长句切分

零、目录

一、任务背景：一逗到底

关于推广电子社保卡的宣传稿广大居民朋友们:为回应群众关切，简化办事流程，提升公共服务，人社部签发了电子社保卡，是社保卡线上应用的有效凭证，优点如下:一、申领便捷渠道多元，手机申请刷脸认证可通过官方APP和社会渠道自愿申领，2,表现形式多元，手机端显示二维码和信息，三，全国统一用途广泛，已在26个省份，230个城市发放，四功能多样智慧化，融入智慧城市建设，实现一码通城，医保移动支付确保信息安全。

随着社保卡的推广线上线下服务体系的建立，持卡群众将获得更为贴心的服务，尽享信息时代的速度与激情。XX市人社局x

xxx年xx月xx日

总得分 25.5/35分

缺点

【主体-表现形式多元方面的优点】对"表现形式多元。手机端显示二维码和信息"这一方面的要点，写得不够全面、准确。

★能力点拨：要提高概括能力。资料[materialid]239389[/materialid]**段3**提到"电子社保卡有两种表现形式……"，说明电子社保卡的表现形式不单一，是电子社保卡的优点，可以写在宣传稿正文部分，概括出要点：**表现形式多元。**

★能力点拨：要提高摘抄能力。资料[materialid]239389[/materialid]**段3**提到"电子社保卡有两种表现形式，一是手机端显示电子社保卡二维码，用于信息系统识别人员身份、缴费结算、办理业务；二是手机端显示的与实体卡一致的电子社保卡信息，用于人工核对并办理业务"，指出电子社保卡的具体表现形式，可视为电子社保卡的优点，写在宣传稿正文部分，从中可摘抄出要点：

