基于 CRF 的药物副作用实体识别实验

2019/3/13 周开银

一、 实验环境配置(Linux or macOS)

1.1 配置虚拟环境

```
conda create -n python3.6 python=3.6 (创建虚拟环境 python3.6)
source activate python3.6 (打开虚拟环境)
source deactivate python3.6 (关闭虚拟环境)
```

Windows 系统请参考：https://conda.io/projects/conda/en/latest/user-guide/tasks/manage-environments.html

注：若 conda 命令不存在，请自行添加环境变量（Windows）

1.2 Wapiti 工具环境配置

```
source activate python3.6
mkdir yourProject (yourProject 是自定义的项目名)
cd yourProject
git clone https://github.com/Jekub/Wapiti.git
cd Wapiti
make
make install
./wapiti (有帮助文档输出表示安装成功)
```

注：若 make 失败，请安装 gcc

“Permission denied”:

vim Makefile

修改 PREFIX = .

1.3 工作环境配置

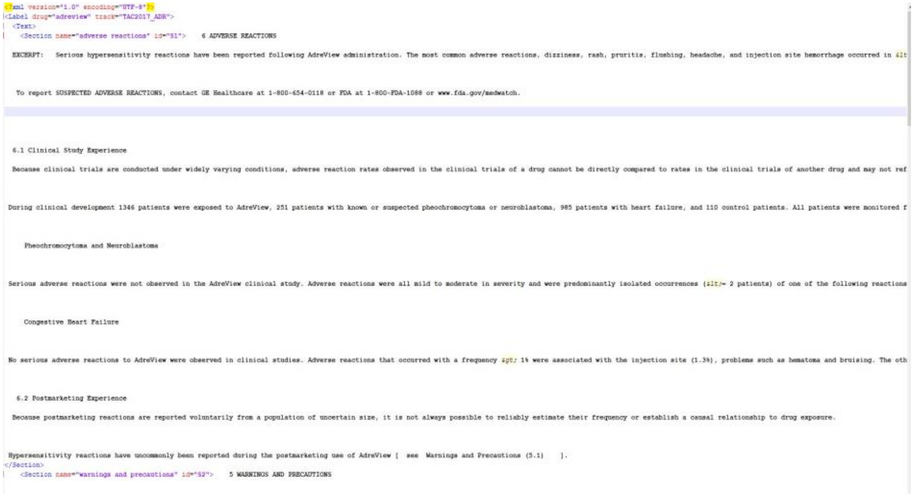
```
cd yourProject
git clone https://github.com/kyzhouhzau/2019SpringTextM.git
```

