姓 名：沙九

个人信息

甘肃甘南 | 中共党员 |1994 年 9 月 |2021 年 7 月硕士毕业| ☎：(+86)18810979033(微信同步)

✉：18810979033@163.com 主要研究方向：自然语言处理 (NLP)、机器学习(ML)

期望岗位：NLP/CV/机器学习/深度学习/软件开发/数据分析/

工作及教育背景

- 北京机械工业自动化研究所 创新产品研发组/后端研发 助理工程师 2021.07-至今
- 北京理工大学 计算机科学与技术专业(平均成绩: 92.64/100.00) 硕士 2018.09-2021.07
- 中央民族大学 计算机科学与技术专业(平均成绩:95.52/100.00) 学士 2014.09-2018.07

IT 技能

- 编程语言: 熟练使用 Python 和 Java,掌握 C,了解 C++、Golang 等。
- **AI 储备：**
  - 掌握 NLP 中机器翻译、文本分类子任务相关概念以及模型；
  - 了解 CV 中的目标检测，分割相关概念以及模型；
- **常用框架：**
  - 熟悉机器学习基本模型并熟练使用scikit-learn、xgboost 等工具；
  - 掌握 Linux 基本语法,能够灵活操作机器学习/深度学习相关流程(特征工程、模型训练以及部署提供服务)。
  - 熟练使用Pytorch 和 tensorflow 深度学习框架，了解 PaddlePaddle 等框架；
  - 熟悉 CNN、LSTM 以及 Transformer 等模型基本结构；
  - 了解 Mysql 和 Oracle 基本语法并结合 spring boot 进行数据库的基本操作；

项目/科研经历

一、 多策略特定领域的智能机器翻译系统 2019.09-2020.05 项目

- **项目描述：**为某军区通信监测研发一套特定领域内多策略融合方式的机器翻译系统。
- **核心难点：**开发英语到汉语自动翻译引擎系统，实现多策略融合的翻译技术，支持篇章和文档翻译；开发翻译引擎系统接口和专业词库接口，实现用户的定制开发。
- **主要工作如下：**
  - 采用开源框架基于 Tensorflow 的 OpenNMT 框架作为基线系统进行训练 NMT;
  - 通过 Back-translation 进行数据增强，通过解码约束方法融入先验知识，从而达到领域迁移的效果；
  - 通过 ElasticSearch 实现站内全文搜索并计算相似度，从而选取最佳翻译引擎(RBMT/NMT)；

- 分别采取基于 Trados 和 Flask—Web 搭建翻译平台提供翻译引擎；
- 针对翻译不充分问题，支持有选择性的二次翻译，并且翻译译文与源端的句子进行对齐高亮显示。

二、 面向法言法语的民族语和外国语机器翻译技术 2018.10-2021.07 项目

- **项目描述：**研究面向法言法语的多语种机器翻译技术与互译便携式设备，通过语言互译解决司法场景的语言障碍难题，推动司法效率大幅提升。
- **所属专向：**“公共安全风险防控与应急技术装备”重点专项。
- **核心难点：**智慧司法智能化认知技术研究，重点在于通过数据增强方法提升低资源翻译模型性能，难点特定领域的数据稀缺，无法将离散的先验知识直接融入到连续的训练模型中。
- **主要工作如下：**
  - 通过半自动数据增强方法构建面向法言法语的稀缺资源多语种平行语料库。
  - 为了提升翻译速度。NMT 翻译模型中引入动态停止机制和段删除机制，此模型将译文拆成多个片段并逐段生成，在每个片段内部采用自回归生成方式，而段间则采用非自回归方式。
  - 在生成译文时，将译文拆分成多个片段，每个片段内自左向右逐词生成，而片段间则并行生成。
  - 为了更好的捕捉目标语言端依赖关系，在生成每个词时，其不仅依赖于所在片段内已经被生成的词，还依赖于其他片段内已经被生成的词；
  - 集成多语种的便携式终端，能够支持 6 个语种 12 个方向的翻译系统。