二、数据集

2.1 正则处理，替换符号

①正则处理，替换符号

高尔基说：“知识是我们这个社会的绝对价值。”知识究竟有什么价值呢？

高尔基说：知识是我们这个社会的绝对价值。知识究竟有什么价值呢。

2.2 数据集

	训练集	验证集	测试集
2014人民日报	9.8w	1.2w	小题(0.15w) 大作文(0.2w)
2017人民日报	107w	13w	

2.3 测试集样例

1） 小题0.15w

一、市民环保意识差，乱扔垃圾。

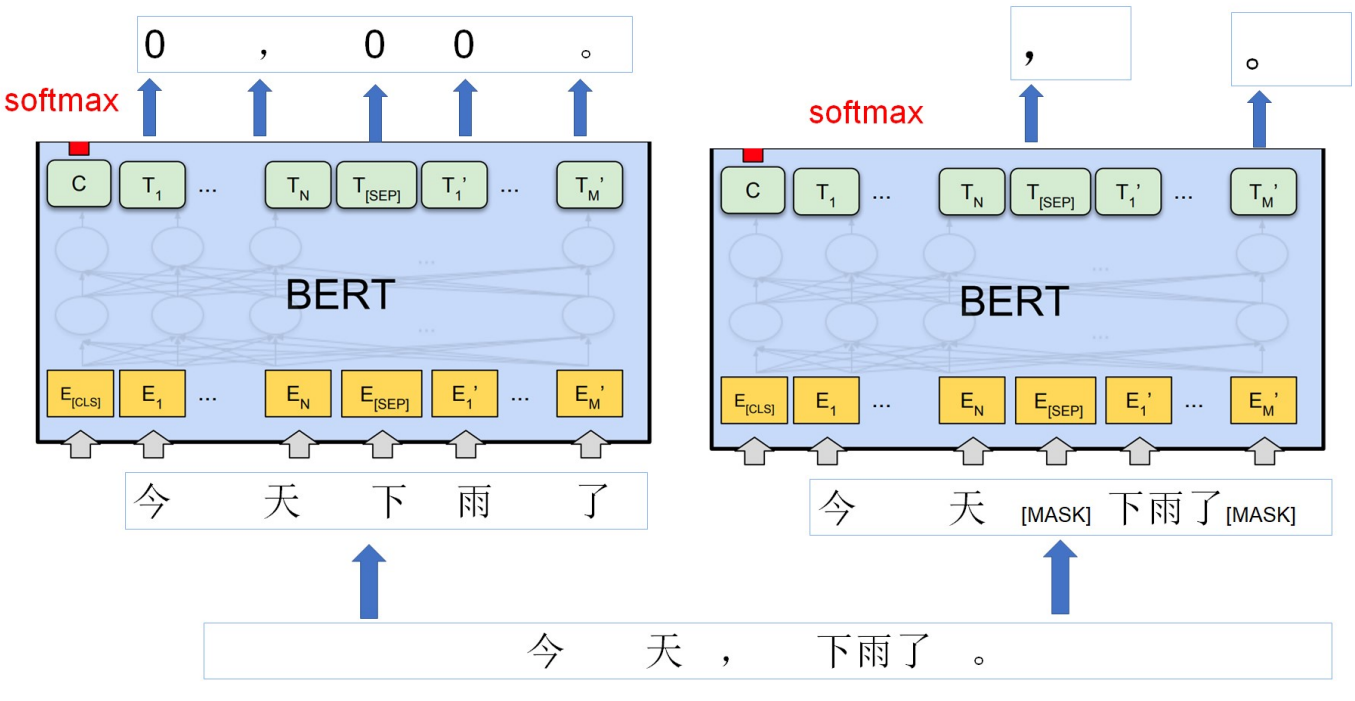
二、垃圾量多，垃圾处理能力弱；大件垃圾多，运输难，焚烧难。

2）大作文0.2w

高尔基说：“知识是我们这个社会的绝对价值。”知识究竟有什么价值呢？

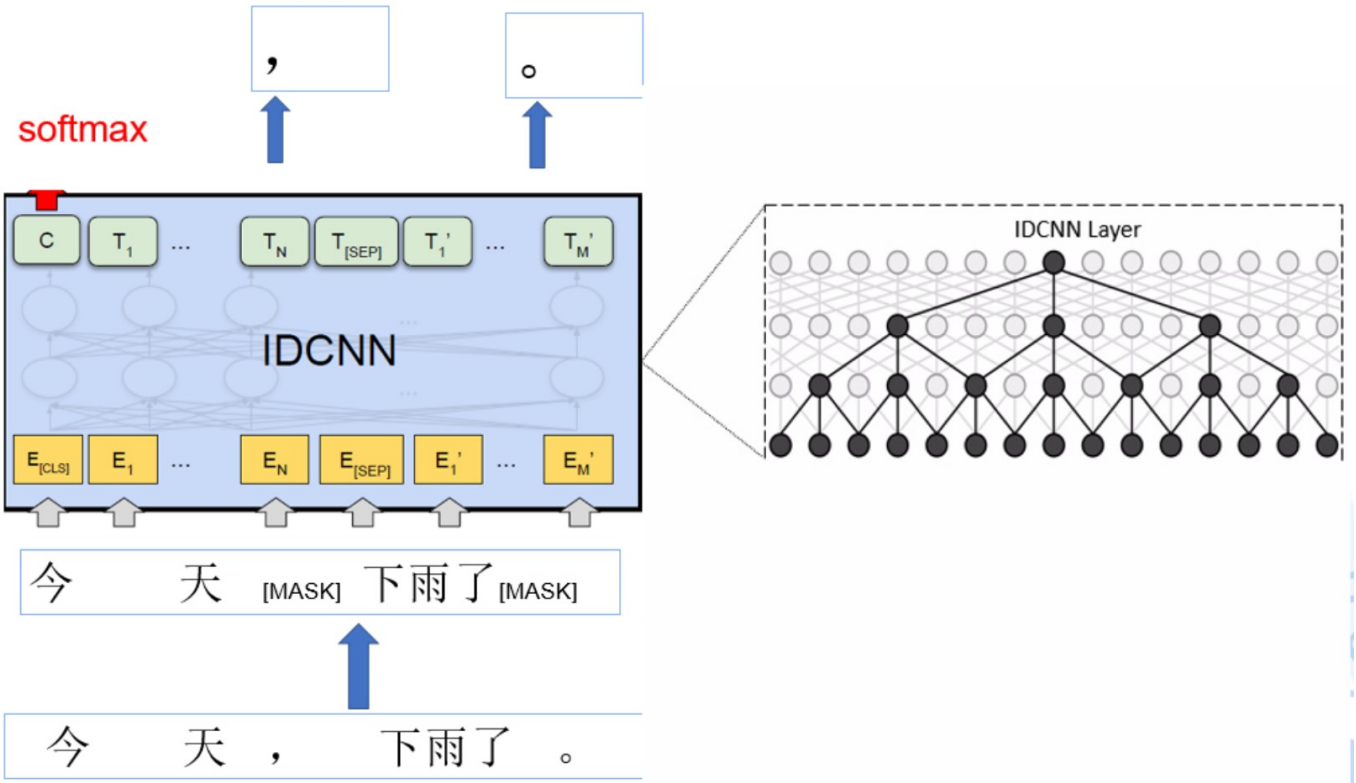
三、模型介绍

3.1 BERT +ASR纠错 vs BERT模型+剪枝



3.2 IDCNN模型

3.2.1 模型结构



3.2.2 膨胀卷积系数选择

为了考虑句子间的关系，膨胀系数不同于NER设置的1,1,2，设置为1，2，5。

3.2.3 预训练

模拟BERT mask机制，对IDCNN也进行了类似的预训练

四、实验结果

4.1 离线测试结果

测试集21-25类型 最佳F1值	BERT+ASR替换纠错	BERT+剪枝	BiLSTM	IDCNN
2014 corpus	0.76	0.81	0.69（768维）	0.68（768）
2017 corpus	0.81	0.85	0.76（768维）	0.78（400维）

4.2 推理速度

IDCNN参数：训练集107w,width=3,lr =0.001 ,输出层dropout=0.5 ,膨胀系数为1，2，5

BiLSTM参数：训练集107w,lr =0.001 ,输出层dropout=0.5

BERT：训练集107w,lr=2e-5,输出层dropout=0.1