二、 数据描述

2.1 训练数据

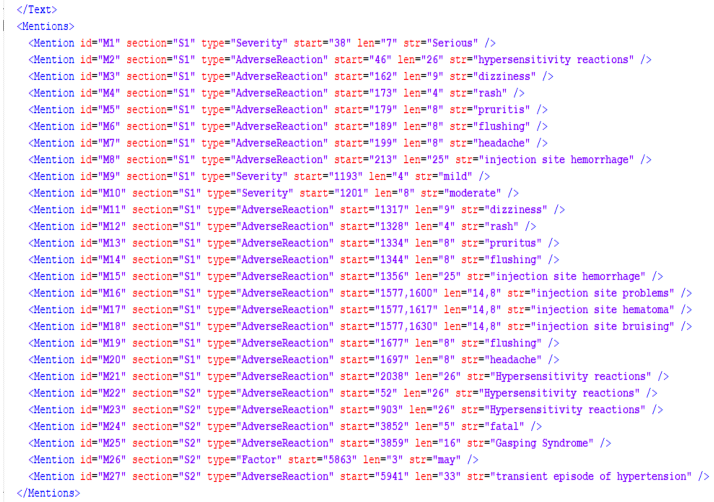
训练数据有 100 个 Drug Label，格式如下面所展示。

Text:



```
<?xml version="1.0" encoding="UTF-8"?>
<Label drug="adreview" track="TAC2017_ADR">
  <Text>
    <Section name="adverse reactions" id="S1">    6 ADVERSE REACTIONS

  EXCERPT:   Serious hypersensitivity reactions have been reported following AdreView administration. The most common adverse reactions, dizziness, rash, pruritis, flushing, headache, and injection site hemorrhage occurred in &lt

   To report SUSPECTED ADVERSE REACTIONS, contact GE Healthcare at 1-800-654-0118 or FDA at 1-800-FDA-1088 or www.fda.gov/medwatch.



  6.1 Clinical Study Experience

  Because clinical trials are conducted under widely varying conditions, adverse reaction rates observed in the clinical trials of a drug cannot be directly compared to rates in the clinical trials of another drug and may not ref


 During clinical development 1346 patients were exposed to AdreView, 251 patients with known or suspected pheochromocytoma or neuroblastoma, 985 patients with heart failure, and 110 control patients. All patients were monitored f


     Pheochromocytoma and Neuroblastoma


 Serious adverse reactions were not observed in the AdreView clinical study. Adverse reactions were all mild to moderate in severity and were predominantly isolated occurrences (&lt;= 2 patients) of one of the following reactions


     Congestive Heart Failure


 No serious adverse reactions to AdreView were observed in clinical studies. Adverse reactions that occurred with a frequency &gt; 1% were associated with the injection site (1.3%), problems such as hematoma and bruising. The oth


   6.2 Postmarketing Experience

  Because postmarketing reactions are reported voluntarily from a population of uncertain size, it is not always possible to reliably estimate their frequency or establish a causal relationship to drug exposure.


 Hypersensitivity reactions have uncommonly been reported during the postmarketing use of AdreView [  see  Warnings and Precautions (5.1)    ].
</Section>
    <Section name="warnings and precautions" id="S2">    5 WARNINGS AND PRECAUTIONS
```

Entity:



