数字海南·智创未来

第二届海南大数据创新应用大赛

智能算法赛道 我想去海南 队伍

队员：张啸宇、徐建

主办单位：海南省大数据管理局

承办单位：数字海南有限公司

海南省大数据产业联盟

支持单位：海南省农业农村厅　海南省教育厅　海南省卫生健康委员会

协办单位：阿里云计算有限公司　中国电信股份有限公司海南分公司

浪潮云信息技术有限公司　太极计算机股份有限公司

任务描述

- 基本信息：姓名、出生年月、性别、电话等
- 教育信息：毕业院校、学位、毕业时间
- 工作信息：工作单位、工作时间、工作内容等
- 项目信息：项目名称、项目实践、项目责任等

电话：[redacted] 出生年月：1962.09 户口:甘肃省陇南市

教育背景

2003.01-2007.01 中央戏剧学院 硕士学位 体育学

工作经历

2003.07-2018.07 熊猫精酿酒业有限公司 应用运维工程师

工作内容:

1、协助项目工程管理人员做好工程维修预结算，报送至甲方；2、负责各个工程项目施工图纸的领取、整理、发放、归档；3、了解国家及各过程所在地的过程验收资料规范，编制验收资料；4、其他项目管理人员指派的工作任务。

2008/12-2012/01 上海德伟思教育培训有限公司 业务跟单

工作内容:

1.负责工程项目概预算编制；2.根据合同及进度负责工程项目资金费用的初审及申请；3.负责根据各类签证费用要求办理符合合同约定的变更签证手续；4.参与合同招标、评标和谈判工作。

问题建模

- 输入：文本序列
- 目标信息类型：长文本、短语片段
- 长文本：句子为单位的序列分类
- 短语片段：BIO信息抽取

[redacted]

电话：[redacted] 出生年月：1962.09 户口:甘肃省陇南市

教育背景

2003.01-2007.01 中央戏剧学院 硕士学位 体育学

工作经历

2003.07-2018.07 熊猫精酿酒业有限公司 应用运维工程师

工作内容:

1、协助项目工程管理人员做好工程维修预结算，报送至甲方；2、负责各个工程项目施工图纸的领取、整理、发放、归档；3、了解国家及各过程所在地的过程验收资料规范，编制验收资料；4、其他项目管理人员指派的工作任务。

