【创新洞见】

“创新洞见”栏目旨在呈现与传播全球交叉研究领域的前沿思考与实践成果，下设“学术文章”和“青瞻未来”子栏目。“学术文章”汇集国内外知名的交叉研究专家、校内相关职能部门负责人等对交叉学科建设、科研创新机制、未来科技发展趋势的见解与建议，推动交叉研究理念在校内外的传播与落地。“青瞻未来”重点展现学生学者在前沿领域的思考与探索实践，为交叉创新注入青年视角与新生力量。

学术文章

毕晓君：民族古籍文献智能分析与机器翻译研究

民族古籍文献是中华民族多元一体文明格局的重要载体，保存着各民族在语言文字、社会历史、宗教哲学、天文地理等方面的知识与经验。东巴文、古彝文、水书、满文、回鹘文等文献，既具有重要的文化遗产价值，也为文字起源发展、书写演化和各民族交往交流交融的研究提供了珍贵材料。

然而，由于专业人才稀缺、文献数量庞大，加之部分文字已不再广泛使用，相关释读人才日益稀缺，大量民族古籍仍处于难以释读、难以整理、难以利用的状态。如何借助人工智能技术推动民族古籍的识别、修复、释读和翻译，成为数字人文与人工智能交叉研究中的重要问题。在此背景下，中央民族大学毕晓君研究团队在“民族语言智能分析与安全治理”教育部重点实验室的平台支撑下，开展了相关研究，并取得了较好的成果。

一、民族古籍的研究价值与现实难题

民族古籍文献的整理与释读，首先面临着保存状态与文字形态的双重挑战。一方面，许多古籍年代久远，存在文字残缺、图像模糊、页面破损等问题，亟需通过图像处理技术进行修复；另一方面，民族古籍中存在大量高相似度字符，部分字符只有局部细节差异，意义却完全不同，这对机器的细粒度特征提取能力提出了极高要求。

更深层的难题在于机器翻译。以东巴文等古老文字为例，即使能够识别单个字符，也并不意味着能够准确理解整句话的含义。许多民族古籍并非逐字翻译型语言，文字组合、上下文省略和非线性排列都会增加整体语义理解的难度。因此，民族古籍智能分析不能停留在“认字”层面，必须进一步解决语义理解和机器翻译问题。

二、数据基础的系统构建

将人工智能应用于民族古籍研究，首先需要建立可靠的数据基础。围绕单字识别，研究团队构建了东巴文、古彝文、水书、满文等多种文字的数据集。为了提升识别模型的泛化能力，每个字符都尽量收集多种书写样本，使模型能够适应不同书写风格和字形变化。目前，东巴文和水书的单字数据集已

面向社会开放使用，为后续研究提供了基础支撑。

围绕机器翻译，团队进一步构建了多民族古籍平行语料库。以东巴文为例，研究团队基于已出版的东巴古籍相关书籍，完成段落级、句子级和单字级的多层标注，并由此建立可用于机器翻译和内容挖掘的语料资源。除东巴文外，古彝文、水书、满文等语料标注也在推进之中，为多民族古籍智能释读奠定了关键基础。

三、图像修复与细粒度识别

在古籍图像修复方面，团队并未简单套用通用图像修复方法，而是结合文字系统自身特点开展研究。对于残缺较小的文字，现有图像修复方法尚可发挥作用；但当文字残缺较大时，单纯依靠图像填补往往难以恢复真实字形。为此，团队利用已构建的字库作为参照，先对残缺字进行粗修复，再从完整字库中匹配候选字，并结合上下文信息进一步判断，从而提高修复结果的可信度。

在单字识别方面，研究重点集中于高相似度字符识别。针对东巴文中许多字符整体结构相近，仅局部笔画或细节上存在差异的特点，团队设计了更加关注局部特征的识别模型。该模型在兼顾全局信息的同时强化细粒度差异提取，成功将东巴文单字识别率提升至较高水平。后续轻量化模型的开发，也将会为移动端部署和实际应用提供了可能。

四、词组挖掘与文字系统研究

东巴文长期被认为是一种弱语法、弱语义的文字系统，缺乏明确句式、句型和词组，这也是其机器翻译难度较大的重要原因。研究团队在实践中提出一个新的问题：东巴文是否真的没有词组，还是传统人工研究难以从海量文献中系统发现？

基于此，团队通过单字识别、上下文关联分析、跨模态图文对齐和人工校对等步骤，从东巴古籍中挖掘出一批稳定出现、语义一致的词组。例如，某些两个字符组合在古籍中反复出现，并稳定表达“东方”“灵魂”等整体含义，而非两个单字意义的简单相加。

