# 目标

## 新词发现

根据一批文本自动分词/发现新词

## 文本规范化

### 自动断句

对没有标点符号的文本，自动检测句子边界。

"*电池耐用待机时间长软件也不错后盖很漂亮* " => "*电池耐用。待机时间长。软件也不错。后盖很漂亮*"

### 词语纠错

对错误的词语进行纠错

"*反映快*" => "*反应快*"

### 无意义文本过滤

无意义内容

"*阿金SOHO定你定了打进去最阿拉搜好萌啊岁哦陪咋咯啊嗲啤酒外婆和jar李子叶*"

无关内容

"*色香味俱全*"

## 归一化优化

完善之前的归一化逻辑

## 句法规则完善

统计句法规则的分布情况，针对占比超过90%的规则特别处理

# 实现情况

## 新词发现

## 文本规范化

### 自动断句

Sentence Bound Detection

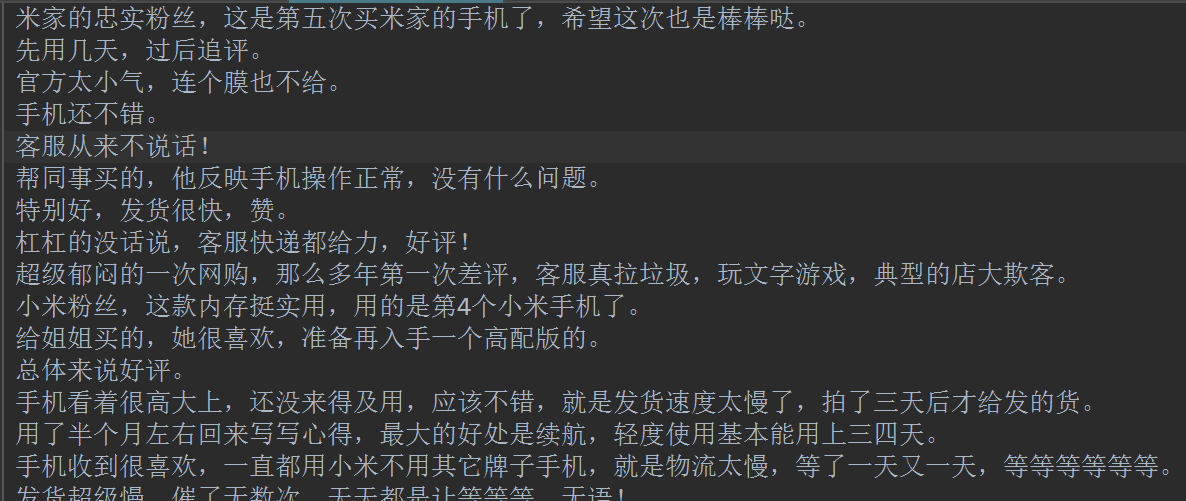
中文在这方面的研究很少；英文多偏向于检测是否是标点，比如MR.J，自动断句的也是很少。

