Gij Yonjgen Rox Hoiz Vahsaw Minzcuz Guek Raeuz Yenz-fat Cingzgvang、Nanzdiemj Caeuq Baenzlawz Gaijndei

我国民族语文智能翻译软件研发的现状、难点及对策建议

□ 卢天友

【提要】目前，国内从事民族语文智能翻译软件研发的机构除了中国民族语文翻译局，还有中国科学院计算所和自动化所、中国信息通信研究院、国家语言资源监测与研究少数民族语言中心、新疆大学、西藏大学等科研院所和高校，以及科大讯飞、捷通华声、腾讯等企业。研发的软件涉及蒙古、藏、维吾尔、哈萨克、朝鲜、彝、壮等国内主要少数民族语种，成果有近百款。与中外语文智能翻译软件的研发相比，我国民族语文智能翻译软件的研发还存在理论研究成果少、基础资源积累不足、技术研发难度大、市场应用前景小、人才队伍规模小等问题。民族语文智能翻译软件在维护国家安全、促进民族团结、铸牢中华民族共同体意识等方面有重要作用，为此需要以问题为导向，提出具体措施，做好民族语文智能翻译软件的研发和推广使用。

【关键词】民族语文软件 研发 现状 难点 对策建议

语言文字是信息最为重要的载体，信息化与语言文字工作本质上是“一体两面”。语言文字工作要满足信息化的需要，促进信息化的发展；同时，也必须借助信息化来全面提升语言文字工作。民族地区信息化的主要任务之一是民族语文的信息化。通过民族语文信息化推动各民族学习国家通用语言文字，科学保护各少数民族语言文字，提升民族地区干部群众的语言能力，铸牢中华民族共同体意识，构建我国和谐语言生活。民族语文智能翻译软件的研发是民族语文信息化的重要路径，民族语文智能翻译软件是实现和谐语言生活的有效载体。当前，随着信息技术和互联网的迅速发展，全球一体化进程加快，文化一体多元化意识加强，少数民族语言文字的使用范围、社会地位、文化功能、战略意义发生了显著变化。一方面，信息化将民族语文推到了国家安全统一与民族团结进步的战略高度。信息技术与互联网使得信息传播更为快捷和隐蔽，各种语言文字处理软件、自动翻译软件，广泛运用的各类新媒体、新型移动终端，都可能成为国家安全统一与民族团结进步的重要的守护者或破坏者。另一方面，支撑引领民族地区经济社会转型发展，提升政府社会治理和公共服务能力、效率，打破各民族间交往交流障碍，助力乡村振兴，迫切需要构建新型民族语文信息技术服务体系。因此，铸牢中华民族共同体意识，助力推广普及国家通用语言文字，为科学保护各民族语言文字、尊重和保障少数民族语言文字学习和使用提供技术支持，占领民族语文信息处理制高点，助推乡村振兴战略实施，是当前民族语文信息化工作的主要任务，是民族语文智能翻译软件研发的题中之义。

一、我国民族语文智能翻译软件研发现状

据不完全了解，目前国内从事民族语文智能翻译软件研发的机构主要有中国民族语文翻译局、中国科学院计算所和自动化所、中国信息通信研究院、国家语言资源监测与研究少数民族语言中心（设在中央民族大学）、中国民族信息技术研究院（设在西北民族大学）、中国朝鲜语言文字信息化基地（设在延边大学）、北京应用科学技术研究院、清华大学、内蒙古大学、西藏大学、新疆大学、青海师范大学、西北师范大学、西南民族大学、广西民族大学等科研院所和高校，以及科大讯飞、捷通华声、腾讯等企业。近年来，中科院计算所和自动化所与新疆大学、西藏大学、内蒙古大学、西北民族大学等高校进行了相关语种的研究，研发出了一些实验室产品，目前尚未见向社会提供翻译服务。国家语言资源监测与研究少数民族语言中心，侧重于理论研究，在民族语文信息化软件研发方面也取得了一定成果，包括语料库管理平台，藏文编码转换软件，藏文输入法软件，民文数据采集系统，朝文词法分析软件，汉、藏、维敏感词发现、管理与跟踪软件，传统蒙古文语料校对及词性标注辅助软件，维吾尔文网站搜集软件，Windows XP仿真Vista民族语文插件中心网络管理平台，等等。中国民族信息技术研究院的主要研究成果有《藏文操作系统》《藏汉双语信息处理系统》《藏汉英多功能组合软件》《藏文智能输入研究》等多项省部级新产品新技术及科技成果。中国朝鲜语言文字信息化基地近年来研发了“朝鲜语发音软件”等软件，目前市面上朝鲜文的智能翻译系统和语音翻译软件的应用相对来说比较成熟。科大讯飞、捷通华声、腾讯等企业通过独立研发或与相关高校联合研发出了“维汉口语即时翻译软件”“藏译通”“腾讯民汉翻译”小程序等实用性民汉智能翻译软件。

