



北京交通大学 计算机科学与技术学院
BEIJING JIAOTONG UNIVERSITY School of Computer Science & Technology

自然语言处理团队

Natural Language Processing Group



■ 团队介绍

北京交通大学自然语言处理研究组（语言智能与大数据处理研究所）成立于 2009 年 8 月，现已发展为具有国际先进国内领先的从事自然语言处理研究的科研团体之一。研究室现有教授 2 人，副教授 1 人、讲师 1 人、博士生、硕士生等 50 余人组成的研究队伍。研究组以自然语言处理与大模型技术为核心，聚焦于前沿人工智能研究与应用。主要研究方向包括：



大模型技术研发与优化：专注多语言大模型、多模态大模型的训练与评估，以及模型轻量化与高效推理技术。

知识增强与认知智能：结合知识图谱与大模型，研究知识驱动的智能对话系统、问答技术与认知推理能力提升。

多模态信息处理与 AIGC：探索文本、图像、语音等多模态数据的融理解与生成，推进 AI 生成内容技术的创新应用。

垂直领域智能应用：聚焦交通、地理等安全关键领域的大模型应用，解决实际场景中的复杂问题。



团队已研发多个具有影响力的开源模型与系统，包括多语言大模型、智能对话平台以及面向交通领域的专业大模型系统，并与多家企业建立了产学研合作关系。团队导师成员在 ACL、EMNLP 等国际顶级会议发表论文百余篇，并获得多项国家级科研项目支持，在国内外学术界和产业界获得广泛认可。

迄今为止，研究组已培养毕业博士和硕士 100 余人。曾先后与中国科学院计算技术研究所、中国科学院自动化研究所、日本情报通信研究机构（NICT）等国际知名企业及研究机构合作研发，与美国、日本、加拿大、欧洲、新加坡、香港、台湾等多个国家或地区建立了合作关系。

研究组在 AI 赋能轨道交通方向有丰富的学科交叉前沿研究经验，在自然语言处理、机器翻译和大语言模型等方面积累了丰富的模型研究与系统研发经验，主持相关领域的国家自然科学基金项目、国家重点研发计划、863 计划专项课题、科技部中日合作项目和科技部中越合作项目等若干项国家项目和国际合作项目，取得了一系列重要的研究成果。

研究组本着“团结、合作、创新、发展”的实验室文化，以精益求精的务实风格和勇于探索的求真精神，精诚合作，开拓创新。今后，我们将以自然语言理解为核心，以提高社会生产力为目标，努力开拓创新，将研究成果在机器翻译、文本大数据处理（云计算）、人机对话系统、语音语言智能控制等应用领域进行成果转化。



■ 导师介绍

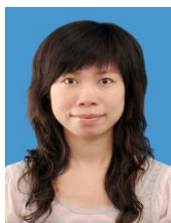


徐金安

教授，博士生导师

北京交通大学语言智能与大数据处理研究所主任 1992 年毕业于北方交通大学，获学士学位；同年被分配到铁道部第四勘测设计院上海分院工作，1997 年任工程师职务。1999 年至 2006 年留学日本，分别于 2003 年和 2006 年获得日本国北海道大学大学院工学研究科电子情报专攻自然语言处理专业工学硕士和工学博士学位；2006 年至 2009 年就职于日本电气株式会社中央研究院研究员。2009 年归国，任教于北京交通大学计算机学院。CCF 杰出会员，中国计算机学会自然语言处理专委会等委员、中文信息学会专委会委员、人工智能学会多语信息处理专委会副秘书长。

研究方向：机器翻译、自然语言处理、人机交互、大数据处理、文本情感分析、对话系统、知识图谱、多语种多模态大模型及其应用等。

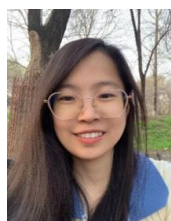


陈钰枫

研究员，博士生导师

1999 年至 2003 年于北京交通大学获得学士学位；2003 年至 2008 年于中科院自动化所获得博士学位；2008 年至 2014 年任职于中科院自动化所模式识别国家重点实验室，2014 年至今任教于北京交通大学。于 2014 年荣获钱伟长中文信息处理科学技术奖一等奖。在国内外权威期刊和会议上发表高质量论文 100 余篇；已申请国家发明专利 24 项；主持国家自然科学基金面上项目 3 项、AI2030 国家重点研发计划任务 1 项及省部级项目 12 项；作为核心成员参加国家自然科学基金重点项目、腾讯校企合作等各类纵向科研项目 30 余项。