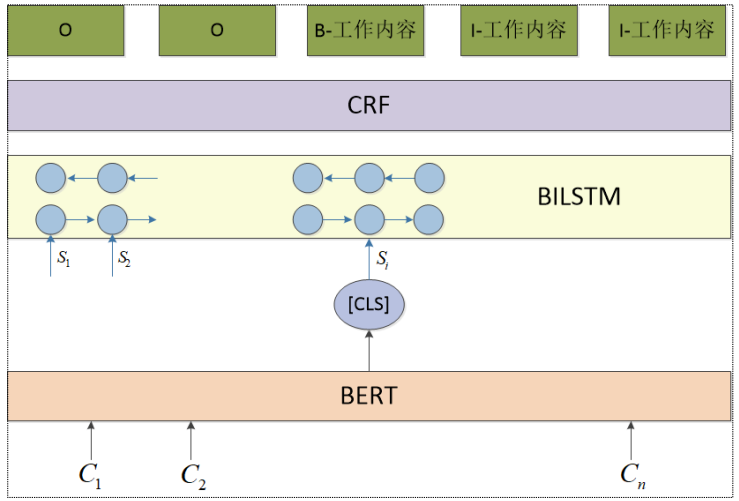
2008/12-2012/01 上海德伟思教育培训有限公司 业务跟单

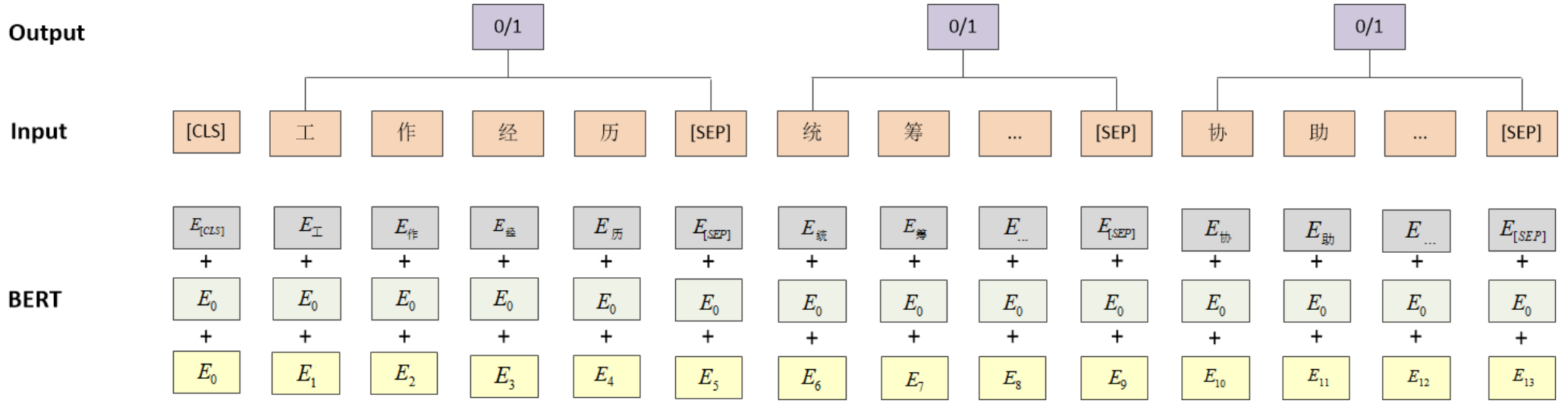
工作内容:

1.负责工程项目概预算编制；2.根据合同及进度负责工程项目资金费用的初审及申请；3.负责根据各类签证费用要求办理符合合同约定的变更签证手续；4.参与合同招标、评标和谈判工作。

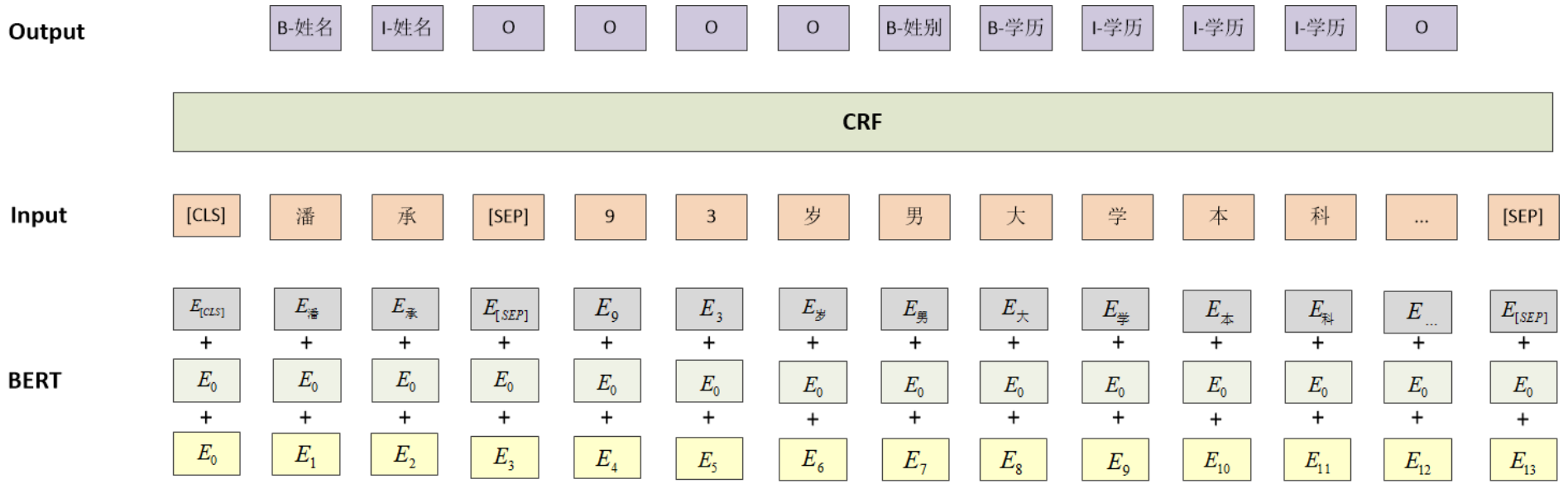
S1： “姓名”
S2：电话:xxxxxxxxxxx出生年月:1962.09户口:甘肃省陇南市
S3：教育背景
S4：2003.01~2007.01 中央戏剧学院 硕士学位 体育学
…