三、 大数据驱动的汉语与英语及中国少数民族语言之间的机器翻译 2020.03-2021.7 项目

- **项目描述：**研究面向有限标注资源和海量非标注资源的半监督和弱监督机器翻译框架，并设计大数据与先验知识相结合的机器翻译模型。
- **所属项目：**大数据驱动的自然语言理解、问答和翻译 (云计算和大数据)。
- **核心难点：**当前双语数据资源不平衡、单语资源丰富、先验知识难以利用等问题；文本大数据中语言和领域资源的不均衡现象；汉语和蒙藏维等我国少数民族语言之间双语平行资源匮乏而单语资富饶的现象；双语词典和知识图谱等基于符号系统的先验知识难以融入神经机器翻译模型的现象。
- **主要工作如下：**
  - 通过迁移学习和深度神经网络的机器翻译框架，在民语的编码设计模型有选择地共享英语的编码参数。
  - 利用小规模双语数据分别训练源语言到目标语言的翻译模型MT_{s2t}和目标语言到源语言的翻译模型MT_{t2s}。
  - 设计一种词向量学习方法在双语数据和单语数据上联合学习源和目标语言的词向量，采用池化方法获得源和目标语言单语句子的向量表示。
  - 采用基于数据合成的解决方案，给定小规模双语平行语料，学习一个统计机器翻译模型SMT_{s2t}便于融入双语词典等先验知识。

四、 融合大数据与人类常识的开放域多语言知识图谱构建

- **项目描述：**研究大数据驱动的世界/常识知识获取与融合方法，建立大规模、高质量、融合常识知识的多语言知识图谱。

- **所属项目：**大数据驱动的自然语言理解、问答和翻译（云计算合大数据）。
- **核心难点：**具有普适性的知识表示体系和结构模型，以支撑大数据环境下的知识的表示和计算；大数据环境下，知识多源、异构、低质的特点给知识获取带来巨大挑战；传统的常识知识获取往往基于逻辑表达的文本语义分析，缺乏语言知识的约束和指导，难以适应知识库大规模、多领域的特点。
- **主要工作如下：**
  - 对于实体指称项词典，利用网络百科文本中的链接信息自动获取实体的指称信息；
  - 对于关系模板，利用知识库中语义关系的参数对文本进行自动回标获得其上下文信息，并利用深度神经网络从中获取特征对关系模板进行语义表示；
  - 融合文本信息的跨语言知识表示学习和语义相关度计算方法，应用于跨语言实例的相关性度量。
  - 研究语言知识单元（字、词、短语等）与世界知识单元（实体、事件、关系）之间的统一语义表示：将语言知识和世界知识的表示学习集成在统一框架中；
  - 为大数据驱动的自然语言理解、问答和翻译提供基础知识资源。

实习/工作经历

一、 **公司名称：北京自动化研究所** 2021.08-至今 工作

- **职位名称：**技术研发工程师
- **工作描述：**支撑工业应用微服务系统实现基于 ElasticSearch 的在线更新机器翻译方法。
- **工作难点：**当前参数众多结构复杂的神经网络模型难以做到在线更新；基于样本的 NMT 系统泛化性较差；稀缺工业领域内的数据。
- **主要工作内容：**
  - 使用一种动态结合样本检索和神经机器翻译的方法；在检索到相似样本的情况下能够提升翻译效果，在检索不到相似样本时，也能保持原有的翻译质量，同时保持在线更新的能力。
  - 采用两个翻译引擎，分别为基于 Transformer 的通用领域神经机器翻译模型和样本检索模块，用于执行相似样本检索、相似度计算和概率估计。
  - 通过键值对构建离线数据库，其中键为目标端语言的句子中一个词出现的上下文的向量表示；值为对应的目标端语言的词。
  - 解码时结合可学习的核函数采用样本检索方法，利用核密度估计根据检索到的样本估计出一个基于样本的分布。
  - 通过自适应分布混合方式，将模型分布和基于样本分布按一定权重进行线性插值，得到混合分布并由混合分布预测出下一个词。

