Japanese Morphology with RNN and CNN Architectures

Austin Blodgett

Georgetown University, Department of Linguistics

Abstract

This research applies several deep learning architectures to the task of Japanese morpheme segmentation. The experiments performed use a stacked bidirectional GRU, a deep 1-dimensional convolutional network, and a transfer learning task. The results in this paper show that a GRU RNN architecture outperforms other experiments, suggesting that long-distance knowledge is important for performing this task.

1. Introduction

The task of (morpheme or word) tokenization of East Asian languages is a necessary process that must be performed before many other NLP tasks. Morpheme tokenization, demonstrated in figure 1, is the process of taking raw text as input and segmenting that text into meaningful tokens, where each token is a root, suffix, or prefix.

そこで、どの店でもすぐに、在庫がなくなる。

⇓

そこ | で | 、 | どの | 店 | で | も | すぐ | に | 、 | 在庫 | が | なくなる | 。

Figure 1: Morpheme tokenization task

This task can be understood as a sequence prediction task. In this approach, demonstrated in figure 2, the input and output are sequences of the same length. A classifier predicts for each character a 0 or 1, where 1 indicates that character precedes a morpheme boundary and 0 is used otherwise. As a benefit to future work, this approach is also easily extendable to morpheme analysis by allowing the class 1 to instead be one of a list of classes for labelling other features of the morpheme.

Email address: `ajb341@georgetown.edu` (Austin Blodgett)

Preprint submitted to ANLY 590: Deep Learning *December 16, 2018*

そ こ で 、 ど の 店 で も す ぐ に 、 在 庫 が な く な る 。

⇓

0 1 1 1 0 1 1 1 1 0 1 1 1 0 1 1 0 0 0 1 1

Figure 2: Morpheme tokenization as sequence prediction: 1 indicates a morpheme boundary; 0 indicates otherwise)

1.1. What makes this task Difficult?

Writing System: Japanese, like other East Asian languages, is written without spaces. Additionally, Japanese relies on 3+ writing systems (kanji, hiragana, katakana) which contributes to the problem of *character sparsity* discussed below..

Synthetic Language: Japanese is a *synthetic* (specifically *agglutinative*) language—a word is often composed of many meaningful parts and word formation is highly productive. Japanese tokenization at the word level can lead to a large percentage of words being out-of-vocabulary because of the process of word formation in Japanese. So, unlike with English or Mandarin, it may be more beneficial in Japanese to tokenize at the **morpheme** level instead of the word level.

Character Sparsity: Japanese text uses 6,000+ characters with a non-uniform, long-tail distribution. These features together contribute to data sparsity at the character level, which can make training deep learning on character input more difficult.

Ambiguous Boundaries: Token boundaries in Japanese can be ambiguous, in which case there is no exact solution.

2. Related Work

Other researchers have relied on Bayesian models (Mochihashi et al., 2009), Conditional Random Fields (Kudo et al., 2004), and Markov Models (Nakagawa, 2004). These tasks tend to rely on dictionary lookup and probabilistic models with necessary locality constraints such as the Markov assumption and its extensions.

To this researchers knowledge, deep learning has not yet been applied to this task.

3. Datasets

The Balanced Corpus of Contemporary Written Japanese or *BCCWJ* (Maekawa, 2007) is a corpus of Japanese text, annotated for morphological features. BCCWJ includes 100 million tokens. This research extracts 900,000+ unique sentences and preprocesses them.

4. Methods

This research focuses on 3 experiments. An RNN model, a CNN model, and an RNN with transfer learning from a related training task. Each architecture uses a similar structure—3 stacked layers plus character embeddings, a vocabulary size of 6,000, and dropout. The layer dimensions are 256, 64, 64, and 1.

One question that was of interest during this research was whether long-distance information or local information was more important for this task—if all that matters for tokenization is the characters immediately to the right or left of the boundary than local information is more important; if resolving boundary ambiguities requires having knowledge about the entire sentence, then long distance information is important.