中国民族语文翻译局经过多年积累，有丰富的多语种平行语料，语料资源储备在国内同行业位居第一；基于业务工作需要既搞科研又做应用，在理论架构与应用研究方面取得了一些成绩，在基础理论研究、系统平台建设和实际应用方面走在了全国前列。目前，翻译局开展的多语种智能翻译系统研发已取得阶段性进展，已经研发出了蒙古、藏、维吾尔、哈萨克、朝鲜、彝、壮7个语种的智能翻译、语音识别和合成、OCR三大系统。在这三大系统上研发了50多款应用软件，包括智能翻译系统、民汉对话通、语音转写通（民汉）、浏览器民汉互译软件、民族语文语音输入法、Windows民族语文语音输入法、民汉智能语音翻译软件、民汉照相翻译、民汉实时翻译等。其中多个民族还实现了智能庭审语音系统的应用。这些软件的成功研发获得了国家版权局颁发的《计算机软件著作权登记证书》，在多个方面填补了我国民族语文翻译软件研发领域的空白，为提高我国民族语文翻译信息化水平，推进民族语文翻译新词术语规范化、标准化、信息化进程作出了贡献，有助于促进各民族的交往交流交融，也将为人工智能、中文信息处理等国家战略基础技术提供有效的技术支撑。

据不完全统计，截至2021年底，国内有关民族语文智能翻译软件产品的研发，涉及蒙古、藏、维吾尔、哈萨克、朝鲜、彝、壮等主要语种，成果有近百款软件，其中以中国民族语文翻译局研发的数量最多、语种最全面，仅在智能翻译、语音识别和合成、OCR三大系统基础上就研发出了50多款应用软件。

二、我国民族语文智能翻译软件研发存在的问题和技术难点

我国民族语文信息化工作始于20世纪80年代。近四十年来取得了长足发展，民族语文信息处理技术领域取得了不少研究成果，产生了积极的社会效益和经济效益，并有力地推动了相应民族语文信息技术的发展。近年来，随着人工智能技术的飞速发展，基于语音识别和语音合成的中文与外文智能翻译系统得到了快速发展。目前，与中外语文智能翻译系统研发进度及所取得的成绩相比，我国的民族语文智能翻译软件的研发还存在理论研究成果转化少、基础资源积累不足、技术研发难度大、市场应用前景小、人才队伍规模小等问题，这些问题制约了我国民族语文信息化的发展。

（一）理论研究成果转化少，合作研发体系不健全

学术力量尚未形成工作联动机制和资源整合模式。在我国民族语文信息化工作中发挥作用的学术力量，以体制内的有关科研机构和高等院校为主。他们通常有独立的部门、专家、相应的研究力量，取得了一定的理论研究成果，很多语料库等资源建设了起来，也研发出了一些实验室产品。但因为没有工作联动机制和研究成果不公开，这些资源与研究成果没能及时、有效转化，没能及时、便利地应用到民族地区的社会文化生活中来。

商业力量薄弱而且各自为战。目前，在少数民族语文的信息化智能化方面，因其市场前景小、投入与产出很难成正比，再加上相关政策扶持力度不大，仅有科大讯飞、捷通华声、百度、腾讯等有一定实力但为数不多的企业参与研发和推广，产品语种单一、缺乏数据支撑，缺少市场驱动力、推广力度不够。同时，企业的商业特点也决定了不同公司间的研究成果与信息资源很难得到共享与整合，企业之间的竞争与利益关系决定了各家企业之间必然是要各自为战。

同时，还存在有些地区和部门在思想上对民族语文信息化工作的重要性认识不到位，在平时工作中还存在着“等靠要”思维，在人员配备和经费投入上明显不足等问题。

（二）工作起步晚，基础薄弱，应用范围狭窄，人才稀缺

工作起步晚，基础薄弱。由于民族语文市场化程度低，投入大产出少，且互联网多以免费方式为用户服务，民族语文的语音识别和合成方面许多技术点没有解决，所以与汉语文的信息化过程相比，民族语文的信息化过程一直都是处于追赶、跟上的状态。就目前的民族语文信息化智能化来看，其工作起步晚、基础薄弱、空白点多。据了解，在2015年之前市面上还没有一款民族语文语音识别和语音合成产品可用，直至中国民族语文翻译局于2015年建立国内首个多语种智能翻译机房，并在最近几年陆续完成了蒙古、藏、维吾尔、哈萨克、朝鲜、彝、壮7个语种的智能语音翻译系统、民族语文神经网络机器翻译软件、多语种语言识别翻译系统等智能翻译软件的研发后，我国的民族语文信息化智能化才能说勉强跟得上汉语文的信息化智能化步伐。

