (§ 4.1),
- convert document-level annotated corpora into sentence-level annotated corpora (§ 4.2 and § 4.3), and
- convert sentence-level annotated corpora into word-level annotated corpora (§ 4.4).

4.1. Compiling a Pronunciation Dictionary

We first compile a large-scale pronunciation dictionary for words from various dictionaries of Japanese morphological analyzers.[8] Specifically, we have compiled surface forms and pronunciations of morphemes included in the following dictionaries.

- MeCab-ipadic[9]
- MeCab-ipadic-neologd[10]
- UniDic for Contemporary Written Japanese[11]
- UniDic for Spoken Japanese[11]
- SudachiDict (full)[12]

These dictionaries are chosen because they are standard open-source dictionaries for morphological analysis in Japanese. In addition, we have collected the surface forms and pronunciations for authors and organizations from the book titles in the National Diet Library. We have ultimately obtained 2.5M words in kanji.

4.2. Preprocessing

We next conduct the following resource-specific preprocessing for individual language resources.

The book titles in the National Diet Library contain titles of all books published in Japan since the late modern era, which include old character forms for kanji (*e.g.*, '櫻' for '桜' (*sakura*, cherry blossom).[13] We thus performed the following preprocessing; we

- convert full-width alphanumeric characters (*e.g.*, Ａ and ５) to half-width (*e.g.*, A and 5),
- convert half-width katakana characters (*e.g.*, ｶﾒﾗ) to full-width (*e.g.*, カメラ),
- remove the titles that consist of only English alphabet, traditional Chinese characters, and Hangul characters, and
- convert kanji characters in old character forms (*e.g.*, 櫻) into new character forms (*e.g.*, 桜).

As a result of this preprocessing, we have collected 18,115,976 book titles from the total 19,633,431 book titles.

Aozora Bunko includes early modern literature written in Japanese. They also include some annotations such as ruby characters (pronunciations) for difficult kanji characters and string decorations. We thus performed the following preprocessing; we

- convert full-width alphanumeric characters to half-width,
- convert half-width katakana characters to full-width,
- convert some code points defined in JIS X 0213[14] into the corresponding kanji characters,
- collect ruby characters for kanji characters as their pronunciations,
- remove ruby characters and string decorations, and
- split data into chapters based on chapter headings.

Ruby characters for some kanji characters are used as gold-standard alignments when word-level alignment is performed later.

[8] We here assume morphemes defined in the morphological dictionary as words; although the definition of morphemes slightly depend on the individual dictionaries, it does not a serious impact on our task of word-level token and pronunciation alignment.

[9] https://taku910.github.io/mecab/

[10] https://github.com/neologd/mecab-ipadic-neologd/

[11] https://clrd.ninjal.ac.jp/unidic/en/

[12] https://github.com/WorksApplications/SudachiDict

[13] The old character forms are used until the Japanese government defined a list of kanji for general use in 1946.

[14] https://ja.wikipedia.org/wiki/JIS_X_0213

Finally, the Braille translations of the Aozora Bunko books are represented in various electric Braille formats such as BES, BSE, and BET, and a single data file sometimes contains multiple books. We thus performed the following preprocessing; we

- convert binary Braille data in BES, BSE, and BET formats into the corresponding hiragana characters (namely, pronunciations),
- split the data into single books by extracting the book titles and page numbers from the table of contents,
- remove cover page, table of contents, explanatory notes, gloss, colophon,
- convert historical kana orthography to modern kana usage (*e.g.*, くぁれ→かれ), and
- split data into chapters based on chapter headings defined by the indent of text.

After the above preprocessing, we can find the corresponding books and chapters for the Aozora Bunko text and its Braille translation.

