Japanese Morphology With RNN and CNN Neural Architectures

Austin Blodgett
Dept. of Linguistics, Georgetown University

Research Problem

Morphological Segmentation and Analysis is often necessary as an NLP pre-processing step for languages like Japanese.

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|。

Task can be understood as predicting (for each character) an end-of-morpheme boundary. The task becomes a sequence classification problem.

0 1 1 1 0 1 1 1 1 0 1 1 1 0 1 1 0 0 0 1
(1 for morpheme boundary, 0 otherwise)

Why is Japanese Morphological Segmentation Difficult?

1. Features of Japanese:
   - Japanese relies on **3+ writing systems** *(kanji, hiragana, katakana)* and is written is written without spaces.
   - Japanese is an **agglutinative** language; a word in Japanese is often composed of many meaningful parts. I.e., segmenting words isn't necessarily useful.
2. Data Sparsity: 6,000+ characters with non-uniform (long-tail) distribution. These features together contribute to data sparsity in the morpheme segmentation task.
3. Ambiguity: Some sentences are ambiguous and have no exact solution.

Data

The Balanced Corpus of Contemporary Written Japanese (**BCCWJ**) 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.

Experiments

This research focuses on 3 experiments. A baseline RNN model, a CNN model, and an RNN with transfer learning from a related training task.

Experiment 1 (RNN): A Bidirectional Stacked RNN with the GRU architecture achieves the best results on test data. Success of RNNs for this task is expected, but leaves room for improvement.

Experiment 2 (CNN): A Deep 1D CNN was trained with parallel features to Experiment 1. one would expect CNN**s** to perform the best if character proximity was the most important factor.

Experiment 3 (RNN+transfer learning): 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.

Each Architecture uses 3 stacked layers plus character embeddings, a vocabulary size of 6,000, and dropout. (layer dimensions: 256; 64; 64; 1)

Github: https://github.com/ablodge/JapaneseMorphy

Related Work

Other researchers have relied on Bayesian models [2], Conditional Random Fields [1], and others. It is unclear whether DL methods will eventually outperform these.

[1] Kudo, T., Yamamoto, K., & Matsumoto, Y. (2004). Applying conditional random fields to Japanese morphological analysis. In *Proc. of the 2004 conference on EMNLP*.
[2] Mochihashi, D., Yamada, T., & Ueda, N. (2009). Bayesian unsupervised word segmentation with nested Pitman-Yor language modeling. In *Proc. of ACL*.
[3] Nakagawa, T. (2004, August). Chinese and Japanese word segmentation using word-level and character-level information. In *Proc. of the 20th international conference on Computational Linguistics* (p. 466). ACL.

Research Contribution

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

These results can still be improved. 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.

Results

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

*Since input to CCNs is padded, scores marked with (*) may be less reliable, because of dummy predictions.

	Accuracy	F1	Precision	Recall
RNN	**0.778**	**0.801**	0.842	0.777
CNN	0.680*	0.719*	0.631*	**0.858**
Transfer	0.766	0.779	**0.870**	0.716

Contact
Austin Blodgett
Georgetown University
Dept. of Linguistics
Email: ajb341@Georgetown.edu