研究方向：多语言大模型、多模态大模型、知识图谱、对话系统等。



韩文娟

副教授，硕士生导师

中国科学院大学博士，加州大学洛杉矶分校访问学者，新加坡国立大学 (NUS) 研究员。EMNLP 2022 研讨会 (UM-IoS Workshop) 主席，ACM SIGAI CHINA 新星奖，铁科院审稿顾问，北京通用人工智能研究院专家顾问。在国内外公认的本学科权威学术刊物上发表 40 余篇。研究方向为多模态大模型，及在综合交通等安全关键领域的应用。

研究方向：自然语言处理、多模态信息处理、认知智能、大模型等、AIGC。



黄锴宇

讲师，硕士生导师

大连理工大学计算机和日语双学士、博士，清华大学智能产业研究院博士后，合作导师为刘洋教授。主持国家自然科学基金青年项目，青英人才计划。长期从事预训练语言模型及自然语言处理研究工作，在多语言信息处理、信息检索、持续学习和科学智能等领域有着多年的研究经验。已在国内外重要学术期刊和会议上发表 50 余篇论文，并获得 ACL 2023 杰出论文奖。

研究方向：多语言信息处理，对话问答，地理科学智能，信息检索。

■ 研究内容

自然语言处理：包括语音识别、自然语言理解、生成、人机交互及所涉及的中间阶段，是人工智能和计算机科学的交叉学科。自然语言处理涉及词法，句法分析等基础任务，机器翻译，自动问答等核心技术，并将其应用于推荐系统，搜索引擎等领域。



随着深度学习的发展，自然语言处理具有诸多研究热点，正处于黄金发展阶段。尤其是 2022 年 ChatGPT、GPT4.0 等多语种多模态大规模预训练模型的诞生，把 NLP 技术的发展上升到了通用人工智能和认知智能的新阶段！

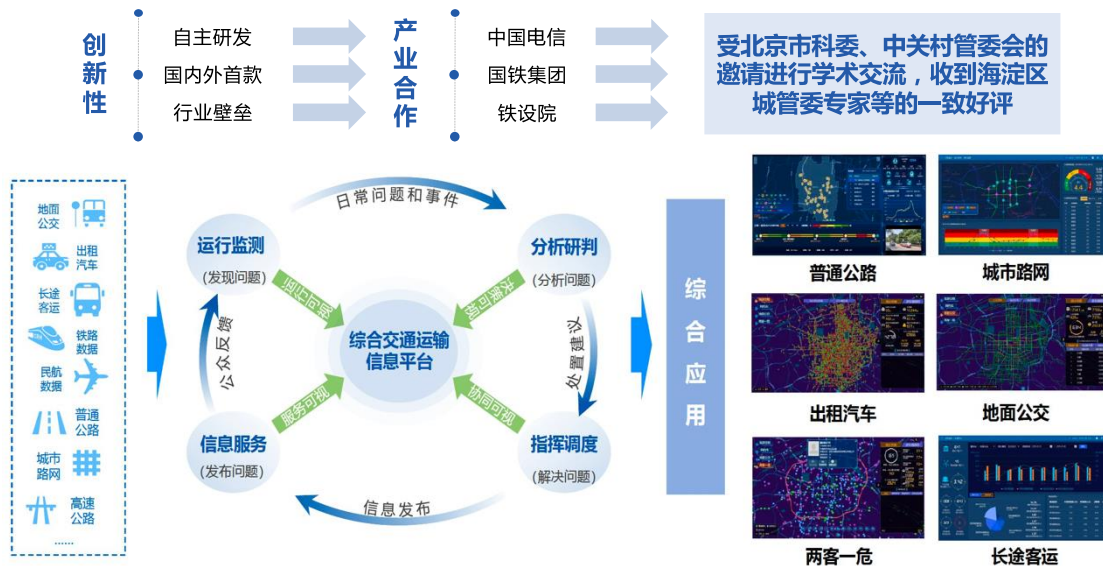
■ 研究成果

本研究组的代表性成果是在日汉机器翻译研究方面，开创性地提出了融合格依存文法的日汉依存树到串统计机器翻译理论并构建了翻译系统，并且我们还自主研发了端到端神经网络机器翻译系统，在国际国内的机器翻译评测任务中，先后夺得 WAT2016 夺得 3 个第一、CCMT2020 夺得 4 个第一、WMT2020 和 WMT2021 分别夺得英汉第一、CCMT2021 夺得 5 个第一名。



