彭思尧

139-7492-8920 | loganpeng1992@gmail.com | logan-siyao-peng.github.io | logan-siyao-peng | 139-7492-8920

个人总结

- **求职岗位：** 篇章方向自然语言处理实习研究员；**实习时间：** 2020 年 11 月-2021 年 8 月；**实习城市：** 不限
- 热衷于研究**多语言**背景下**语义**和**篇章**任务，并将**语言学理论**和**语料库特征分析**融入到**深度学习**的模型里。

实习经历

腾讯 Tencent 中国北京

应用算法实习生 *2020 年 5 月 – 2020 年 9 月*

- 基于 Tranformer-based 和 XLM-based 的 Predicator-Estimator Models 的聚类模型，对训练数据进行实体增强，在 WMT2020 的句级别英中机器翻译 Quality Estimation Shared Task – Post-editing Effort 上获得并列第一。
- 通过长短文本匹配的方法对《腾讯看点》的新闻进行话题和概念的识别，用以增强用户画像，从而提高新闻推送的针对性、有效性。
- 对新闻文本进行粗粒度的实体识别和细粒度的实体分类，在多个人工标注的《腾讯看点》测试集上，识别效果 (F1) 优于 Texsmart。

培生集团 Pearson 美国科罗拉多州博尔德

自然语言处理实习生 *2019 年 6 月 – 2019 年 8 月*

- 为大学多类题材的短作文设计一个对于论点论据逻辑评估的通用标注系统，并进行三轮的内部评测和有声思维实验。
- 提取并计算自动篇章解析中的相关特征，例如树形 N-gram 的 KNN 和树编辑距离，并将这些特征嵌入一个用来预测文章组织结构分数的随机森林模型；实验发现自动篇章解析系统的不可靠性阻碍了随机森林模型的效果提升。

文章发表

M. Kranzlein, E. Manning, **S. Peng**, S. Wein, A. Arora, and N. Schneider. PASTRIE: A Corpus of Prepositions Annotated with Supersense Tags in Reddit International English. In *Proc. of LAW@COLING2020*

H. Wu, Z. Wang, Q. Ma, X. Wen, R. Wang, X. Wang, Y. Zhang, Z. Yao, and **S. Peng**. Tencent submission for WMT20 Quality Estimation Shared Task. In *Proc. of WMT@EMNLP2020*

S. Peng, Y. Liu, Y. Zhu, A. Blodgett, Y. Zhao, and N. Schneider. A Corpus of Adpositional Supersenses for Mandarin Chinese. In *Proc. of LREC2020*

L. Gessler, **S. Peng**, Y. Liu, Y. Zhu, S. Behzad, and A. Zeldes. AMALGUM – A Free, Balanced, Multilayer English Web Corpus. In *Proc. of LREC2020*

Y. Yu, **S. Peng**, and G. Yang. Modeling Long-Range Context for Concurrent Dialogue Acts Recognition. In *Proc. of CIKM2019*

Y. Yu, Y. Zhu, Y. Liu, Y. Liu, **S. Peng**, M. Gong, and A. Zeldes. GumDrop at the DISRPT2019 Shared Task: A Model Stacking Approach to Discourse Unit Segmentation and Connective Detection. In *Proc. of DISRPT@NAACL2019*, pages 133–143

Y. Zhu, Y. Liu, **S. Peng**, A. Blodgett, Y. Zhao, and N. Schneider. Adpositional Supersenses for Mandarin Chinese. In *Proc. of SCiL@LSA2019*, pages 334–337

S. Peng and A. Zeldes. All Roads Lead to UD: Converting Stanford and Penn Parses to English Universal Dependencies with Multilayer Annotations. In *Proc. of LAW-MWE-CxG@COLING2018*, pages 167–177

科研经历

篇章分析 Discourse Analysis 美国华盛顿特区

指导老师: Dr. Amir Zeldes *2019 年 1 月 - 至今*

- 对两个不同的篇章标注系统 – 修辞结构理论 (Rhetorical Structure Theory, RST) 和宾大篇章树库 (Penn Discourse Treebank, PDTB) – 下的同组文本进行隐含关系的关联分析，以便于标注的转换与生成。
- 为切分基础修辞单位 (Elementary Discourse Unit) 的聚合系统构建基于句法特征的回归模型，在小数据集上弥补神经网络的欠拟合。

多语言、多文体语义标注 Multi-lingual & Multi-genre Semantic Annotation 美国华盛顿特区

指导老师: Dr. Nathan Schneider *2018 年 5 月 - 至今*

- 指导标注团队对英语非母语者所发的英文 Reddit 帖子进行语义标注并分析母语对于英语介词选择的影响，以助于母语识别等项目。
- 通过将英文介词标注应用到中文平行语料上并对双语标注进行定性和定量的分析，证实介词语义标注框架的跨语言适应性。

教育背景

乔治城大学 Georgetown Univ. 美国华盛顿特区

计算语言学 博士 *2017 年 8 月 - 预计 2021/2022 年*

石溪大学 SUNY - Stony Brook 美国纽约州石溪

语言学 博士转学 *2016 年 8 月 - 2017 年 5 月*

莱顿大学 Leiden Univ. 荷兰莱顿

类型语言学 硕士 *2015 年 8 月 - 2016 年 8 月*

加州大学伯克利分校 UC Berkeley 美国加州伯克利

应用数学 & 语言学 & 法语 本科 *2011 年 8 月 - 2015 年 5 月*

相关课程

NLP	自然语言处理、语料库语言学、篇章建模、对话系统、机器翻译、机器学习
计算机	数据结构和算法、计算机程序结构和解析
数学	概率、线性代数、离散数学、微分方程、微积分

专业技能

NLP	Scikit-learn, Numpy, Pandas, StanfordNLP, Flair, Keras, Pytorch, Tensorflow
编程	Python, R, Bash, MATLAB, SQL, Haskell
其他	Linux, Bash, Google Cloud, Git, LaTeX

助教经历

2020 秋	语言学导论 Introduction to Languages
2020 春	语言数据解析 Analyzing Language data with R
2020 春	机器翻译 Statistical Machine Translation
2019 春	语言数据解析 Analyzing Language data with R
2018 秋	自然语言处理导论 Intro: Natural Language Processing
2017 春	美国中的语言 Languages of the USA
2016 秋	世界上的语言 Languages of the world