	BERT+剪枝	BiLSTM	IDCNN
F1值	0.85	0.76	0.78(400维,膨胀系数1，2，5)
P100 GPU 并发20 单次预测(ms)	9	6	2.5

五、业务结果

	正确率	速度
长句模型切分	22%->78%	1->1/3（提升3倍）

数字断句

一、任务介绍

旧方案

```
            {"1", "2", "3", "4", "5", "6", "7", "8", "9", "10"},
            {"1.", "2.", "3.", "4.", "5.", "6.", "7.", "8.", "9.", "10."},
            {"1、", "2、", "3、", "4、", "5、", "6、", "7、", "8、", "9、",
"10、"},
            {"1)", "2)", "3)", "4)", "5)", "6)", "7)", "8)", "9)", "10)"},
            {"一", "二", "三", "四", "五", "六", "七", "八", "九", "十"},
            {"A.", "B.", "C.", "D.", "E.", "F.", "G.", "H."},
            {"a.", "b.", "c.", "d.", "e.", "f.", "g.", "h."},
            {"首先", "其次", "再次"},
            {"首先", "其次", "最后"},
            {"①", "②", "③", "④", "⑤", "⑥", "⑦", "⑧", "⑨", "⑩"},
```

二、总体方案介绍

2.1 原有方法

1. 先以句号断开
2. 查看每个序列

a. 遍历组别如图A 中的10行组别，识别序列中的数字

b. 若数字从0开始且顺序，那么断开

c.

