State-of-the-art Chinese Word Segmentation with Bi-LSTMs

Michał Ostyk-Narbutt (1854051)
Prof. Roberto Navigli
Natural Language Processing Homework 1

April 10, 2019

SAPIENZA
UNIVERSITÀ DI ROMA

Contents

1 Introduction

2 Model

3 Results

4 Conclusion

The student has to deliver via the Google form: a link to the Gitlab shared project with the source code and any additional data needed to run the software a paper of up to 4 pages (+infinite pages for references, images, tables, graphs, etc.) including: a brief introduction to the project problem, a brief state of the art, an illustration of the methods/approach/techniques (min. 1 page), a quantitative (and ideally a small qualitative) evaluation of the system, some analysis of the results.

Algorithm 1 depicts the overall procedure of training.

Algorithm 1 PPO with Clipped Objective

1: Collect a batch of N (multiple of the mini-batch size) transitions from parallel environments (state, action, log-probabilities,a reward, done-mask (0 if terminal), V(s) (value of the state for each state).
2: Calculate the returns for the batch using GAE
3: Calculate: advantage = returns - values
4: For e epochs: *loop*:
    1: Sample through enough random mini-batches to cover all data.
    2: Pass state into network, obtain action, value of state', entropy and new-log-probabilities.
    3: Calculate the surrogate policy loss and MSE value loss.
    4: Backpropagate the loss through the network using SGD.
5: 5. Repeat above until converged.

References

[1] `https://github.com/Ostyk/Chinese-LSTM`