何盛源

邮箱：he.she1314@gmail.com 性别：男 年龄：29 籍贯：广东省佛山市
电话：18664232096 领英：ShengyuanHe Github：https://github.com/Jugglecomemid

数据分析硕士(统计建模、深度学习方向)。专注于自然语言深度学习和大模型互动对话等领域。熟悉 transformer 和GPT生成系模型框架，如 ChatGpt、ChatGLM、LLAMA等 LLM 语言大模型，曾参考并复现 transformer 源代码。擅长大模型应用开发，互动对话，对话管理追踪等方向。亦拥有虚拟数字人互动，机器翻译、金融商业、知识图谱等方向的互动对话及深度学习项目经验。同时也涉猎 CV 领域，如 OCR、图片文本和表格识别等算法。熟练使用网络爬虫和正则匹配算法。对 NLP 大模型和数据科学等知识有浓厚的兴趣，对数据敏感，逻辑思维能力强，善于从数据中发现规律，一直保持对世界主流AI科技的关注和探索。寻求自然语言处理或数据科学研究相关的工作。

教育经历

美国东北大学 | 分析学硕士 2018.1- 2019.6

- 课程: 数据可视化，数据库和 SQL ，数据挖掘与应用，模型预测与分析，大数据与数据管理等。

中国华南农业大学 | 国际经济与贸易经济学学士 2013.9 - 2017.6

- 课程: 统计基础，数据库应用，资本管理，投资学等。

技术与编程语言

开发平台：Pycharm、IDEA、VScode
编程语言：Python、R、Bash
神经网络框架：Pytorch、Tensorflow
外语：TOEFL(95), GRE(314)，英语六级

工作经验

NLP算法工程师 | **广州赛灵力科技有限公司**，中国广州 2023.3 - 至今

专注于大模型应用框架的开发和优化，具备丰富的AI互动对话系统设计与实现经验。熟练运用小样本学习和Prompt设计，擅长模型微调和强化学习，具备多个语言大模型部署和管理能力。对现主流大模型如 ChatGLM4、Chatgpt等语言大模型（LLM）有密切的关注和丰富的应用经验。

• **大模型应用框架开发** 主导设计并实现了达尔文大模型应用框架。达尔文互动对话框架能支持用户配置不同的对话角色，不同的对话角色也可以拥有不同的技能，知识库等属性，为用户提供智能的交互能力。框架涵盖意图分类、可自由配置的 Agent 智能体、推荐系统、用户画像抽取、多轮知识问答等功能。同时支持流式文本清洗、材料引用展示和多模态回复生成（如含有音频，视频，图片）。

• **可配置的 Agent 功能框架开发** 支持用户自定义外部api技能（如导航，查询天气等），实现了对话状态管理和追踪功能，能够精确追踪用户设定的参数并根据参数结果执行相应的技能；此外，框架亦能够根据所执行技能缺失的参数向用户反问获取，能分多轮获取某一技能的参数并在收集完毕后执行技能。此 Agent 功能执行速度快，互动性强，用户可在自然语言交互中实现其设定的技能。

• **小样本学习与Prompt设计** 熟练运用小样本学习技术，设计有效的Prompt提示词，指导大语言模型（LLM）执行任务，用极低的延时（少于0.5秒）和高准确率（高于95%准确率）完成意图分类，实体识别等传统NLP任务。同时，也设计了高效的提示词来满足不同用户配置的不同角色。

• **大模型微调** 旨在增强大模型知识问答的能力，降低模型在缺失参考材料时编造答案的概率，使用了20万条指令数据对 ChatGLM2 进行过 Lora 微调。实验效果对比在固定指令下的模型回复有了显著的提升，模型编造答案的概率降低至5%以下。

算法工程师 | **深圳来觅数据信息科技有限公司**，中国深圳 2021.6 - 2023.3

• **构建上市公司融资历程知识图谱** 旨在建立易查询、可读性高的上市公司融资历史，如融资类型、金额等，基于上万份证监会披露的上市公司招股书等非结构性数据来抽取。开发实体识别算法界定抽取候选句的实体内容，如融资方和投资方，融资金额等。再者，利用文本分类方法判断候选句的融资关系和映射关系（如股转和增资）。上述算法组合能构建出（公司，关系，公司）的三元组。以此三元组为基础延展相关属性，导入neo4j图数据库，达到简易查询效果。

• **新闻话题聚类算法** 旨在能聚集相似的新闻为客户提供相关资讯，在一百多万条新闻标题和其正文上，以无监督 skip-gram 方法训练了一个 word embedding。每个词均能在此矩阵上找到相应的特征数据。根据此特征数据用余弦相似度来构建两篇新闻的相似得分。除外，为构建多种多样的话题而非单纯的几个聚类，开发了 single_pass 算法，通过调节相似得分阈值能决定新闻能否编入到同一个话题中。