这一发现不仅提升了机器翻译效果，更为理解东巴文的文字系统属性提供了重要支持，证明其并非单纯的图画记事符号，而是具有较为明确的文字组织规则。由此可见，人工智能不仅可以作为古籍整理工具，也可能为人类文字起源和书写演化研究提供新的证据。

五、机器翻译的大模型探索

在机器翻译方面，东巴文的难点集中体现在三个方面：结构规律较少，单字语义丰富，且上下文省略现象突出。由于部分主语或语义对象可能在相隔数句之后才再次出现，单纯以句子为单位进行翻译容易丢失上下文信息。为此，团队采用段落级翻译思路，并通过句子随机组合的方式增强语料，缓解低资源语言语料不足的问题。

在大模型应用方面，团队以多模态大模型为基础，结合东巴文图像与中文释义信息进行训练。实验表明，直接使用现有大模型效果有限，但在引入段落级语义增强、句子组合策略和分隔符设计后，翻译效果明显提升。研究团队也强调，民族古籍机器翻译并不以完全替代语言学家为目标，而是为专家提供辅助工具，助力其在机器初译基础上开展校订、考证和深入研究，从而提升古籍释读效率。

六、应用拓展与交叉启示

多民族古籍智能分析的应用价值远不止于古籍保护。一方面，古籍机器翻译可帮助研究者处理大量尚未释读的文献，使散落于国内外图书馆和收藏机构的民族古籍重新进入研究视野，推动语言学、历史学、民族学和人类学研究。另一方面，相关技术也可迁移至现实语言服务场景。例如，团队与中

国民族语文翻译中心（局）合作，探索基于大模型的民族语文机器翻译系统，并在朝鲜语、壮语、彝语等语种的实际翻译场景中提升了工作效率。

从研究范式来看，这项工作体现了人工智能与人文研究的深度交叉。人工智能提供了图像修复、文字识别、语料标注、词组挖掘和机器翻译等技术手段；人文研究则提供了问题意识、文献依据和解释框架。二者的结合，不是简单地以技术替代专家判断，而是在尊重专业知识的基础上，用计算方法扩大研究材料、提高处理效率，并发现传统人工研究难以观察到的结构性规律。

七、面向未来的研究方向

未来，多民族古籍智能分析仍有广阔的空间。随着数据集和语料库的不断完善，研究可从文字识别和机器翻译进一步走向内容理解、知识挖掘与文明研究。例如，在完成大规模释读后，研究者可进一步分析古籍中关于仪式、医学、天文、地理、战争和社会组织的记载，探索不同民族文化知识的形成、传承与演变。

总体而言，多民族古籍文献智能分析与机器翻译，不仅是人工智能技术在低资源语言场景中的应用探索，也是数字时代保护、激活和阐释中华民族文化遗产的重要路径，进而有助于铸牢中华民族共同体意识。同时，也使沉睡的古籍重新进入可读、可研、可用的知识体系，为文理交叉回应真实学术问题提供了有益启示。

杨东：构建中国自主的数字法学知识体系

一、构建中国自主数字法学知识体系的时代意义

构建中国自主的数字法学知识体系，是深入贯彻习近平法治思想，推进法治中国建设、服务中国式现代化的重大理论与实践任务，是新时代数字革命背景下法学研究的核心重点。该体系是中国特色哲学社会科学知识体系的关键组成部分，作为数字文明时代全新的法学形态，立足我国数字法治建设实际，突破传统法学的研究边界，搭建起专属的理论范式、学科范畴与话语体系，有效填补了数字领域的法理研究空白。同时，其以学术自主、学科自立、话语自信为核心，与经济学、社会学等领域的自主知识体系协同联动，极大丰富了我国哲学社会科学的内涵，提升了原创性、前沿性知识体系的完整性。

在法治轨道上建设社会主义现代化国家的时代要求下，该体系是数字领域落实中国式现代化的重要支撑。其以辩证唯物主义和唯物史观为指导，深度结合我国数字经济、数字社会、数字政府建设实践，将马克思主义法治方法论与数据治理、算法规制、平台治理等前沿法治工作融合，突破西方技术实证理论的局限。通过统筹学科、学术、话语三大体系建设，规范法学研究与学科发展方向，实现数字法学体系化、自主化发展。此外，该体系坚持立德树人，搭建交叉学科培育体系，对接算法伦理、数据合规、网络安全等实务领域，有效破解数字法治人才供需矛盾。同时聚焦数据安全、平台监管等治理难题，推动法治与数字技术融合，统筹发展与安全、国内与涉外法治，全面提升国家治理现代化水平。

二、中国自主数字法学知识体系的历史逻辑

我国法学知识体系的发展，是一部从全