This researcher expects a CNN model to better represent local information and an RNN to better represent long-distant information, since that is how the architectures are designed. The results in the next section demonstrate that the RNN architecture performs the best, suggesting that long distance information is important for this task.

4.1. RNN Experiment: Stacked Bidirectional GRU

Experiment 1 uses a Bidirectional Stacked RNN with the GRU architecture achieves the best results on test data. Success of RNN suggests the importance of long-distance information in this task.

4.2. CNN Experiment: 1-dim CNN

Experiment 2 uses a deep 1D CNN was trained with parallel features to Experiment 1. Additionally, for this experiment the input is padded to keep sequences the same length. One would expect CNNs to perform the best if character proximity was the most important factor.

4.3. Transfer Learning Experiment: Use of embeddings from a token level task

Experiment 3 uses transfer learning from a simpler, related task. To transfer learn features for the Japanese morphology task, a simple GRU model was trained with embeddings on a simple, related task (the same task at the word level, instead of sentence level). The embeddings were then transferred to the same architecture from Experiment 1.

5. Results

The results of the experiments, shown in table 1, reveal the types of structure that is necessary for this task. Experiment 1 outperforms Experiments 2 and 3 in F1 score, suggesting that some long distance dependency (as opposed to local information) is necessary to succeed in the task.

Results from Experiment 1 show the best results on test, while Experiments 2 and 3 show the best recall and precision respectively.

	Accuracy (%)	F1 (%)	Precision (%)	Recall (%)
RNN	**77.8**	**80.1**	84.2	77.7
CNN	68.0*	71.9*	63.1*	**85.8**
Transfer	76.6	77.9	**87.0**	71.6

Table 1: Results for RNN (§4.1), CNN (§4.2), and Transfer (§4.3) learning experiments. The max of each column is bold. Since input to CCNs is padded, scores marked with (*) may be less reliable, because of dummy predictions.

6. Discussion of Results

As discussed above, Experiment 1 (§4.1) using a stacked bidirectional GRU outperforms the other two experiments in terms of F1 score. These results suggest the importance of long-distance information in Japanese morpheme tokenization. RNN architectures are designed to capture long-distance information in sequences. Additionally, a bidirectional model means that long-distance information from either direction is used. Intuitively, importance of long-distance information in this task suggests that many morpheme boundaries are ambiguous and that identifying morpheme boundaries requires some knowledge of an entire sentence.

7. Conclusions

Long-distance knowledge is important for morpheme tokenization of Japanese, though more experiments can be done to learn more of the trade-off between local and long-distant information.

In future work, it would be a good idea to incorporate other types of transfer learning, and to perform joint learning of segmentation with other morphological analysis tasks. As described in section 1, it is easy to extend morpheme tokenization to morpheme classification or analysis. Trying an ensemble model between RNN and CNN architectures to capture multiple types of information would be a good area of research. It would also be interesting to experiment with other transfer learning tasks, such as training on data balanced by character type to fight character sparsity.

References

Kudo, T., Yamamoto, K., & Matsumoto, Y. (2004). Applying conditional random fields to japanese morphological analysis. In *Proceedings of the 2004 conference on empirical methods in natural language processing*.

Maekawa, K. (2007). Kotonoha and bccwj: development of a balanced corpus of contemporary written japanese. In *Corpora and Language Research: Proceedings of the First International Conference on Korean Language, Literature, and Culture* (pp. 158–177).

Mochihashi, D., Yamada, T., & Ueda, N. (2009). Bayesian unsupervised word segmentation with nested pitman-yor language modeling. In *Proceedings of the Joint Conference of the 47th Annual Meeting of the ACL and the 4th International Joint Conference on Natural Language Processing of the AFNLP: Volume 1-Volume 1* (pp. 100–108). Association for Computational Linguistics.

Nakagawa, T. (2004). Chinese and japanese word segmentation using word-level and character-level information. In *Proceedings of the 20th international conference on Computational Linguistics* (p. 466). Association for Computational Linguistics.