研究成果

依托“交通大数据与人工智能”教育部重点实验室、联合中国计算机学会智慧交通分会，发布了国内外首款交通大模型 TransGPT·致远。将城市中与交通相关的人、车、路、环境数据汇集到一个平台，准确、实时、精准的掌握交通态势，主动发现交通异常、分析原因、参与调度。



TransGPT·致远应用落地

完成于2023年7月

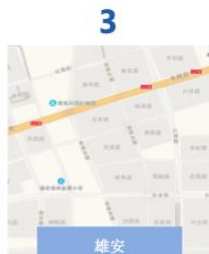
<https://github.com/DUOMO/TransGPT>



在大模型赋能铁路智能化升级方面，与中国铁路设计集团有限公司展开深度合作，研究BIM模型的自动化审核，作为中国国家铁路集团和中国铁建的第一个大模型项目，起到标杆性示范效应。



2023年由中国电信携手国内外众多知名企业举办的“2023数字科技生态大会”在广州拉开大幕。作为中国电信一年一度的盛会，中国电信与本单位共促数实新生的合作展望，发布“道驰公路大模型”，全面展现了各行各业数智化的最新成果“道驰公路大模型”。



构建基于TransGPT·致远的超长时流量预测系统，通过该系统的应用，可以实现对交通流量的精准预测和有效管理，为雄安新区的智能交通建设和城市发展提供有力技术支持。

企业名	申请使用时间	用途
邢台会联达司法鉴定所	2023-07-23	鉴定所小程序
中煤电气	2023-08-02	商用联系人：叶春阳
北京冀都科技发展有限公司	2023-08-02	商用
广西交投科技股份有限公司	2023-08-07	预研联系人：尹杰
泰安市公安局交警支队	2023-12-05	用于高速公路交通管理数据分析。联系人：王磊
其他企业、高校		
上海建信信息	2024-01-25	应用，用于智能交通业务，联系人：CEO陈树兴
上海智能交通有限公司	2024-03-04	商业使用，用于“资源统筹+全程服务”的一站式整体解决方案。联系人：王雪燕
中国人民公安大学警学院	2024-03-21	上课使用
中国电建集团华东勘测设计研究院有限公司	2024-03-21	预研联系人：AI实验室负责人殷丹鹏
苏文利集团股份有限公司	2024-04-07	预研联系人黄宜亮

TransGPT·致远所获奖励和荣誉



获2024全国人工智能应用场景创新挑战赛优胜奖



论文《交通大模型对智慧交通教育的机遇与挑战》获CCF举办2024中国高校计算机教育大会二等奖



获由成都政府举办的2024人工智能大模型基准测试科创大会暨中西部数字经济峰会交通领域“垂直大模型创新应用”奖



获由CCF举办的NLPC2024会议 Outstanding Paper 杰出论文奖

研究成果

研究组累计翻译出版译著 4 部、发表学术论文 200 余篇，其中包括 EMNLP、ACL、AAAI、IJCAI、COLING、TALLIP、TPAMI、TASLP、AI 等国际顶会和期刊 50 余篇，申报国家级发明专利 50 余项。连续 11 年校级优秀硕士毕业论文（赖思羽等）。其中包括 2021 年发表顶会论文 17 篇、2022 年 21 篇，2023 年 ACL 录用 7 篇，投稿 EMNLP2023 累计 19 篇，2024 年 ACL 录用 5 篇，2025 年 ACL 录用 5 篇。申报发明专利 50 余项。



近期工作

ACL 2025

Less, but Better: Efficient Multilingual Expansion for LLMs via Layer-wise Mixture-of-Experts

Xue Zhang, Yunlong Liang, Fandong Meng, Songming Zhang, Yufeng Chen, Jinan Xu, Jie Zhou