4.3. Building Sentence-level Parallel Corpora

We next extract sentence-level pronunciation annotations from chapter-level annotations. We perform this step only for the Aozora Bunko text, since the book titles compiled in National Diet Library are short enough to directly perform word-level pattern matching.
We extract parallel sentences from the parallel chapters by using periods to segment sentences in the text and using a morphological analysis to find pronunciations corresponding to the resulting sentences via pattern matching. We perform a morphological analysis of the Aozora Bunko text using a Japanese morphological analyzer, MeCab,[15] to associate guessed pronunciations for automatically-segmented words (morphemes) in the text. Because the edition of the target book can be different in the Aozora Bunko text and its Braille counterpart and the morphological analyzer can provide wrong pronunciations for heteronyms, there are several mismatches in both texts. We therefore perform an approximate pattern matching between the guessed pronunciation of the original text and gold pronunciation converted from the Braille data based on Levenshtein distance to resolve this mismatch while considering punctuations in the original text to segment the text into sentences. As a result, we obtain, for each sentence, a gold pronunciation that matches with the guessed pronunciation of that sentence. Although this procedure may associate wrong pronunciations for some sentences due to the difference in the editions of the target book, these noisy data will be removed in finding word-level alignments.

[15] https://taku910.github.io/mecab/

4.4. Building Word-Level Parallel corpora

Finally, we obtain corpora with word-level pronunciations from pairs of a sentence and its pronunciation obtained in Section 4.2 and 4.3. Since the pronunciations in the book titles and the Braille translation of Aozora Bunko text are manually tokenized, we follow this tokenization when obtaining pairs of words (tokens) and its pronunciations, with the exception for tokens that consist of different character types (*e.g.*, 崩す) such as kanji (崩) and hiragana (す). For these tokens, we further split them into character sequences with the same character types (崩␣す) to associate pronunciations for kanji tokens (here, 崩).
We first tokenize the original text (*e.g.*, すぐ着崩す) by using the morphological analyzer, MeCab, to obtain tokens in the text (すぐ␣着␣崩す) and further split the resulting tokens into character sequences with the same character types (すぐ␣着␣崩␣す). We next concatenate successive kanji tokens (着 and 崩) in the resulting text (すぐ␣着崩␣す), since the guessed tokenization for kanji sequences can be inconsistent with the tokenization in the pronunciation. We then compare the resulting tokenized text (すぐ␣着崩␣す) with its manually-tokenized pronunciations (すぐ␣き␣くずす (*sugu ki kuzusu*)) to find pairs of a token and its pronunciation (すぐ and すぐ (*sugu*), 着崩 and き␣くず (*kikuzu*), and す and す (su)). There will be some mismatches between pronunciations defined in the pronunciation dictionary and the provided pronunciations. To resolve this, we

- regard Chinese numerals as Arabic numerals, and
- handle iteration marks (*e.g.*, 'ゝヽゞヾ'),

Finally, we build pronunciation lattices for kanji sequences (here, 着崩) using the pronunciation dictionary compiled in Section 4.1, and perform a depth-first matching between the resulting lattice and the corresponding tokenized pronunciations to obtain text corresponding to each tokenized pronunciation (き (*ki*) for 着, くず (*kuzu*) for 崩).
We removed noisy parallel sentences from the final corpora when we could not have word-level matching of pronunciations. For Aozora Bunko text, we removed all sentences in a book when we could not have alignments for 90% of characters in the book to guarantee the quality of the resulting corpora.

Postprocessing We finally perform corpus-specific postprocessing to reduce the matching failure. For example, for the book titles, we modified '-' to 'ー' to match 'コ-ヒ-' with 'コーヒー.' Since the early Braille data have inconsistent format, we corrected the table of contents for some books.
We have ultimately obtained the Book Title corpus with 336,586,111 characters (16,460,687 book titles) and the Aozora Bunko corpus with 52,385,928 characters (1,618,222 sentences from 2044 books written by 120 authors).

