词性标注及其算法比较

项目报告

学 院： 信息工程学院

指导教师： 孙 媛

班 级： 19大数据班

学生姓名： 宋特 刘溪月

蒋子文

学 号：19011464 19011404

19011465

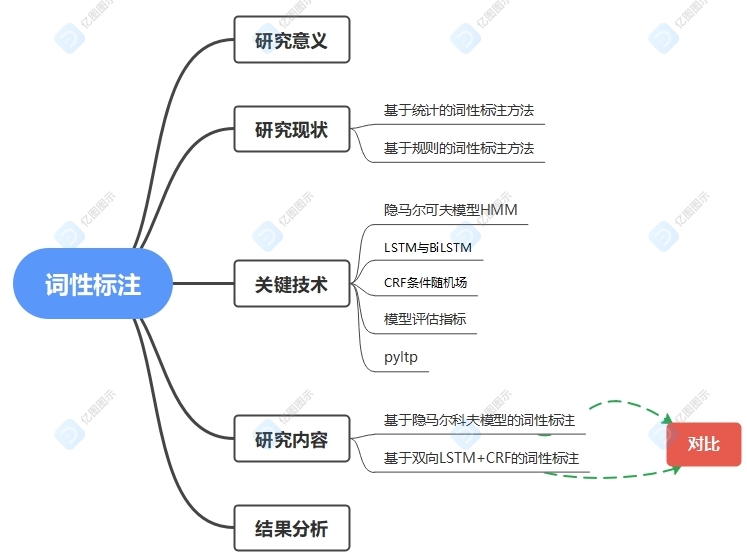
日期：2021 年 12 月 26 日

摘 要

**词性**指以词的特点做为划分同类的根据，其可分为实词，虚词两大类。实词包括：名词，代词，动词，形容词，数词，量词；虚词包括：副词，介词，连词，冠词，助词，叹词，拟声词。

**词性标注（Part-Of-Speech tagging, POS tagging）**也被称为语法标注（grammatical tagging）或词类消疑（word-category disambiguation），是语料库语言学（corpus linguistics）中将语料库内单词的词性按其含义和上下文内容进行标记的文本数据处理技术，其在本质上是分类问题，将语料库中的单词按词性分类。一个词的词性由其在所属语言的含义、形态和语法功能决定。本作业将通过两种典型方法实现词性标注，并对其标注结果进行分析与比较。

# 一.研究框架：



# 二.研究意义

词性标注是文本数据的预处理环节之一，原始文本在NLP或文本挖掘应用中，首先通过字符分割（word segmentation）和字符嵌入（word embedding）被向量化，随后通过词性标注得到高阶层特征，并输入语法分析器执行语义分析（sentiment analysis）、指代消解（coreference resolution）等任务

# 三.研究现状

就整个词性标注的研究而言，中文词性标注方法主要可分为类：基于统计的方法和基于规则的方法。

## 基于统计的词性标注方法：

基于统计的自然语言处理方法在消除歧义和句法分析等方面得到越来越广泛的应用,是近年来兴起的一种新的也是最常使用的方法。 对于给定的输入词串,该方法先确定其所有可能的词性串,选出得分最高的作为最佳输出。 其中应用比较广泛的主要有隐马尔可夫模型(HMM)方法和条件随机场(CRF)的方法。

HMM 模型是一种特别适合处理随机序列数据的统计模型。 它需要大量训练语料来达到较高的标注准确率,还存在标记偏置问题。CRFs 模型是在给定输入节点条件下计算输出节点条件概率的无向图模型,是序列标注和切分的统计模型。 CRFs 虽然能克服最大熵和 HMM 等有向图模型的标记偏置[21] 问题,但它倾向选择概率较高的词进行标记,这可能影响标注的准确率。 因此在实际应用中研究人员多对其进行改进以此来提高准确率。 2010 年,Moon T 等提出了一种利用边界条件的 HMM 模型,利用文本内容和功能词之间的不同对局部文本快速进行词性标注,提高了无监督词标注的准确率。 2011 年,孙静等提出了一种基于 CRF 模型的无监督词性标注方法,先利用词典对已分词文本进行词性标注,再利用 CRF 对语料进行迭代标注,逐步优化标注结果。 2012 年,袁里驰提出了一种改进的 HMM 词性标注方法,把马尔可夫

族模型和句法分析结合进行词性标注,在相同测试条件下,马尔可夫族模型性能明显优于 HMM 模型。

基于统计的词性标注方法是目前应用较多的方法,通过对大规模语料库进行训练得到,覆盖面很广,标注结果有很好的一致性和较高的覆盖率,因此被广泛用于自然语言处理领域。 但当训练语料达到一定规模后,通过扩充语料规模来提高正确率也变的不实际。需要设计更为严格精细的特征统计模板,以便参考更多的特征信息来进行词性标注。 另外在进行标注的过程中要处理好概率参数的获取以及如何应用所获得的概率参数对文本进行词性标注等问题。

## 基于规则的词性标注方法：

其是一种传统的方法,获取的规则集精度直接影响标注结果的优劣。 该方法能充分利用现有语言学成果,总结出许多有用的规则。 先利用词典对语料进行基本切分和标注,列出该对象所有可能的词性,然后依据上下文信息,结合规则