二、 **公司名称：华为技术有限公司** 2020.06-2020.12 实习

- **职位名称：**AI 算法研究实习生。
- **工作描述：**通过 web 入侵检测技术，研究注入类的 web 攻击，自动化恶意流量家族提取。基于异常检测的 web

入侵识别，训练阶段通常需要针对每个 url，基于大量正常样本，抽象出能够描述样本集的统计学或机器学习模型(Profile)。检测阶段，通过判断 web 访问是否与 Profile 相符，来识别异常。

- **工作难点：**标签数据的缺乏，web 入侵样本稀少，且变化多样，对模型的学习和训练造成困难。
- **主要工作如下：**
  - 首先，将异常访问从日志中剥离，标记为异常流量；然后，后期目标是对异常流量进行攻击分类统计；最后，从攻击中溯源，检测出是否被成功入侵等；
  - 分别尝试了基于统计学习模型、基于文本分析的机器学习模型、基于单分类模型和基于聚类模型进行 Profile 的建立；
  - 通过使用基于文本序列模型(HMM 的状态序列)对参数进行了序列化建模，通过优化模型，使得接收参数的多个形式，减少模型数，提高效率。

三、 **公司名称：北京智源人工智能研究院** 2020.01-2020.05 实习

- **职位名称：** NLP 算法研究实习生。
- **工作描述：**基于用户个性化信息来生成评论的方法研究，构建从多维度获取特征并能够自动生成相应评论。
- **工作难点：**非结构化数据转为结构化数据；数据分布不均匀且为离散；用户自定义特征的 embedding 为静态影响 decode 生成回复的语法性。
- **主要工作如下：**
  - 基于 seq2seq 模型+attention 机制，encoder 和 decoder 都采用 lstm；
  - 应用了一种基于门记忆的特征 embedding; 将用户的个性化特征属性经过一个全连接层，得到向量表示，表明用户的个人特征；设计了一个 gated memory 来动态表达用户的个人特征；
  - 基于博客和用户个人描述的联合 attention 计算；
  - 把用户各种个性化信息，属性知识进行不同的编码处理，并用 co-attention 联合在一起的处理；
  - 通过已提供的学者信息特征及外部嵌入描述，构建一套能够自动生成对应学者研究爱好的系统。

竞赛/获奖/论文情况

竞赛

- 十一届中国计算机博弈锦标赛冠军;
- CWMT 2018 藏汉翻译第二名；
- CCMT 2019 藏汉翻译第三名。

获奖

- “第十三届中国大学生年度人物”候选；
- 北京市三好学生；宝钢优秀学生奖；国奖；专一；
- 优秀学生干部等，北京市优秀毕业生，优秀硕士生。

- **论文**

- 2022 年 ACM《Integrating Pre-training model into NMT with Bi-Directional Feature Transformation》。
- 2021 年 普通发明专利《一种利用半自回归融合领域术语的低资源机器翻译方法》。
- 2020 年 CCL 《面向司法领域的高质量开源藏汉平行语料库构建》并获最佳论文奖。
- 2020 年 《Revisiting Back-Translation for Low-Resource Machine Translation Between Chinese and Vietnamese》 发表于 IEEE Access。
- 2019 年 《多策略切分粒度的藏汉双向神经机器翻译研究》 被 CCMT2019 会议录用并发表在《厦门大学学报(自然科学版)》。
- 2016 年 《藏文自动分词与词性标注研究》 。

自我评价

- 本人性格开朗，为人诚恳，乐于沟通，有团队精神。喜欢弹唱、足球。热爱钻研，心态良好，能积极面对工作生活中的困难。始终秉承“学以致用”的准则，学习能力强。希望能够在实践中持续学习，发挥自己的主动性、创造性，为公司的发展竭尽全力。

个人站点

博客	**Github**	**微信公众号**