Heteronyms covered as subwords in the pre-trained BERT used in experiments
'表$_2$', '角$_4$', '大分$_2$', '国立$_2$', '人気$_3$', '市場$_2$', '気質$_2$', '役所$_2$', '上方$_2$', '上手$_3$', '下手$_3$', '人事$_2$', '金星$_2$', '仮名$_2$', '内面$_2$', '礼拝$_2$', '遺言$_3$', '口腔$_2$', '後世$_2$', '骨$_2$', '一途$_2$', '一言$_3$', '最中$_3$', '一目$_2$', '係$_3$', '足跡$_2$', '今日$_2$', '明日$_3$', '生物$_3$', '変化$_2$', '大事$_2$', '水車$_2$', '一見$_2$', '一端$_2$', '大家$_3$', '心中$_2$', '書物$_2$', '一角$_2$', '一行$_3$', '一時$_3$', '一定$_2$', '一方$_2$', '一夜$_2$', '下野$_3$', '化学$_2$', '火口$_2$', '花弁$_2$', '玩具$_2$', '強力$_3$', '金色$_2$', '経緯$_2$', '故郷$_2$', '紅葉$_2$', '行方$_3$', '根本$_2$', '左右$_3$', '山陰$_2$', '十分$_2$', '上下$_5$', '身体$_2$', '水面$_2$', '世論$_2$', '清水$_3$', '大手$_2$', '大人$_4$', '大勢$_3$', '中間$_5$', '日向$_4$2', '日時$_3$', '夫婦$_2$', '牧場$_2$', '末期$_2$', '利益$_2$', '工夫$_2$', '一味$_2$', '魚$_3$', '区分$_2$', '施行$_4$', '施工$_2$', '転生$_2$', '博士$_2$', '法華$_2$', '真面目$_3$', '眼鏡$_2$', '文字$_2$', '文書$_3$', '律令$_2$', '現世$_2$', '日中$_2$', '夜中$_3$', '前世$_2$', '二人$_2$', '立像$_2$'

Heteronyms not covered as subwords in the pre-trained BERT used in experiments
'教化$_3$', '見物$_2$', '清浄$_2$', '谷間$_2$', '追従$_2$', '墓石$_2$', '大文字$_2$', '漢書$_2$', '作法$_2$', '兵法$_2$', '大人気$_2$', '半月$_2$', '黒子$_2$', '外面$_2$', '競売$_2$', '開眼$_2$', '求道$_2$', '血脈$_2$', '施業$_2$', '借家$_2$', '頭蓋骨$_2$', '法衣$_2$', '昨日$_2$', '氷柱$_2$', '風車$_2$', '寒気$_2$', '背筋$_2$', '逆手$_2$', '色紙$_2$', '生花$_3$', '白髪$_2$', '貼付$_2$', '一回$_2$', '一期$_2$', '一月$_3$', '一所$_2$', '一寸$_2$', '一声$_2$', '一石$_2$', '一日$_4$', '一分$_3$', '一文$_3$', '一片$_3$', '何時$_3$', '何分$_2$', '火煙$_2$', '火傷$_2$', '火床$_3$', '火先$_2$', '火筒$_2$', '芥子$_3$', '気骨$_2$', '銀杏$_3$', '元金$_2$', '五分$_2$', '後々$_2$', '後生$_2$', '御供$_4$', '細々$_3$', '細目$_2$', '三位$_2$', '疾風$_3$', '菖蒲$_2$', '世人$_2$', '世路$_2$', '船底$_2$', '早急$_2$', '相乗$_2$', '造作$_2$', '他言$_2$', '東雲$_2$', '頭数$_2$', '二重$_2$', '日供$_2$', '日次$_4$', '日暮$_3$', '日来$_3$', '梅雨$_2$', '風穴$_2$', '仏語$_3$', '分別$_2$', '面子$_2$', '木目$_2$', '目下$_2$', '夜直$_2$', '夜来$_2$', '夜話$_2$', '野兎$_2$', '野馬$_3$', '野分$_2$', '野辺$_2$', '野面$_3$', '野立$_3$', '冷水$_2$', '連中$_2$', '飛沫$_2$', '翡翠$_2$', '餃子$_2$', '一足$_2$', '意気地$_2$', '一昨日$_3$', '一昨年$_2$', '十八番$_2$', '十六夜$_2$', '明後日$_2$', '石綿$_2$', '公文$_2$', '読本$_3$', '仏国$_3$', '古本$_2$', '町家$_2$', '遊行$_2$'

