Modeling Native Chinese Speakers' Acquisition of Japanese Kanji Pronunciation

Zilong Li, Alexis Palmer

Motivation

Kanji refers to Chinese characters used in Japanese. Because of the long history of communication, Chinese and Japanese share a large kanji vocabulary. People native in one language may infer the pronunciation of a kanji word in another language.

漢 : h → k + an → an

字 : z → z + i → i

Kanji is a kind of logogram characters. Characters with the same radical have similar pronunciations. Orthographic information and phonological both play roles in kanji acquisition.

	反	饭	返	贩	阪
Chinese:	fan	fan	fan	fan	ban
Japanese:	han	han	hen	han	han

Background

Xing et al. (2010) taught pronunciation correspondence rules to Japanese learners who are native Chinese speakers at the beginning level. They found that students achieved higher accuracy in the exam when educated with these rules.

Previous works (Matsumi et al., 2014; Tanaka, 2015; Fei et al., 2022b) recorded reaction time and accuracy when Chinese students with different proficiency in Japanese encountered kanji words in visual and auditory lexical decision task. The orthographic information is believed to have facilitation, while it is still unclear whether the phonological knowledge of Chinese is positive or negative to Japanese learning.

Reasearch Questions

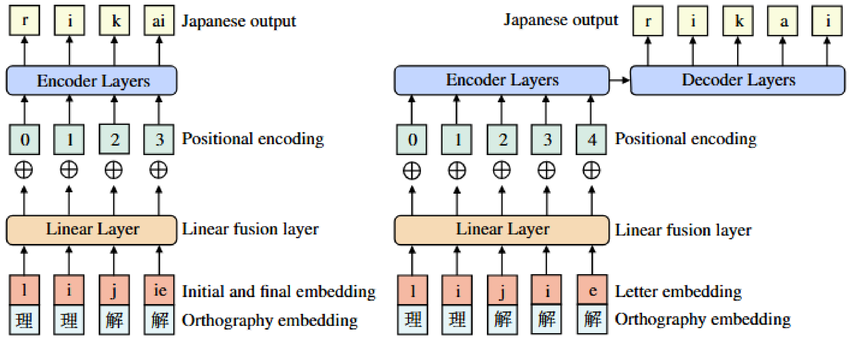
- In this project, we build deep learning models as **proofs-of-concepts.** Provide different input, explore the learnability of Kanji pronunciation.
- Answer the following research questions:
  1. Are summarized Chinese-Japanese correspondence rules effective in Japanese learning?
  2. Are these rules learnable in the regular Japanese learning process?
  3. How is the effect of orthographic information, positive or negative?
  4. What happens if Chinese phonological knowledge is associated with orthographic knowledge in Japanese learning?

Experiments

- Task: input kanji word pronunciation (in either initial-final or romanized letters) and orthography, output Japanese pronunciation

xian xiang
现 象 – – [models] –> gen syou

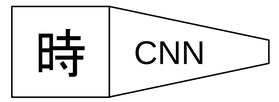
- Data: Japanese Kanji Vocabulary in Contrast to Chinese (JKVC) sorted in the order of difficulty
- Learn by rules (Encoder-only model with initials and finals of Chinese as input)
- Learn implicitly (Encoder-decoder model with romanized letters as input)



- Learning from Scratch (randomly initialized CNN)

vs.

- Language Transfer (pretrained phonological decision task)



Result

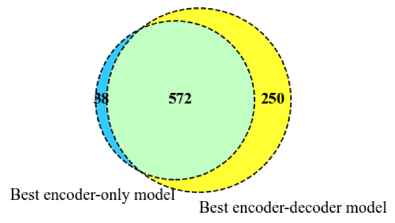
- Evaluation Metrics: Levenshtein edit distance

Models	Beg	Int	Adv
Encoder (phonological rules limitation)	2.56	2.06	0.79
En+CNN (orthographic information)	2.89	1.26	0.78
En+pretrained CNN (language transfer)	2.73	1.95	0.74
Encoder-Decoder (no rules limitation)	3.68	2.64	0.50
En-De+CNN (orthographic information)	3.81	**0.97**	**0.36**
En-De+pretrained CNN (language transfer)	3.75	1.36	0.41
Bart-base-chinese (pretrained LM)	**1.68**	1.10	0.42

- Overlap between words correctly predicted by the best encoder-only model and by the best encoder-decoder model

kan
ken
) 1 substitution

ko
kou
) 1 insertion



Conclusion

1. Summarized pronunciation rules are effective especially under limited training data setting.
2. The correspondence in kanji pronunciation can be covered in a regular learning process.
3. Orthographic ability is important for learning vocabulary belonging to the intermediate or higher level.
4. Language transfer from Chinese is not always helpful.