Task 6. Recognize Ontology in AGAC corpus by BERT and CRF

Github link: https://github.com/Allen-ZKW/NLP_HZAU/tree/task6

Author: Kewei Zhao
Date: 2021-5-20

Abstract

In this task, we build virtual environment to fit codes which is supported by teacher Xia. We mainly have two targets when we doing this task: 1) Run scripts successfully. 2) Read these scripts to understand BERT-model and CRF-model

Principle

BERT model is a pre-training model which will return a result (one word matches one vector) which is base on training data. But training the model and getting the vectors will cost a huge sum of time and calculation. So in this task, we just import "Transformers" module to use BERT-model which is already trained by other researchers.

Measure

Install Mini Conda

```
wget https://repo.anaconda.com/miniconda/Miniconda3-latest-Linux-x86_64.sh
sh Miniconda3-latest-Linux-x86_64.sh
```

Create and Activate Virtual Environment

```
conda create -n NLP_task6 python=3.8
conda activate NLP_task6
conda config --add channels
https://mirrors.tuna.tsinghua.edu.cn/anaconda/cloud/msys2/
conda config --add channels
https://mirrors.tuna.tsinghua.edu.cn/anaconda/cloud/conda-forge/
conda config --add channels
https://mirrors.tuna.tsinghua.edu.cn/anaconda/pkgs/free/
```

Download Required Modules

Because we need to download pytorch to fit our environment, we choose to download pytorch independently and remove the part of pytorch in requirments.txt

```
conda install pytorch torchvision torchaudio cpuonly -c pytorch
conda install --yes --file requirements.txt
```

Run the Script

```
python main.py
```

Abstract Information From Conlleval.log

```
cat conlleval.log|grep "accuracy.*precision" > result.txt
sz result.txt
```

Result

The first table is our result, the second one is suggested results.

Model	accuracy	precision	recall	FB1
BERT+CRF	95.77%	54.63%	56.46%	55.53

Model	Accuracy	Precision	Recall	F1-score
BERT+CRF	**95.7730**	**54.6274**	**56.4596**	**55.5284**

Discussion

In the result above, we successfully run the scripts. In this task, we also feel the advantages of BERT, BERT model will give every words in text a very good vector instead of a random vector. Because of init-vector, our CRF model can be trained faster and the accuracy of result is better than before.

Of course, in this task, we also meet many problems. These problems mainly exist in environment and module part. First, in conda environment, we cannot use "pip install" but use "conda install". Secondly, relation between pytorch and environment is very complicated, so we visit https://pytorch.org/ and get command we need.