Table 1: 203 common heteronyms in Japanese. The subscript shows the number of pronunciation candidates.

heteronym	BERT	# counts	pronunciation (meaning)	Book Title	Aozora Bunko	pronunciation (meaning)	Book Title	Aozora Bunko
変化	✓	88322	*henka* (change)	86365	1612	*henge* (embodiment)	281	64
市場	✓	85723	*ichiba* (marketplace)	592	179	*shijou* (market)	84899	54
国立	✓	19445	*kokuritsu* (national)	19718	24	*Kunitachi* (city name)	243	0
口腔	✓	12051	*koukou* (mouth orifice)	6459	16	*koukuu* (mouth orifice)	5573	3
表	✓	6052	*omote* (outside)	544	2829	*hyou* (table)	2679	0
大分	✓	4421	*daibu* (farily)	7	1079	*Oita* (prefecture name)	3318	17
競売		1253	*kyobai* (auction)	305	8	*keibai* (auction)	938	2
礼拝	✓	944	*reihai* (Christian worship)	780	85	*raihai* (Buddhism worship)	12	67
後世	✓	743	*kousei* (after ages)	486	226	*gose* (afterlife)	4	27
日供		0	*nichigu* (altarage)	0	0	*nikku* (altarage)	0	0

Table 2: The number of occurrences of 203 common heteronyms in our pronunciation-annotated corpora; the column titled BERT shows whether each word is included in the vocabulary of the pre-trained Japanese BERT model used later to predict pronunciations (✓) or not.

5. Analysis

To see the difficulty in predicting pronunciations, we have investigated distributions of pronunciations for common heteronyms on our corpora. We first extract 203 heteronyms from applied rules for characters and "Yomi" (National Diet Library, 2021) (Table 1). We then counted the number of occurrences of each pronunciation of these heteronyms in each corpus. Here, we exclude the cases where the heteronyms appear in compound expressions (*e.g.*, '国立駅' where '国立' is a heteronym), since the pronunciation disambiguation for such cases is rather trivial. Due to the space limitations, we here analyzed a part of the heteronyms with two alternative pronunciations in Table 2. The most frequent heteronym was '変化,' which has two pronunciations '*henka* (change)' and '*henge* (embodiment)', while the least frequent heteronym is '日供,' which has two pronunciations '*nichigu* (altarage)' and '*nikku* (altarage).' 197 of the 203 heteronyms appeared more than 30 times in the entire corpus.

We can also observe that the pronunciation distributions vary across the two domains. For example, the number of '国立' in the Aozora Bunko corpus is much less than that in the Book Title corpus for both pronunciations (*kokuritsu* (national) and *Kunitachi* (city name)), and one pronunciation of '表,' *hyou* (table), does not appear in the Aozora Bunko corpus. The rare pronunciation of '国立' (*Kunitachi*) in the Aozora Bunko corpus can be explained by the fact that it was introduced in 1926. Meanwhile, '大分' rarely appears in the Book Title corpus as *daibu* (fairly), and this is because the book titles rarely include adverbs. Since frequent pronunciations vary across domains, a classifier for predicting pronunciations will suffer from the risk of overfitting to the the domain used in training the classifier. In the future, we will explore a method of collecting additional examples for rare pronunciations to augment our corpora; for example, we will use contextualized word embeddings of the rare pronunciations in our corpora to collect examples from the Web.