提出了基于层感知的专家分配算法（LayerMoE）。首先通过对不同层的语言表征相似性进行分析，发现了存在的相似度差异性，进而根据相似度为各层分配新专家：相似性越高的层需要的新专家越少，相似性越低的层需要的新专家越多，实现了资源高效的新语言扩展。此外，在新旧语言相似性较高的层中，在路由网络前添加分类器，判断当前 token 属于旧语言还是新语言，引导旧语言 token 流向原始专家，进一步缓解了灾难性遗忘。

AlignDistil: Token-Level Language Model Alignment as Adaptive Policy Distillation

Songming Zhang, Xue Zhang, Tong Zhang, Bojie Hu, Yufeng Chen, Jinan Xu

提出了一种 token 级的大模型对齐方法，显著优于 DPO、PPO 等算法，且有更快的收敛速度。

ACL 2025 Findings

A Law Reasoning Benchmark for LLM with Tree-Organized Structures including Factum Probandum, Evidence and Experiences.

Jiaxin Shen, jinan Xu, Huiqi Hu, Luyi Lin, Guoyang Ma, Fei Zheng, Fandong Meng, Jie Zhou, Wenjuan Han

我们首次提出了法律领域的证据推理基准，该基准融合了层次化的犯罪事实、犯罪证据和隐含的人类经验常识，防止最终的事实认定产生偏见。这一基准的提出为可信人工智能法律系统的构建指明了方向。

Migician: Revealing the Magic of Free-Form Multi-Image Grounding in Multimodal Large Language Models

You Li, Heyu Huang, Chi Chen, Kaiyu Huang, Chao Huang, Zonghao Guo, Zhiyuan Liu, Jinan Xu, Yuhua Li, Ruixuan Li, Maosong Sun

关注于多模态大模型的细粒度感知能力，将传统视觉定位任务拓展到多张图像的场景下，推出 Migician 大模型，支持任意形式多图视觉定位任务。

Enhancing Cross-Tokenizer Knowledge Distillation with Contextual Dynamical Mapping

Yijie Chen, Yijin Liu, Fandong Meng, Yufeng Chen, Jinan Xu, Jie Zhou

该工作提出了 CDM，一个跨分词器的知识蒸馏方法，它使用上下文信息动态优化序列层面和词汇层面的对齐，以提高蒸馏的效果。实验在五个主流的大语言模型上进行，涵盖指令遵循、代码生成、数学等任务，结果显示 CDM 比现有跨分词器蒸馏方法有明显的优势。

COLING 2025

Multilingual Knowledge Editing with Language-Agnostic Factual Neurons

Xue Zhang, Yunlong Liang, Fandong Meng, Songming Zhang, Yufeng Chen, Jinan Xu, Jie Zhou

本文提出了 LU-LAFNs，即通过定位并更新语言无关事实神经元来实现多语言知识编辑。该方法基于一项关键发现：LLMs 中不同语言的同一事实知识会激活一组共享神经元，即语言无关事实神经元（LAFNs），这些神经元代表跨语言共享的事实知识，隐含多语言知识的语义关联。通过建模多语言知识的语义关联，避免了知识冲突，在 Bi-ZsRE 和 MzsRE 基准上的编辑性能（如可靠性、泛化性）优于现有方法。

NLPCC 2024 Best Paper

Model-Agnostic Knowledge Distillation Between Heterogeneous Models

Jiaxin Shen, Yanyao Liu, Yong Jiang, Yufeng Chen, and Wenjuan Han.

传统的知识发展的师生范式不能适用于具有不同输出表征的教师模型和学生模型（称为异质模型）。为了克服这个限制，我们提出了一个模型无关的 KD 方法叫做 MAKD。MAKD 包括信息样本生成和知识转移两个阶段。其核心思想是生成能够反映教师模型与学生模型之间差异的信息样本，从而使生成的样本具有有效性和高效性。

在研及完成的科研项目主要包括：科技部科技创新 2030 重大专项课题、国家重点研发计划项目子任务、国家自然科学基金面上项目、北京市自然科学基金、科技部中日国际合作项目，科技部中越合作项目等重要科研项目 30 余项。

■ 学长风采



刘宜进

2021 年毕业，以大咖 OFFER 就职于腾讯微信模式识别研究组，学硕在读期间共发表 ACL、AAAI、EMNLP 等顶级会议 3 篇，夺得 WMT2020 机器翻译评测英汉方向第一名。