库消除歧义,最终保留唯一合适词性。

基于规则的方法表达清晰,应用范围较广,但不能方便地通过机器学习来自动获取规则,人工构造又是一项艰难耗时的任务,如果把规则描述过细,规则的覆盖面就会大大减小,很难根据实际情况进行调整。 如果不根据上下文仅根据规则判断词性又可能会出现歧义。 它不属于统计模型,所以适应性也较差。 而统计方法正好能弥补这个缺点,所以在实际应用中常把两者结合起来或对基于规则的方法进行改进。 这样既能充分利用现有语言学的成果,还能利用统计模型来增强方法的适应性。 标注的准确率也比单一使用一种要高。 1995 年,Eric Brill提出了基于转换的错误驱动的方法。 利用初始标注器来标注训练语料库,得到的结果与正确结果进行比较,从中选出效果最好的变换模式作为系统的标注规则重新标注语料库,重复该过程直到获得所有规则,再用这些规则对待标注语料进行标注。 2008 年,王广正等提出基于规则优先级的词性标注方法,对每条词性标注规则加上优先级,通过控制优先级来完成兼类词的词性标注。 2010 年,姜尚仆等提出了一种基于规则和统计的分词和词性标注方法,使用基于单一感知器的联合分词和词性标注算法作为基本框架,以基于规则的词语邻接属性为特征。 虽然该方法是针对日语进行词性标注的,但该方法的思想和技术同样值得中文词性标注研究人员借鉴。 2010 年,陈小芳等提出了一种基于统计和规则相结合的汉语术语语义分析方法。 首先以词、词性、距离信息、上下文信息、词语的第一个原义信息为特征,基于这些特征得到支持向量机分析模型,在此基础上利用统计和规则相结合的方法进行术语语义分析。

# 四.关键技术

## （一）隐马尔科夫模型

隐马尔可夫模型（Hidden Markov Model，HMM）是统计模型，它用来描述一个含有隐含未知参数的马尔可夫过程。最初是在20世纪60年代后半期Leonard E. Baum和其它一些作者在一系列的统计学论文中描述的。HMM最初的应用之一是开始于20世纪70年代中期的语音识别。其难点是从可观察的参数中确定该过程的隐含参数。然后利用这些参数来作进一步的分析。其算法可以描述如下:

如果一个系统有n个有限状态*S*={*s*1,*s*2,…*sn*},随着时间推移，该系统将从某一状态转移到另一状态，*Q*={*q*1,*q*2,…*qn*}位一个随机变量序列，该序列中的变量取值为状态集S中的某个状态，其中*qt*表示系统在时间t的状态。那么：系统在时间t处于状态*sj*的概率取决于其在时间1,2, …

 t-1的状态，该概率为：

*P*(*qt*=*sj*|*qt*−1=*si*,*qt*−2=*sk*…)

如果在特定条件下，系统在时间t的状态只与其在时间t-1的状态相关，即：

*P*(*qt*=*sj*|*qt*−1=*si*,*qt*−2=*sk*…)=*P*(*qt*=*sj*|*qt*−1=*si*)

则该系统构成一个离散的一阶马尔可夫链。  
进一步，如果只考虑上述公式独立于时间t的随机过程：

*P*(*qt*=*sj*|*qt*−1=*si*)=*aij*,1≤*i*,*j*≤*N*

该随机过程为马尔可夫模型。其中，状态转移概率aij 必须满足以下条件：

*aij*≥0,∑*j*=1*Naij*=1

而相对于马尔可夫模型，在隐马尔可夫模型中，我们不知道模型经过的状态序列，只知道状态的概率函数，即，观察到的事件是状态的随机函数，因此，该模型是一个双重的随机过程。其中，模型的状态转换过程是不可观察的，即隐蔽的，可观察事件的随机过程是隐蔽的观察状态转换过程的随机函数。隐马尔可夫模型可以用五个元素来描述，包括2个状态集合和三个概率矩阵：

（1） 隐含状态 S

这些状态之间满足马尔可夫性质，是马尔可夫模型中实际所隐含的状态。这些状态通常无法通过直接观测而得到。（例如S1,S2,S3等等)

