DIAN CHEN (陈点)

+86 18883383965 | ✉ okcd00@qq.com
Homepage: okcd00.tech/about
LinkedIn: linkedin.com/in/okcd00
Github: github.com/okcd00

教育经历

博士研究生在读 | 计算机科学与技术 Sep. 2016 – Jun 2023

中国科学院 · 计算技术研究所 (保研推免直博) 中国科学院大学

- 课题组：中国科学院智能信息处理重点实验室 机器学习与数据挖掘课题组
- 研究领域：自然语言理解 & 自然语言处理，包括信息抽取，文本校对，文本预训练等
  Chinese Spelling Error Correction (CSC, 中文拼写校正)
  Natural Language Interface for Databases (NLI, 数据库自然语言接口)
  Nested Relation Extraction (NRE, 嵌套关联提取)
  Text Information Extraction (IE, 文本信息抽取)

工学学士 | 物联网工程 Sep. 2012 – Jul 2016

计算机学院 · GPA3.60 (专业第二) 重庆大学

- 获得中国计算机学会 (CCF) 2016年度优秀大学生的荣誉，受邀于 CNCC2016 接受颁奖。
- 重庆大学 ACM 集训队初创队员，在亚洲区域赛中获1银2铜；在省级竞赛中，获1金2银。
- 优秀学生、优秀毕业生、校级优秀毕设、国家奖学金、长江电力奖学金、多次获学业奖学金甲等。

项目经历

金融文档全面复核系统 | **AutoDoc** 2016 – Now

Theano, Tensorflow, Pytorch, Pytorch-Lightning / Regex, etc. 庖丁科技

- 该项目在金融文档中提取不同种类的信息，从多角度做正确性复核，并协助用户完善文档内容。
- 负责其中关于数值错误、用字错误、主客体错误、原因披露等任务的模型及标注系统的设计与迭代。
- 该项目已为超过40家金融机构部署。与港交所合作于 Regulation Asia 获得中国首次 **Outstanding Award**。

智能数据库检索系统 | **Text2SQL** 2019 – 2022

Pytorch / PostgreSQL, MySQL 庖丁科技

- 该项目用于通过自然语言的方式调用数据库，系统根据询问生成可执行的SQL，返回所需信息给用户。
- 带领一个 3 名开发和 2 名前端同事的小团队，完成了该项目的完整架构设计与原型实现，包括知识库的设计。
- 于深交所技术大会获得 2019 年度研究课题二等奖，智能数据检索系统也获得第七届证券期货科学技术优秀奖。
- 智能语音数据库查询项目于中金所中标(竞品来自包括BAT的多个大厂)，现第一期已成功交付。

智能金融刷报系统 | **Glazer** 2018 – 2020

Tensorflow, Pytorch / 庖丁科技

- 协助金融从业人员完成日常刷报工作，即使用既有研报和新数据源，自动完成新的研报中绝大部分的工作。
- 进行了需求分析和调研后，提出这个市场需求的产品设计。后续在该项目的实现中，也负责了命名实体识别模型与嵌套因果模型的实现与迭代。
- 该项目目前已为中信证券在内的多家金融机构及券商服务，客户反馈刷报工作的时间节约平均超过90%。

数据爬取与分析 | **CDSpider** 2015 – 2017

Sklearn / PySpider, BeautifulSoup4, Baidu-Map-Reduce 百度研究院 大数据实验室 (BDL)

- 主职工作为设计无人值守的增量式数据爬取方法。实习期间实现了两套方案，设计了自维护的代理池，成功为实验室的金融研报分析任务获取到数十万篇的完整研报信息，也参与了研报情感分析研究。
- 参与了百度与大悦城合作的优惠券推荐项目，通过跨模态数据分析用户画像。经过 AB-Test 验证了提出的方法使得用户购买金额较之传统LDA方法平均提升12.5%，研究形成论文发表于会议 KDD2016。

个人技能

Deep Learning Framework: Pytorch, Pytorch-Lightning, Tensorflow 1.x
Languages: 英语(CET-6)，普通话等级考试(二级甲等)
Open-Source Experience:

- **BBCM** 209 ☆ (在最大的中文错字校正开源社区 **PyCorrector** 中，被认为是效果最好的开源CSC 模型)。
- **Graftr** 90 ☆ (模型参数的可交互文件树操作命令行工具，是微调实验和多阶段预训练的实用工具)；
- **CDSelector** 17 ☆(国科大选课脚本，多年来作为选课脚本示例，在CSDN中的介绍文章有超过22.4k的阅读量)。

Programming Languages: Python (主要语言), C++ (本科ACM竞赛基础)
Programming Environment: VSCode, PyCharm, DevC++, JupyterLab, Zsh-Vim
Shell Environment: Termius, SecureCRT, Git Bash, Terminals

论文成果

Nested Relation Extraction with Iterative Neural Network CCF-B 会议+ CCF-B 期刊

Yixuan Cao, **Dian Chen**, Hongwei Li, Ping Luo. CIKM 2019: 1001-1010

Yixuan Cao, **Dian Chen**, ···, Ping Luo. Frontiers Comput. Sci. 15(3): 153323 (2021)

- 我们率先提出了用于嵌套关系提取的迭代式通用网络，
- 以此网络设计为基础，构建了文本信息提取一体式框架UTIE，
- 公司现有的金融实体、事件等信息抽取都基于此框架，为诸多项目提供保障。

Nested Causality Mining on Financial Statements CCF-C 会议

Dian Chen, Yixuan Cao, Ping Luo. NLPCC 2020: 725-737

- 提出了等效嵌套因果数据结构，用于文本中提取嵌套因果的方法，
- 该方法实际落地于智能金融刷报系统 Glazer，用于展示需要修改的原因位置，
- 该项目目前已运用于多家金融机构及券商，客户反馈刷报工作的时间节约超过90%。

The Contexts Deserve: Explicit Modeling the Context for Chinese NER 核心期刊

Dian Chen, Yixuan Cao, Ping Luo. High Technology Letters. 2024(7).

- 我们提出，当NER模型仅凭字面难以判断时，可以依靠实体上下文中的信息做出判断，
- 这一方法在公开数据集中，可令模型对未知新实体的识别错误明显下降(9.9%)，
- 实际运用于行文规范的金融语料中，在涉及多领域的金融命名实体上提升明显(>5%)。

Span-based Chinese NER with Span Filtering EI检索期刊

Dian Chen, Yixuan Cao, Qingping Yang, Ping Luo. High Technology Letters (English). 2023: Accepted.

- 引入基于跨度的提取方法，缓解中文虚词省略带来的实体边界识别问题。
- 提出基于打分模型的采样策略，训练中有效减少计算(>50%)的同时提升模型泛化能力。

Towards Natural Language Interfaces to Databases in (已投递至 KDD)

Professional Applications: Reliability Prior to Generalizability.

Yixuan Cao, Chaoxu Pang, **Dian Chen**, ···, Juyao Liu, Ping Luo. *KDD 2023: Pending.

- 我们提出了基于自设计知识库、模板对匹配、建模询问改写的 Text2SQL 解决方案，
- 与海通证券合作，于深交所技术大会获得2019年度研究课题二等奖，
- 与海通证券合作，智能数据检索系统获第七届证券期货科学技术优秀奖，
- 智能语音数据库查询项目于中金所中标(竞品来自包括BAT的多个大厂)，已交付。