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

Fumikazu Sato (U. Tokyo & National Diet Library), Naoki Yoshinaga (IIS, U. Tokyo), and Masaru Kitsuregawa (NII & IIS, U. Tokyo)

Overview

- Incorrect reading with a screen reader confuses visually impaired persons and children with reading difficulties in understanding written text
- We have built two large-scale pronunciation annotated corpora in Japanese: Book Title Corpus (336M char) & Aozora Bunko Corpus (52M char.)
- We used the obtained corpus to train and evaluate BERT-based pronunciation classifiers, and obtained macro accuracy of 0.939

Research Background

Screen readers are essential for visually impaired persons to read text via speech

- [Japan] a law on *act to further the improvement of reading environments for visually impaired persons*

Challenge: heteronymous logograms (kanji) causes serious troubles in reading text by screen readers

表	に	出る	
omote	*ni*	*deru*	(listed in a table)
hyou	*ni*	*deru*	(go outside)

Goal: improve the accuracy of a machine learning (ML) classifier for predicting pronunciations

Japanese writing systems

Japanese sentences consist of phonograms (hiragana & katakana) and **logograms (kanji)**

パリ	に	立ち	寄る	
Pari	*ni*	*tachi*	*yoru*	(stop off at Paris)

Two difficulties in reading kanji (logograms)

- Idiomatic reading for common and proper nouns
  東風 (kochi, a wind from east in spring)
- Heteronymous logograms
  表 (*hyou* (table) vs. *omote* (outside))

Related work

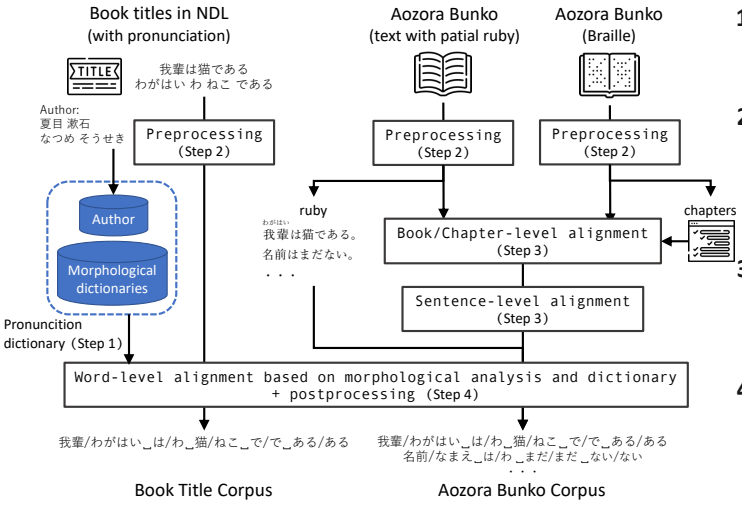
- Expand a morphological **dictionary for proper nouns** from Wikipedia [T. Sato 2017]
- Obtain a **pseudo (noisy) corpora** from:
  - Web for reading proper nouns [Sumita+, 2006]
  - speech data [Kurata+ 2007, Sasada+ 2008]
  - Wikipedia [Hatori+, 2011b, 2011a]
  - kana-kanji conversions logs [Takahashi+ 2014]
  - contexts of synonyms [Nishiyama+ 2018]

The largest corpus with manual annotation (BCCWJ) has just 60k sentences

Method to obtain corpora with word-level pronunciation annotations

We exploit existing resources with sentence and document-level pronunciation annotations to obtain word-level annotations

- Book titles compiled by National Diet Library (NDL), which cover all books and magazines since the late modern era (sentence-level annotations)
- Fiction and non-fiction books in Aozora Bunko (Japanese digital library) and its Braille translations (document-level annotations)



1. **Compiling a dictionary** to enumerate pronunciation candidates in Step 4
   - **Exploit morphological dictionaries**: MeCab (ipadic, ipadic-neologd, unidic for Contemporary written Japanese and Spoken Japanese) and SudachiDict (full)
2. **Preprocessing** (mostly normalization, see our paper for details)
   - Normalize katakana, alphanumeric, and kanji
   - (Book titles) use authors (its pronunciations) to expand the dictionary in Step 1
   - (Aozora Bunko) collect ruby annotations as gold pronunciation for Step 4
   - (Braille) convert Braille (BES, BSE and BET formats) to hiragana (pronunciation)
3. **Sentence-level alignment (Aozora Bunko only)**
   1. Use chapter info extracted from Braille to obtain chapter-level alignments
   2. Run a morphological analyzer on the text to obtain pseudo pronunciations
   3. Split text into sentences by periods while matching pseudo/gold pronunciations
4. **Word-level alignment** (すぐ着崩す/すぐ␣き␣くずす as running example)
   1. Split text into tokens in the same char types via morphemes (すぐ␣着崩␣す)
   2. Find **pairs of a token in the text and its pronunciation** in the tokenized pronunciation (すぐ/すぐ␣着崩/き␣くず␣す/す)
   3. Find **pronunciations for kanji tokens** (着/き, 崩/くず) in the kanji sequence (着崩/き␣くず) by building pronunciation lattices with the dictionary in Step 1

Statistics of the obtained corpora

- Book Title corpus (https://github.com/ndl-lab/huriganacorpus-ndlbib): 336,586,111 characters (16,460,687 book titles)
- Aozora Bunko corpus (https://github.com/ndl-lab/huriganacorpus-aozora): 52,385,928 characters (1,618,222 sentences, 2044 books, 120 authors)

Analysis on pronunciation distributions

We see pronunciation distributions in the obtained corpora to find difficulties of the pronunciation prediction task

- **Target**: 203 heteronymous logograms (kanji) extracted from *applied rules for characters and "Yomi"* [National Diet Library, 2021]
  - Exclude compound nouns such as (国立駅, *Kunitachi eki*), since their pronunciations can be unique (if a dictionary covers)

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

- 197 out of 203 heteronymous kanji appear more than 30 times
- Pronunciation distributions vary across domains (*e.g.*, 表, 大分), suggesting the risk of overfitting to the domain used in training
  - Is there a way to collect contexts for individual readings?

Experiments on predicting pronunciation using BERT

We evaluate the utility of our corpora on pronunciation prediction

- **Train/dev/test data**: 456,223, 152,095, and 152,180 sentences
- **Model**: Pretrained BERT (https://huggingface.co/cl-tohoku) for sequence labeling
- **Target**: 93 heteronyms in the subword vocabulary of the pretrained BERT

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

- We obtained macro accuracy of 0.939 (majority class baseline: 0.884)
- Semantically-distinguishable heteronyms (大分, 心中, 表) are easy to read, while heteronyms depending on style/domain (故郷, 金色) are hard to read
- Some examples need more contexts for disambiguation