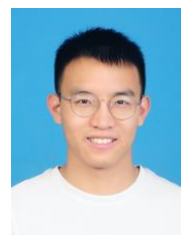


姓名: 戴剑波 出生年月: 1997.08
学历: 硕士 毕业院校: 爱丁堡大学
住址: 浙江金华 电话: 13486963560
邮箱: 503478374@qq.com
求职意向: 寻找自然语言算法或机器学习算法研发全职岗位



教育背景

爱丁堡大学 人工智能硕士 2019.09-2020.10

GPA: 3.96/4.00 (70.9/100), 排名约前 15%, 一等学位 (Distinction)。

相关课程: 机器学习与模式识别, 加速自然语言处理, 深度学习, 概率图模型与概率推断, 自然语言理解、生成与机器翻译, 博弈论算法及应用, 编程技巧。

大连理工大学; 曼彻斯特大学 化学工程本科双学位 2015.09-2019.06

GPA: 3.83/4.00 (88.3/100), 排名约前 3%; GPA: 3.82/4.00 (75.9/100), 排名约前 7%, 一等学位。

相关课程: C 语言程序设计, 数据结构与算法, 大学计算机基础, 计算机网络原理, 操作系统, 计算机组成原理, 高等数学, 概率论与统计, 线性代数, 微分方程数值方法, 离散数学, 电工与电子技术。

获奖情况: 全国大学生数学竞赛市一等奖; 大连理工大学优秀毕业生, 中国大学生 Chem-E-Car 竞赛全国第 8。

个人技能

编程技能: 熟悉 Python, Shell, C, LaTeX, 有 SQL, C++, Qt 编程经验
框架工具: 熟悉 Pytorch, Numpy, Docker, Matplotlib, OpenNMT, Pandas, SciPy
自然语言: 中文(母语), 英语(无障碍商务沟通)

项目经历

语义分析: 不完整用户问题建议 (2020.06-2020.11) 暑期研究, 耶鲁大学语言、信息与学习实验室

1. 基于 Spider 等自然语言与 SQL 的数据库, 设计不同类别的不完整问题并建立它们与相应完整问题, SQL 对应的数据集。定义了基于特定数据库的根据用户输入推荐完整问题与 SQL 并通过最小的用户反馈保证正确性的新任务。
2. 提出一种基于 Constituency Parsing 语法树的对不完整问题与完整问题的语义编辑距离并用该距离为标准自动化生成一系列的与一个不完整问题对应的完整问题, 从而将数据量从 4 千扩增到 3 万。
3. 提出一种基于 Albert 的基准模型, 训练得到不完整问题与完整问题的匹配度, 计算召回率 (top10 召回率约为 46%)。
4. 论文 2021 年 NAACL-HLT 在投。

语境化的神经词形还原 (2020.05-2020.08) 硕士毕业设计, 爱丁堡大学

1. 在 9 种语言上 11 个语料库上分别训练了一个叫 Lematus 的模型作为基准模型和一个叫 COPYNET 的用复制机制增强的 seq2seq 模型。以准确率与莱文斯坦距离为指标评估比较了 Lematus 与 COPYNET 在不同语言上的表现。
2. 发现由复制机制增强后的模型相对基准模型在词形还原任务上的表现与语言的形态学特征无关。该机制能在部分语言上发挥作用, 尤其是其将日语的词例准确率提高了 5.5% (达到 97.8%), 词型准确率提高了 13.3% (达到 95.3%)。
3. 复制机制能发挥作用的原因主要是因为提高了在未曾见过的数据上的表现, 该机制未被发现能在分歧数据上发挥作用。该机制对词语长度较敏感, 数据中 copied 词语比例和语言中的字符数量也能大大影响提升模型效果 (约 15%)。

神经机器翻译 (2020.02-2020.03) 自然语言理解, 生成与机器翻译课程项目, 爱丁堡大学

1. 训练基于双向 LSTM 的编码-解码结构的基准翻译模型来实现英语到德语的翻译, 实现束搜索并改变网络结构来增强模型并比较在不同数据集上表现。该基准模型在验证集上的困惑度为 26.8, 在测试集上的 BLEU 分数为 11.22。
2. 在 Transformer 框架上继续实现 lexical module 和多头注意力机制, 输出注意力的热图来研究他们对翻译效果的影响, 该改进后的模型在验证集上达到了 23.5 的困惑度并在测试集上取得了 9.78 的 BLEU 分数。

使用推特数据探索分布性假设及单词的相似度 (2019.11) 自然语言处理课程项目, 爱丁堡大学

1. 实现多种矢量语义模型来表示单词相似度, 包括 PPMI 及带 Laplace 平滑的变体等。将结果排序并用 Spearman's Footrule 距离来量化并和其他词相似度数据集中做比较。带平滑的 PPMI 方法有最小的相对排序误差 (约 15.35/80)。
2. 实现不同的单词相似度测量方法如余弦相似, Jaccard 相似度和 Dice 相似度并用皮尔森相关系数研究其词频对这些方法的影响, 其中带平滑的 PPMI 方法最不容易被词频影响 (系数为 -0.02)。