应用场景：语音转文字

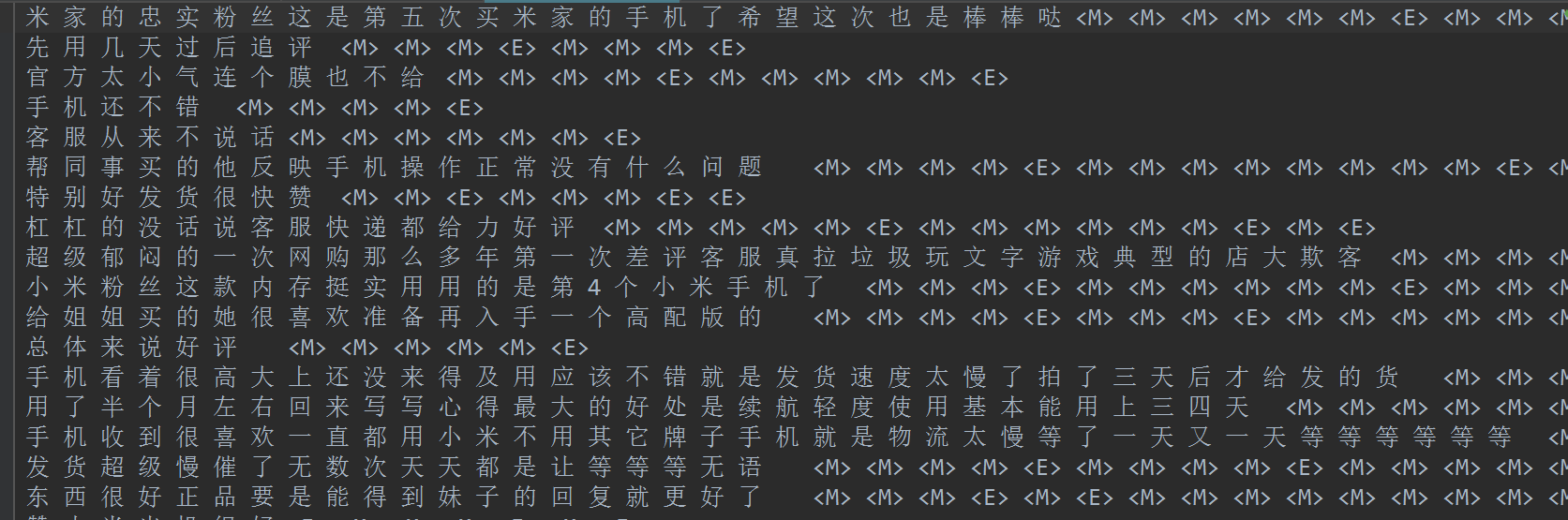
实现方法：深度学习，sequence2sequence模型

#### 训练

筛选评论文本，提取出带有[，。！？?]的句子



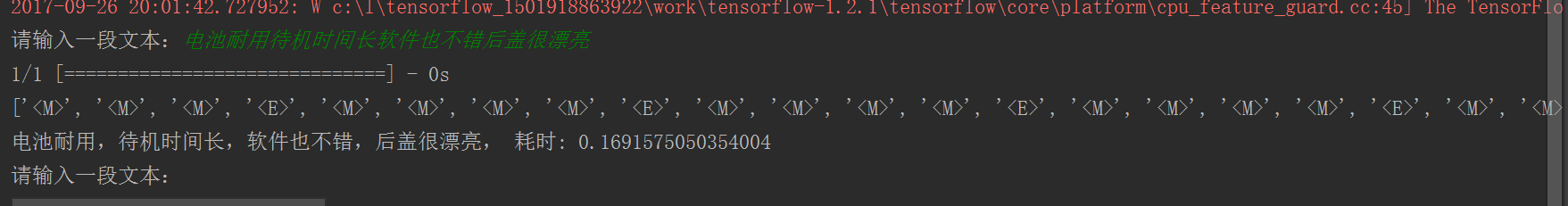
生成标注数据，<M>=句子前面的字 <E>句子最后一个字



训练，Keras LSTM，sequence2sequence模型

1.7W+数据

#### 预测



#### 结果

总量：10058

完全正确：3808 37.9%

错误，与原句不同的即认为是错的，其中有些是可以接受的

test\_sbd.py

### 词语纠错

英文里有拼写错误，中文有书写错误，但现在都是通过计算机输入，不会出现书写错误。多为同音字错误，这里我们处理的就只有同音字错误。

#### 方案一

方式1

训练两个语言模型Hanzi Language Model、Pinyin Language Model

prob1 = hanziModel.predict\_prob(xxx)

pob2 = pinyinModel.predict\_prob(xxx)

prob2/prob1的值与阈值比较

方式2

Hanzi Language Model

文本 -> 拼音 -> 找出所有同音字，取概率最大的组合

手机不错

-> shou ji bu cuo

-> [手 收 受 ...] [机 即 级 ...] [不 部 ...] [措 错 挫 ...]

