李大为

Data & Applied Scientist • AI/Data/Back-End

(+86) 156-5225-0910 | dwlee@pku.edu.cn | daviddwlee84 | Da-Wei Li

Summary

我目前在 Microsoft 担任 Data & Applied Scientist。包括在 MSRA 与 STCA 的实习经历已经在微软工作超过三年。自诩为一个 **maker** ——也就是将任何创意的点子加以实践者，从动手中学习是我的哲学。我也是一个 Vim 的重度使用者，喜欢探索新科技、新技术，以及解决问题与挑战带来的成就感。较熟悉于以下这些领域：**AI 全栈开发、自然语言处里与深度学习、推荐系统、大数据处里与分析、后端开发**。

Experiences

微软互联网工程院 **Bing Multimedia** 组

Suzhou, Jiangsu, China

Data & Applied Scientist — *Jul. 2021 - Present*

- 为 Bing、Edge、MSN 提供更好的视频推荐质量以提升 DAU 与用户的 CTR。
- 在多个场景包含 MSN Article、Bing Super Caption、Edge Underside 中上线了 **Relevance Model** 来控制推荐视频之相关性以便提升更大的 coverage 同时保证推荐质量，整体降低了人工标注的 **defect rate -2.2%**。
- 在 MSN Article 的 37 个 market 中上线了 **CTR Model** 带来 **Video-CI/UU +2.4%** 与 **Video-CI Ratio +2.33%** 的提升。
- 为 Bing VDP 提供基于 co-info 的 **CF Recall**，带来 **Traffic Coverage +23.0%** 与 **CTR/UU +1.33 %** 的提升。
- 为提升开发效率搭建线上 **Related Video Debug Tool** 与可复用的通用部件例如数据预处理模块、自动化标注流程、流程脚本视觉化等。

微软互联网工程院 **Bing NLP** 组

Beijing, China

Algorithm Intern — *Jul. 2020 - Jun. 2021*

- 搭建 **Numeric Information Extraction System** 用于金融研报的数值信息提取. 针对任务需求设计 annotation guideline，训练 MRC-based Sequence Labeling 模型，与搭建基于 docker 的后端服务。
- 搭建 **AI Writer** 的 NLP 模型用来自动生成文章与改写文章。

微软亚洲研究院 **Knowledge Computing** 组

Beijing, China

Research Intern — *Dec. 2019 - May. 2020*

- 在 AAAI 2021 上发表关于 "**学术文章的简报自动生成**" 的论文，并受邀在微软官方 Bilibili 上进行直播分享。

Publications

Towards Topic-Aware Slide Generation For Academic Papers With Unsupervised Mutual Learning

AAAI 2021 (CCF A)

Da-Wei Li, Danqing Huang, Tingting Ma, Chin-Yew Lin — *May. 2021*

- 对学术论文自动生成简报，方法结合了 extractive summarization 以及 unsupervised mutual learning。
- 在 mutual learning 的框架中利用两个 extractors （一个 Log-Linear Classifier 与一个 Neural Sentence Selection Model）来互相学习，藉由两种模型各自的优势在 unsupervised 的限制之下学出更好的结果。

Open Relation Extraction with Non-Existent and Multi-Span Relationships

KR 2022 (CCF B)

Huifan Yang, **Da-Wei Li**, Zekun Li, Donglin Yang, Bin Wu — *Feb. 2022*

- 设计 Query-based Multi-head Open Relation Extractor (QuORE) 来抽取 single/multi-span relations 并且判断 non-existent relationships。
- QuORE 是一个 multi-head 的框架其中由两个 sub-modules 构成，SSE (Single-Span Extraction) 与 QASL (Query-based Sequence Labeling)，他们在训练时共享 loss 并在预测时有一个 selector 来选择应该用哪个模型的输出作为最终结果。

Open Relation Extraction via Query-Based Span Prediction

KSEM 2022 (CCF C)

Huifan Yang, **Da-Wei Li**, Zekun Li, Donglin Yang, Jinsheng Qi, Bin Wu — *May. 2022*

- 设计 Query-based Open Relation Extractor (QORE)，是一个多语言的 Transformers-based language model 用来抽取 arguments 与 context 中所包含的关系。
- 其中实验包括 Multilingual ORE、在 non-query/query-based and non-LM/LM models 上的 ablation study、zero-shot domain transferability 以及 few-shot learning 能力。

Education

北京大学

Beijing, China

软件工程硕士 — *Sep. 2018 - Jul. 2021*

- 主要关注于 NLP 与 Knowledge Graph 相关的应用。参与北大开源协会，是 ML/NLP 组的核心成员。
- 毕业论文为《面向中文文本的数值抽取与理解方法设计与实现》，内容是设计一套符合中文特性的标注方法来训练抽取模型，并基于此模型搭建数值抽取的应用。

台湾科技大学

Taipei, Taiwan

电子工程学士 (辅系财金) — *Sep. 2013 - Jun. 2017*

- 主要关注于嵌入式系统设计及其他工程类项目如 App, Web 等。
- 毕业设计为《基于模组化架构之四轴飞行器设计及其于影像辨识之应用》，从零构建一台无人机，设计与调整基于 PID 的平衡算法，并结合机器视觉做物件追踪的应用展示。
- 拥有 4 次个人接案经验，其中较有趣的案子为利用 Android-based 的 AR 眼镜来远程控制视频云台再将影像直播串流到眼镜中，其中涉及云台电路设计与模型 3D 打印。
- 参加 5 次不同工程领域的竞赛，包含设计 2048 游戏 AI、App 设计竞赛、LED 设计竞赛与 MCU 设计竞赛等。