2019年度自治区科技奖励提名项目公示信息
（科技进步奖）

一、**项目名称：**复杂场景下维吾尔语文本发现与机器翻译关键技术研发与应用

二、提名单位

中国科学院新疆分院

三、提名单位意见：

该项目紧密结合国家和自治区的应用需求，针对多源、多通道、多形态、多格式、多错误（非标准）等复杂场景下海量维吾尔语文本分析处理研究，建立了资源稀缺语言高质量语料资源库和知识库，攻克了复杂形态语言的形态特征分析、层次化语言模型、命名实体识别、维汉机器翻译建模等系列关键技术，研发了词语形态切分、形态还原、词性标注及词对齐等系列工具，构建了复杂场景下海量文本的相似语种间语种识别、同根关键词检索、实体关系抽取、维汉机器翻译全流程应用平台。项目采用的技术方法与路线合理，数据、资料可靠，分析论证科学，难度、规模和工作量很大。

该项目提出了资源稀缺语言的语料质量自动评价模型及领域适应性分析方法、多特征融合的文本相似性度量方法，建立了大规模高质量的维吾尔语语料知识库，为开展维吾尔语自然语言处理奠定了坚实基础。项目提出了面向资源缺乏语言的实体链接方法，建立了主动预见式的话题事件检测模型及融合复杂语言形态特征的维汉机器翻译等模型，突破了形态复杂语言同根词检索、命名实体识别、实体关系抽取、机器翻译等系列关键技术。自主研发了形态相似语言语

种识别、维吾尔语关键词检索、维吾尔语同根关键词检索、维吾尔语-汉语双语语料对齐等 9 个工具软件，形成了复杂场景下海量维吾尔语文本处理的工程技术体系。

该项目研究成果获发明专利 2 项、软件著作权 27 项，培养硕博士研究生 31 人，发表学术论文 74 篇。项目系列技术成果已在疆内外的反恐维稳、公共安全、电子政务、教育、电子取证、保密检查等领域广泛应用。项目技术成熟，产品质量很好，应用价值重大，经济和社会效益很大。

该项目材料内容真实、完成单位及个人排名不存在异议，提名该项目为自治区科技进步一等奖。

四、项目简介：

新疆是“一带一路”核心区，也是我国反恐维稳主战场，维吾尔族是新疆的主体民族之一，维吾尔语使用人口超过 1300 多万人，网络环境下维吾尔语文本发现与机器翻译技术研发与应用，对促进民族间文化交流，提升信息获取掌控能力，维护社会稳定和长治久安总目标具有重大的意义。在中国科学院、自治区政府的支持下，中国科学院新疆理化技术研究所（以下简称新疆理化所）集中优势技术力量，历时 10 年，针对多源、多通道、多形态、多格式、多错误（非标准）等复杂场景下海量维吾尔语文本分析处理开展研究，建立了资源稀缺语言高质量语料资源库和知识库，攻克了复杂形态语言的形态特征分析、层次化语言模型、命名实体识别、维汉机器翻译建模等系列关键技术，研发了词语形态切分、形态还原、词性标注及词对齐等系列工具，构建了复杂场景下海量文本的相似语种间语种识别、同根关键词检索、实体关系抽取、维汉机器翻译全流程应用平台。

1、 提出了资源稀缺语言的语料质量自动评价模型及领域适应性分析方法和多特征融合的文本相似性度量方法，建立了大规模高质量的维吾尔语语料知识库，研发了一系列文本智能处理工具库，为开展维吾尔语自然语言处理奠定了坚实基础和技术支撑。

2、 构建了资源稀缺型语言的事件抽取模型及话题事件检测模型。提出了面向资源缺乏语言的实体链接方法，通过构建实体指称项的上下文特征和全局特征，实现候选实体的扩充和排序，扩大事件检测覆盖率，有效缓解数据稀疏问题，提高事件抽取准确率；

3、 针对非标准维吾尔语文本与标准维吾尔语文本的差异性，构建非标准维吾尔语的纠错模型，结合黏着语的复杂语言形态生成机理，适当粒度的词语表示及建模，融入命名实体识别与翻译、基于多特征的调序等，构建了基于复杂形态语言非标准文本特征分析的维汉机器翻译系统；

4、 突破了形态复杂语言同根词检索、命名实体识别、实体关系抽取、机器翻译等系列关键技术，构建了相似语种间文本语种识别系统、同根关键词检索系统、实体关系抽取系统、到非标准文本特征分析的维汉机器翻译系统全流程应用平台，形成了复杂场景下海量维吾尔语文本处理的工程技术体系。