2.2 缺点

1. 句号断开的序列，缺失1,有2，3，4无法断开
2. 句号断开后的序列，
3. 有些1，2，4分在不同的部分无法断开

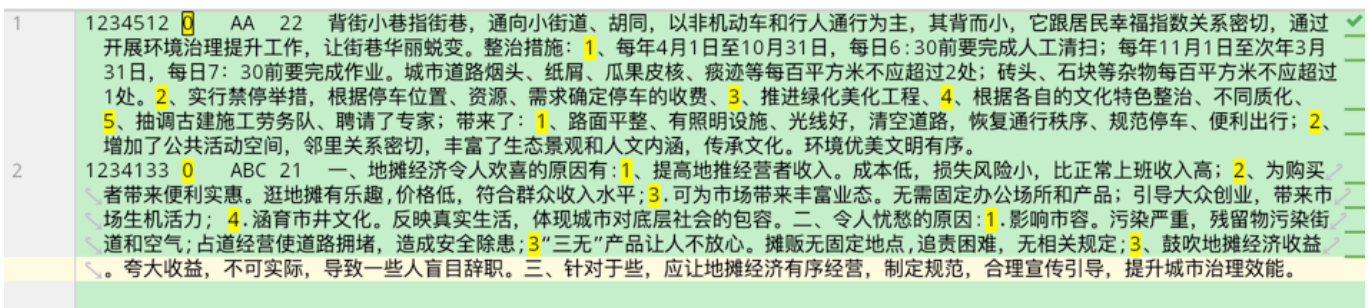
2.3 新的方法

1. 数字断开位置：
   a. 识别整句所有的数字
   b. 依据数字【是否顺序、相邻字符个数、相隔字符和】等将数字序列划分为ABC3个等级，如下图的1234512划分为2个等级，AA
   c. A，B等级中的数字均可断开，C等级中的数字不断开
2. 汉字数字断开位置，同1
3. 其他断开位置，同1
4. 标点断开位置，同1
5. 合并上述所有的断开位置，一起断开

2.4 新方法例子介绍

新的正则：

```
("(?<![a-z0-9A-Z])(?<![0-9][:：.])[0-9](?![a-z0-9A-Z处位个万月日次岁至到~～-])(?![:：.][0-9])");
```



图B

2.4.1 正则处理如上图，抽取所有符合条件的数字

【数字切分图】中，如第1行抽取的是1234512序列，第2行抽取的是1234133序列

2.4.2 对数字的组合分类

A:

①12

②3个数字以上，要求顺序紧密，比如456

B：3种情况

①1，3

②顺序紧密的2个数字，比如56

③顺序不紧密，比如125，对于K个字符，相邻的gap要求<k

C:

以上其他情况

图B中所示：

- 1234512->AA序列
- 1234133->ABC序列

2.4.3 处理办法

1.对于A,B按照里面数字的下标直接断开

2.C不考虑

三、其他例子和办法

3.1 数字(1->9)

3.1.1 识别的数字没有歧义

旧方法

```
1234512 0 AA 22 背街小巷指街巷，通向小街道、胡同，以非机动车和行人通行为主，其背而小，它跟居民幸福指数关系密切，通过开展环境治理提升工作，让街巷华丽蜕变。整治措施：1、每年4月1日至10月31日，每日6:30前要完成人工清扫；每年11月1日至次年3月31日，每日7：30前要完成作业。城市道路烟头、纸屑、瓜果皮核、痰迹等每百平方米不应超过2处；砖头、石块等杂物每百平方米不应超过1处。2、实行禁停举措，根据停车位置、资源、需求确定停车的收费、3、推进绿化美化工程、4、根据各自的文化特色整治、不同质化、5、抽调古建施工劳务队、聘请了专家；带来了：1、路面平整、有照明设施、光线好，清空道路，恢复通行秩序、规范停车、便利出行；2、增加了公共活动空间，邻里关系密切，丰富了生态景观和人文内涵，传承文化。环境优美文明有序。
```

缺点：

1. 月，日的数字不应该包含
2. 7：30分的7不应该包含
3. 1处的处不应该包含

新方法

识别[.\()]等格式

```
1234512 0 AA 22 背街小巷指街巷，通向小街道、胡同，以非机动车和行人通行为主，其背而小，它跟居民幸福指数关系密切，通过开展环境治理提升工作，让街巷华丽蜕变。整治措施：1、每年4月1日至10月31日，每日6:30前要完成人工清扫；每年11月1日至次年3月31日，每日7：30前要完成作业。城市道路烟头、纸屑、瓜果皮核、痰迹等每百平方米不应超过2处；砖头、石块等杂物每百平方米不应超过1处。2、实行禁停举措，根据停车位置、资源、需求确定停车的收费、3、推进绿化美化工程、4、根据各自的文化特色整治、不同质化、5、抽调古建施工劳务队、聘请了专家；带来了：1、路面平整、有照明设施、光线好，清空道路，恢复通行秩序、规范停车、便利出行；2、增加了公共活动空间，邻里关系密切，丰富了生态景观和人文内涵，传承文化。环境优美文明有序。
```

