**第七周**

**分词研究工作汇报**

**核心：提高法律文书的分词效果**

**怎么提高=》通过完善词典**

**大概流程&论文写作思路：**

**1.文书数据预处理**

文书格式统一

统一文书编码

统一UTF\_8编码

停用词过滤

对法律文书的语言进行词性标注

对法律文书的主要内容字段进行提取

**2.训练数据，基于裁判文书网，人工标记，提取出的关键词**

**3.jieba涉及的方法**

前缀词典方法，高效进行词图扫描

动态规划，查找最大概率路径，找出词频的最大切分组合

对于未登录词，采用汉字成词能力的HMM模型，结合维特比算法

基于上述方法，进行文本切分

1. 全模式，把句子中所有在词库中出现的词都找出来

2. 不使用隐马尔科夫模型的精确模式，基于最大概率路径, 找出基于词频的最大切分组合

3. 同时使用最大概率路径和隐马尔科夫模型，对于未登录词也有比较好的切分效果

会根据字典生成句子中汉字构成的有向无环图（DAG）

以“但也并不是那么出乎意料或难以置信”这句话作为输入，生成的DAG如下，简单的讲就是把句子中词的位置标记出来

0 [0] 但

1 [1] 也

2 [2] 并

3 [3, 4] 不是

4 [4] 是

5 [5, 6] 那么

6 [6] 么

7 [7, 8, 10] 出乎意料

8 [8] 乎

9 [9, 10] 意料

10 [10] 料

11 [11] 或

12 [12, 13, 15] 难以置信

13 [13] 以

14 [14, 15] 置信

15 [15] 信

一、全模式 现在看一下全模式的代码：

def \_\_cut\_all(sentence): dag = get\_DAG(sentence) old\_j = -1 for k, L in iteritems(dag): if len(L) == 1 and k > old\_j: yield sentence[k:L[0] + 1] old\_j = L[0] else: for j in L: if j > k: yield sentence[k:j + 1] old\_j = j

就是把上面生成的DAG中的所有的组合显示出来，还是以上面的句子为例，全模式切分的结果如下，是不是觉得这么非常的easy。

但/也/并/不是/那么/出乎/出乎意料/意料/或/难以/难以置信/置信

二、 不使用隐马尔科夫模型的精确模式，用一个比较简单的句子为例，输入的句子是“难以置信”，按照全模式会输出：难以/置信/难以置信 三个组合。作为一个将汉语的人，我们明显知道最佳的分词结果就是“难以置信”这一种结果。下面用最大概率的方法解释为什么是这个结果。子啊字典中我们可以查询“难以置信”所有的组合以及它们出现的概率（除以所有词出现的总次数）如下：

词次数出现概率

难185053.08\*10-4

以1361362.27\*10-3

置111451.85\*10-4

信111881.86\*10-4

难以56819.45\*10-5

置信641.06\*10-6

难以置信1692.81\*10-6

根据上面各词的概率，可以算出“难以置信”所有分词方式的概率，而最大出现的可能就是“难以置信”,而且它的概率相比其他组合高的可不是一倍两倍。

分词方式出现概率

难 以 置 信1.37\*10-14

难以 置 信3.65\*10-14

难 以 置信7.41\*10-13

难以 置信1.01\*10-10

难以置信2.81\*10-6

三、 同时使用最大概率路径和隐马尔科夫模型

由于这一部分设计到的理论知识比较多，建议先看一下相关的论文和博客。然后再通过代码验证对理论算法的知识，可以更深刻的理解隐马尔科夫模型。在这个地方中文句子作为可观测序列，对应的隐藏状态值集合为(B, M, E, S): {B:begin, M:middle, E:end, S:single}。分别代表每个状态代表的是该字在词语中的位置，B代表该字是词语中的起始字，M代表是词语中的中间字，E代表是词语中的结束字，S则代表是单字成词。