五、推广应用情况

项目系列技术成果已在疆内外的反恐维稳、公共安全、电子政务、教育等领域广泛应用。项目成果获国家发明专利2项、软件著作权27项，发表学术论文74篇，培养硕博士

研究生31人。项目整体成果在第三方评价中评为“国际领先”水平，其中非受限维汉机器翻译系统在2013年、2015年、2017年的全国机器翻译评测中连续获得第一名，成果在上海、广州、杭州、温州、厦门等地的国安和公安部门部署应用，产生直接和间接经济效益5100多万元。项目成果拓展了信息文化沟通交流渠道，提升了信息获取与掌控能力，显著推动了自然语言信息处理领域的技术进步，为维护新疆地区社会稳定和长治久安提供了强有力的技术支撑。

六、主要知识产权证明目录：

1）共获得国家发明专利2项

2017年10月，获得1项发明专利，专利名称：多特征融合的文本相似性度量系统,专利号：ZL 2015 1 0072955.2。

2018年3月，获得1项发明专利，专利名称：面向资源缺乏语言的实体联接方法,专利号：ZL 2015 1 0304943.8。

2）共获得软件著作权27项，主要包括：

序号	软件著作权名称	编号
1	少数民族语言词典软件V1.0	2017SR188888
2	文本语料管理系统V1.0	2017SR189149
3	基于多语种网站的爬虫系统V1.0	2016SR203310
4	基于多语种网页抽取语料系统V1.0	2016SR203321
5	机器翻译语料覆盖度选取系统V1.0	2016SR203913
6	面向多语种的语料筛选软件V1.0	2016SR203918
7	维吾尔语词性标注软件V1.0	2016SR206385
8	面向谷歌浏览器（Chrome）的维汉机器翻译插件软件V1.0	2016SR206400
9	维吾尔语-汉语双语语料对齐软件V1.0	2017SR210739
10	维吾尔语（词级搜索）关键词检索系统[简称：维吾尔语关键词检索系统]V1.0	2017SR634649

11	维吾尔语（词根级搜索）关键词检索系统[简称：维吾尔语词根关键词检索系统]V1.0	2017SR634658
12	基于网页的维汉机器翻译系统 V1.0	2017SR649088
13	维汉机器翻译系统融合软件 V1.0	2017SR698691
14	非标准拉丁维吾尔文转现行维吾尔文软件	2018SR1030794
15	维吾尔语文本摘要系统	2018SR1030102
16	维汉机器翻译系统	2018SR1031920
17	电子物证检验中维汉文本智能处理软件	2018SR1091303
18	多特征融合词向量训练系统	2019SR0187259
19	多语言语种识别系统	2019SR0181376
20	多语种语料库标注平台	2019SR0181385
21	基于深度迁移学习的维吾尔语文本摘要系统	2019SR0194675
22	基于神经网络的相似语言短文本语种识别软件	2019SR0187265
23	基于子词信息的维吾尔语词项规范化	2019SR0182534
24	平行语料管理平台	2019SR0187268
25	维吾尔谚语识别系统	2019SR0187262
26	维吾尔语关键词翻译及生成软件	2019SR0183165
27	维吾尔语新闻话题检测系统	2019SR0187270

七、主要完成人情况：

排名	姓名	职务/职称	工作及完成单位	对成果创造性贡献
1	蒋同海	所长/研究员	中国科学院新疆理化技术研究所	组织协调、总体设计
2	杨雅婷	副研究员	中国科学院新疆理化技术研究所	组织协调、框架设计
3	王　磊	研究员	中国科学院新疆理化技术研究所	组织协调、设计、推广
4	董　瑞	助理研究员	中国科学院新疆理化技术研究所	模型设计、算法设计、开发、
5	马　博	副研究员	中国科学院新疆理化技术研究所	算法设计、开发
6	吐尔洪·吾司曼	助理研究员	中国科学院新疆理化技术研究所	资源库建设、模型训练、开发

7	丁景全	处长/副研究员	中国科学院新疆理化技术研究所	资源库建设、推广
8	马玉鹏	副主任/研究员	中国科学院新疆理化技术研究所	开发
9	赵　凡	副研究员	中国科学院新疆理化技术研究所	开发
10	王晓博	副研究员	中国科学院新疆理化技术研究所	资源库建设、推广
11	艾孜麦提·艾尼瓦尔	助理研究员	中国科学院新疆理化技术研究所	资源库建设
12	刘香玉	工程师	中国科学院新疆理化技术研究所	测试