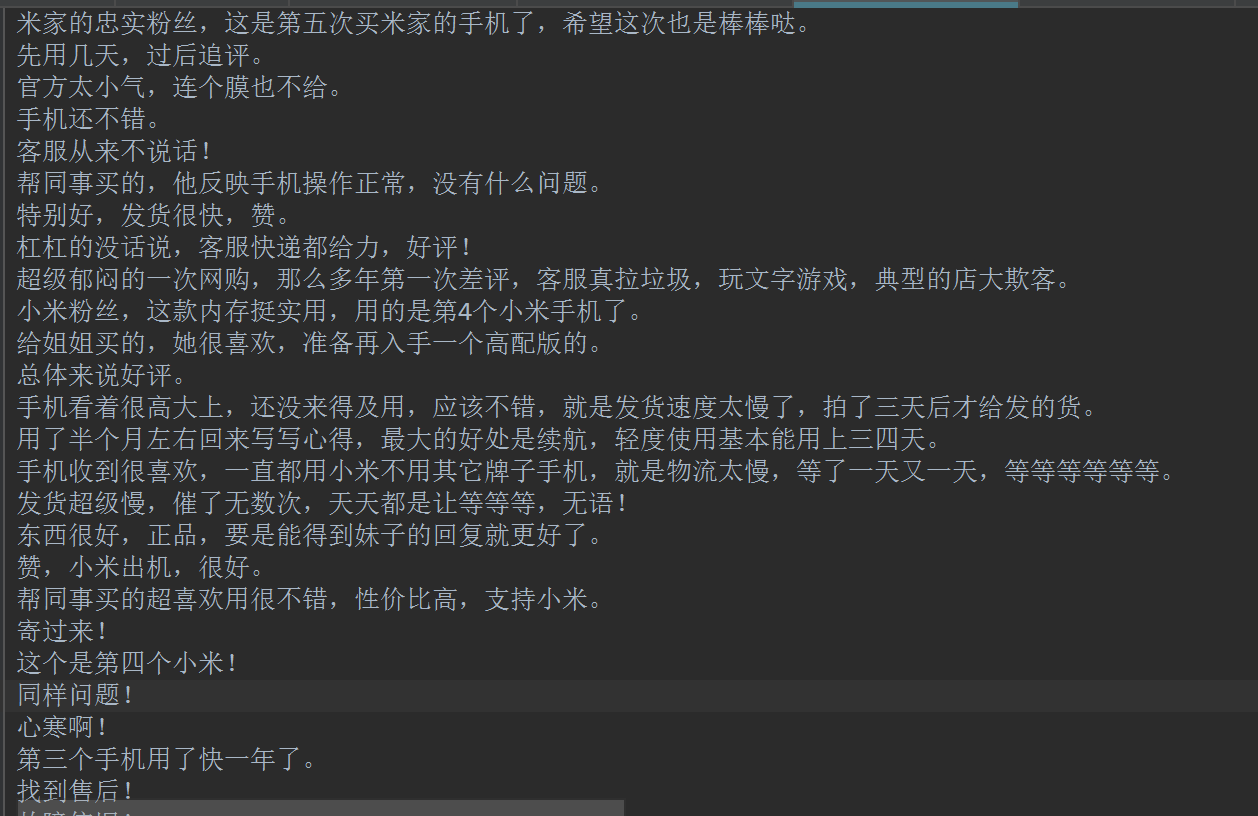
-> prob(手机不错)的概率最大

#### 方案二

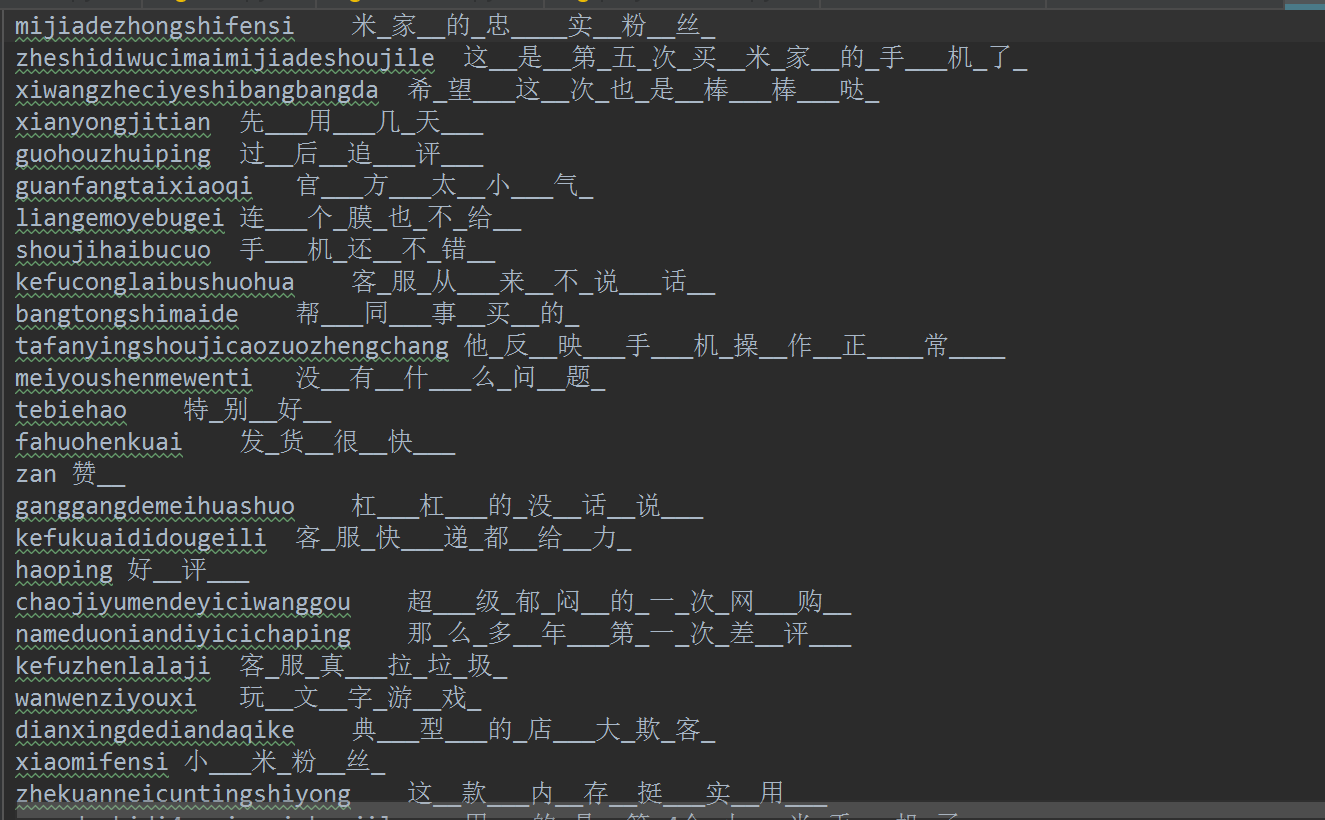
Sequence2Sequence Model

拼音 -> 汉字

**原始文本**



**构建标注数据**

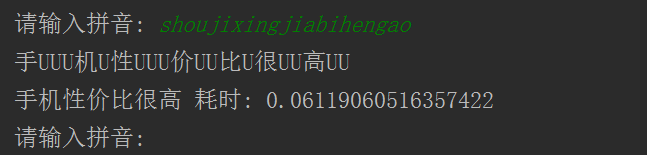


**训练**

Floyd平台，1个GPU，训练了10+小时，epoch\_15，

训练数据100W+