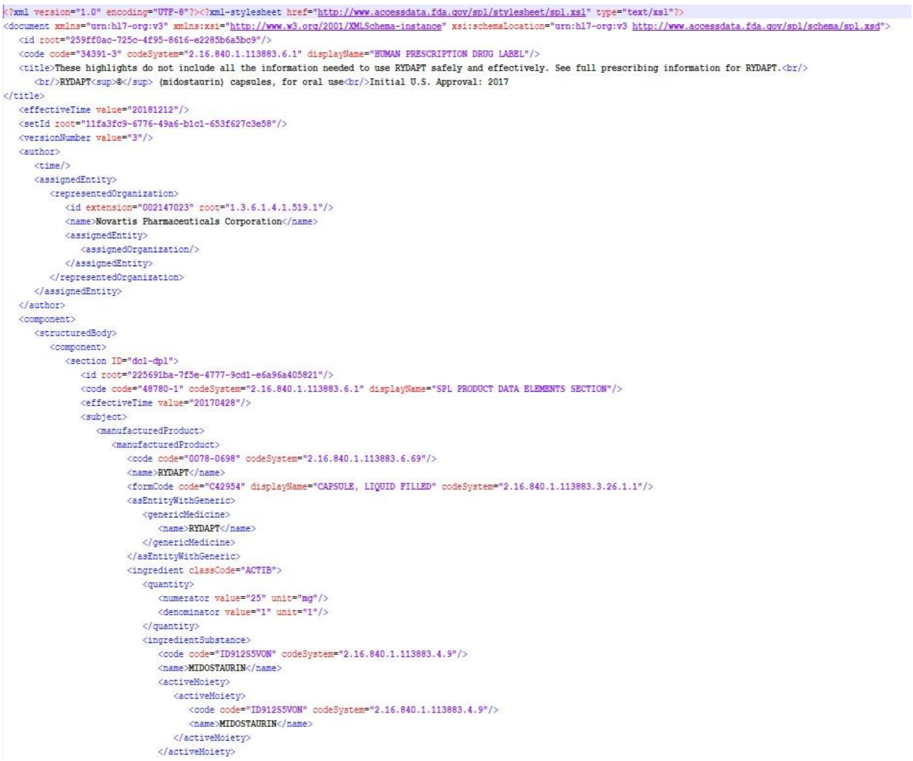
```
  </Text>
  <Mentions>
    <Mention id="M1" section="S1" type="Severity" start="38" len="7" str="Serious" />
    <Mention id="M2" section="S1" type="AdverseReaction" start="46" len="26" str="hypersensitivity reactions" />
    <Mention id="M3" section="S1" type="AdverseReaction" start="162" len="9" str="dizziness" />
    <Mention id="M4" section="S1" type="AdverseReaction" start="173" len="4" str="rash" />
    <Mention id="M5" section="S1" type="AdverseReaction" start="179" len="8" str="pruritis" />
    <Mention id="M6" section="S1" type="AdverseReaction" start="189" len="8" str="flushing" />
    <Mention id="M7" section="S1" type="AdverseReaction" start="199" len="8" str="headache" />
    <Mention id="M8" section="S1" type="AdverseReaction" start="213" len="25" str="injection site hemorrhage" />
    <Mention id="M9" section="S1" type="Severity" start="1193" len="4" str="mild" />
    <Mention id="M10" section="S1" type="Severity" start="1201" len="8" str="moderate" />
    <Mention id="M11" section="S1" type="AdverseReaction" start="1317" len="9" str="dizziness" />
    <Mention id="M12" section="S1" type="AdverseReaction" start="1328" len="4" str="rash" />
    <Mention id="M13" section="S1" type="AdverseReaction" start="1334" len="8" str="pruritus" />
    <Mention id="M14" section="S1" type="AdverseReaction" start="1344" len="8" str="flushing" />
    <Mention id="M15" section="S1" type="AdverseReaction" start="1356" len="25" str="injection site hemorrhage" />
    <Mention id="M16" section="S1" type="AdverseReaction" start="1577,1600" len="14,8" str="injection site problems" />
    <Mention id="M17" section="S1" type="AdverseReaction" start="1577,1617" len="14,8" str="injection site hematoma" />
    <Mention id="M18" section="S1" type="AdverseReaction" start="1577,1630" len="14,8" str="injection site bruising" />
    <Mention id="M19" section="S1" type="AdverseReaction" start="1677" len="8" str="flushing" />
    <Mention id="M20" section="S1" type="AdverseReaction" start="1697" len="8" str="headache" />
    <Mention id="M21" section="S1" type="AdverseReaction" start="2038" len="26" str="Hypersensitivity reactions" />
    <Mention id="M22" section="S2" type="AdverseReaction" start="52" len="26" str="Hypersensitivity reactions" />
    <Mention id="M23" section="S2" type="AdverseReaction" start="903" len="26" str="Hypersensitivity reactions" />
    <Mention id="M24" section="S2" type="AdverseReaction" start="3852" len="5" str="fatal" />
    <Mention id="M25" section="S2" type="AdverseReaction" start="3859" len="16" str="Gasping Syndrome" />
    <Mention id="M26" section="S2" type="Factor" start="5863" len="3" str="may" />
    <Mention id="M27" section="S2" type="AdverseReaction" start="5941" len="33" str="transient episode of hypertension" />
  </Mentions>
```