软件应用范围狭窄。目前，研发成果推广力度不够，应用范围狭窄，影响了软件的完善和更新换代。一款新生软件能否顺利发展成熟，最主要是靠市场来检验和受众来评判。只有加大推广力度和扩大使用范围，在使用过程发现问题、解决问题，进而才能完善软件。有关民族语文智能翻译软件在民族地区的推广力度不够，应用范围不够广泛。目前除了维吾尔、蒙古、藏这几个语种的智能翻译系统在新疆、内蒙古、西藏、青海等省区，在国家安全和社会公共服务方面得到较好应用外，在双语教育教学和涉及民生等经济公共服务方面还需加大力度推广和应用，才能更好促进软件的完善，加快相关领域的信息化智能化步伐。

研发人员稀缺。目前，在有关科研机构、高校、企业里从事民族语文智能翻译软件研发的人员中，既懂相关民族语文又掌握了相关软件技术的研发人员非常稀缺。从国家民委有关部门了解到，各民族院校在现有专业设置中开设有民族语文信息化专业；因为要求学生既学习某一种民族语文又学习计算机相关知识的，故在本科阶段没有，研究生阶段极少。据调查了解，这一方面的硕博研究生毕业后，因对口单位少、福利待遇低，绝大部分进入了企业从事与民族语文信息化工作无关的研发工作。其它有关科研机构，在人才队伍的质量与规模方面，也与其所从事的科研工作强度和工作重要性方面不相匹配。

（三）我国民族语文智能翻译软件研发主要技术难点

规范化标准化是信息化智能化的基础，标准化是推动信息化进步的基础，也是信息系统有效运行的保证。没有相关的标准作为基础和保障，少数民族语文信息化的发展就很难真正实现。目前国内很多部门和民族地区都在加强信息化建设。但模型的创建、信息的采集和资源的开发缺乏统一标准，重复建设，造成信息资源的浪费和闲置。在人工智能和大数据时代，在深度网络神经算法开源的情况下，研发一款高效实用的民族语文软件产品，需要有大量规范、准确的数据作为支撑。规范化标准化是研发民族语文智能翻译软件的基础性工程。

在深度网络神经算法开源的情况下，民族语文智能翻译软件研发的难点与算法及语言本身无关，难在“文本分析”。以智能语音翻译软件为例，基于HMM/RNN的语音合成技术框架、文本分析技术框架和语音数据库这3个模块的成熟才能构建少数民族语言语音合成系统。经过优化后的系统，能够完整提取语音结构和句法结构，能够形成针对黏着语的文本分析技术方案，并对文本、语音进行双模态建模。目前，数据充分的语言，基本可以直接采用技术软件进行分析，但资源受限语言仍然需要使用传统的技术路线，需要参考专家知识。

具体地讲，民族语文智能翻译软件研发的技术难点主要有5个方面：一是建立平行语料库。这个工作包括段段对照、句句对齐、基础词汇校对等步骤和内容。目前靠人海战术来完成，效率低。拥有内容丰富、文本正确的大数据是增加深度学习训练和提高译文准确率的基础。以壮文智能翻译软件为例，截至2022年3月，壮文语料库大约有15万个词汇（词条）、40万个句子，大部分内容为政论性文章和法律法规，涉及文学和科技方面的内容比较少。因此政论性文章和法律法规翻译准确率可以达到80%以上，但文学和科技类翻译准确率就比较差。二是新词术语的规范化。新词术语的翻译和接受有个过程，一开始可能会有多种译法，个别译法甚至存在打架的现象，这一定程度上造成了规范化的滞后，直接影响了翻译的准确率。三是建立分词规则和分词库。分词是文本处理的基础步骤，也是人机自然语言交互的基础模块。智能翻译系统的良好性能需建立在科学系统的分词库基础之上，分词的好坏直接影响到翻译效果。蒙古文、维吾尔文、哈萨克文、朝鲜文、壮文与英文相似，不需要构建专门的分词库，但藏文、彝文则需要建立较大规模的分词库。四是处理民族语方言词问题。民族地区地域广阔，同一语种的方言语音差异很大，处理得好坏直接影响语音输出质量。为解决同一语种的方言语音差异问题，需对不同地域、不同方言人群进行大样本人工采集语音数据，平均采集2000小时以上。将后台应用软件收集的语音数据精细标注，再反复进行模型训练。五是解决编码缺陷问题。这个问题主要存在于个别文种，不规范的编码规则容易造成语音与字符在进行匹配转换时出现混乱和差错。