（2）可观测状态 O

　　在模型中与隐含状态相关联，可通过直接观测而得到。(例如O1,O2,O3

等等，可观测状态的数目不一定要和隐含状态的数目一致。

（3）初始状态概率矩阵 π

表示隐含状态在初始时刻t=1的概率矩阵，(例如t=1时，P(S1)=p1,P(S2)=P2,P(S3)=p3，则初始状态概率矩阵 π=[ p1 p2 p3 ]。

（4）隐含状态转移概率矩阵A

　 描述了HMM模型中各个状态之间的转移概率。其中Aij=P(Sj|Si),1≤i,,j≤N

， 表示在 t 时刻、状态为 Si 的条件下，在 t+1 时刻状态是 Sj的概率

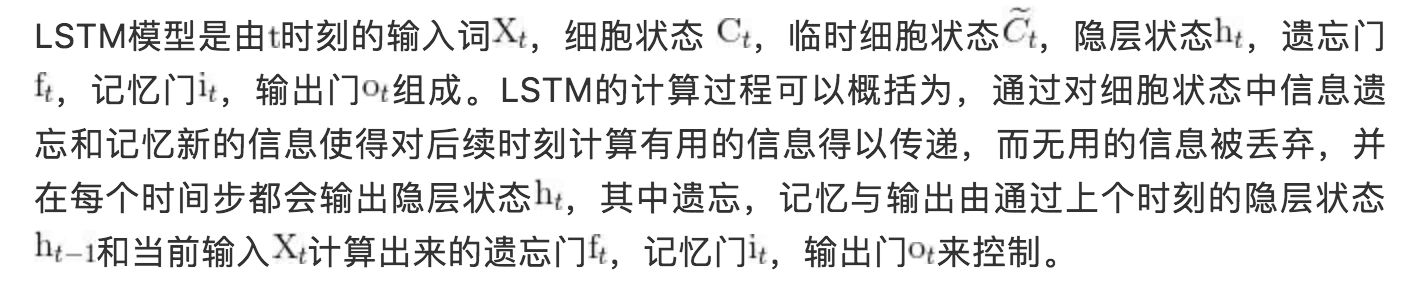
（5） 观测状态转移概率矩阵B

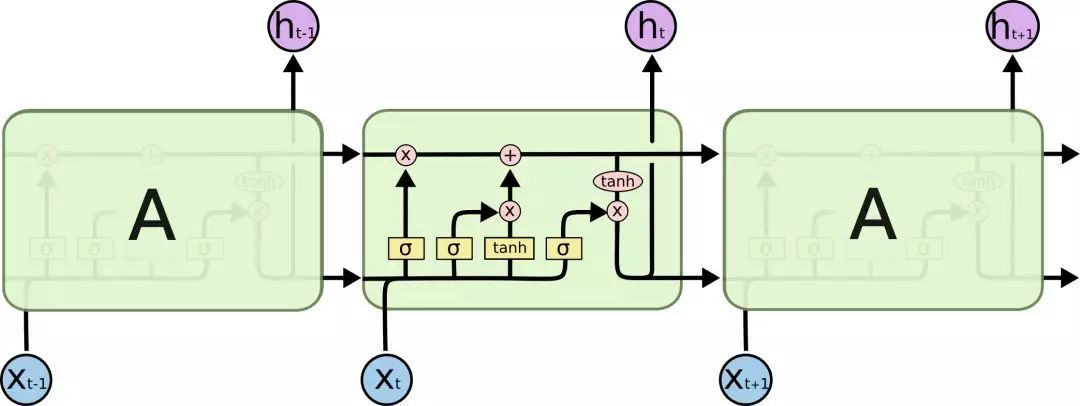
　 令N代表隐含状态数目，M代表可观测状态数目，则 Bij=P(Oi|Sj),1≤i≤M,1≤j≤N表示在 t 时刻、隐含状态是 sj条件下，观察状态为 Oi 的概率。

## （二）LSTM与BiLSTM

长短期记忆网络（LSTM，Long Short-Term Memory）是一种时间循环神经网络，是为了解决一般的RNN（循环神经网络）存在的长期依赖问题而专门设计出来的。使用LSTM模型可以更好的捕捉到较长距离的依赖关系。因为LSTM通过训练过程可以学到记忆哪些信息和遗忘哪些信息。