2.2 测试数据

从 DailyMed (https://dailymed.nlm.nih.gov) 中下载药物标签 34 个，格式为 XML。

（已经下载好:2019SpringTextM/need_drug）

例：



```
<?xml version="1.0" encoding="UTF-8"?><?xml-stylesheet href="http://www.accessdata.fda.gov/spl/stylesheet/spl.xsl" type="text/xsl"?>
<document xmlns="urn:hl7-org:v3" xmlns:xsi="http://www.w3.org/2001/XMLSchema-instance" xsi:schemaLocation="urn:hl7-org:v3 http://www.accessdata.fda.gov/spl/schema/spl.xsd">
   <id root="259ff0ac-725c-4f95-8616-e2285b6a5bc9"/>
   <code code="34391-3" codeSystem="2.16.840.1.113883.6.1" displayName="HUMAN PRESCRIPTION DRUG LABEL"/>
   <title>These highlights do not include all the information needed to use RYDAPT safely and effectively. See full prescribing information for RYDAPT.<br/>
      <br/>RYDAPT<sup>®</sup> (midostaurin) capsules, for oral use<br/>Initial U.S. Approval: 2017
</title>
   <effectiveTime value="20181212"/>
   <setId root="11fa3fc9-6776-49a6-b1c1-653f627c3e58"/>
   <versionNumber value="3"/>
   <author>
      <time/>
      <assignedEntity>
         <representedOrganization>
            <id extension="002147023" root="1.3.6.1.4.1.519.1"/>
            <name>Novartis Pharmaceuticals Corporation</name>
            <assignedEntity>
               <assignedOrganization/>
            </assignedEntity>
         </representedOrganization>
      </assignedEntity>
   </author>
   <component>
      <structuredBody>
         <component>
            <section ID="dcl-dpl">
               <id root="225691ba-7f5e-4777-9cd1-e6a96a405821"/>
               <code code="48780-1" codeSystem="2.16.840.1.113883.6.1" displayName="SPL PRODUCT DATA ELEMENTS SECTION"/>
               <effectiveTime value="20170428"/>
               <subject>
                  <manufacturedProduct>
                     <manufacturedProduct>
                        <code code="0078-0698" codeSystem="2.16.840.1.113883.6.69"/>
                        <name>RYDAPT</name>
                        <formCode code="C42954" displayName="CAPSULE, LIQUID FILLED" codeSystem="2.16.840.1.113883.3.26.1.1"/>
                        <asEntityWithGeneric>
                           <genericMedicine>
                              <name>RYDAPT</name>
                           </genericMedicine>
                        </asEntityWithGeneric>
                        <ingredient classCode="ACTIB">
                           <quantity>
                              <numerator value="25" unit="mg"/>
                              <denominator value="1" unit="1"/>
                           </quantity>
                           <ingredientSubstance>
                              <code code="ID912S5VON" codeSystem="2.16.840.1.113883.4.9"/>
                              <name>MIDOSTAURIN</name>
                              <activeMoiety>
                                 <activeMoiety>
                                    <code code="ID912S5VON" codeSystem="2.16.840.1.113883.4.9"/>
                                    <name>MIDOSTAURIN</name>
                                 </activeMoiety>
                              </activeMoiety>
```

三、 数据处理

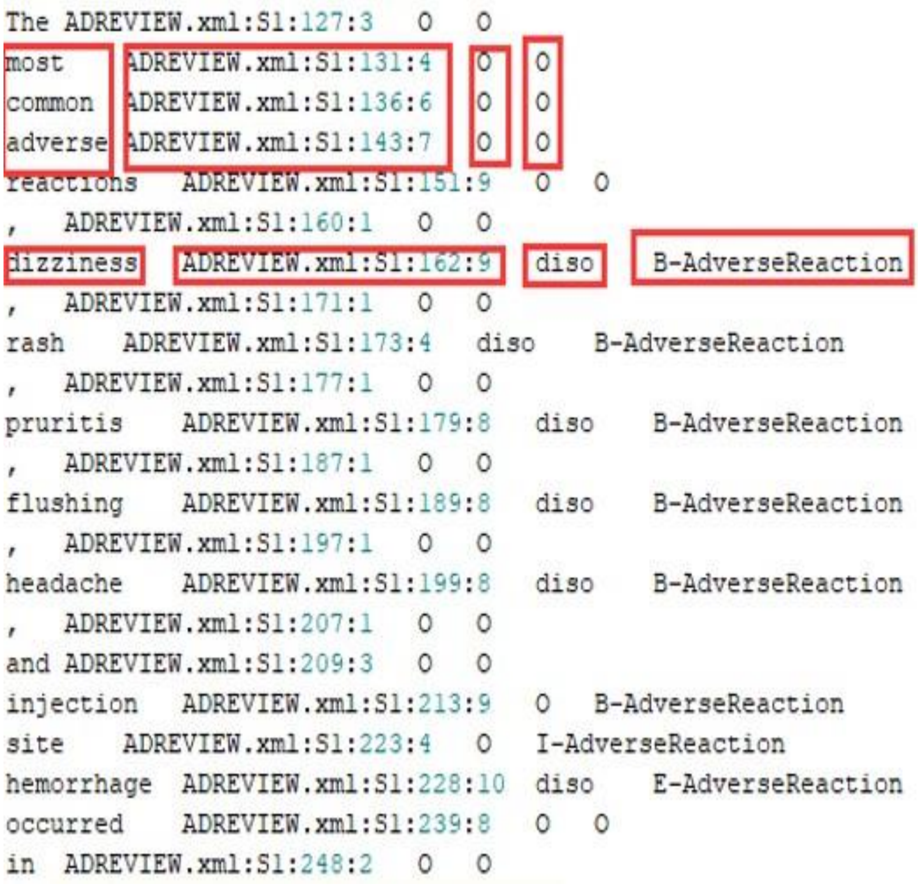
训练数据和测试数据的数据格式均为 XML，但不完全相同。在训练模型，和测试前均需要对数据进行预处理。将数据处理成”.tab”格式。

