GlyphFormer: Improving Japanese Language Models with Sub-character Tokenization

Koichiro Kanno
artisan.baggio@gmail.com

August 18, 2024

Abstract

This paper examines the effectiveness of using sub-character tokenization for Japanese language processing by utilizing the ALBERT [1] model. I focused on radical and element-based sub-character tokenization and compared the results with traditional character-based tokenization. The evaluation was conducted on a dataset derived from the Japanese novel "Botchan," containing 500 sentences. The results indicate that sub-character tokenization significantly improves the model's perplexity, especially when using radical and element-based approaches.

1. Introduction

Natural language processing (NLP) using models like Transformers has predominantly been researched in languages based on alphabets, with limited studies focusing on multibyte character languages such as Japanese. Unlike English, where words are clearly separated by spaces, Japanese text often lacks explicit delimiters, making word segmentation a complex task. Tools like MeCab [2] and SentencePiece [3] are commonly used for tokenizing Japanese text into words or subwords before applying Transformer models.

However, Japanese text heavily utilizes kanji characters, which are composed of various components that convey semantic and phonetic information. This study focuses on how dividing kanji into smaller units, or sub-characters, can affect the performance of the ALBERT model in Japanese language processing.

Kanji characters are traditionally classified into four types based on their formation:

- **Pictographs:** Characters that are visual representations of objects (e.g., 山 (mountain), 川 (river)).
- **Ideographs:** Characters that represent abstract concepts through lines or dots (e.g., 上 (up), 下 (down)).
- **Compound Ideographs:** Characters formed by combining multiple characters to convey a new meaning (e.g., 林 (forest), composed of two 木 (tree) characters).
- **Phono-semantic Characters:** Characters formed by combining a phonetic component with a semantic component, known as a radical (e.g., 語 (word), where 言 (speech) is the radical).

2. Previous Work

A previous study using LSTM models demonstrated that breaking down kanji into sub-characters could improve model performance on Japanese NLP tasks [4]. However, this study did not explore the use of more advanced Transformer-based models like ALBERT. My research aims to extend these findings by applying sub-character tokenization techniques to the ALBERT model.

---

3. Methodology

The experiments were conducted using the ALBERT model, focusing on two sub-character tokenization methods: radical-based and element-based [5]. These methods were compared against traditional character-based tokenization using a dataset derived from the novel "Botchan," which contains 500 sentences out of a total of 2715.

The model was trained using the following parameters:

- Model: ALBERT without pre-training
- Batch size: 1
- Gradient accumulation steps: 1
- Learning rate: 5 × 10^{-5}
- Epochs: 3
- Optimizer: AdamW

Perplexity was used as the evaluation metric, calculated as follows:

$$\text{Perplexity} = \exp\left(\frac{1}{N}\sum_{i=1}^{N}\log_2\frac{1}{P(w_i)}\right)$$

Where N is the number of words in the dataset, and $P(w_i)$ is the probability of word w_i.

3.1. Sub-character Tokenization Examples

The sentence "親譲りの無鉄砲で小供の時から損ばかりしている。" was tokenized as follows:

- **Character-based:** 親 譲 り の 無 鉄 砲 で 小 供 の 時 から 損 ばかり し て いる 。
- **Radical-based:** 立 木 見 言 襄 り の 丿 一 灬 金 失 石 包 で 小 亻 共 の 日 寺 から 扌 員 ばかり し て いる 。
- **Element-based:** 立 木 見 亠 目 言 六 衣 八 一 亠 口 襄 三 り の 灬 一 丿 金 大 夫 失 丿 己 包 石 丿 口 勹 で 小 八 共 亻 の 日 寸 土 寺 から 員 口 目 貝 扌 ばかり し て いる 。

---

4. Results

The results of the experiments are summarized in the following table. As shown, sub-character tokenization methods significantly improved perplexity compared to traditional character-based tokenization.

Table: Comparison of Perplexity for Different Tokenization Methods

Tokenization Method	Perplexity	Improvement (%)
Character-based	3.85	-
Radical-based	2.90	24.68
Element-based	2.82	26.75

---

5. Discussion

In this study, I investigated the impact of sub-character tokenization on Japanese language processing using the ALBERT model. By decomposing kanji into radicals and elements, the model's ability to capture semantic and phonetic nuances improved, as reflected in the perplexity scores. This effect was particularly pronounced for kanji that are composed of multiple meaningful components, such as compound ideographs and phono-semantic characters.

The findings suggest that the choice of sub-character type plays a significant role in enhancing model performance. The radical and element-based tokenizations yielded better results than traditional character-based methods, highlighting the importance of selecting appropriate sub-character units for training models on Japanese text.

---

6. Conclusion

This research demonstrated that using sub-character tokenization methods, particularly those based on radicals and elements, can enhance the performance of language models like ALBERT in Japanese NLP tasks. The study's results indicate a notable improvement in perplexity, suggesting that sub-character information contributes to more effective learning.

Future research should explore the application of these findings to larger datasets and other NLP tasks. Additionally, further investigation is needed to determine the optimal sub-character tokenization method for different types of kanji and linguistic tasks.

Moreover, these methods may also prove effective not only in Japanese but also in other languages such as Chinese, Arabic, and possibly even English.

References

[1] ALBERT details on HuggingFace: https://huggingface.co/albert/albert-base-v2

[2] MeCab official site: http://taku910.github.io/mecab/

[3] SentencePiece GitHub repository: https://github.com/google/sentencepiece

[4] V. Nguyen, J. Brooke, and T. Baldwin, "Sub-character Neural Language Modelling in Japanese," in Proc. of the First Workshop on Subword and Character Level Models in NLP, Copenhagen, Denmark, Sep. 2017, pp. 148-153. https://aclanthology.org/W17-4122.pdf

[5] KanjiVG-radical: https://github.com/yagays/kanjivg-radical/tree/master