张镭镧

github.com/zll17 · zhangleilan@gmail.com · (+86) 188-1139-3706

教育经历

清华大学 计算机科学与技术 硕士 2017 – 2020

- GPA: 3.41 / 4.0
- 研究机构：清华大学语音语言技术实验室 (CSLT)
- 研究领域：自然语言处理，文本聚类，主题建模，深度学习
- 修读课程：机器学习，优化方法，计算语言学，泛函分析，计算语言学，算法分析与设计

四川大学 数学基地班 学士 2012 – 2016

- GPA: 3.2 / 4.0
- 四川大学单项一等奖学金（2013-2014）
- 全国大学生数学竞赛专业组二等奖（2014）
- 修读课程：数学分析，概率论，统计学，实分析，复分析，抽象代数，微分几何，偏微分方程，数值分析，C++ 程序设计，数据结构，计算机组成，操作系统，计算机网络，数据库基础

专业技能

- 编程语言：C++，Python，Java，x64 汇编，Mathematica
- 机器学习框架：PyTorch，Sklearn，Pandas，Keras，TensorFlow，Gensim
- 通用工具：Linux，Git，LaTeX，SQL，Electron
- 数学基础 (研究生入学考试-数学成绩: 147/150)
- 英语：CET-6（517），能流利阅读和写作英文论文。
- 其他：2013 年四川大学自行车环校赛 男子组 第二名；2014 年大九湖全国大学生自行车环湖赛 男子组 第 15 名。

发表成果

- **Leilan Zhang**, Qiang Zhou. *Automatically Annotate TV Series Subtitles for Dialogue Corpus Construction*. 2019 Asia-Pacific Signal and Information Processing Association Annual Summit and Conference (APSIPA ASC), Lanzhou, China, 2019, pp. 1029-1035.
- **Leilan Zhang**, Qiang Zhou. *Topic Segmentation for Dialogue Stream*. 2019 Asia-Pacific Signal and Information Processing Association Annual Summit and Conference (APSIPA ASC), Lanzhou, China, 2019, pp. 1036-1043.
- 周强, **张镭镧**. 基于剧本的字幕场景和说话人信息自动标注方法和系统. 中国专利: 201811633842.5, 2019.07.09.
- 周强，**张镭镧**. 字幕对话流的主题分割方法及装置. 中国专利：201910906359.8，2020.02.21.
- 硕士论文：基于神经主题模型的字幕对话文本聚类研究.

项目经历

- **神经主题模型开源库** [*Github*] *Mar 2020*
  - 基于 PyTorch 实现的神经主题模型库，以 VAE 和 WAE 为框架实现了一系列神经主题模型，包括 NVDM-GSM、W-LDA、WTM-GMM-std、WTM-GMM-ctm、ETM、GMNTM 和 BATM，提供了完整的训练脚本和包含 6 个指标的性能评测模块，接口统一做到开箱即用。在短文本上的评测结果显著优于经典主题模型 LDA，模型数量在进一步扩充中。独立开发，目前已开源，在 Github 上收到 103 个 Star 和 25 个 Fork。
  - 针对短文本语料主题集中的特点，提出了两个改进的神经主题模型 GMNTM 和 WTM-GMM，在隐空间分别采用自适应高斯混合分布和标准高斯混合分布作为先验分布，相较于基线模型 W-LDA 的一致性和多样性有明显提高。
  - 技术报告：基于 VAE 的 Neural Topic Mode 研究进展概述.

- **基于对话的音乐推荐系统** [*Doc*] [*Demo*] *May 2019*
  - 基于从网易云音乐中爬取的 7000 余首歌曲的歌词，使用主题模型对其进行聚类分析，构造了以对话形式对用户进行音乐推荐的系统，该系统由意图检测模块、主动推荐模块、被动检索模块以及开放闲聊模块组成，由两个成员协作完成。
  - 音乐意图检测模块基于 TextCNN 实现，用以评估用户查询中的音乐意图关联度，为解决数据匮乏问题，该模块训练数据由若干种子用户数据经模板化后采用增强手段得到，检测准确率达到 97%。
  - 被动检索模块用以回应意图明确的音乐查询，在多轮对话中利用 LSTM+CRF 进行槽位填充，并基于 Levenshtein 距离检索复合要求的音乐。
  - 主动推荐模块用以回应意图模糊的音乐查询，以数据集中歌曲及用户评论的对应关系构建图，采用 node2vec 对歌曲和用户结点进行表征，以集体用户均值作为当前用户的冷启动向量，基于协同过滤算法进行推荐，并根据用户的反馈更新当前用户的音乐偏好。
  - 开放闲聊模块为基于 Bi-GRU 的 Seq2Seq 架构，并实现了 attention 机制，用以对不包含音乐意图的开放域查询进行回复。

- **字幕说话人自动标注系统** [*Data*] *Nov 2018*
  - 以 4 部美剧的 779 集字幕和剧本为基础，设计实现了为字幕自动标注说话人的程序，开放了包含 260674 个话语消息的中英多轮对话语料。
  - 实现解析器从无结构剧本中提取元素，并基于 BM25 算法评测话语消息之间的相似度，建立字幕与对应剧本中的话语的初步对齐。
  - 实现基于时序卷积网络 TCN 的错误检测模型，并通过统计真实数据分布特征生成模拟数据以构造训练数据集，采用启发式算法进行后处理，有效检测和纠正初步对齐中的错误，说话人标注准确率约为 94%，比此前同类算法提升了约 12%。

- **基于 GNN 的高能粒子轨迹重构** [*Demo*] *Mar 2021*
  - 使用 TrackML 数据集，通过传感器的撞击信号重构粒子运动轨迹。以 r 邻域对撞击信号的隐空间嵌入构造图，基于 hinge loss 进行度量学习，使得相同轨道的粒子彼此趋近，不同轨道粒子彼此疏离，并通过 MLP 过滤简单冗余边，减小计算复杂度（acc:35%@recall:99%）。
  - 基于 AGNN 对连接图进行建模，通过 flow 模型 MaskAF 对 AGNN 隐空间进行规整化，并基于 MLP 训练分类器判断是否对撞击点对的边进行裁剪形成 doublet。
  - 以 doublet 为结点，对含有公共点的 doublet 进行连接构造图，再次基于 AGNN 对图进行建模，并训练分类器判断是否构成三邻接点 triplet，通过 triplet 重构整个粒子运动轨迹，达到 acc:82%@recall:98%，较于此前基于几何的启发式重构法 (acc:5%@recall:45%) 有显著的提升。

学术活动

- **志愿工作**

  亚太信号与信息处理会议 志愿者 (APSIPA 2019 ASC) *兰州, Nov 2019*

  中国计算语言学会议 志愿者 (CCL 2018) *长沙, Oct 2018*

- **实习经历**

  多益网络 AI 研究院，Chatbot 组，实习生 *广州, May 2016*

  IRIS-HEP, EXA.TRKX Collaboration, Remote Intern CERN, Mar 2021