3.1 训练数据预处理

通过以下脚本预处理训练数据。

```
python tac2brat.py -d train_xml -o outtrain -F TokenDict:diso:diso-DISO.dic -t conll -s OBBEI

-d 训练文件夹
-o 预处理结果输出文件夹
-F 字典特征
-t 输出格式
-s 标签格式
```

输出格式：



```
The ADREVIEW.xml:S1:127:3    O    O
most    ADREVIEW.xml:S1:131:4    O    O
common  ADREVIEW.xml:S1:136:6    O    O
adverse ADREVIEW.xml:S1:143:7    O    O
reactions    ADREVIEW.xml:S1:151:9    O    O
,    ADREVIEW.xml:S1:160:1    O    O
dizziness    ADREVIEW.xml:S1:162:9    diso    B-AdverseReaction
,    ADREVIEW.xml:S1:171:1    O    O
rash    ADREVIEW.xml:S1:173:4    diso    B-AdverseReaction
,    ADREVIEW.xml:S1:177:1    O    O
pruritis    ADREVIEW.xml:S1:179:8    diso    B-AdverseReaction
,    ADREVIEW.xml:S1:187:1    O    O
flushing    ADREVIEW.xml:S1:189:8    diso    B-AdverseReaction
,    ADREVIEW.xml:S1:197:1    O    O
headache    ADREVIEW.xml:S1:199:8    diso    B-AdverseReaction
,    ADREVIEW.xml:S1:207:1    O    O
and ADREVIEW.xml:S1:209:3    O    O
injection    ADREVIEW.xml:S1:213:9    O    B-AdverseReaction
site    ADREVIEW.xml:S1:223:4    O    I-AdverseReaction
hemorrhage    ADREVIEW.xml:S1:228:10    diso    E-AdverseReaction
occurred    ADREVIEW.xml:S1:239:8    O    O
in    ADREVIEW.xml:S1:248:2    O    O
```