- **句子为单位的序列分类**
  - **方案1： BERT + BILSTM + CRF**
  - **方案2： BERT + 0/1分类**





- **BIE信息抽取**


Output
B-姓名
I-姓名
O
O
O
O
B-姓别
B-学历
I-学历
I-学历
I-学历
O
CRF
Input
[CLS]
潘
承
[SEP]
9
3
岁
男
大
学
本
科
...
[SEP]
BERT
$E_{[CLS]}$
$E_{潘}$
$E_{承}$
$E_{[SEP]}$
E_9
E_3
$E_{岁}$
$E_{男}$
$E_{大}$
$E_{学}$
$E_{本}$
$E_{科}$
$E_{...}$
$E_{[SEP]}$
E_0
E_1
E_2
E_3
E_4
E_5
E_6
E_7
E_8
E_9
E_{10}
E_{11}
E_{12}
E_{13}

- **数据处理**
  - **pdf解析器：pdfminer，tika**

- **数据处理**

  - pdf解析器： pdfminer，tika
  - **处理重叠，过滤繁体和兼容性汉字**

- **数据处理**
  - pdf解析器：pdfminer，tika
  - 处理重叠，过滤繁体和兼容性汉字
  - 模板读取顺序混乱

意向岗位：前端开发

出生日期：1949.12

籍贯：广东省广州市

工作年限：3年

电话：

邮箱：

兴趣爱好

编程、看电影、音乐

教育背景

2007.11 -- 2011.11　北京市海淀区职工大学 学士学位

工作经历

2006/04-2016/09　浙江机有网络科技有限公司　司机

1、具备视频拍摄脚本的策划能力；2、根据公司要求及脚本内容制作视频拍摄及制作；3、负责公司视频录制、拍摄制作；4、负责视频剪辑、字幕、音效及画面处理与合成输出；5、熟练使用摄像机及相应的配套附件、能独立完成剪辑特效合成；6、完成领导交办的其他任务。（需携带相关作品面试）

1999.12-2013.04　中山市广弘集团有限公司　人事助理

1.对儿童健康状况进行检测，给儿童建立健康档案；2.根据儿童的检测结果及体征情况，进行分析评估，给予个性化膳食营养、运动指导及健康促进干预方案；3.解答家长关于儿童营养健康问题的提问，提供有关儿童营养健康方面的建议；4.负责儿童营养知识教程的编写和培训；5.完成公司领导安排的工作，及配合公司各部门的相关业务；

项目经历

2009/08-2018/01

项目介绍：量子信息技术的认识论研究

项目内容：

1、接听电话，接收传真，按要求转接电话或记录信息，确保及时准确。2、对来访客人做好接待、登记、引导工作，及时通知被访人员。3、负责公司快递、信件、包裹的收发工作。4、负责办公用品的管理及采购。5、负责复印、传真和打印等设备的使用与管理工作，合理使用，降低材料消耗。6、做好会前准备、会议记录和会后内容整理工作。7、做好公司相关资料、档案管理工作。8、日常费用的申请，公司钉钉的维护管理。9、领导交办的其他人事行政工作。

数据处理

- pdf解析器： pdfminer，tika
- 处理重叠，过滤繁体和兼容性汉字
- 模板读取顺序混乱
- **解析器读取顺序混乱**

数据处理

- pdf解析器：tika，pdfminer
- 繁体、重叠（兼容性汉字）
- 模板读取顺序混乱
- 解析器读取顺序混乱
- 错误断行

主要经历
Project Experience

工作经历:

2006 年 07 月-2016 年 11 月　上海星舷信息技术有限公司　平面/视觉经理

工作内容:

1、数据的维护、更新及汇总，各类报表包括但不限于日报、周报、月报、佣金结算表、汇总表等的收集、汇总。2、协助部门经理做好各类文档工作，并建立项目档案，负责项目客户信息统计、更新。3、有较强的沟通能力，能独立协调与其他部门之间的工作。

