陆俊如

兴趣方向：自然语言处理与机器学习

junru.lu@warwick.ac.uk Coventry, UK github.com/LuJunru lujunru.github.io

教育经历

MPhil/Ph.D. candidate in Computer Science

华威大学

2019 年 11 月 – 2023 年 11 月　　考文垂, 英国

- **导师**: **何瑜岚教授**
- **学院**: Computer Science
- **研究兴趣**: 文本表示、机器问答和文本生成。目前我的工作主要围绕真实世界场景下的机器问答。

M.S. in Applied Urban Science & Informatics

纽约大学

2018 年 9 月 – 2019 年 9 月　　纽约市, 美国

- GPA: 3.67/4.00
- **研究中心**: Center for Urban Science + Progress
- **部门**: Tandon School of Engineering
- **项目** 1: 使用 Twitter API 收集纽约市内带有签到信息的推文，并通过情感分析创建情绪地图 [**代码**]。
- **项目** 2: 基于随机森林和 LGBM 模型，使用 Yelp 评论和相关用户、商户历史信息来预测评论打分 [**代码**]。
- **毕业设计**: 使用双层差分模型和贝叶斯网络来推断 Uber/Lyft 对纽约市停车罚单数量的影响 [**代码**]。

工学学士，信息管理与信息系统专业

国际关系学院

2014 年 9 月 – 2018 年 7 月　　北京, 中国

- GPA: 90.6/100
- **导师**: **李斌阳教授**
- **学院**: 信息科技学院
- **毕业论文**: 一种融合多类注意力机制的两阶段机器阅读理解模型 (A-Reader)。在 A-Reader 中，文本表示通过自注意力编码实现，而文章和问题之间的语义交互则由自注意力和双向注意力共同完成。这里的“两阶段”指的是先用最终的语义交互矩阵训练一个二分类模型来筛选最佳段落，再通过语义交互矩阵和指针网络生成答案。

经济学双学位

北京大学

2015 年 9 月 – 2018 年 7 月　　北京, 中国

- GPA: 84.2/100
- **学院**: 国家发展研究院
- **主要课程**: 财务会计、计量经济学、微观经济学和金融学
- **备注**: 该**双学位**是北京大学专为非经济学本科的北大和外校学生设置的项目。

奖项荣誉

SMP 全国第六名

- 第六届全国社会媒体信息处理大赛 – CSDN 用户画像挖掘评测全国第六名。该比赛要求完成三个子任务：博文关键词提取、用户兴趣标注和用户黏性预测。
- 对于任务一，我们使用 Tf-Idf、Textrank、LDA 和人工规则来抽取关键词。而任务二使用了基于文档向量的 Stacking 模型（后期在实验室中改用 TextCNN 和 self-attention 得到了 3% 的提升）。任务三采用 Stacking 回归模型实现 [**代码**]。

奖学金

2015、2016 和 2017 年度国际关系学院校级奖学金（前 5%）

本科荣誉

2018 年度国际关系学院优秀毕业论文暨优秀毕业生

Enactus 创行世界杯

2018 年 Enactus 创行世界杯中国区决赛三等奖、华北区域赛一等奖。Enactus 是一家国际非营利非政府组织，致力于通过大学生力量为第三方受众设计可持续盈利的商业模式以改善受众生计。

Like a Boss 产品设计大赛

获全球总冠军，奖金 $30,000。这项赛事由 Boss 直聘 APP 发起，旨在邀请全球留学生帮助该 APP 进行针对留学生市场的产品优化设计。[**官网**]

相关技能

自然语言处理　机器学习　网络爬虫

Pytorch　Pandas　SkitLearn

Numpy　PySpark

编程语言

Python
Sql
Javascript

工作经历

自然语言处理实习工程师

北京深思考人工智能

2018 年 3 月 – 2018 年 6 月　　北京, 中国

- **项目** 1: 针对句子相似度判断模型 TextCNN 和 Siamese-LSTM 的研究。这里的 TextCNN 主要是改进其输入：将句子 1 和 2 切割成若干子单位，两两计算子单位之间的相似度，最终填充成特征矩阵。而 Siamese-LSTM 则是：先用同一个 LSTM 训练句子表示，然后直接计算句子 1 和 2 在 LSTM 最后一层隐层向量间的曼哈顿距离 [**代码**]。
- **项目** 2: 在 Dureader 数据集上实现一个两阶段的 BiDAF 模型。BiDAF 是一种使用双向 LSTM 和双向注意力分别实现文本表示和语义交互的传统机器阅读理解模型。这里的" 两阶段" 指的是使用人工指定的特征选取最佳段，再通过指针网络生成答案。

数据挖掘实习工程师，文本挖掘部

北京百分点信息科技

2017 年 10 月 – 2018 年 3 月　　北京，中国

- **项目**: 研发一个基于社区问答的单轮中文问答系统。该系统主要包括构建一个（问题，答案）形式的知识库、基于 Elasticsearch 的初选模块、考察语义的精选模块：针对（新问题，老问题）的句子相似度判断模型和针对（新问题，候选答案）的问答质量评估模型，以及空回答时启用的在线搜索模块 [**代码**]。

学术论文

期刊论文

- Lu, Junru et al. (2019). "Identifying User Profile by Incorporating Self-Attention Mechanism based on CSDN Data Set". In: Data Intelligence 1.2, pp. 160–175.

会议论文

- 【Accepted by COLING 2020】Lu, Junru et al. (2020). "Chime: Cross-passage Hierarchical Memory Network for Generative Review Question Answering". In: The 28th International Conference on Computational Linguistics.
- Cai, Junjie et al. (2020). "Estimating the effect of Uber & Lyft on parking violation in NYC". in: Transportation Research Board Annual Meeting.