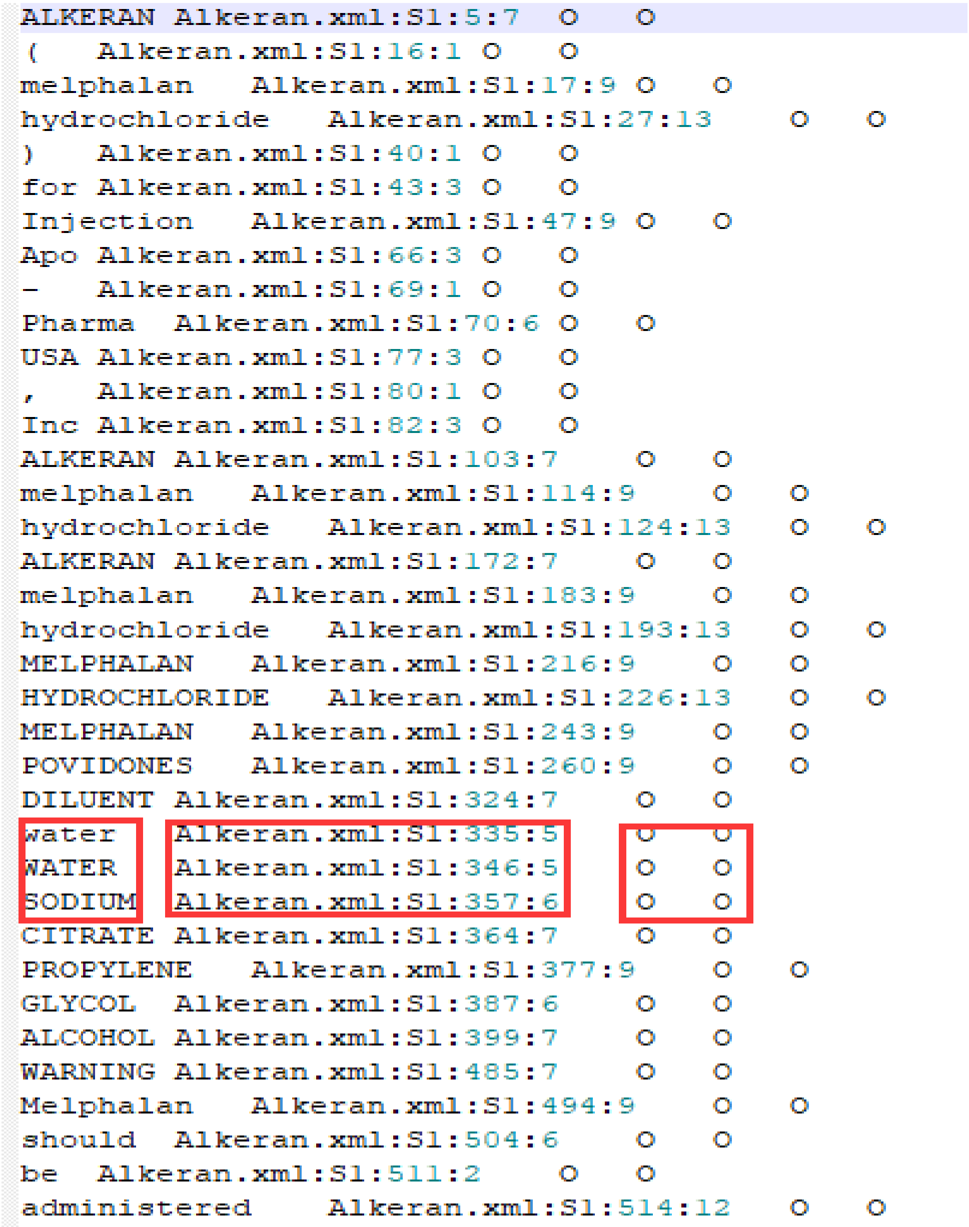
注：同一横行中第一个红框中是分词后的单词，第二个红框是文件名：所属段落：起始位置：单词长度，第三个红框是 字典特征，第四个红框是实体标签。在实体标签中对于某类标签如：标签 AdverseReaction 。若某个单词 “pruritis” 属于该标 签则被标注为"B-AdverseReaction";若某个词组“injection site hemorrhage”属于该标签，则该词组被标注为“B- AdverseReaction I-AdverseaeAction E-AdverseReaction”。于是，我们将该标签方式称为"BIEO"，此处 B-type 表示 begin, I-type 表示 Inside, E-type 表示 End, O 表示不属于该标签。

3.2 测试数据预处理

通过以下脚本处理测试数据。

```
cd 工作目录
python  to_xml_needed.py  测试文件夹  output  （注：在这里测试文件夹是 need_drug,表示我们从 dailyMed 下载的 XML 药物标签就放在该文件夹中。）
python tac2brat.py -d output  -o outtest -F TokenDict:diso:diso-DISO.dic -t conll -s OBBEI
```