heteronym	pronunciation (meaning)	count total	count (corr.)	pronunciation (meaning)	count total	count (corr.)	acc.
大分	*daibu* (farily)	218	216	*Oita* (prefecture name)	664	663	0.997
身体	*shintai* (system)	4016	3998	*karada* (body)	847	770	0.980
一目	*hitome* (glance)	335	332	*ichimoku* (respect)	49	36	0.958
心中	*shincyuu* (feelings)	59	51	*shinjyuu* (joint suicide)	345	336	0.958
表	*omote* (outside)	662	603	*hyou* (table)	526	522	0.947
玩具	*omocha* (toy)	52	47	*gangu* (toy)	280	266	0.943
博士	*hakushi* (doctor)	3585	3374	*hakase* (expert)	535	479	0.935
礼拝	*reihai* (Christian worship)	174	168	*raihai* (Buddhism worship)	17	9	0.927
故郷	*kokyou* (hometown)	784	755	*furusato* (hometown)	106	28	0.880
今日	*kyou* (today)	3682	3403	*kon'nichi* (nowadays)	1471	1045	0.863
現世	*gensei* (this life)	36	25	*gense* (this life)	49	48	0.859
金色	*kin'iro* (golden)	200	197	*konjiki* (golden)	104	57	0.836
上方	*kamikata* (Kyoto-Osaka area)	291	238	*jouhou* (upper)	128	112	0.835
口腔	*koukou* (mouth orifice)	1300	1000	*koukuu* (mouth orifice)	1113	873	0.776

Table 3: Results of pronunciation prediction using BERT; the columns corr. refers to the number of correctly classified examples for each pronunciation.

6. Experiments

This section evaluates the utility of our corpora on the pronunciation prediction task. We use the pre-trained Japanese BERT[16] to solve the pronunciation prediction task as a sequence labeling task. Although we can also use our corpora to solve the pronunciation prediction task by generation (Hatori and Suzuki, 2011a) instead of classification, here we adopt the classification-based approach commonly used in the literature. This is because i) we can assume a large-scale pronunciation dictionary to enumerate pronunciation candidates for kanji tokens, ii) we target on heterogeneous logograms in this study.

In what follows, we first explain the experimental settings, and then report the accuracy of pronunciation prediction. For brevity, in this experiment, we focus on heteronyms included in the subword vocabulary of the pre-trained Japanese BERT. Among the 203 heteronyms in Table 1, 93 heteronyms (223 pronunciations) are covered by the subword vocabulary of the pre-trained Japanese BERT.

6.1. Settings

Data We first collect sentences that include the target heteronyms from both the Book Title and Aozora Bunko corpora. We then split the resulting corpora into training, development and test split with a ratio of 6:2:2; the training, development, and test data included 456,223 (9,246,160), 152,095 (3,079,925), and 152,180 (3,074,577) sentences (tokens), respectively.

Model We implemented the BERT-based sequence labeling using PyTorch Lightning[17] and huggingface-transformers.[18] Since it is too costly to train independent disambiguation models for individual heteronyms, we cast the prediction task as sequence labeling. We provide heteronym-specific pronunciation tags (namely, 223 tags in total) for individual heteronym-pronunciation pairs. For subword tokens other than the target 93 heteronyms, we give a single dummy tag as the OTHER tag in the named entity recognition.[19]

6.2. Results

The macro average of prediction accuracy with the BERT-based classifier was 0.939, while that of the majority class baseline is 0.884.[20] Table 3 lists the detailed experimental results for some heteronyms with two pronunciations. We can see that our classifier successfully predicted the correct pronunciation for heteronyms that have semantically-distinguishable pronunciations (*e.g.*, *daibu* (fairly) and *ooita* (Oita prefecture) for '大分,' and *shinchuu* (feelings) and *shinjuu* (joint suicide) for '心中.' Meanwhile, it is difficult to distinguish pronunciations with similar meanings; *kyou* (today) and *kon'nichi* (Nowadays) for '今日,' and *koukuu* and *koukuu* (mouth orifice) for '口腔.'

