P8-20

NLP 2024 KOBE

Exploring the Potential of Prompt-Based Method for Kanji-Kana Conversion in Japanese Braille Translation

Micah Kitsunai[1] Deborah Watty[1] Shu-Kai Hsieh[1]

[1]Graduate Institute of Linguistics, National Taiwan University

National Taiwan University 國立臺灣大學

Introduction

About Japanese Braille

- Braille operates using 6 dots in a single cell, providing 64 unique combinations depending on the raised or unraised status of each dot.
- When translating from text to Braille, it is necessary to convert all Kanji to Kana first.
- **Kanji have multiple readings.**
- A key challenge in Braille translation is **accurately predicting the appropriate readings of Kanji.**

今日		は	晴れ	
today		TOPIC	sunny	
キョ (kyo)	ウ (u)	ハ (ha)	ハ (ha)	レ (re)
キョ (kyo)	ウ (u)	ワ (wa)	ハ (ha)	レ (re)

Figure 1. Example of Kana Conversion.

Word	Reading	Type	Note
日	hi	On vs. Kun	On'yomi
日	nichi	On vs. Kun	Kun'yomi
会社	**ka**isha	voiced vs. unvoiced	Single use
会社	**ga**isha	voiced vs. unvoiced	Compound word (e.g. 株式会社)
辛い	karai	polyphone	(Meaning) spicy
辛い	tsurai	polyphone	(Meaning) sadness
昨日	kinou	synonym	Soft expression
昨日	sakujitsu	synonym	Formal expression

Figure 2. Classification of Multiple Readings in Japanese.

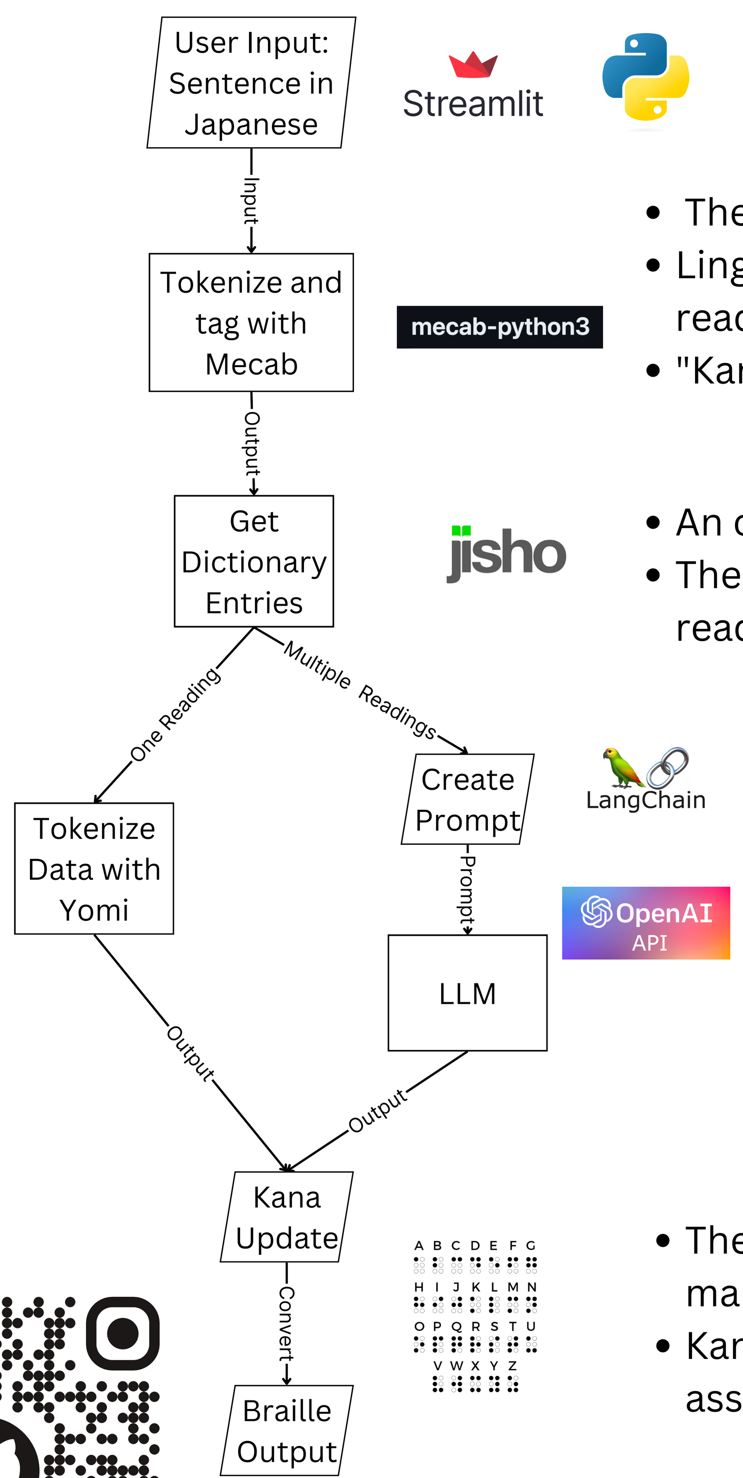
Other Approaches

- Rule-based methods, such as Liblouis [1], exist but require the continuous addition of new rules.
- Using Neural Machine Translation Technology [2, 3] or BERT for polyphone disambiguation [4].
- Using pre-trained language models for **polyphone disambiguation** in text-to-speech [5], but they require extensive Braille data and training.
- [6] applied **a prompt-based method for disambiguating polyphone characters in Taiwanese Mandarin, incorporating syntactic analysis and external dictionary data into LLM prompts**, achieving higher accuracy than traditional rule-based approaches.

