Bi-LSTM Chinese POS Tagging

https://github.com/tracy2811/chinese-pos-tagging-bi-lstm

Trang Nguyen - BS18-DS01

Table of Contents

Bi-LSTM Chinese POS Tagging

https://github.com/tracy2811/chinese-pos-tagging-bi-lstm

Trang Nguyen - BS18-DS01

Why Chinese Part-of-Speech (POS) Tagging?

- A prerequisite task to simplify many different NLP problems
  - Text to speech conversion
  - Word sense disambiguation
  - ...
- An intriguing task
  - Chinese word is not demarcated
  - Word Segmentation (WS) task is vital
- Helpful for language learners
  - “...常常他们的盐一样咸的咸菜...“《活着》余华
  - To taste their salt or their pickles?
- We already have Jieba, kcws? Educational Purpose!

Dataset

UD_Chinese-GSDSimp dataset from Universal Dependencies

15 UPOS tags (out of 17 possible)

Open class word	Closed class word	Other
ADJ, ADV, NOUN, PROPN, VERB	ADP, AUX, CCONJ, DET, NUM, PART, PRON	PUNCT, SYM, X

	Train	Test	Total
#Sentences	3997	500	4997

Preprocessing

Mapping: token - index, POS tag - index

- For POS-tagging-only models (LSTM, Bi-LSTM)
  - No extra step before mapping
  - No new POS tag
  - One token may contain one or more characters.
- For Joint WS and POS tagging model (Joint Bi-LSTM)
  - One character = one token
  - NC tag = this character joins with the previous character

Preprocessing - Example

	然	而	,	这	样	的	处	理	也	衍	生	了	一	些	问	题	。
Or	ADV		PUNCT	PRON		PART	NOUN		ADV	VERB		AUX	ADJ		NOUN		PUNCT
Bi	ADV	NC	PUNCT	PRON	NC	PART	NOUN	NC	ADV	VERB	NC	AUX	ADJ	NC	NOUN	NC	PUNCT

Models

3 models:

1. LSTM
2. Bi-LSTM
3. Joint Bi-LSTM

Same structure - 3 layers:

1. Embedding
2. LSTM (bidirectional or not)
3. Linear

EMBEDDING_DIM = 64

HIDDEN_DIM = 64

BiLSTMTagger(

(word_embeddings): Embedding(3514, 64)

(lstm): LSTM(64, 64, bidirectional=True)

(hidden2tag): Linear(in_features=128, out_features=16, bias=True)

)

Train and Test Results

Loss function: negative log likelihood loss NLLLoss
Optimizer: gradient descent optimizer SGD
Metric: F1 score
Number of Epochs: 10

Model	F1 average	F1
LSTM	0.7933	0.7930
Bi-LSTM	0.8185	0.8140
Joint Bi-LSTM	0.8461	0.8471

Train and Test Results - Example LSTM, Bi-LSTM result

他/担任/多/媒体/音乐/剧/《/琥珀/》/以及/儿童/狂欢/剧/《/魔山/》/的/音乐/总监/。

['PRON', 'VERB', 'PART', 'NOUN', 'NOUN', 'PART', 'PUNCT', 'NOUN', 'PUNCT', 'CCONJ', 'NOUN', 'VERB', 'PART', 'PUNCT', 'NOUN', 'PUNCT', 'PART', 'NOUN', 'NOUN', 'PUNCT']

LSTM

['PRON', 'VERB', 'NUM', 'NOUN', 'NOUN', 'NOUN', 'PUNCT', 'NOUN', 'PUNCT', 'CCONJ', 'NOUN', 'NOUN', 'NOUN', 'PUNCT', 'PART', 'PUNCT', 'PART', 'NOUN', 'NOUN', 'PUNCT']

0.7458333333333333

Bi-LSTM

['PRON', 'VERB', 'NUM', 'NOUN', 'NOUN', 'NOUN', 'PUNCT', 'NOUN', 'PUNCT', 'CCONJ', 'NOUN', 'NOUN', 'PART', 'PUNCT', 'NOUN', 'PUNCT', 'PART', 'NOUN', 'NOUN', 'PUNCT']

0.85625

Train and Test Results - Example Joint Bi-LSTM result

Sentence: 由于认为德义军队战力已就绪，隆美尔决定再发动攻势。

Tags: ['ADP', 'NC', 'VERB', 'NC', 'PROPN', 'PROPN', 'NOUN', 'NC', 'NOUN', 'NC', 'ADV', 'VERB', 'NC', 'PUNCT', 'PROPN', 'NC', 'NC', 'VERB', 'NC', 'ADV', 'VERB', 'NC', 'NOUN', 'NC', 'PUNCT']

Predicted: ['ADP', 'NC', 'VERB', 'NC', 'PROPN', 'NC', 'NOUN', 'NC', 'NOUN', 'NC', 'ADV', 'VERB', 'NC', 'PUNCT', 'PROPN', 'NC', 'NC', 'NOUN', 'NC', 'ADV', 'VERB', 'NC', 'NOUN', 'NC', 'PUNCT']

F1 score: 0.916952380952381

Train and Test Results - Tuning Hyperparameters

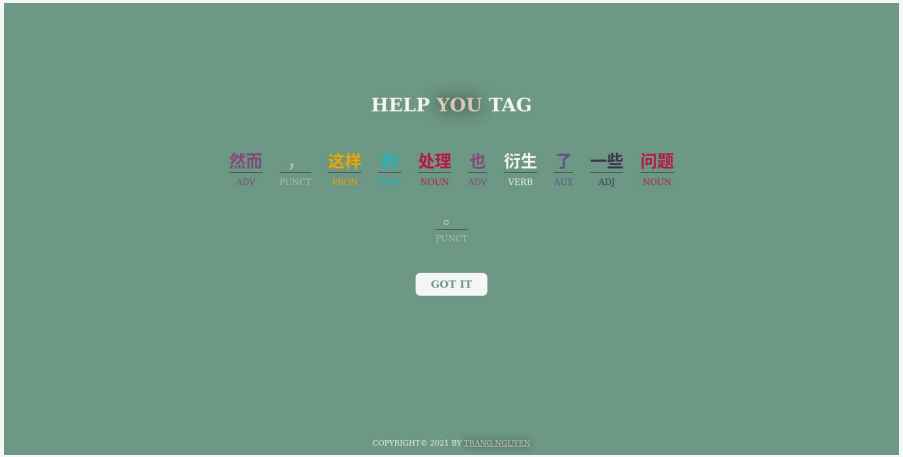
Hyperparameter	Value
Embedding dimension	256
Hidden dimension	256
Learning rate	0.0145183

#Epochs	20
F1 average	0.8685
F1	0.8698

Web Demo

Joint Bi-LSTM model + Flask

Enter a sentence



Get result

Web Demo

https://youtu.be/3Jd_Q6_Zi3M

Future work

- Tuning: embedding dimension, hidden dimension
- Adding Attention
- Adding more features to the Web app, such as showing definition of each token

References

- https://www.kaggle.com/krishanudb/lstm-character-word-pos-tag-model-pytorch
- https://towardsdatascience.com/joint-khmer-word-segmentation-and-pos-tagging-cad650e78d30
- https://github.com/crownpku/awesome-chinese-nlp

Thank you!