● 处理细节1---区域界定

➢ **区域：**

- □ **基本属性**
- □ **教育经历**
- □ **工作经历**
- □ **项目经历**

毕业院校：北京科技经营管理学院　　学 历：博士研究生
年 龄：67 岁　　政治面貌：中国民主促进会会员
性 别：男　　籍 贯：新疆省塔城市
联系电话：
邮 箱：

工作经验

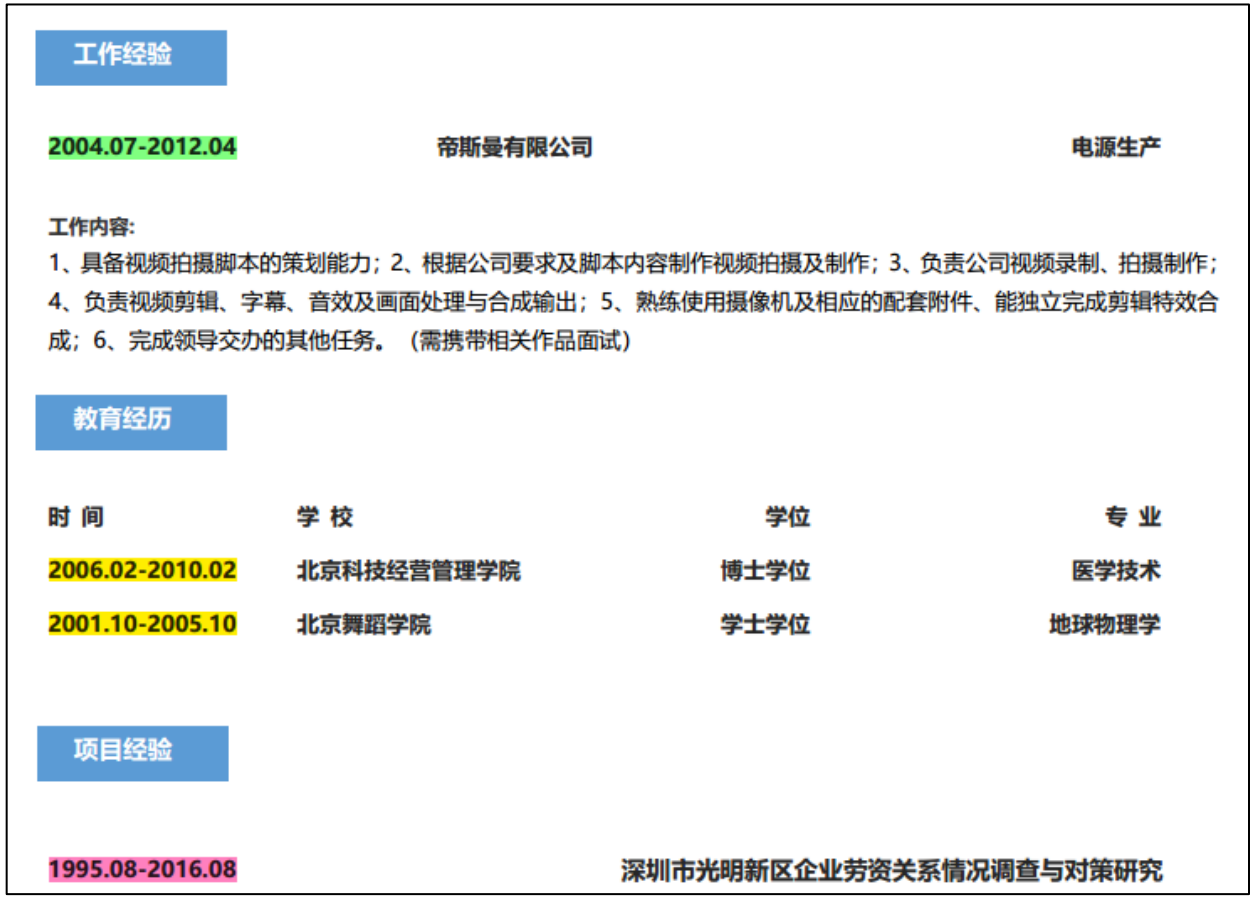
2004.07-2012.04	帝斯曼有限公司	电源生产

工作内容:
1、具备视频拍摄脚本的策划能力；2、根据公司要求及脚本内容制作视频拍摄及制作；3、负责公司视频录制、拍摄制作；4、负责视频剪辑、字幕、音效及画面处理与合成输出；5、熟练使用摄像机及相应的配套附件、能独立完成剪辑特效合成；6、完成领导交办的其他任务。（需携带相关作品面试）

教育经历

时 间	学 校	学位	专 业
2006.02-2010.02	北京科技经营管理学院	博士学位	医学技术
2001.10-2005.10	北京舞蹈学院	学士学位	地球物理学

- **处理细节2---同类合并**
  - **时间类：工作时间、毕业时间、项目时间**
  - **内容类：工作内容、项目责任**



工作经验

2004.07-2012.04　　帝斯曼有限公司　　电源生产

工作内容:
1、具备视频拍摄脚本的策划能力；2、根据公司要求及脚本内容制作视频拍摄及制作；3、负责公司视频录制、拍摄制作；4、负责视频剪辑、字幕、音效及画面处理与合成输出；5、熟练使用摄像机及相应的配套附件、能独立完成剪辑特效合成；6、完成领导交办的其他任务。（需携带相关作品面试）