3.1.2 识别不同形式的数字

旧方法：

```
1234512 0 AA 22 背街小巷指街巷，通向小街道、胡同，以非机动车和行人通行为主，其背而小，它跟居民幸福指数关系密切，通过开展环境治理提升工作，让街巷华丽蜕变。整治措施：1、每年4月1日至10月31日，每日6:30前要完成人工清扫；每年11月1日至次年3月31日，每日7：30前要完成作业。城市道路烟头、纸屑、瓜果皮核、痰迹等每百平方米不应超过2处；砖头、石块等杂物每百平方米不应超过1处。2实行禁停举措，根据停车位置、资源、需求确定停车的收费、3）推进绿化美化工程、4、根据各自的文化特色整治、不同质化、5.抽调古建施工劳务队、聘请了专家；带来了：1、路面平整、有照明设施、光线好，清空道路，恢复通行秩序、规范停车、便利出行；（2）增加了公共活动空间，邻里关系密切，丰富了生态景观和人文内涵，传承文化。环境优美文明有序。
```

缺点：

1. 原来的方法只能识别3）

新方法：

```
1234512 0 AA 22 背街小巷指街巷，通向小街道、胡同，以非机动车和行人通行为主，其背而小，它跟居民幸福指数关系密切，通过开展环境治理提升工作，让街巷华丽蜕变。整治措施：1、每年4月1日至10月31日，每日6:30前要完成人工清扫；每年11月1日至次年3月31日，每日7：30前要完成作业。城市道路烟头、纸屑、瓜果皮核、痰迹等每百平方米不应超过2处；砖头、石块等杂物每百平方米不应超过1处。2实行禁停举措，根据停车位置、资源、需求确定停车的收费、3）推进绿化美化工程、4、根据各自的文化特色整治、不同质化、5.抽调古建施工劳务队、聘请了专家；带来了：1、路面平整、有照明设施、光线好，清空道路，恢复通行秩序、规范停车、便利出行；（2）增加了公共活动空间，邻里关系密切，丰富了生态景观和人文内涵，传承文化。环境优美文明有序。
```

3.2 汉字数字的【一->九】

3.2 1增加对第一、第二的识别

旧方法

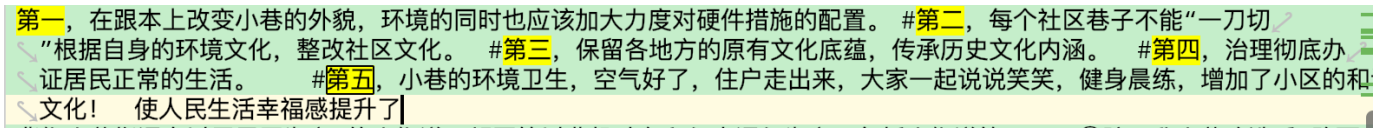
```
第一，在跟本上改变小巷的外貌，环境的同时也应该加大力度对硬件措施的配置。 #第二，每个社区巷子不能“一刀切”根据自身的环境文化，整改社区文化。 #第三，保留各地方的原有文化底蕴，传承历史文化内涵。 #第四，治理彻底办证居民正常的生活。 #第五，小巷的环境卫生，空气好了，住户走出来，大家一起说说笑笑，健身晨练，增加了小区的和文化！ 使人民生活幸福感提升了
```