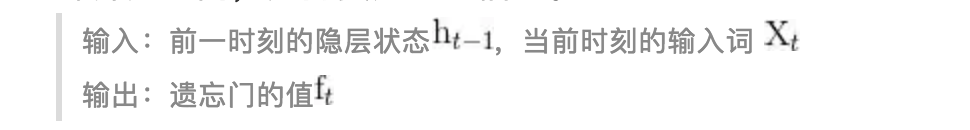
LSTM整体框架：

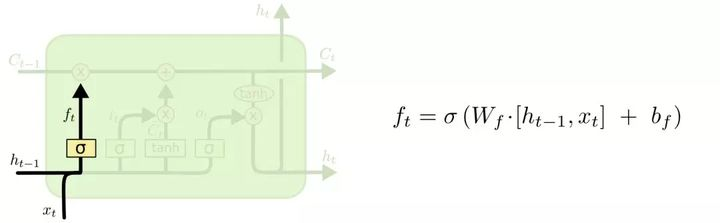




LSTM详细计算过程：

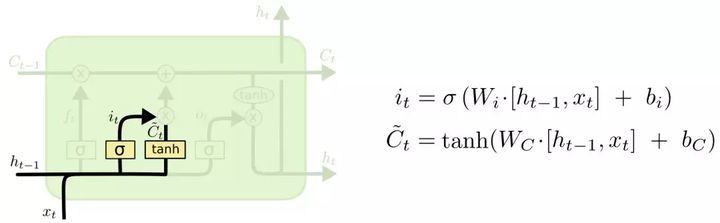
计算遗忘门，选择要遗忘的信息。





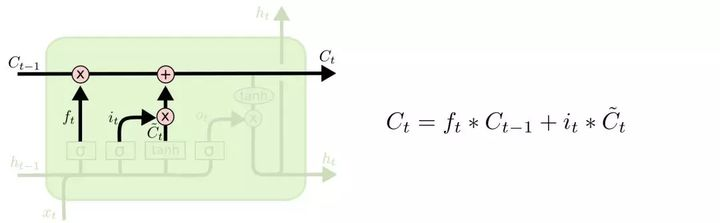
计算记忆门，选择要记忆的信息。





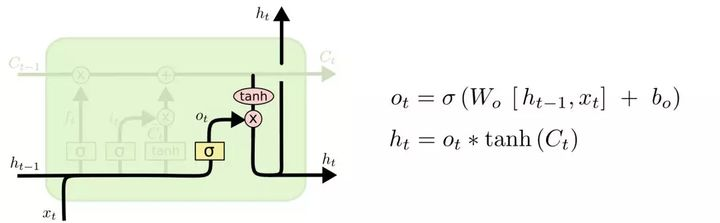
计算当前时刻细胞状态





计算输出门和当前时刻隐层状态

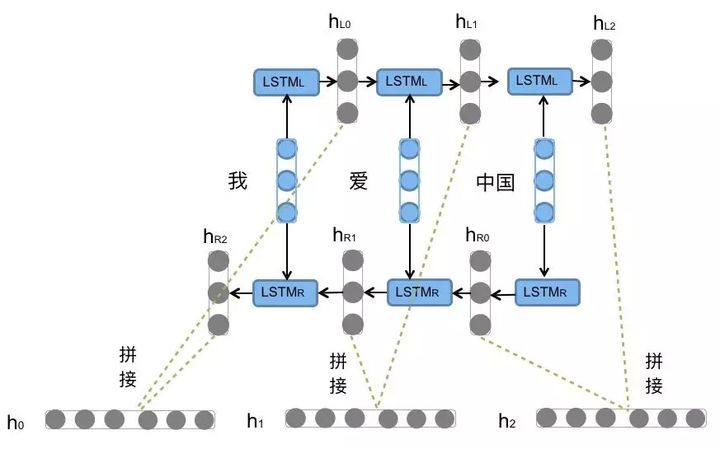




最终，我们可以得到与句子长度相同的隐层状态序列{ }

但是利用LSTM对句子进行建模还存在一个问题：无法编码从后到前的信息。BiLSTM是Bi-directional Long Short-Term Memory的缩写，是由前向LSTM与后向LSTM组合而成。通过BiLSTM可以更好的捕捉双向的语义依赖。

前向的LSTM与后向的LSTM结合成BiLSTM。比如，我们对“我爱中国”这句话进行编码，模型如图所示。



无论是LSTM还是BiLSTM两者在自然语言处理任务中都常被用来建模上下文信息。

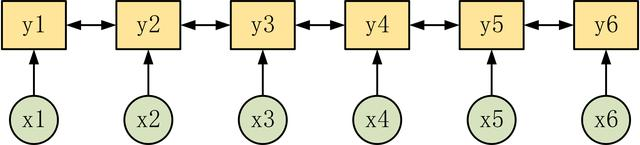
## （三）CRF（条件随机场）

随机场：随机场是一种图模型，包含结点的集合和边的集合，结点表示一个随机变量，而边表示随机变量之间的依赖关系。如果按照某一种分布随机给图中每一个结点赋予一个值，则称为随机场。

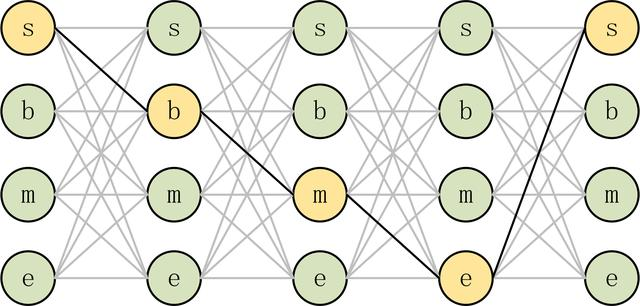
马尔科夫随机场：马尔科夫性质指某一个时刻 t 的输出值只和 t-1 时刻的输出有关系，和更早的输出没有关系。马尔科夫随机场则是一种特殊的随机场，其假设每一个结点的取值只和相邻的结点有关系，和不相邻结点无关。

而条件随机场 CRF：CRF 是一种特殊的马尔科夫随机场，CRF 假设模型中只有 X (观测值) 和 Y (状态值)。在 CRF 中每一个状态值 yi 只和其相邻的状态值有关，而观测值 x 不具有马尔科夫性质。注意观测序列 X 是作为一个整体影响 Y 计算。

用 LSTM 进行序列标注时，假设我们的输入序列为 n，则 LSTM 会进行 n 次预测，分别预测 n 个时刻的输出 yi。但是 LSTM 不能考虑输出的每一个元素的相关性，例如在分词时，s 表示单个字的词，则 s 之后是不能接 m 或者 e 的，而 LSTM 逐元素分类是不能考虑这种相关性的。但是 CRF 可以考虑输出元素的前后关联性 (CRF 可以通过特征函数学习状态间的关联)，如下图所示。



CRF 会通过特征函数集计算出一个输出序列的分数，并用所有可能序列的分数之和进行归一化，如下图所示，sbmes 就是其中一个序列路径，CRF 会找出概率最大的路径作为预测序列。



## （四）PRF模型评估指标

### （1）精度（P）

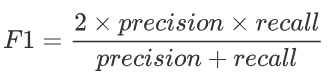
精度（precision） 精度是精确性的度量，表示被分为正例的示例中实际为正例的比例，precision=TP/(TP+FP)，在本实验中即被算法正确标注的词数/算法标注的所有词数。

### （2）召回率（R）

召回率（recall） 召回率是覆盖面的度量，度量有多个正例被分为正例，recall=TP/(TP+FN)=TP/P=sensitive，在本实验中即被算法正确标注的词数/pyltp库标注的所有词数。

