李大为

Master Student · Natural Language Processing

(+86) 156-5225-0910 | dwlee@pku.edu.cn | dwlee-personal-website.netlify.com | daviddwlee84

Summary

目前就读于北京大学软件工程。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。在本科时熟悉嵌入式系统设计，也很喜欢做一些课馀有趣的项目。目前在研究所，接触较多的是 Knowledge Graph 和 Extractive Summarization 相关领域的 NLP Deep Learning 研究与应用。

Education

北京大学

软件工程硕士 *Sep. 2018 - Present*

- 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。

台湾科技大学

电子工程学士 *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。
- 拥有 4 次个人接案经验与参加 5 次不同工程领域的竞赛。

Experience

微软亚洲研究院 Knowledge Computing 组

Research Intern *Dec. 2019 - Present*

- 学术文章之简报/投影片自动生成。

北京大学软件工程国家工程研究中心

Research Intern *Jul. 2019 - Present*

- 参与医保反欺诈与医疗记录分析与抽取。

台湾工业技术研究院资通所嵌入式系统软体组

Intern *Jul. 2016 - Aug. 2016*

- 在组里参与自动驾驶之项目，主要负责将组里设计之算法使用在 STV0991 开发板上。

ML/NLP Project

GCAKE: Graph and Context Attentional Knowledge Embedding

Side Project *Oct. 2019 - Jan. 2020*

- 是一个利用 self-attention 的 knowledge graph representation learning 架构。使用不仅限于三元组的信息，同时考虑了上下文与图结构之关系。

Similar Cases Recommendation via Legal Knowledge Graph Construction and Representation

Side Project *Aug. 2019 - Oct. 2019*

- 提出了一个将法律案件利用 knowledge graph embedding 进行表示并利用其来推荐之 pipeline。
- 其中包含子任务有 Named-entity Recognition、Relation Extraction、Knowledge Graph Embedding。主要模型是使用 jointly-trained multitask 的 fine-tuned BERT。

Sentence Similarity in Intelligent Question Answering

Side Project *Aug. 2019 - Sep. 2019*

- 主要基于 Enhanced RCNN model (BiLSTM + Attention + CNN) 来优化句子相似度的计算。
- 背景设立于智能问答之应用，以同时考量 performance 与 complexity 之间的权衡为出发点。

Stanford CS224n: Natural Language Processing with Deep Learning

Online Course *Jul. 2019 - Dec. 2019*

- 实作项目包含 Neural Dependency Parsing、Neural Machine Translation、Question Answering。

Jigsaw Unintended Bias in Toxicity Classification

Kaggle Competition *Feb. 2019 - May. 2019*

- 此竞赛之目的是要能辨别什么留言是在骂人亦或是无恶意之叙事。我们小组分别设计了多个基于不同模型（如 BERT, ELMo）的 classifier，并将其 ensemble。在竞赛期间，我们组曾达到 Top 1%。

SemEval-2013 Task 13: Word Sense Induction for Graded and Non-Graded Senses

Course Project *Jun. 2019 - Jul. 2019*

- 这是一个 word disambiguation 的任务，考虑相同词在不同句子中可能代表不同的含意。其中有两个子任务，一个是寻找该词最相近的 WordNet 解释，另一者则是将句子间相似的词进行 clustering。

SemEval-2018 Task 7: Semantic Relation Extraction and Classification in Scientific Papers

Course Project *May. 2019 - Jun. 2019*

- 这是一个 relation classification 的任务。主要文本是来自于多篇论文之 abstract 段落。对于句子中所标示之 instance，来判断两者间是属于所给定的哪种 relation 之一。

Chinese Word Segmentation, Part-of-speech Tagging, Named-entity Recognition

Course Project *Apr. 2019 - May. 2019*

- 经典的中文 sequence labeling 任务。主要基于 BiLSTM-CRF 与其他 baselines 进行比较。

混泥土汞车砼活塞故障预警

Digital China Innovation Contest 2019 *Jan. 2019 - Mar. 2019*

- 这个比赛中，数据是针对混泥土汞车在运行期间的一组 time-series 数据。目标是要预测某组序列运行后最终潜在故障的概率。我主要使用 LightGBM 并最终达到 Top 5%。

Engineering Project

Raspberry Pi 集群

Side Project *Sep. 2018*

- 我利用 4 个 Raspberry Pis 来搭建集群。自制了一个能快速部属 Hadoop、Spark 等 ecosystem 的脚本。

Leapsy AR 眼镜影像串流遥控摄像云台

个人接案 *Jul. 2017 - Oct. 2017*

- 我收集来自一个搭载 Android 系统的 AR 眼镜的 sensor 数据，用以捕捉当前使用者的姿态，并将此讯号传递给 Raspberry Pi 来同步遥控摄像云台之面向，最终将影像透过 Wi-Fi 来实时串流回眼镜中。
- 我设计 3D print 的模型来组合摄相头与两个伺服马达，同时也针对马达与 Raspberry Pi 设计了其电源电路。

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

本科毕业专题 *Apr. 2016 - Sep. 2016*

- 从零打造一台四轴飞行器，包含马达、扇页、sensor 等各晶片的挑选，与整体电路的设计。目标设计理念是可快速将整体框架套用在任何开发板上（只需对 IO 进行特化）。
- 在 control board 使用 PID 平衡算法，并在 CV board 上搭载基于 OpenCV 的 object detection。
- 分别在 HOLTEK MCU Design Contest 2016 获得佳作与在 ARM Design Contest 2016 获得 Top 10。同时也证明了我们的设计理念。