八、主要完成单位及创新推广贡献

单位名称	中国科学院新疆理化技术研究所				
排　　名	1	法定代表人	蒋同海	所 在 地	新疆乌鲁木齐
单位性质	事业	传　　真	0991-3838957	邮政编码	830011
通讯地址	新疆乌鲁木齐市北京南路 40-1 号				
联 系 人	盖敏强	单位电话	0991-3838931	移动电话	18709919732
电子邮箱	gaimq@ms.xjb.ac.cn				
对本项目科技创新和推广应用情况的贡献：					

项目成果成熟稳定，已在新疆、上海、广州、杭州、温州、厦门等地的国安、公安部门及企业部署应用，直接和间接应用人员17500多人，取得了很好的应用效果。开拓展了信息文化沟通交流渠道，提升了信息获取与掌控能力，为维护新疆地区社会稳定和长治久安提供了技术支撑。

新疆是国家“一带一路”战略的核心区，也是我国反恐维稳斗争的最前沿，全国维吾尔语使用人口超过1300多万人，网络环境下维吾尔语文本发现与机器翻译技术研发与应用，对促进民族间文化交流，提升信息获取掌控能力，维护社会稳定和长治久安总目标具有重大的意义。项目的研究成果应用于多领域多层面，均取得显著的应用效果和社会意义。

1、为维护新疆地区社会稳定和长治久安提供技术支撑

项目成果在公共安全领域，实现了对敏感信息、突发事件主动发现以及该主题事件在后续时间序列上的检测与内容理解，有效的辅助实际业务的开展。截至目前，项目成果已在新疆、上海、广州、杭州、温州、厦门等地的国安、公安部门及企业部署应用，直接和间接应用人员17500多人，取得了很好的应用效果，减轻了工作负担、提高了工作效率、提升了信息获取与掌控能力，为维护新疆地区社会稳定和长治久安提供了技术支撑。

2、显著推动了自然语言信息处理领域的技术进步。

项目开展复杂场景下海量维吾尔语文本分析处理研究，建立了资源稀缺语言高质量语料资源库和知识库，为开展维吾尔语自然语言处理奠定了坚实基础；攻克了复杂形态语言的形态特征分析、层次化语言模型、命名实体识别、维汉机器翻译建模等系列关键技术，研发了词语形态切分、形态还原、词性标注及词对齐等系列工具，为开展维吾尔语自然语言处理提供了技术支撑；构建了复杂场景下海量文本的相似语种间语种识别、同根关键词检索、实体关系抽取、维汉机器翻译全流程应用平台，形成了复杂场景下海量维吾尔语文本处理的工程技术体系，并广泛应用，显著推动了自然语言信息处理领域的技术进步。

3、促进了民族间的文化交流

在经济文化领域的应用，促进了民族间的文化交流，对将独具特色的民族文化向外宣传推广起到积极促进作用。项目成果的应用对维护我国民族地区的社会稳定和反对分裂活动、加强各民族交流、传承并发展民族文化等起到了极为积极的作用。

4、培养了人才，提高了新疆科技研究开发能力和水平

项目的研发攻克了一系列技术难关，取得系列产品，非受限维汉机器翻译系统在2013年、2015年、2017年的全国机器翻译评测中获得第一名，项目整体成果在第三方评价中评为“国际领先”水平，项目的实施与应用过程中，培养锻炼了一支扎根新疆的科研团队，同时培养硕博士研究生30多人。

声明：本单位同意完成单位排名，遵守《自治区科学技术奖励条例》及其实施细则的有关规定，承诺遵守评审工作纪律，保证所提供的有关材料真实有效，且不存在任何违反《中华人民共和国保守国家秘密法》和《科学技术保密规定》等相关法律法规及侵犯他人知识产权的情形。如有材料虚假或违纪行为，愿意承担相应责任并接受相应处理。如产生争议，保证积极配合调查处理工作。

法定代表人签名：　　单位（盖章）

年　月　日　　年　月　日

九、完成人合作关系说明

蒋同海研究员是成果总体主持人，是“面向网络信息采集的维汉机器翻译系统开发与应用”的项目负责人，王磊、杨雅婷、董瑞、马博、吐尔洪是项目主要骨干，丁景全、马玉鹏、赵凡、王晓博是该项目参与人员；杨雅婷是“维汉机器翻译中复杂语言形态模型的研究”、“基于多特征融合的复杂形态语言建模研究”的项目负责人，董瑞、艾孜麦提·艾尼瓦尔是该项目的主要参与人员、王磊是“维汉机器翻译关键技术研究及示范”项目负责人，杨雅婷、董瑞、刘香玉是该项目参与人员。