### （3）F1测试值

F1值又叫做综合分类率



用一个F1值来综合评估精确率和召回率，它是精确率和召回率的调和均值，当精确率和召回率都高时，F1值也会高，其为了综合多个类别的分类情况，用于评测系统整体性能。

***因为本实验中，每个词性类别都有一个P、R、F1值，所以取Macro值，即综合取各类别分类结果的平均值MacroP、MacroR、MacroF1.***

## （五）pyltp库

LTP是哈工大开发的中文NLP开源工具，LTP提供了一系列中文自然语言处理工具，用户可以使用这些工具对于中文文本进行分词、词性标注、命名实体识别等等工作。本实验将用其词性标注结果作为语料标注建立HMM模型：

安装方法：

（1）pip install pyltp(大概率报错)

（2）copy以下文件到代码目录：



pip install pyltp-0.2.1-cp36-cp36m-win\_amd64.whl(需要wheel包)

下载模型：

LTP库需要搭配模型来使用，在 [LTP 的模型页面](http://ltp.ai/download.html) 下载模型，这里要注意 pyltp 库版本需要和模型的版本对应。



# 五.主要研究内容

## （一）基于隐马尔科夫模型（HMM）的词性标注

隐马尔可夫模型和序列标注问题是天然适配的，所以自然而然的，早期很多做词性标注，命名实体识别等的算法，都采用了这个模型，是属于非常典型的词性标注方法之一。

## （二）基于双向LSTM+CRF的词性标注

在如今常用的词性标注算法当中，BiLSTM+CRF可谓新起之秀，是目前最主流最有效的方法之一。其中CRF部分相当于通过特征函数添加了一个转移矩阵，而其他的特征提取则是交由神经网络来完成。

