陈贝多

求职意向：自然语言处理算法工程师

籍贯：湖北宜昌　出生日期：1998.03.26　个人主页：http://home.ustc.edu.cn/~beiduo/
手机/微信：+86 18356503893　邮箱地址：beiduo@mail.ustc.edu.cn

科研/竞赛经历

SemEval 2022 Task 11 MultiCoNER 三赛道冠军

2021.10 – 2022.2　科大讯飞

- 主导参与著名自然语言处理比赛 SemEval 2022 多语言复杂命名体识别任务的全部十三个赛道，拿下三个**冠军**（中文、孟加拉语、多语种混合赛道），在其他十个赛道获得亚军。参赛队伍包括阿里达摩院、网易互娱、中科院自动化所、浙大、国科大以及众多国内外高校和科研机构。
- 以 XLM-RoBERTa 为预训练模型，选取 Span、CRF、Softmax 等分类器构建命名体识别基线系统。分析数据来源，通过 Wikepedia 构造对应实体库，实现初步实体库融合网络，在此基础上提出**实体库适应性整合网络**方法 (gazetteer-adapted integration network)，降低实体库网络与预训练模型的割裂性，进一步提升性能。通过实体替换等数据增强策略，针对每个语种赛道训练多个模型并整合，达到每个赛道上的最优效果。

基于对比学习的多语言预训练研究

2021.5 – 2021.10　语音及语言处理国家工程实验室

- 设计了一种**多层次对比学习**方法 (multi-level contrastive learning)，将句子级别与单词级别信息整合到同一个预训练框架下，显著提升了多语言模型 mBERT 的跨语言对齐能力；同时设计了一个**过零对比损失函数** (cross-zero noise contrastive estimation)，减轻了在低计算资源下浮点误差对对比学习训练的负面影响，进一步提升模型在小批次训练下的效果。
- 以 mBERT 为基础预训练模型，在 Google 的著名多语言基准 Xtreme Benchmark 中的 6 个任务上获得**性能的显著提升**，其中 BUCC 2018 的多语言句子对齐任务效果提升 80%。本工作发表在 ICASSP 2022。

基于多层次信息融合的多语言 NLP 研究

2020.9 – 2021.5　语音及语言处理国家工程实验室

- 设计了一种基于注意力机制的**双层特征融合模块** (double layers feature aggregation)，使多语言模型 mBERT 上下层信息互补以提升跨语言性能。
- 以 mBERT 为基线，在多语言基准 Xtreme Benchmark 上获得 4 个任务超过 40 个语言的**一致性性能提升**，并依据大量实验经验性分析得到 mBERT 中有关语种和语法信息分布的结论。本工作发表在 ICPR 2022。

基于混合短通道蒸馏学习的跨语言命名体识别研究

2022.3 – 2022.6　语音及语言处理国家工程实验室

- 设计了一种**混合短通道蒸馏**框架，以 mBERT 为基线利用多层信息进行师生模型蒸馏，辅以 MMD 域拉近方法，显著提升跨语言 NER 性能。
- 在经典多语言命名体识别数据集 CoNLL02/03 上取得目前为止在 zero-shot 条件下**最优的平均效果 81.53**，超越 ACL 2022 最新文章 MTMT 的 sota 80.68。同时在另外两个公开学术数据集上同样取得了远超之前工作的效果。本工作发表在 EMNLP 2022。

教育背景

- 2020.9 – 2023.6　中国科学技术大学（硕士）
  信息科学与技术学院　GPA: 3.93/4.3　Top 1%
- 2016.9 – 2020.6　中国科学技术大学（本科）
  信息科学与技术学院　GPA: 3.77/4.3　Top 2%

实习经历

- 2022.6 – now　微软亚洲研究院（研究实习生）
- 2021.6 – 2022.3　科大讯飞（研究实习生）

学术论文/专利（均为第一作者）

- Multi-Level Contrastive Learning for Cross-Lingual Alignment (2022 IEEE International Conference on Acoustics, Speech and Signal Processing, **ICASSP 2022**)
- Wider & Closer: Mixture of Short-channel Distillers for Zero-shot Cross-lingual Named Entity Recognition (The 2022 Conference on Empirical Methods in Natural Language Processing, **EMNLP 2022**)
- USTC-NELSLIP at SemEval-2022 Task 11: Gazetteer-Adapted Integration Network for Multilingual Complex Named Entity Recognition (The 16th International Workshop on Semantic Evaluation at 2022 Annual Conference of the North American Chapter of the Association for Computational Linguistics, **SemEval at NAACL 2022**)
- Feature Aggregation in Zero-Shot Cross-Lingual Transfer Using Multilingual BERT (26TH International Conference on Pattern Recognition, **ICPR 2022**)
- 专利：基于数据增强训练的多特征融合神经机器翻译检错方法（专利号 CN112926345A，国家知识产权局）

获奖经历

- 2022 研究生国家奖学金
- 2020 苏州育才奖学金（年度 GPA 最高）
- 2019 中科院电子所奖学金
- 2018 中国大学生数学建模竞赛三等奖
- 2021/2022 中国科学技术大学研究生一等奖学金

个人技能

- 编程语言：熟悉并会使用 Python、C/C++、Shell 等
- 工具框架：熟悉并会使用 Pytorch、Matlab 等
- 模型方法：熟悉并会使用和改进 Transformer、BiLSTM 等
- 英语能力：CET4、CET6 级证书，托福总分 98