梁云龙

2020 级计算机科学与技术专业博士研究生，中共党员，师从徐金安教授。曾荣获 2 次国家奖学金和 1 次知行奖学金提名奖。科学研究方面能力突出，专注学科前沿，在导师的指导下，致力于攻克对话机器翻译课题中译文不流畅、译文不能充分传达对话特性等难题，已发表 16 篇国际顶级期刊或会议论文（均为一作或共一，其中 CCF A: 10 篇），Google 论文引用次数 792；曾荣获 WMT 国际机器翻译评测对话翻译 2 个世界冠军、国家级人工智能类竞赛亚军以及企业级亚军等成绩。



赖思羽

2020 级学硕，曾获国家奖学金、北京市优秀毕业生和校级优秀毕业论文。在校期间一作发表 nlp 领域顶会论文 3 篇，专利和软著各一项。目前就职中国人寿研发中心，从事算法相关工作。



梁晨

2021 级硕士，硕士期间从事信息抽取的研究，参与文心一言推理能力建设，共发表包含 ACL、EMNLP 以及 SCI 1 区期刊在内的 4 篇论文，获得北京交通大学硕士优秀学位论文，现已入职百度。



陈怡洁

2022 级学硕，曾获国家奖学金与校优秀毕业生，硕士期间研究大语言模型的训练优化与应用，并在 ACL 顶级会议上发表论文两篇，在 ICML workshop 上中稿论文一篇。目前就职腾讯微信，继续从事 NLP 与大模型相关工作。



毛郁龙

2025 年毕业，获得国家奖学金、北京市优秀毕业生、优秀学位论文；在校期间以第一作者在 NLP 顶级会议 ACL 发表论文，并获评做口头报告；硕士期间累计实习超一年，参与或主导十余个算法项目，影响亿级别用户；毕业后在字节跳动从事算法工作。



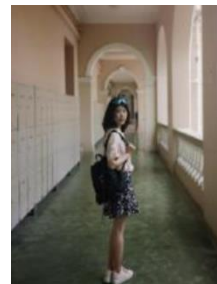
学长风采



吴培昊
2015 年毕业
字节跳动 AI lab



刘江鸣
2015 年毕业
爱丁堡大学读博



明芳
2017 年毕业
铁科院



李少童
2018 年毕业
腾讯



刘方旭
2019 年毕业
腾讯

更多去向

毕业年份	去向单位
2023 年	百度、字节跳动、创新工场、中国交通银行上海分行
2022 年	百度、字节跳动、中国工商银行、京东研究院
2021 年	腾讯微信、字节跳动、中国农业银行北京分行、创新工场
2020 年	腾讯 AI 研究院、腾讯 CSIG 智能平台部、字节跳动 AI lab、国家开发银行上海分行、中国银行北京分行、中国工商银行上海分行
2019 年	腾讯 CSIG 智能平台部、中国移动通信有限公司研究院、字节跳动 AI lab、360 算法中心、航天信息股份有限公司、北京日报社技术部、京东 AI 平台与研发部、中国银行国际金融研修院
2018 年	腾讯 AI 研究院、搜狗 AI 研究部、苏州移动研究院、航天信息股份有限公司、国家电网、中国银行
2017 年	华为诺亚方舟实验室、建信基金、北京市专利局、中国铁道科学研究院、北京市建筑设计研究院有限公司、中国银行



■ 团队瞬间



欢迎优秀的你加入



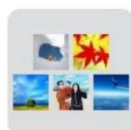
嗨! 🤝 欢迎加入北交 NLP 团队!

BJTU-NLP 欢迎你 (中文)

BJTU-NLP ᠪᠢᠵᠤ-ᠨᠯᠣᠫ ᠲᠡግᠦᠨ ᠮᠤᠩᠭᠤᠯ (蒙文)

BJTU-NLP ཁྱོད་ལ་དགའ་བསུ་ཞུ། (藏文)

سبزنی



群聊：2025年-保研夏令营-NLP
团队群



该二维码7天内(7月14日前)有效, 重新进入将更新





北京交通大学 计算机科学与技术学院
BEIJING JIAOTONG UNIVERSITY School of Computer Science & Technology

知行

北京交通大学自然语言处理团队制作

Made by BJTU-NLP

2025 年 7 月