# 六.过程，遇到的问题与部分代码分析

## （一）HMM：

导入文本，预处理后得到训练集

分词，plt库进行词性标注得到训练集标注

构建（训练）HMM模型

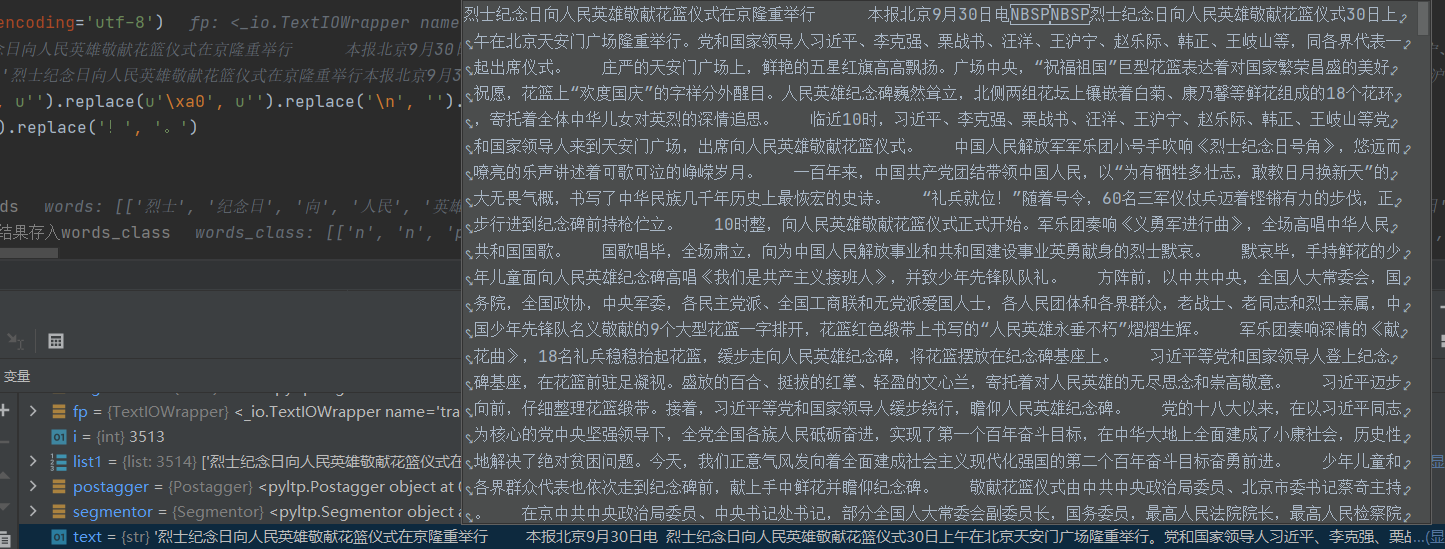
导入测试集文本并预处理，通过viterbi算法回溯得到标注结果

写入文本，并评估模型

### 1.导入文本与数据预处理：

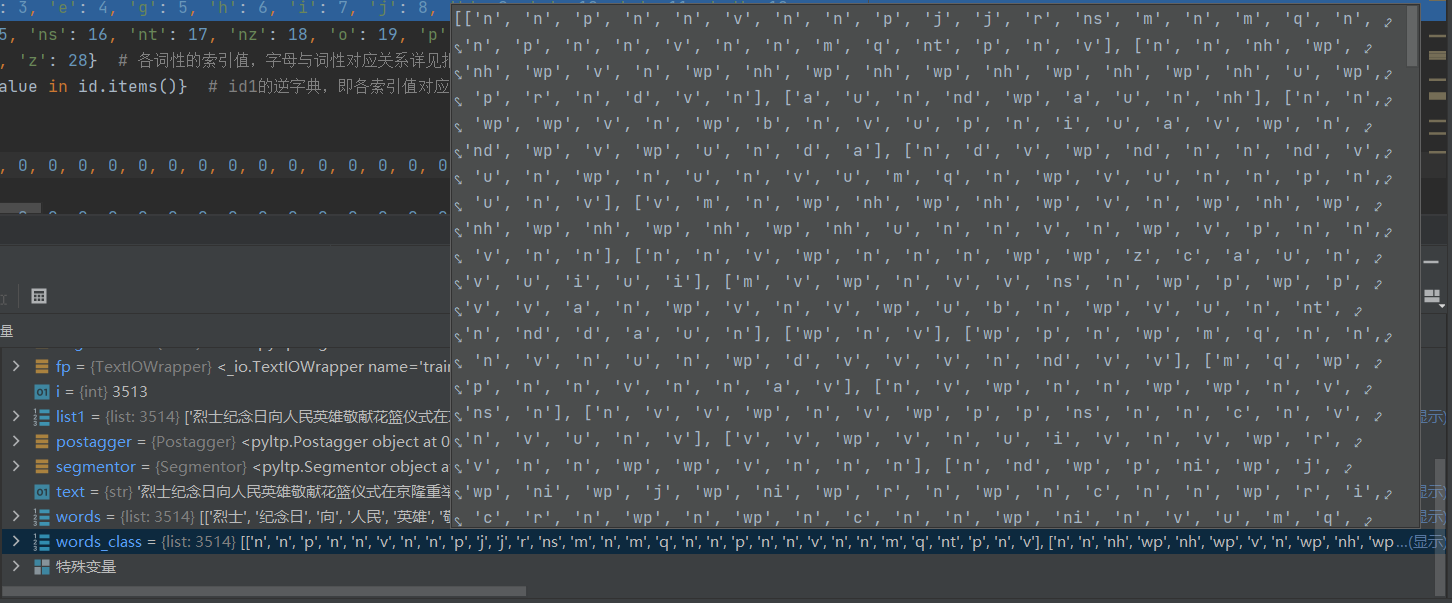
本次大作业我们选择了一周的人民日报文本（2021.10.01~2021.10.07）做为HMM的训练语料，节选其后一日文本（约1000字）作为测试语料。