缺点

1. 一刀切被识别进来

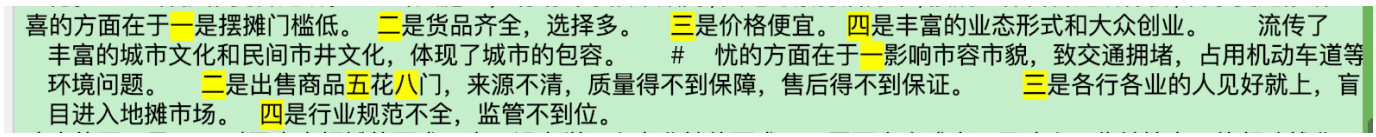
2. 一起也被识别了

新方法



3.2.2 增加一是，二是等识别

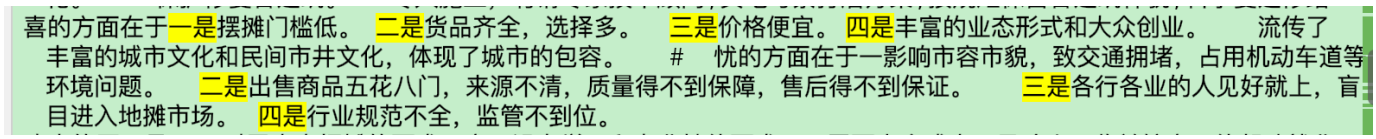
旧方法



缺点

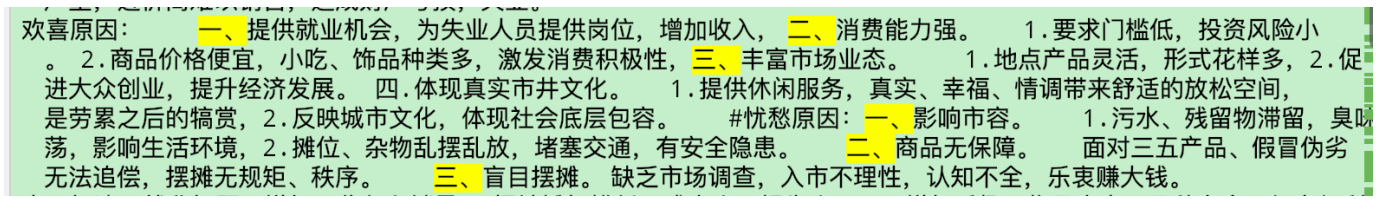
1. 五花八门的八不应该被识别出来

新方法



3.2.3 识别不同形式的一二三四

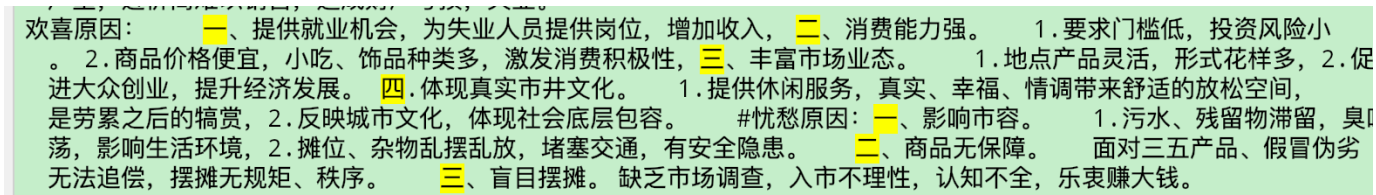
旧方法：



缺点：

1. 四.没有识别出来

新方法：



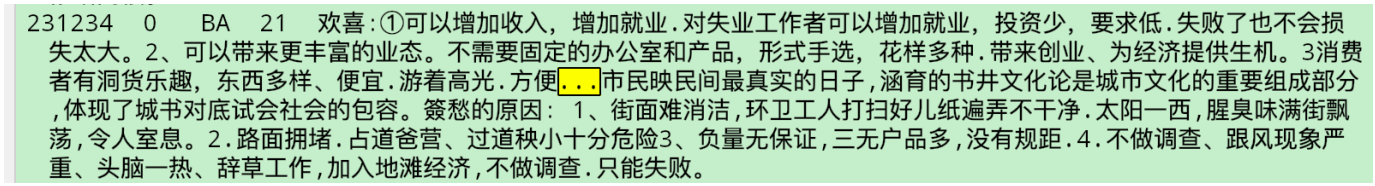
3.3 标点断开[！ !?? ； ；。 …]

旧方法：

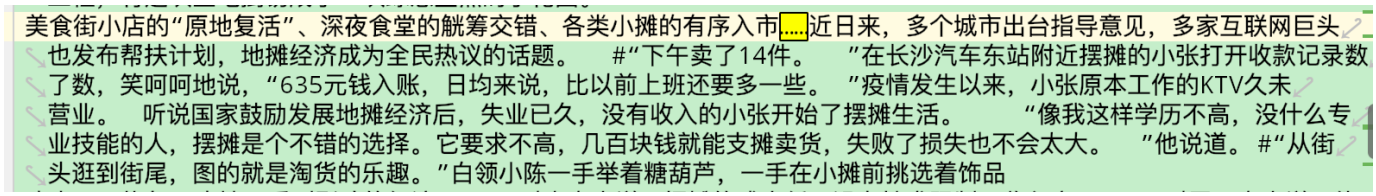
没有处理

新方法：

1.3个点



2.连续的6个点



四、实验结果

	正确率	速度
数字切分长句	81%->92%	1->1/10（提升10倍）