在HMM模型中文分词中，我们的输入是一个句子(也就是观察值序列)，输出是这个句子中每个字的状态值。比如:“小明硕士毕业于中国科学院计算所”，隐藏序列为

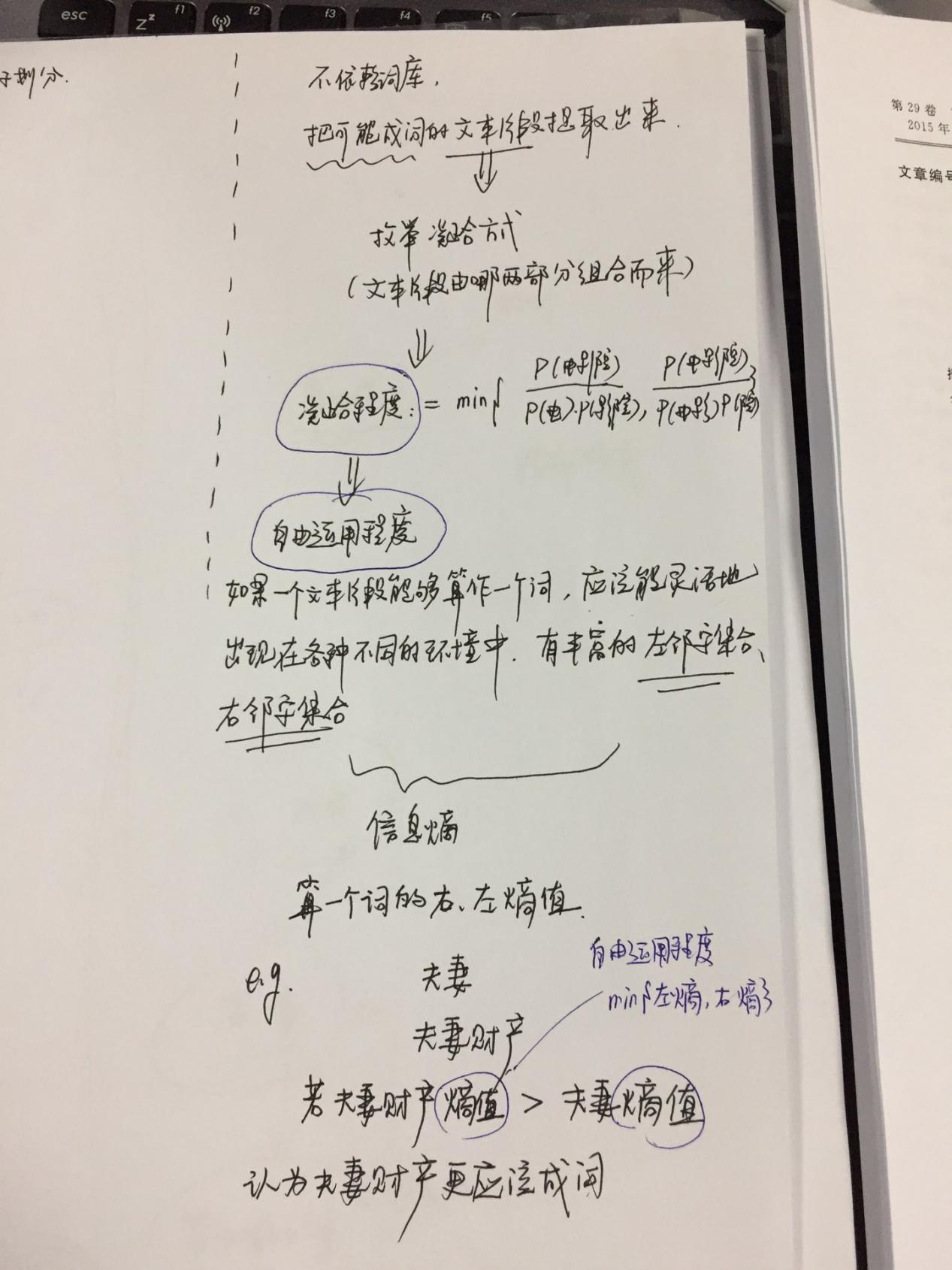
BEBEBMEBEBMEBES，根据这个状态序列我们可以进行切词:BE/BE/BME/BE/BME/BE/S，所以切词结果如下:小明/硕士/毕业于/中国/科学院/计算/所。

finalseg中有BMES各个状态间的转移概率以及隐藏状态对应于各个中文的发射概率，再根据维特比算法，计算分词就相当的容易了。由于对隐马尔科夫理解有限，也就不瞎写了，有兴趣的可以看一下http://blog.csdn.net/likelet/article/details/7056068，写的非常的赞。