输出格式：



```
ALKERAN Alkeran.xml:S1:5:7  O   O
(   Alkeran.xml:S1:16:1 O   O
melphalan   Alkeran.xml:S1:17:9 O   O
hydrochloride   Alkeran.xml:S1:27:13    O   O
)   Alkeran.xml:S1:40:1 O   O
for Alkeran.xml:S1:43:3 O   O
Injection   Alkeran.xml:S1:47:9 O   O
Apo Alkeran.xml:S1:66:3 O   O
-   Alkeran.xml:S1:69:1 O   O
Pharma  Alkeran.xml:S1:70:6 O   O
USA Alkeran.xml:S1:77:3 O   O
,   Alkeran.xml:S1:80:1 O   O
Inc Alkeran.xml:S1:82:3 O   O
ALKERAN Alkeran.xml:S1:103:7    O   O
melphalan   Alkeran.xml:S1:114:9    O   O
hydrochloride   Alkeran.xml:S1:124:13   O   O
ALKERAN Alkeran.xml:S1:172:7    O   O
melphalan   Alkeran.xml:S1:183:9    O   O
hydrochloride   Alkeran.xml:S1:193:13   O   O
MELPHALAN   Alkeran.xml:S1:216:9    O   O
HYDROCHLORIDE   Alkeran.xml:S1:226:13   O   O
MELPHALAN   Alkeran.xml:S1:243:9    O   O
POVIDONES   Alkeran.xml:S1:260:9    O   O
DILUENT Alkeran.xml:S1:324:7    O   O
water   Alkeran.xml:S1:335:5    O   O
WATER   Alkeran.xml:S1:346:5    O   O
SODIUM  Alkeran.xml:S1:357:6    O   O
CITRATE Alkeran.xml:S1:364:7    O   O
PROPYLENE   Alkeran.xml:S1:377:9    O   O
GLYCOL  Alkeran.xml:S1:387:6    O   O
ALCOHOL Alkeran.xml:S1:399:7    O   O
WARNING Alkeran.xml:S1:485:7    O   O
Melphalan   Alkeran.xml:S1:494:9    O   O
should  Alkeran.xml:S1:504:6    O   O
be  Alkeran.xml:S1:511:2    O   O
administered    Alkeran.xml:S1:514:12   O   O
```

注：格式同训练数据

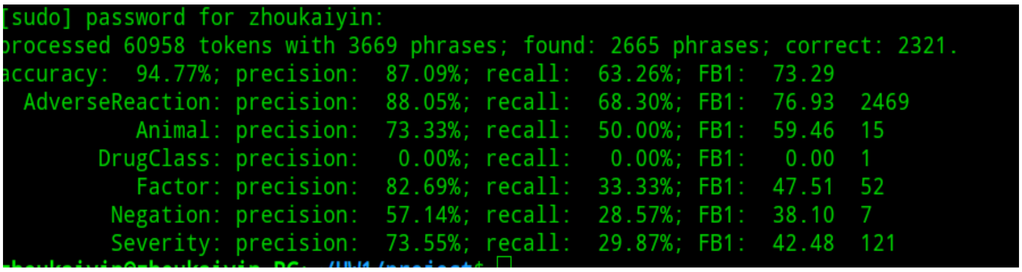
四、 模型训练

将训练数据随机按找 7:3 划分，在训练模型过程中，7 份用作实际训练模型，3 份用作开发集调整参数。当模型最优后用该参数训练所有 10 份数据获得模型，并对测试数据进行预测。