三、对我国民族语文智能翻译软件研发的几点建议

民族语文工作是民族工作的重要内容，涉及我国的反分裂、反恐怖、反渗透，涉及我国构建多语和谐的语言生活、繁荣发展少数民族文化、促进民族地区经济社会发展，事关公共服务。民族语文信息化智能化是落实国家民族语文政策的技术途径。为此，要紧跟时代发展步伐，做好民族语文信息化智能化工作。鉴于民族语文智能翻译软件在维护国家安全、促进民族团结、铸牢中华民族共同体意识等方面的重要性，提出以下几点建议。

（一）建立多方合作研发和推广机制，有效促进区域合作和跨界合作

目前，研发一款民族语文智能翻译软件需要基础理论支持、前沿技术引领和文本大数据支撑。为此，需要有关科研院所、高校以及企业进行全方位合作和深度合作，合作过程明确中心、突出重点，有效进行分工协作。国家民委委属高校和民族地区的地方性大学结合自身情况进行重点研究，突出地域特点和学科优势，在立项阶段与企业建立项目合作机制，加大技术难点攻关，及时转化科研成果。民族地区的各级民族语文行政管理部门加强对本地民族语文信息化工作的指导和支持，建立健全新闻出版、广播电视、教育等民族语文工作部门日常业务交流合作机制，加强民族语文语料库建设。

（二）加大力度培养民族语文信息化智能化人才队伍

人才是所有行业的基础，人才兴则行业兴，人才强则行业强。为此，需要加大力度培养民族语文信息化智能化人才队伍。一是深化人才体制机制改革。鼓励引导政府部门、重点企业完善信息通信业人才培养机制，支持民族语文信息化领域的人才培养。二是深化人才教育改革。通过行政主管部门协调和支持，在有关科研院所和高校建立学科人才培养基地，培养既懂计算机知识又懂少数民族语文的复合型、应用型人才。

（三）切实发挥民族语文智能翻译软件在推广国家通用语言文字工作中的特殊作用

党的十九届五中全会指出，提高民族地区教育质量和水平，加大国家通用语言文字推广力度。2021年政府工作报告中明确指出，加大国家通用语言文字推广力度。大力推广国家通用语言文字，是党和国家站在实现中华民族伟大复兴、推动民族地区同步迈进社会主义现代化强国新征程的战略部署。民族语文智能翻译软件作为一种即时、交互、便捷、高效的软件，在推广国家通用语言文字工作中具有特殊作用。首先，是加大国家通用语言文字社会使用的技术载体。当前，随着智能手机等新型移动终端的广泛使用，民族地区干部群众主要通过智能手机等新型移动终端上网来获取更多的生产生活资源和进行日常信息交往。其次，是加强国家通用语言文字学校教育的有益补充。在推进国家通用语言文字教育过程中，民族语文智能翻译软件作为一种信息技术智能化数字化的载体和手段，可以通过双语教育教学智能化在这方面进行补充和加强。最后，是科学保护各少数民族语言文字的创新载体和技术手段。语言文字是文化传承的载体，民族语文承载着少数民族的传统文化，是中华文化的重要组成部分，为此要坚持加大国家通用语言文字推广力度与保障少数民族语言文化传承并行不悖。民族语文智能翻译软件在智能化数字化的现代信息社会里，有利于推动和实现科学保护各少数民族语言文字，传承和发展各少数民族传统文化，保持中华文化多元一体的文化优势。

四、结束语

习近平总书记在第五次中央民族工作会议上强调，要推广普及国家通用语言文字，科学保护各民族语言文字，尊重和保障少数民族语言文字学习和使用。民族语文是少数民族交际和思维的工具，也是中华民族的共同文化财富。新征程上，我们要在以习近平同志为核心的党中央坚强领导下，准确把握和全面贯彻习近平关于加强和改进民族工作的重要思想，牢牢把握住新时代新要求，用新思想来指引新战略，做好民族语文信息化工作；本着“民族语文智能翻译软件助力国家民族工作”的原则，进一步研发适应民族地区需要的民族语文智能翻译软件产品，不断提高民族语文信息化智能化自主创新能力，让民族语文信息化智能化成果更多应用到民族地区和少数民族干部群众中，为铸牢中华民族共同体意识、助力民族地区乡村振兴发挥其应有作用。

参考文献

[1] 李宇明.2011.《中国少数民族语言文字规范化信息化报告》，北京：民族出版社.

[2] 李宇明. 2021.《新世纪20年的中国语言规划》，《北华大学学报（社会科学版）》第22卷第1期.

[3] 李旭练、樊焓希. 2019.《以民族语文信息化成果助力民族地区脱贫攻坚》，《中国民族报（理论周刊）》第5版.