Misclassified Examples We finally report some misclassified examples that highlight the difficulty of the pronunciation prediction task. The classifier some-

[16] https://github.com/cl-tohoku/bert-japanese

[17] https://github.com/PyTorchLightning/pytorch-lightning

[18] https://github.com/huggingface/transformers

[19] The hyperparameters of the model were as follows: learning rate= $1e-5$, batch size= 32, number of epochs= 5, and max token length= 128. We used Adam (Kingma and Ba, 2015) as an optimizer and minimizes the cross-entropy loss function.

[20] There were a few cases in which BERT classifiers the pronunciation of the target heteronym as the single dummy tag for tokens other than the target heteronym (less than 6 times for heteronyms in Table 3). These errors are excluded from the results.

times misclassifies pronunciations that depend on the specific style of text. The following examples are taken from the Book Title corpus and Dogura Magura written by the Kyusaku Yumeno in the Aozora Bunko corpus, respectively.

(1) 青年 よ 故郷 に 帰って 市長 に なろう
*seinen yo *kokyou ni kaette shichou ni narou*
'Boys, return to your hometown to be a mayor.'

(2) そこ の 閻魔 は 医学 の 博士 で。
*soko no enma ha igaku no *hakase de*
'Yama there is a doctor of medicine.'

In (1), 故郷 should be pronounced as '*furusato*' instead of '*kokyou*.' Although both pronunciations mean hometown, *furusato* is preferred in spoken language (as in this example), while *kokyou* is preferred in written language. In (2), 博士 should be pronounced as '*hakushi*' instead of '*hakase*.' *Hakushi* is preferred especially when 博士 mentions a doctoral degree (here, doctor of medicine), while *hakase* is preferred in a more casual context.

There are several cases where we need more contexts for classification. The following example is taken from Kaso Jinbutsu written by Shusei Tokuda in the Aozora Bunko corpus.

(3) 先生 は 大家 よ。
*sensei wa *ooya yo*
'You are a great master.'

In this example, 大家 means a great master and should be pronounced as *taika* instead of *ooya*, which means a landlord. We need more context into consideration to handle this kind of examples.

7. Conclusions

We have developed large-scale Japanese corpora whose words are annotated with pronunciations, exploiting existing language resources including i) book titles in National Diet Library and ii) books in Aozora Bunko, a Japanese digital library and their Braille translations. After converting existing resources into sentence-level aligned corpora, we performed word-level alignment using a pronunciation dictionary compiled from various morphological analyzer dictionaries.

We finally obtained two large-scale corpora with word-level pronunciation annotations: the Book Title corpus with 336M characters (16.4M titles) and the Aozora Bunko corpus with 52M characters (1.6M sentences). We have fine-tuned the pre-trained Japanese BERT on the pronunciation prediction task, and confirmed the utility of our corpora in improving the pronunciation prediction. We have released our Book Title corpus corpus[21] and Aozora Bunko[22] to promote research on pronunciation prediction in Japanese.

[21] https://github.com/ndl-lab/huriganacorpus-ndlbib

[22] https://github.com/ndl-lab/huriganacorpus-aozora

Acknowledgments

We thank the National Association of Institutions of Information Service for Visually Impaired Persons, Japan and the Japanese Braille Library for their permission to distribute the Aozora Bunko corpus. We thank Toru Aoike for his help in using book titles and releasing our corpora. The research (second author) was partially supported by NII CRIS collaborative research program operated by NII CRIS and LINE Corporation.

8. Bibliographical References

Devlin, J., Chang, M.-W., Lee, K., and Toutanova, K. (2019). BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4171–4186, Minneapolis, Minnesota, June. Association for Computational Linguistics.

Hadiwinoto, C., Ng, H. T., and Gan, W. C. (2019). Improved word sense disambiguation using pre-trained contextualized word representations. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 5297–5306, Hong Kong, China, November. Association for Computational Linguistics.