基于3这个的实验结果

准确率，召回率，F值

**4.基于统计方法，基于信息熵，计算左右熵**



提取出关键词，添加入词典中

单纯基于4的实验结果

准确率，召回率，F值

3，4结合的实验结果对比

准确率，召回率，F值

**5.对法律文书制定规则标准**

再进行规则过滤

单纯用5的实验结果

准确率，召回率，F值

3，5结合

准确率，召回率，F值

4，5结合

准确率，召回率，F值

3，4，5结合

准确率，召回率，F值

**具体规则：**

**1）.以关键名词作为提取关键：**

思路：法律文书中，n.的重要性占首位。提取这样的词：文档中可能无意义，和其他词结合，有很大意义

根据词性统计出名词信息熵（或者其频率）见电子表格

最基本的基于词性标注结果，提取出名词，查看相关名词的信息熵，

adj.+n.

v.+n

一对多（1:n）找noun.+\*的组合

找出单个的，依着单个的，找出右熵

多对一 （n:1）找\*+noun.的组合

找出单个的，依着单个的，找寻左熵

**2）.通过正则可以解决的**

自动识别，添加入词典

Eg：

第+num+条

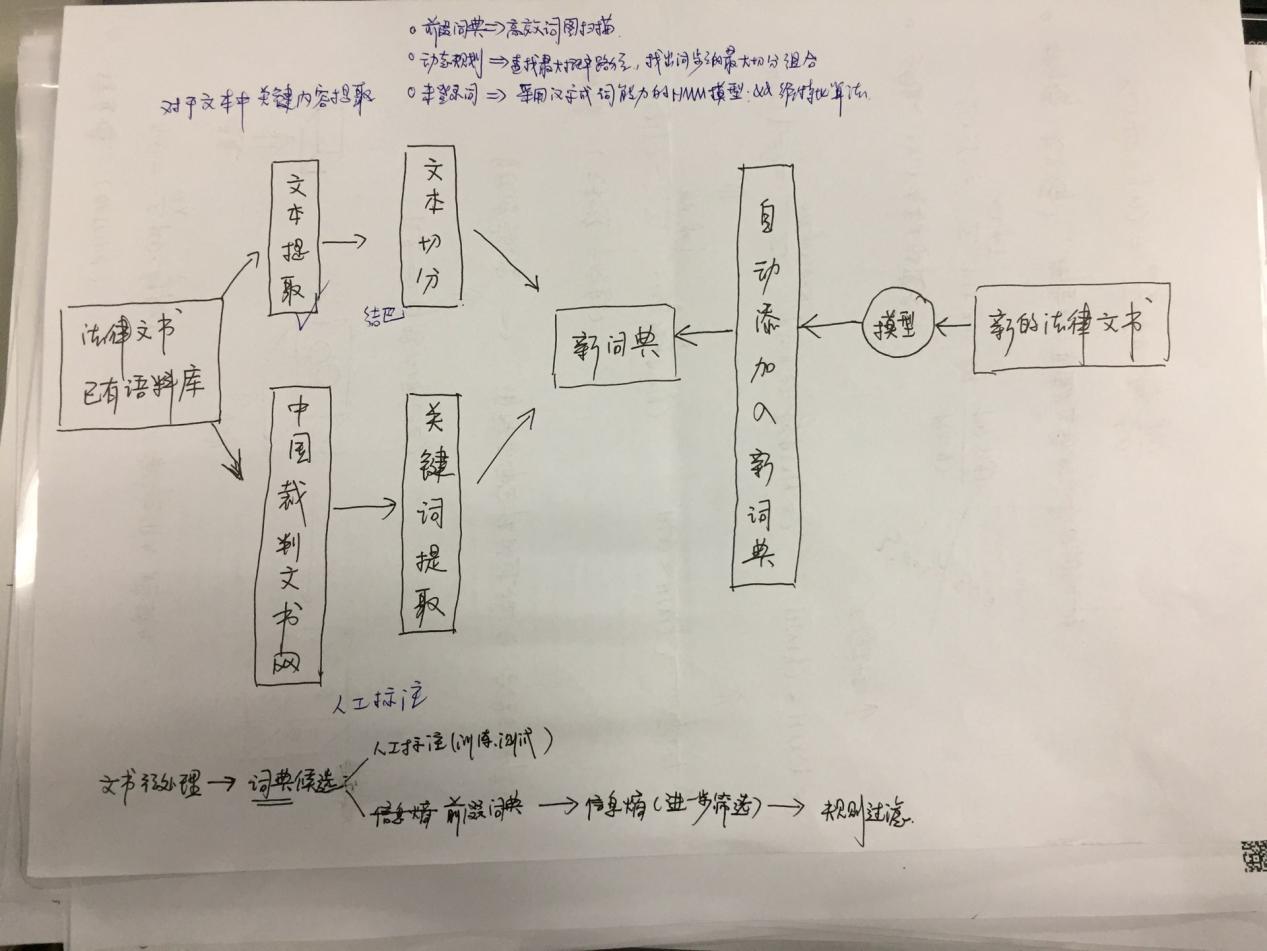
否定词

数量词

。。。

**6.条件随机场**

以目前抽取出的词，作为训练数据，利用条件随机场，对训练语料进行建模，对于新的文书，可以完成自动词典添加的功能。



=========================================================

准确率，召回率，F值

[SIGHAN](http://sighan.cs.uchicago.edu/)是国际计算语言学会（ACL）中文语言处理小组的简称，其英文全称为“Special Interest Group for Chinese Language Processing of the Association for Computational Linguistics”，又可以理解为“SIG汉“或“SIG漢“。

1) 介绍（Introduction）：

　　本目录包含了训练集、测试集及测试集的（黄金）标准切分，同时也包括了一个用于评分的脚本和一个可以作为基线测试的简单中文分词器。(This directory contains the training, test, and gold-standard data used in the 2nd International Chinese Word Segmentation Bakeoff. Also included is the script used to score the results submitted by the bakeoff participants and the simple segmenter used to generate the baseline and topline data.)