**预测**



#### 结果



总评论数：2000

不能转换的：106 5.3%

能转换的：1894 94.7%

认为有错的：124

拼音转汉字准确率：> 93.5% (1894-124)/1894

Jiucuow

test\_wed.py

### 无意义文本过滤

语言模型

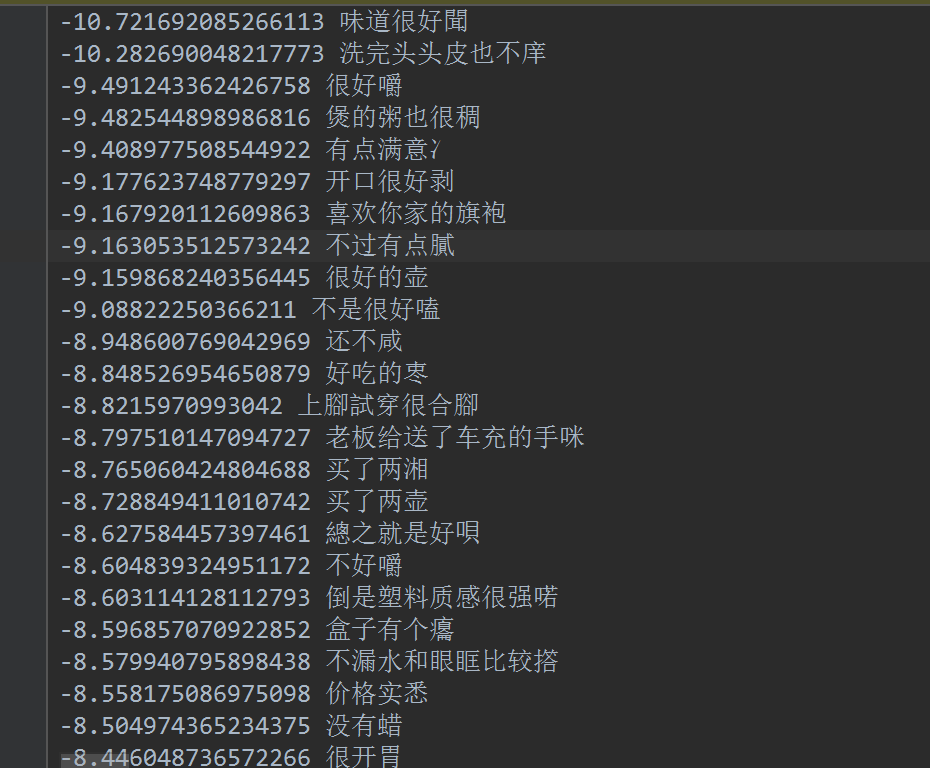
prob = lmModel.predict\_prob(xxxx)

prob低于阈值即为无意义文本

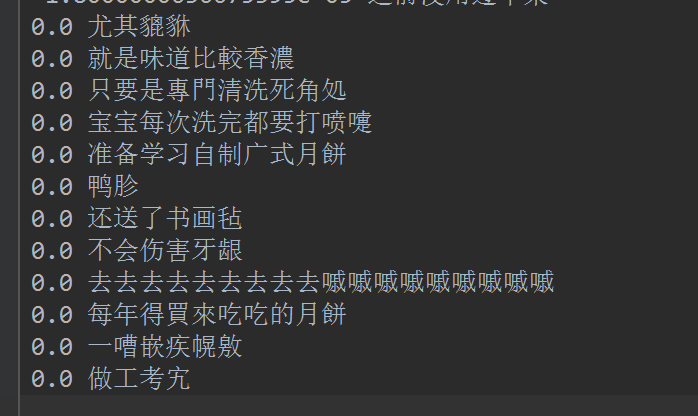
#### 结果



1. 无关内容



1. 无意义文本



## 归一化优化

重新训练word2vec模型，改善聚类效果

## 句法规则完善



根据已有的特征词/评价词，统计它们之间的关系

# 结果



目前：

total: 182, extract: 149, right: 132

准确率:88.6%, 召回率:72.5%, F1: 79.8%

之前：

准确率:90%, 召回率:59%, F1: 71%

1. SBD增大训练数据
2. 无人工下应用到新品类
3. 固定一版评估库用于最后效果评测

准确率：90% 召回率：80-85%