Hatori, J. and Suzuki, H. (2011a). Japanese pronunciation prediction as phrasal statistical machine translation. In *Proceedings of 5th International Joint Conference on Natural Language Processing*, pages 120–128, Chiang Mai, Thailand, November. Asian Federation of Natural Language Processing.

Hatori, J. and Suzuki, H. (2011b). Predicting word pronunciation in Japanese. In *CICLing 2011, Lecture Notes in Computer Science (6609)*, pages 477–492.

Huang, L., Sun, C., Qiu, X., and Huang, X. (2019). GlossBERT: BERT for word sense disambiguation with gloss knowledge. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 3509–3514, Hong Kong, China, November. Association for Computational Linguistics.

Kingma, D. P. and Ba, J. (2015). Adam: A Method for Stochastic Optimization. In *International Conference for Learning Representations*.

Kudo, T., Yamamoto, K., and Matsumoto, Y. (2004). Applying conditional random fields to Japanese morphological analysis. In *Proceedings of the 2004 Conference on Empirical Methods in Natural Language Processing*, pages 230–237, Barcelona, Spain, July. Association for Computational Linguistics.

Kurata, G., Mori, S., Itoh, N., and Nishimura, M. (2007). Unsupervised lexicon acquisition from speech and text. In *Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing*, pages 421–424.

Loureiro, D., Rezaee, K., Pilehvar, M. T., and Camacho-Collados, J. (2021). Analysis and evaluation of language models for word sense disambiguation. *Computational Linguistics*, 47(2):387–443, June.

Maekawa, K. (2008). Balanced corpus of contemporary written japanese. In *Proceedings of the sixth Workshop on Asian Language Resources (ALR-8)*, pages 101–102.

National Diet Library, J. (2021). Applied rules for characters and "yomi".

Neubig, G. and Mori, S. (2010). Word-based partial annotation for efficient corpus construction. In *Proceedings of the Seventh International Conference on Language Resources and Evaluation (LREC'10)*, Valletta, Malta, May. European Language Resources Association (ELRA).

Nishiyama, K., Yamamoto, K., and Nakajima, H. (2018). Dataset construction method for word reading disambiguation. In *Proceedings of the 32nd Annual Conference of the Japanese Society for Artificial Intelligence*. (In Japanese).

Ranzato, M., Chopra, S., Auli, M., and Zaremba, W. (2016). Sequence level training with recurrent neural networks. In *Proceedings of the fourth International Conference on Learning Representations*.

Sasada, T., Mori, S., and Kawahara, T. (2008). The improvement of predicting pronunciation by acquiring lexicons from speech and text. In *Proceedings of the Annual Meeting of Natural Language Proceeding*, pages 420–423. (In Japanese).

Sumita, E. and Sugaya, F. (2006). Word pronunciation disambiguation using the Web. In *Proceedings of the Human Language Technology Conference of the NAACL, Companion Volume: Short Papers*, pages 165–168, New York City, USA, June. Association for Computational Linguistics.

Takahashi, F. and Mori, S. (2014). Improving the accuracy of word segmentation and pronunciation predction using kana-to-kanji conversion logs. In *IPSJ SIG Technical Report*. (In Japanese).

Toshinori Sato, T. H. and Okumura, M. (2017). Implementation of a word segmentation dictionary called mecab-ipadic-neologd and study on how to use it effectively for information retrieval (in japanese). In *Proceedings of the Twenty-three Annual Meeting of the Association for Natural Language Processing*, pages NLP2017–B6–1. The Association for Natural Language Processing.

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, L. u., and Polosukhin, I. (2017). Attention is all you need. In I. Guyon, et al., editors, *Advances in Neural Information Processing Systems*, volume 30. Curran Associates, Inc.

Yap, B. P., Koh, A., and Chng, E. S. (2020). Adapting BERT for word sense disambiguation with gloss selection objective and example sentences. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 41–46, Online, November. Association for Computational Linguistics.