NLP
Assignment 1
Chinese Word Segmentaion

Hossam Arafat - Matricola: 1803850

24 April 2019

SAPIENZA
UNIVERSITÀ DI ROMA

1 Introduction

This report details the process of building a Bi-LSTM for Chinese Word Segmentation. The model is trained on the "ICWB2" Dataset. [1].

The Network's hyper-parameters, π^{LSTM}, namely, the learning rate μ, Batch Size b, and finally, E the size of the Embedding Layers, were found using a Grid Search procedure.
Grid Search sweeps μ through the values $0.035 \rightarrow 0.04$, b in between $32 \rightarrow 320$ and finally, E through the range between $32 \rightarrow 128$.
At each iteration of Grid Search, the SVM is trained on the current values of the hyper-parameters and the optimal values are are simply the ones which yield the minimum training error.

1.1 Model

Un-Stacked Bi-Directional LSTM was used. The unigrams and Bigrams of each character are fed to the network's embedding layer.

2 Conclusion

The results were in par with the finidings discussed in the original paper "SOTA Chinese Word Segmentation using BiLSTMs".

[1] *Dataset described in detail in Appendix A*

3 Appendix A

Ex.	μ	*Optimizer*	b	Train Acc	Dev Acc	Time per Batch
1	0.04	*ADAM*	32	94.2%	91.3%	60 min.
2	.004	*SGD*	200	95.2%	91.2%	13 min.

Table 1: Settings and Results for Questions 1,2 and 3

HiddenUnits	*UnigramsVocabSize*	*BigramsVocabSize*	*EmbeddingSize*	*BatchSize*
$2e-3$	6592	1042976	32	200

Table 2: Final Settings and Hyper-parameters of the Network

Name	*AS*	*MSR*	*CITYU*	*PKU*
Training Set	1005	731	542	11
Dev Set	2640	1980	1660	1200

Table 3: ICWB-2 Dataset