作业五. Wapiti序列标注的CRF实现和.pat模板测试

生信1802-余思克-2018317220208

实验目的

Wapiti是一个用于训练和使用具有弹性罚分的具有各种算法的判别性序列标记模型的工具。 目前，它实现了maxent模型，最大熵马尔可夫模型（MEMM）和线性链条件随机场（CRF）模型。本次实验将通过Wapiti工具进行序列标注的CRF模型实现。

实验步骤

1. Wapiti工具的安装

本次实验的CRF模型实现是通过Wapiti工具完成的，故首先需要安装该工具，Wapiti的安装有两种方法，第一种是通过`git clone`命令从GitHub获取Wapiti文件，然后手动进行构建与安装。使用`wapiti`命令，若可以出现帮助文档则说明安装成功。

```
git clone https://github.com/Jekub/Wapiti.git
cd Wapiti
sudo make
sudo make install # 使用git clone命令从GitHub获取Wapiti文件并手动安装
```

在Wapiti的安装过程中，可能出现许多报错信息，如果使用`git clone`命令时，可能出现无法连接GitHub服务器的情况，此时使用git config命令重置代理即可。

```
git config --global  --unset https.https://github.com.proxy
git config --global  --unset http.https://github.com.proxy # 重置代理
```

若Wapiti不能正常安装，可能是因为缺乏GCC编译器，需要使用`sudo apt install`命令进行安装。

```
sudo apt install gcc
```

2. 数据准备

待训练的数据是AGAC语料库中的数据，使用`git clone`命令从Github获取数据，并解压文件夹中的AGAC压缩包。

```
git clone https://github.com/ouyangsizhuo/2021Spring_CRF_AGACtask1.git # 获取数据文件
```

3. 训练模型

使用`wapiti train`命令进行模型训练。其中，-a参数用于指定训练模型使用的算法；-t参数用于指定训练模型使用的并行线程数；-i用于指定训练算法的最大迭代次数；-p参数用于指定pat模板文件的地址。

```
wapiti train -a sgd-l1 -t 8 -i 50 -p pat/Tok321dis.pat <(cat
AGAC/train_split/*.txt) AGAC/mod/AGAC_train.mod # 模型训练
```

4. 预测标签

通过训练出的模型，对数据进行标签预测。其中-m参数用于指定上一步训练出来的模型。

```
wapiti label -c -m AGAC/mod/AGAC_train.mod <(cat AGAC/test_split/*.txt)
AGAC/train_out.tab # 标签预测
```

5. 评估结果

通过将训练出来的模型预测的标签与正确结果进行对比，从而评估模型的预测结果。

```
perl conlleval.pl -d $'\t' <AGAC/train_out.tab | tee AGAC/train_out.eval # 结果评
估
```

6. 参数调节.pat模板测试

尝试修改模型训练参数以及修改.pat模板中的参数从而提高模型的评估能力。

```
vim Tok321dis.pat # 修改.pat模板中的参数
```

实验结果

通过对模型训练命令中不同参数的尝试，我们发现训练迭代次数超过50次之后，模型的错误率明显降低并趋于稳定，于是使用50次迭代作为后续分析的标准参数。

```
[  45] obj=NA act=47591    err= 0.73%/ 8.54% time=1.52s/94.23s
[  46] obj=NA act=47590    err= 0.76%/ 8.81% time=1.53s/95.75s
[  47] obj=NA act=47585    err= 0.73%/ 8.50% time=1.51s/97.26s
[  48] obj=NA act=47584    err= 0.76%/ 8.81% time=1.51s/98.78s
[  49] obj=NA act=47574    err= 0.79%/ 9.09% time=1.51s/100.29s
[  50] obj=NA act=47574    err= 0.76%/ 8.81% time=1.52s/101.81s
Save the model
Done
```

1. 默认.pat模板的模型预测结果

默认.pat模板参数如下。

```
U:tok:1:-1:%X[-1,0]
U:tok:1:+0:%X[0,0]
U:tok:1:+1:%X[1,0]

U:tok:2:-1:%X[-1,0]/%X[0,0]
U:tok:2:+0:%X[0,0]/%X[1,0]

U:tok:3:-2:%X[-2,0]/%X[-1,0]/%X[0,0]
U:tok:3:-1:%X[-1,0]/%X[0,0]/%X[1,0]
U:tok:3:+0:%X[0,0]/%X[1,0]/%X[2,0]


U:pre:1:+0:4:%M[0,0,"^.?.?.?.?"]

U:suf:1:+0:4:%M[0,0,".?.?.?.?$"]

U:dis:1:-1:%X[-1,2]
U:dis:1:+0:%X[0,2]
U:dis:1:+1:%X[1,2]
```