4.1 调整参数优化模型

```
sudo bash dev-wapiti.sh

or bash dev-wapitiv2.sh
```

结果打印：

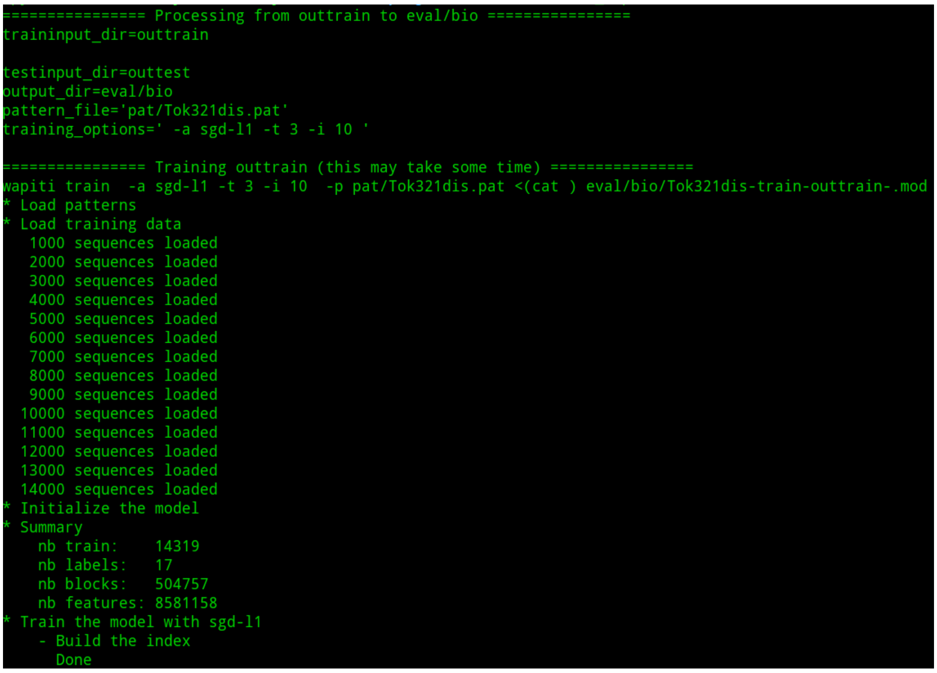


```
[sudo] password for zhoukaiyin:
processed 60958 tokens with 3669 phrases; found: 2665 phrases; correct: 2321.
accuracy:  94.77%; precision:  87.09%; recall:  63.26%; FB1:  73.29
  AdverseReaction: precision:  88.05%; recall:  68.30%; FB1:  76.93  2469
           Animal: precision:  73.33%; recall:  50.00%; FB1:  59.46  15
        DrugClass: precision:   0.00%; recall:   0.00%; FB1:   0.00  1
           Factor: precision:  82.69%; recall:  33.33%; FB1:  47.51  52
         Negation: precision:  57.14%; recall:  28.57%; FB1:  38.10  7
         Severity: precision:  73.55%; recall:  29.87%; FB1:  42.48  121
```

模型被存储在 eval/bio 中。

4.2 预测测试数据

```
sudo bash test_wapiti.sh
```



```
================ Processing from outtrain to eval/bio ================
traininput_dir=outtrain

testinput_dir=outtest
output_dir=eval/bio
pattern_file='pat/Tok321dis.pat'
training_options=' -a sgd-l1 -t 3 -i 10 '

================ Training outtrain (this may take some time) ================
wapiti train  -a sgd-l1 -t 3 -i 10  -p pat/Tok321dis.pat <(cat ) eval/bio/Tok321dis-train-outtrain-.mod
* Load patterns
* Load training data
   1000 sequences loaded
   2000 sequences loaded
   3000 sequences loaded
   4000 sequences loaded
   5000 sequences loaded
   6000 sequences loaded
   7000 sequences loaded
   8000 sequences loaded
   9000 sequences loaded
  10000 sequences loaded
  11000 sequences loaded
  12000 sequences loaded
  13000 sequences loaded
  14000 sequences loaded
* Initialize the model
* Summary
    nb train:    14319
    nb labels:   17
    nb blocks:   504757
    nb features: 8581158
* Train the model with sgd-l1
    - Build the index
      Done
```

出现 Finished!说明序列标注完成。序列标注结果存储在 eval/bio/Tok321dis-train-test-outtrain.tab 中。