2) 文件列表（File List）

　　在gold目录里包含了测试集标准切分及从训练集中抽取的词表（Contains the gold standard segmentation of the test data along with the training data word lists.）

　　在scripts目录里包含了评分脚本和简单中文分词器（Contains the scoring script and simple segmenter.）

　　在testing目录里包含了未切分的测试数据（Contains the unsegmented test data.）

　　在training目录里包含了已经切分好的标准训练数据（Contains the segmented training data.）

　　在doc目录里包括了bakeoff的一些指南（Contains the instructions used in the bakeoff.）

3) 编码（Encoding Issues）

　　文件包括扩展名”.utf8”则其编码为UTF-8(Files with the extension ".utf8" are encoded in UTF-8 Unicode.)

　　文件包括扩展名”.txt”则其编码分别为（Files with the extension ".txt" are encoded as follows）:

　　前缀为as\_，代表的是台湾中央研究院提供，编码为Big Five (CP950)；

　　前缀为hk\_，代表的是香港城市大学提供，编码为Big Five/HKSCS；

　　前缀为msr\_，代表的是微软亚洲研究院提供，编码为 EUC-CN (CP936)；

　　前缀为pku\_，代表的北京大学提供，编码为EUC-CN (CP936)；

　　EUC-CN即是GB2312（EUC-CN is often called "GB" or "GB2312" encoding, though technically GB2312 is a character set, not a character encoding.）

4） 评分（Scoring）

　　评分脚本“score”是用来比较两个分词文件的，需要三个参数（The script 'score' is used to generate compare two segmentations. The script takes three arguments)：

　　1. 训练集词表（The training set word list）

　　2. “黄金”标准分词文件（The gold standard segmentation）

　　3. 测试集的切分文件（The segmented test file）

准确率(Precision)和召回率(Recall)