通过50次训练算法迭代，预测结果如下。

```
processed 62559 tokens with 2416 phrases; found: 822 phrases; correct: 299
accuracy:  92.14%; precision:  36.37%; recall:  12.38%; FB1:  18.47
              CPA: precision:  13.64%; recall:   5.36%; FB1:   7.69  22
          Disease: precision:  23.81%; recall:   9.57%; FB1:  13.65  168
           Enzyme: precision:   0.00%; recall:   0.00%; FB1:   0.00  4
             Gene: precision:  41.30%; recall:   4.08%; FB1:   7.42  46
      Interaction: precision:  25.00%; recall:  14.29%; FB1:  18.18  4
              MPA: precision:  14.71%; recall:   4.98%; FB1:   7.43  68
           NegReg: precision:  58.93%; recall:  25.78%; FB1:  35.87  56
          Pathway: precision:   0.00%; recall:   0.00%; FB1:   0.00  2
           PosReg: precision:  28.89%; recall:  15.85%; FB1:  20.47  45
          Protein: precision:  16.67%; recall:   1.67%; FB1:   3.03  6
              Reg: precision:  67.16%; recall:  11.31%; FB1:  19.35  67
              Var: precision:  40.12%; recall:  23.63%; FB1:  29.74  334
```

2. 修改.pat模板参数后的模型预测结果

修改后的.pat模板参数如下，相较于默认模板，增加U:tok:4，U:tok:5:两层模板，并将U:tok:1与U:dis:1在原有基础上各增加了两个步长。

```
U:tok:1:-1:%X[-1,0]
U:tok:1:+0:%X[0,0]
U:tok:1:+1:%X[1,0]

U:tok:2:-1:%X[-1,0]/%X[0,0]
U:tok:2:+0:%X[0,0]/%X[1,0]

U:tok:3:-2:%X[-2,0]/%X[-1,0]/%X[0,0]
U:tok:3:-1:%X[-1,0]/%X[0,0]/%X[1,0]
U:tok:3:+0:%X[0,0]/%X[1,0]/%X[2,0]

U:tok:4:-3:%X[-3,0]/%X[-2,0]/%X[-1,0]/%X[0,0]
U:tok:4:-2:%X[-2,0]/%X[-1,0]/%X[0,0]/%X[1,0]
U:tok:4:-1:%X[-1,0]/%X[0,0]/%X[1,0]/%X[2,0]
U:tok:4:+0:%X[0,0]/%X[1,0]/%X[2,0]/%X[3,0]

U:pre:1:+0:4:%M[0,0,"^.?.?.?.?"]

U:suf:1:+0:4:%M[0,0,".?.?.?.?$"]

U:dis:1:-1:%X[-1,2]
U:dis:1:+0:%X[0,2]
U:dis:1:+1:%X[1,2]
```

经过50次训练算法迭代，预测结果如下。

```
processed 62559 tokens with 2416 phrases; found: 1123 phrases; correct: 427.
accuracy:  91.91%; precision:  38.02%; recall:  17.67%; FB1:  24.13
             CPA: precision:   5.00%; recall:   3.57%; FB1:   4.17  40
         Disease: precision:  31.22%; recall:  17.70%; FB1:  22.60  237
          Enzyme: precision:   0.00%; recall:   0.00%; FB1:   0.00  2
            Gene: precision:  46.72%; recall:  13.73%; FB1:  21.23  137
     Interaction: precision:  50.00%; recall:  14.29%; FB1:  22.22  2
             MPA: precision:  12.36%; recall:   5.47%; FB1:   7.59  89
         NegReg: precision:  47.87%; recall:  35.16%; FB1:  40.54  94
        Pathway: precision:   0.00%; recall:   0.00%; FB1:   0.00  6
         PosReg: precision:  27.59%; recall:  19.51%; FB1:  22.86  58
        Protein: precision:  20.00%; recall:   3.33%; FB1:   5.71  10
            Reg: precision:  66.28%; recall:  14.32%; FB1:  23.55  86
            Var: precision:  42.82%; recall:  27.34%; FB1:  33.37  362
```

讨论

通过对.pat模板参数的修改，适当增加了相关的参数之后，对比两次预测结果可以发现：被正确识别的词组也有了明显的提高（由299个提高到427个），预测非-O准确率也有所提高。通过观察预测结果可以发现，在大多数标签上的预测结果的精确率，召回率以及FB1值都有着明显的提高，少数指标出现了下降。总体来说，综合预测词组数目以及预测准确率指标上来看，修改.pat模板参数后的模型预测结果较之前在有着不错的提升，其模型更适用于序列标签的预测。

虽然通过参数的修改，预测结果出现了明显的提升，但是其预测效果仍然不太好，综合非-O准确率没能超过40%，说明.pat模板还需要进一步的增加与优化参数以使其训练出更好的模型。但是.pat模板中的参数不能无限随意增加，因为随着步长的拓展，其需要的运算量将会呈爆炸增长，很快就会使计算机算力无法达到其训练模型所需的计算要求，因此，为了提高预测效果，除了在本方法中调整调参策略，还应该尝试新颖的、想过更好的技术与方法（例如神经网络等）来完成序列标注任务。