• **实体识别算法框架** 为处理日常需要实体识别的文本数据，如交易事件双方、金额、种类和公司经营范围的领域、业务等，开发了以 bert_crf 和 bert_crf_bilstm 模型框架的实体识别算法。结合标注平台 docaano 输出的标注数据，开发了将这些数据一键转换成训练集测试集和模型训练评估的框架。根据项目和数据来选择更优模型， 各项目下模型 F1 平均值超过 92% ，最高能达到 97.5%。

• **表格识别和抽取算法**　为解决在 PDF 和图片上抽取表格信息的困难，基于 pdfplumber 和 camelot 两个 PDF 文本抽取算法，通过 CV2 计算表格线条汇聚焦点来构建表格雏形，并使用递归算法和计算表格上下文相似得分开发了一 个具备自动识别跨页，处理跨页(合并或分离) 等功能的提取完整表格算法。该算法支持精准表格提取和全局表格提取，召回率达到 90%以上，精确率达到 96%以上。

• **图片旋转识别和数据增强**　为解决图片旋转而无法很好地使用OCR 识别技术，以 Paddleocr 框架开发了能够识别图片旋转的分类算法，该算法能识别其旋转角度并加以调正。同时，以对抗性训练 GAN 框架方法生成了一批新的图片数据来弥补少量的初始数据。

自然语言处理工程师 | **Apple,** 中国北京　　2019.12 - 2021.4

• **机器翻译模型开发**　先后分别以 transformer、Marian 为核心框架，训练英粤双向(EN-YUE)神经机器翻译(NMT)，训练时应用预训练模型与多 GPU 并行加速。借助柱搜索算法来扩展预测 3 个翻译结果(Bleu: 41.2、39.8、35.6)，并成功部署到 Torchserve 和docker 上。该模型已被采用为 Siri Intel 部门的数据增强及翻译辅助系统，并通过正反馈不断优化模型。

• **Siri 数据各项分类优化**　为提高 Siri 的模型精度和表现，从数据结构、标注层面上来对文本数据进行批量清洗和规范化。同时进行数据隐私保护，利用正则算法来匹配命名实体并使用虚拟生成的数据取代。

• **网络数据爬取**　负责编写豆瓣、简书、微博、知乎等网页内容的提取逻辑和规则。

数据科学实习生 | **Analytics Consult LLC,** 美国波士顿　　2019.04 - 2019.10

• **电子病历文本挖掘**　构建一个机器学习模型能对乳腺癌患者的治疗和诊断结果进行聚类和分类，数据包括首席外科医生对台湾乳腺癌患者的注释。

• 开发模糊匹配算法，从电子病历 Word 文档所包含的表格中提取 90%的关键文本。并创建关系型数据库来存储和处理非结构化文本数据。

• 应用正则表达式来定义提取模式，例如癌症期数、癌症位置以及专业术语。基于这些模式来清洗和保留关键信息。

• 通过数据可视化和使用无监督学习算法（例如 LDA）来探索和识别关键事件，以聚类结果和人工检验来为数据标注。

• 以 GBDT 算法构建分类模型。成功通过使用纵向医疗记录，预测肿瘤的初始阶段和位置。外科医生能够识别癌症分期以及哪些患者可能会发展到最严重的阶段，因此能够及时采取更积极的治疗方案。

项目经验

基于文本挖掘的深度学习模型　　2019.03- 2019.05

• 数据是由 Jigsaw Ai 小组在推特收集的 200 万条评论组成，小组对评论是否具有攻击性进行了概率标记。该项目的目的是基于这些概率建立深度学习模型来预测评论的是否具有攻击性。

• 进行了诸如正则表达式的文本清洗技能，并通过使用了单字符对每个单词进行了字符化。导入词嵌到矩阵，组合矩阵和词并将其转换为数组作为预训练模型。

• 以 Keras 作为框架，通过使用 python 导入预训练模型并定义 LSTM 层，Dropout 层等，开发了双向 LSTM 神经网络。设置 4 个历元和 20%的验证数据集以减少过度拟合。

• 在测试数据集后，神经网络的最终准确性为 90.31%，能相对准确地预测推特中评论是否具有攻击性以及攻击性的程度，以便推特可以使用该模型过滤那些侮辱性评论。

基于决策树和随机森林对药物引起的肝损伤水平进行预测　　2019.04- 2019.06

• 该项目的目标是为 Evidence-Based Toxicology Collaboration(EBTC)预测药物引起的肝损伤类别。

• 从 EBTC 提供的数据集中提取，清洗并合并信息，例如各个细胞测试的参数如细胞类型，药峰浓度等。

• 开发了决策树模型和预修剪方法以减少过度拟合。另外，构建了随机森林模型，应用超参数调整来优化模型。

• 两个模型均达到预期效果。EBTC 能够使用此模型作为参考，并重新考虑从测试中收集的信息的合理性和完整性， 从而可以提高预测的准确性。

荣誉及奖励

作为交换生前往美国阿肯色州立大学交流与深造国际经济与贸易。　　2015.08- 2016.04