李震

16621660628 | 15726723275@163.com

求职意向：自然语言处理、机器学习

个人总结

- 软件工程专业研三在读，2021 年毕业。
- 熟悉使用 Python 开发，熟悉 pytorch、sklearn 等深度学习机器学习工具，熟悉爬虫技术。
- 有自然语言处理开发经验，了解文本分类、序列标注任务及机器学习常用算法，主要研究事件抽取相关内容。

教育经历

2018.09-2021.07	**东南大学**	**软件工程（硕士）**
2014.09-2018.07	**江苏理工学院**	**软件工程（本科）**

项目经历

2019.06 - 2020.9　　军事语料样本集构建　　项目成员

项目简介：构建大型军事语料样本集，由数据处理组件、数据标注组件、军事语料组件、自然语言处理组件组成。

主要工作：数据处理方面，负责网络数据信息的获取，清洗 。自然语言处理方面，使用同类别实体替换的方式，增强文本数据，提升 ner 模型约 2%(f1)的性能。

2018.12 - 2019.03　　银行信用卡活动信息抽取　　项目负责

项目简介：面向信用卡行业，从新卡、新活动的新闻、活动条款及细则文本中，抽取出结构化的知识。比如从官网活动公告中抽取活动时间、活动对象、活动的条件细节及相应奖励等。

主要工作：负责网络数据获取，清洗，规则及模型的构建，实验等工作，由于活动内容的细节描述变化性很强，导致活动事件 schema 难以定义，针对这个问题，通过统计高频标题的方式总结合并常见的 slot 名，并定义对应 slot 抽取方法。针对描述性变化强的数据，主要想法是利用模型进行学习，在标注过程中总结其语法特征，并编码进模型，另外通过同义词替换等方式增强文本数据。最后针对指标表现不好的 slot 的对应抽取方法进行修正完善，迭代进行，逐步提高模型性能。

2018.06 - 2018.09　　用于人脸比对的身份证网纹照片修复技术　　项目成员

项目简介：公安提供的数据是带网纹遮挡的人脸照片，这导致应用 app 通过用户所上传的人脸照片来验证用户真实身份的时候往往准确率不高。需要对包含网纹的人脸照片进行消除网纹处理，然后再与用户所上传的真实人脸照片进行比较，以提升用户身份识别的准确度。

主要工作：负责算法模型的实现，采用自编码结构和对抗训练两种方式，消除照片上的网纹，提升了应用 app 对用户身份识别的准确率。

论文工作

Hierarchical Chinese Legal event extraction via Pedal Attention Mechanism（COLING2020）

简介：从法律文书中抽取结构化的信息，为类案推送和辅助判决等应用提供支持。

主要工作：负责数据获取及处理，事件结构的定义，模型设计及实验。传统事件抽取方法无法建立论元和论元之间的关系以及事件之间的因果关系，而这些元素通常会直接影响量刑结果，为了建立论元和论元之间的关系以及事件之间的因果关系，定义了层级事件类型，设计了动态事件结构。其次，法律文书对于被告人的犯罪事实描述较为详实，造成语句序列长的问题，因此，抽取过程中不可避免的需要解决长依赖问题和指代消解问题，利用踏板注意力机制解决了长依赖问题和指代消解的问题。

实习经历

2020.06 – 2020.09　　平安科技智能疾病管理团队

信息抽取工作：

- 负责从手术并发症、适应症、禁忌症的医疗文本描述中抽取疾病实体；
- 负责从检查检验医疗文本中抽取疾病实体以及与疾病实体对应的文本介绍；
- 负责从医疗检查检验文本及诊断文本中抽取危急值，并判断病患是否达到危急值标准；

研究探索性工作：

主要任务：

在 mimic3 数据上给病患的病历描述匹配对应的 ICD code，是一个文本多标签分类任务。任务的主要难点是标签空间大（有 8921 个不同类别标签），尝试通过检索排序减少候选标签个数来解决该问题。

遇到的问题：

1. 病历表述文本平均长度 2500 左右 ,远超预训练模型 512 的限制。尝试了修改 position embedding 和 longformer 两种方式来解决该问题。
2. 文本过长，预训练模型参数量大，难以在显存设备上训练。使用 gradient checkpointing、混合精度、跨层权重共享等技术来降低模型对显存的需求。

开源贡献

2019.06 – 2019.09　　文本分类　　https://github.com/jeffery0628/text_classification　　127star

项目简介：项目使用深度学习模型进行文本分类(微博情感分类、新闻十分类)，对常用文本分类模型进行复现，调参，比较。通过这个项目对数据及模型的选择有了更深的理解，对 pytorch 的掌握更加熟练。

竞赛经历

2019.09 – 2019.12　　知乎看山杯专家发现大赛 top 5%

竞赛简介：比赛使用知乎一个月的邀请数据作为训练数据，未来一周的邀请数据作为测试数据。任务预测用户是否会接受某个新问题的邀请。比赛数据包含邀请数据、用户历史回答数据、用户画像数据及各种维度的词向量数据。评估指标是 AUC。

主要工作：1. EDA，发现用户回答具有一定规律性，如回答问题的时间，近期回答等。

2. 挖掘用户特征、问题特征、问题-用户交互特征： cosine 相似度、构造关键词、时间差，构建 attention 等特征。
3. 集成：对 LightGBM、XGboost、CatBoost、RandomForest 模型采用加权融合的方式进行集成。

荣誉奖项

- **2016-2017　　国家励志奖学金**
- **2015-2016　　国家励志奖学金**

其他

- **语言：**大学英语四/六级（CET-4/6），良好的听说读写能力，快速浏览英文论文及开发文档。
- **兴趣爱好：**游泳、网球、滑板、美剧。
- **个人主页：**http://www.jeffery.ink/