首先，导入文本并将数据预处理得到训练集。



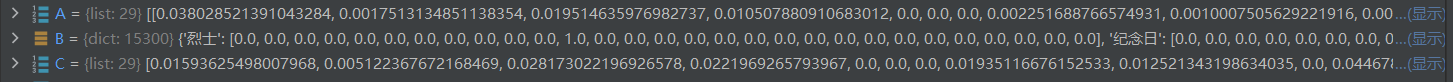
### 2. 分词，plt库进行词性标注得到训练集标注

将训练集通过jieba.lcut分词，然后通过pyltp的postagger函数得到语料标注。



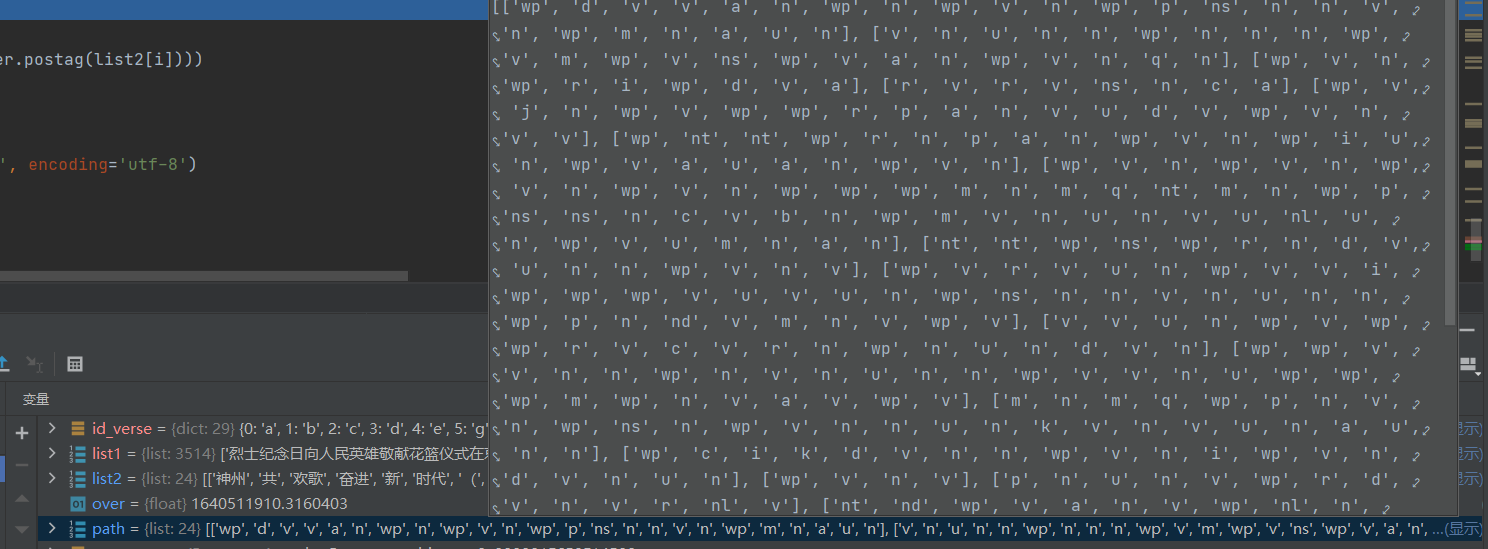
### 3. 构建（训练）HMM模型

HMM模型的训练实际上就是各概率矩阵的计算与获取，这里仅展示矩阵的部分数据（其中A为概率矩阵，B为发射矩阵，C为初始矩阵），详细计算过程请见附录代码。



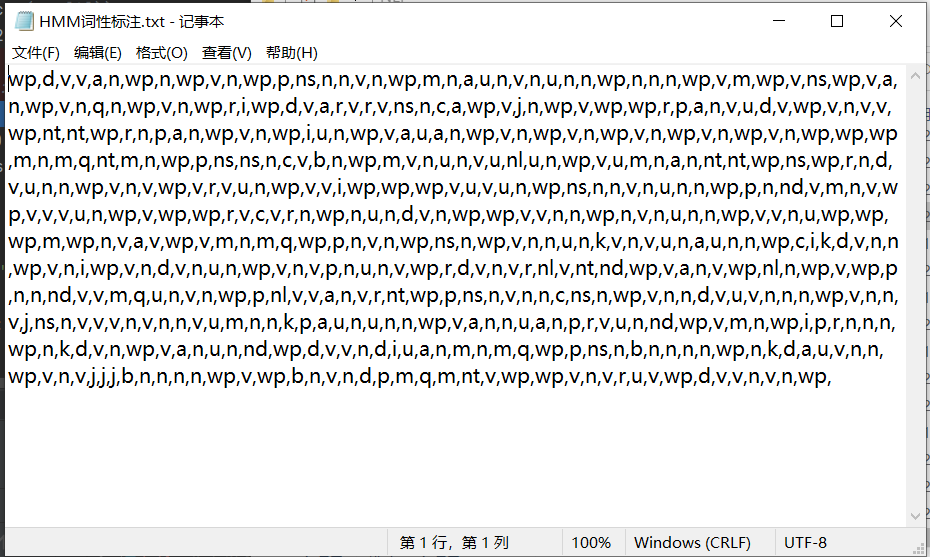
### 4. 导入测试集文本并预处理，通过viterbi得到测试序列

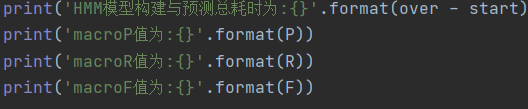
将测试集导入并进行文本预处理后存入Liste2，之后通过viterbi算法并调用递归函数findpath得到测试集的标注序列。