Objectives

- Addressing the need for manual corrections in existing automatic Braille translation tools, which are not perfectly accurate.
- Proposing a cost-effective method for applying Large Language Models (LLMs) to automatic Braille translation, without the need for fine-tuning.

Methodology



- The open-source morphological analysis engine.
- Linguistic details for each morpheme, including POS, reading, and phonetics.
- "Kana as spoken form" , transcribing sounds as heard.

- An online Japanese-English dictionary.
- The focus is on words containing Kanji with multiple readings or multiple dictionary entries.

- Library for developing applications with LLMs.
- Incorporated meanings and possible readings from JishoAPI and create prompt.

- The readings data are combined with Braille data and mapped.
- Kana and Braille characters are almost exclusively associated as one - to-one pairs.

```
I have a Japanese sentence："人気の店". I want the Kanji converted to Katakana.
Target Kanji："['人気', '店']". Here are the relevant dictionary entries.
=================================================
Word information for '人気'
Reading: にんき
Meaning:
- En:popularity, public favor
- En:condition (e.g. market), tone, character, nature
Reading: ひとけ or ひとげ
Meaning:
- En:sign of life, human presence
- En:humanity, humaneness

Please output the reading of the target Kanji, referring to the dictionary, and the context of
the sentence itself.
```

Figure 5. An Example of a Prompt

Evaluation and Results

- Constructed a dataset of 67 short sentences, each including one polyphone from a list of 18 different polyphone characters, sourced from the Balanced Corpus of Contemporary Written Japanese [7].

Mecab:

- Identified instances of inadequate context-based judgment.
- Certain characters were consistently assigned the same reading across all examples, and instances were observed where technically correct but uncommon readings were selected.

OpenAI Models:

- Observed random On'yomi and Kun'yomi combinations or entirely random, non-existent readings in both GPT-3.5 and GPT-4 models.
- **GPT-3.5's low Kanji reading accuracy is linked to a lack of sufficient Kanji data in the model.**
- **GPT-4 shows significant improvement, indicating that prediction accuracy is highly dependent on data quality and volume.**

Mecab + OpenAI Models:

- Observed improved accuracy.
- Referencing Mecab and dictionaries prevented the generation of non-existent readings by GPT.
- **This methodology reduces the risk of creating non-existent words by using Mecab for morphological analysis and dictionary data.**
- **Accuracy for synonyms with different readings remains low, possibly due to inadequate semantic categorization for these cases.**

Model	Score	Accuracy
Mecab	41/67	61.19%
GPT-3.5	29/67	43.28%
GPT-4	48/67	71.64%
Mecab ＋ GPT-3.5	43/67	64.17%
Mecab ＋ GPT-4	50/67	74.63%

Figure 6. Comparison of Scores and Accuracy for Each Model

Word (Reading)	Error	Category	Method
人気 hitoke	*ninnki	Unable to judge the context	Mecab
私 watashi	*watakushi	uncommon pronunciation	Mecab
人気 hitoke	*jinnke	non-existent reading	LLM
人事 jinnji	*ninnji	non-existent reading	LLM
昨日 kinou	*sakujitsu	synonym	all

Figure 7. Examples and Classification of Errors

Conclusions

- Compared to the standalone use of morphological analysis tools like Mecab and lanuage models such as GPT, **the method proposed in this research — incorporating dictionary data for polyphones into prompts and then inputting these into an LLM — demonstrated higher accuracy.**
- **Future Works:** There is potential in expanding the dictionary database and incorporating comprehensive textual context and word usage examples into the prompts.

References

[1] C. Egli, "Liblouis–a universal solution for braille transcription services," in *Proceedings of Daisy 2009 Conference*, 2009.

[2] Y. Shimomura, H. Kawabe, H. Nambo, and S. Seto, *Braille Translation System Using Neural Machine Translation Technology I - Code Conversion*, p. 335–345. Springer International Publishing, June 2019.

[3] H. Kawabe, Y. Shimomura, and S. Seto, "Braille translation system using neural machine translation technology II – code conversion of kana-kanji mixed sentences," in *Proceedings of the Fifteenth International Conference on Management Science and Engineering Management: Volume 1 15*, pp. 417–426, Springer, 2021.

[4] F.Satou, M.Kitsuregawa, *et al.*, "Improvement of reading of heteronym by a word embedded vector on pre-trained bert," *AAC*, vol. 2020, pp. 1–5, 2020.

[5] R. Hida, M. Hamada, C. Kamada, E. Tsunoo, T. Sekiya, and T. Kumakura, "Polyphone disambiguation and accent prediction using pre-trained language models in japanese tts front-end," 2022.

[6] D. Watty, M. Kitsunai, and S.-K. Hsieh, "Prompt-based translation of chinese into taiwanese mandarin braille," in *2023 International Conference on Asian Language Processing (IALP)*, pp. 56–61, IEEE, 2023.

[7] K.Maekawa, "Kotonoha development of balanced corpus of contemporary written japanese," *Study of Japanese*, vol. 4, no. 1, pp. 82–95, 2008.