教育经历

时 间	学 校	学位	专 业
2006.02-2010.02	北京科技经营管理学院	博士学位	医学技术
2001.10-2005.10	北京舞蹈学院	学士学位	地球物理学

项目经验

1995.08-2016.08　　深圳市光明新区企业劳资关系情况调查与对策研究

- **处理细节3---全体标注**

  - **标注所有目标片段**

毕业院校：北京科技经营管理学院　　学 历：博士研究生
年 龄：67 岁　　政治面貌：中国民主促进会会员
性 别：男　　籍 贯：新疆省塔城市
联系电话：
邮 箱：

工作经验

2004.07-2012.04	帝斯曼有限公司	电源生产

工作内容:
1、具备视频拍摄脚本的策划能力；2、根据公司要求及脚本内容制作视频拍摄及制作；3、负责公司视频录制、拍摄制作；4、负责视频剪辑、字幕、音效及画面处理与合成输出；5、熟练使用摄像机及相应的配套附件、能独立完成剪辑特效合成；6、完成领导交办的其他任务。（需携带相关作品面试）

教育经历

时 间	学 校	学位	专 业
2006.02-2010.02	北京科技经营管理学院	博士学位	医学技术
2001.10-2005.10	北京舞蹈学院	学士学位	地球物理学

- **讨论**

  - **验证集 vs 测试集： 0.99+ vs 0.85+**
    - **数据分布不一致**
    - **区域划分**

讨论

- 验证集 vs 测试集： 0.99+ vs 0.85+
- 泛化性：
  - 对抗扰动 FGM： 0.8553 vs 0.8492 ↓
  - 5模型融合： 0.8549 vs 0.8553 ↑

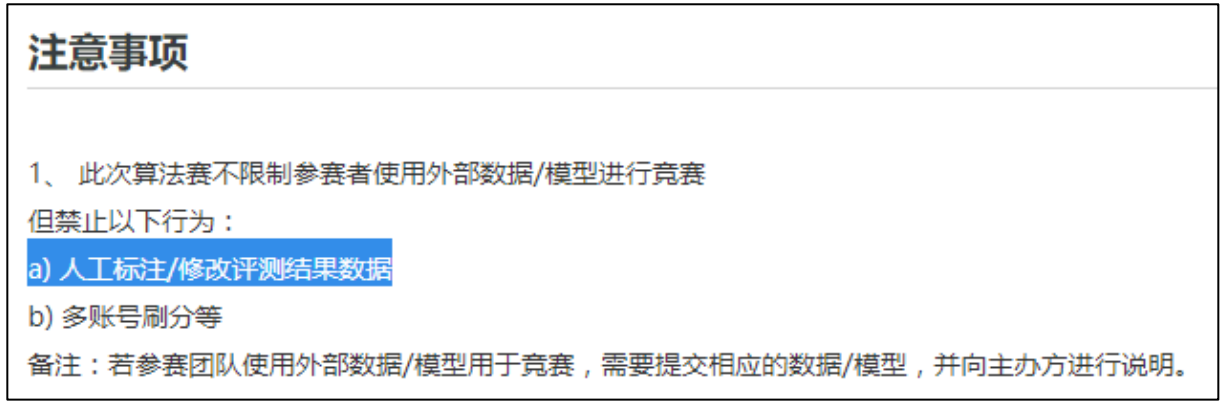
训练数据信息已经得到了充分学习

讨论

- 验证集 vs 测试集： 0.99+ vs 0.85+
- 泛化性
- 未来改进：
  - 区域划分
  - 人工标注

讨论

- 验证集 vs 测试集： 0.99+ vs 0.85+
- 泛化性
- 未来改进：
  - 区域划分
  - 人工标注
    - 模板10种，2000条人造简历
    - 90%训练数据 vs 100%训练数据： 0.8423 vs 0.8549


注意事项

1、 此次算法赛不限制参赛者使用外部数据/模型进行竞赛

但禁止以下行为：

a) 人工标注/修改评测结果数据

b) 多账号刷分等

备注：若参赛团队使用外部数据/模型用于竞赛，需要提交相应的数据/模型，并向主办方进行说明。

格式和内容高度相似的额外200条数据在线上提升 1.2个百分点；人工标注200条真实简历会可能会带来更多提升！

Thank you

汇报人：徐建
2020.06.06