### 5.写入文本，评估模型：

写入文本：



调用模型评估指标函数，得到模型评估结果，函数算法详见附录代码：  


## （二）BiLSTM+CRF:

导入文本，预处理后得到训练集

分词，plt库进行词性标注得到训练集标注

构建并训练LSTM模型，并计算CRF

导入测试集文本并预处理，通过模型得到预测标注结果

写入文本，并评估模型

### 1.导入文本与数据预处理：

开始时为了产生一定程度的对照效果，我们选择了与HMM相同的测试语料。

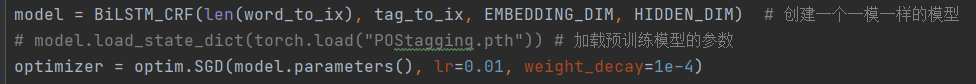
这一步操作同HMM是一样的，这里不做赘述。

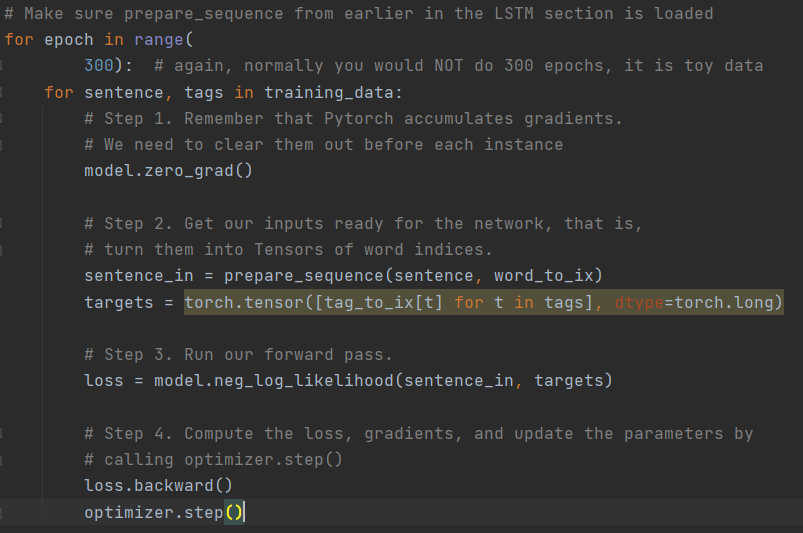
### 2.分词，plt库进行词性标注得到训练集标注

同HMM。

### 3.构建并训练LSTM与CRF模型：

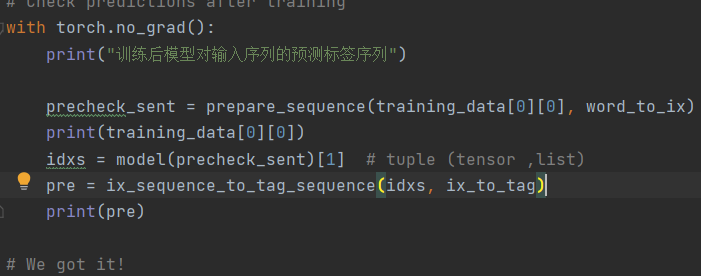
调用类并构建一个新的LSTM模型，其类定义与函数请详见附录代码：



但在训练模型时，出现了因为样本规模过大导致维度爆炸而电脑算力不足无法计算而报错的情况，于是只能调整样本规模到1000字（节选人民日报）。

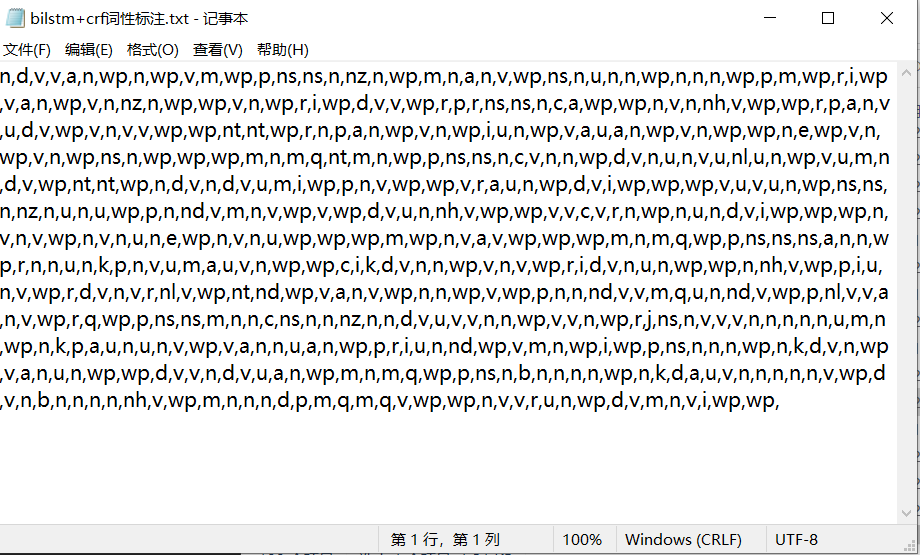
### 4.导入测试集文本并预处理，通过模型得到预测标注结果

通过训练好的BiLSTM+CRF模型，预测并得到标注序列。

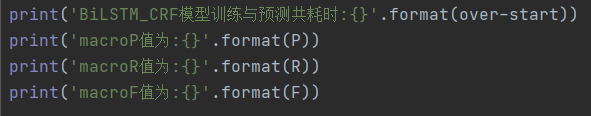


### 5.写入文本，并评估模型

写入文本：

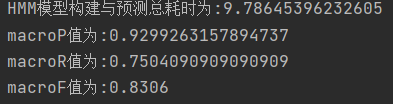


调用模型评估指标函数，得到模型评估结果，函数算法详见附录代码：



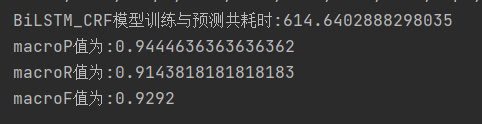
# 七.结果分析：

## （一）HMM：



模型整体训练速度与运算速度均比较快，模型效果也不错，其结果P值相对较高。

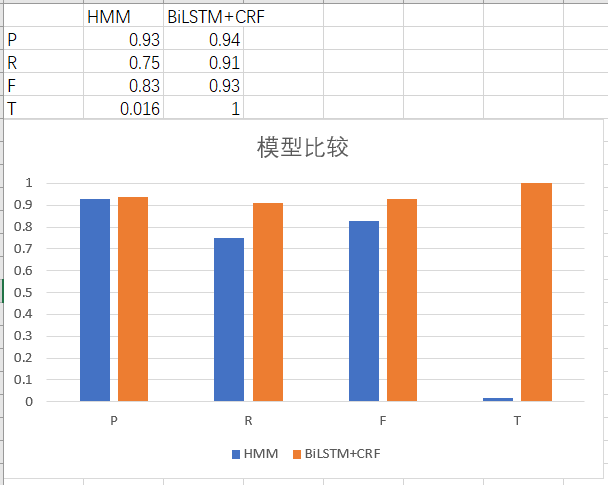
## （二）BiLSTM+CRF



为了让模型基本收敛，我们进行了300次迭代，耗时约10分钟14秒，且占用内存很大。可见LSTM模型训练速度慢，运算量大的弊端。但其模型效果特别好，精度，RECALL和F1值都在0.95左右。

## （三）模型比较：

在可视化模型比较结果时，为了较为直观的比较时间，我们将BiLSTM+CRF收敛所花费的时间（T）作为参数1，将时间标准化。



通过比较不难看出，传统的基于统计的HMM在词性标注上模型的整体效果要比BiLSTM+CRF略逊一色，但其准确率却可以基本达到与B+C相同的水平，且其模型搭建时间要远远快于B+C。

由此我们不难得到结论，在文本规模较小时，LSTM更高的精度与更好的模型效果，并且计算量与训练时间也将在我们的可接受范围内；但当我们面临大规模的文本词性标注问题时，选择HMM算法可以更快的得到标注结果，并且得到一个效果较为不错的模型。