Chinese Word Segmentation using Bi-LSTMs

Amr Aly (1848399)

Problem Definition

Chinese word segmentation is the task of splitting Chinese text (a sequence of Chinese characters) into words.

Baseline Model

The baseline model described in [1] was implemented using keras and the suggested hyperparameters in Table 1, except for the optimizer, Adam, which showed faster convergence. A dictionary was created using all *unique* unigrams and bigrams extracted from all datasets. Sentences were stripped from spaces and vectorized − using unigrams and bigrams, while labels were OHE. Padding was performed on per-batch-basis to support training/predicting on variable-length sentences. Padding zeros were masked on all model layers.

Language Model

As suggested by [1], the use of a pre-trained language model would improve the model's performance. Hence, BERT [2], was used. Example codes from the authors were modified to obtain character-based contextualized embeddings and support sentences longer than 512 characters. However, more work should be done to fine-tune the model on our datasets and extract bigram embeddings from the model.

I was only able to train MSR and CITYU datasets on BERT, because of the large number of model parameters. Pruning the computation graph is feasible solution to the problem.

Extra Work

It was noticed when training and testing the model on different datasets, the performance suffers. One reason to this, is the different segmentation criteria used by each dataset. As suggested by [3] [4], one solution is to use *multitask learning*, a *task* here is defined to be: learning the segmentation according to some criteria. This can be done by introducing a shared domain projection layer which embeds common knowledge across all datasets, and a private domain layer that captures different segmentation criteria of each dataset, and combine the output of both layers. Unfortunately, there was not enough time available to implement this method.

References

[1] Ma, J., Ganchev, K. and Weiss, D., 2018. State-of-the-art Chinese word segmentation with bi-lstms. arXiv preprint arXiv:1808.06511.
[2] https://github.com/google-research/bert
[3] Chen, X., Shi, Z., Qiu, X. and Huang, X., 2017. Adversarial multi-criteria learning for chinese word segmentation. arXiv preprint arXiv:1704.07556.
[4] Huang, W., Cheng, X., Chen, K., Wang, T. and Chu, W., 2019. Toward Fast and Accurate Neural Chinese Word Segmentation with Multi-Criteria Learning. arXiv preprint arXiv:1903.04190.

Embedding Layer − Unigrams	128	Hidden layer size	256
Embedding Layer − Bigrams	128	Learning rate	0.001
Dropout (LSTM & Recurrent)	0.2	Batch size	32

Table 1 Baseline model hyperparameters

Dataset	PKU	MSR	AS (Simplified)	CITYU (Simplified)
Precision	0.9389	0.9705	0.9530	0.9566

Table 2 Baseline model results
(trained and tested on each dataset separately)

Dataset	PKU	MSR	AS (Simplified)	CITYU (Simplified)
Precision	0.9253	0.9530	0.8952	0.9257

Table 3 Combined baseline model results
(PKU + MSR + CITYU)

Dataset	PKU	MSR	AS (Simplified)	CITYU (Simplified)
Precision	0.8914	?	?	0.9426

Table 4 BERT model results
(trained and tested on each dataset separately)