Precision = 正确切分出的词的数目/切分出的词的总数

Recall = 正确切分出的词的数目/应切分出的词的总数

综合性能指标F-measure

Fβ = (β2 + 1)\*Precision\*Recall/(β2\*Precision + Recall)

β为权重因子，如果将准确率和召回率同等看待，取β = 1，就得到最常用的F1-measure

F1 = 2\*Precisiton\*Recall/(Precision+Recall)

未登录词召回率(R\_OOV)和词典中词的召回率(R\_IV)

R\_OOV = 正确切分出的未登录词的数目/标准答案中未知词的总数

R\_IV = 正确切分出的已知词的数目/标准答案中已知词的总数

=======================================================

另附词性标注表如下：

1. 名词 (1个一类，7个二类，5个三类)

名词分为以下子类：

n 名词

nr 人名

nr1 汉语姓氏

nr2 汉语名字

nrj 日语人名

nrf 音译人名

ns 地名

nsf 音译地名

nt 机构团体名

nz 其它专名

nl 名词性惯用语

ng 名词性语素

2. 时间词(1个一类，1个二类)

t 时间词

tg 时间词性语素

3. 处所词(1个一类)

s 处所词

4. 方位词(1个一类)

f 方位词

5. 动词(1个一类，9个二类)

v 动词

vd 副动词

vn 名动词

vshi 动词“是”

vyou 动词“有”

vf 趋向动词

vx 形式动词

vi 不及物动词（内动词）

vl 动词性惯用语

vg 动词性语素

6. 形容词(1个一类，4个二类)

a 形容词

ad 副形词

an 名形词

ag 形容词性语素

al 形容词性惯用语

7. 区别词(1个一类，2个二类)

b 区别词

bl 区别词性惯用语

8. 状态词(1个一类)

z 状态词

9. 代词(1个一类，4个二类，6个三类)

r 代词

rr 人称代词

rz 指示代词

rzt 时间指示代词

rzs 处所指示代词

rzv 谓词性指示代词

ry 疑问代词

ryt 时间疑问代词

rys 处所疑问代词

ryv 谓词性疑问代词

rg 代词性语素

10. 数词(1个一类，1个二类)

m 数词

mq 数量词

11. 量词(1个一类，2个二类)

q 量词

qv 动量词

qt 时量词

12. 副词(1个一类)

d 副词

13. 介词(1个一类，2个二类)

p 介词

pba 介词“把”

pbei 介词“被”

14. 连词(1个一类，1个二类)

c 连词

cc 并列连词

15. 助词(1个一类，15个二类)

u 助词

uzhe 着

ule 了 喽

uguo 过

ude1 的 底

ude2 地

ude3 得

usuo 所

udeng 等 等等 云云

uyy 一样 一般 似的 般

udh 的话

uls 来讲 来说 而言 说来

uzhi 之

ulian 连 （“连小学生都会”）

16. 叹词(1个一类)

e 叹词

17. 语气词(1个一类)

y 语气词(delete yg)

18. 拟声词(1个一类)

o 拟声词

19. 前缀(1个一类)

h 前缀

20. 后缀(1个一类)

k 后缀

21. 字符串(1个一类，2个二类)

x 字符串

xx 非语素字

xu 网址URL

22. 标点符号(1个一类，16个二类)

w 标点符号

wkz 左括号，全角：（ 〔 ［ ｛ 《 【 〖 〈 半角：( [ { <

wky 右括号，全角：） 〕 ］ ｝ 》 】 〗 〉 半角： ) ] { >

wyz 左引号，全角：“ ‘ 『

wyy 右引号，全角：” ’ 』

wj 句号，全角：。

ww 问号，全角：？ 半角：?

wt 叹号，全角：！ 半角：!

wd 逗号，全角：， 半角：,

wf 分号，全角：； 半角： ;

wn 顿号，全角：、

wm 冒号，全角：： 半角： :

ws 省略号，全角：…… …

wp 破折号，全角：—— －－ ——－ 半角：--- ----

wb 百分号千分号，全角：％ ‰ 半角：%

wh 单位符号，全角：￥ ＄ ￡ ° ℃ 半角：$